The Effects of Climate Change on Natural Disasters and Global Development

Project Group #2
Members: Safa Sajid, Sadia Khan Durani, Mariam Virk

Project Idea and Importance

Project Idea: To analyze the influence of climate change on natural disasters and its effects on global development. The research focuses on how countries are economically and socially affected by natural disasters, investigating the link between these disasters and climate change.

The motivation behind this idea comes from the need to develop effective disaster preparedness strategies, formulate robust economic policies, and create social support systems to mitigate the adverse impacts of natural disasters and climate change on global development. By establishing the correlation between climate change and the severity/frequency of natural disasters, this project advocates for urgent global action on climate change. The increasing frequency of natural disasters, such as hurricanes and floods, underscores the necessity for enhanced disaster planning and prevention measures. The economic impacts, including costs such as rebuilding infrastructure are key to creating effective economic policies and support systems. As disasters exacerbated by climate change pose challenges to sustainable development, particularly in developing nations with limited resources, we aim for this project to serve as a guiding compass for international development and aid strategies. We hope that we can utilize its findings to facilitate disaster risk reduction, climate change mitigation, and adaptation strategies.

Ultimately, our project outcomes include:

(1) offering a comprehensive analysis of the economic and social impacts of natural disasters in a changing climate in North American countries
(2) stressing the need for decisive action against climate change to mitigate its disastrous impacts.

Explanation and Justification of the Selected Datasets

Climate Dataset:
www.kaggle.com/datasets/goyaladi/climate-insights-dataset?select=climate_change_data.csv
The climate insights dataset was chosen for its specific focus on climate variables like temperature and CO2 emissions across North America's major countries, this dataset is key for analyzing climate change over the past two decades.
Why? This dataset gives us detailed temperature and CO2 data, essential for examining climate change's role in natural disasters.
Factors Considered: We looked for data covering our target region and period (2000-2020) and essential climate metrics that directly relate to our research on natural disasters.

Disasters Dataset:

https://www.kaggle.com/datasets/jnegrini/emdat19002021/data
This natural disasters dataset was chosen because it offers an extensive record of natural disasters, letting us explore the trends and effects within the same countries and timeframe as our climate data.
Why? It's critical for cross-referencing with climate trends to explore potential links to disaster events.
Factors considered: The dataset's extensive record of disasters, credibility of the source,, and the ability to filter by country and date range were key decision points.

Human Development/Population Dataset:
https://ourworldindata.org/grapher/augmented-hdi-vs-gdp-per-capita
Lastly, the human development and population dataset was chosen because it includes socio-economic indicators (AHDI and GDP per capita) for our specified countries and timeframe, crucial for understanding the broader impacts of climate-influenced natural disasters.
Why? It helps us connect climate change, disaster impact, and socio-economic development.
Factors Considered: We needed a dataset that offers a multidimensional perspective on development that aligns with our climate and disaster data for both geography and time frame.

Each dataset was chosen based on its alignment with our geographic areas of focus, time frame, and relevance to our research objectives. This was to ensure that our datasets work together to provide a comprehensive analysis of the effects of climate change on natural disasters and socio-economic progress.

Data Cleaning Steps Prior to the Analysis

The following data cleaning steps were made to ensure data were correctly referenced across the 3 datasets. Our data cleaning processes utilized Python libraries such as pandas, numpy for data manipulation.

Climate Dataset

1. Filtered for data specific to Canada, the USA, and Mexico.
2. Converted date information to just the year for annual analysis.
3. Limited the dataset to the years 2000-2020.
4. Removed the 'Location' column as it was unnecessary.
5. Extracted 'Year' data from the datetime column to focus on yearly data at a country level.
6. Standardized all variations of "United States of America" to "United States".
7. Averaged temperature and CO2 emissions for duplicate (Year, Country) entries.
8. Selected the maximum value for other columns with duplicates.
9. Reintegrated averaged data back into the main dataset.
10. Grouped data by 'Country' and 'Year' to ensure unique entries.
11. Verified there were no missing entries in the dataset.

Human Development/Population Dataset

1. Loaded and filtered data for the target countries (Canada, USA, Mexico)
2. Limited the data to the 2000-2020 range.
3. Renamed columns for clarity.
4. Ensured a unique year-country entry for consistency with primary keys.
5. Checked for any missing data.

Disasters Dataset

1. Loaded and filtered data for Canada, the USA, and Mexico.
2. Restricted the dataset to the years 2000-2020.
3. Selected columns relevant to the analysis.
4. Renamed columns to enhance clarity.
5. Reordered columns logically.
6. Standardized the country name for the US ("United States") using a mapping dictionary.
7. Translated numerical month encodings to month names.
8. Checked for missing data entries.

The following steps were taken to create our DDL file needed to populate the Oracle Database.

Database Operations:

1. Hard-coded the 'CREATE TABLE' statements corresponding to our ER diagram
2. Implemented a function to generate 'INSERT' statements using our datasets
    - Included the conversion of null values in the form of 'nan' to 'NULL' to ensure consistency with Oracle Database.
3. Implemented a function to write the 'INSERT' statements to a sql file
4. Generated an SQL file with the statements for our 5 tables based on our relational schema

Recall: Relational Schema

Climate_Metrics(*Country_Name*, *Year_Recorded*, Max_Temperature, Avg_Temperature, Max_CO2_Emissions, Avg_CO2_Emissions, Sea_Level_Rise, Humidity, Wind_Speed, Precipitation)

Socio_Economic_Indicators(*Socio_Country*, *Socio_Year*, AHDI, GDP_Per_Capita, Population)

Natural_Disaster(*Disaster_No*, **Disaster_Country, Disaster_Year**, Disaster_Month, Disaster_Subgroup, Disaster_Type, Disaster_Subtype)

- Disaster_Country, Disaster_Year references Climate_Metrics(Country_Name, Year_Recorded)

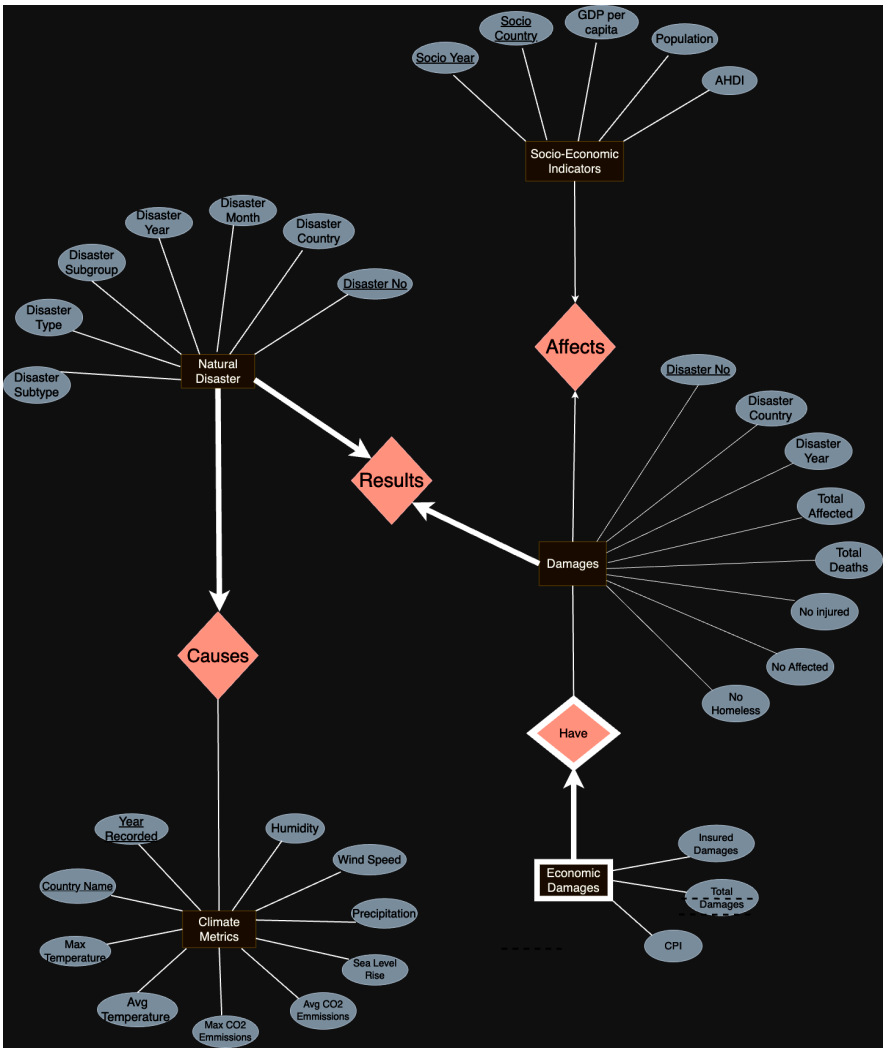
Damages(***Disaster_No*, Disaster_Country, Disaster_Year**, Total_Affected, Total_Deaths, No_Injured, No_Affected, No_Homeless)

- Disaster_No references Natural_Disaster(Disaster_No)
- Disaster_Country, Disaster_Year references Socio_Economic_Indicators(Socio_Country, Socio_Year)

Economic_Damages(***Disaster_No***, *Total_Damages*, Insured_Damages, CPI)

- (Disaster_No, Total_Damages) is the primary key for the weak entity where Disaster_No comes from its super entity and Total_Damages is its partial key

FINAL ENTITY RELATIONSHIP (ER) DIAGRAM:



In our initial version of the ER diagram, we had separate entities for Country, Natural Disaster, Climate Metrics, Socio-Economic Indicators, Damages, and Economic Damages as a weak entity. This version included 5 relationships between the entities and an additional weak entity relationship between Damages and Economic Damage. After some reflection, we came up with a revised version (shown above) that omitted the Country entity and instead included it as an attribute for the remaining entities. This decision was made to better align our ER diagram with our data records used as each climate metric data, natural disaster data, and socio-economic indicator would be recorded for a single country.

- These revisions reduced the number of tables needed in our relational schema.
- Although this resulted in duplicate attributes across different entities, we used foreign keys to reduce this duplication. E.g., Country_Name, Disaster No, etc.

Cardinality Constraints:
Natural Disaster to Climate Metrics: Many to One
Natural Disaster to Damages: One to One
Socio-Economic Indicators to Damages: One to One
Weak Entity ⇒ Damages to Economic Damages: One to Many

Participation Constraints:
Participation of Natural Disaster in 'Causes' must be total
Participation of Climate Metrics in 'Causes' can be partial

Participation of Natural Disaster in 'Results' must be total
Participation of Damages in 'Results' must be total

Participation of Socio-Economic Indicators in 'Affects' can be partial
Participation of Damages in 'Affects' can be partial

Participation of Damages in 'Have' is partial but participation of Economic Damages in 'Have' must be total as it is the weak entity.

Relationships Discussion:
Our choice to represent "Natural Disaster" as a distinct entity, rather than having it as an attribute under 'Damages' entity, was influenced by our aim to encapsulate a wide range of information specific to each Natural Disaster. By assigning the "Natural Disaster" entity attributes like 'Disaster Type', 'Disaster Subtype', etc., allows us to categorize disasters while maintaining a focus on the essential data related to their impacts. This allowed us to create visualizations by disaster types further in our analysis. Additionally, based on a conceptual level, it made sense to keep "Natural Disaster" separate from "Climate Metrics" and "Damages."

We encoded a many to one relation between "Natural Disaster" and "Climate Metrics" as a natural disaster must be caused by a climate metric. A natural disaster can be caused by one climate metric (since we have one set of climate metrics data per year for a country), but a climate metric can cause multiple natural disasters.

We encoded a one to one relation between "Natural Disaster" and "Damages" as a natural disaster can cause one set of damages (since this comes from the same dataset in our project) and one damage is caused by one natural disaster.

We encoded a one to one relation between "Socio-Economic Indicators" and "Damages" since one socio-economic indicator should be mapped to one set of damages that occurred in a country in a particular year.

Data Analysis Methodology

For our project, we adopted a comprehensive data analysis approach. From starting with researching for reliable datasets, using Pandas and Numpy for data cleaning and preparation, creating a SQL file to populate our database, then using Oracledb for data extraction, and finally, Altair and statsmodels.API for generating visualizations and running a regression model.

Our chosen methodology was designed to help uncover insights into how climate change influences climate metrics such as temperature, and how those factors can cause natural disasters and inhibit global development. Beginning with a graphical analysis approach provided us with an overview of our data structure and distribution, and from there we plotted relevant variables to derive relationships, finally running 2 regression models in an attempt to solidify our findings and ensure statistical rigour and interpretability of our findings.

By representing data visually, we wanted to be able to easily identify trends, outliers, and patterns that might be hidden in data tables. The visualizations aided in conveying complex data to a broader audience, so it could be meaningful and accessible. We implemented line plots, area charts, bar charts, and pie charts to capture the multifaceted nature of our data. Line and area charts were particularly effective in visualizing time series data, revealing trends and changes over time, especially in situations where there was a serial lag in association between variables, which can't easily be discovered through datasets or regression models. We used bar and pie charts to provide a clear and immediate understanding of the distribution and impact of different disaster types, specifically allowing us to see the most impactful disaster that we could focus our efforts in exploring.

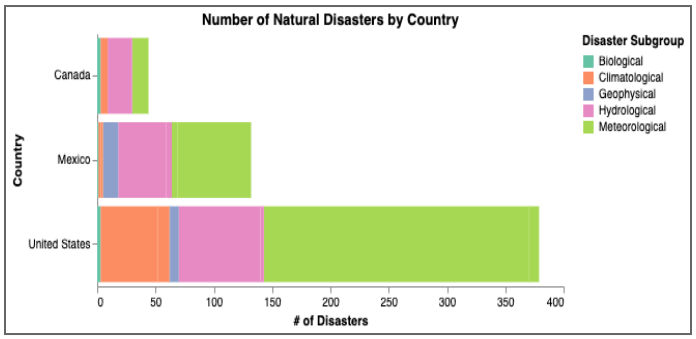
Linear regression models can be used to understand associations between quantitative independent variables to a dependent variable. Since our data consisted of many quantitative variables, we chose this statistical technique to look for a relationship between sea level rise, temperature, CO2 emissions, etc. To learn the best variables that are related to sea level rise, we used a technique called *forward selection.* This algorithm begins by fitting a linear regression model with no predictor variables and adds one variable at a time, keeping the significant variables when moving forward. This provided us with the best performing model for predicting sea level rise and ensured statistical rigour of our findings.

Visualizations, Results and Final Discussion

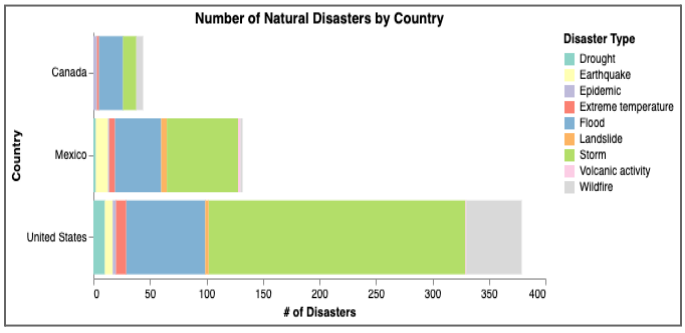
As our project question had multiple parts - exploring the relationship between climate change and natural disasters, and then exploring the relationship between damages from natural disasters versus a country's socio-economic status, our analysis and results are also in multiple parts.

To begin, we initially explored the most common natural disasters and climate change indicators in our countries of interest, to focus our project on those factors. Through some exploratory data analysis, we found that in Canada, United States, and Mexico (our 3 countries of choice), the most common natural disasters were meteorological (Visualization #1), and more specifically, floods and storms (Visualization #2). These are also the disasters that caused the most damages and most deaths in all the countries (Visualization #3, #4). It is clear that these disasters were crucial to focus on, especially after seeing the total cost in damages was over $400,000,000 USD. Since these disasters are caused by sudden and destructive changes in Earth's water/distribution of water, it is clear that the nature of these disasters is easily affected by climate change, and we further explored this.

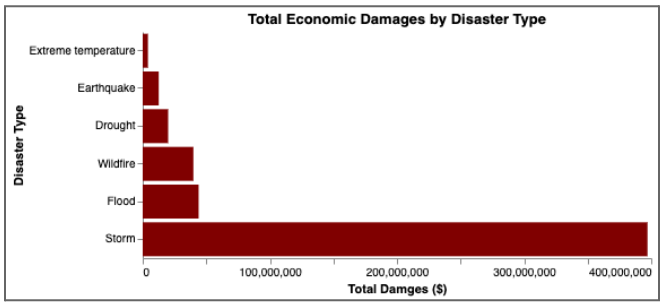
Visualization #1



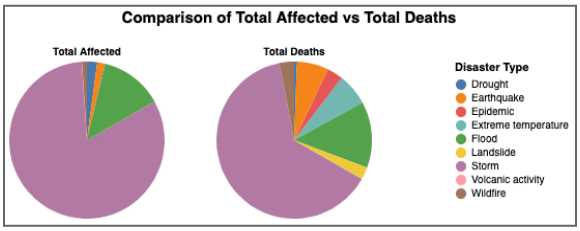
Visualization #2



Visualization #3

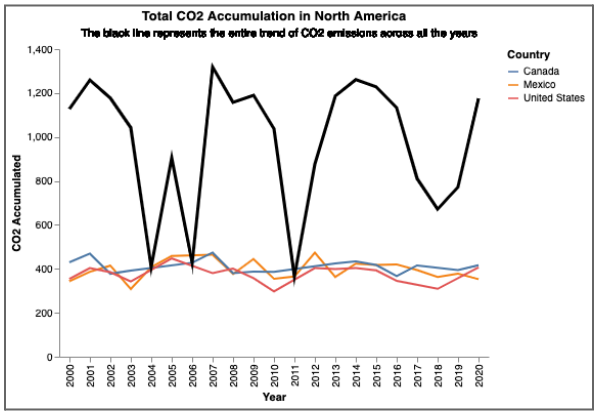


Visualization #4



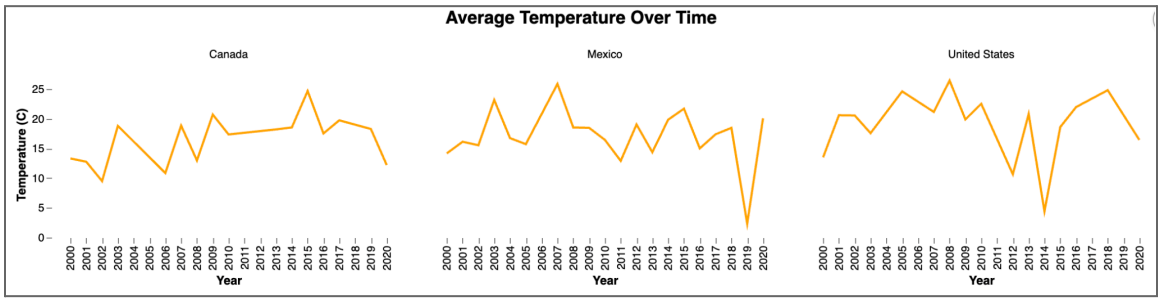
We began our analysis by exploring whether C02 is being emitted at a consistent or increasing rate by the countries in question. By plotting a line graph of Canada, United States, and Mexico's yearly C02 emissions, as well as the combined total C02 emissions (in black), we see that although there may be a few dips in the amount of emissions, it is still being produced at a consistent rate. (Visualization #5). This is extremely concerning, because it is clear that the amount of C02 being produced and accumulating in the atmosphere has been increasing over the past 20 years. Having confirmed this through our visualization, we moved forward to see if the emissions were affecting the countries' temperature.

Visualization #5



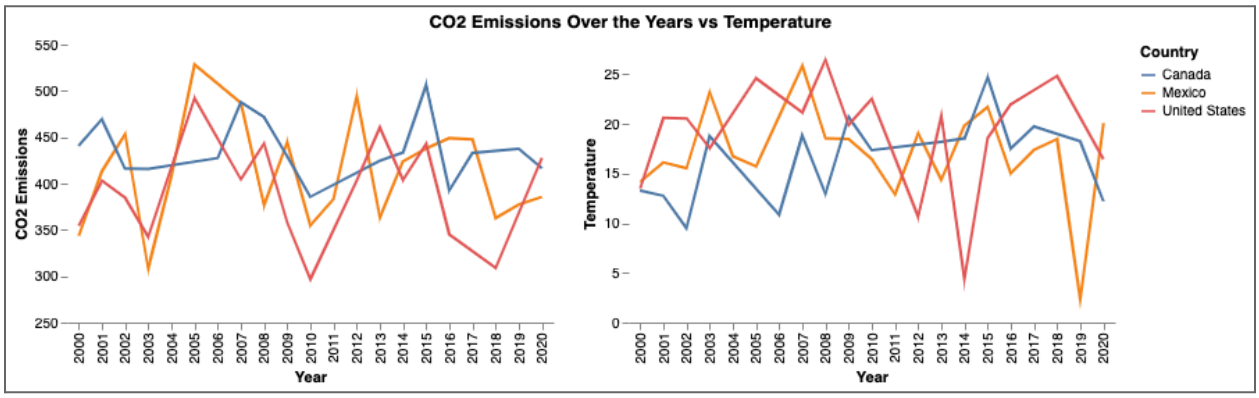
We plotted the annual average temperature for each country as a line graph, to assess the serial changes over 20 years (Visualization #6). Although there are some dips in temperature, for example, in Mexico 2019, we realized this resulted from the way we removed duplicated (County, Year) pairs from our Climate's dataset as we focused on yearly data from a country level rather than specific locations. This caused some unexpected spikes and drops in temperature, as the temperature may have been recorded in a warmer/colder city, but when the average was taken, it didn't separate by warmer/colder locations. We also need to remove duplicates in order to have a valid primary key for the Climate_Metrics entity. From averaging out the dips and spikes in temperature, it is clear that there is a slight upward trend, and overall the temperatures have been increasing. Having this visual confirmation, we continued on to see if the temperature increase was correlated with the C02 emissions.

Visualization #6



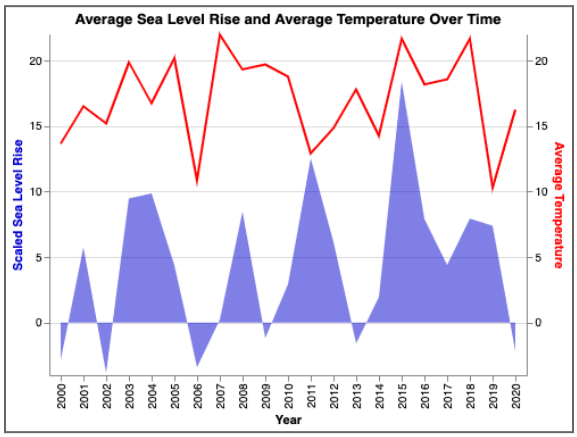
We created an interactive line graph of C02 emissions per country, with a move-able window that allows the user to see the corresponding temperature in a side-by-side format (Visualization #7). When comparing in this manner, we can see that when there is a rise in C02 emissions, the temperature also starts to rise about 2-3 years after. This implies that there is a relationship between C02 emissions and average temperature - as C02 emissions increase, the average temperature also increases, and this begins to intensify storms, raise sea levels, and more. This relationship has a 2-3 year lag, and was only apparent once we plotted the graphs side by side. A method to improve this diagnosis would be to manually adjust one of the graphs by 2-3 years, to directly compare them. Now that we could see a relationship between C02 and temperatures, implying that C02 affects temperature, we moved on to explore whether temperature increases would affect sea level rising (which is one of the more relevant climate change factors in North America)

Visualization #7



We plotted an area chart to show the overall sea levels rise and fall for North America, along with a line graph layered over that showed the temperature over time (Visualization #8). From the layered temperature red line over the sea level rise, we see that in general, when temperature rises, it is followed by sea level rising 1-2 years after. Similarly, when temperature falls, it is followed by sea level falling 1-2 years after. Additionally, note that the general trend of sea level rising is increasing, which makes sense given that we established that the C02 in the atmosphere is also increasing over time.

Visualization #8



After visually confirming that sea levels were rising, we used forward selection to derive a regression model that would allow us to see exactly which factors aside from only temperature were affecting sea level rise (Regression Model 1). After running the algorithm, we found that the factors were maximum/average temperature and c02 emissions, humidity, and GDP per capita. We also plotted a residual plot to see how well this model performed,

and although the residuals were relatively random, we noticed the Adjusted R^2 value was lower than expected (0.335). This is perhaps due to the fact that these explanatory variables would take a few years to affect sea levels, and in the future we would confirm this by adjusting the years to account for this delay.

Regression Model 1:

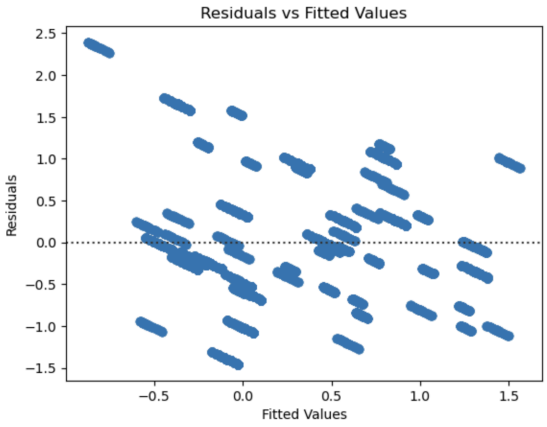
OLS Regression Results

Dep. Variable:	Sea_Level_Rise	R-squared:	0.335
Model:	OLS	Adj. R-squared:	0.335
Method:	Least Squares	F-statistic:	2969.
Date:	Thu, 28 Mar 2024	Prob (F-statistic):	0.00
Time:	17:38:33	Log-Likelihood:	-39676.
No. Observations:	35340	AIC:	7.937e+04
Df Residuals:	35333	BIC:	7.943e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.7031	0.058	63.932	0.000	3.590	3.817
Humidity	0.0012	0.000	6.810	0.000	0.001	0.002
Max_Temperature	0.1129	0.002	49.323	0.000	0.108	0.117
Avg_Temperature	-0.1207	0.002	-51.106	0.000	-0.125	-0.116
GDP_Per_Capita	-1.559e-05	2.96e-07	-52.669	0.000	-1.62e-05	-1.5e-05
Avg_CO2_Emissions	-0.0216	0.000	-62.332	0.000	-0.022	-0.021
Max_CO2_Emissions	0.0135	0.000	49.149	0.000	0.013	0.014

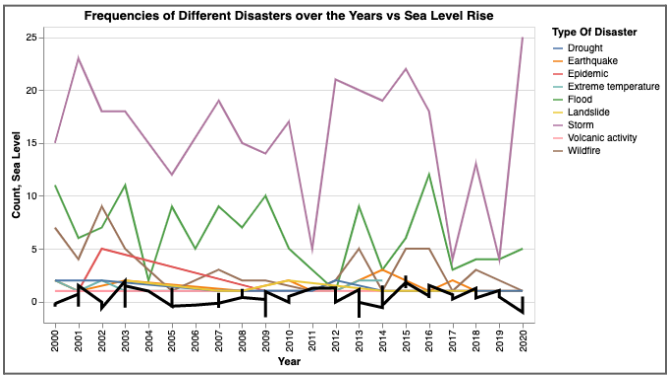
Omnibus:	1039.368	Durbin-Watson:	0.097
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1132.639
Skew:	0.438	Prob(JB):	1.12e-246
Kurtosis:	2.943	Cond. No.	6.65e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.65e+05. This might indicate that there are strong multicollinearity or other numerical problems.



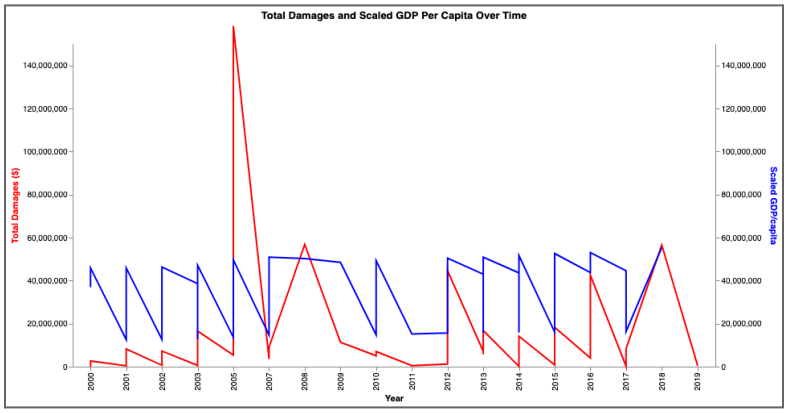
Having seen that sea level rises are correlated with temperature as well as other variables, we further explored to see if the sea levels changes were related to the natural disasters, through a line graph (Visualization #9). Once again, we see a relationship in the rise and falls of sea levels with the rise and falls of storms and floods, but it is delayed by 2-3 years. Having established this, we finally explored whether the damages caused by these natural disasters affected the GDP of the North American countries.

Visualization #9



By plotting a line graph of GDP per Capita overlaid with the Total Damages ($ USD), we can see that there is an upward trend in both GDP and Total Damages, which is concerning (Visualization #10). This indicates that over the past 20 years, the cost of damages is increasing, so even if the GDP Per Capita is increasing, the damages are still impacting the economy of a country. This may be bearable for North American countries with our increasing GDP, but it is obvious that the same safety net doesn't exist for developing countries.

Visualization #10



To show that all of these factors and findings are worth exploring more, we created a regression model to see if these factors of Temperature, Sea Level Rise, GDP Per Capita, and Total Damages were affecting the population through the Total Deaths variable (Regression Model 2). This model performed decently well with an adjusted R^2 of 0.872, however there may be some multicollinearity in the model which can be explored further in an extension of the project. Additionally, a residual plot shows that there may be some missing trends or underfitting in the model, and in an extension of the project we would run a stepwise selection algorithm to find any significant variables that may have gone undetected.

The model shows that sea levels rising, avg temperature (caused from C02 emissions), the country's gdp, and the total damages from disasters is significantly correlated with the total deaths from natural disasters. This shows that the damages and effects from climate change are negatively impacting a country's economy and population, highlighting the significance of climate change.

Regression Model 2:

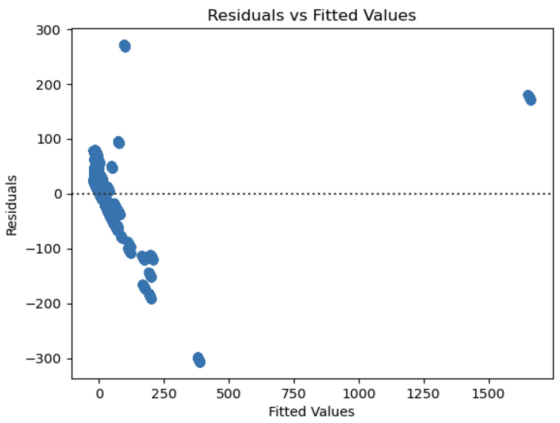
OLS Regression Results

Dep. Variable:	Total_Deaths	R-squared:	0.872
Model:	OLS	Adj. R-squared:	0.872
Method:	Least Squares	F-statistic:	6.008e+04
Date:	Wed, 27 Mar 2024	Prob (F-statistic):	0.00
Time:	04:33:17	Log-Likelihood:	-1.9595e+05
No. Observations:	35340	AIC:	3.919e+05
Df Residuals:	35335	BIC:	3.919e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	32.8497	1.479	22.214	0.000	29.951	35.748
Sea_Level_Rise	0.0548	0.368	0.149	0.882	-0.667	0.777
Avg_Temperature	0.0277	0.066	0.420	0.675	-0.102	0.157
GDP_Per_Capita	-0.0010	2.46e-05	-40.902	0.000	-0.001	-0.001
Total_Damages	1.34e-05	2.73e-08	490.100	0.000	1.33e-05	1.35e-05

Omnibus:	7288.434	Durbin-Watson:	0.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	106202.977
Skew:	-0.585	Prob(JB):	0.00
Kurtosis:	11.412	Cond. No.	5.66e+07

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.66e+07. This might indicate that there are strong multicollinearity or other numerical problems.



To conclude, we have shown that over the past 20 years, North American countries have been steadily producing C02 emissions, and the accumulation of this has been shown to correlate with increasing temperatures as shown in our line graphs. We further explored how both of these factors affect the seas and oceans surrounding North America, and that they're serially correlated with an increase in the frequency of North America's most common natural disasters - Storms and Floods. Finally, we saw that the increasing economy of North American countries has allowed them to survive the increasing expenses and damages cost from natural disasters - but this won't be the same for developing countries, which is why climate change is a serious concern and rapid measures must be taken to slow it down.

Future Project Directions

1. We would analyze data over a longer time period, ideally across 30-40 years, from approximately 1980 to 2020. This is because we found that using only a 20 year span didn't provide extremely significant results, and after researching more, we found that factors like temperature only increase about 0.06 degrees every decade - so a time span of 40 years would allow us to confirm a temperature increase as a result of C02, as well as other factors like sea levels rising. This wasn't feasible during our current project due to the short duration, because once we realized we needed older data we didn't have enough time to find it, clean, and integrate it into our database.

2. Given more time, we would supplement any missing rows in our climate metrics entity table using external data from the internet that had similar attributes. This would help our results significantly, because our other entity tables depend on the climate metrics table, so missing rows cause a cascading

effect of rows not being added into the database. This resulted in having a small dataset to make visualizations and regression models from, so by supplementing the missing rows we would have a thorough dataset and more reliable results.

3. We would showcase the lag in serial correlation between our variables by adjusting for year in our line graphs to more clearly look for association. This was currently causing some difficulty for us to implement, but would be a useful addition in the future to showcase our results more clearly to a reader.

4. We would also include data from other countries to gain a global perspective on climate change impacts, more specifically, countries that aren't as developed. This is mainly because climate change is a global issue, and now that we have an idea of how developed countries in the west are affected, we can compare the results to undeveloped countries, and build a full picture of the global changes.

5. We would take a closer look at how disasters affect economies and societies to help us understand their true cost and long-term consequences. Specifically, we would find more data about socio-economic factors like unemployment rates and housing insecurity, to see how an impact on the economy can affect other factors. We could use this detailed information to produce a case on the practical reasons why governments should care and put resources towards stopping climate change.

Long Term Data Maintenance Concerns

What concerns do you have about long term data storage?

- Since our data focuses on yearly data for each of the entities like climate metrics, natural disasters, et., data will grow substantially in volume over time. More natural disasters per year for each country will be recorded. This can pose some challenges as more data will require larger storage and more advanced systems to handle it.
- Another concern, which we have seen already, is the increase in missing values in the data across different entities.
- In our relational schema, for the Climate_Metrics table, we considered Country_Name and Year_Recorded as the primary key, thus to ensure the primary key is still valid, data must be added at yearly level for each country to maintain data integrity.
- For further analyses on long term data, it can be difficult to manage large volumes which could cause data loss, therefore, regular backups/system checks need to be implemented to prevent data from getting lost.

How will you preserve data provenance?

- Date provenance ensures the validity of data through keeping specific changes made to the data documented.
- We will preserve data provenance through maintaining a clear history of the data set origins, how it has been processed over time. For example, if multiple Climate_Metrics data was collected for Canada in different regions, we will record how we aggregated the data to meet the requirements of our relational schema.
- Additionally, changes made to the database through INSERT, UPDATE, or DELETE SQL methods will be logged, given the resources needed to implement this. One example is through a version control system showing who made each change and why.
- We can maintain a document with information on our relational schema, data definitions and additional steps to take to handle the database to help anyone using it understand its background and how the data was handled.

Advantages of a Database vs. A Series of Data Files

Using a database compared to a series of data files offers many advantages. Databases offer to handle a collection of data in a tidy and clean way and help organize large volumes of data. Given that if we wanted to expand our project and include more data long term, using a database is better. A database allows us to run complicated data tasks better, especially when we need to keep the data organized, secure and well managed.

- Our analysis requires complex queries, for example, joining multiple tables from our database to extract different columns, filtering for specific requirements, and aggregating large amounts of data. Although these may require critical thinking to write queries, once a query is made, databases can easily return accurate tuples.
- Given our current relational schema and database created, we organized the data in separable entities efficiently which made sense conceptually. For example, the Natural Disaster entity stores all the disasters that occured, the Climate Metrics entity stores the climate data from different years for the 3 countries, and the Socio-Economic Indicators entity stores all country economic related data. This structure allowed us to retrieve data from specific entities and also map the data between entities to make analyses.
- As previously mentioned, our data is focused yearly, and as it grows in volume, a database will seamlessly add new tuples into each table (given the knowledge of how to use the database) handling large datasets more efficiently compared to just files which might be hard to manage. Our data is also interconnected data, therefore, a database can keep everything in order much better than separate files.
- Lastly, databases allow multiple users to run separate queries to run analyses without causing each other issues.

If we were to use a series of data files, some issues we would encounter are:

- Trouble organizing large data files as it grows in volume.
- Longer processes as we would need to write our programs to map different data files to each other in order to draw connections. In some cases, it would even be impossible. But, this can be easily done with a database. For example, we ran a query to extract the Sea_Level_Rise, Avg_Temperature columns from the Climate_Metrics table, GDP_Per_Capita from the Socio_Economic_Indicators table, and the Total_Damages column from the Economic_Damages table.
- Higher risk of data duplication, which we were able to reduce using foreign keys.
- Difficulty maintaining data accuracy across separate files leading to inconsistencies.
- Lacks robustness, therefore, could easily crash when trying to extract specific information or large amounts of data.
- Data integrity and security being at risk as individual data files will be available, whereas, databases provide abstract views of the data, enhancing security measures.

References:

Investopedia. (n.d.). Stepwise Regression. Accessed Mar 29th, 2024, from https://www.investopedia.com/terms/s/stepwise-regression.asp#:~:text=The%20forward%20selection%20approach%20starts,importance%20relative%20to%20overall%20results.