DS5110 - Final Project
Iteration 4

Julian Baker
Konstantinos Theodoropoulos

Novemeber 2025

1 Dataset Description

1.1 Dataset of 2.3 million US wildfires

The dataset we use can be found at this link.

This dataset is the 5th update to a dataset used for National Fire Program Analysis (FPA) system, and contains records of wildfires from 1992 to 2020. It includes 2.3 million records of wildfires over the 29 year span. This appears to be the most up-to date records of wildfires in the US, and critically has had effort spent to remove redundant records - something we felt could be an issue with our plan of combining data. Structurally, the dataset contains many variables, but of key importance (for analysis of where they occur) are County, State, FOD-ID, Discovery Date (and Year), Containment date, and Fire Size. There are more variables that we expect to use, but the jist of it is that we can identify when and where they occur and were contained, and the size of the wildfire. We should also be able to calculate how many days a wildfire is alive. Simply put, this appears to be the best, well-kept and most recently up-to date dataset of wildfires, and an important baseline for a project designed around wildfires.

Source:
Short, Karen C. 2022. Spatial wildfire occurrence data for the United States, 1992-2020 [FPA_FOD_20221014]. 6th Edition. Fort Collins, CO: Forest Service Research Data Archive. `https://doi.org/10.2737/RDS-2013-0009.6`

1.2 Canada's National Fire Database records

The dataset we use can be found at this link

The second dataset that will be used is the Canadian National Fire Database (NFDB) fire point dataset, which compiles forest fire locations reported by provincial, territorial, and federal fire management agencies across Canada. The dataset is provided as a shapefile and includes key attributes such as latitude and longitude, fire start and end dates (where available), fire size, cause, and fire type.

This dataset is well suited to the project's objectives because it contains many of the same features present in the U.S. wildfire dataset, allowing the two sources to be merged into a unified SQL database.

Source:
Natural Resources Canada. 2024. *National Fire Database: Fire Point Data (NFDBPNT).* Canadian Wildland Fire Information System. `https://cwfis.cfs.nrcan.gc.ca/datamart/download/nfdbpnt`

2 Tools and Methodologies

For this project, we will need software to analyze and store the data, and then use models to make predictions. For the analysis and merging and cleaning, Google Colab will be used as the software of choice - as it is the most familiar option with these capabilities. Google Colab also makes it easier to work on the same code if need be, and we really just need some software to run python well for this project. The models made will be determined after data exploration, but likely will feature some form of regression, linear or otherwise (logistic regression in particular seems promising if we wish to classify counties as having a large wildfire or not). We should justify the models used prior to running the them, though.

We will use SQLite to create and manage the SQL database, working directly from the terminal. SQLite is lightweight and easy to set up, making it a good choice for storing and querying the merged dataset. Its simplicity means that the database can be accessed and used reliably without extra configuration.

3 Preliminary Timeline

The schedule for this project runs as follows:

- Select the individual dataset to use (11/12)
- Finish Iterations 2-4 (11/14)
- Clean said datasets (11/15)
- Merge all datasets (11/16)
- Creation of SQL database (11/23) (Konstantinos)
- Data exploration (11/23) (JJ)
- Predictive Models of wildfires (11/26) (JJ)
- SQL queries (11/26) (Both)

- Final Project Presentation (12/1, finish day prior)
- Final Project Report (12/8)

4 Team Member Contributions

4.1 Julian Baker

So far, Julian has worked as the management for this team, setting deadlines for this project and communication within the team. He identified the US wildfire dataset early in the process as being a great starting point for merging dataset. He has looked ahead to ensure that the dataset can be used for analysis has done a bit of work on that front, though for is required for the exploratory analysis. He has contributed writing to all of the iterations. As team, we have collaborated via text and talking in class, and Julian has helped assign tasks fairly to ensure an even workload.

4.2 Konstantinos Theodoropoulos

So far, Konstantinos has focused on data preparation and backend development. He found and evaluated datasets to make sure they are compatible and have the necessary information for merging into a single SQL database. He has also looked into SQL methods for creating the database and contributed to writing and updating the project reports for each iteration. The team has collaborated through regular check-ins, sharing resources and updates, and coordinating tasks.

5 Progress and Next Steps

In summary, the following tasks have been completed: gathered the datasets, light exploration, defining the timeline, and completed the iterations. We wish we could be further along, but with the loss of a team member and sickness, we are just happy to have a plan to finish this project on time while still achieving most of our desired goals from the start. Despite the loss of a team member and many unforeseen setbacks, we are confident that we can complete this project on time and achieve most of the goals that we set out to do from the start.

Julian:
The way this project's work is split, we each will handle one dataset and one large task. For Julian, his dataset will be the US wildfire dataset - he is tasked with cleaning it. Both team members will be responsible for merging the datasets. Once the datasets have been cleaned, Julian will be responsible for exploratory analysis and models of the merged dataset.

Konstantinos:
In the upcoming week, Konstantinos will work with the Canadian wildfire dataset, focusing on cleaning and preparing the data. He will also build an SQL

database to combine the U.S. and Canadian datasets and start identifying SQL queries that could provide useful insights for the analysis.