Visualisation Semester Project: GHG Emissions and Natural Disasters

Ashutosh Maurya

M.Sc. Data Science, 2021-23, Chennai Mathematical Institute

Abstract

This is the final report of the semester project of the Visualisation course taken during September-December, 2021 at M.Sc. Data Science, Chennai Mathematical Institute.

The purpose of the project is to study and analyse the greenhouse emissions of countries, and compare it with the occurrences of natural disasters around the globe. This can shed some light on the effect of harmful emissions on the planet's ecosystems and their stability, which in turn leads to all kinds of natural disasters, from droughts to floods.

The visualisation is done in `R` using the `ggplot` package. An RShiny dashboard is also made using the `shiny` package, and the application is deployed as a web app hosted on `shinyapps.io`. Additional visualisation is also done using Tableau.

1 Dataset

The dataset used is the **Global Environment Indicators** dataset, supplied by the United Nations Statistics Division. The dateset has data on several environment indicators which help us to understand and analyze the health of the planet, and take actions based on these statistics. The indicators include Energy and Minerals, Forests, Waste, and many more. For this project, the data on Greenhouse Emissions from Air and Climate indicators, and data on Occurrences of Natural Disasters was taken.

1.1 Air and Climate

From Air and Climate indicators, `GHG_Emissions.csv` and `GHG_Emissions_by_Sector.csv` were taken, from which the variables `Country ID`, `Country`, `GHG total without LULUCF, latest year`, `% change since 1990` and `GHG emissions per capita, latest year`, `GHG from energy, as percentage to total`, `GHG from industrial processes and product use, as percentage to total`, `GHG from agriculture, as percentage to total`, `GHG from waste, as percentage to total` were taken. Description of variables are given in the table.

1.2 Natural Disasters

From Natural Disasters indicator, `Climatological disasters.csv`, `Geophysical disasters.csv`, `Hydrological disasters.csv` and `Meteorological disasters.csv` are taken, from which the variables `CountryID`, `Countries or areas`, `Occurrence 1990-1999`, `Occurrence 1999-2009` and `Occurrence 2009-2019` are taken from each file. Description of variables are given in the table.

From the documentation, Natural Disasters are classified as:

- Climatological disasters: Hazards caused by long-lived, meso- to macro-scale atmospheric processes ranging from intra-seasonal to multi-decadal climate variability. They are sub-classified as Drought, Glacial Lake Outburst and Wildfires. Drought is an extended period of unusually low precipitation that produces a shortage of water for people, animals and plants. Glacial lake outburst is a flood that occurs when water dammed by a glacier is suddenly released.

- Geophysical disasters: Hazards originating from solid earth. They are classified as: Earthquakes, Mass Movements and Volcanic Activities.
  Earthquake is sudden movement of a block of the Earth's crust along a geological fault and associated ground shaking. Mass movement is any type of downslope movement of earth materials. Volcanic activity is a type of volcanic event near an opening/vent in the Earth's surface including volcanic eruptions of lava, ash, hot vapour, gas, and pyroclastic material.

- Hydrological disasters: Hazards caused by the occurrence, movement, and distribution of surface and subsurface freshwater and saltwater.They are further classified as: Flood, Landslide and Wave Action.
  Flood is a general term for the overflow of water from a stream channel onto normally dry land in the floodplain, higher-than-normal levels along the coast and in lakes or reservoirs, as well as ponding of water at or near the point where the rain fell. Landslide is the movement of soil or rock controlled by gravity and the speed of the movement usually ranges between slow and rapid, but not very slow.

- Meteorological disasters: Hazards caused by short-lived, micro- to meso-scale extreme weather and atmospheric conditions that last from minutes to days. They include extreme temperature, fog, and storm.

2 Data Loading and Cleaning

The required variables were extracted from the respective tables. In the disasters data, there are missing values present as ". . . " in the data, and since the variable is

Variable Name	Type of Variable	Description
`Country ID`	Nominal categorical	Unique number to identify country
Country	Nominal categorical	Name of country
`GHG total without LULUCF, latest year`	Continuous numerical	Total emissions of greenhouse gases in latest year (whose data is available), excluding emissions from LULUCEF (commercial uses, land-use change, and forestry) activities (in 1000 tonnes of CO_2 equivalent)
`% change since 1990`	Continuous numerical	Percentage change in total emissions from 1990 to latest year
`GHG emissions per capita, latest year`	Continuous numerical	Per capita emissions of greenhouse gases (in 1000 tonnes of CO_2 equivalent)

Table 1: Variables from `GHG_Emissions.csv`

Variable Name	Type of Variable	Description
`Country ID`	Nominal categorical	Unique number to identify country
`Country`	Nominal categorical	Name of country
`GHG from energy, as percentage to total`	Continuous numerical	Percentage of GHG gases emissions in energy production
`GHG from industrial processes and product use, as percentage to total`	Continuous numerical	Percentage of GHG gases emissions in industrial processes and product use
`GHG from agriculture, as percentage to total`	Continuous numerical	Percentage of GHG gases emissions in agriculture
`GHG from waste, as percentage to total`	Continuous numerical	Percentage of GHG gases emissions in waste disposal and management

Table 2: Variables from `GHG_Emissions_by_Sector.csv`

Variable Name	Type of Variable	Description
`CountryID`	Nominal categorical	Unique number to identify country
`Countries or areas`	Nominal categorical	Name of country
`Occurrence 1990-1999`	Discrete numerical	Number of occurrences of disaster in the decade 1990-1999
`Occurrence 1999-2009`	Discrete numerical	Number of occurrences of disaster in the decade 2000-2009
`Occurrence 2009-2019`	Discrete numerical	Number of occurrences of disaster in the decade 2010-2019

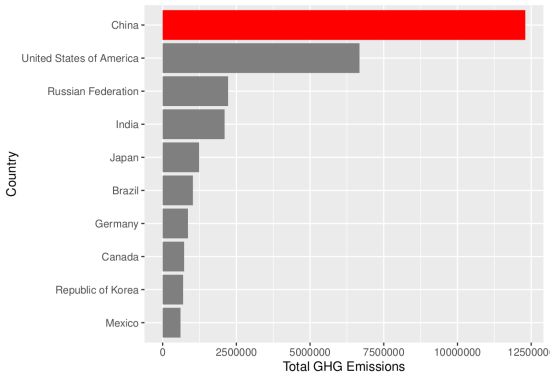
Table 3: Variables from each of `Climatological disasters.csv`, `Hydrological disasters.csv`, `Meteorological disasters.csv` and `Geophysical disasters.csv`

the number of occurrences of disasters, the missing values can be handled by assuming them as zero. For the greenhouse gas data, the data with sector wise emissions was merged with the original data using Country ID as the key to facilitate visualisation. Numerical variables were stored as characters in the dataset, so they were converted to integers or decimals in the cleaning process.

3 Visualisations and Analysis

3.1 Univariate Analysis

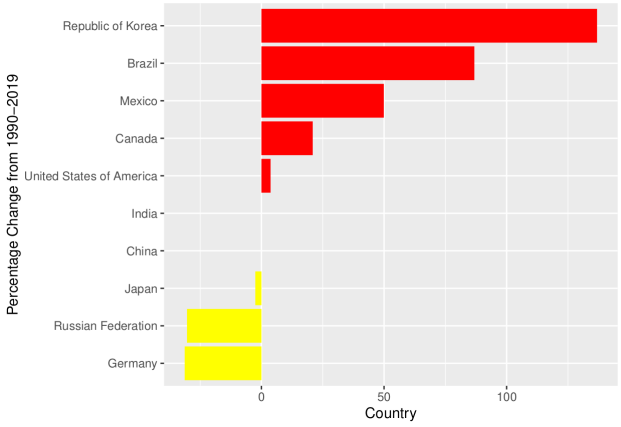
Each variable was visualised and analysed individually to understand the structure and behaviour of the data. Bar plots and box plots were used for these visualisations, and careful use of colours was done to highlight important features of the variables rather than for mere beautification. Since there are more than 190 countries and areas, most of the visualisations were limited to the top 10 contributors with respect to individual variables. Bar plots were used to compare that subset, while box plots provide insight into the entire distribution.



From the visualisations, we can make the following conclusions:

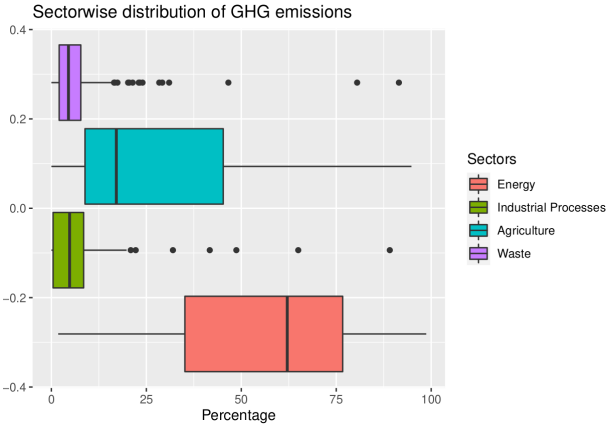
- China's emissions far surpasses emissions from any other country. Highly developed nations and fast growing nations are in the top 10 contributors.

- Developing nations seem to have a positive percentage changes, inferring that their emissions have greatly increased over the decades. Some developed countries have managed to get a negative percentage changes.



- Per capita emissions seems to have a better distribution, indicating that total emissions may be a biased variable because the difference in population of countries is very large. It is apparent that highly developed nations have the highest per capita emissions, while developing countries have lower per capita emissions.

- It is to be noted that on an average, more than half of the emissions are due to energy generation processes. Highly developed countries and fast growing developing countries have more the 75% of their emissions due to energy generation processes.

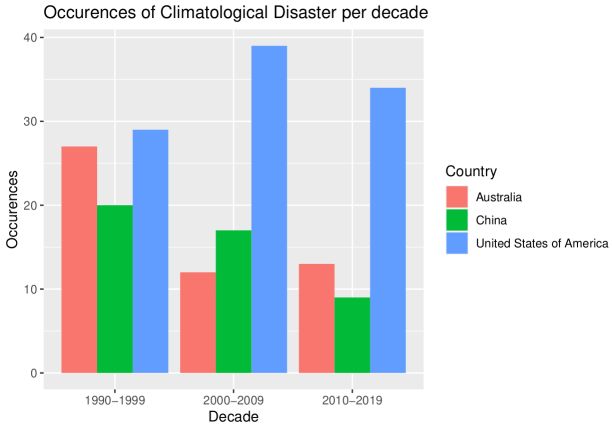


3.2 Multivariate Analysis

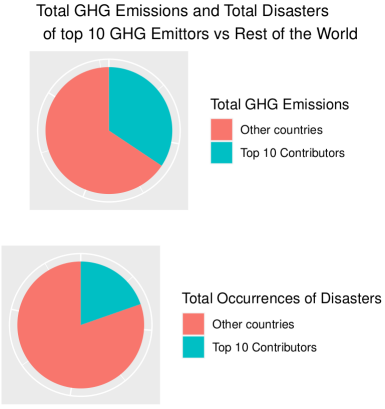
Multivariate analysis must be carried out to understand the relationships, or independence, or different variables. Towards the purpose of trying to analyse the effect of emissions on the number of occurrences of natural disasters, bar plots and pie plots were used.

From the visualisations, we can make the following conclusions:

- Energy generation is the biggest contributor across all countries, Agriculture seems to be the second, and Industrial and Waste follows next.



- Top 10 GHG emitters contribute to around one-third of the total GHG emissions, but only suffer around one-sixth of the total disasters.



4 Summary

From the visualisation of Greenhouse Gases Emissions, it can be seen clearly that the variables are heavily skewed and there are too many outliers. The top contributors for greenhouse emissions seem to be the developed nations and those developing nations whose growth is very high, like China, India and Brazil.

From the visualisation of Natural Disasters, it can be seen that different countries suffered from different disasters, but some countries are frequently hit by every kind of disasters, while some don't ever get struck by disasters. This may be due to a bias in data towards accurate reporting of developed nations compared to possible misrepresentation of poorly developed nations.

From the entire visualisation process, the most important findings are noted below.

- Top 10 contributors of greenhouse gases emissions emit around half the the amount emitted by the rest of 180 countries, but only suffer from less than one-fifth of the total disasters.

- Even though fast growing developing countries have high emissions, developed nations have higher per capita emissions. But developing nations are lowering their emissions, while developing nations have high positive percentage change.

- It is clear that energy generation creates the most emissions across all countries.
- There may be a bias in the disasters data in the form of misrepresentation of poorly developed countries, due to which the grave reality of their emaciation is not apparent.

From these findings, a few steps can be suggested for the improvement of the health of our environment.

- Developing countries have to be assisted to adopt improved technologies for energy generation, since those countries will continue to contribute heavily to GHG emissions in this aspect.
- Developing and under developed countries have to be better prepared for disasters, as due to poor infrastructure and laggy preventive measures, they will suffer more casualties even if number of occurrences is low compared to developed countries.
- Collection of accurate data from every single country has to be prioritised so that decisions can be made to assist under represented countries and regions.

Important Links

- Kaggle Dataset: `https://www.kaggle.com/ruchi798/global-environmental-indicators`
- Complete Visualisation and Analysis: `https://drive.google.com/drive/folders/1AHDyzFIDQFwKJSReJySV0HnUv06ZvAS2?usp=sharing`
- RShiny Dashboard: `https://ashutoshm.shinyapps.io/SemesterProject/`
- YouTube Presenation: `https://youtu.be/bcicSr62wDA`