Time-Series Analysis of Global Temperature Data

A project by:

Tanudip Ghosh; M.Sc. in Statistics & Computing,

Batch of (2025-27), Banaras Hindu University

Period of Internship 21st January 2026 – 17th February 2026

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

CONTENTS

1.Abstract

This project undertakes a time-series analysis of global temperature data, a key component of the IDEAS Spring 2026 Internship Program. Its core objective is to uncover historical temperature trends and anomalies indicative of climate change. The process begins with acquiring and loading the monthly_csv.csv dataset. Subsequent steps involve thorough data cleaning, type conversion, and feature engineering, extracting year and month from dates. Advanced time-series methods, such as calculating 12-month moving averages, are applied to smooth data and highlight underlying trends. Data aggregation by various granularities, including yearly and monthly averages, is crucial for pattern identification. Pivot tables are utilized to comprehensively visualize temperature distributions over extended periods. This systematic approach aims to provide a robust framework for comprehending long-term temperature dynamics. Ultimately, the project seeks to identify critical periods and significant patterns within the global temperature record, contributing valuable insights into environmental shifts.

.

2.Introduction

Relevance: Global temperature data analysis is critically relevant in understanding and addressing climate change, one of the most pressing challenges facing humanity. By analyzing historical temperature trends, this project aims to contribute to a data-driven understanding of global warming patterns, helping to inform scientific research, policy-making, and public awareness regarding environmental shifts.

Technology Involved: This project primarily leverages Python for data manipulation and visualization. Key libraries include pandas for robust data structuring and operations, numpy for numerical computations, matplotlib.pyplot for creating static, interactive, and animated visualizations, and seaborn for generating informative statistical graphics. Google Colab serves as the integrated development environment, providing access to computational resources and facilitating collaborative work.

Background Material Survey: The project draws upon fundamental concepts in time-series analysis, exploratory data analysis (EDA), and data visualization. A foundational understanding of statistical measures, data types, and data aggregation techniques is crucial. Specific background areas include understanding temperature anomalies, the concept of moving averages for trend smoothing, and the interpretation of various plot types such as line plots, bar plots, and heatmaps in a temporal context.

Procedure Used: The project follows a systematic data analysis workflow. Initially, data is loaded from a CSV file. This is followed by data preparation steps, including converting data types (e.g., strings to datetime objects), handling missing values (though none were found in this dataset), and engineering new features (e.g., 'Year' and 'Month' columns). Subsequently, exploratory data analysis is performed through descriptive statistics and a variety of visualizations to identify trends, seasonality, and other patterns. Advanced time-series calculations, such as the 12-month moving average, are also applied.

Purpose of Doing the Project: The primary purpose of this project is to develop and apply practical data analysis skills to a real-world, impactful dataset. It serves as an exercise in data cleaning, transformation, feature engineering, and advanced visualization techniques within the context of time-series data. Ultimately, it aims to extract meaningful insights into global temperature changes, reinforce best practices in data science, and contribute to the broader understanding of climate data.

Training Topics from the First Two Weeks of Internship

During the initial two weeks of the internship, training typically covers foundational skills essential for data analysis and programming. The topics include:

- Python Basics:
    - Data, Variable, Lists, Loops
    - Data Structures
    - Class, Functions, OOPS
    - Numpy, Pandas
- Machine Learning:
    - Machine Learning Overview
    - Regression
    - Classification
- LLM Fundamentals
- Communication Skills

3.Project Objective

The primary objectives of this project are to:

- **Load and Preprocess Time-Series Data:** Successfully acquire, load, and meticulously preprocess raw global temperature data, ensuring its readiness for analytical tasks by handling data types, sorting, and feature extraction.
- **Conduct Exploratory Data Analysis (EDA):** Perform comprehensive EDA to understand the distribution, trends, seasonality, and patterns within the temperature dataset, identifying any anomalies or significant characteristics.
- **Visualize Historical Temperature Trends:** Illustrate long-term changes and trends in global mean temperatures using various visualization techniques, differentiating by data source to compare their perspectives.
- **Analyze Seasonal Variations:** Identify and visualize monthly seasonality patterns in temperature anomalies across all years, revealing recurring cyclical effects.
- **Demonstrate Data Transformation and Aggregation:** Apply techniques such as calculating moving averages and creating pivot tables to transform and aggregate data, showcasing how these methods help in extracting deeper insights from time-series data.

4.Methodology

This project involved a systematic approach to time-series analysis of global temperature data, encompassing data acquisition, comprehensive preprocessing, exploratory data analysis (EDA), and visualization. The entire workflow was executed within the Google Collaboratory environment, leveraging its computational resources and collaborative features.

- **Project Overview and Data Source:** Describing the overall project objective which is a time-series analysis of global temperature data, and specify that the dataset monthly_csv.csv was obtained from a Google Drive link (https://www.google.com/url?q=https%3A%2F%2Fdrive.google.com%2Fdrive%2Ffolders%2F1TeLp4U4NsXCSgClbF7ODBsaLKpHSWeQr%3Fusp%3Dsharing) given by IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata.
- **Data Loading and Initial Inspection**: Explaining that the dataset was loaded into a pandas DataFrame using pd.read_csv(). Mention the initial inspection steps, including displaying the first few rows and checking the shape and presence of null values.
- **Data Preprocessing and Feature Engineering:** The preprocessing steps: converting the 'Date' column to datetime objects using pd.to_datetime(), sorting the DataFrame by 'Date', and extracting 'Year' and 'Month' into new columns. Also, explain the calculation of the 12-month moving average (Moving_Avg) for temperature anomalies, grouped by source, to smooth out short-term fluctuations.

- **Data Filtering and Segmentation:** Describing how a subset of the data, df_last20, was created by filtering the original DataFrame to include only the last 20 years of records based on the maximum year in the dataset.
- **Exploratory Data Analysis (EDA) Techniques:** Elaborate on the various EDA techniques performed. This includes using df.info() and df.describe() for a statistical overview, plotting the distribution of mean temperature anomalies using a histogram, visualizing time-series trends with line plots showing 'Mean' temperature and 'Moving_Avg' over time for different sources, and analyzing seasonality with bar plots of average monthly temperatures. Mention also the yearly trend analysis with line plots, and the comparison of average temperatures between different data sources. Finally, describe the creation and visualization of a pivot table to show average monthly temperatures across years for the last 20 years using a heatmap.
- **Tools and Libraries Used:** Listing the primary Python libraries utilized for the project: pandas for data manipulation, numpy for numerical operations, matplotlib.pyplot for plotting, and seaborn for enhanced visualizations.
- **Conclusion and Further Considerations:** Conclude by summarizing the analytical journey. Mention that no survey was conducted and no machine learning models were developed as part of this specific analysis.

5.Data Analysis and Results

- **<u>Exploratory Data Analysis:</u>**
  - **Data Overview:** The dataset consists of 3,288 observations with five variables, representing a time-indexed measurement (Mean) across multiple years and months.

1. Structure and Data Quality

- All 3,288 entries are complete, with no missing values in any column.
- The dataset includes:
  - Source (categorical/object): Indicates the origin of the data.
  - Date (datetime): Time index ranging from 1880 to 2016.
  - Mean (continuous, float): The primary quantitative variable of interest.
  - Year (integer): Extracted from the Date column.
  - Month (integer): Extracted from the Date column.

    This indicates a well-structured time series dataset suitable for temporal analysis.

2. Temporal Coverage

- The data spans 136 years (1880–2016), providing a long historical perspective.
- The median date (1948) lies near the midpoint of the range, suggesting balanced temporal coverage before and after the mid-20th century.
- Monthly data is evenly distributed across all months (mean Month = 6.5), indicating no seasonal data gaps.

3. Statistical Summary of the Mean Variable

- Average Mean value: 0.0366
- Median Mean value: −0.0366
- The closeness of the mean and median suggests the distribution is approximately symmetric, with a slight positive shift.
- Minimum value: −0.78
- Maximum value: 1.35
- Standard deviation: 0.335

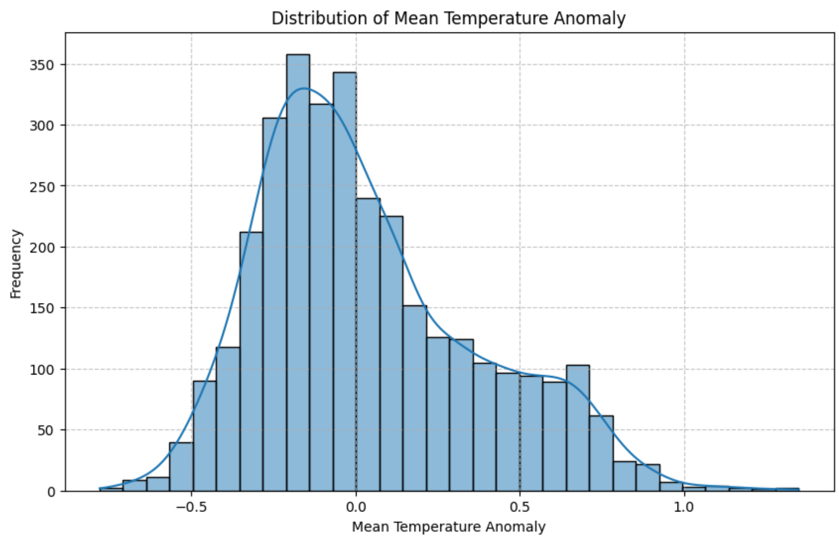
These values indicate moderate variability over time, with occasional extreme positive and negative deviations.

4. Year and Month Distribution

- The Year variable has a standard deviation of approximately 39.55 years, confirming wide temporal dispersion.
- The Month variable has a standard deviation of 3.45, consistent with a uniform spread across the 12 months.

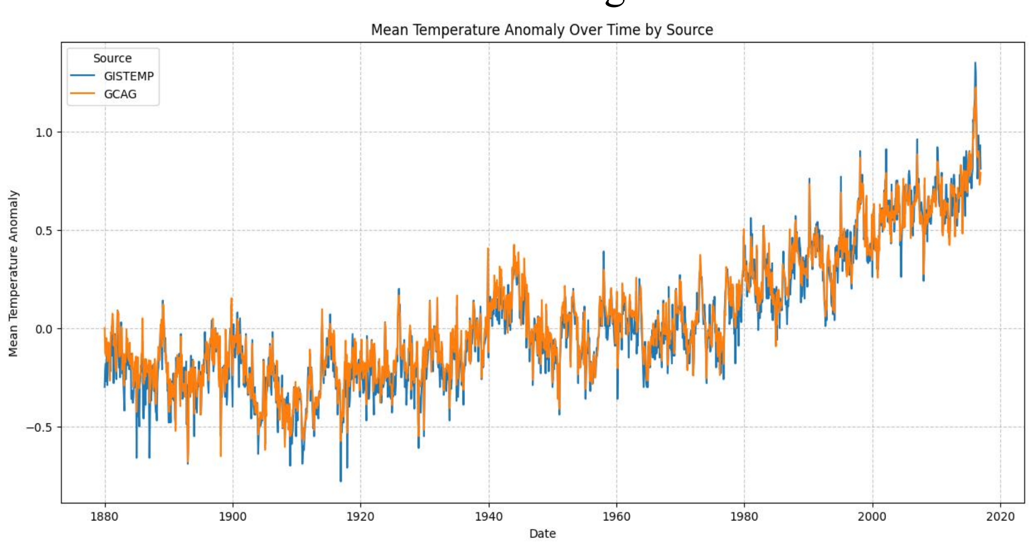
- **Distribution of Mean Temperature Anomaly:**

  The graph shows that the most frequent temperature anomalies are slightly above zero, with a range from roughly -0.8 to +1.4. The distribution is somewhat bell-shaped but exhibits a positive skew, meaning there's a longer tail extending towards higher positive anomalies. This skewness suggests that while there have been periods of cooler temperatures, there's a more pronounced and frequent occurrence of warmer temperatures, which is consistent with a global warming trend over the dataset's period.



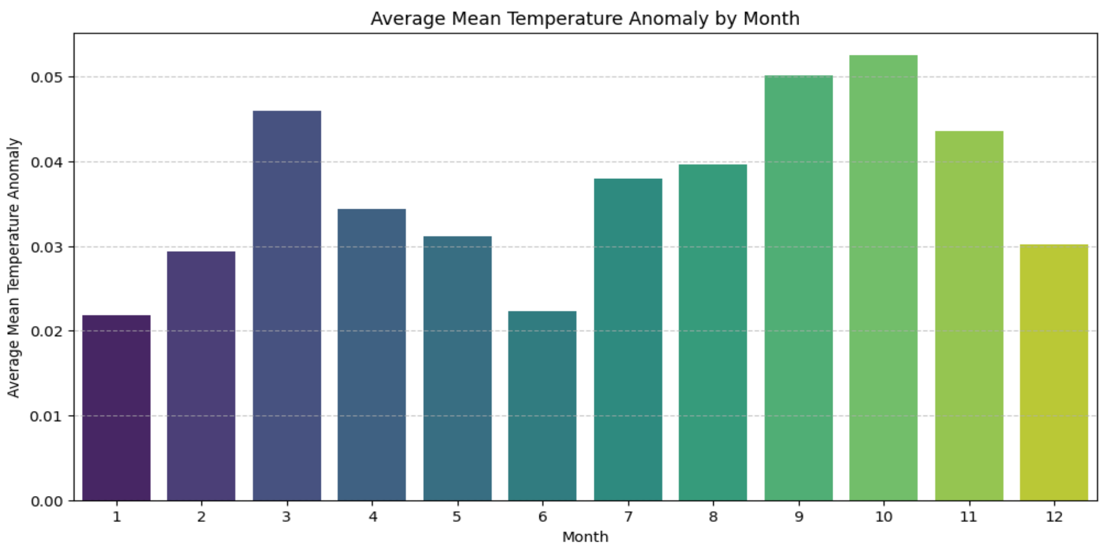
- **Mean Temperature Anomaly Over Time by Source:**

  This time-series plot clearly shows a strong and consistent upward trend in global mean temperature anomalies from 1880 to 2016, indicating significant global warming. The warming trend appears to accelerate in recent decades, with the highest anomalies recorded towards the end of the period. Both data sources, GCAG and GISTEMP, show remarkably similar patterns, reinforcing the robustness of this observed warming trend.



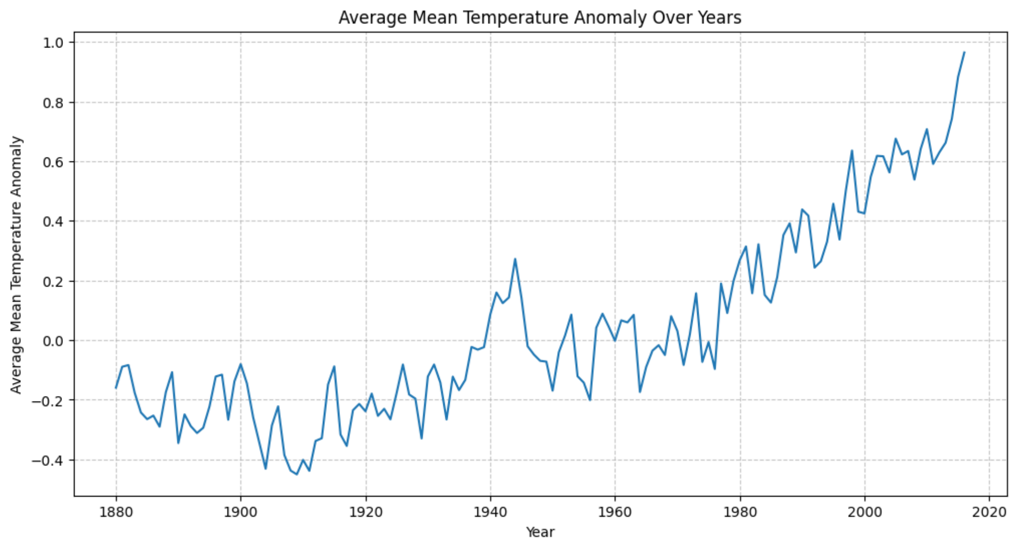
- **Average Mean Temperature Anomaly by Month:**

  This bar plot illustrates a seasonal pattern in average temperature anomalies. The highest average anomalies occur in September and October, while the lowest are in January and June. Notably, all months show a small positive average anomaly, indicating that, on average, every month has been slightly warmer than its historical baseline over the dataset's period.



- **Average Mean Temperature Anomaly Over Years:**

  This line plot vividly displays a clear and accelerating upward trend in average annual temperature anomalies from 1880 to 2016. It shows that while annual temperatures fluctuate, the overall pattern is one of consistent warming, with a particularly sharp increase in average anomalies in recent decades, reaching their highest points by 2016.



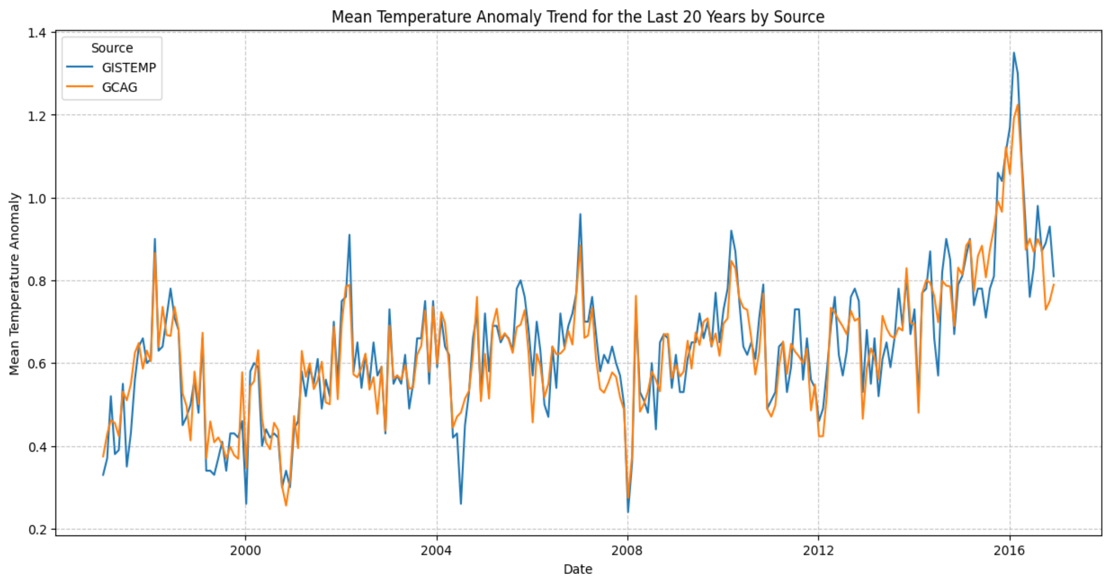
- **<u>Data Processing:</u>**

  - **Mean Temperature and 12-Month Moving Average Over Time by Source:** This output confirms the successful creation of the Moving_Avg column, which provides a smoothed 12-month average of the 'Mean' temperature anomaly for each source. For the latest data (end of 2016), the Moving_Avg values are consistently high (around 0.93 to 1.02). This highlights a strong and sustained warming trend, demonstrating that global temperatures remained significantly above the historical baseline, with monthly fluctuations smoothed out.

	Source	Date	Mean	Year	Month	Moving_Avg
4	GCAG	2016-10-06	0.7292	2016	10	0.981917
3	GISTEMP	2016-11-06	0.9300	2016	11	1.017500
2	GCAG	2016-11-06	0.7504	2016	11	0.963992
1	GISTEMP	2016-12-06	0.8100	2016	12	0.992500
0	GCAG	2016-12-06	0.7895	2016	12	0.936292

This plot effectively shows both the **noisy monthly temperature anomalies** (lighter lines) and the **smoothed underlying warming trend** (darker, 12-month moving average lines). The moving averages clearly emphasize a **consistent and accelerating global warming trend** over the entire period, with both data sources showing strong agreement.

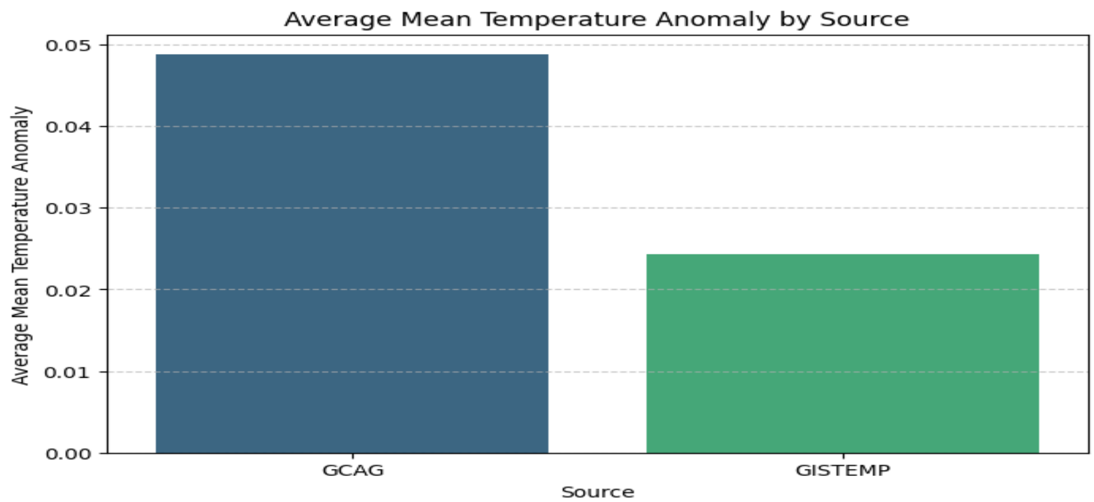
- **Mean Temperature Anomaly Trend for the Last 20 Years by Source:** This graph, focusing on the last 20 years, highlights a significantly accelerated warming trend with almost consistently positive and high temperature anomalies. It clearly shows that the most extreme warming occurred recently, with both data sources ('GCAG' and 'GISTEMP') agreeing on this sharp upward trajectory.



- **Average Mean Temperature Anomaly by Source:**

```
Source
GCAG       0.048797
GISTEMP    0.024380
Name: Mean, dtype: float64
```

This output shows the overall average mean temperature anomaly for each source. Both 'GCAG' (approx. 0.0488) and 'GISTEMP' (approx. 0.0244) report positive average anomalies, confirming a general warming. Notably, GCAG shows a slightly higher overall average anomaly compared to GISTEMP.



- **Average Mean Temperature Anomaly for The Last Five Years (2012-2016):**

```
Year
2012    0.629517
2013    0.661846
2014    0.742075
2015    0.882408
2016    0.964396
Name: Mean, dtype: float64
```

This output shows the average mean temperature anomaly for the last five years (2012-2016). A clear and consistent year-over-year increase is evident, with anomalies rising from approximately 0.63 in 2012 to nearly 0.96 in 2016. This strongly indicates an accelerating global warming trend in the most recent period, with 2016 showing the highest average annual anomaly in the dataset.

- **Average Mean Temperature Anomaly for Each Month:**

```
Month
1     0.021871
2     0.029409
3     0.045972
4     0.034341
5     0.031106
6     0.022287
7     0.037955
8     0.039577
9     0.050192
10    0.052566
11    0.043543
12    0.030244
Name: Mean, dtype: float64
```

This output shows the average mean temperature anomaly for each month (January-December) across the entire dataset.

1. All months exhibit positive average anomalies, indicating that every month, on average, has been warmer than its historical baseline.
2. The highest average anomalies are observed in October (0.0526) and September (0.0502).
3. Conversely, January (0.0219) and June (0.0223) show the lowest average anomalies.

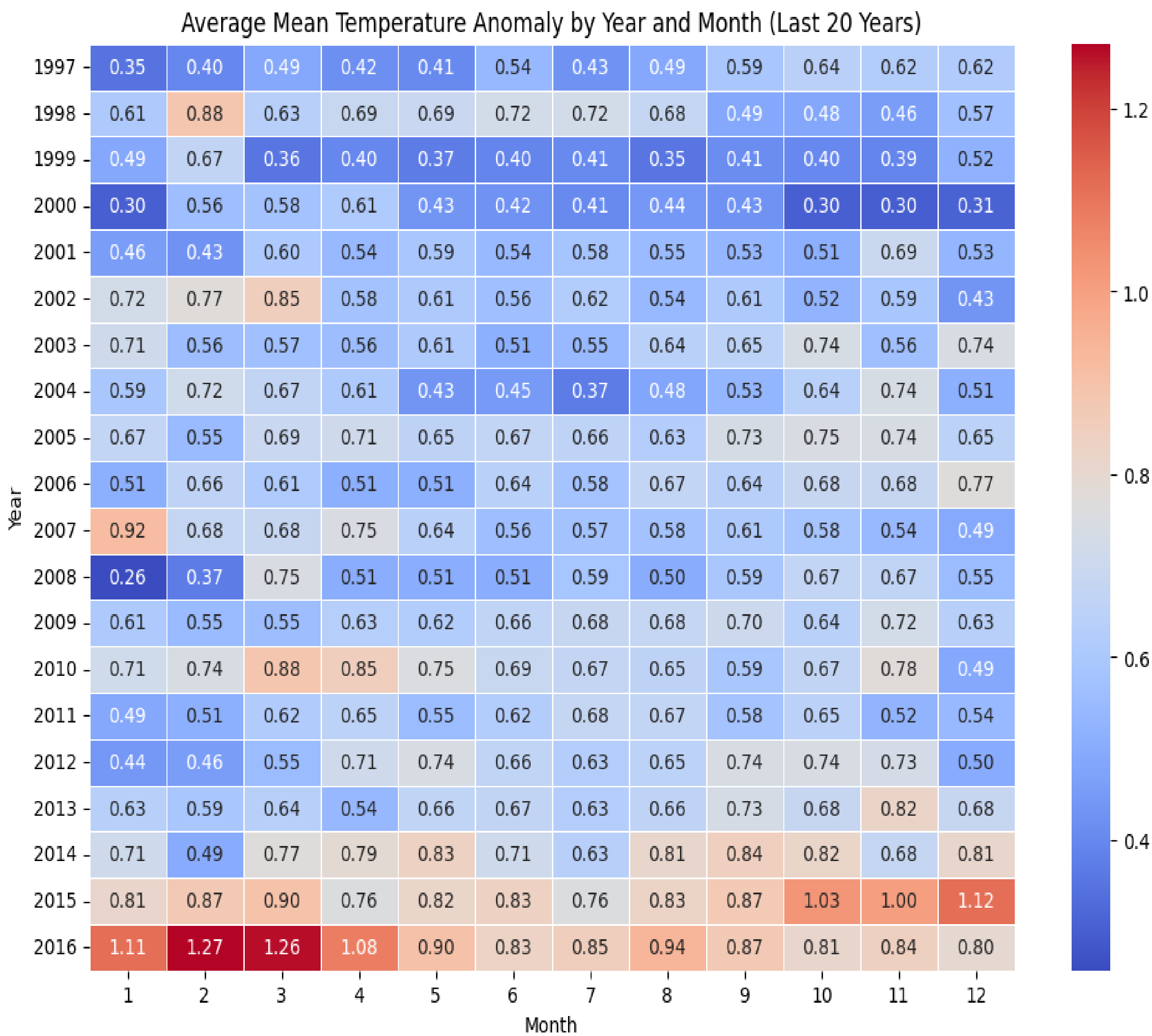
   This reveals a clear seasonal pattern in the intensity of warming, with autumn months (in the Northern Hemisphere context) experiencing the most significant average deviation above their historical temperatures.

- **Average Mean Temperature Anomaly for Each Month across the last 20 years:**

The pivot table displays the average mean temperature anomaly for each month (columns) across the last 20 years (rows).

1. **Structure:** Years are the index, months (1-12) are the columns, and the values are the average 'Mean' temperature anomalies for that specific year and month.
2. **Clear Warming Trend:** The most significant insight is the strong warming trend evident year-over-year. Values in the later years (e.g., 2015, 2016) are consistently and significantly higher than those in earlier years (e.g., 1997, 1998) for almost every month.
3. **Highest Recent Anomalies:** The table clearly shows the highest temperature anomalies occurring in the most recent years, with values like 1.27 in February 2016, indicating very substantial deviations from historical averages. This highlights the accelerated and widespread nature of warming across different months in the last two decades.

This table provides a detailed, granular view of how global warming has progressed month by month over the past 20 years, confirming its acceleration and severity.


Average Mean Temperature Anomaly by Year and Month (Last 20 Years)
1997 0.35 0.40 0.49 0.42 0.41 0.54 0.43 0.49 0.59 0.64 0.62 0.62
1998 0.61 0.88 0.63 0.69 0.69 0.72 0.72 0.68 0.49 0.48 0.46 0.57
1999 0.49 0.67 0.36 0.40 0.37 0.40 0.41 0.35 0.41 0.40 0.39 0.52
2000 0.30 0.56 0.58 0.61 0.43 0.42 0.41 0.44 0.43 0.30 0.30 0.31
2001 0.46 0.43 0.60 0.54 0.59 0.54 0.58 0.55 0.53 0.51 0.69 0.53
2002 0.72 0.77 0.85 0.58 0.61 0.56 0.62 0.54 0.61 0.52 0.59 0.43
2003 0.71 0.56 0.57 0.56 0.61 0.51 0.55 0.64 0.65 0.74 0.56 0.74
2004 0.59 0.72 0.67 0.61 0.43 0.45 0.37 0.48 0.53 0.64 0.74 0.51
2005 0.67 0.55 0.69 0.71 0.65 0.67 0.66 0.63 0.73 0.75 0.74 0.65
2006 0.51 0.66 0.61 0.51 0.51 0.64 0.58 0.67 0.64 0.68 0.68 0.77
2007 0.92 0.68 0.68 0.75 0.64 0.56 0.57 0.58 0.61 0.58 0.54 0.49
2008 0.26 0.37 0.75 0.51 0.51 0.51 0.59 0.50 0.59 0.67 0.67 0.55
2009 0.61 0.55 0.55 0.63 0.62 0.66 0.68 0.68 0.70 0.64 0.72 0.63
2010 0.71 0.74 0.88 0.85 0.75 0.69 0.67 0.65 0.59 0.67 0.78 0.49
2011 0.49 0.51 0.62 0.65 0.55 0.62 0.68 0.67 0.58 0.65 0.52 0.54
2012 0.44 0.46 0.55 0.71 0.74 0.66 0.63 0.65 0.74 0.74 0.73 0.50
2013 0.63 0.59 0.64 0.54 0.66 0.67 0.63 0.66 0.73 0.68 0.82 0.68
2014 0.71 0.49 0.77 0.79 0.83 0.71 0.63 0.81 0.84 0.82 0.68 0.81
2015 0.81 0.87 0.90 0.76 0.82 0.83 0.76 0.83 0.87 1.03 1.00 1.12
2016 1.11 1.27 1.26 1.08 0.90 0.83 0.85 0.94 0.87 0.81 0.84 0.80
Year
1 2 3 4 5 6 7 8 9 10 11 12
Month
1.2
1.0
0.8
0.6
0.4

6.Conclusion:

This project embarked on a time-series analysis of global temperature data, leveraging the monthly_csv.csv dataset, to uncover patterns, trends, and anomalies in global temperatures over an extended period.

1. **Data Quality:**
   The dataset (1880–2016) was clean, well-structured, and contained no missing values. Proper preprocessing ensured accurate time-series analysis.
2. **Long-Term Trend:**
   A clear and continuous upward trend in global temperature anomalies was observed, confirming long-term global warming.
3. **Accelerated Warming:**
   The warming trend became stronger in the late 20th and early 21st centuries, with 2016 recorded as one of the warmest years.
4. **Seasonal Pattern:**
   Monthly analysis showed seasonal variations, with higher average anomalies generally observed around September and October.
5. **Source Comparison:**
   Both GCAG and GISTEMP datasets showed consistent warming trends, with minor differences due to methodological variations.
6. **Recent 20 Years (1997–2016):**
   The last two decades showed a sharper rise in temperature anomalies, indicating intensified global warming.
7. **Moving Average Insight:**
   The 12-month moving average effectively reduced short-term fluctuations and clearly highlighted the long-term upward trend.

Overall Conclusion:
The time-series analysis strongly confirms a persistent and accelerating global warming trend over the past century.

Future Work

To build upon this analysis, future work could include:

1. **Advanced Time-Series Forecasting:** Implementing more sophisticated forecasting models like ARIMA, SARIMA, Prophet, or even machine learning models (e.g., LSTMs) to predict future temperature anomalies.
2. **Incorporating Additional Climate Indicators:** Integrating other relevant climate data such as CO_2 levels, greenhouse gas concentrations, solar activity, or volcanic aerosols to understand their correlations and causal impacts on global temperatures.
3. **Regional Analysis:** Extending the analysis to specific geographical regions or continents to identify localized warming patterns and their unique drivers.
4. **Anomaly Baseline Period Analysis:** Investigating the impact of different baseline periods for calculating temperature anomalies on the observed trends.

7.Appendices

References:

"Python for Data Analysis" by Wes McKinney

"Applied Time Series Analysis with Python: An Introductory Guide" by Benjamin Lowe

Data Set: https://drive.google.com/file/d/1IxuRnoAfV-LNVcpCt0Ya2CDsOP_tMhuz/view?usp=drive_link

Code: https://github.com/Tanudip572/Time-Series-Analysis-of-Global-Temperature-Data/blob/main/07_Time_Series_Analysis_of_Global_Temperature_Data_Spring_2026%20.ipynb