GlobalBlaze: Analyzing Global Trends in Wildfire Burned Areas and Emissions(2002-2023)

---

1. Introduction

Wildfires are one of the natural disasters with wide range of impacts on human health, ecosystems and the climate. It is important to understand the trends and patterns of the wildfires and emission gases released from that is crucial for effective wildfire management and environmental protection. This project aims to analyse the global trends in wildfire burned areas and the emissions from 2002 to 2023 almost 2 decades data.

1.1. Main Analytical Questions:

1. What are the long-term trends in wildfire burned areas and emissions globally, and are there significant changes over time?
2. How do seasonal variations affect wildfire activities and emissions, and are there specific months or quarters with higher prevalence or levels?
3. Which countries or regions experience the most wildfires, and what factors contribute to their prevalence?
4. Is there a correlation between wildfire area burned and emissions gases?

---

2. Datasets

The Project utilizes two datasets from the Global Wildfire Information System, which includes historical data from 2002 to 2023. These datasets are in the form of Tabular format(CSV) and licensed under Open Data Source. Moreover, it's noted that these datasets are well-maintained, with minimal missing values, ensuring data integrity and reliability for analysis. Python is chosen over Jayvee due its extensive library ecosystems, providing efficient data handling , analysis and integration.

2.1. Global Monthly Burned Area Dataset(GMBAD) [2002-2023]

- Metadata URL:https://gwis.jrc.ec.europa.eu/apps/country.profile/downloads
- Data URL:https://effis-gwis-cms.s3.eu-west-1.amazonaws.com/apps/country.profile/MCD64A1_burned_area_full_dataset_2002-2023.zip
- Data Type: Zipped CSV
- License: Creative Commons Attribution 4.0 International

The Dataset contains monthly burned area data(in hectares) globally, representing countries affected by wildfires. It contains attributes such as the year, month, country , region , forest, savannas, shrubs_grasslands, croplands and other burned areas.

2.2. Global Monthly Emission Dataset(GMED) [2002-2023]

- Metadata URL:https://gwis.jrc.ec.europa.eu/apps/country.profile/downloads

- Data URL:https://effis-gwis-cms.s3.eu-west-1.amazonaws.com/apps/country.profile/emission_gfed_full_2002_2023.zip
- Data Type: Zipped CSV
- License: Creative Commons Attribution 4.0 International

The Dataset contains monthly emitted gases data (in tons) globally, representing countries affected by wildfires. It provides information such as the year, month, country , region , and various emission gases such as CO2:Carbon Dioxide, CO: Carbon Monoxide, TPM: Total Particulate Matter, PM25: Particulate Matter less than 2.5 micrometres in diameter, TPC: Total Particulate Count, NMHC: Non-Methane Hydrocarbons, OC: Organic Carbon, CH4: Methane. SO2: Sulphur Dioxide, BC: Black Carbon, NOx: Nitrogen Oxides.

---

3. Data Pipeline

The data pipeline we've built is essential for our project's goal of analysing wildfire and emissions data to uncover insights. This section gives a thorough overview of the pipeline, explaining its main parts, the tools used, the changes made to the data, and how we deal with problems or changes in the data's format.

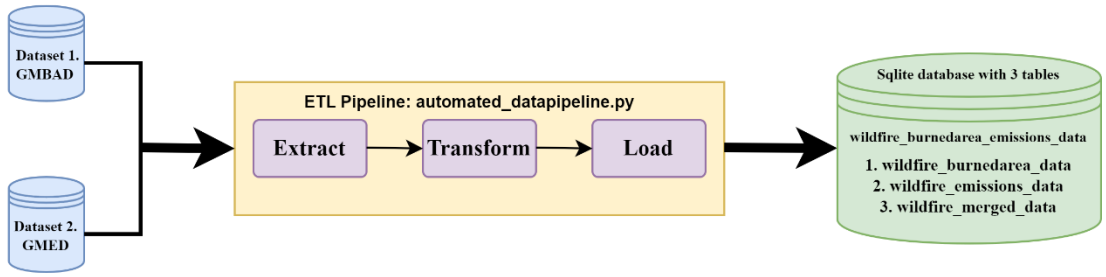

Figure 1: ETL Pipeline Flow Chart

Data Pipeline Overview:

The data pipeline follows a classic Extract, Transform, Load (ETL) paradigm, encompassing the following stages:

3.1. Extraction:

Data extraction involves retrieving wildfire and emissions datasets from remote sources. Utilizing Python's requests library, the pipeline downloads ZIP files containing the required data from specified URLs. Upon retrieval, the pipeline utilizes the zip file module to extract the contents of the ZIP files into memory, preparing the data for subsequent processing.

3.2. Transformation:

The transformation phase encompasses various cleaning and manipulation operations aimed at preparing the raw data for analysis. For wildfire data, irrelevant columns are dropped, rows with zero burned area values are removed, columns are renamed for consistency, and a new column representing quarters is created. Similarly, emissions data undergoes similar cleaning and transformation steps to ensure data integrity and consistency. These transformations are crucial for standardizing the data and making it conducive to meaningful analysis.

3.3 Loading:

The transformed data is then loaded into pandas Data Frames, facilitating efficient data manipulation and analysis. Following transformation, the pipeline merges the wildfire and emissions datasets into a unified Data

Frame based on common attributes such as year, month, and country. The merged Data Frame then persisted into a SQLite database, allowing for easy storage, retrieval, and querying of the processed data.

4.Results and Limitations

4.1. Chosen Data Format for the Output of the Pipeline

The pipeline output is saved in a database ensures data integrity, enabling atomic operations and providing robust security measures. It facilitates seamless querying and analysis alongside structured data, leveraging the database's scalability and performance optimizations, while also simplifying data backup and recovery processes.

	Year	Month	Quarter	Country_Code	Country_Name	Region_Name	Forest_BA	Savannas_BA	Shrubs_Grasslands_BA	Croplands_BA	Other_BA
0	2002	1	Q1	AGO	Angola	Huíla	0.0	0.000	407.851	0.0	0.000
1	2002	1	Q1	AGO	Angola	Lunda Norte	0.0	64.398	42.932	0.0	0.000
2	2002	1	Q1	AGO	Angola	Malanje	0.0	0.000	21.466	0.0	0.000
3	2002	1	Q1	AGO	Angola	Moxico	0.0	279.056	21.466	0.0	64.398
4	2002	1	Q1	AGO	Angola	Namibe	0.0	0.000	236.125	0.0	0.000

Table 1: Sample Output data table for GMBAD dataset

	Year	Month	Quarter	Country_Code	Country_Name	Region_Name	CO2	CO	TPM	PM25	TPC	NMHC	OC	CH4	SO2	BC	NOx
0	2002	1	Q1	AFG	Afghanistan	Jawzjan	139.786	5.549	0.738	0.591	0.250	0.333	0.216	0.191	0.039	0.034	0.319
1	2002	1	Q1	AFG	Afghanistan	Kunar	14828.511	554.090	74.758	63.061	26.385	29.903	23.043	17.062	4.222	3.254	34.301
2	2002	1	Q1	AFG	Afghanistan	Takhar	492.090	31.668	3.850	1.944	0.947	3.074	0.714	1.807	0.124	0.233	0.966
3	2002	1	Q1	AGO	Angola	Huíla	16875.859	630.593	85.080	71.767	30.028	34.032	26.225	19.418	4.805	3.703	39.037
4	2002	1	Q1	AGO	Angola	Moxico	20483.095	765.383	103.266	87.108	36.447	41.306	31.830	23.569	5.831	4.495	47.381

Table 2: Sample output data table for GMED dataset

4.2. Data Quality

- **Accuracy:** highly accurate datasets from GWIS with consistent records from 2002 to 2023.
- **Completeness:** complete datasets with no gaps in global wildfire burned area and emission records.
- **Consistency:** consistent data with standardized formats for each month and country.
- **Uniqueness:** unique records without duplicated records ensuring reliable analysis.
- **Relevance:** the datasets are directly relevant to the project's objectives.

4.3. Limitations

- **Global Monthly Burned Area Dataset(GMBAD)** [2002-2023]: The dataset is of high quality with no missing values. However, the dataset contains zero values with the burned area attributes.

- **Global Monthly Emission Dataset(GMED)** [2002-2023] The dataset had missing values that were filled with zeros or 'Unknown', which could introduce bias or inaccuracies. Additionally, The dataset may underreport economic and human impacts due to incomplete records or reporting discrepancies, limiting the comprehensiveness of the analysis.
