Data Cleaning and Question Definition

Members: Michael Rodriguez, Matthew Wu, Colton Lapp

Overview:

We have merged together many datasets related to climate change into 3 final dataframes.

1. **Dataframe 1:** *Global x Time* Dataframe
    a. This data frame contains 115 time series variables on the global level related to climate change. We got these variables from 9 different datasets. Examples include: sea level rise, glacier size, total CO2 in the atmosphere, etc. We standardized the time units for every dataset and joined these datasets together by this 'Date' variable.
2. **Dataframe 2:** *Country x Time* Dataframe
    a. This data frame contains 128 time series variables on the country level related to climate change. We got these variables from 6 different datasets. Examples include: emissions by country, natural disasters by country, etc. We standardized the time units and the country names and joined these datasets together by Date AND country. We standardized the country names by converting them to ISO-3 codes.
3. **Dataframe 3:** *Sectoral* cross sectional Dataset
    a. We had a sectoral breakdown of "total emissions" which was not a time series variable, but instead, was simply a current cross section. We did not merge this with any dataset but instead kept it separate.

1. **Explore if your dataset/s has NAs and deal with them (or provide an explanation for keeping them and how you will deal with them moving forward)**
    a. Because different variables are available at different time frequencies and for different time horizons, our final datasets have a lot of NaN's in them where variable coverage doesn't exist. For example, if emissions data is on a monthly basis but sea level data is on a yearly basis, then sea level data will be NaN for 11 out of 12 months. Because everything is a time series, we left these NA values in place and will interpolate them linearly later if we need to for data visualization purposes. Otherwise, we don't

really see a point in trying to fill in these missing values as they aren't really "missing" but are just at a lower time frequency.

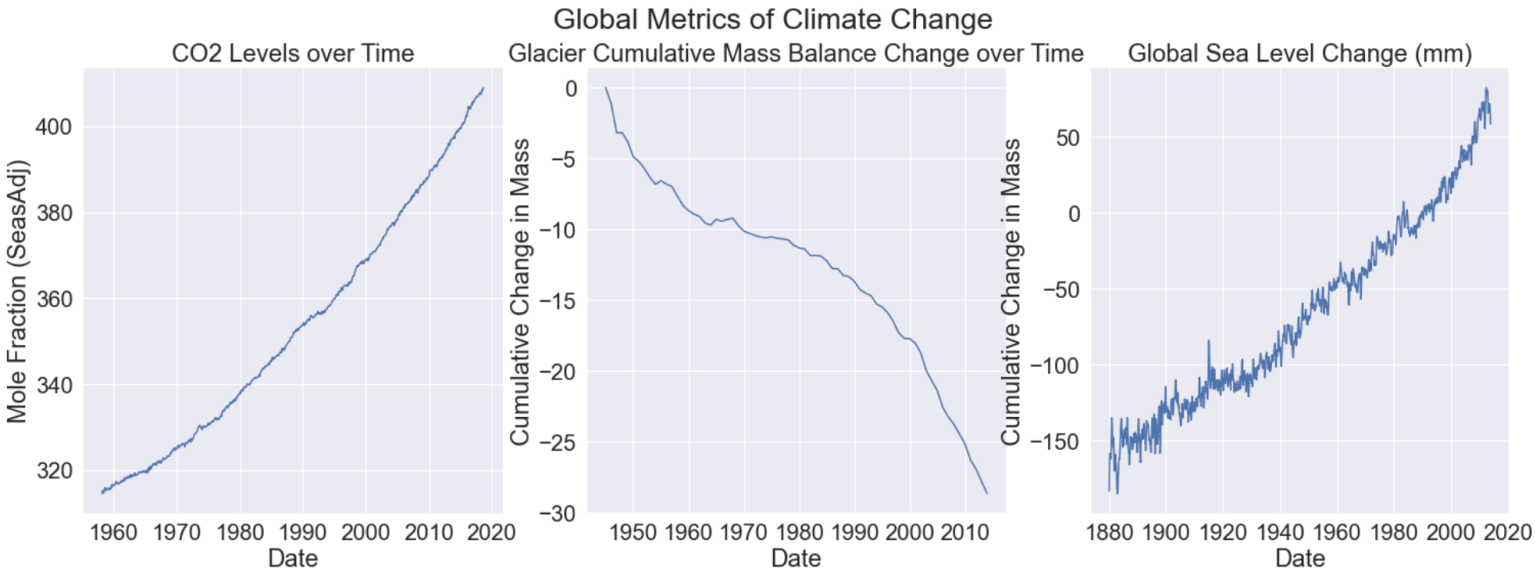
2. **Perform any necessary data conversation, such as dealing with categorical variables, and verifying each variable is the right type.**
    - **a.** We have converted all date variables into datetime objects, all country codes into categorical variables, and ensured everything else is a float or integer.

3. **If you have more than one dataset make sure to clean all of your datasets!**
    a. We have cleaned all of our datasets in a uniform fashion

4. **Decide whether you are merging your datasets or keeping them separate.**
    a. We have decided to join our datasets into three larger datasets - a global time series dataset, a country time series dataset, and a cross-sectional dataset.

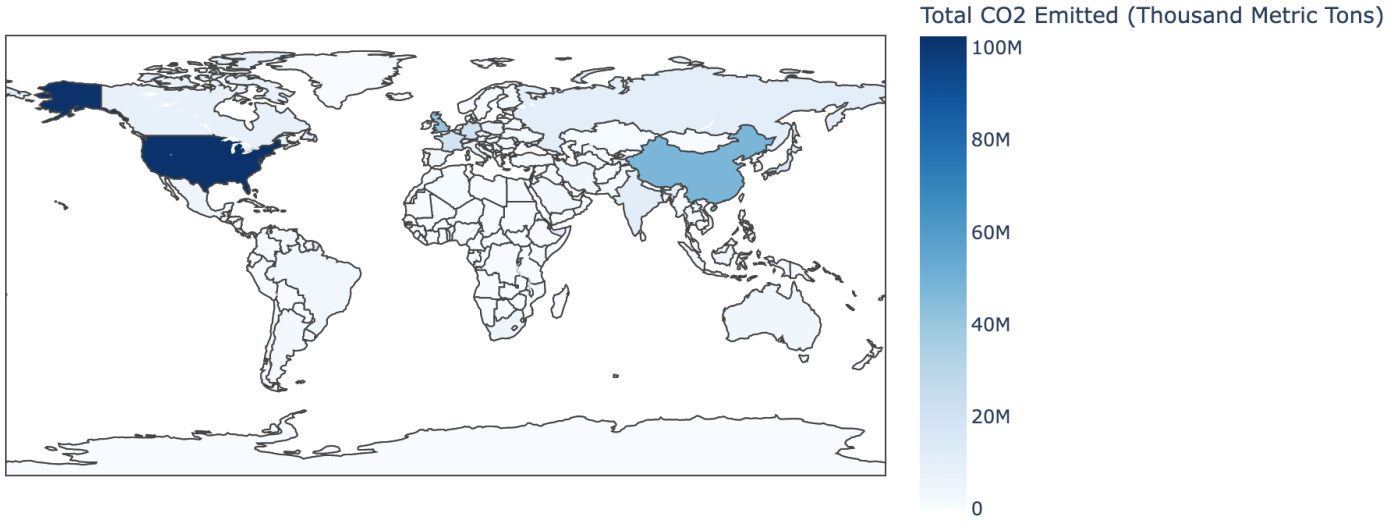
5. **If you are keeping your datasets separate provide a brief explanation of how you are going to use each dataset.**
    a. We are going to use the global datasets to show the overall trends of climate change. We are going to use the country dataset to show how impacts/causes vary by country. We are going to use the cross sectional dataset to show a current picture of emissions by sector.

6. **Provide an initial visualization of your dataset that will serve as an overview of your story.**

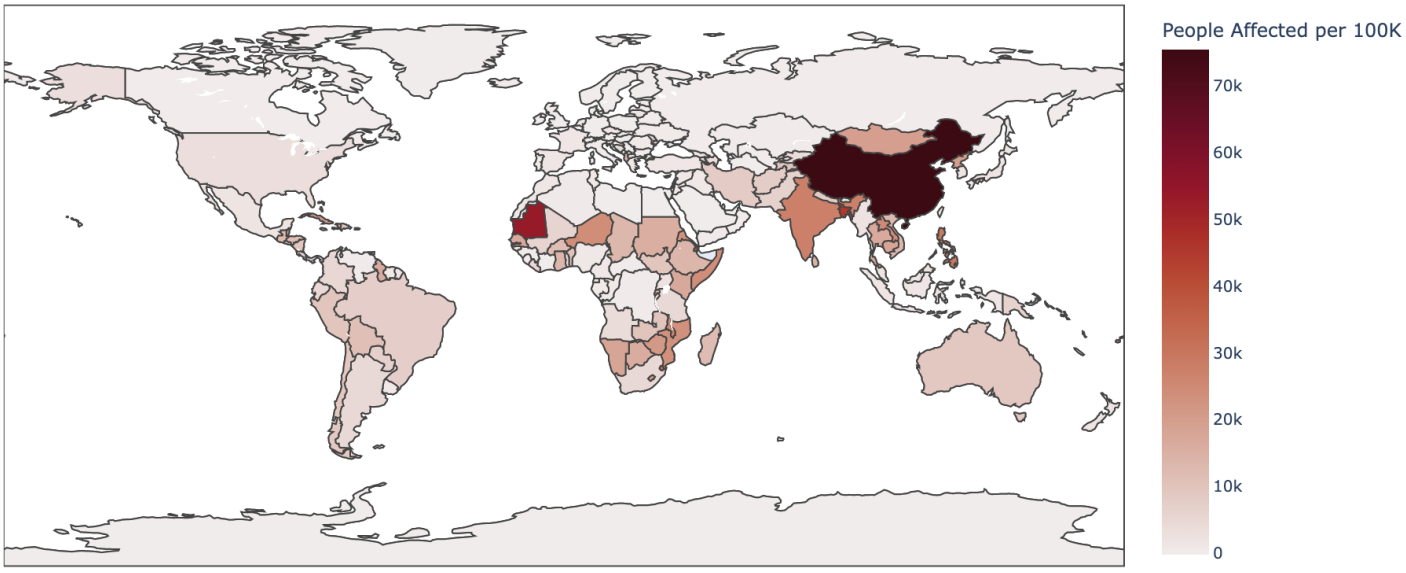


4) Provide two more initial visualizations that describe your data. These can be a correlation plot of all the variables, a boxplot or density plots of the variables, or even a plot specific to any of the questions you are trying to tackle.

Total Net Emissions Over Time For All Countries



Number of people affected by disasters per 100,000



5) Provide 1 summary statistic for each question (as a preliminary test).

Q1: How severe is climate change/what has the historical trend been?

The average yearly rise in sea level has been: **1.7mm**

Q2: Which countries are most responsible for climate change?

The United States has emitted the most emissions cumulatively of any country. A total of **24.3%** of all emissions

Q3: Which countries are currently suffering/predicted to suffer from climate change?

Mongolia has experienced the most per capita extreme temperature events at **7,447 per 100K people**

Q4: Where is the future of climate change initiatives headed?

China has the highest amount of renewable energy production for any country, at a value of **2452.5 TWh**

6) Refine the 3-4 questions you want to answer with your datasets (based on the feedback you received from your previous assignment)

Because your feedback was positive, our 4 questions remain largely the same:

1. How severe is climate change/what has the historical trend been?
    a. Specifically, we will look at trends in sea level rise, carbon dioxide concentration, glacier size, and maybe a few more variables
2. Which countries are most responsible for climate change?
    a. We will analyze historical emissions data from countries, making sure to cross reference with population and possibly GDP to add context to our data
3. Which countries are currently suffering/predicted to suffer from climate change?
    a. We will use data on people being affected by natural disasters, floods, droughts, etc to answer this question
4. Where is the future of climate change initiatives headed?

a. We have data on current usage of renewable energy by country which we plan to use to answer this question