1. Introduction:

1.1 Background

Rising CO2 emissions fuel Earth's warming trend, Greenhouse gasses trap heat, causing harm without end. Fossil fuels burn, releasing carbon in the air, Atmospheric changes trigger weather beyond repair.

Ice caps melt, contributing to sea levels high,Extreme weather events become frequent as storms amplify. Forests shrink, oceans acidify, ecosystems feel the strain, Wildlife struggles to adapt in a world of changing rain.Human activities drive this relentless climb,Global surface temperatures peak, time after time.

1.2 Research Question and Motivation

This project therefore aims to investigate whether there are conclusions that can be drawn out of a possible correlation between the CO2 emissions and global temperature anomalies.Therefore, global temperature change from year 1950 to 2024 are observed and also in this time period, the CO2 emissions by different industries of each country. The result of this project will provide the useful insight like how the global increase in CO2 emissions increase the surface temperature of the world.

2. Data

2.1 Data Sources

2.1.1 Data source 1: NASA GISS Surface Temperature Analysis (GISTEMP)

- Data URL: https://data.giss.nasa.gov/gistemp/tabledata_v4/GLB.Ts+dSST.csv
- Data Type: CSV
- This data set addresses the problem of global surface temperature measurement and analysis. It provides a comprehensive record of temperature anomalies over time, which is crucial for understanding trends in global warming and climate change.

2.1.2 Datasource2: Our World in Data CO2 dataset

- Data URL: https://github.com/owid/co2-data/raw/master/owid-co2-data.csv
- Data Type: CSV
- This data set addresses the problem of tracking global carbon dioxide emissions. It provides detailed records of CO2 emissions from different countries , which is critical for understanding the sources of greenhouse gasses contributing to climate change.

2.2 Data Structure and Meaning:

- **NASA GISS Surface Temperature Analysis (GISTEMP):** The data is structured in a tabular CSV format. It is a global time series data with monthly frequency, where there are year and months columns. This data shows temperature change of the world in degree- celsius from 1880 to 2024. Column J-D shows the temperature change from January to December for each year. The data contains missing values till 1950, accurate, it is up-to-date and it reflects the real world

```
Downloaded data from https://data.giss.nasa.gov/gistemp/tabledata_v4/GLB.Ts+dSST.csv:
   Year   Jan   Feb   Mar   Apr   May  ...   Aug   Sep   Oct   Nov   Dec   J-D
0  1950 -0.26 -0.27 -0.07 -0.21  -.11  ...  -.16  -.11  -.20  -.34  -.21  -.17
1  1951 -0.34 -0.41 -0.20 -0.14   .00  ...   .06   .05   .08  -.01   .16  -.07
2  1952  0.11  0.11 -0.08  0.03  -.03  ...   .05   .07   .00  -.13  -.02   .01
3  1953  0.07  0.15  0.11  0.19   .11  ...   .05   .05   .08  -.03   .05   .08
4  1954 -0.24 -0.10 -0.15 -0.14  -.20  ...  -.18  -.10  -.02   .08  -.18  -.13

[5 rows x 14 columns]
```

Figure 1: First five rows of global surface temperature changes.

- **Our World in Data CO2 dataset:** The data is structured in tabular csv format. It is the global time series data. It shows the CO2 emission of the countries by each year from 1850 to 2024. There are different types of CO2 emissions. The indicator value is expressed in billion tons. The data contains missing values till 1950, accurate, up-to-date and it reflects the real world.

```
Downloaded data from https://github.com/owid/co2-data/raw/master/owid-co2-data.csv:
       country  year  cement_co2  ...  gas_co2  oil_co2  share_global_co2
0  Afghanistan  1950         0.0  ...      0.0    0.063             0.001
1  Afghanistan  1951         0.0  ...      0.0    0.066             0.001
2  Afghanistan  1952         0.0  ...      0.0    0.060             0.001
3  Afghanistan  1953         0.0  ...      0.0    0.068             0.002
4  Afghanistan  1954         0.0  ...      0.0    0.064             0.002

[5 rows x 9 columns]
```

Figure 2: First five rows of carbon dioxide emissions by countries

3. Analysis

3.1 Methods:

3.1.1 Data pipeline

pipeline.sh

This shell script runs the data pipeline(madepipeline.py).

madepipeline.py

This python script downloads data from both URLs and delete rows before 1950 because of inconsistencies. It drops the irrelevant columns from the first URL and only keeps the relevant columns from second URL. It than saves the data into sqlite database, in the climate table.

tests.sh

This shell script first cleans any pre-existing output database files and verify clean up. After this, it runs the pipeline. If pipeline does not fail it than runs the system_tests.py

system_test.py

This python script first reads the output_files_info.json to read the expected filenames in database. It defines the data directory and than function checks if the file exists in any sub folders. If all expected files exists it gives the system test passed else it gives system test failed.

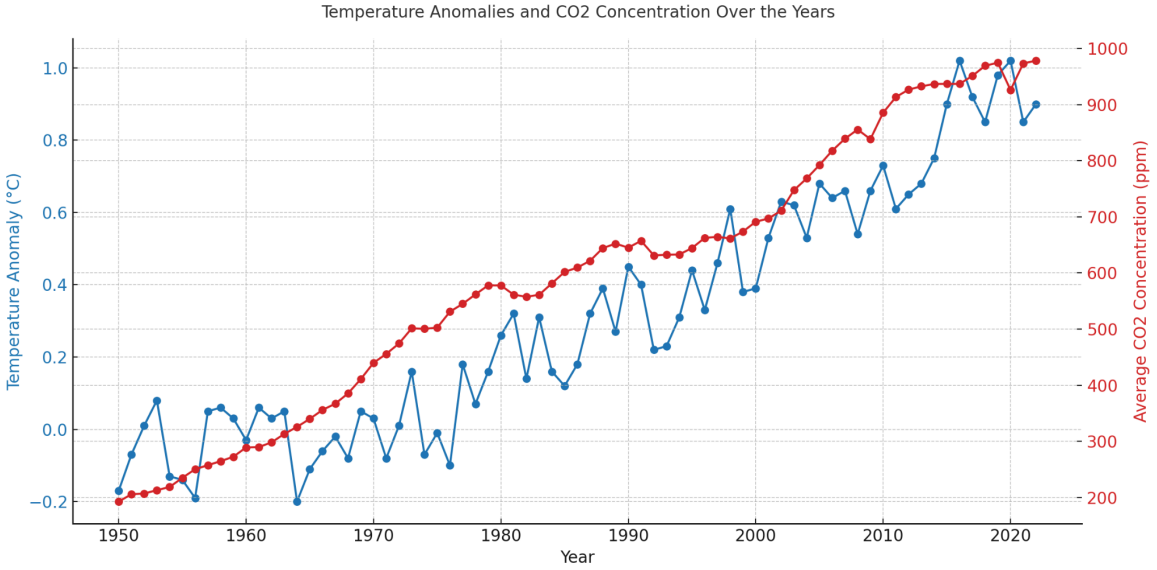
3.2 Result and Interpretation

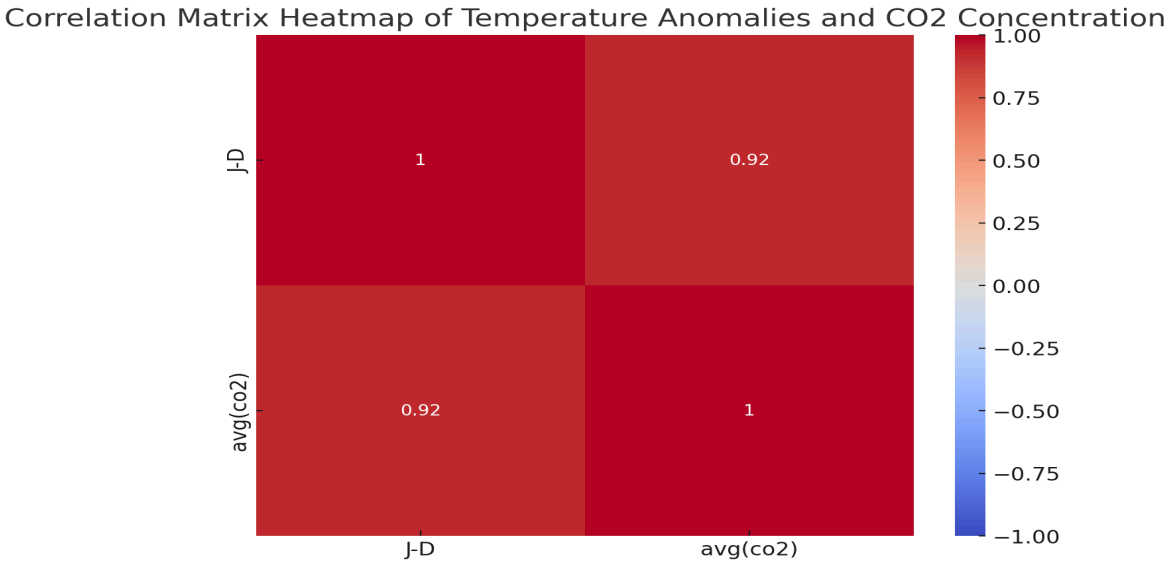
To gain insights into possible answers to the research question, It was necessary to deeply understand the overall temperature anomalies of the world for each year from 1950 to 2024 and also the monthly temperature anomalies for each year. It was also equally important to understand what are the causes of these anomalies. According to well known institutions of the world regarding climate change, CO2 is the major cause. So the global CO2 release and different types of CO2 industries are studied for each country.

3.2.1 variables studied:

CO2	**Temperature**
Share global CO2	Yearly
Cement CO2	Monthly
CO2	
CO2 growth percentage	

Coal CO2	
Gas CO2	
Oil CO2	





The plot and correlation matrix are showing that there is a strong relationship between the temperature anomalies and CO2 concentration.

4. Conclusion and Critical Analysis:

The data about average yearly temperature anomaly shows clear trend of rising anomalies over the years. CO2 emissions from various sources (cement, coal, gas, oil, and total CO2) have generally increased over the years.The trend lines for each source of CO2 emissions show a consistent rise, especially notable from industrial activities.

The analysis is limited by the scope of the provided data. The plot shows that while the temperature anomalies rises with the increase in CO2 emissions but it is not the only cause of the rising temperature, additional variables would provide a more robust analysis.