UCD Analytics with Python Project report

GitHub URL

https://github.com/killiantos/UCDPA_KillianDavis

Abstract

Analysis of datasets demonstrating how the world has been affected more and more by climate change in the past 60 years with regards to indicators such as natural disasters, global land temperatures and CO2 emissions, with a focus on their correlation to each other

Introduction

The topic that I have decided to investigate for the purpose of this project is the indicators of global warming and its effect on climate change across the world. Climate change is a subject that I follow with great interest as it affects all of our futures and I wanted to use this project as an opportunity both to display my knowledge in Python but also to learn more about this topic for my own sake.

In this project I will analyse and try to prove a correlation between carbon dioxide emissions, rising global air temperatures and the frequency and scale of natural disasters and what these trends are predicting for the future of our planet. Human caused global warming has been proven many times over by scientists across the globe so this project will seek to confirm the claims of the scientific community through quantitative analysis of the datasets I acquired.

Climate change as an idea was first theorised in an 1896 paper by Swedish scientist Svante Arrhenius. In this paper Arrhenius claimed that changes in the carbon dioxide composition of our atmosphere could lead to a drastic change in the surface temperature of the earth due to the “Greenhouse” Effect (Arrhenius, 1896). However it was not until 1938 that the first alarm bell was rung to indicate that not only was this change possible, but that it had begun already.Guy Callendar was the name of that scientist who first warned us about climate change due to human action (Callendar, 1938). Since the 1970’s the broader scientific community has accepted climate change, not as a theory, but fact (NASA webpage on climate change evidence). Despite this there are still many politicians and their supporters who do not accept climate change as real. Through data analytics projects such as this one we can present the evidence to these policymakers in the hope of changing their mind and compelling them to action.

Dataset

The datasets that I have chosen to use for this project were found on Kaggle, a free online platform for sharing datasets around the world. However, I have ensured the provenance of the datasets are worthy of study by ensuring the sources of the datasets are reputable, from either internationally recognised non profit organisations, NASA or the United Nations. I have included links to the specific datasets in the references section at the end of this report

Dataset 1: Global Temperatures

This dataset is one that was obtained from a reputable source named Berkeley Earth, an independent, non-profit organisation focused on environmental data science and analysis based in the USA. It provides a global view of rising temperatures on earth as it only reports the global average temperature of land and sea for any given date. It contains 9 attributes (shown with their data types below);

```
dt                                          object
LandAverageTemperature                     float64
LandAverageTemperatureUncertainty          float64
LandMaxTemperature                         float64
LandMaxTemperatureUncertainty              float64
LandMinTemperature                         float64
LandMinTemperatureUncertainty              float64
LandAndOceanAverageTemperature             float64
LandAndOceanAverageTemperatureUncertainty  float64
```

Dataset 2: Natural Disasters

This dataset comes from EOSDIS (The Earth Observation System Information and Data System). This is a branch of NASA's Earth Sciences division that uses earth observing satellites to compile its data. This dataset gives information on all natural disasters that have occurred on Earth since 1900. With this data I hope to prove a correlation with rising temperatures and natural disasters. It contains 45 attributes, many of which are not useful for analysis purposes so the attributes that I will be using are shown with their data types below;

```
Year                        int64
Disaster Subgroup          object
Disaster Type              object
Country                    object
Region                     object
Continent                  object
Start Month               float64
Total Deaths              float64
Total Affected            float64
Total Damages (000 US$)   float64
```

Dataset 3: CO2 emissions by country

This final dataset is an aggregate of data collected by the United Nations Framework Convention on Climate Change (UNFCCC) and the International Energy Agency (IEA). This dataset is small with only 4 attributes;

```
country_code              object
Country                   object
year                       int64
co2 emissions(kilotons)  float64
```

.

Despite this the co2 emissions per year will provide very valuable insight when appended to another data frame as I do in this project. One limiting factor of this dataset is it only goes back to 1960 whereas the first dataset goes back to the 1700's and the second to the year 1900 so its scope is considerably smaller.

Implementation Process

1. Data Importing

In order to utilise these datasets for my analysis the first step is to call pandas .read_csv function on the csv file and assign it to a data frame (GlobalTemp, Natural_Disasters and CO2 are the three data frames respectively). I then call .head(), .shape and .describe() on each of the data frames in order to get an initial understanding of what these datasets look like and how I will begin to analyse them

2. Data Cleaning

Now that the data frames are imported into my notebook they must be cleaned before any analysis can commence. To do this I first drop any duplicate entries using .drop_duplicates() and sort the data frames by date using .sort_values as the change over time will be the key aspect of my analysis. Following on from this I got the count of null entries in each data frame in order to see how much of the dataset was missing. Two of the data frames GlobalTemp and Natural_Disasters had many missing values so I used a combination of the .fillna and .mean functions in order to assign the average value of the column to any cell that contained a null value, this will make for much cleaner analysis without skewing the results.

2.1. Merging of the Global Temp and CO2 data sets

Between these 2 datasets I have 13 columns but we do not need this many for the purposes of this analysis. I subsetted each of the datasets to only take the columns that are needed, then I extracted the year from dt in Global Temp and set it as a new column, this will be our key for merging. Then I called pd.merge on the year column for both data sets. The resulting data set was very large (over 600,000 rows) so I had to further subset this new data frame to only include data from 1960 onwards, I did this because many of the null values in the data set were before the 20th century and the co2 emissions data set begins in 1960 so it made sense to start from there.

2.2. Cleaning the data using SQL

In order to demonstrate the SQL usage learning milestone I used sqlite3 to create an empty database “co2_emissions_kt_by_country.db”. I then created a new table called CO2_emissions with the appropriate column headings and data types. I inserted the values of the 'co2_emissions_kt_by_country.csv' into this new database and performed some data cleansing by deleting entries that had a null value for Country_Code or Country and also filled any null values in the CO2_Emissions_kilotons column with the mean value of the rest of the column. I then used the .to_csv method to create a csv file containing the database

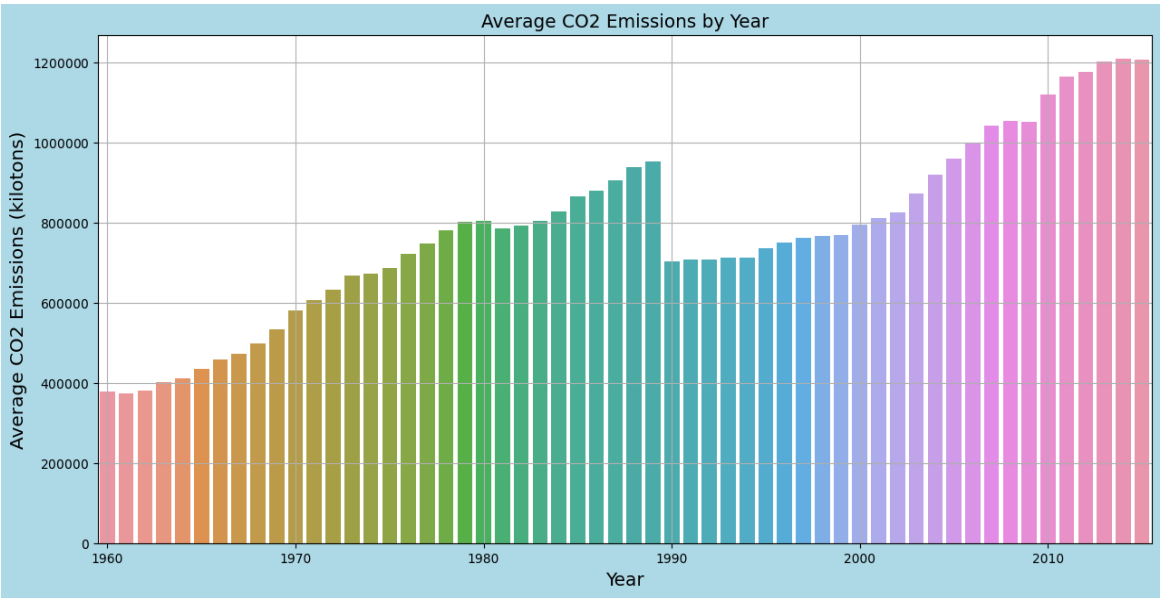
3. Analysis, Insights and Results

3.1. CO2_Temp

3.1.1. *CO2 emissions over time*

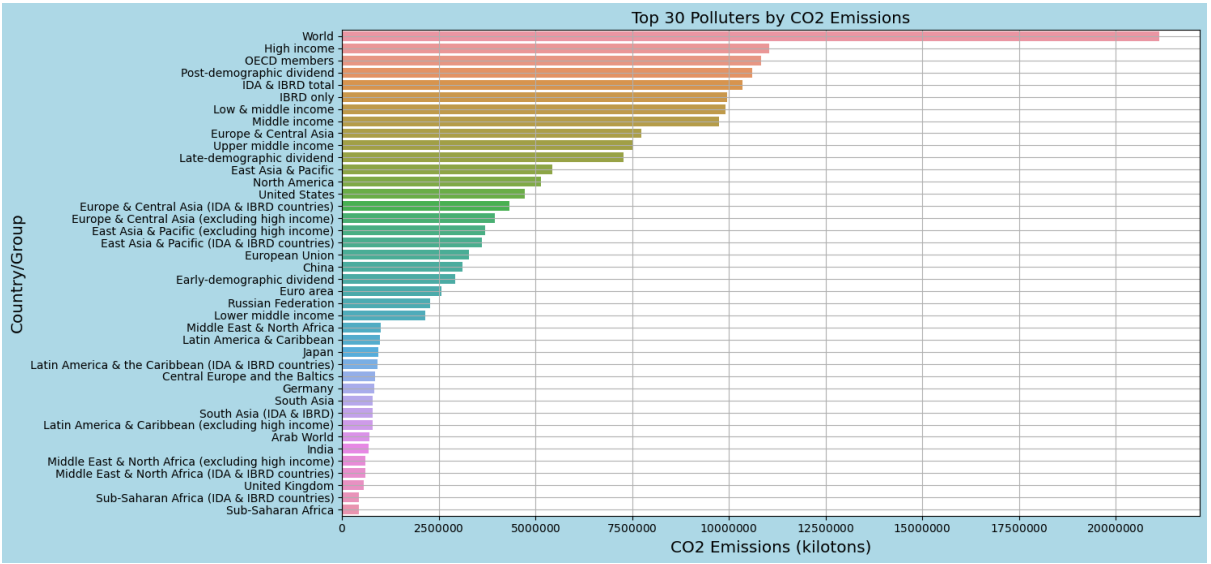
The first piece of information that I wanted to visualise was to plot the average CO2 emissions of the whole world since 1960. In order to do this I grouped all of the entries in my CO2_Temp data frame by year and got the mean of CO2 emissions. I wanted the mean and not the sum to see how the average country’s pollution over time looked. I set the xticks in increments of 10 years in order to minimise noise on the

graph, set the y axis to display in full numbers rather than scientific notation and also added colour to the background of this and every plot to help catch the eye. As can be seen in the below graph there is a clear upwards trend in average CO2 emissions year on year



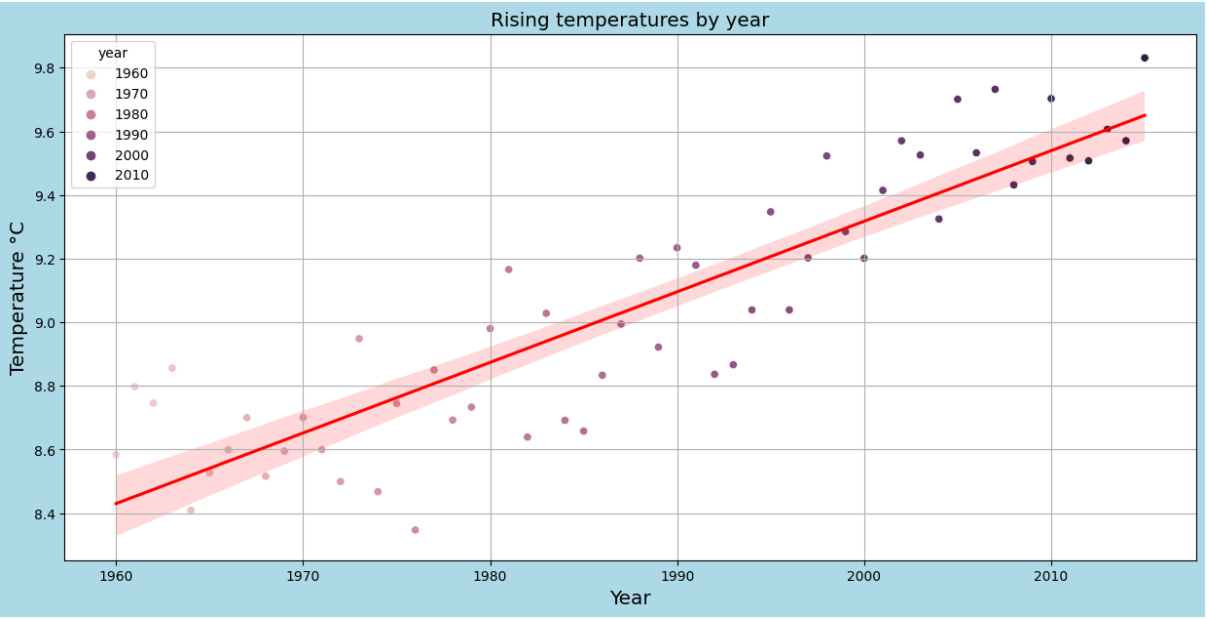
3.1.2. *CO2 emissions by top polluters*

Like the above plot this one is also a bar plot but I flipped the axes to have CO2 emissions on the x axis and top polluter groups on the y axis. This plot shows the top polluter groups and countries across the globe. High income countries are the biggest polluters by far making up more than 50% of emissions and the United States is by far the biggest single polluter with more emissions than the entire European Union



3.1.3. *Temperature increase over time*

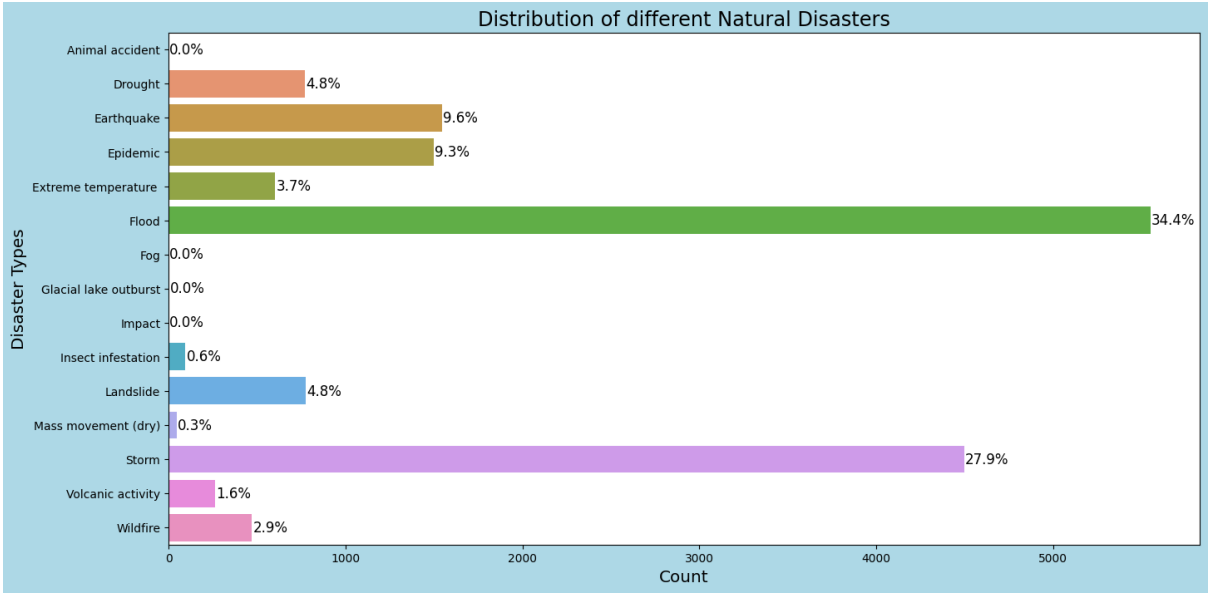
This plot is a scatter plot where each dot is the average global temperature for any given year. I used hue to show which decade each dot is from and added a line of best fit using the .regplot method in order to demonstrate the rising trend of global temperatures as can be seen below



3.2. Natural Disasters

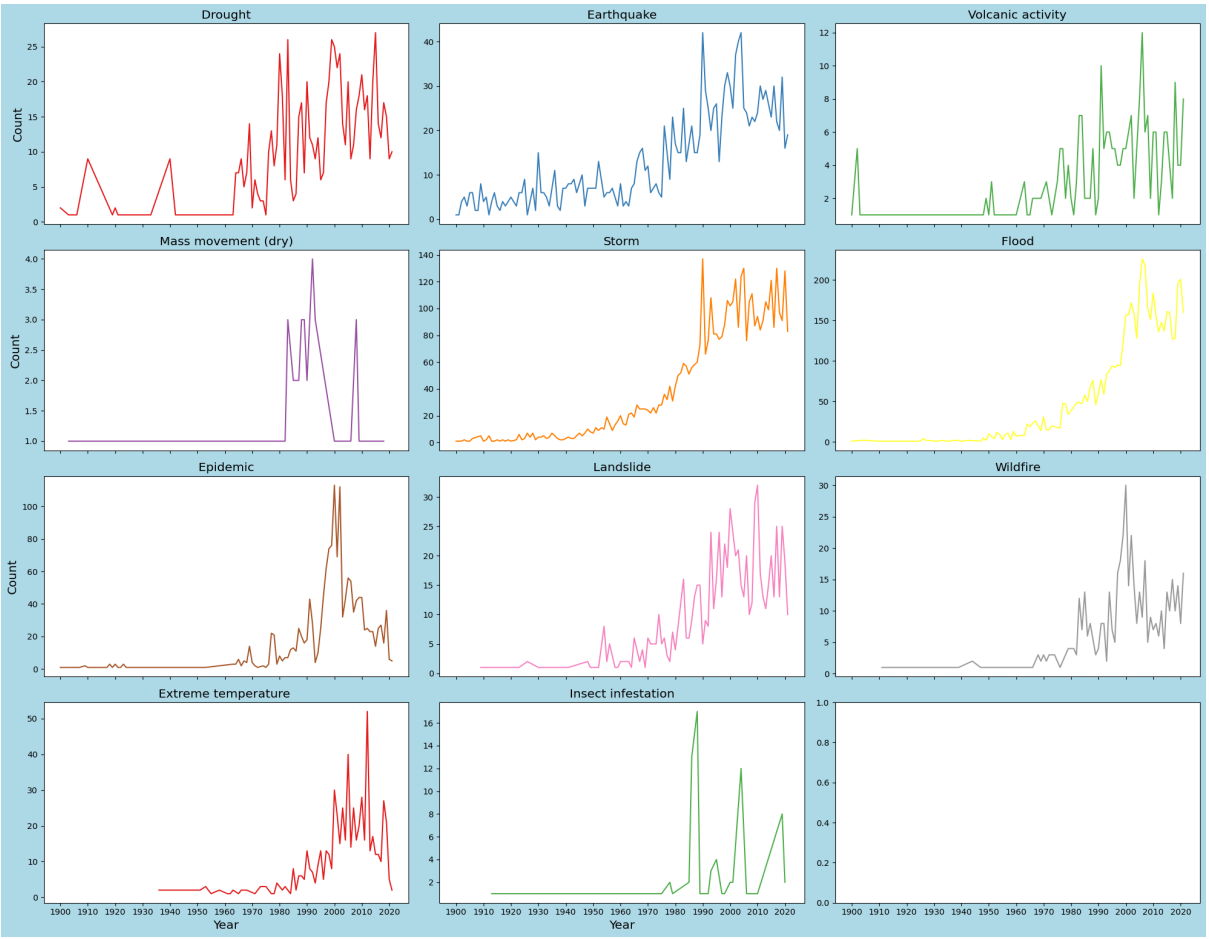
3.2.1. *Distribution of natural disasters occurrences*

The first plot using the Natural Disasters data frame shows a breakdown of each different type of natural disaster and the frequency of each one as a percentage label. I got the count of each disaster type by using .value_counts on the Disaster Type attribute. In order to generate different % values and add them as a label I created a for loop that iterates over each value count of disaster type and divides its count by the total sum and used ax.text to add it as a label. Storms and Floods make up the vast majority of all natural disasters with the two combined being more than 60% of all disasters across the world



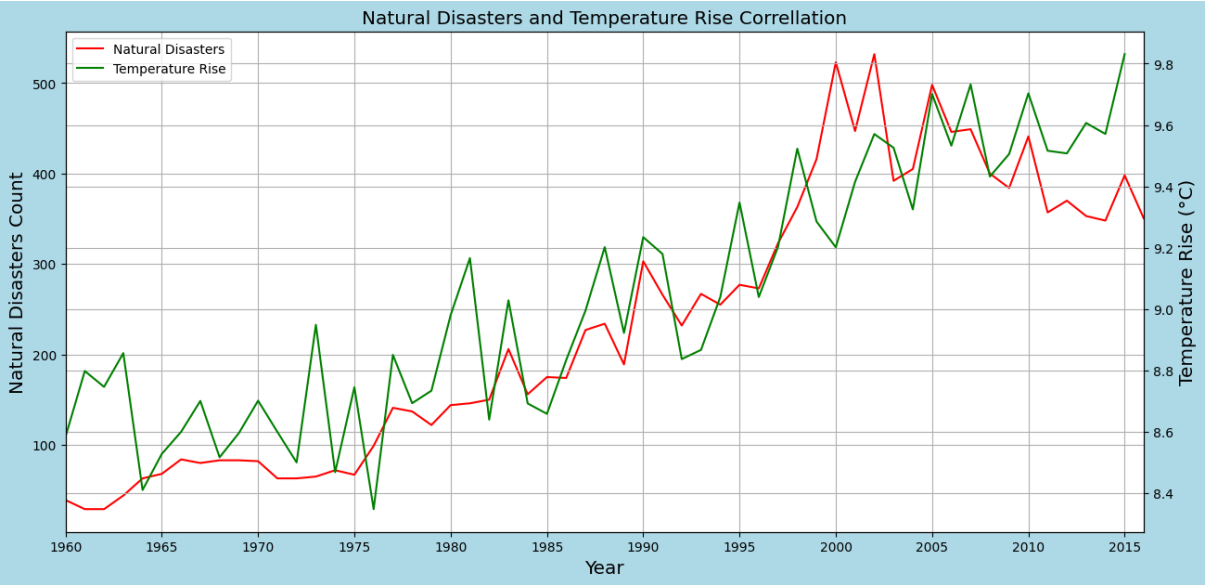
3.2.2. ***Count of different natural disasters by year***

For this visualisation I wanted to demonstrate the use of subplots by breaking each disaster type out individually and tracking how their frequency has changed over the past 60 years. In order to do this I created a custom function ‘plot_disaster_types’ that takes in as arguments two lists (unique disaster types and a list of colours). The function would then use a for loop to iterate over each element of both lists and create a subplot using each one. I also included a conditional statement to check if the number of data points in each plot was less than 2 because 4 of the 15 disaster types did not have enough data to be useful for analysis. I set the x axis labels to only appear under the bottom row of subplots in order to minimise noise on the visualisation. My reasoning for splitting out the disaster types into subplots like this was to see which of the disasters seem to be most strongly correlated with the trend in CO2 emissions and rising temperatures. More than half seem to be loosely trending upwards but what is interesting is that the two most common disasters (flooding and storms) show the clearest upward trend indicating their frequency has increased alongside global temperatures and emissions



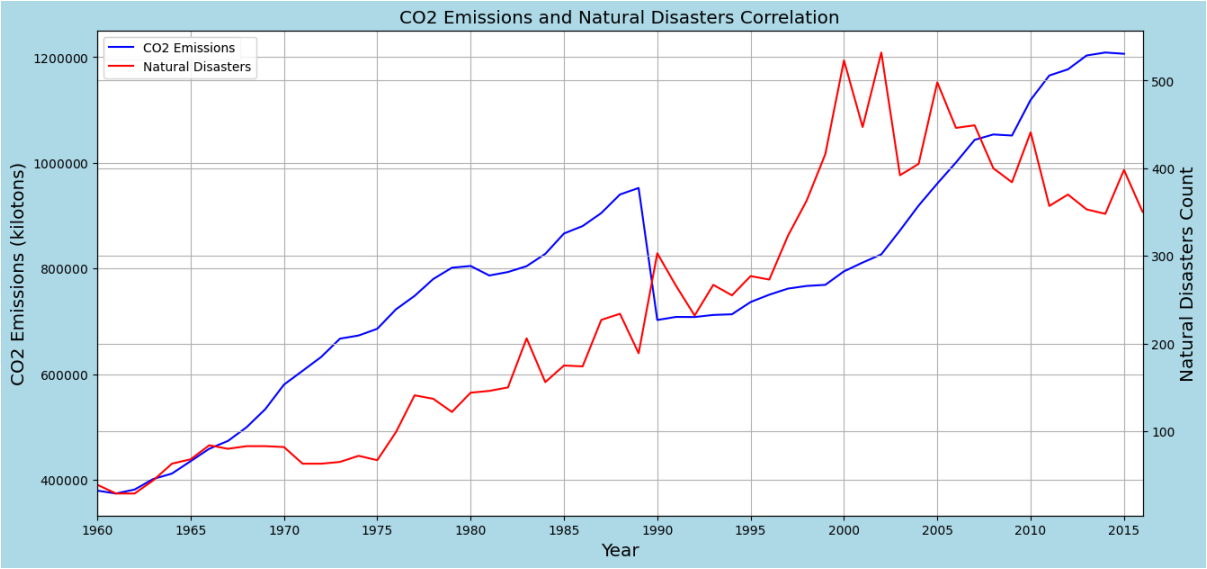
3.2.3. ***Disaster occurrences, temperature rise correlation***

For these last 2 plots I wanted to demonstrate if there was any correlation between the frequency of natural disasters, CO2 emissions and the rise in global temperatures. However, there is no functionality in matplotlib or seaborn to have three plots sharing an axis. However, there is a way to allow two plots to share an axis. To do this I had to use the .twinx function which overlays the line generated by plot 2 on top of plot 1 and allows the two to share the x axis as year. I got the value_counts of each natural disaster and used this as ax1, I then grouped the average land temperature values by years and set this to be ax2. As natural disasters data begins in 1900 and not 1960 like our subsetted CO2_Temp data frame I used .xlim to set the range and avoid large data gaps on the plot. I also used the .get_legend_handles_labels() for both axes and summed them in order to create 1 shared legend for both plots. As can be seen below there is clear correlation between the two plots



3.2.4. ***CO2 Emissions and Natural Disasters Correlation***

Similar to the above graph I wanted to plot the correlation between Natural Disaster Frequency and CO2 emissions. To do this I first had to get the CO2 emissions average grouped by year and also the value counts of natural disaster types by year. I then used the twinx function as in the above plot. As CO2 emissions initially displayed in scientific notation on the left y axis I converted this to whole numbers using the .set_major_formatter function. These two plots do not seem to be as closely correlated as natural disasters and temperature but there is still a definite loose correlation



Insights

- There is a clear rising trend across all 3 of the main variables tracked by this report; Natural Disaster Frequency, CO2 emissions and global temperatures
- High income countries, in particular the USA, contribute a vast amount to CO2 emissions to the globes total carbon footprint
- Storms and Floods are the most common natural disasters and also show the strongest parity with the trend in rising temperatures/emissions.
- There is a strong positive correlation between natural disaster frequency and rising temperatures. This indicates that as our planet gets warmer there will be more frequent and greater devastation in the form of natural disasters
- There is a positive correlation between CO2 emissions and natural disasters. From our previous insight this indicates a domino effect of CO2 emissions leading to higher temperatures which creates more natural disasters

Machine Learning

Were I to incorporate Machine Learning into this project I believe the best use for it would be to use this training data to create a regression algorithm (this data is not suited to classification) that could predict the number of natural disasters that are likely to occur in any given future time period based on new inputs of global temperature and CO2 emissions. If the model were sophisticated enough it could even possibly be used to predict a specific time period where disasters are most likely and what type of disaster should be expected next for any given region. This could be an extremely valuable tool not only in saving lives but also in alleviating some of the destruction through preparation for these disasters. It could finally be used to convince policymakers in government of the need for greater action on limiting emissions allowances by corporations and reducing our carbon footprint.

References & Bibliography

Arhenius, 1896, "*On the influence of carbonic acid in the air upon the temperature of the ground*", The London, Edinburgh and Dublin Philosophical magazine and journal of science

Callendar, 1938, "*The artificial production of carbon dioxide and its influence on temperature",* Quarterly Journal of the Royal Meteorological Society

https://climate.nasa.gov/evidence/

https://berkeleyearth.org/about/

Datasets;

https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data
https://www.kaggle.com/datasets/brsdincer/all-natural-disasters-19002021-eosdis
https://www.kaggle.com/datasets/ulrikthygepedersen/co2-emissions-by-country