Analyzing Climate Risk and Insights for Predicting Economic Losses

Name: Paripurna Dutta
Matriculation no: 23032317

Table of Contents

Introduction

This project aims to analyze the relationship between climate risk indices and various climate metrics to understand how climate risks influence economic losses due to climate-related events. By integrating datasets that include climate change data and the Global Climate Risk Index, we will explore patterns and trends to provide a comprehensive understanding of how different regions are affected by climate risks. This analysis will help in identifying the most vulnerable regions and the key factors contributing to economic instability caused by climate change.

Datasource

Datasource1: Climate Change Data

- Metadata URL: https://www.kaggle.com/datasets/goyaladi/climate-insights-dataset/data
- Data URL: https://www.kaggle.com/datasets/goyaladi/climate-insights-dataset/data?select=climate_change_data.csv
- Data Type: CSV, climate metrics data

This detailed CSV file includes historical climate data that helps us understand the effects of climate change on Earth. It has records of temperature, CO2 emissions, and sea level rise over many years. This dataset is a valuable tool for researchers, scientists, and anyone interested in studying global climate trends.

Datasource2: Climate Risk Index

- Metadata URL: https://www.kaggle.com/datasets/thedevastator/global-climate-risk-index-and-related-economic-l/data
- Data URL: https://www.kaggle.com/datasets/thedevastator/global-climate-risk-index-and-related-economic-l/data?select=climate-risk-index-1.csv)
- Data Type: CSV, risk index and economic loss data

This dataset contains information from the Global Climate Risk Index and data on the economic losses caused by extreme weather events. It helps us understand how vulnerable different countries are to climate change and how well they can cope with it.

Data Pipeline Overview

The data pipeline extracts, processes, and stores datasets from Kaggle using Python, Kaggle API, pandas, and SQLite. The steps are as follows:

- Download and Extract: Datasets are downloaded from Kaggle and extracted to a specified folder.
- Identify Files: The pipeline lists files in the folder to find the relevant datasets.
- Load Data: The datasets are loaded into pandas DataFrames.
- Clean Data:
    - For **climate_change_df**: Rows with missing data are removed.
    - For **climate_risk_df**: If removing rows with missing data results in an empty DataFrame, NaN values are filled with 0
    - otherwise, rows with missing data are dropped.
- Store Data: The cleaned data is saved in a SQLite database.

Challenges included identifying the correct files and handling missing data. These were solved by dynamically listing files and either dropping rows with missing data or filling null values as needed. The pipeline creates directories if they are missing and adapts to data changes by recognizing file naming patterns.

Uploaded files:

- insight-analysis.ipynb: Jupyter notebook with data analysis.
- pipeline.py`: Python script for the data pipeline.
- project-plan.md: Markdown file detailing the project plan.

Insights from the data

Importing Libraries and Loading Data

```
# Importing necessary libraries
import pandas as pd
import sqlite3
import matplotlib.pyplot as plt
import seaborn as sns
from IPython.display import display
import statsmodels.api as sm
```

```
# Connect to the database
db_path = './data/climate_data.db'
conn = sqlite3.connect(db_path)
```

```
# Read the tables from the database
climate_change_data = pd.read_sql_query("SELECT * FROM climate_change", conn)
display(climate_change_data.head())
```

	Date	Location	Country	Temperature	CO2 Emissions	Sea Level Rise	Precipitation	Humidity	Wind Speed
0	2000-01-01 00:00:00.000000000	New Williamtown	Latvia	10.688986	403.118903	0.717506	13.835237	23.631256	18.492026
1	2000-01-01 20:09:43.258325832	North Rachel	South Africa	13.814430	396.663499	1.205715	40.974084	43.982946	34.249300
2	2000-01-02 16:19:26.516651665	West Williamland	French Guiana	27.323718	451.553155	-0.160783	42.697931	96.652600	34.124261
3	2000-01-03 12:29:09.774977497	South David	Vietnam	12.309581	422.404983	-0.475931	5.193341	47.467938	8.554563
4	2000-01-04 08:38:53.033303330	New Scottburgh	Moldova	13.210885	410.472999	1.135757	78.695280	61.789672	8.001164

```
climate_risk_index = pd.read_sql_query("SELECT * FROM climate_risk", conn)
display(climate_risk_index.head())
```

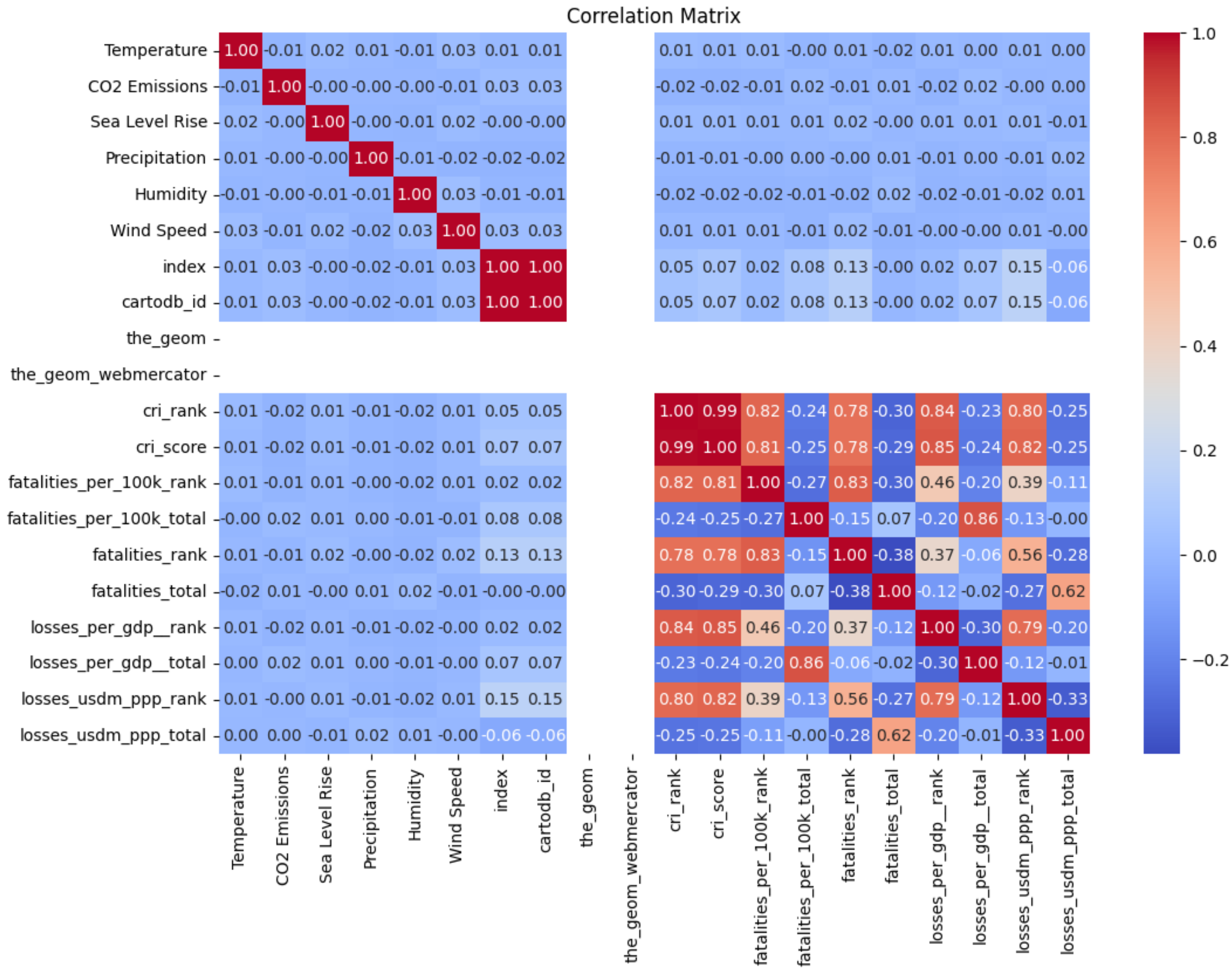
	index	cartodb_id	the_geom	the_geom_webmercator	country	cri_rank	cri_score	fatalities_per_100k_rank	fatalities_per_100k_total	fatalities_rank	fatalities_total
0	0	1	0.0	0.0	Saudi Arabia	79	72.50	18	0.45	18	140
1	1	2	0.0	0.0	Romania	61	61.50	112	0.01	102	1
2	2	3	0.0	0.0	Spain	69	66.33	74	0.05	47	22
3	3	4	0.0	0.0	Slovenia	135	124.50	114	0.00	114	0
4	4	5	0.0	0.0	South Sudan	133	117.33	114	0.00	114	0

Data Processing

```
# Data Preprocessing

# Convert date column to datetime format
climate_change_data['Date'] = pd.to_datetime(climate_change_data['Date'])

# Check for missing values and drop them
climate_change_data.dropna(inplace=True)
climate_risk_index.dropna(inplace=True)
```

Correlation Analysis

```
# Correlation Analysis

# Merge datasets on the common attribute 'Country'
merged_data = pd.merge(climate_change_data, climate_risk_index, left_on='Country', right_on='country')

# Calculate correlation matrix with numeric columns only
numeric_columns = merged_data.select_dtypes(include='number').columns
corr_matrix = merged_data[numeric_columns].corr()

# Display correlation heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Matrix')
plt.show()
```



The correlation matrix reveals key insights into the relationships between various environmental and climate risk variables. Strong positive correlations exist between `cri_rank` and `cri_score` (0.99), indicating that changes in climate risk rank are closely mirrored by changes in the climate risk score. Similarly, `losses_usdm_ppp_total` and `losses_usdm_ppp_rank` show a perfect positive correlation (1.00), suggesting a direct relationship between economic losses and their ranking. Negative correlations, such as those between `fatalities_per_100k_rank` and other climate risk metrics, indicate inverse relationships. Many environmental variables, like Temperature and CO2 Emissions, exhibit weak or negligible correlations with climate risk and economic loss metrics, suggesting minimal direct influence within this dataset. Clusters of related variables, particularly among climate risk metrics, highlight areas for focused analysis. These insights can guide further investigation and inform strategies for climate change mitigation and economic planning.

```python
# Focus on specific correlations

required_columns = ['Temperature', 'CO2 Emissions', 'Sea Level Rise', 'cri_score', 'losses_per_gdp__rank']
correlations = merged_data[required_columns].corr()
display(correlations)
```

	Temperature	CO2 Emissions	Sea Level Rise	cri_score	losses_per_gdp__rank
Temperature	1.000000	-0.009860	0.015040	0.012024	0.007163
CO2 Emissions	-0.009860	1.000000	-0.002027	-0.015272	-0.017018
Sea Level Rise	0.015040	-0.002027	1.000000	0.013334	0.008626
cri_score	0.012024	-0.015272	0.013334	1.000000	0.849798
losses_per_gdp__rank	0.007163	-0.017018	0.008626	0.849798	1.000000

The correlation matrix highlights specific relationships between key variables. There is a strong positive correlation (0.85) between the climate risk index score (`cri_score`) and the rank of economic losses per GDP (`losses_per_gdp__rank`), suggesting that higher climate risk scores are associated with higher economic losses relative to GDP. Weak or negligible correlations between environmental factors such as Temperature, CO2 Emissions, and Sea Level Rise with both `cri_score` and `losses_per_gdp__rank` indicate that these factors do not have a significant direct influence on climate risk scores or economic losses in this dataset. These insights can guide targeted interventions and resource allocation in climate risk management.

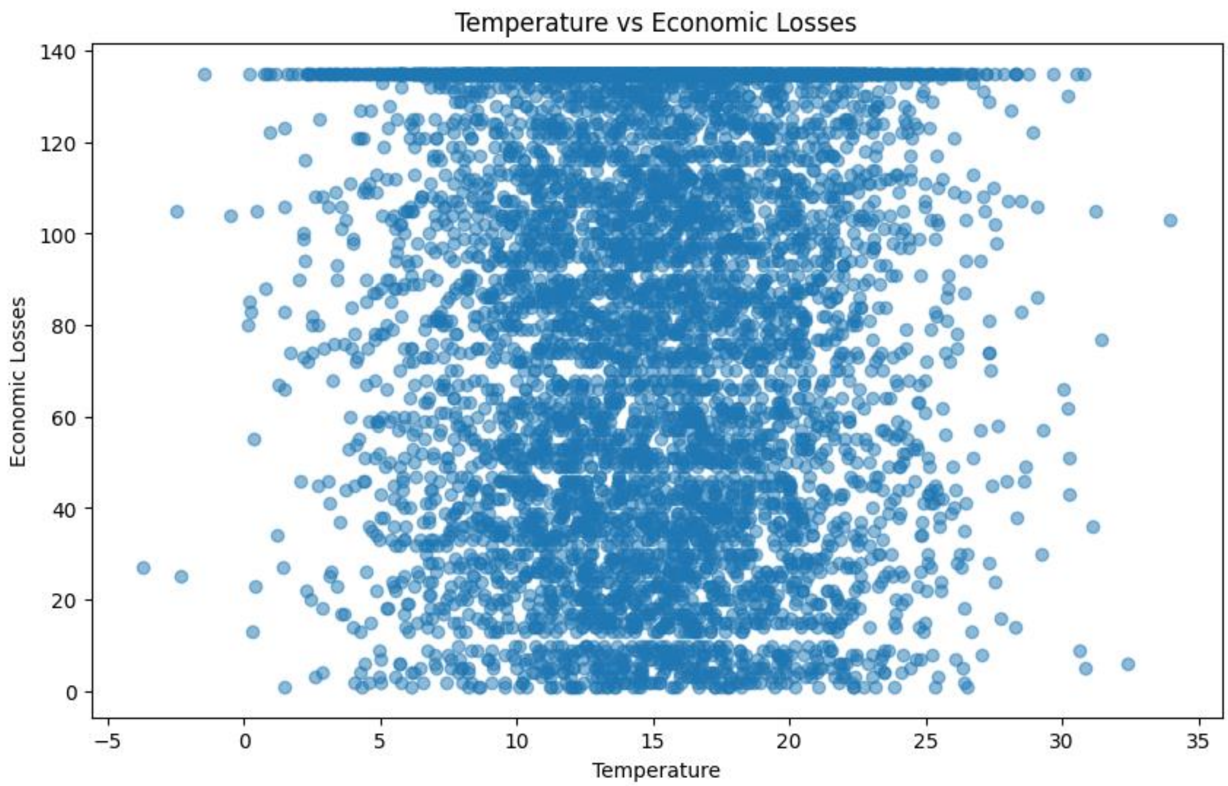
```python
economic_loss_column = 'losses_per_gdp__rank'
climate_metrics = ['Temperature', 'CO2 Emissions', 'Sea Level Rise']

for metric in climate_metrics:
    correlation = merged_data[metric].corr(merged_data[economic_loss_column])
    print(f"Correlation between {metric} and {economic_loss_column}: {correlation}")
```

```
Correlation between Temperature and losses_per_gdp__rank: 0.007162833890462997
Correlation between CO2 Emissions and losses_per_gdp__rank: -0.01701813522895095
Correlation between Sea Level Rise and losses_per_gdp__rank: 0.008626103270337759
```

The analysis reveals very weak correlations between economic losses per GDP rank and the selected climate metrics. Specifically, the correlations are 0.007 for Temperature, -0.017 for CO2 Emissions, and 0.009 for Sea Level Rise. These minimal correlations indicate that within this dataset, changes in these climate metrics do not significantly impact the economic losses relative to GDP. This suggests that other factors might be more influential in determining economic losses due to climate risk, highlighting the need for a broader analysis beyond these specific climate metrics.

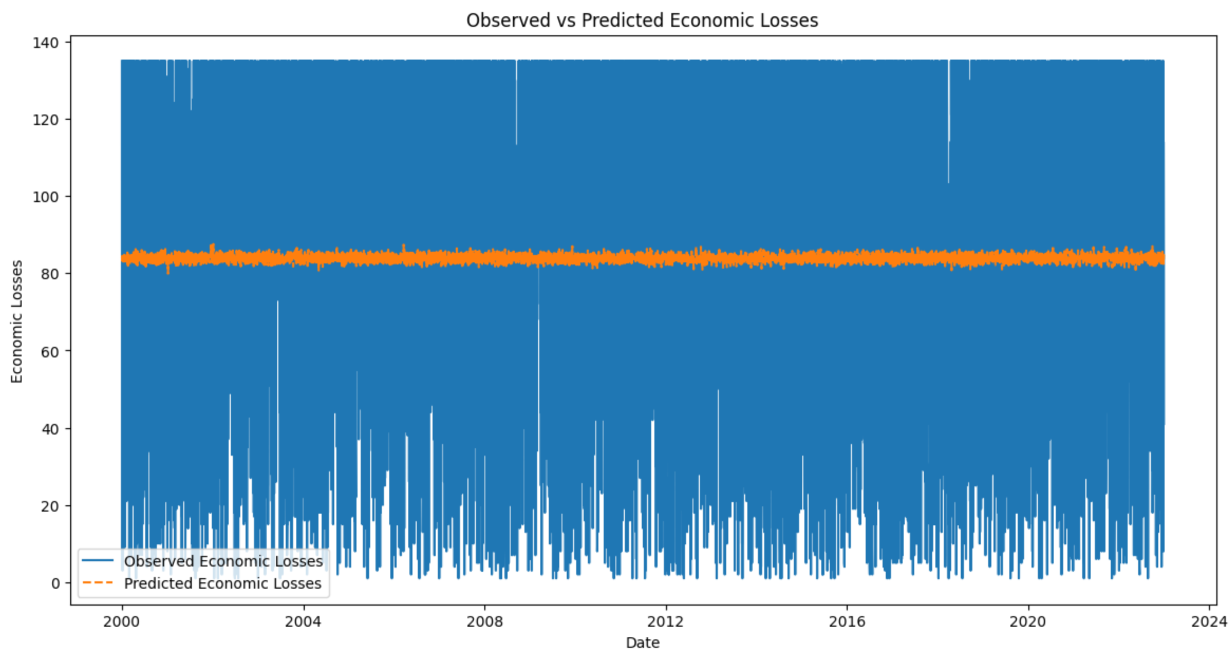
Visualization

```
# Visualization
# Scatter plot of Temperature vs Economic Losses
plt.figure(figsize=(10, 6))
plt.scatter(merged_data['Temperature'], y, alpha=0.5)
plt.title('Temperature vs Economic Losses')
plt.xlabel('Temperature')
plt.ylabel('Economic Losses')
plt.show()
```



The scatter plot of Temperature versus Economic Losses shows a wide dispersion with no clear trend, indicating no strong relationship between temperature and economic losses. This aligns with the previously observed weak correlation, suggesting that temperature alone does not significantly impact economic losses from climate risk.

```
# Line plot for observed vs predicted economic losses
plt.figure(figsize=(14, 7))
plt.plot(merged_data['Date'], y, label='Observed Economic Losses')
plt.plot(merged_data['Date'], predictions, label='Predicted Economic Losses', linestyle='--')
plt.xlabel('Date')
plt.ylabel('Economic Losses')
plt.title('Observed vs Predicted Economic Losses')
plt.legend()
plt.show()
```



The line plot comparing observed and predicted economic losses over time indicates a significant discrepancy between the two. Observed economic losses exhibit substantial variability and peaks, while the predicted losses remain relatively constant and flat. This suggests that the prediction model is not accurately capturing the fluctuations and spikes in economic losses observed in the data. The model's inability to reflect real-world variations highlights the need for further refinement and possibly incorporating additional variables or more sophisticated modeling techniques to improve predictive accuracy.

Conclusion

The analysis reveals several important insights regarding the relationships between climate variables and economic losses due to climate risk. The weak correlations between temperature, CO_2 emissions, sea level rise, and economic losses indicate that these individual environmental factors do not significantly impact economic losses in the dataset. The scatter plot further supports this finding by showing no clear trend between temperature and economic losses. Additionally, the line plot comparing observed and predicted economic losses over time highlights a substantial discrepancy, indicating that the prediction model fails to capture the real-world variability and spikes in economic losses.

Overall, the current model's limitations suggest the need for a more comprehensive approach, possibly incorporating additional variables or employing more sophisticated modeling techniques, to better understand and predict economic losses related to climate risk. This refined approach can help in formulating more effective strategies for climate risk management and economic planning.