Data Report

Question

How does the trend in global mean surface seawater acidity correlate with greenhouse gas emissions from agriculture over the past two decades?

Data Sources

Data Set 1: Global Mean Surface Seawater Acidity

- **Source:** Eurostat, available at Link

- **Reason for Selection:** This dataset provides comprehensive and updated information on seawater acidity, which is crucial for studying oceanic environmental changes.

- **Data Content:** Annual measurements of the global mean surface seawater pH levels, reflecting acidity trends over time.

- **Data Structure and Quality:** Structured in a yearly time series format, the data is high quality, collected and verified by reputable scientific institutions.

Data Set 2: Greenhouse Gas Emissions from Agriculture

- **Source:** Eurostat, available at Link

- **Reason for Selection:** This dataset offers detailed information on greenhouse gas emissions specific to agriculture, essential for analyzing environmental impacts.

- **Data Content:** Annual greenhouse gas emissions from agricultural activities, broken down by types of gases (e.g., methane, nitrous oxide).

- **Data Structure and Quality:** Presented in a time series format, the data is detailed and reliable, following stringent reporting and verification standards.

Licenses

- **Eurostat Data:** Typically available under public domain or open-data licenses, permitting use with proper attribution.

- **Compliance Plan:** All data will be cited appropriately, adhering to the terms of use specified by Eurostat.

Data Pipeline

Technology Used

- **Tools:** Jayvee for Pipeline.
- **Data Storage:** SQLite DB with each table for the input.

Transformation and Cleaning Steps

- **Cleaning:** Removed Unwanted Data and applied Proper Data Type.
- **Data Merging:** Create 2 Separate pipelines for the inputs and save data into one DB File to allow easy Normalization and Querying.

Problems Encountered

- **Inconsistent Timeframes:** Resolved by aligning datasets to common time periods.
- **Jayvee:** Adding New identity columns for mapping tables and indexing

Pipeline Flow