Data Report for Project: "How has the adoption of CO2 emissions by sectors in the USA affected global temperatures?"

Question

What is the impact of sector-specific CO2 emissions in the USA on global surface temperatures?

Data Sources

1. Sector CO2 Emissions in the USA

Source: Kaggle - US CO2 Emissions by Sector
https://www.kaggle.com/datasets/alistairking/u-s-co2-emissions

Description: This dataset contains annual CO2 emissions data for the USA, broken down by various sectors such as transportation, energy, industry, and agriculture from 1970 to recent years. The data includes the amount of CO2 emitted (in million metric tons) by each sector.

Structure and Quality:

- **Columns**: Year, Sector, State, fuel-name, CO2 Emissions value

```
df = pd.read_csv(filepath_or_buffer='../data/GlobalLandTemperaturesByState.csv')
```

	year	state-name	sector-name	fuel-name	value
1	1970	Alabama	Industrial carbon dioxide emissions	Coal	26.721507
2	1970	Alabama	Industrial carbon dioxide emissions	Petroleum	3.577779
3	1970	Alabama	Industrial carbon dioxide emissions	Natural Gas	8.944097

- **Quality**: The data is clean and well-structured, with clear labels and no missing values.

License: The Creative Commons Attribution 4.0 International (CC BY 4.0) license. Public Domain.
https://www.usa.gov/government-copyright

2. Global Surface Temperature Data

Source: Kaggle - Climate Change: Earth Surface Temperature Data
https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

Description: This dataset provides global surface temperature data, including monthly average temperatures for various countries and regions from the early 19th century to recent years. I will specifically focus on data from 1970 onwards to align with the CO2 emissions data.

Structure and Quality:

- **Columns**: Date, AverageTemperature, AverageTemperatureUncertainty, Country

```
df = pd.read_csv(filepath_or_buffer='../data/emissions.csv')
```

	dt	AverageTemperat...	State	Country	Year
1970-01-01	3.739	Alabama	United States	1970
1970-02-01	7.2189999999999985	Alabama	United States	1970
1970-03-01	11.957	Alabama	United States	1970

- **Quality**: The dataset contains some missing values and inconsistencies that need to be addressed during the data cleaning process.

Data Pipeline

High-Level Overview

The data pipeline for this project consists of several key stages: data extraction, data transformation and cleaning, and data storage. The pipeline is implemented using Python, leveraging libraries such as pandas for data manipulation and SQLite for data storage.

Implementation and Technologies

- **Data Extraction**: The data is downloaded from Kaggle using the Kaggle API.
- **Data Transformation and Cleaning**:
    - **CO2 Emissions Data**: Filter the data for the years 1970 to the most recent year available. Ensure consistency in sector naming.
    - **Temperature Data**: Filter the data for the years 1970 to the most recent year available. Handle missing values by interpolating or using the mean of neighboring values. Remove unnecessary columns.
- **Data Storage**: The cleaned data is stored in an SQLite database for easy querying and analysis.

Transformation and Cleaning Steps

1. **CO2 Emissions Data**:
    - Load the CSV file.
    - Filter rows to include data from 1970 onwards.
    - Handle any missing values if present (though the dataset is generally clean).
2. **Temperature Data**:
    - Load the CSV file.
    - Filter rows to include data from 1970 onwards.
    - Handle missing values by interpolation or using mean values.
    - Drop unnecessary columns such as 'AverageTemperatureUncertainty'.

Problems Encountered and Solutions

- **Missing Values**: Some temperature records had missing values. This was addressed by using interpolation methods to fill gaps.
- **Data Integrity**: Data quality is ensured by applying transformations that guarantee that data loaded into the database is clean, correctly formatted, and includes only relevant data.
- **Data Alignment**: Ensuring both datasets covered the same time period required filtering and aligning the data appropriately.

Error Handling and Changing Input Data

The pipeline includes checks to handle missing data and validate data types. If new data is added or the structure changes, the pipeline is designed to log errors and provide descriptive messages to facilitate debugging.

Results and Limitations

Output Data

The output of the data pipeline is a cleaned and combined dataset stored in an SQLite database. This database contains two main tables: `emissions` and `temperature`, each with data from 1970 to the present.

Data Structure and Quality

- **CO2 Emissions Table**: Contains columns for year, sector, and CO2 emissions (MMT).
- **Temperature Table**: Contains columns for year and average temperature.

Output Data Format

The cleaned data is stored in an SQLite database because it provides efficient querying capabilities and is well-suited for our analysis needs. This format allows easy access to specific subsets of data and facilitates further statistical analysis.

Critical Reflection on Data and Potential Issues

- **Data Gaps**: Although missing values were handled, interpolation may introduce some bias.
- **Temporal Alignment**: Differences in data collection periods and methods between datasets may affect the accuracy of the results.
- **Sectoral Aggregation**: The way sectors are aggregated in the emissions data may not perfectly align with the impacts on temperature.

Conclusion

This report outlines the data sources, the structure and quality of the data, and the automated data pipeline created to prepare the data for analysis. The final cleaned dataset will be used to investigate the impact of sector-specific CO2 emissions in the USA on global surface temperatures.