Understanding and Predicting Climate Change Metrics

1. Project Question

The main question guiding this project is: How can we understand and predict the impact of climate change metrics, specifically CO2 levels and temperatures, over time?

This involves analyzing historical climate data to identify trends and using machine learning models to predict future changes in CO2 concentrations and temperature averages.

2. Data Sources

The data sources for this project were chosen based on their comprehensive coverage of climate-related metrics and their credibility. The datasets used include:

- Global Land Temperatures by City: Provides historical temperature data for various cities around the world.
- Temperatures by Country: Contains historical temperature data aggregated by country.
- Temperatures by Major City: Focuses on temperature data for major global cities.
- Temperatures by State: Includes temperature records for states within countries.
- Global Temperatures: A global dataset for temperature records.

These datasets are sourced from Kaggle's climate change datasets and contain various climate metrics such as average temperature, CO2 levels, and timestamps. Each dataset was chosen for its relevance and the granularity of the data it provides, which is essential for detailed analysis and prediction.

Licenses: Most of these datasets are available under open-data licenses, which allow for usage and distribution with proper attribution. We ensured compliance with these licenses by acknowledging the sources in our documentation and not using the data for any commercial purposes.

3. Data Pipeline

Overview

The data pipeline is implemented using Python, leveraging libraries such as Pandas for data manipulation, Matplotlib and Seaborn for visualization, and Scikit-Learn for machine learning tasks. The pipeline involves several stages:

1. Data Loading: Importing the datasets into Pandas DataFrames.
2. Data Cleaning: Dropping missing values to ensure data integrity.
3. Date Format Modification: Converting date columns to extract and use the year for aggregations.
4. Data Transformation: Aggregating temperature data by year and region for more manageable and insightful analysis.
5. Data Visualization: Creating plots to visualize trends in temperature and CO2 levels.
6. Prediction: Using linear regression models to predict future CO2 levels based on historical data.

Transformation and Cleaning Steps

- Dropping NaN Values: All datasets underwent a cleaning process where rows with NaN values were removed to prevent inaccuracies in analysis.
- Date Handling: Date columns were converted to extract the year, which simplified the aggregation of temperature data on a yearly basis.
- Aggregation: For instance, temperature data was aggregated by averaging temperatures per year and filtering specific data, like focusing on the United States.

Problems and Solutions

- Missing Values: The primary issue was handling missing values, which was resolved by dropping

any rows with NaN values.

- Date Format Consistency: Different date formats were standardized to facilitate easier manipulation and analysis.

- Data Aggregation: Aggregating large datasets required efficient handling and computation, which was managed by utilizing Pandas powerful groupby and aggregation functions.

Error Handling

The pipeline includes basic error handling mechanisms to manage:

- Missing Data: Handled by dropping NaN values.

- Changing Input Data: The pipeline can be re-run with new data without requiring significant modifications, making it robust to changes in input data.

fig 2

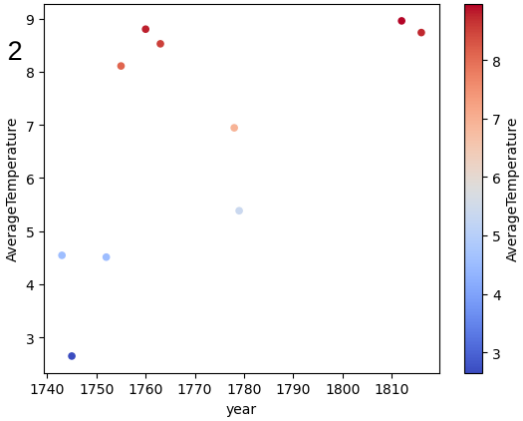


Figure 2 This scatter plot also shows the average temperature over the years but for an earlier time period.The temperatures vary over this time period, with some years showing relatively higher temperatures and others showing lower.

fig 1

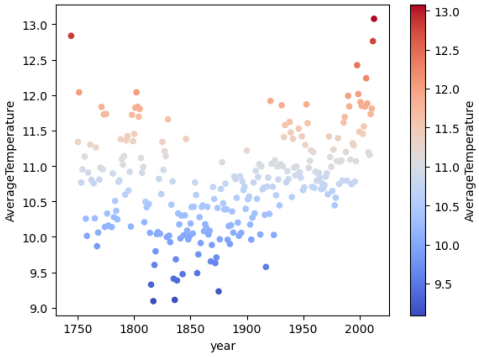


Figure 1 This scatter plot shows the average temperature over the years. There is a visible trend of increasing average temperatures over time, particularly in the most recent years.

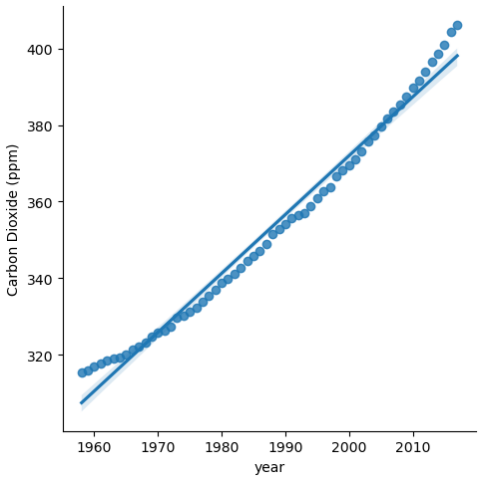


fig 3

Figure 3

This scatter plot shows the carbon dioxide (CO2) levels over the years, with a linear regression line indicating the trend.

The CO2 levels show a consistent increase over the years, with a steep upward trend.

4. Results and Limitations

Output Data

The output of the data pipeline consists of:

- Cleaned and Aggregated Datasets: Ready for analysis and visualization.

- Predictive Models: Linear regression models predicting future CO2 levels.

Data Structure and Quality

The resulting datasets are structured in a tabular format, with columns representing years, average temperatures, and CO2 levels. The quality of the data is high, given the extensive cleaning and aggregation processes applied.

Data Format

The output data is stored in CSV format due to its simplicity and compatibility with various tools for further analysis and visualization.

Reflections on Data and Potential Issues

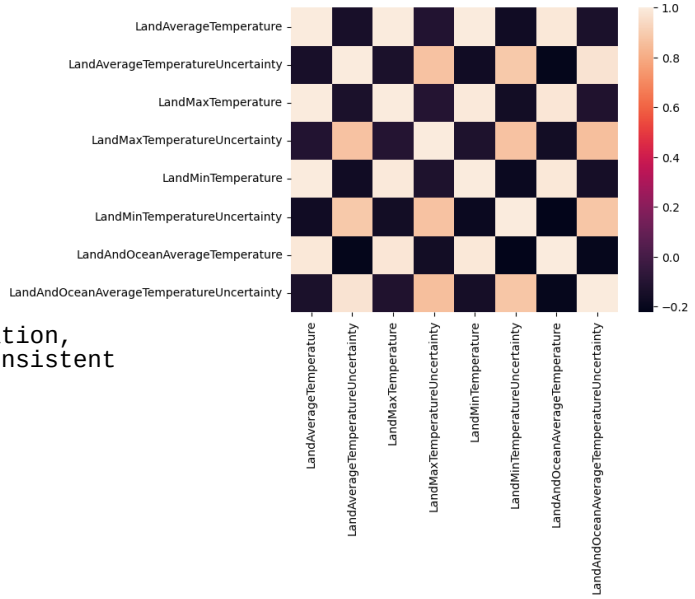
While the data and predictions provide valuable insights, there are limitations:

- Historical Data Limitations: Predictions are based solely on historical data, which may not account for unprecedented future events or policy changes affecting climate change.

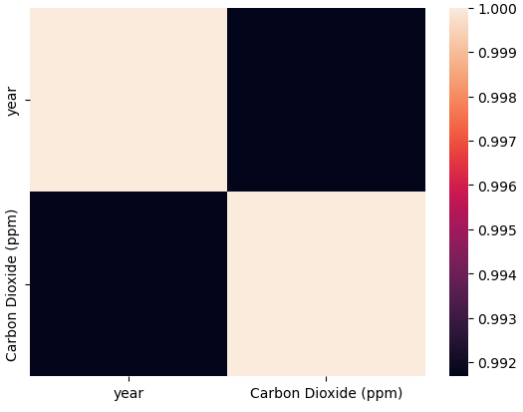
- Model Accuracy: The linear regression model, while useful, has its limitations and may not capture all the complexities of climate dynamics.

- Data Completeness: Despite efforts to clean the data, some inherent limitations in the original datasets (e.g., gaps in data collection) may impact the analysis.

Explaining visuals

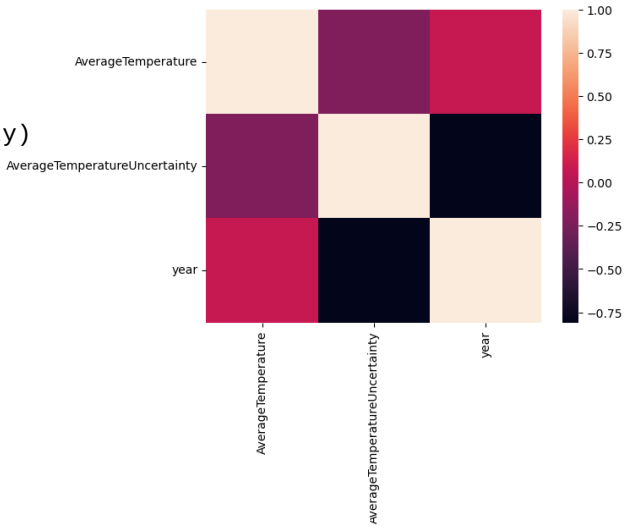


The heatmap reveals significant correlations between the different temperature metrics and their uncertainties.
Land average temperature and land max temperature show a strong positive correlation.
Temperature uncertainties also show a significant correlation, which indicates that uncertainties in measurements are consistent across different types of temperature data.
Understanding these correlations helps in comprehending the interdependencies between various climate metrics.

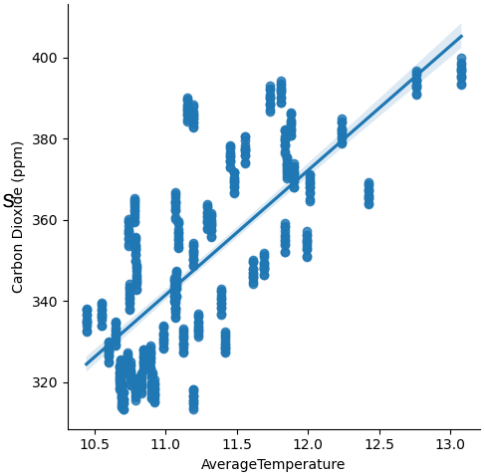


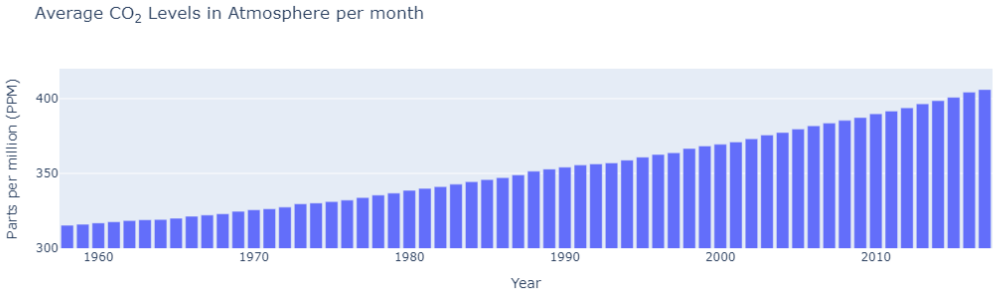
The x-axis represents the year.
The y-axis represents the CO2 concentration in ppm.
The color scale indicates the correlation values, with dark colors representing higher correlations.
There is a very high correlation between the year and CO2 levels, indicating that as time progresses, the CO2 levels increase.

Strong positive correlations (near 1.0) are evident among similar types of temperature measurements and their uncertainties. For example, LandAverageTemperature is strongly correlated with LandMaxTemperature and LandMinTemperature. The uncertainty measures (e.g., LandAverageTemperatureUncertainty) do not exhibit strong correlations with actual temperature measures, as expected, since uncertainties should be less dependent on the measurements themselves.

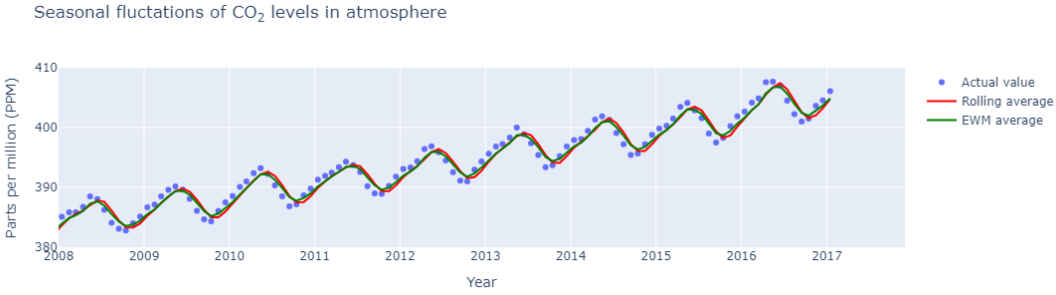


There is a positive linear relationship between AverageTemperature and Carbon Dioxide levels. As AverageTemperature increases, CO_2 levels also tend to increase. This supports the hypothesis that higher temperatures are associated with higher concentrations of CO_2 in the atmosphere, consistent with the greenhouse effect.

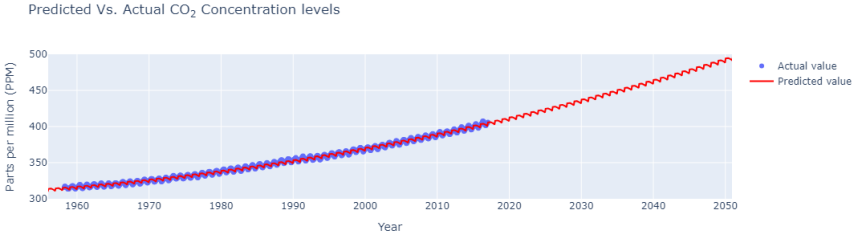




As the average temperature increases, the CO2 levels also tend to increase. This suggests that higher temperatures might be associated with higher CO2 concentrations, potentially due to increased emissions from various sources.



The chart reveals a cyclical pattern in CO2 levels, with regular peaks and troughs each year. This seasonality is likely due to natural processes such as plant growth and decay cycles. Despite the seasonal fluctuations, the overall trend shows an increase in CO2 levels over time.



The predicted values closely follow the actual values, indicating that the model is performing well. The model predicts a continued increase in CO2 levels into the future, which aligns with historical trends. If the current trend continues, CO2 levels are expected to reach significantly higher levels by 2050, which could have severe implications for the climate.