Data Report

Prepared by: Redwanul Karim [**23426184**]

1 Introduction

The project aims to explore the possible connection between solar activity - like solar flares, and climate change on Earth. Through the statistical analysis of historical data, our goal is to uncover whether there exists a clear relationship between solar events and the observable shift in global climate patterns. This investigation holds significant importance as it provides insights essential for refining climate models and improving our capacity to forecast and address the consequences of climate change.

2 Research Question

"Is there any relationship between solar activity (Solar Flares) and climate change on Earth?"

3 Data Sources

3.1 Solar Flare Data

- **Source:** Zenodo [Dataset Link]
- **Rationale:** The Solar Flare Dataset from Zenodo, comprising 8,874 records spanning from May 2010 to December 2019, provides crucial insights derived from vector magnetic field data collected by Joint Science Operations Center (JSOC) and the Space Weather Prediction Center (SWPC) using Python's Sunpy library.
- **Data Content:** Recording Time, Flare Number, Lattitute, Longitude, Vector Magnetic Field data, such as - USFLUX, TOTPOT, TOTBSQ, ABSNJZH, and SAVNCPP etc.
- **Data Structure and Quality:** Tabular format (CSV), sourced from reliable JSOC and SWPC
- **Licensing:** MIT License

3.2 Temperature Change Data

- **Source:** Annual Surface Temperature Change from IMF [Dataset Link]
- **Rationale:** The Surface Temperature Change dataset, obtained from Food and Agriculture Organization Corporate Statistical Database (FAOSTAT), shows how Earth's average surface temperature has changed from 1961 to 2021 compared to temperatures between 1951 and 1980, using data from NASA GISS. It helps analyze temperature changes in different countries and is essential for understanding global temperature patterns.
- **Data Content:** Date, Country, ISO3, Temperature change etc.
- **Data Structure and Quality:** Tabular format (CSV), sourced from FAOSTAT
- **Licensing:** IMF License

3.3 CO_2 Concentration Data

- **Source:** Atmospheric CO_2 Concentrations from IMF [Dataset Link]
- **Rationale:** The Atmospheric CO_2 dataset offers monthly and yearly records of carbon dioxide levels in the air dating back to 1958, enabling users to track changes over time. Sourced from the National Oceanic and Atmospheric Administration Global Monitoring Laboratory, that provides dependable data crucial for climate research and analysis.
- **Data Content:** Date, CO_2 Concentration in PPM etc.
- **Data Structure and Quality:** Tabular format (CSV), sourced from NOAA
- **Licensing:** IMF License

4 Data Pipeline

In this section, we describe the ETL (Extract, Transform, Load) data pipeline implemented for our project. The pipeline is designed to systematically download, process, and store data from various sources, ensuring it is clean, structured, and ready for analysis. Below, we detail each step of the ETL Data Pipeline.

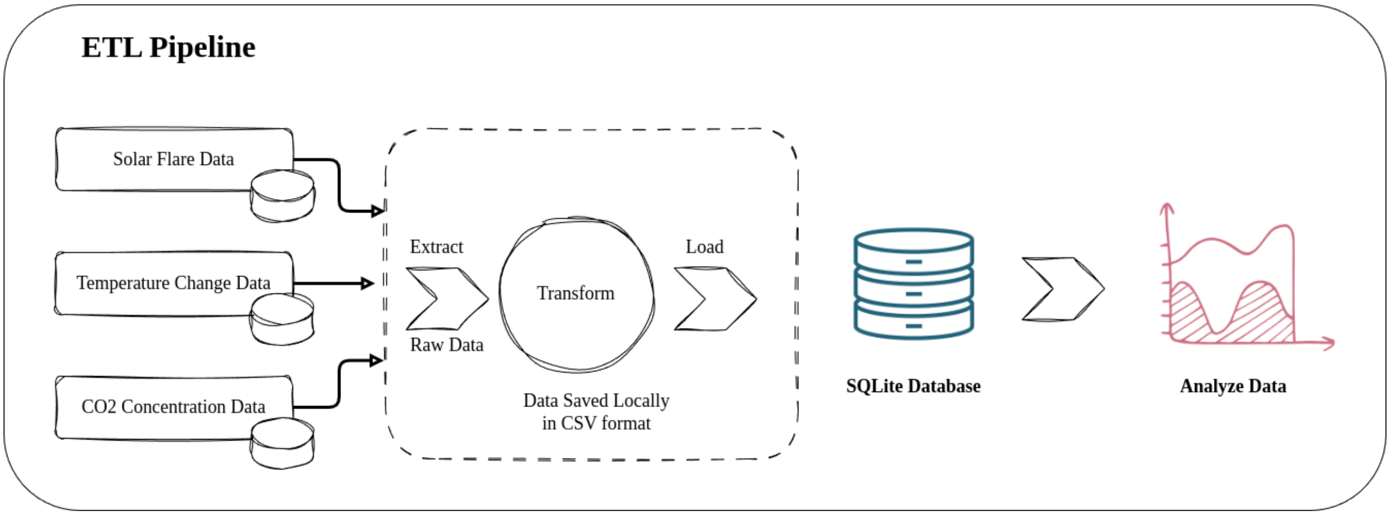

Figure 1: ETL Data Pipeline

4.1 Extract

The extraction phase involves sourcing datasets from their respective origins. The datasets used include Solar Flare Data, Temperature Change Data, and CO_2 Concentration Data. Using Pandas' CSV reader, we downloaded these datasets and saved them locally as CSV files. This method ensures data is captured in a structured and accessible format for subsequent processing.

4.2 Transform

The transformation phase involves several key steps to clean and prepare the data for analysis:

- **Data Ingestion and Cleaning:** The locally saved CSV files are read into Pandas DataFrames, as it provides a flexible structure for data manipulation. Unnecessary columns such as - Indicator, Unit, Source, CTS Code, CTS Full Descriptor etc, are removed as these data do not contribute to the analysis phase, and rows with missing values are dropped to ensure data integrity.
- **Data Type Conversion:** Columns representing dates and times are converted to appropriate datetime formats, and columns are renamed for clarity and consistency, for example column 'T_REC' in Solar Flare data was renamed to 'Date'.
- **Data Reshaping:** Using Pandas' melt method, wide-format temperature change data is converted into long-format, transforming columns into rows. This normalization step makes the data easier to work with for analysis and storage.
- **Saving Transformed Data:** The cleaned and transformed data points are saved back to local CSV files, that provides a checkpoint before loading into the SQLite database.

4.3 Load

The loading phase involves importing the transformed data into a structured database format. A SQLite database is created to store the transformed data due to its simplicity and ease of integration with Pandas. The transformed CSV files are read into Pandas DataFrames again, and each DataFrame is loaded into its own table within the SQLite database. This separation ensures organized and efficient data storage, enabling straightforward querying and analysis.

By following these ETL steps, the data pipeline effectively extracts, transforms, and loads datasets, making them ready for further analysis and ensuring high data quality and consistency throughout the process.

5 Data Analysis

We conducted a statistical analysis to investigate the relationship between solar activity (Solare Flare events) and climate change (Temperature Change and CO_2 concentration) on Earth. Using a subset of our data extracted from the sink SQLite database prepared with the ETL Data Pipeline.

We began by determining the common time window for all datasets and then employed a Random Forest Regressor to identify the top 5 significant features from the Solar Flare dataset, namely 'TOTBSQ', 'TOTPOT', 'ABSNJZH', 'SAVNCPP', and 'USFLUX'. Subsequently, we performed hypothesis testing using t-tests for Temperature change and CO_2 concentration, revealing significant p-values for both variables.

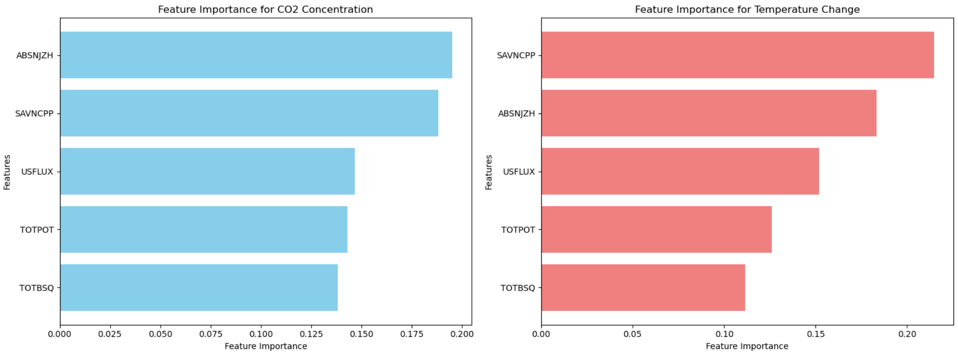

Figure 2: Five most important feature of Solar Flare data in relation to Temperature Change and CO_2 Concentration on Earth

6 Result

The hypothesis testing (t-tests) revealed notable p-values for both CO_2 concentration and Temperature change. However, a deeper examination on data distribution and auto-correlation of the data variables uncovered that the data points we are using violates the underlying assumptions of normality and homogeneity of variances, which casts a doubt on the validity of our hypothesis testing results. As a result, based on our data analysis, at this point we cannot conclude with certainty that there is relationship between solar flare events and climate change on earth.

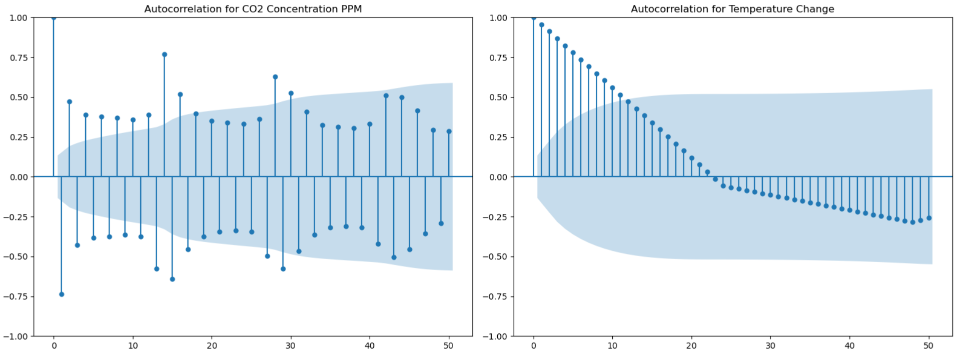

Figure 3: Auto-correlation plot of Temperature Change and CO_2 Concentration data

7 Limitations and Future Direction

While hypothesis testing offers useful information about statistical correlations, it doesn't prove cause and effect. Additionally, the violations of assumptions like normality and equal variances mean we should be careful with our interpretations. Moving forward, we need to explore domain-specific machine learning algorithms to handle the complexities of non-normally distributed data in order to better understand the relationship between solar activity and climate change on Earth.