Andy Li

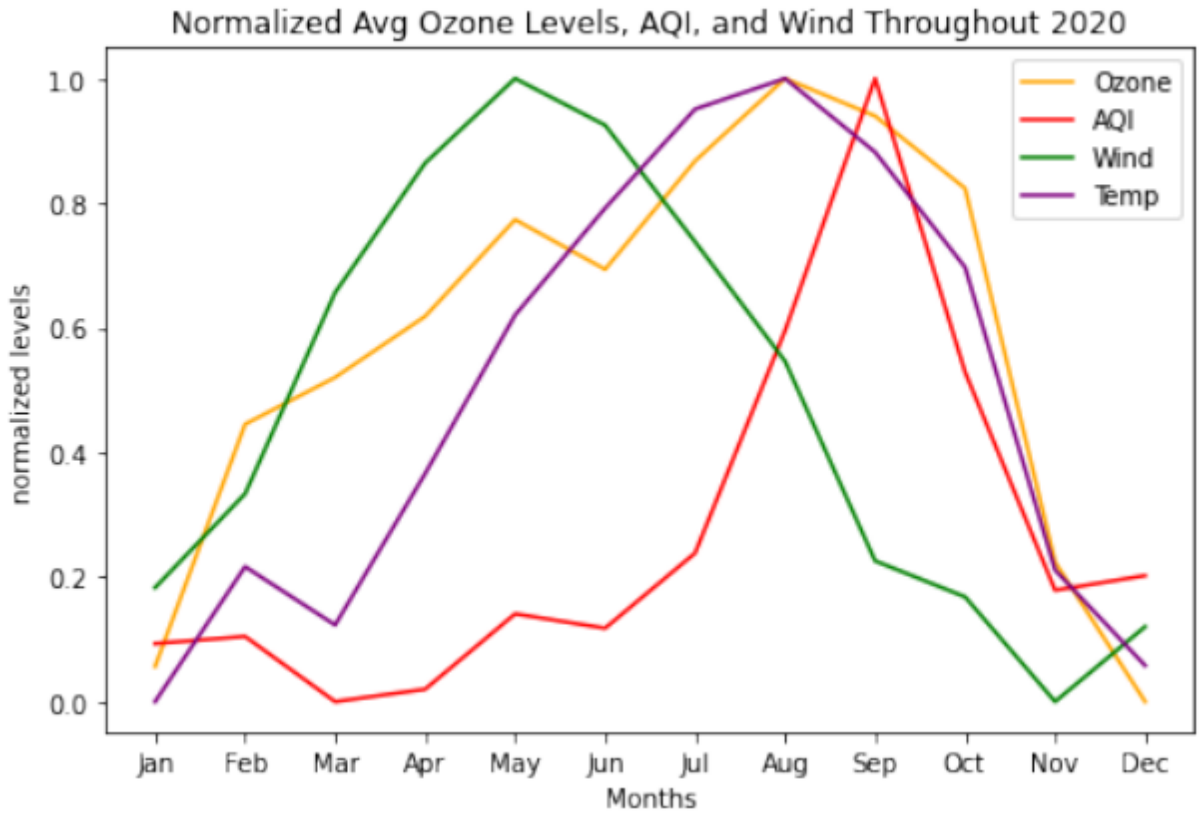
Giani Kurniawan

Patrick Pan

Final Open Ended Modeling Report

I. Open-Ended EDA:

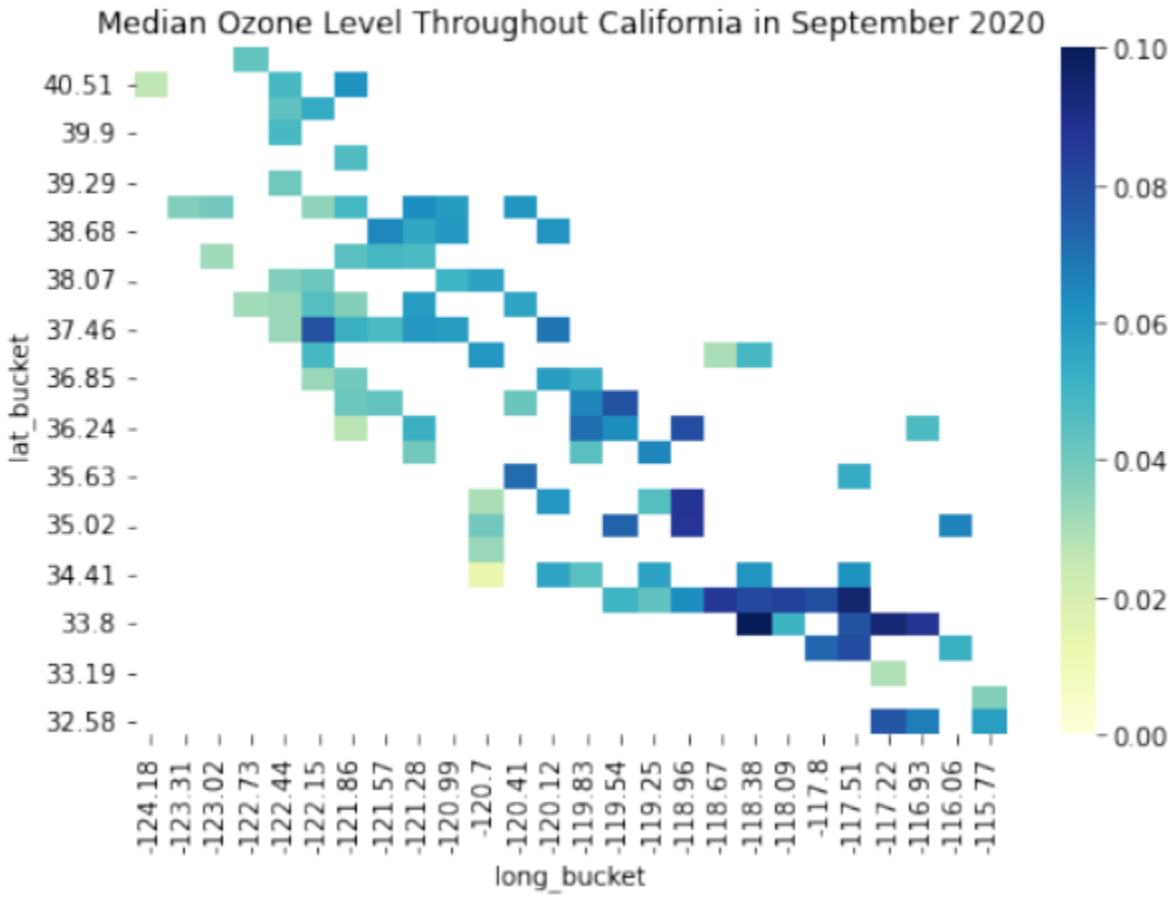
i) Plot 1



We were initially curious about how the ozone levels and wind speed in 2020 (especially in September) looked in comparison to the average AQI in 2020. To build this visualization, we merged the daily_ozone DataFrame and daily_wind DataFrame onto our existing epa_merged_CA DataFrame, which already has the latitudes and longitudes of different locations. With the data we need, we first wanted to visualize the average ozone levels and

average wind speeds throughout the months of 2020. By graphing a line plot of the normalized average ozone, normalized average wind speed, and normalized average AQI over the months of 2020 in California, we can compare the three overtime. As can be seen on the first graph, the levels for normalized average ozone levels and normalized average wind speeds have already been increasing since January, but for AQI it lags behind and doesn't really increase until July. Ozone and AQI increase a good amount until September, and then both start to drop off and decrease until December. However, starting July, wind temperatures start dropping rapidly before starting to increase in late September. Although there is a lag in the AQI, all three show a positive correlation until July, when wildfire season starts. After July, ozone and AQI are still positively correlated, but wind speed becomes negatively correlated with ozone and AQI. This makes sense since wildfire season will increase AQI and air pollutants such as ozone levels, but should decrease the wind speed due to the thick atmosphere created by wildfires.

ii) Plot 2



The second visual we created to understand the relationship between ozone and AQI is by creating a heatmap for ozone. We tried creating a heatmap for wind speed but were unable to overlay two heatmaps on top of one another. The wind speed heatmap also did not contain enough variation in locations, often being just a few blobs which would not add much value. To create the ozone heatmap similar to what we did in 6a and 6b, we built a function that creates a pivot table of the coordinate buckets and the median ozone level values. From what we can see from the ozone heatmap of California in September 2020, it is very similar to the AQI heat map of California in September 2020. Areas with high median AQI because of the wildfires such as LA and Fresno county have dark spots indicating high levels of ozone. This supports our idea

that ozone levels and AQI have a positive correlation. This intuitively makes sense as ground-level ozone is a hazardous air pollutant that may be the result of industrial production and wildfires. Due to the large amounts of California wildfires in September 2020 (especially in LA and Fresno county) the ground-level ozone levels will definitely be very high. Through our analysis and data visualizations, we can see a clear correlation between ozone level and wind speed with AQI, especially during wildfire season. These correlations indicate that ozone level and wind speed may be good features to use in predicting AQI.

Something we may want to look further into may be certain elevation and also their density. Wildfires usually occur in rural areas with high elevations so we can look deeper into this question and explore further in our hypothesis/model. These additional features can be acquired through our AQI dataset and also external datasets. These features may end up being good predictive features of AQI levels/categories.

iii) Open-ended questions

What causes AQI to increase? How quickly is climate change affecting AQI levels? How does elevation play into this? How much does AQI vary between different locations? What can we conclude about the relationship between seasonality and locations with respect to AQI? How exactly does global warming affect AQI levels, if at all? Are there any trends or predictable patterns regarding wildfires? How well does our analysis carry over to states or areas other than California? What are some other features that could be useful predictors of AQI? Since global warming causes both extreme heat and extreme cold, how does the extreme cold element affect

the AQI levels? Are there other elements of poor air quality that are caused by global warming but are not captured by AQI?

II. Problem

i) New hypothesis: Although most high-elevation forests in California have not been subjected to fire suppression, human activities, and because trees at these high elevations are in wetter forests, higher-elevation forests are more prone to wildfires due to the increased effects of global warming (measured with particulate matter concentrations such as ozone). We hypothesize that there is a positive correlation between elevation and AQI levels in California due to the effects of global warming and thus the higher threat of wildfires. Thus, features relating to elevation and wildfires such as general location (county and site), exact location (latitude and longitude), day and month, elevation, ozone level, wind speed, temperature, wildfire data (avg acres burned), land use description, AADT, and county population density can all be good predictive features of AQI levels/categories.

We decided to change our hypothesis because we wanted to look into elevation, a feature that might be overlooked when analyzing correlation with AQI levels. Since our initial EDA was on ozone levels, wind speed, and temperature, all features relating to a relative location's elevation, we included these as features in our model as well. Elevation may play a big part as an important feature in predicting AQI levels that we were previously unaware of. We don't usually see wildfires in low elevation areas like San Francisco because it is urbanized with a high population density. Diving into elevation, along with its related features like population density, location, wildfire data, etc was much more interesting compared to our initial hypothesis which

was just looking at the correlation between particulate matter concentrations such as ozone and AQI levels.

Although our initial hypothesis regarding lagged ozone levels and AQI is certainly very interesting, we found our new one to be more incisive and nuanced. We were very much inspired by the climate change guest lectures in this class and wanted to explore the dataset on a deeper level.

III. Answer

From our analysis and modeling, we confirm our hypothesis. This analysis and modeling has definitely helped us successfully answer our research question and evaluate our hypothesis. We hypothesized that there is a positive correlation between elevation and AQI levels in California due to the effects of global warming and thus higher threat of wildfires. By adding additional features from the AQI dataset and from external datasets to make our model more refined towards our hypothesis, we saw huge improvements in the new model's performance to predict AQI levels/categories. This improvement shows that by refining our model towards our hypothesis, our validation of our hypothesis also increases. In summary, our hypothesis seems to be correct due to the improvements in accuracy we see in our model when we refine it towards the hypothesis. Thus, features relating to elevation and wildfires such as: general location (county and site), exact location (latitude and longitude), day and month, elevation, ozone level, wind speed, temperature, wildfire data (avg acres burned), land use, AADT, and county population density can all be good predictive features of AQI levels/categories.

IV. Modeling

Our baseline model uses just three features relating to the elevation that can affect AQI. These three features are elevation, temperature, and County Code. These three baseline features were chosen because it serves as a good foundation relating to our hypothesis.

Our baseline and improved final model will be a Random Forest Classifier model because we are using multiple features to predict AQI categories (categorical) and we want to avoid overfitting. Another reason why we chose a random forest model was because we tried linear regression initially for the guided modeling part and it underfit horribly compared to using a decision tree or random forest model. Even with all the final features we added, we cannot get the binary error for the linear regression model to be under 0.4 for the guided model. However, by switching to a decision tree for the guided modeling part, with just the three features of ozone, wind speed, temperature, and AADT, we were able to get the binary error down to a consistent 0.2. The big reason why we decided to switch to a random forest for our final improved model was to decrease the binary error even more and also to avoid overfitting since we will be adding a lot more internal and external features.

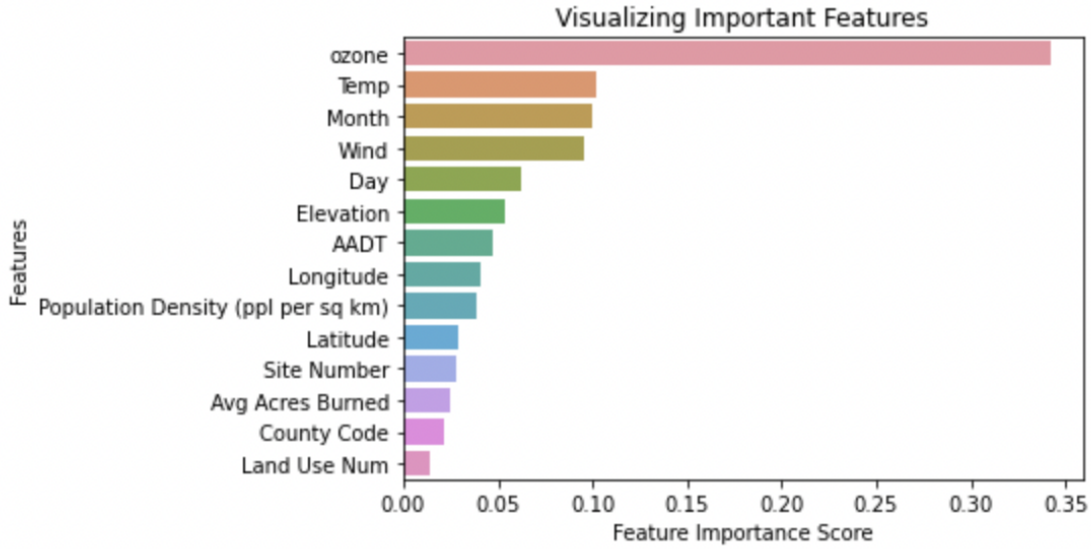
The inputs are ‘features’ which are in the form of a DataFrame and represent our X matrix and ‘targets’ a NumPy array that corresponds to the AQI category for each of the data points in features and is essentially our vector of outputs.

Our improved final model required adding additional features from the AQI dataset, as well as the three external datasets including traffic, population density, and wildfires (cited at the end of report). The baseline model was not fitting properly because there is a high bias. In order to achieve better pinpoint locations of certain elevation, we needed more defined features such as

site number, latitude and longitude, month and day, ozone levels, wind speed, temperature, average acres burned from wildfires, land use, AADT, and population density. The final set of features we used in our improved final model were general location (county and site), exact location (latitude and longitude), day and month, elevation, ozone level, wind speed, temperature, wildfire data (average acres burned), land use, AADT, and county population density. These are all good features because they relate to elevation and wildfires and have some sort of correlation. We chose location because certain locations have higher and lower elevations. We also chose day and month because wildfires usually occur during the last few months of the year which is considered ‘wildfire season.’ To get a better idea of whether elevations are more prone to wildfires, we looked at the average acres burned by wildfires for each county and thus differing elevations. Ozone level, wind speed, and temperature are all features related to wildfires and can differ depending on the elevation of the location. We also chose density features such as land use description, AADT, and population density because higher elevations tend to be less densely populated and we wanted to dive deeper into this analysis.

V. Model Evaluation and Analysis

Visualization 1:

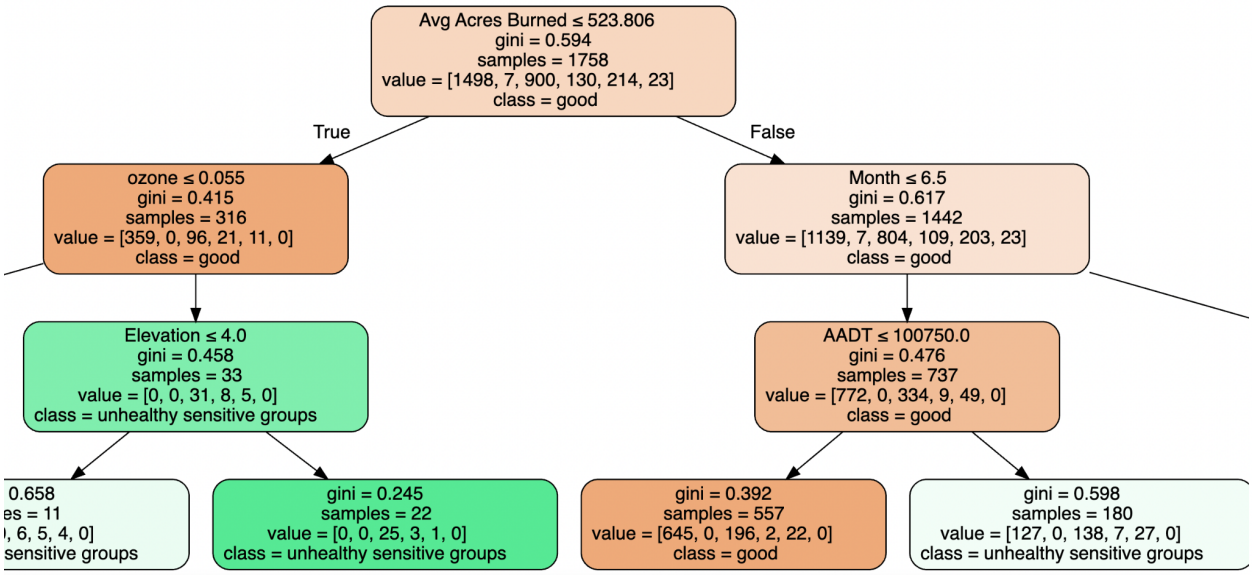


This first visualization is an analysis on the features of our model. By using the RandomForestClassifier's built in function feature_importances_, we are able to construct a bar graph to visually rank the importance of each feature in our model. As we can see, ozone has the highest importance in our model. This is because ozone is a particulate matter that is directly related to AQI and thus should have a very high correlation with AQI. Thus, ozone will make a great feature for predicting AQI and is therefore ranked high in terms of importance. Although all features have some sort of feature importance in our model, some features rank very low and may not affect the model as much. Two of our external features, AADT and population density, have a good importance score and definitely improves the accuracy of our model. By adding just AADT and population density as additional features from our baseline model to our improved model, we saw a consistent improvement which lowered the binary error by about 0.08 (improved accuracy of 0.08). However, some features such as land use, county code, and

external feature average acres burned by wildfires have minimal effects on improving our model. By adding average acres burned by wildfires as an additional feature to our baseline model, we saw a decrease of our binary error of about 0.002 consistently. This was surprising because we thought wildfire data would be a good feature relating to elevation that can predict AQI since higher elevations are usually where more wildfires occur. This may just be an issue with the data since we used average acres burned from wildfires (averaging from years 2013-2019) for California counties and not 2020 wildfire data. Unfortunately we cannot find any 2020 wildfire data that has location/counties that we can merge with our current dataset.

Visualization 2:

A Small Part of One Decision Tree from our Random Forest model



Another visualization to show the success of our model is a picture of one branch of a decision tree in our random forest model. This visualization essentially shows how a decision tree works. AQI categories, which are represented by the different groups of colors, and categorized by going down the decision tree. It is simply a tree of questions that must be answered in sequence

to yield a predicted classification. By going down each branch, we get closer to our classification by looking at conditions of different features. This visualization is important as it shows how our model is working and what it is doing to predict AQI categories using our chosen features.

A way that we evaluated the performance of our improved final model was by comparing its validation errors (binary error and cv error) with that of the baseline model and seeing if there was a decrease in errors. A decrease in binary and cv error means that our model has improved by lowering its validation errors and increasing its accuracy. By adding additional features from our AQI dataset and features from 3 external datasets (traffic, population density, wildfires) we were able to drastically improve our model. Our hypothesis was that there is a positive correlation between elevation and AQI levels in California due to the effects of global warming and thus higher threat of wildfires. Our initial baseline model which only used 3 features (temperature, county, and elevation) was not enough and did not fit well due to high bias. To better pinpoint locations of certain elevations to support our hypothesis and fit our model, we added additional features such as site number, lat/long, month and day, ozone levels, wind speed, temperature, avg acres burned from wildfires, land use, AADT, and population density. With our improved model, we got a binary error of around 0.11 which is a lot lower than our baseline model which had a binary error of around 0.36. Utilizing our mean squared absolute distance error loss function from Q9, our improved model resulted in a cv error on the validation set of around 0.16 which is also a lot lower than the cv error of the baseline model which was around 0.6. By comparing the validation errors (binary and cv), we can clearly see that our new model has improved from our baseline model by a good margin.

VI. Model Improvement

Improvement 1: We saw that the baseline with 3 features model has a binary error of around 0.36 or accuracy of 0.64 and a cv error of 0.6. We were not satisfied and believed our model was still underfitting to our standards. We wanted to get the binary error of our improved final model to be around 0.1-0.15. To improve our model by getting a lower binary error and thus higher accuracy, we included additional features to better pinpoint certain elevations to better support our hypothesis. To do this, we will use additional features from our AQI dataset and external datasets such as the Traffic dataset and California WildFires dataset that have a correlation to AQI levels. To implement new qualitative features into our random forest classifier, we will map numerical values to each qualitative variable. To see if our improved model has improved inaccuracy, we will see if the binary error decreased. After adding additional features from our AQI dataset and features from external datasets, our binary error decreased from 0.36 (baseline model) to around 0.11 (improved final model) and our cv error decreased from 0.6 to 0.16.

Improvement 2: Another improvement we made in our improved final model was that we increased the training-test split from 70-30 to 80-20. We believed that by allocating 10% more of the data to training, we can get an even more accurate model. When we initially used a 70-30 split, we got a binary error of around 0.12 consistently. With a 80-20 split instead, we were able to improve marginally about 0.01 to get a binary error of around 0.11 consistently. Since our dataset is large, we can use a 80-20 split to get an accurate model and also have enough data to test/validate on.

VII. Future Work

For our future work, we would like to further investigate the trends and importance that ozone plays towards affecting AQI, especially with relation to elevation. Throughout our EDA and model-construction, we consistently noticed that ozone levels were an incredibly important feature. Intuitively, this makes sense because ozone is a particulate matter that is directly related to AQI and thus should have a very high correlation with it. In fact, our original hypothesis involved studying how lagged ozone levels could be a predictor of AQI, before we pivoted towards focusing on elevation.

Interestingly, ozone is simultaneously a danger to human health but also necessary to protect it. Atmospheric ozone serves to protect our planet from the harmful effects of global warming by reflecting solar radiation. However, low-level ozone is a pollutant that harms human health. From our model and EDA, we can see that elevation is another important feature towards predicting the AQI levels and its changes. This caused us to consider whether places with greater elevation (closer to the atmosphere) could be affected differently by the ozone because then the ozone in that area could theoretically take on more properties of the benign atmospheric ozone and less of the malignant ground-level ozone. If ozone produced from wildfires leaves an area by dissipating into the atmosphere, then intuitively one could imagine that places of varying elevations could have different levels of ozone dissipation, causing them to be affected by wildfires to differing degrees.

We want to really focus on exploring that relationship in our future work. We would start by studying the localized effects that ozone has on various places (by County Code and elevation), and uncovering the correlation levels. In doing so, we could feel more confident that there is some definitive link between various geographical features such as elevation and ozone,

which is what we conjecture. Next, we would want to focus on producing visualizations showing the duration and magnitude of ozone level "spikes" following major wildfires around the state.

If our future work supports the hypothesis that AQI is affected by elevation mainly through the mechanism of changes via ozone levels, this could help emergency planners better prepare to deal with the impacts of climate change in a way specifically tailored to places of varying elevation. If our future work's data does not support that hypothesis, then we could potentially investigate what other mechanisms are driving how elevation plays a role in AQI changes.

VIII. Citations

External Dataset 1: Traffic dataset provided by course

External Dataset 2 (Annual County Population Density in California 2020) : https://covid19.census.gov/datasets/USCensus::average-household-size-and-population-density-county/explore?showTable=true

External Dataset 3 (California Wildfire Data (AVG Acres Burned for Each County): https://www.kaggle.com/ananthu017/california-wildfire-incidents-20132020