Data Visualization : Milestone 1 - Dataset Selection and EDA

Baciu, Daniel-Mihai [1]*; Ben Ahmed, Mohamed Amine* [2]*; De Vries, Frederik Gerard* [3]

[1] *CS 369808, daniel.baciu@epfl.ch*
[1] *RO 300371, mohamed.benahmed@epfl.ch*
[2] *IN 369939, rik.devries@epfl.ch*

1 Introduction

The last five decades have witnessed a remarkable era of global transformation, marked by profound shifts in climate patterns, population dynamics, and energy consumption trends. These changes, driven by a confluence of natural processes, human activities, and socio-economic factors, have reshaped our planet's landscape and profoundly impacted ecosystems, societies, and economies worldwide. Against the backdrop of escalating concerns over climate change, resource depletion, and demographic transitions, understanding the intricate interplay between these interconnected phenomena has emerged as a paramount imperative for sustainable development and collective well-being.

In this ambitious endeavor, we embark on a comprehensive exploration of the multifaceted relationships between climate change, population migrations, and energy utilization across nations and regions worldwide. By harnessing extensive datasets spanning diverse temporal and spatial scales, we seek to unravel the underlying dynamics driving these phenomena, identify emerging trends, and illuminate pathways towards fostering resilience, equity, and environmental stewardship in an increasingly interconnected world.

2 Dataset

The datasets underpinning this project constitute a rich and diverse corpus of information, meticulously curated to provide a comprehensive understanding of the complex interdependencies shaping our global trajectory. Here, we present a detailed overview of each dataset, elucidating their sources, scope, temporal coverage, geographic extent, key variables, granularity, and data quality assurance measures:

1. **Global Daily Climate Data** [*]:
- **Source**: The climate data utilized in this study are sourced from Kaggle, specifically from the dataset titled "Global Daily Climate Data," curated by Guillem Servera.
- **Scope**: This dataset encompasses an extensive archive of daily climate records, capturing vital metrics such as temperature and precipitation, aggregated over a span of many decades.
- **Temporal Coverage**: With its longitudinal coverage spanning the past many decoades, this dataset affords a panoramic view of climate trends, variability, and extremes, enabling robust analyses of temporal patterns and anomalies. Some locations provide historical data tracing back to January 2, 1833. We have decided to take into account only the past 50 years.
- **Geographic Coverage**: Encompassing a diverse array of regions, countries, and climatic zones, the dataset provides insights into both global-scale phenomena and localized climatic nuances.
- **Variables**: Key variables of interest include daily temperature readings, precipitation levels, humidity indices, and potentially supplementary climate indicators, fostering a holistic understanding of climatic dynamics.
- **Granularity**: Data granularity is maintained at a daily resolution, facilitating fine-grained analyses and trend discernment at various spatial and temporal scales.
- **Data Quality**: Guillem Servera's curation processes likely encompass rigorous quality assurance protocols, ensuring the reliability, accuracy, and consistency of the dataset for robust scientific inquiry.

2. **World Energy Consumption and Population Data** [*]:
- **Source**: The energy consumption and population data are sourced from Kaggle, specifically from the dataset titled "World Energy Consumption," curated by Pralabh Poudel.
- **Scope**: This dataset offers a comprehensive portrayal of global energy usage patterns, demographic transitions, and societal dynamics, encapsulating the intricate interplay between human activities and environmental impacts.
- **Temporal Coverage**: Mirroring the temporal span of the climate data, this dataset spans the past century, capturing pivotal shifts in energy consumption trends and population dynamics. Also here, we have decided to take into consideration only the last 50 years.
- **Geographic Coverage**: With its expansive coverage encompassing nations and territories across the globe, the dataset facilitates cross-national comparisons, regional analyses, and insights into localized socio-economic contexts.
- **Variables**: Key variables encompass energy consumption metrics (e.g., fossil fuel consumption, renewable energy penetration rates) and demographic indicators (e.g., population size, migration flows), providing a comprehensive lens through which to assess societal transitions and environmental pressures.
- **Granularity**: The dataset likely offers data granularity at both continental, national and potentially sub-national levels, enabling nuanced analyses tailored to specific geographical contexts and administrative units.
- **Data Quality**: Pralabh Poudel's data curation efforts are presumed to uphold stringent standards of accuracy, completeness, and reliability, ensuring the robustness and credibility of analyses conducted using this dataset.

In synthesis, the convergence of these meticulously curated datasets engenders an unparalleled opportunity to delve into the intricate dynamics underpinning global environmental change,

[*] *https://www.kaggle.com/datasets/guillemservera/global-daily-climate-data*

[*] *https://www.kaggle.com/datasets/pralabhpoudel/world-energy-consumption*

demographic shifts, and energy transitions. By harnessing the wealth of information encapsulated within these repositories, we endeavor to unravel the complexities of our evolving planet, glean insights vital for evidence-based policymaking, and foster a collective ethos of sustainability, resilience, and stewardship towards safeguarding the future of generations to come. Through rigorous analysis, interdisciplinary collaboration, and visionary foresight, we aspire to chart a course towards a more equitable, harmonious, and sustainable coexistence with our planetary home.

3 Problematic

As climate change is such an important topic, it has been well-researched and there are vast amounts of data available. However, all this data can make it difficult to get a clear picture of what is happening. We would therefore like to visualize it for the general public and scientists alike. Our main goals are to show:

- what the causes of climate change are
- what the effects of the process over the years are
- what countries are doing in order to combat the problem.

We are not planning to make any future predictions and instead will show data up until now. Our visualizations aim to enable our users to draw their own conclusions on whether or not we are doing enough as a society to combat this global problem.

4 Exploratory Data Analysis

In this report, we analyze a comprehensive dataset containing daily climate data spanning several decades, alongside a secondary dataset detailing world energy consumption, population, and GDP by country. Our objective is to explore the relationship between climate change and energy consumption, considering population growth and economic development as additional factors.

4.1 Weather Reports Analysis

The initial examination of our climate dataset reveals a rich historical range, primarily concentrated from the 1950s through to 2023, as shown in 1. This extensive temporal coverage provides a valuable foundation for analyzing long-term climate trends.

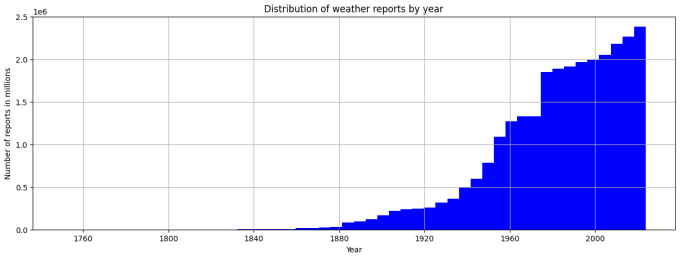

Fig. 1: Distribution of weather reports by year

Given the importance of geographical diversity in our study, we assessed the number of reporting stations by year. Setting a threshold of 1000 stations to ensure broad geographical representation, we determined that our analysis should commence from the year 1973. This decision is supported by 2, marking the beginning of our 50-year range of climate data.

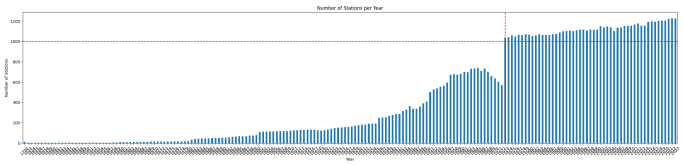

Fig. 2: Number of Stations per Year

To validate the uniformity of data reporting across stations, we analyzed the distribution of weather reports per station, finding a satisfactory consistency as illustrated in 3. This uniformity is crucial for the reliability of our subsequent analyses.

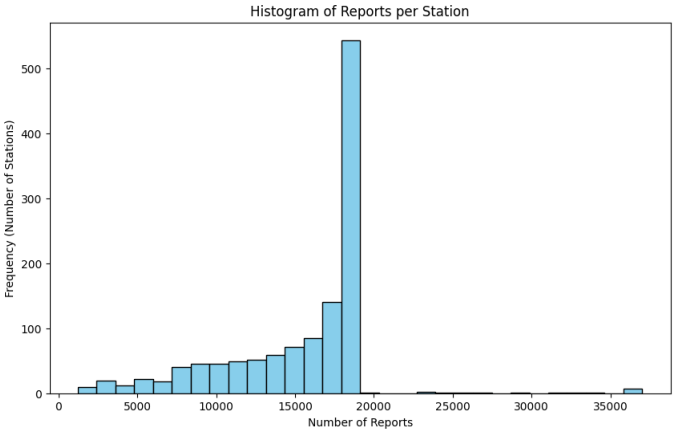

Fig. 3: Histogram of Reports per Station

4.1.1 Missing Data Assessment

A key step in our data preparation involved assessing missing data by year for various climate features. 4 indicates that essential variables, such as average temperature and precipitation, are predominantly well-reported, which fortifies the core of our climate analysis.

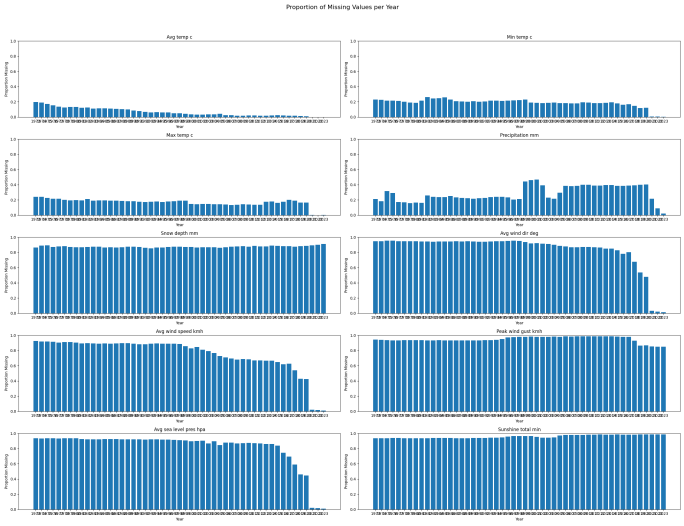

Fig. 4: Proportion of Missing Values per Year

However, when investigating the geographical distribution of missing data for additional variables like snow depth, wind speed, sea level pressure, and sunshine duration, we identified a significant

limitation. As depicted in 5, comprehensive data coverage is primarily limited to the USA, prompting us to focus a detailed case study on this region later. For broader analyses, we decided to exclude the aforementioned variables due to insufficient global data coverage.

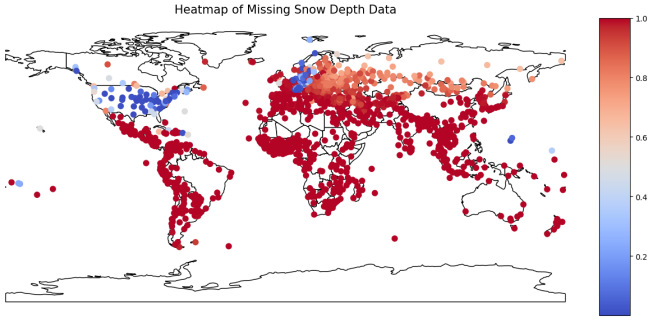

Fig. 5: Heatmap of Missing Snow Depth Data

4.2 Energy Consumption and Climate Change

Integrating a dataset on world energy consumption, we aim to explore its correlation with climate change, factoring in yearly population and GDP as potential influencers. Given the prevalence of missing values in energy consumption data, we aligned this dataset's temporal range with our climate data, starting from 1973.

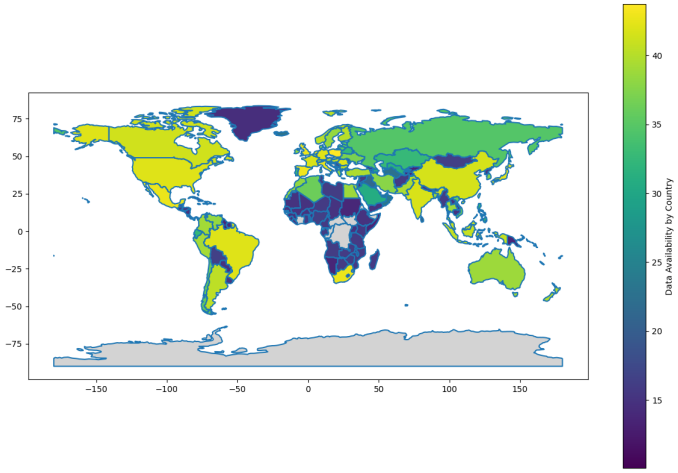

Fig. 6: Data Availability by Country

A geographical overview of available energy consumption data (6) reveals adequate coverage across numerous countries, validating the dataset's utility for our analysis despite some gaps. Specifically, for the critical last two decades, which have experienced accelerated industrial growth, our data completeness assessment (7) shows that over 70% of the data is available, ensuring a robust foundation for examining the recent dynamics between energy consumption and climate change.

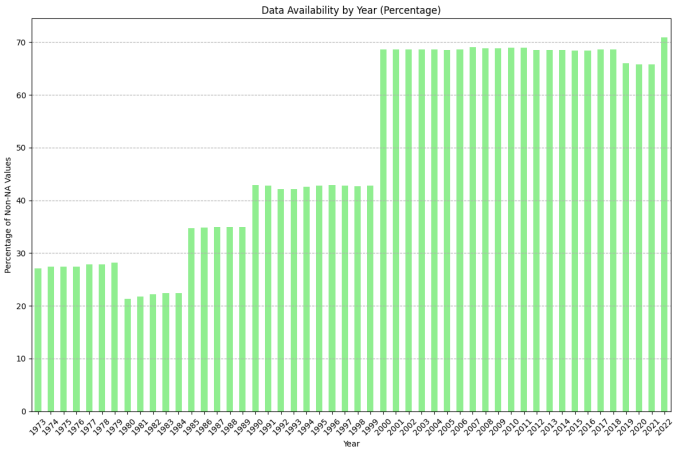

Fig. 7: Data Availability by Year (Percentage)

5 Related work

Climate change and weather conditions are a very widely researched topic as it is one of the largest issues of our time. There have also been numerous efforts to visualize the process that is going on. NASA, for example, has a lot of excellent visualizations such as their Climate Spiral video * and the Climate Time Machine †. The Climate Time Machine is an interactive website that lets users visualize data regarding sea levels, global temperatures, and so on over time. The website works by letting the user navigate through a set of images. If the user wants to see the development over time, they can choose to play the images one after the other to create something that resembles a video.

There exist also many visualizations that are mainly US-focused. The Climate Explorer ‡ and the Impact Lab Map § show regional temperatures in the United States along with predictions for the future.

If we also consider more static, less interactive visualizations of climate data, the International Peace Institute Global Observatory published a good article depicting a variety of plots and maps regarding climate change ¶.

* *https://science.nasa.gov/resource/video-climate-spiral-1880-2022/*
† *https://climate.nasa.gov/interactives/climate-time-machine/*
‡ *https://crt-climate-explorer.nemac.org/*
§ *https://impactlab.org/map/*
¶ *https://theglobalobservatory.org/2023/12/2023-a-year-in-climate-data-visualizations/*