**Analysis of Dataset, Use Cases and Function Calling**

**Dataset Name: Emissions from Drained Organic Soils**

**Dataset URL:** https://datasource.kapsarc.org/explore/dataset/emissions-from-drained-organic-soils/information/

**Can we access Dataset?** Yes, it is publicly available (Public Domain License).

**Dataset Topic:** This dataset provides information on nitrous oxide (N2O) and carbon dioxide (CO2) emissions associated with the drainage of organic soils for agricultural use (cropland and grassland). The emissions data are available by country and computed geospatially using IPCC guidelines. The dataset covers the period from 1990 to 2021.

**Type of Data Available (Chart/Values/...):**
Values (numerical emissions data).

**Is it live data? Frequency of data updates?**
No, this is not live data. The dataset extends only to 2021 and is updated annually.

**How can we access the data (Json/Download/View in site)?**
Data can be accessed through:

- Viewing directly on the website.
- Download options (e.g., CSV, Excel).
- API calls.

**Dataset Pre-processing Need? List any Steps:**

Handle missing data, if any.

Convert the timestamp to a human-readable format if needed.

Dropping unnecessary columns for analysis.

**Data Summary:**

- Time Period: 1990 – 2021.
- Granularity: Country-level data.
- Rows and Columns: The dataset contains 60,077 rows and 16 columns. It is not live data and is discontinued, organized in a tabular format for easy manipulation and analysis.

**Data Structure:**

**Rows:** There are multiple entries per country per year, resulting in a time-series structure.

**Columns:** The dataset consists of the following columns:

**Year:** Year of recorded data (e.g., "2010").

**Domain Code:** Code representing the domain (e.g., emissions from drained organic soils).

**Area Code (M49):** Standard country code as per the M49 classification.

**Area:** Name of the country or region (e.g., "Canada").

**Element Code:** Code for the type of greenhouse gas emitted.

**Element:** Description of the emission element (e.g., "Emissions from drained organic soils").

**Item Code:** Specific item related to emissions (e.g., "N2O", "CO2").

**Item:** Detailed description of the item (e.g., "Nitrous Oxide", "Carbon Dioxide").

**Year Code:** Code for the year of the data entry.

**Source Code:** Code identifying the data source.

**Source:** Name of the source of the data (e.g., "FAO").

**Unit:** Standard measuring unit for the greenhouse gas emitted (e.g., "kt" for kilotonnes).

**Value:** Amount of greenhouse gas emitted in the specified unit.

**Flag:** Any flags associated with the data (e.g., indicating estimations).

**Flag Description:** Explanation of the flags used in the dataset.

**Note:** Additional notes or comments related to the data entry.

**CRUD Operations**

**1.Create**

Add new records (e.g., emissions data for new countries).

Upload new dataset versions.

**2.Read**

Query specific records (e.g., by country or year).

Analyze data (trends, statistics).

Export data in various formats (e.g., CSV, JSON).

**3.Update**

Modify existing records (e.g., correct emissions values).

Update metadata (e.g., description, keywords).

**4.Delete**

Remove inaccurate or irrelevant records.

Archive older dataset versions if replaced.

This framework helps maintain the dataset's accuracy and usability.

**Supported Data**

**1.Country Demographics**

Population data, GDP, and land use statistics to provide context for emissions data.

**2.Agricultural Practices**

Data on types of crops grown and agricultural practices used in different countries, which can impact emissions from drained organic soils.

**3.Climate Data**

Historical climate data (temperature, precipitation) to analyze correlations between climate variables and emissions.

**4.Soil Types and Quality**

Information on soil composition and quality in different regions, which can affect organic soil drainage and emissions.

**5.National Greenhouse Gas Inventories**

Data from national inventories submitted to the UNFCCC, which could provide more detailed emissions breakdowns.

**6.Policy and Regulation Data**

Information on environmental policies, regulations, and initiatives aimed at reducing greenhouse gas emissions in agriculture.

**7.Land Use Changes**

Historical data on land use changes (e.g., deforestation, urbanization) that may influence organic soil drainage and emissions.

**8.Research Studies**

Academic papers and reports on organic soil management, greenhouse gas emissions, and mitigation strategies.

**9.Regional Emissions Data**

Data from regional organizations or initiatives that focus on emissions reductions in specific areas or sectors.

**10.Economic Data**

Economic impact assessments related to agriculture and emissions, which can provide insights into the trade-offs of different practices.

**Key Features**

**1.Comprehensive Coverage:**

Data includes emissions from 1990 to 2021 for various countries worldwide.

**2.Multiple Greenhouse Gases:**

Emissions data for both Nitrous Oxide ($N_2O$) and Carbon Dioxide ($CO_2$).

**3.Country-Level Detail:**

Granular data allows for analysis by individual country, facilitating comparisons and trends.

**4.Standardized Codes:**

Uses standardized area codes (M49) for consistency in country identification.

**5.Extensive Record Count:**

Contains 60,077 rows, providing a rich dataset for analysis.

**6.Variety of Metadata:**

Includes multiple columns of metadata (16 in total), offering context and details about emissions.

**7.Public Domain License:**

Freely available for use and redistribution, encouraging widespread access and research.

**8.Historical Context:**

Enables analysis of trends over time, contributing to understanding the impacts of agricultural practices on emissions.

**9.Facilitates Research:**

Supports environmental research, policy-making, and sustainability initiatives by providing essential data on soil emissions.

**10.Data Organization:**

Organized in a tabular format for easy manipulation and analysis, compatible with various data analysis tools.

These key features make the dataset a valuable resource for researchers, policymakers, and anyone interested in the environmental impact of agricultural practices.

**Example Queries**

**1.Basic Query**

Description: Retrieve overall emissions data across all countries, substances, and years.

**2.Country-Specific Query**

Description: Retrieve emissions data for a specific country (e.g., "Brazil").

**3.Year-Specific Query**

Description: Filter emissions data for a specific year (e.g., 2018).

**4.Substance-Specific Query**

Description: Aggregate emissions data for a specific substance (e.g., "CO2").

**5.Combination of Country, Year, and Substance**

Description: Show N2O emissions for a specific country (e.g., "Canada") between given years (e.g., 2015 to 2020).

**6.Comparison Query**

**Description:** Compare emissions between two countries (e.g., "Brazil" and "Argentina") for a specified time period.

**7.Highest Emissions Query**

Description: List areas with the highest CO2 emissions over a defined time range (e.g., 2016 to 2021).

**8.Handling Missing Data**

Description: Check for missing data and handle cases where data is unavailable or invalid.

**Overview of the Organic Soil Emissions Tool**

The Organic Soil Emissions Tool is designed to provide users with comprehensive insights into greenhouse gas emissions (CO2 and N2O) associated with cropland organic soils. Leveraging detailed datasets, this tool allows for in-depth analysis, trend identification, and comparison of emissions data across various geographic regions and time periods. Users can interact with the tool to uncover critical information that can aid in environmental assessments and policy-making decisions.

This document outlines five key use cases for the Organic Soil Emissions Tool, showcasing its capabilities in tracking and analyzing emissions data.

**Use Cases**

**1.Emissions Trend Analysis**

**Use Case**: Analyze the trend of $CO_2$ and $N_2O$ emissions over a specified period for a specific country or area to identify emissions patterns.

**Function**: calculate_trends(area, start_year, end_year)

**Description**: This function filters data for the specified area and time range, calculates annual emissions, and identifies the trend as "increasing," "decreasing," or "stable."

**Return Value**: A dictionary with the area name, yearly emissions data, and the trend (e.g., increasing or decreasing) for $CO_2$ and $N_2O$ emissions.

**2.Area-Specific Emissions Data**

**Use Case**: Retrieve $CO_2$ or $N_2O$ emissions data for a specific area over a defined time range for focused analysis.

**Function**: get_summary_data(area, start_year, end_year)

**Description**: Filters emissions data for a particular area and time range. Aggregates $CO_2$ and $N_2O$ emissions to provide a summary of emissions data for the specified region and period.

**Return Value**: A dictionary with area, emission type ($CO_2$ or $N_2O$), and total emissions over the selected date range.

**3.Comparison of Emissions Between Areas**

**Use Case**: Compare emissions between two or more areas to identify differences in emissions levels over a specific time range.

**Function**: compare_emissions(areas, start_year, end_year)

**Description**: Filters data for the specified areas and date range, then aggregates emissions values for $CO_2$ and $N_2O$. Provides a comparison between the selected areas over the defined period.

**Return Value**: A dictionary with area names and respective cumulative emissions data for $CO_2$ and $N_2O$.

**4. Highest Emissions Ranking**

**Use Case**: Identify the areas with the highest $CO_2$ or $N_2O$ emissions over a specified period.

**Function**: list_highest_emissions(element, start_year, end_year)

**Description**: Filters the dataset by emissions type ($CO_2$ or $N_2O$) and the specified time range, calculates total emissions for each area, and ranks areas by their emissions values.

**Return Value**: A dictionary with area names and emissions values ranked from highest to lowest for the specified emission type and period.

**5. Emission Alerts**

**Use Case**: Set an alert for emissions exceeding a specific threshold in a particular area.

**Function**: set_emission_alert(area, element, threshold)

**Description**: Monitors emissions data for a specific area and emission type ($CO_2$ or $N_2O$). If emissions exceed the defined threshold, the function triggers an alert.

**Return Value**: A dictionary indicating whether the threshold was exceeded, including details of the emissions level and threshold for comparison.

**6.Identifying Missing Data Points**

**Use Case**: Identify missing years of emissions data for $CO_2$ or $N_2O$ for a specific area, supporting data completeness checks.

**Function**: find_missing_data(area, element)

**Description**: Filters the dataset by area and emission type, then identifies years where data is missing for the specified emission type.

**Return Value**: A dictionary listing the missing years of data for $CO_2$ or $N_2O$ for the specified area.

**7.Forecasting Future Emissions**

**Use Case**: Project future emissions levels for $CO_2$ or $N_2O$ in a specific area based on historical data trends.

**Function**: long_term_forecast(area, element, projection_years)

**Description**: Uses an ARIMA model to forecast emissions for the specified area and emission type ($CO_2$ or $N_2O$) based on historical data.

**Return Value**: A dictionary with the projected emissions values for each forecasted year, enabling long-term emissions planning.

**8.Emissions Correlation Analysis**

**Use Case**: Determine the correlation between $CO_2$ and $N_2O$ emissions for a specific area, helping to reveal patterns between these emissions.

**Function**: analyze_correlation(area, start_year, end_year)

**Description**: Filters data for the specified area and time range, then calculates the correlation coefficient between $CO_2$ and $N_2O$ emissions.

**Return Value**: A dictionary with the correlation coefficient, indicating the strength and direction of the relationship between $CO_2$ and $N_2O$ emissions in the selected area and time frame.

**9.Cumulative Emissions Calculation**

**Use Case**: Calculate total cumulative emissions for $CO_2$ or $N_2O$ in a specific area over a selected time range.

**Function**: cumulative_emissions(area, element, start_year, end_year)

**Description**: Filters emissions data for a specific area and emission type ($CO_2$ or $N_2O$) within a specified date range, then calculates the total cumulative emissions.

**Return Value**: A dictionary with the cumulative emissions value for the specified area, element, and date range.

**10. Use Case: Compare $CO_2$ emissions across multiple countries with specific conditions.**

**Function:** compare_emissions_with_conditions(areas, element, item, source, unit, start_year, end_year)

**Description:** Filters emissions data for specified countries, element type ($CO_2$), item (e.g., "Cropland organic soils"), source (e.g., "FAO TIER 1"), and unit (kt) over a defined time range. Aggregates cumulative emissions for each country to provide a comparative summary.

**Return Value:** A dictionary with the selected countries as keys and their respective cumulative emissions as values. Additional metadata includes the element, item type, source, unit, and the date range for the comparison.

**Setting Up the Organic Soil Emissions Tool and Performing Integration and Unit Testing**

**1. Load Files into the Environment**

Extract the contents of the emissions-from-drained-organic-soils zip folder into your development environment. Ensure that the following files are included:

custom_tools.py

unit_test_emissionorganic.py

organic_soil_emissions_tool.py

**2. Install Required Dependencies**

Open your terminal or command prompt and navigate to the directory containing your files. Install the necessary dependencies by referring to the requirements.txt file that should be included in your project.

**3. Install the Llama Stack Client**

Since the Organic Soil Emissions Tool utilizes the SingleMessageCustomTool from the CustomTool class, ensure that the Llama Stack Client is installed in your environment.

**4. Run the Custom and Organic Soil Emissions Tool Scripts**

Execute the scripts to initialize the necessary functions. This step ensures that your emissions tool is loaded and ready for use.

**5. Run the Unit Test Script**

After confirming that the main Organic Soil Emissions Tool script runs correctly, proceed to run the unit test script. This will validate the individual functions of your tool.

**Section 3: Datasets**

**3.1 Overview**

**Emissions from Drained Organic Soils**

**Topic:** This dataset provides emissions data on carbon dioxide ($CO_2$) and nitrous oxide ($N_2O$) associated with the drainage of organic soils for agricultural purposes, specifically for cropland and grassland. The data is computed geospatially using IPCC guidelines, offering insights into emissions across different countries. This dataset aids in understanding emissions trends related to soil management practices and their environmental impacts.

**Time Period: 1990 – 2021**

Data Structure: This dataset is structured at the country level with multiple entries per country for each year, forming a time-series format. It contains 60,077 rows and 16 columns, allowing for a detailed analysis of emissions by year and region. The columns include identifiers like "Domain Code," "Area Code (M49)," and "Year Code," along with metrics on emissions by type (e.g., "Emissions from drained organic soils").

**Access and Format:** The dataset is publicly available under a Public Domain License. It can be accessed:

By viewing directly on the website.

Through download options in CSV and Excel formats.

Via API calls for integration with analysis tools.

**Use Cases:** This dataset supports a range of use cases, including:

Tracking and analyzing $CO_2$ and $N_2O$ emissions trends for cropland and grassland.

Comparative emissions analysis across countries.

Sector-specific emissions reporting.

Environmental impact assessments related to agricultural soil management.

Data Pre-processing: Minimal pre-processing is required. Necessary steps include:

Handling any missing data.

Converting timestamps to a human-readable format if needed.

Dropping unnecessary columns for focused analysis.

**Source:** KAPSARC.

**3.2 Field Descriptions**

Here is the detailed understanding of the fields present in each of the revised datasets:

**Dataset 2: Emissions from Drained Organic Soils**

**Year:** Year of recorded data (e.g., "2010").

**Domain Code:** Code representing the domain (e.g., emissions from drained organic soils).

**Area Code (M49):** Standard country code as per the M49 classification.

**Area:** Name of the country or region (e.g., "Canada").

**Element Code:** Code for the type of greenhouse gas emitted.

**Element:** Description of the emission element (e.g., "Emissions from drained organic soils").

**Item Code:** Specific item related to emissions (e.g., "N2O", "CO2").

**Item:** Detailed description of the item (e.g., "Nitrous Oxide", "Carbon Dioxide").

**Year Code:** Code for the year of the data entry.

**Source Code:** Code identifying the data source.

**Source:** Name of the source of the data (e.g., "FAO").

**Unit:** Standard measuring unit for the greenhouse gas emitted (e.g., "kt" for kilotonnes).

**Value:** Amount of greenhouse gas emitted in the specified unit.

**Flag:** Any flags associated with the data (e.g., indicating estimations).

**Flag Description:** Explanation of the flags used in the dataset.

**Note:** Additional notes or comments related to the data entry

### 3.3 Data Context

**Dataset: Emissions from Drained Organic Soils**

The "Emissions from Drained Organic Soils" dataset provides emissions data for nitrous oxide ($N_2O$) and carbon dioxide ($CO_2$) specifically tied to the drainage of organic soils for agricultural activities such as cropland and grassland. Spanning from 1990 to 2021, this dataset allows for a deep analysis of emissions related to soil management practices at a country level, offering valuable insight into how land-use changes in agriculture contribute to greenhouse gas emissions. The data was collected and computed geospatially using IPCC (Intergovernmental Panel on Climate Change) guidelines, ensuring standardized and reliable measurements across different regions.

In the context of global climate change and emissions reduction efforts, this dataset serves as an essential resource for understanding the environmental impacts associated with agricultural soil management. As countries strive to meet emissions reduction targets under agreements like the Paris Agreement, this dataset supports the assessment of emissions trends from soil management, which is often an overlooked contributor to national emissions inventories. By providing detailed, country-specific data, the dataset enables policymakers, researchers, and environmental organizations to track emissions from soil drainage and develop effective mitigation strategies that address emissions from agriculture and land use.

The dataset's sectoral focus on cropland and grassland organic soils helps identify specific agricultural practices and regions contributing most to greenhouse gas emissions. This information is critical for developing tailored strategies to reduce emissions, such as improved soil management techniques, changes in land-use policies, and sustainable agriculture practices. Moreover, the dataset supports the monitoring of historical emissions trends, helping to evaluate the long-term environmental impacts of soil drainage and providing a basis for predictive modeling and policy planning for sustainable land management practices.

### 3.4 Data Conditioning

**Dataset: Emissions from Drained Organic Soils**

**Handling Missing Values:** Missing or null values were examined, particularly in fields like "Flag" and "Note," which provide context or indicators about data reliability. Where appropriate, missing values were either ignored or filled with placeholders to maintain data structure consistency.

**Standardization:** Standardized country names and region codes using M49 classification to ensure consistency across analyses. This allows for easier integration with other datasets that use standardized country codes.

**Date Formatting:** Converted all year entries to a consistent integer format to streamline temporal analysis. Although dates are provided in year format only, all entries were standardized to ensure consistency.

**Data Type Consistency:** Ensured all relevant columns were in the correct format: "Year" as an integer, "Value" as a float (representing emissions), and categorical fields such as "Area" and "Element" as strings. This step helps maintain data integrity during processing and analysis.

**Unit Consistency:** Standardized emissions units to kilotonnes (kt) across the dataset for both $CO_2$ and $N_2O$ emissions. This uniformity ensures consistency when aggregating or comparing data across different areas and timeframes.

**Outlier Detection:** Reviewed emissions values for extreme outliers, especially for unexpected negative values or sudden spikes, which could indicate data entry errors or anomalies in the collection process. Outliers were flagged for further review and validation where necessary.

### 3.5 Data Quality Assessment

**Emissions from Drained Organic Soils**

**Completeness:** The dataset provides extensive emissions data for $CO_2$ and $N_2O$ associated with the drainage of organic soils, covering a wide timeframe from 1990 to 2021. Data entries are available for multiple countries and years, though occasional missing values may exist, particularly in the "Value" column, which might require attention during analysis to ensure full temporal coverage.

**Uniqueness:** Each entry is uniquely defined by a combination of columns, including country, year, element type ($CO_2$ or $N_2O$ emissions), item type (cropland or grassland organic soils), and source. Duplicate entries are minimal, ensuring that each record represents distinct emissions data for a specific location, emission type, and time period.

**Accuracy:** The dataset is sourced from reliable databases that follow IPCC guidelines for emissions calculation, which adds credibility and accuracy. However, minor deviations may arise from differences in reporting standards across countries or estimation methods used for emissions calculations. Cross-referencing with other emissions datasets could provide an additional accuracy check.

**Atomicity:** Each data record is atomic, representing a single emissions measurement for a particular country, year, element type, and item. Each row stands alone and does not combine multiple data points, which ensures high granularity and facilitates targeted analysis of emissions data by country, year, or emission type.

**Conformity:** The dataset generally conforms to standard data formats for country names, emission elements, units (e.g., kt for kilotonnes), and item types (e.g., cropland or grassland organic soils). Some minor conformity adjustments may be required, such as standardizing the names in the "Area" or "Source" columns, to maintain uniformity across all entries.

**Overall Quality:** The dataset is of good quality and is well-suited for detailed emissions analysis related to organic soils across countries and years. While completeness and conformity could benefit from minor enhancements, the dataset remains robust and valuable for tracking and analyzing greenhouse gas emissions from drained organic soils, providing insights that support climate and agricultural policy assessments.