Prepared by Parisa Fathian (23069523)

1.1 Question

What is the impact of rising temperatures, as a result of climate change, on the premature mortality rate due to non-communicable diseases (NCDs) in various countries between the years 2000 and 2019?

1.2 Data Sources

To analyze the impact of rising temperatures due to climate change on the premature mortality rates from non-communicable diseases (NCDs) in various countries, I have selected two comprehensive data sets: one for temperature data and another for mortality rates. The analysis will cover the years from 2000 to 2019.

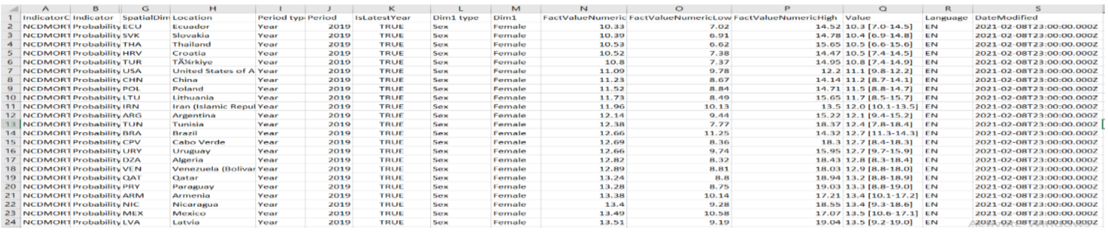
Data source-01: Berkley Earth Surface Temperature Data

- **Metadata:** Berkley Earth Surface Temperature Data
- **Sample Data:** The dataset can be accessed in CSV format, containing columns such as `Year`, `Country`, `monthly Temperature`, and `monthly Temperature anomaly`.
- **Description:** This data source provides historical temperature records from Berkley Earth, offering a detailed account of monthly surface temperatures across various countries over time. The dataset includes key attributes such as the average monthly temperature, monthly variations, and temperature anomalies. For our analysis, we will calculate the average annual temperature and temperature anomalies for each country from 2000 to 2019, which are essential for understanding long-term climate trends and their potential health impacts.

Data source-02: WHO Global Health Observatory (GHO) Data

- **Metadata:** WHO Global Health Observatory Data
- **Sample Data:** The dataset can be accessed in CSV format, containing columns such as `Country`, `Year`, and `Value`, where `Value` represents the probability of dying from NCDs between the ages of 30 and 70.
- **Description:** This data source from the World Health Organization (WHO) provides mortality statistics related to non-communicable diseases (NCDs) for various countries. It includes the probability of premature death from cardiovascular diseases, cancer, diabetes, and chronic respiratory diseases. The dataset is crucial for analyzing health outcomes in the context of changing environmental conditions.

Both datasets are critical for this analysis as they offer a comprehensive view of the variables involved: the temperature data helps in understanding climate trends, while the mortality data sheds light on health outcomes. Together, they provide the necessary information to investigate the relationship between climate change and public health.



	A	B	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	IndicatorC	Indicator	SpatialDim	Location	Period typ	Period	IsLatestYear	Dim1 type	Dim1	FactValueNumeric	FactValueNumericLow	FactValueNumericHigh	Value	Language	DateModified
2	NCDMORT	Probability	ECU	Ecuador	Year	2019	TRUE	Sex	Female	10.33	7.02	14.52	10.3 [7.0-14.5]	EN	2021-02-08T23:00:00.000Z
3	NCDMORT	Probability	SVK	Slovakia	Year	2019	TRUE	Sex	Female	10.39	6.91	14.78	10.4 [6.9-14.8]	EN	2021-02-08T23:00:00.000Z
4	NCDMORT	Probability	THA	Thailand	Year	2019	TRUE	Sex	Female	10.53	6.62	15.65	10.5 [6.6-15.6]	EN	2021-02-08T23:00:00.000Z
5	NCDMORT	Probability	HRV	Croatia	Year	2019	TRUE	Sex	Female	10.52	7.38	14.47	10.5 [7.4-14.5]	EN	2021-02-08T23:00:00.000Z
6	NCDMORT	Probability	TUR	TÃ¼rkiye	Year	2019	TRUE	Sex	Female	10.8	7.37	14.95	10.8 [7.4-14.9]	EN	2021-02-08T23:00:00.000Z
7	NCDMORT	Probability	USA	United States of A	Year	2019	TRUE	Sex	Female	11.09	9.78	12.2	11.1 [9.8-12.2]	EN	2021-02-08T23:00:00.000Z
8	NCDMORT	Probability	CHN	China	Year	2019	TRUE	Sex	Female	11.23	8.67	14.14	11.2 [8.7-14.1]	EN	2021-02-08T23:00:00.000Z
9	NCDMORT	Probability	POL	Poland	Year	2019	TRUE	Sex	Female	11.52	8.84	14.71	11.5 [8.8-14.7]	EN	2021-02-08T23:00:00.000Z
10	NCDMORT	Probability	LTU	Lithuania	Year	2019	TRUE	Sex	Female	11.73	8.49	15.65	11.7 [8.5-15.7]	EN	2021-02-08T23:00:00.000Z
11	NCDMORT	Probability	IRN	Iran (Islamic Repul	Year	2019	TRUE	Sex	Female	11.96	10.13	13.5	12.0 [10.1-13.5]	EN	2021-02-08T23:00:00.000Z
12	NCDMORT	Probability	ARG	Argentina	Year	2019	TRUE	Sex	Female	12.14	9.44	15.22	12.1 [9.4-15.2]	EN	2021-02-08T23:00:00.000Z
13	NCDMORT	Probability	TUN	Tunisia	Year	2019	TRUE	Sex	Female	12.38	7.77	18.37	12.4 [7.8-18.4]	EN	2021-02-08T23:00:00.000Z
14	NCDMORT	Probability	BRA	Brazil	Year	2019	TRUE	Sex	Female	12.66	11.25	14.32	12.7 [11.3-14.3]	EN	2021-02-08T23:00:00.000Z
15	NCDMORT	Probability	CPV	Cabo Verde	Year	2019	TRUE	Sex	Female	12.69	8.36	18.3	12.7 [8.4-18.3]	EN	2021-02-08T23:00:00.000Z
16	NCDMORT	Probability	URY	Uruguay	Year	2019	TRUE	Sex	Female	12.66	9.74	15.95	12.7 [9.7-15.9]	EN	2021-02-08T23:00:00.000Z
17	NCDMORT	Probability	DZA	Algeria	Year	2019	TRUE	Sex	Female	12.82	8.32	18.43	12.8 [8.3-18.4]	EN	2021-02-08T23:00:00.000Z
18	NCDMORT	Probability	VEN	Venezuela (Bolivar	Year	2019	TRUE	Sex	Female	12.89	8.81	18.03	12.9 [8.8-18.0]	EN	2021-02-08T23:00:00.000Z
19	NCDMORT	Probability	QAT	Qatar	Year	2019	TRUE	Sex	Female	13.24	8.8	18.94	13.2 [8.8-18.9]	EN	2021-02-08T23:00:00.000Z
20	NCDMORT	Probability	PRY	Paraguay	Year	2019	TRUE	Sex	Female	13.28	8.75	19.03	13.3 [8.8-19.0]	EN	2021-02-08T23:00:00.000Z
21	NCDMORT	Probability	ARM	Armenia	Year	2019	TRUE	Sex	Female	13.38	10.14	17.21	13.4 [10.1-17.2]	EN	2021-02-08T23:00:00.000Z
22	NCDMORT	Probability	NIC	Nicaragua	Year	2019	TRUE	Sex	Female	13.4	9.28	18.55	13.4 [9.3-18.6]	EN	2021-02-08T23:00:00.000Z
23	NCDMORT	Probability	MEX	Mexico	Year	2019	TRUE	Sex	Female	13.49	10.58	17.07	13.5 [10.6-17.1]	EN	2021-02-08T23:00:00.000Z
24	NCDMORT	Probability	LVA	Latvia	Year	2019	TRUE	Sex	Female	13.51	9.19	19.04	13.5 [9.2-19.0]	EN	2021-02-08T23:00:00.000Z

Figure 01: Raw Probability of dying between age 30 and exact age 70 from any of cardiovascular disease, cancer, diabetes, or chronic respiratory disease

Data Structure & Quality: Both datasets selected for this project are in tabular structure and CSV format, making them easy to handle and analyze using common data processing tools.

- **Accuracy**: The datasets have been sourced from reputable and authoritative sources, ensuring high accuracy and reliability:

    - **Berkley Earth Surface Temperature Data** provides accurate historical temperature records that are widely used in climate research.
    - **WHO Global Health Observatory Data** offers trustworthy mortality statistics derived from standardized health reporting practices.

    These datasets reflect real-world data, ensuring that the analysis is based on authentic and precise information.

- **Consistency**: All data is presented in a consistent tabular format, facilitating straightforward data manipulation and analysis. The temperature data includes monthly records of surface temperatures, while the health data provides annual mortality rates. This consistency in format allows for seamless integration and comparison across different datasets.
- **Relevancy :** The datasets cover the years 2000 to 2019, providing a relevant time frame to study long-term trends and impacts of climate change on health:

    - **Temperature Data:** Monthly temperature records from various countries, aggregated to annual averages.
    - **Health Data:** Annual mortality rates due to non-communicable diseases (NCDs) for the same countries.

    This period is chosen to capture significant climate variations and their potential impacts on public health over time.

- **Validity Checks :** Extensive data cleaning procedures have been applied to ensure the validity of the datasets:

    - **Temperature Data:** Checked for missing, invalid, and duplicate values. Any anomalies or inconsistencies have been addressed through rigorous data cleaning processes.
    - **Health Data:** Similarly, cleaned to remove any missing or invalid entries and to ensure that the data is complete and accurate for analysis.

By ensuring the data's accuracy, consistency, relevancy, and validity, we can confidently use these datasets to explore the relationship between climate change and premature mortality rates due to NCDs.

Data sources licenses & obligations: Both datasets used in this project are from public sources and are free to use, subject to specific licenses and obligations.

Berkley Earth Surface Temperature Data:

- **Access:** Berkley Earth Surface Temperature Data
- **License:** Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

- **Terms:** Allows sharing and adapting with attribution, for non-commercial purposes. CC BY-NC 4.0 Terms

WHO Global Health Observatory (GHO) Data:

- **Access:** WHO Global Health Observatory Data
- **License:** Open data license permitting non-commercial use.
- **Terms:** Free for research, analysis, and reporting, with proper attribution.

1.3 Data Pipeline

Technology : The data pipeline is built using Python with the following libraries: Numpy ,Pandas ,Requests, SQLite3.

Data Transformation Steps :

- Data Fetching**:** Both datasets are downloaded using the `requests` library.
- Temperature Data Transformation**:** Filter data for the years 2000 to 2019.Calculate the average annual temperature and temperature anomalies for each country.
- Health Data Transformation**:** Select relevant columns and Clean and preprocess the data.
- Data Storage**:** Store the transformed datasets in a SQLite database for further analysis.

Problems Encountered and Error Handling :

- **Data Download Issues:** Encountered challenges with downloading large datasets. Solved by implementing retry mechanisms in the `requests` library.
- **Missing and Invalid Values:** Addressed by thorough data cleaning and validation processes.
- **Dynamic Data Handling:** The pipeline is designed to adapt to changing data structures or formats, ensuring robustness and flexibility.

This pipeline ensures that data is accurately fetched, transformed, and stored, facilitating subsequent analysis to explore the relationship between climate change and health outcomes.

1.4 1.4 Result and Limitations

Data Output :

The data pipeline outputs two tables in a SQLite database : **Temperature Data:** Contains annual average temperatures and anomalies for each country from 2000 to 2019**. HealthData**: Contains annual mortality rates due to non-communicable diseases (NCDs) for each country over the same period.

Data Structure and Quality

- **Consistency and Accuracy:** Data has been thoroughly cleaned and validated to ensure consistency and accuracy. Each table is structured with clear, relevant columns, facilitating accurate analysis.
- **Accessibility:** The use of SQLite ensures efficient data retrieval and easy integration with various data analysis tools. This format supports seamless data manipulation and reporting.

Limitations and Potential Issues

- **Outliers and Anomalies:** Outliers or anomalies may still be present in the data. Rigorous profiling and validation techniques are necessary to identify and address these issues.
- **Data Completeness:** Some countries or years may have incomplete data, which can affect the comprehensiveness of the analysis.

- **Changing Data Formats:** While the pipeline is designed to handle changes in data structure, unforeseen changes can pose challenges.

Implementing robust statistical methods and anomaly detection algorithms will help ensure the integrity of the analysis. Continuous monitoring and updates to the data pipeline are essential to effectively manage these limitations.