

DS5110 - Final Project

Iteratation 2

Julian Baker
Konstantinos Theodoropoulos

Novemeber 2025

1 Project Kickoff

The specific goals and outcomes of this project are to produce a report and presentation on Wildfires in the continental US. Specifically, we will gather two datasets on wildfires and make predictions and analyses based on them. Our previous experience tells us that it may require a lot of data cleaning. After gathering the necessary data, we will be able to move on and make predictions and analyses, including models predicting future wildfires.

The scope of the project can be defined by the following key deliverables:

- Creation of a dataset that is formed by merging two wildfire datasets
- Creation of an SQL database based on said dataset
- Examples (8) of queries for the SQL database, showing how it can be used
- Development of a predictive model for future wildfires
- A presentation and report showcasing our findings and our work

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, unsure if this is correct)

I was unable to find the exact due dates for them, but I will tentatively assume they are due December 1st for the presentation and December 8th for the report, though this is speculation.

2 Team Discussions

Unfortunately for us, we share a very similar skillset; great experience in programming languages to analyze data and good experience in cleaning datasets, but lacking in experience with SQL and databases.

As such, one team member will need to focus on the SQL portion of this, and the other will need to take on more work to compensate. This is where the loss of a team member really hurts - much more could be done with the SQL portion if we had a third teammate, and originally we had planned for that.

Instead, Konstantinos will be responsible for designing and making the SQL database alone. In return, Julian will be responsible both for the data exploration and the data modeling. Once the SQL database is made, both will be required to work on SQL queries, namely those provided in the feedback of iteration 1 (though more is always welcomed, within reason).

Technologically, Julian is comfortable with using Google Colab and Python to run said exploratory analysis and develop models. If need be, he can use R with RStudio to develop graphs (they believe those look nicer), but the code should mainly, if not all, be done in Python as per the specification in the final project's description.

Although Konstantinos is most familiar with Python, he has prior exposure to both SQLite and PySpark and is prepared to strengthen his skills in whichever platform best supports the project. To effectively manage the backend data preparation, he may need to review certain SQL tooling and workflow options, but he has sufficient foundational experience to quickly learn any additional techniques required.

3 Skills and Tools Assessment

The team has access to a range of external resources that can provide support when needed. The course professor and teaching assistants are available to assist with SQL workflows, database integration, geospatial data handling, and general troubleshooting. Making use of these resources aligns with the project's emphasis on problem-solving and independent learning.

Given the project's scope and the nature of the datasets, a combination of SQL and Python-based tools is appropriate. One dataset is already structured as a database containing 2.3 million U.S. wildfire records, while the Canadian dataset is provided as a shapefile. Merging these into a unified SQL database can be handled effectively with SQLite. Python will then be used for analysis and modeling, drawing on libraries such as pandas, NumPy, and the relevant visualization or machine learning tools. This toolset balances capability with familiarity, ensuring the team can work efficiently and adapt as needed.

To stay proficient with the selected tools, the team will reference online documentation, tutorials, and examples throughout development. Whenever questions arise, they will seek clarification from the professor or teaching assistants to avoid bottlenecks and maintain steady progress.

Roles were assigned to make the best use of each member's strengths while still encouraging collaboration. Both team members will participate in SQL query development and overall decision-making, ensuring the workload is shared and that each person contributes meaningfully to the project's core components.

4 Initial Setup

The project will be developed using a combination of local and cloud-based tools. SQL work will be carried out using SQLite through the command line, while Python development will take place primarily in Google Colab. The shared codebase will be maintained in a GitHub repository, allowing for easy access and modifications. Additional geospatial libraries may be utilized later if needed to process the Canadian shapefile.

Version control has been set up through GitHub, and both team members have full access to the repository. With one member working locally and the other working exclusively in Colab, GitHub will serve as the central storage medium for all project files.

5 Dataset Links

5.1 Dataset of 2.3 million US wildfires

<https://www.kaggle.com/datasets/behroozsohrabi/us-wildfire-records-6th-edition>

5.2 Canada's National Fire Database records

<https://cwfis.cfs.nrcan.gc.ca/datamart/download/nfdbpnt>