# Data 512 Wildfire Project Report

Author: Ashwin Naresh
Repository: nanoash7/data512-final

## Introduction

With the ever increasing rate of climate change, the risk and prevalence of wildfires is something that demands attention. Fires can cause millions in property damage and can lead to environmental degradation through the physical burning and smoke emissions. In this project, I will be looking at the impact of wildfires on the city of **Savannah, GA** and offer some insight on what can be done to limit their effect on the people that live there.

Savanna is the oldest city in the state of Georgia and is well known for its natural beauty. The long list of natural parks and beaches is a major draw for the millions of tourists that the city experiences each year. Given that tourism is such a massive aspect of Savannah, this project could have a major impact on how the city and the state should deal with the rising threats of air pollution and wildfires. And on a personal note as someone who is very involved with the outdoors, I have a personal stake in this project to try and ensure that the city of Savannah deals with air quality and wildfires in the best way possible.

This project contains two major sections of analysis:

Part 1: Common Analysis
- Introductory analysis of wildfire frequency and size near Savannah, GA.
- Creation of a smoke estimate model to predict future air quality for the city.
- Compare the custom model with Savannah's AQI metrics.

Part 2: Project Extension
- Gather external tourism dataset for Savannah, GA.
- Create custom human impact metric and model.
- Recommended action and next steps.

## Background and Prep Work

### Part 1: Common Analysis

For the common analysis (which was performed by the entire class), we were provided with the wildfire data.

Wildfire Dataset: https://www.sciencebase.gov/catalog/item/61aa537dd34eb622f699df81

In particular, we will deal with the USGS_Wildland_Fire_Combined_Dataset.json file which contains json formatted data for millions of wildfires across the country. We were provided with basic code to parse this file and I modified this code to extract the exact features I needed from each JSON entry, namely the id, year, size, and distance (from Savannah) for each fire. Given the enormous size of this file, it is not feasible to run this code multiple times. As I parsed the json data file, I write the parsed data to a csv file which I can load at a much higher speed.

Parsed Wildfire Data: parsed_data.csv
Columns:

- Id: The assigned id for the fire which is a unique value.
- Year: The year that the fire occurred.
- Size: The size of the fire in acres.
- Distance: The distance of the fire in miles from Savannah, GA.

The other piece of data I needed to collect was the Air Quality Index measurements for Savannah, GA. Again, this API was provided to us and I used it to gather the AQI particulate data for the city from 1986 to 2021. I used this API to construct a csv file that contains a list of AQI measurements from the various monitoring stations for each year.

AQI API Root: https://aqs.epa.gov/data/api

Parsed AQI Data: parsed_aqi_data.csv
Columns:

- Year: Year of the AQI measurement.
- AQI: Particulate AQI value for a specific monitoring station.

## Part 2: Project Extension

My initial extension plan ended up not being feasible. My original plan was to analyze the economic impact of worsening air quality on the city of Savannah. However, to my surprise, this type of economic data was not publicly accessible. I even tried to look for housing and rental data to try and make the connection to the economy, but this data was either missing a lot of values, or was collected over too short of a time period to be meaningful. As a result, I ended up switching the focus of my extension plan to expanding upon the smoke estimate I created in part 1.

The dataset I ended up using was provided to me by the city. To get this data, I emailed the Savannah Area GIS Open Data helpline found on SAGIS Open Data Site and inquired about their tourism data. They provided me with the input file that I used for analysis in part 2.

Tourism Data: tourist_data.csv
Columns:

- Year: Year of the tourism measurement.
- Tourists: Total number of estimated tourists. Sum of the following 2 fields.
- Overnight-visit: Estimated number of single day tourists.

- Day-visit: Estimated number of multi-day tourists.

# Methodology

## Part 1: Common Analysis

The wildfire data we acquired from the GeoJson wildfire dataset contains fires from across the country. However, we want to keep our analysis local to the city of Savannah and as a result, will be excluding fires that occur more than 650 miles from the city in the models we create. This ensures that our model excludes wildfires that have a negligible impact on Savannah.

Another step I needed to do was clean and aggregate the AQI data. The AQI data I collected spans multiple different monitoring stations, and as it turns out, not all of the stations collect data at the same time. As a result, there were a lot of null values that needed to be removed, and after removing them, I was able to aggregate the entries to get the average AQI estimate for each year.

### Smoke Estimate Rationale

With the data parsed and cleaned, I created the smoke estimate model. To do this, I needed to quantify the impact of a given wildfire on Savannah's air quality. Smoke roughly follows the inverse square law, where the intensity of the smoke is inversely proportional to the square of the distance. Furthermore, the intensity of the smoke is also roughly proportional to the size of the fire, which we are measuring in acres. Putting this together gives us the following naive smoke estimate metric:

$$Smoke\ Estimate\ = \frac{Wildfire\ Area}{Distancee^2}$$

### Model Selection

When trying to model a phenomenon, it is important to select the correct type of model. A well selected model is fine tuned for the type of data at hand and given the vast amount of options available, this will take some extra time to decide.

The smoke estimate data we have contains missing values, has seasonal/periodic sub-trends, and involves time series data. These features line up perfectly with the Prophet modeling library, developed by Meta. Prophet comes pre-prepared with a variety of training techniques and I ended up going with the standard configuration.

## Part 2: Project Extension

The tourism data I received from SAGIS Open Data was already cleaned for me, which meant I could immediately start on the model creation.

### Human-Impact Metric

To gauge the impact of rising smoke levels, we will use a new metric that combines the smoke estimate from part 1 with the tourism data. The first thing we do is join the smoke estimate metric with the tourism data and select “year” as the join key. We can discard any rows where the year field has no match. The next step is to normalize the inputs seeing as their current ranges are drastically different: Smoke Estimate ~0 to 170 and Tourism ~2 million to 17 million. We can use the MinMaxScaler library to normalize the input fields and create a new column in the data frame for that. Finally, we can add weights to the input fields and use a linear combination to create our human-impact estimate:

$$Human\ Impact\ Estimate\ =\ \alpha(Normalized\ Smoke\ Estimate)\ +\ \beta(Normalized\ Tourism\ Count)$$

We can select alpha and beta in accordance with how impactful each of these factors are. The air quality plays a greater factor than the number of people, so an alpha value of 0.7 and a beta value of 0.3 is what we will use.

Just like with part 1, we will be using the Prophet modeling library to create our time series model for the human impact estimate.

## Ethical Considerations

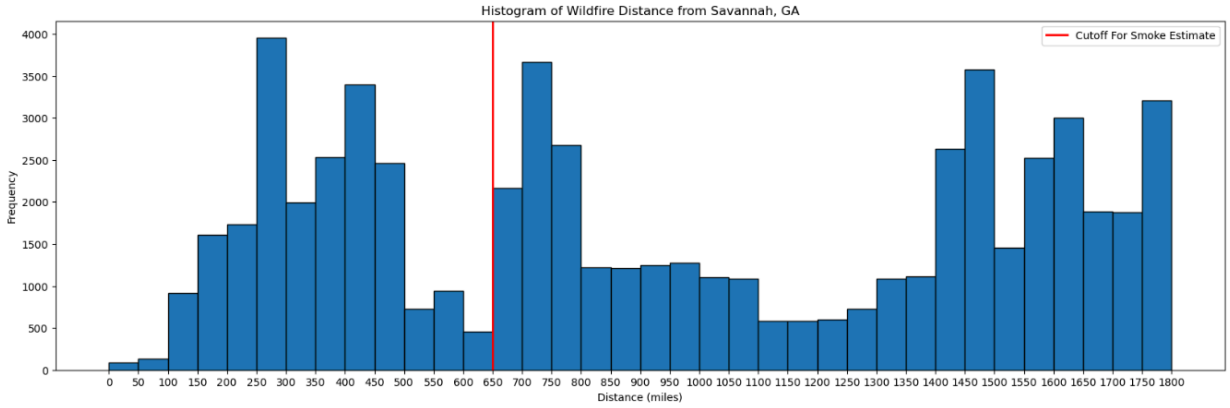
An important aspect of a real world data science project like this one is to make sure that we consider the ethical implications. In this project, we make sure to only use anonymous and non-personally identifiable information like total tourists counts. This means that an individual’s data is never stored or used. Furthermore, it is important to make it clear that the results of this project are provided in the proper context and that all the possible limitations and shortcomings are laid out properly.

# Results

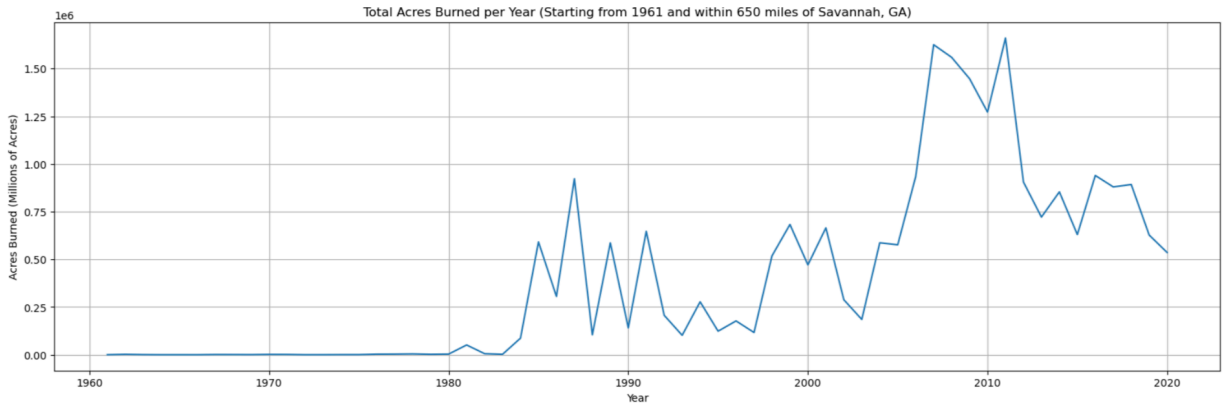
## Part 1: Common Analysis

### Exploratory Analysis

The following histogram shows the numbers of wildfires that occurred in 50 mile increments from Savannah, GA. The data is constrained to 1800 miles and the red line at the 650 mile marker is the cutoff for fires that will be considered for the upcoming smoke estimate model.

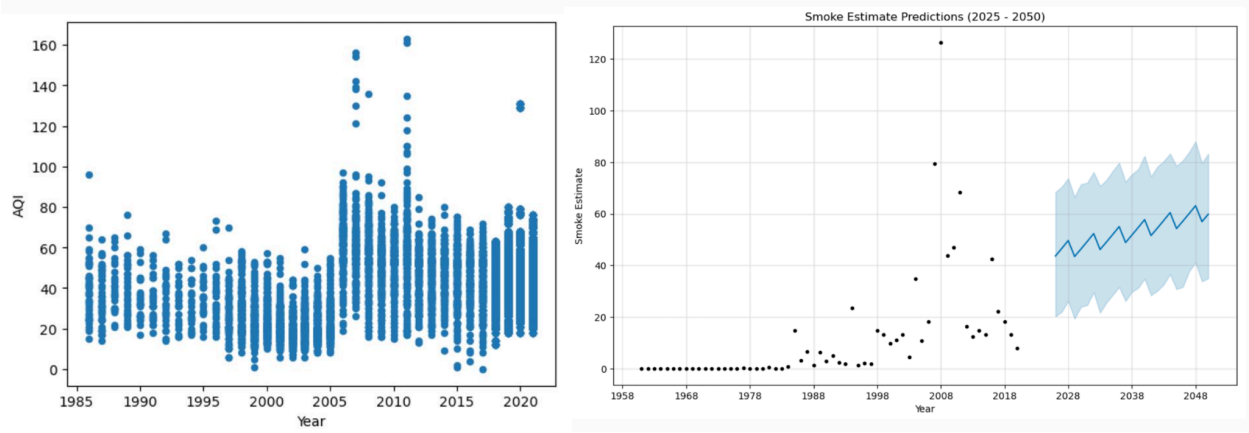



The following graph shows the number of acres that were burned per year. We discard all fires that occurred outside the 650 mile radius from Savannah, GA. We see a first spike around 1986-1987 and a larger peak around 2008-2011. Overall, the amount of acres burned is increasing.




## Smoke Estimate Analysis

The plot on the left is the actual AQI measurements from monitoring stations around Savannah, GA. In particular, I am considering the particulate AQI value and plotting it alongside my custom smoke estimate on the right. The graph on the right is my custom smoke estimate model, where the section in the blue highlight is the predicted error bound for 2025-2050 according to the prophet model I created.

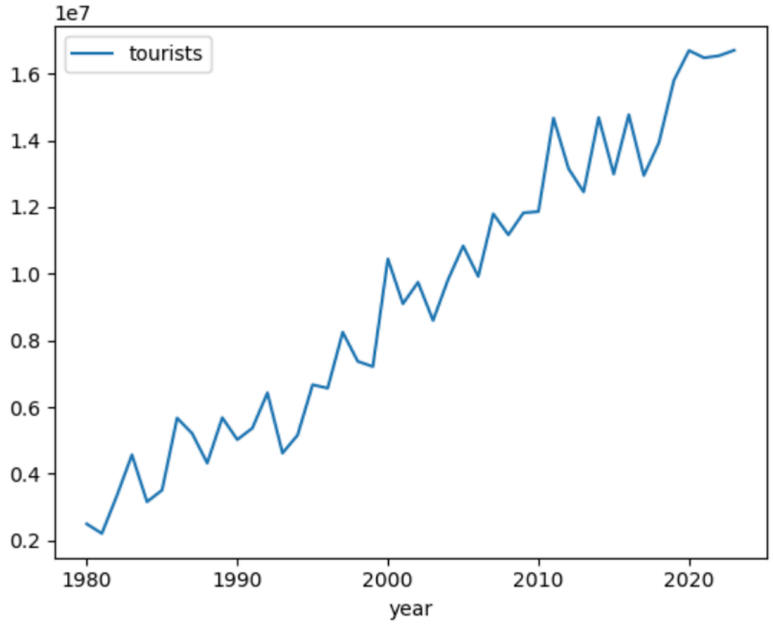



The plots share a similar shape, with a roughly flat start, a peak around 2008, followed by a drop, and an upwards trend around 2020. This is supporting evidence that the custom smoke estimate model is doing a good job at modeling the air quality (particulate aqi) around Savannah, GA.

## Part 2: Project Extension

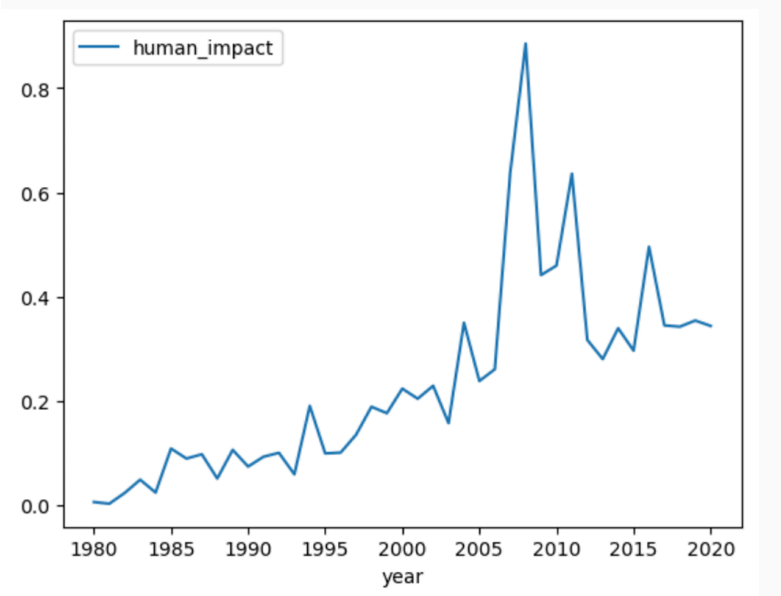
### Exploratory Analysis:

The following graph is the plot of the tourism data I received from the SAGIS website. The data provided has a few different columns but for the purposes of this project, I will only be considering the total tourism value. The data starts in 1980 with a value around 2 million and rises to a value of around 17 million at 2023. Savannah is a quickly growing tourism destination, and this is reflected through the steady rise in tourism as seen in the graph. There are a few spikes and dips which I am unable to reconcile, but the overall trend is a steady upwards increase.

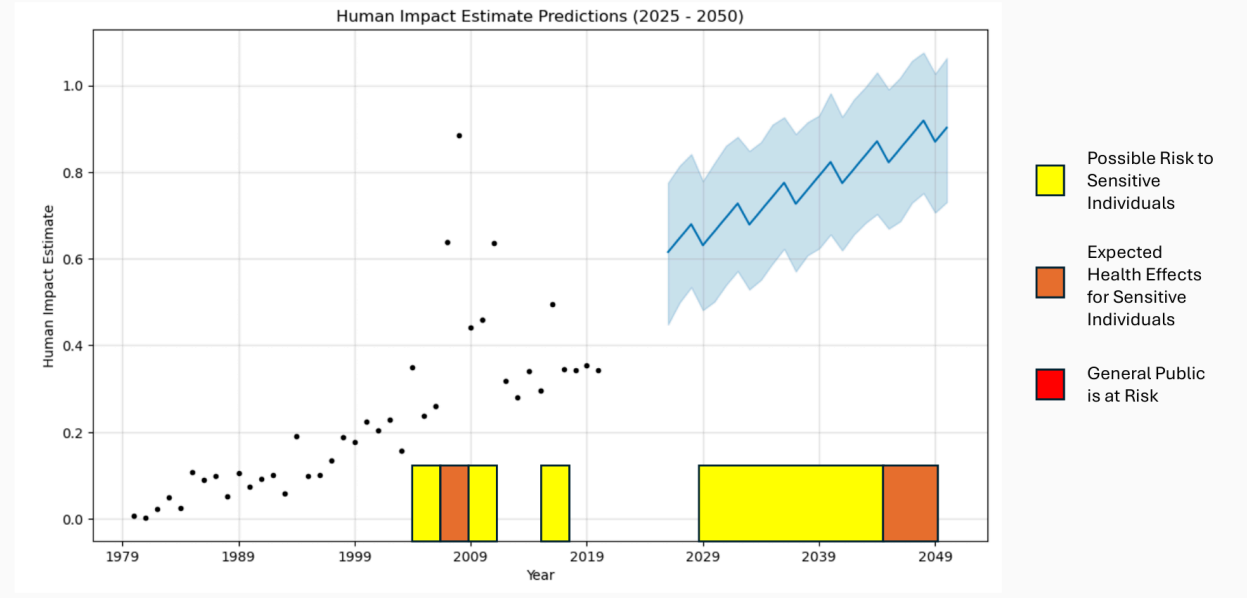


## Human Impact Estimate

Like mentioned in the methodology section, we needed to normalize and weight the data before combining the inputs to create our new metric. The following graph is the result of those processing steps, with the y-axis being the new human impact metric. We see a steady increase from 1980 to 2005, followed by a huge spike from 2005 to 2012. After the spike dies down, the steady increase resumes. This chart is showing us that the effect of the rising smoke levels is also increasing and that the rising tourism levels will only accelerate the impact on public health.




The following plot displays the human impact metric, along with the predicted values according to the prophet model. Just like the last model, the blue highlight represents the projected error bounds from 2025-2050. I have also included a color coding to indicate what the air quality ratings are for all the points which cross an AQI value of 50. THis helps us see that a substantial portion of the human impact metric is coming from the smoke estimate, meaning that this impact will be accelerated even further by an increasing tourism rate.




Given how the metric is calculated, a value of 0.8 is approximately the threshold for slight concern with a value of 1.0 being the threshold for immediate action. The metric crosses the 1.0 mark sometime in the 2050's but action needs to be taken before that in order to sufficiently prepare for it. Sensitive individuals will start being affected in the late 2030s and early 2040s, which means that state level and city level actions need to be taken (which will be discussed in the upcoming sections).

## Limitations

One limitation from part 1 is the structure of our naive smoke estimate. We use the inverse square law to create a very simple relationship between fire size, distance, and smoke, but the relationship is likely to be more complex. Improving this smoke estimate will give us a better human-impact estimate, which will give us better insight into the future air quality of Savannah.

The biggest limitation in the accuracy of the human-impact estimate model is the assumption that tourism and air quality and independent inputs. It's reasonable to expect that tourism numbers will drop if the air quality gets bad enough. The state and city will likely put out warnings, rendering these variables dependent. If we could find a way to model the relationship between air quality and tourism, then we could further refine the human impact metric.

Apart from these larger limitations, there are a few smaller ones that if resolved, could lead to even more accuracy on the various models we created. First, we were only able to get smoke estimate data through 2020. Second, the "length of stay" of tourists can help us refine the human impact. However, this distinction is binary in the data we were provided. If we had data to mark exactly how long a tourist visited, we could get a much more accurate human impact estimate. Finally, we could incorporate wind patterns to get a much more accurate smoke estimate for part 1, leading to a better human impact estimate in part 2.

## Implications and Recommendations

It is important to preface this section with the limitations presented in the previous section. I want to establish that these findings are not free from error and the following recommendations should be taken with the appropriate amount of caution. This is a fundamental tenet of human centered data science and I want to make sure the results of this project are consumed with the right context.

The results we collected indicate a constant and consistent worsening of Savannah's quality. While the quality at the current moment in time is acceptable, the conditions worsen through the late 2030s and 2040s reaching potentially problematic levels by the 2050s. Particularly sensitive individuals or those with pre-existing health conditions will likely be affected first, closer to the end of the 2040s. With the steadily rising tourism levels, this worsening air quality will have an outsized impact on public health in Savannah, GA.

I am recommending a call to action to try and tackle this issue. At the city level, I recommend things like green infrastructure to try and reduce the carbon footprint of the city in an attempt to isolate the source of the worsening air quality. Another local recommendation is new energy policy and the creation of public health campaigns to try and inform the public about the potential threats that wildfire and smoke can have on the environment. This will increase the public's awareness about the issue and serves a key stepping stone for larger actions. At the state level, I am recommending an overhaul of the wildfire prevention and fire fighting procedures. These fires need to be contained quicker which also requires monitoring systems to be revamped.

## Conclusion

In this project, we looked at the effects of wildfires on the air quality and people of Savannah, GA. I started with exploratory analysis, looking at the frequency and size of fires near the city. I used this data to create a smoke estimate metric and compared it to the actual particulate air quality index (AQI) values collected from monitoring stations near the city. To finish part 1, we took this smoke estimate data and fit a prophet model, forecasting the smoke values through the year 2050.

Next, I took a dataset containing yearly tourism numbers of the city of Savannah and combined it with the smoke estimate data to create a human impact estimate. To do this, I needed to normalize and weight the data, finally feeding the resulting data into another prophet model, forecasting the human impact of wildfires through the year 2050.

Through this analysis, we learned that the air quality in Savannah, GA is steadily worsening. Sensitive individuals and those with pre-existing conditions will start getting affected around the year 2050. Lastly, I highlight ways at both the city and state level that we can try and fight this problem.

In conclusion, the analysis highlights the significant impact of wildfires on Savannah's air quality, emphasizing the urgent need for proactive mitigation strategies and robust public health measures to address the rising frequency and severity of such events.

## References

SAGIS Open Data Website: SAGIS Open Data Site
Sciencebase Wildfire Data: Combined wildland fire datasets for the United States and certain territories, 1800s-Present (combined wildland fire polygons) - ScienceBase-Catalog
Inverse Square Law Background: Inverse Square Law

## Data Sources

The data sources used for this project were collected from the ScienceBase Wildfire website and the SAGIS Open Data website. The tourism data can be found in the linked report at the top of this report. Unfortunately, the raw wildfire data is too large to store in the repository but the parsed and cleaned wildfire csv files can be found in the repository. All intermediate data files and their formats and provided in the repository along with the code for using them.