Data 512 Project Part 1 Reflection and Analysis

Author: Ashwin Naresh
Assigned City: Savannah, GA

Implementation Details

This project utilizes 2 different external sources. The first is the wildfire dataset, a massive GeoJson file which we parse to acquire our fire data. The other is the air quality api (AQI). To avoid having to repeatedly use these resources, I chose to create intermediate csv files to store the data so that I would only need to undertake the data collection process a single time. All subsequent tasks can be performed using just the csv files, which saves a lot of time.

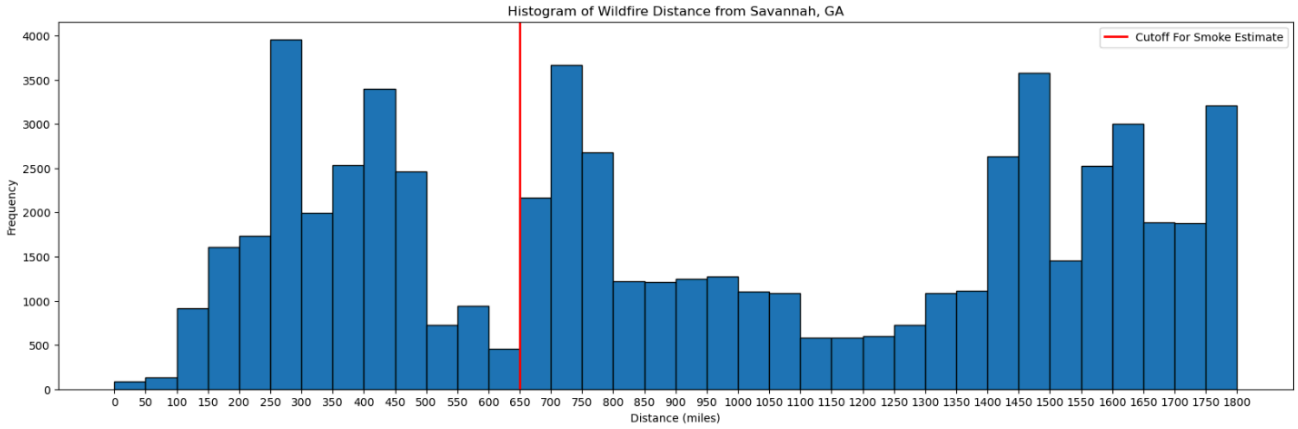
In this project, I also created a custom Smoke Estimate metric which aims to capture the amount of smoke that my assigned city receives from fires each year. For this smoke estimate, we are restricting our data set to fires that happened no earlier than 1961 and fires that happened no further than 650 miles away from the city. This custom smoke estimate is simple: Smoke Estimate = (Acres Burned) / (Distance from city)^2. Smoke diffuses as a function of the square of the distance and the number of burning acres is proportional to the smoke generated. This gives us a number (which we will treat as dimensionless) that represents the cumulative amount of smoke that the city receives each year.

Lastly, we compare our custom smoke estimate with existing air quality metrics, notably particulate AQI. This API lets us gather daily air quality data over a period of time but this data was only available starting from 1986.

After collecting the custom metric and the AQI data, I used the Prophet library to create a prediction model. Prophet, developed by Facebook, is well suited for time series predictions due to its ability to effectively capture seasonal and periodic variation. Furthermore, this library is extremely easy and quick to set up, making it ideal for a project like this.

Visualization Breakdown

Visualization 1:

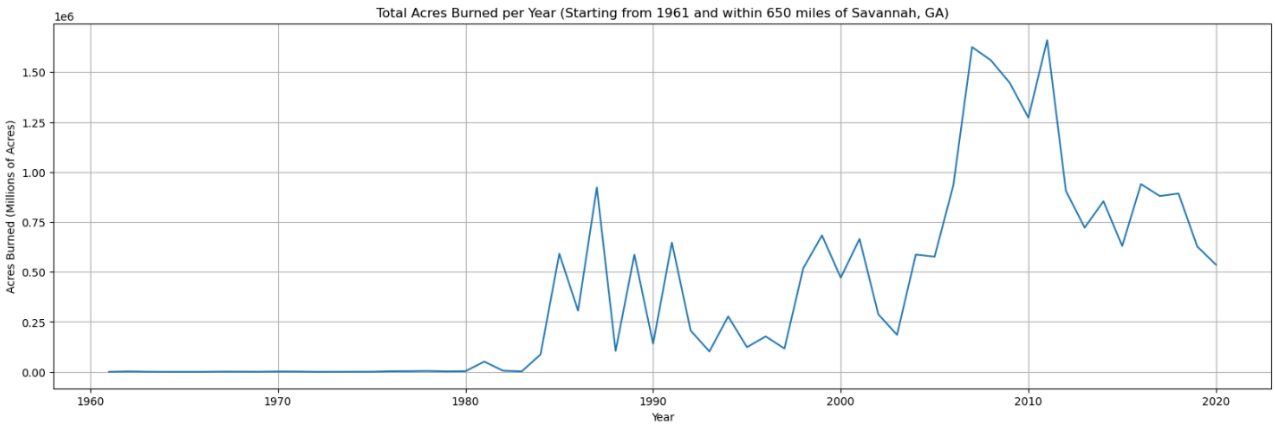


This histogram shows the number of wildfires that occurred at various distances from Savannah, GA. The red line at the 650 mile mark indicates the cutoff for the data that was used to make smoke estimates using the custom metric. This histogram considers fires up to 1800 miles away from the city and all axes are labeled for easy understanding. The x-axis represents bins for distance from the city and the y-axis represents frequency.

What we see is that there are more data points at the ends of the histogram. This probably happens because the middle of the country is likely devoid of major fires. As a result, the East Coast fires populate the left of the histogram and the West Coast fires populate the right of the histogram.

To create this visualization, I parsed the GeoJSON file for fires within the distance and year criteria and created a data frame with the following columns: id, year, size, and distance. With this data frame, I used matplotlib to create a histogram of the distance column.

Visualization 2:

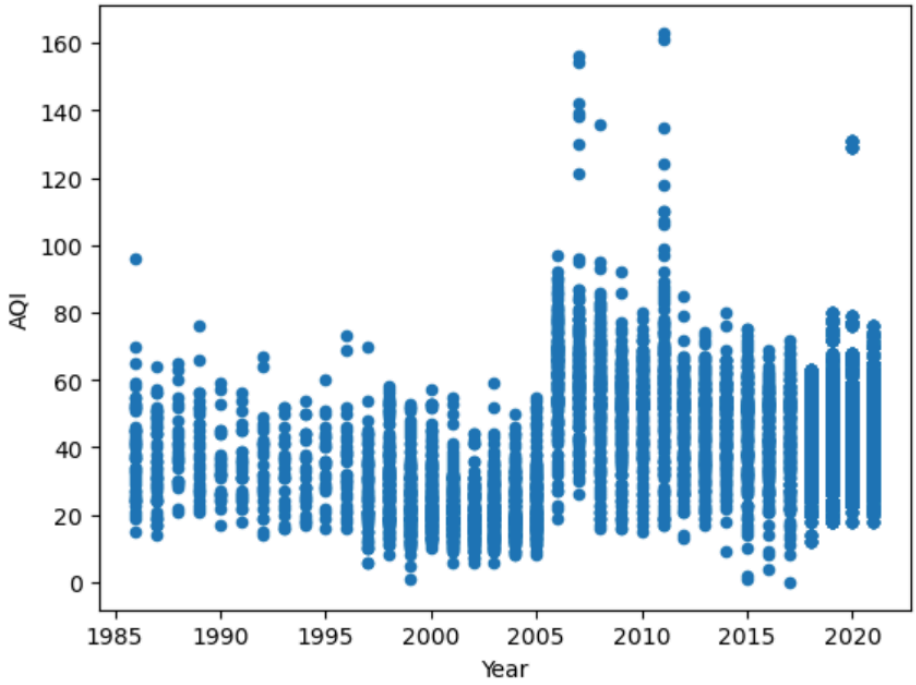


This time series chart shows the total number of acres burned over time. The data starts in 1961 and ends around 2020. The x-axis represents the year and the y-axis represents the total acres burned, in millions of acres. For this chart, we are restricting the data to fires that occur within the 650 mile distance to Savannah, GA. Time series charts are intuitive to understand because humans are good at following the data when the x-axis is time.

This graph shows that the total number of burning acres is rising over time. We are currently in a minor low point, but the overall trend is upwards and we should expect to see an increase in the acres burning if we had data up to the present.

To create this visualization, I took the data frame I created for the prior visualization, and filtered it to only include fires that occurred within 650 miles of the city. Then I grouped by the year, and aggregated the size column to get the total acres burned per year.

Visualization 3:

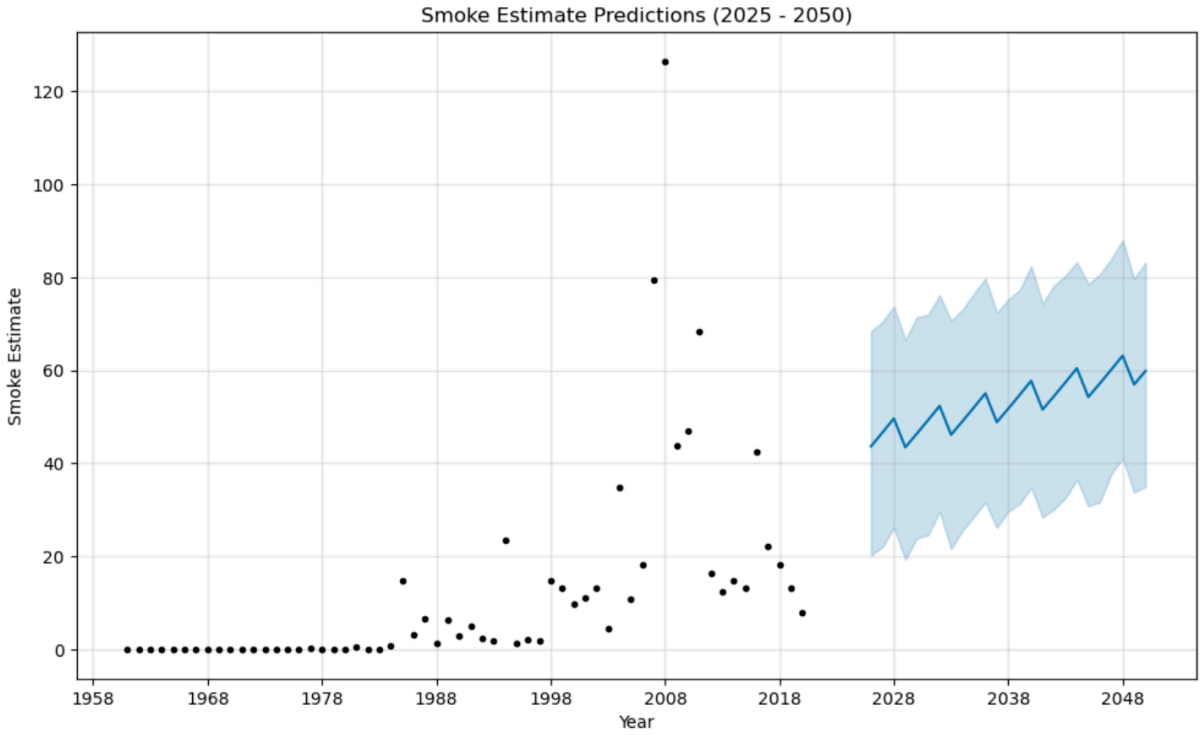


This visualization is a scatter plot of the Air Quality Index (AQI) estimates. This data was collected from the EPA API and contains data collected from multiple different collection stations. For this plot, we consider all AQI measurements that happened between May 1st and October 31st, which we will consider as the fire season. Furthermore, we were only able to get AQI data starting from 1986. The x-axis is the year and the y-axis is the AQI measurement.

The data appears to be sinusoidal in nature with the exception of the large spike between 2006 and 2012. There also seems to be a slight upwards drift in the mean which is reflected in the projections we create in a later visualization.

To create this visualization, I called the AQI api for each year from 1986 to 2021 and recorded the AQI estimate from each monitoring station. I loaded this data into a data frame with the following columns: year and AQI. Then I plotted this data frame to get the visualization.

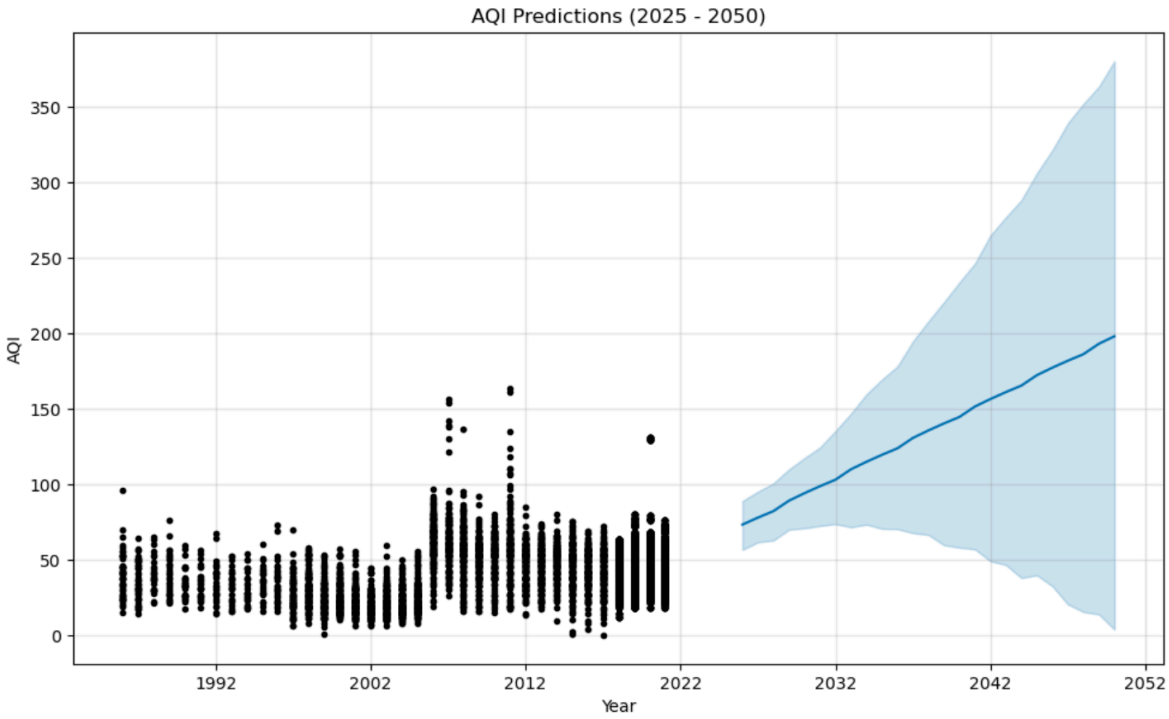
Visualization 4:



This visualization is the result of attempting to fit a model on our custom smoke estimate. We used Prophet to model the data and the solid blue line is the predicted mean smoke estimate for the next 25 years. The light blue shaded region is the uncertainty that the model provides. Unfortunately, our data is quite sporadic and as such, it is hard to effectively model.

However, we can still trust that the general upward trend is accurate. From the data we have, I would say that the predicted smoke estimate values are probably not too far off and the band of uncertainty establishes that there is a lot of variance at play here.

Visualization 5:



Lastly, this visualization is the result of fitting a model to the AQI data. Similarly to the prior visualization, we used Prophet to model the data. We see a clear upwards trend in the predicted mean AQI but this time, the band of uncertainty is much larger. The original data has a high variance and this results in an extremely high uncertainty.

When we try to compare this graph to the custom smoke estimate graph, we see some similarities. We need to keep in mind that the units and scale of the graphs will be much different because they are calculated in very different ways. What we need to look at is the general shape of the plots and when we do that, they are actually quite similar. They are roughly flat and have spikes at similar points on the x-axis. Furthermore, the predicted data also has an upward trend in both graphs. This leads me to believe that our custom smoke estimate is doing a decent job at modeling the smoke that the city receives from wildfires.

There are most certainly other factors at play here that could influence air quality, but the wildfires play a large part and this custom metric captures a good chunk of it.

Reflection

I really enjoyed the collaborative aspect of this project. Due to the nature of the uniquely assigned cities, we needed to strike a balance between collaboration and our own private work. Personally, I did not seek out or share any code with any other students directly. Rather, I talked through the data cleaning and API call processes with others to make sure I was understanding the data flow properly.

One thing I learned from collaborating on this project is that creating a custom metric is challenging. There are so many different ways to make a smoke estimate and I needed to find a balance between a metric that was relatively simple but also accurate. After talking to multiple students, we collectively decided on an inverse square relationship between smoke and distance. This was the key to our metric and we combined it with a proportional relationship with area.

The data gathering and API code is borrowed from the notebooks that were provided to us. The attributions can be found in the main notebook. Overall, I think this assignment was a good challenge and communicating with my peers helped me understand how to go about the data collection process.