1 Data Report

1.1 Question

How has the extent of sea ice changed over time in relation to global temperature changes?

1.2 Data Sources

1. **Global Temperature Change Data**
   - **Source**: "Environment_Temperature_change_E_All_Data_NOFLAG.csv"
   - **Content**: This dataset contains global temperature changes over time.
   - **Structure and Quality**: The dataset includes columns such as `Year`, `Country`, and temperature anomalies. It is structured in a tabular format and includes yearly data from various countries.
   - **License**: Open Data License - This data is under an open data license which allows for free use with proper attribution.
2. **Sea Ice Extent Data**
   - **Source**: "seaice.csv"
   - **Content**: This dataset contains data on the extent of sea ice over time.
   - **Structure and Quality**: The dataset includes columns such as `Year`, `Month`, `Day`, and `Extent`. It is also structured in a tabular format with monthly data.
   - **License**: Open Data License - This data is also under an open data license which allows for free use with proper attribution.

1.3 Data Pipeline

1.3.1 Overview

The data pipeline gathers, purges, transforms, and analyzes data in order to provide insights into the relationship between variations in sea ice extent and global temperature.

1.3.2 Steps

1. **Data Collection**: Load the datasets from the provided sources using a custom pipeline script.
2. **Data Cleaning**: Handle missing values, standardize date formats, and ensure data consistency.
3. **Data Transformation**: Aggregate data to match temporal granularity as yearly averages, normalize values, and merge datasets for combined analysis.
4. **Data Analysis**: Calculate correlations and visualize trends between temperature anomalies and sea ice extent.

1.3.3 Challenges and Solutions

- **Data Inconsistency**: Resolved by standardizing date formats and ensuring both datasets align temporally.
- **Missing Values**: Handled through interpolation or omission depending on the extent of missing data.
- **Changing Input Data**: Implemented automated checks and alerts for data updates or changes.

1.4 Results and Limitations

1.4.1 Output Data

- **Structure**: The output data includes aggregated and cleaned data for global temperature anomalies and sea ice extent, ready for analysis.
- **Format**: The output is stored in SQlite format for easy access and compatibility with various analysis tools.

1.4.2 Reflection

- **Data Quality**: High-quality data has been processed, and it has undergone rigorous cleansing and consistency tests.
- **Potential Issues**: The analysis may be impacted by regional disparities and temporal mismatch. Robust results might require more refinement and cross-validation using multiple data sources.

1.4.3 Disclaimer

Its original goal was to examine how heart disease and climate change are related. Unfortunately, it was discovered that this decision was poorly thought out because there was insufficient information to analyze and determine a significant association. Since this choice is better supported by the data at hand, it was made to investigate the relationship between sea ice extent and global temperature variations using the existing datasets.

1.5 Conclusion

This data pipeline offers a thorough method for examining the relationship between variations in sea ice extent and changes in global temperature. Reliable and perceptive analysis is made possible by the automated processes that guarantee data consistency and integrity.

2 Data Report - Implementation

2.1 Imports

2.2 Data Collection and Preparation from Kaggle

This section details the process of collecting the necessary datasets from Kaggle, preparing them for analysis, and ensuring the data is ready for further processing.

2.3 Data Transformation and Cleaning

2.3.1 Global Temperature Data

Transform the global temperature data to extract relevant information and calculate yearly averages.

```
   Year  Temperature_Anomaly
0  1961                0.207
1  1962                0.034
2  1963                0.159
3  1964               -0.253
```

```
4  1965                    -0.226
```

2.3.2 Sea Ice Data

Clean and aggregate the sea ice data to yearly averages.

```
   Year  Temperature_Anomaly  Sea_Ice_Extent
0  1978                0.004       12.976838
1  1979                0.225       12.019349
2  1980                0.330       11.785219
3  1981                0.439       11.776148
4  1982                0.078       12.048755
```

2.4 Calculate the correlation

Determine whether there is a linear relationship between Temperature_Anomaly and Sea_Ice_Extent by doing a correlation calculation. If so, this will help us create a predictive model for longer periods of time in the future.

```
                     Temperature_Anomaly  Sea_Ice_Extent
Temperature_Anomaly             1.000000       -0.723884
Sea_Ice_Extent                 -0.723884        1.000000
```

2.5 Overview of the Data with Plots