Title: The Impact of CO2 and Greenhouse Gas Emissions on Sea-Level Rise

Main Question
How do rising CO2 and Greenhouse Gas Emissions correlate with changes in mean sea levels?

Data sources

1. Data on CO2 and Greenhouse Gas Emissions by Our World in Data
    - **Metadata URL:** https://climatedata.imf.org/datasets/9c3764c0efcc4c71934ab3988f219e0e/explore
    - **Data URL:** https://opendata.arcgis.com/datasets/9c3764c0efcc4c71934ab3988f219e0e_0.csv
    - **Data Type:** CSV
    - **Description:** This comprehensive CO2 and Greenhouse Gas Emissions dataset, curated by Our World in Data, offers a wide range of essential metrics.
    - **Data Quality:**

Dimensions	
Accuracy	**True**
Completeness	**True**
Consistency	**True**
Timeliness	**True**
Relevancy	**True**

    - **License**
      All visualizations, data, and code created by Our World in Data are fully open access under the Creative Commons BY license.

2. Global Average Absolute Sea Level Change, 1880-2014 from the US Environmental Protection Agency using data from CSIRO, 2015; NOAA, 2015.
    - **Metadata URL:** https://climatedata.imf.org/datasets/b84a7e25159b4c65ba62d3f82c605855/explore
    - **Data URL:** https://opendata.arcgis.com/datasets/b84a7e25159b4c65ba62d3f82c605855_0.csv
    - **Data Type:** CSV
    - **Description:** This data contains "cumulative changes in sea level for the world's oceans since 1880, derived from a combination of long-term tide gauge measurements and recent satellite data. It presents the average absolute sea level change, which indicates the height of the ocean surface irrespective of land movement.
    - **Data Quality:**

Dimensions	
Accuracy	**True**

Dimensions	
Completeness	**True**
Consistency	**True**
Timeliness	**False**
Relevancy	**False**

- **License:**
  This is free and unencumbered software released into the public domain. (Details)

Reason:
I have selected these datasets for a comprehensive and focused study on the impact of CO2 and greenhouse gas emissions on sea-level rise. These datasets offer essential information to analyze and understand the correlations between increasing CO2 levels and their real-world effects, making them ideal for my project.

High-Level Description

- **Data Retrieval**: Downloads two datasets from given URLs.
- **Data Transformation**: Filters and merges the datasets based on specific criteria.
- **Data Saving**: Saves the transformed data to a CSV file.

Technologies Used

- **Python**: The entire pipeline is implemented in Python.
- **Pandas**: Used for data manipulation and transformation.
- **Requests**: Used to download data from URLs.

Transformation and Cleaning Steps

1. **Filtering Data1**: The data from the first CSV (CO2 data) is filtered to include only rows where the country is "World".
2. **Selecting Columns from Data1**: Only the "year", "co2", and "co2_growth_prct" columns are retained from the first dataset.
3. **Selecting Columns from Data2**: Only the "Year" and "CSIRO Adjusted Sea Level" columns are retained from the second dataset.
4. **Merging Datasets**: The two datasets are merged on the "year" (from the first dataset) and "Year" (from the second dataset) columns.
5. **Dropping Redundant Columns**: The redundant "Year" column from the merged dataset is dropped.

Reason for Transformations

These transformations are performed to align and integrate data from two different sources into a single cohesive dataset that can be used for analysis.

Problems Encountered

- **Column Name Mismatches**: The datasets have different column names for the year, which need to be aligned for merging.
- **Data Downloading Issues**: Ensuring the files are downloaded correctly and saved before reading.

Solutions

- **Column Renaming or Mapping**: Mapped "year" from the first dataset to "Year" in the second dataset during the merge operation.
- **Error Handling in Downloads**: Added code to write the downloaded content to files before reading them with Pandas, ensuring that the data is correctly downloaded and saved.

The Output Data

The output data of the pipeline is a CSV file containing combined and transformed data from two different sources. Specifically, the data includes:

- **year:** The year of the data point.
- **co2:** The CO2 emissions for the "World".
- **co2_growth_prct:** The percentage growth of CO2 emissions for the "World".
- **CSIRO Adjusted Sea Level:** The CSIRO adjusted sea level data.

This data is created by merging the CO2 emissions data (filtered for the "World") and the CSIRO-adjusted sea level data on the year column.

Data Structure

- The final dataset is a Pandas DataFrame that is saved as a CSV file.
- It has four columns: year, co2, co2_growth_prct, and CSIRO Adjusted Sea Level.

Data Format

The output of the pipeline is saved as a CSV file. This format is chosen because it is widely supported, easy to read, lightweight, and efficient to store and transfer.

Potential Issues

- **Data Mismatch**: There may be years present in one dataset and not in the other, leading to missing data after the merge.
- **Different Timeframes**: If the two datasets cover different time periods, this could limit the analysis to the common years available in both datasets.

Improvements

- **Documentation**: Thorough documentation of the data sources, transformation steps, and assumptions made can help in understanding and justifying the final analysis.
- **Regular Updates**: If the source data is regularly updated, the pipeline should be set to run periodically to incorporate new data.