ALBERTA WILDFIRES

A preliminary data analysis of wildfires across Alberta for the purpose of early detection and/or possible prevention

Saad Usmani
Saadusmani1993@gmail.com

Table of Contents

1. Introduction

1.1 Purpose

The purpose of this document is to provide a detailed description of the technical aspects of the project. It lays out the technical thought process for users to gain a better understanding of the methodologies and processes used to design the Power BI dashboard and populate it with relevant data.

This document is an optimal reference point for any upgradation and changes that may be required for the project.

1.2 Scope

This wildfires project is completely reliant on the data provided by the Government of Alberta on an annual basis. Hence, the initial boundary faced in this analysis is a limitation of historical data that allows conclusions to be drawn up till a year prior to the current. In this case, the wildfires that have been studied are up till the year 2023.

The Power BI dashboard is set up in a way that allows key point indicators to be updated automatically when new data is fed into the system.

The project will cover geospatial, causal, and time-based elements within the sphere of wildfires across Alberta. It provides generalized answers to the following objectives:

1) Determine and mark locations where wildfires occur most frequently and analyze other features of those locations for better understanding.
2) Determine the most common causes of wildfires and attempt to analyze the trends.
3) Determine trends and patterns of when most wildfires happen and see if an approximation of prediction is possible.

Inferences and actionable conclusions that have been drawn from this dataset are solely within the realm of data analytics and not data science. This means that the project lays the groundwork for statistical analytical and machine learning techniques but does not venture into them.

However, as this project is being published and presented, our team is at work implementing scientific approaches to the dataset, involving machine learning algorithms.

2. System Overview

2.1 System Architecture

The system architecture of the project is divided into the following components:

- **Data Sourcing:**
    - Data for the project is sourced from a public website that lists a dataset which is updated periodically.
    - This data is in the form of a CSV file that needs to be downloaded.
    - An automatic API connection to the data is not available.
- **Data Storage:**

  - For the purpose of this project, the dataset is downloaded onto a portable computer and stored on an SSD.
  - In later iterations, a system will be set up that will automatically update a live dataset stored on cloud. This new dataset will then be periodically pushed into the Power BI dashboard for updated insights.

- **ETL Process:**
  - For the most part, this dataset is extracted into a DataFrame and manipulated in that form for the purpose of analysis.
  - 80% of the ETL process is done using Jupyter Notebook, Python, Pandas, and Numpy. After this process, the new data is loaded onto Power BI using the available Python plugin.
  - Some transformations like removing columns and creating measures are done on Power BI itself.

- **Data Modelling:**
  - After understanding and loading the data, a star-schema for the dataset is designed in Power BI.
- **Visualization:**
  - The dataset and the insights it provides are visualized using Power BI and its tools.
  - Appropriate visualization methods like bar charts, histograms, pie and line charts, are used to address the 3 objectives listed above.

2.2 System Components

Break down the major components of the system. For instance:

- **User Interface:**

The user interface in Power BI is divided into 5 parts:

  - The project introduction and insights interface are integral to any new user attempting to understand and utilize the project. It gives a quick overview of what it is about and what the major findings are. Insights can be edited to include new data that arrives over time.
  - A One Glance Overview and One-window Navigation page is included to provide high-level information about the current dataset. A sort of preliminary understanding before they dive deeper into the analysis. The One-window navigation page lets a user quickly find a particular fire/s that they are searching for.
  - The rest of the 3 parts address the objective questions of where, why, and when wildfires occur.

- **Backend Processing:**

Most backend processing, as mentioned earlier, is performed on Jupyter Notebook with the help of the Pandas and Numpy libraries. Machine learning libraries are to be included in a later iteration of the same project. This includes the following:

  - Creation and separation of data into objective-based DataFrames.
  - Cleaning data of nulls using various techniques. Addressing outliers and anomalies via research in order to include or exclude them.
  - Removing unnecessary columns and rows.
  - Preparing the DataFrames for export into Power BI.

3. Data Management

3.1 Data Sourcing

There is only one source of data for this project in its current state. However, other datasets that can improve or advance the analysis may be included later.

- Alberta Government > Open Government > Open Data: This website is owned and updated by government agencies. This particular dataset was uploaded by the Forestry and Parks body.
- Potential datasets:
  - Topographical data of Alberta outlining geographies that experience little to no wildfires.
  - Weather data to fill up time-series gaps in current dataset when wildfires are not occurring.

3.2 ETL Process

The ETL process which was summarized above, goes as follows:

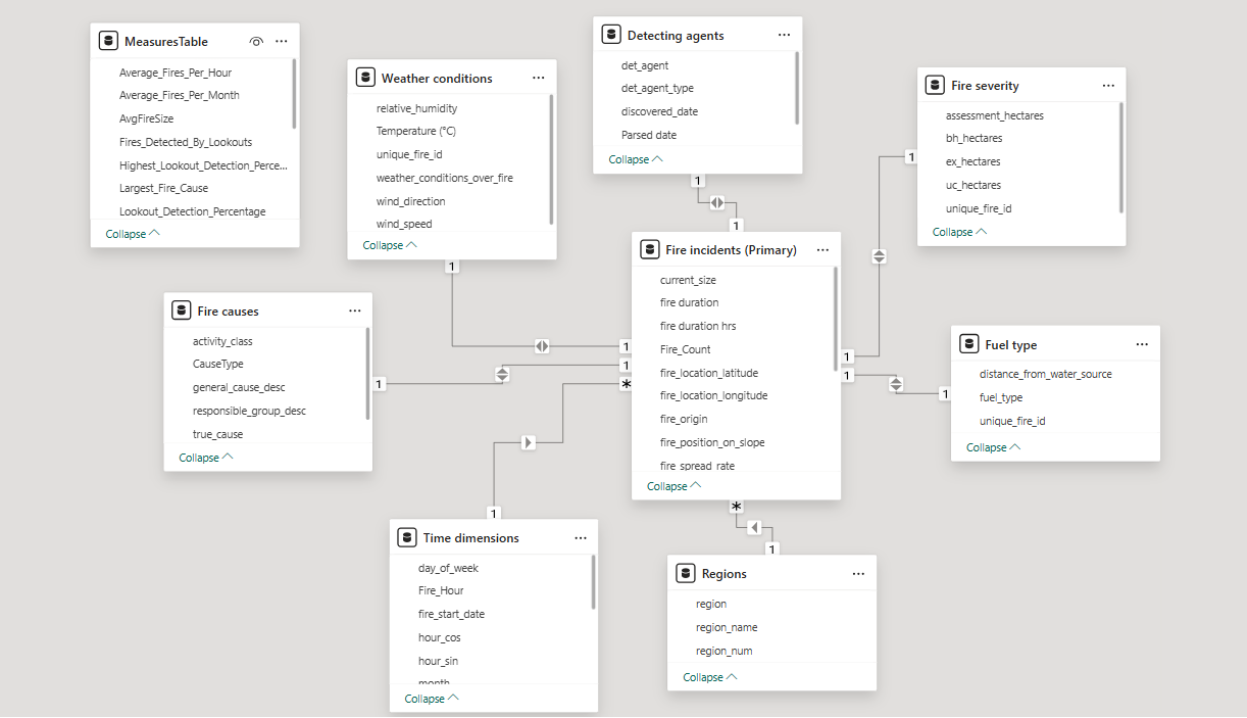
- **Extract**: The data is extracted downloaded from the Government of Alberta Open Data Portal and saved on an SSD. It is then imported on a Jupyter Notebook instance in the form of a DataFrame using the Pandas library.
- **Transform**:
  - Initially, the entire dataset is imported using Pandas, with over 40 rows.
  - The dataset is cleansed of erroneous inputs in columns such as date and time.
  - Datatypes across columns are fixed. Several date and time columns were in String format and had to be changed.
  - Datetime components such as year, month and time were extracted and given separate rows.
  - Features such as fire duration were engineered to enhance analysis.

- The dataset is then scrubbed for outliers and anomalies. These were addressed with measures of central tendency or removed entirely.
- Null values in date and time columns were given values from other date and time columns. For example, null values in fire start time were given values from fire reported time.

- **Load**: Finally, the cleaned DataFrame is loaded into Power BI using the Python plugin that allows the user to insert an entire script. We used our code in Jupyter Notebook by copying and pasting it into the plugin.

4. Data Modelling

4.1 Entity-Relationship Diagram



4.2 Schema Design

Describe the database schema, for example, Tables:

1) **Fire incidents (Primary)**: The primary table in this schema that reports all the fire incidents that occurred.
    a. Unique_fire_id: An engineered column that attaches a unique ID to every fire incident in the dataset using the year and fire name columns previously present in the dataset.
    b. Current_size: The size of the fire when this datapoint was added to the dataset in hectares.
    c. Size_class: A letter class given to the fire based on how large it is. Refer to appendix A for details.
    d. Fire_location_latitude: The latitude in decimal degrees of the wildfire.
    e. Fire_location_longitude: The longitude in decimal degrees of the wildfire.

f. Fire_origin: The class of land where the wildfire originated. Refer to appendix A for details.
g. Fire_start_date: The date and time when the fire started.
h. Fire_spread_rate: The rate of spread of the wildfire at the time of initial assessment, captured in metres per minute.
i. Fire_type: This is the predominant fire behaviour that was observed when the fire was originally assessed. Refer to Appendix A for details.
j. Fire_position_on_slope: The position of the wildfire relative to the slope it is travelling on at the time of initial assessment. Refer to Appendix A.
k. Fire_duration: Column engineered by subtracting fire_uc_date by fire_start_date to get a duration in datetime for the wildfire.
l. Fire_duration hrs: An engineered column providing the duration of the wildfire in hours until it was under control.
m. Region_num: A unique number given to the different regions across Alberta where wildfires occur.
n. Parsed_date: A date column parsed from fire_start_date that was initially imported as a String datatype.
o. Month Name: Name of month when wildfire incident occurred. Used for measures.
p. Year: Year when wildfire occurred. Used for measures.

2) **Fire causes:** A table that shows the causes of all wildfire incidents.
   a. Unique_fire_id: An engineered column that attaches a unique ID to every fire incident in the dataset using the year and fire name columns previously present in the dataset.
   b. General_cause_desc: Classification of the wildfire cause according to the general group, individual industry, or ignition source (for lightning) that started the wildfire. See Appendix A for details.
   c. Responsible_group_desc: For the general cause of Recreation, identify the recreational group responsible for causing the wildfire.
   d. Activity_class: Identifies the specific activity that was going on when the wildfire was started.
   e. True_cause: Identifies the specific reason why the wildfire started.

3) **Weather conditions:** A table that shows the different weather patterns for each unique wildfire.
   a. Unique_fire_id: An engineered column that attaches a unique ID to every fire incident in the dataset using the year and fire name columns previously present in the dataset.
   b. Weather_conditions_over_fire: Description of the weather conditions over the wildfire at the time of initial assessment. See Appendix A for details.
   c. Temperature C: The temperature at the wildfire site at the time of initial assessment, captured in degrees Celsius.
   d. Relative_humidity: The relative humidity at the wildfire site at the time of initial assessment.
   e. Wind_direction: The wind direction at the wildfire site at the time of initial assessment. See Appendix A.

   f. Wind_speed: The wind speed at the wildfire site at the time of initial assessment, captured in kilometers per hour.

4) **Fuel type:** A table that separates the vegetation for every fire incident.
   a. Unique_fire_id: An engineered column that attaches a unique ID to every fire incident in the dataset using the year and fire name columns previously present in the dataset.
   b. Fuel_type: The predominate fuel type (vegetation cover) in which the wildfire is burning, at the time of initial assessment. See Appendix A.
   c. Distance_from_water_source: If a helicopter with a bucket was used during initial action of the wildfire, records the distance in kilometers (to the nearest tenth) that the rotor wing flew from the water source to the wildfire.
5) **Regions:** A table that divides Alberta into regions with unique region names and numbers.
   a. Region_num: A unique number given to a region.
   b. Region: The first letter of each region parsed from the unique identifier column of wildfire incidents.
   c. Region_name: An engineered column for clarity and visualizations showing the actual full name of the region.
6) **Time dimensions:** A table that shows the different time dimensions of each fire incident using the fire_start_date column. Used mostly to create measures and date time axis in visualizations.
   a. Fire_start_date: The date and time when the fire started.
   b. Year: Year when the fire happened.
   c. Month: The month number when fire happened.
   d. Week_of_year: The number of the week within a year when the fire happened. To be used for further time-based analysis.
   e. Day_of_week: The day of the week when the fire happened.
   f. Parsed date: Date parsed from the fire_start_date column.
   g. Month Name: Name of the month when the fire occurred.
   h. Parsed Time: Time of day when the fire incident occurred, parsed from the fire_start_date column.
7) **Fire severity:** A table that shows the damage caused by the fire at different points in time.
   a. Unique_fire_id: An engineered column that attaches a unique ID to every fire incident in the dataset using the year and fire name columns previously present in the dataset.
   b. Assessment_hectares: The size of the wildfire at the time of assessment is recorded to the nearest hectare or hundredth (0.01) of a hectare.
   c. Bh_hectares: The size of the wildfire at the time the Incident Commander changed the status to Being Held (BH), recorded to the nearest hectare or hundredth (0.01) of a hectare.
   d. Uc_hectares: The size of the wildfire at the time the Incident Commander changed the status to Under Control (UC), recorded to the nearest hectare or hundredth (0.01) of a hectare.
   e. Ex_hectares: The size of the wildfire at the time the Incident Commander changed the status to Extinguished (EX), recorded to the nearest hectare or hundredth (0.01) of a hectare.

8) **Detecting agents:** A table that records who detected the fire and when.
    a. Unique_fire_id: An engineered column that attaches a unique ID to every fire incident in the dataset using the year and fire name columns previously present in the dataset.
    b. Det_agent_type: The general type of detection agent responsible for discovering the wildfire. See Appendix A.
    c. Det_agent: A more specific description of the detection agent that discovered the wildfire. See Appendix A.
    d. Discovered_date: The time the detection agent first discovered the wildfire.
    e. Parsed date: A date column parsed from the discovered_date column.
9) **Measures:** A table that calculates multiple measures using different columns from different tables.

5. Visualization & User Interface

5.1 Layout

The user interface in Power BI is divided into 5 parts:

- The project introduction and insights interface are integral to any new user attempting to understand and utilize the project. It gives a quick overview of what it is about and what the major findings are. Insights can be edited to include new data that arrives over time.
- A One Glance Overview and One-window Navigation page is included to provide high-level information about the current dataset. A sort of preliminary understanding before they dive deeper into the analysis. The One-window navigation page lets a user quickly find a particular fire/s that they are searching for.
- The rest of the 3 parts address the objective questions of where, why, and when wildfires occur.

5.2 Features & Functionalities

All pages in the Power BI dashboard have different elements that show different data points:

- **Interactive Elements:** Interactive elements in the various pages of the dashboard include:
    - Visualizations that depict categorical values like types of vegetation, causes, and more. These categories are clickable and can manipulate the data in a page to show KPIs that represent the category selected.
    - Main insights on a page get updated automatically as the dataset is changed or added to.
    - Multiple interactive maps that show different elements like causes of wildfires, regions of wildfires and more.

    - Data tables that change with user selections in the interactive maps. This is a drill-through feature that allows the user to view details of the wildfire they selected in the interactive maps.
- **Filter & Search Option:** Each page has a set of filters that are relevant to the objective being addressed on that page:
    - Regions filters are one of the most important filters across most pages.
    - Date filters that change the KPIs of a page. They also change the wildfires shown on interactive maps.
    - Weather and geospatial filters if the user is searching for isolated wildfires.
    - Clickable button filters for causes and types of vegetation.
- **Navigation:** The dashboard has been set up in an order that takes the user through a journey:
    - The user goes through the overview pages and dives into dashboards addressing objectives.
    - For the most part, filters are kept at the top of each page while KPIs are placed to the right or left corners for visibility.

6. Technical Requirements

6.1 Software & Tools

The following tools and software were used to develop this project:

- Jupyter Notebook: Used to import, transform, and clean the data.
- Power BI: Used to load the data and visualize the findings and create a dashboard.
- Python: The programming language used to transform and load the data.
- Pandas/Numpy: Python libraries used to transform the data within Jupyter Notebook.

6.2 Hardware Requirements

The following hardware requirements are the minimum standard to view the dashboard without issues:

- Multicore processor (at least 4 cores with hyperthreading)
- At least 16 GB of RAM to handle interactive features
- An integrated or dedicated GPU to load visuals and maps

7. Milestones & Timeline

- **Finished understanding the dataset**: [2025/02/06]
- **Cleaned and transformed the data for analysis**: [2025/02/07]
- **Completed overview pages and addressed first objective on Power BI**: [2025/02/08]
- **Completed other objectives pages on Power BI**: [2025/02/09]

8. Conclusion

This dataset was rather extensive and holds significance in Alberta's geographic, economic, and political landscape. Wildfires are a part of the lives of Albertans and studying/understanding them to a greater extent is key to reducing the adverse impact they have.

This project started off with a rather ambitious scope well beyond the boundaries stated at the beginning of this document. A lot of insight has been gathered from this analysis and dashboard but there is more yet to be done. As mentioned, the team is already working on implementing certain machine learning algorithms to fill gaps in the dataset. This will be an ongoing task that will take a deeper understanding of the domain and dataset.

The goal is to set up dashboards that utilize statistical inferencing and machine learning to aid in the detection and/or prevention of wildfires.

Appendix A

Fuel types refer to the type of vegetation found in a particular area:

Coniferous

C-1 Spruce-Lichen Woodland
C-2 Boreal Spruce
C-3 Mature Jack or Lodgepole Pine
C-4 Immature Jack or Lodgepole Pine

Slash

S-1 Jack or Lodgepole Pine slash
S-2 White Spruce-Balsam slash

Mixedwood

M-1 Boreal Mixedwood-Leafless
M-2 Boreal Mixedwood-Green

Deciduous

D-1 Leafless Aspen

Grass

O-1a Matted Grass
O-1b Standing Grass

0 Other fuel type/Unknown