Part 1 - Common Analysis

The Course Project

The course project consists of four parts:

- Part 1 - Common Analysis sets the stage for the subsequent assignments. In Part 1 you conduct a base analysis. All of the students in the class will conduct the same analysis, but with a slightly different data subset.
- Part 2 - Extension Plan will require you to ask a human centered data science question that extends the work in Course Project Part 1 - Common Analysis.
- Part 3 - Presentation will require you to give a modified (shorter) PechaKucha presentation of your completed project.
- Part 4 - Project Repository, creation of a fully documented repository and also requires the submission of a written project report.

Common Analysis

More and more frequently summers in the western US have been characterized by wildfires with smoke billowing across multiple western states. There are many proposed causes for this: climate change, US Forestry policy, growing awareness, just to name a few. Regardless of the cause, the impact of wildland fires is widespread. There is a growing body of work pointing to the negative impacts of smoke on health, tourism, property, and other aspects of society.

The course project will require that you analyze wildfire impacts on a specific city in the US. The end goal is to be able to inform policy makers, city managers, city councils, or other civic institutions, to make an informed plan for how they could or whether they should make plans to mitigate future impacts from wildfires.

Sharing and Collaboration is Allowed

For **PART 1 ONLY** all students in the class **MAY SHARE CODE SNIPPETS**, **STATISTICAL APPROACHES**, and **VISUALIZATION TECHNIQUES**.

SNIPPETS are OK. Students are **NOT ALLOWED** to share **THE SOLUTION**. That is, you may not share a specific coded application or collection of subroutines that comprises a mostly complete solution.

We are encouraging **SHARING** but we want sharing to include **COMPREHENSION** of the **METHOD, APPROACH, and TECHNIQUE**. Your mantra for sharing is “I can help you understand this, but I won’t do this for you.” When sharing a snippet, a statistic, or a technique, it is very helpful to explain what it does. One advantage is that you are all

working on the same data with the same structure. The context of your explanation is relatively fixed for this assignment. You are all working in the same context.

When you borrow or reuse a code snippet, or a statistical approach, or some technique that was provided or outlined by one of your classmates, you should keep track of WHO provided it so you can make an appropriate **ATTRIBUTION** in your submission of Part 1.

Step 0: Data acquisition

The common analysis research question is based on one specific dataset. You should get the Combined wildland fire datasets for the United States and certain territories, 1800s-Present (combined wildland fire polygons) dataset. This dataset was collected and aggregated by the US Geological Survey. The dataset is relatively well documented. Fire polygons are available in ArcGIS and GeoJSON formats.

You have been assigned a specific US city as the focus of your analysis. You are **NOT** analyzing the entire dataset. You have been assigned one US city that will form the basis for your individual analysis. You can find your individual US city assignment from this Google spreadsheet.

Step 1: Create fire smoke estimates

The common research question that you are to answer is:

- What are the estimated smoke impacts on your assigned city for the last 60 years?

You are to create an annual estimate of wildfire smoke in your assigned city. This estimate is just a number that you can eventually use to build a predictive model. Technically, smoke impact should probably be considered the health, tourism, economic or other social problems that result from the smoke. For this we'll generically call your estimate the wildfire smoke impact. You will consider other potential social and economic impacts during Part 2 of the course project. For now, you need some kind of number to represent an estimate of the smoke your city saw during each annual fire season.

Why is this an estimate of fire smoke? These are estimates because of a number of problems that are not easy to resolve and simplifications to make this course project reasonable for just a few weeks of work. One example is that actual smoke impact is based on wind direction over a course of several days, the intensity of the fire, and its duration. However, the fire polygon data only gives a year for each fire - it does not provide specific start and end dates for the fire.

Your smoke estimate should adhere to the following conditions:

1. The estimate only considers the last 60 years of wildland fires (1963-2023).
2. The estimate only considers fires that are within 1250 miles of your assigned city.
3. An annual fire season will run from May 1st through October 31st.

You have a number of different options for working with the wildfire data. Python has GeoJSON readers and ArcGIS file readers as distribution modules you can install. If you are a pandas pro, it supposedly has GeoJSON and ArcGIS data readers. As another option, I wrote a relatively simple GeoJSON reader that I used to facilitate my own exploration of the dataset. Use the option that works for you.

Once you are able to read the wildfire data, you need to be able to find all of the fires within a specified distance from your city. This is a geodetic distance computation. If you are a GIS wizard, you'll know how to do this in your sleep. However, for most of us, who have never done this before, an example is really helpful.

The third problem you need to resolve is how to estimate the smoke for your city. It seems reasonable that a large fire, that burns a large number of acres, and that is close to a city would put more smoke into a city than a small fire that is much further away. One task is to define your smoke estimate and then apply it to every fire for your city. Should your smoke estimate be cumulative during each year or somehow amortized over the fire season? Document what goes into your decision making and what you did to create your smoke estimate.

Another problem is trying to understand how good or bad your smoke estimate might be. Once you have developed your smoke estimate, you should compare your estimate to available AQI (Air Quality Index) data from the US EPA. I hear you asking, "Why we don't simply use AQI as the estimate?" That's a reasonable question. First, the US EPA was only created in 1973, and did not really begin installing air quality monitoring stations until the early 1980s. Further, of 3000+ counties in the US, the EPA has vetted monitoring stations in only 2000 of them. This means that US EPA AQI measures for any one city will need to be some kind of estimate based on monitoring stations that are nearby. There are a few additional considerations that I have glossed over.

I have provided sample code for accessing the US EPA Air Quality System API. You will need to request an API key to access the API. You should test access to the API to understand whether you can get data for your city based on the US County where your city is located, or whether you need to create a geodetic bounding box to access station data. In either case, it is highly unlikely that your city has an AQS monitoring station directly in the city bounds - although it is possible.

When comparing your smoke estimate to the estimate you create from the EPA AQS, you should consider the nature of AQI measures, what they represent, how they are computed, and how they might be (or not be) related to fire smoke.

Lastly, you should develop a predictive model based on the fire data and smoke estimate for your assigned city. Your model should predict smoke estimates for every year for the next 25 years (i.e., 2024-2049). You should be careful to make sure your predictions convey appropriate levels of uncertainty in the prediction.

Step 2: Visualize aspects of your analysis

You will illustrate the work of your analysis with a few time series graphs. The wildfire data is annual, so your time series will be on an annual basis. All the time series should cover the analysis range defined above, but not the prediction range.

1. Produce a histogram showing the number of fires occurring every 50 mile distance from your assigned city up to the max specified distance.
2. Produce a time series graph of total acres burned per year for the fires occurring in the specified distance from your city.
3. Produce a time series graph containing your fire smoke estimate for your city and the AQI estimate for your city.

Step 3: Write and reflect

This step has two objectives.

First, for the visualizations you created in Step 2, you should write up an explanation of each visualization. Some of the important things you might need to explain include: What does the figure show? How does the viewer "read" the figure? What are the axes, and what do they represent? What is the underlying data and how was it processed? You might think of this explanation as an extended figure caption. Each explanation should be no more than one written page. Making a good effort now will make it easier to write your final report for Part 4.

Second, we would like to understand what you got out of the collaborative activities in this assignment. You should write a reflection statement that highlights one or two specific things that you learned from answering the research question posed in this assignment. How did the possibility of collaboration help, hinder, or change your thinking about the problem? Your reflection statement should include specific attributions for all code, methods, and techniques that you reused. Your reflection statement should be no more than two written pages.

Step 4: Submission

The submission for Course Project Part 1 will include the following items:

1. A link to a snapshot of your current repository that should minimally include:
    a. code/notebook,
    b. an appropriate README including information on the data used for this part of the project, and
    c. a license file.

    Your repository link should be a stable copy of your repository that can be used for grading. Even though this submission is only part of the course project, your repository at this point should still adhere to best practices for documentation and reproducibility.

2. Submit a PDF file or google doc (link) containing your visualizations (specified in Step 2), your figure descriptions (extended figure captions), and your reflection statement on the collaborative aspects of the assignment.

Note, all linked documents should be shared with and viewable by instructional staff so that we may grade your Course Project Part 1 submission.

At this point, we are not asking for an extensive write-up of everything you did in this part, Part 1. Course Project Part 4 requires that you submit a full write-up of this part, Part 1, as well as the work you do for your Extension (Part 2). However, you are encouraged to start drafting the relevant sections of your report now, as that will make it easier to complete the Course Project.