Gapminder Data Analysis

Introduction

Gapminder has collected a lot of information about how people live their lives in different countries, tracked across the years, and on a number of different indicators In this project, I will choose 5 indicators to investigate

- Agricultural land coverage = Refers to the share of land area that is arable, under permanent crops, and under permanent pastures (scala from 0-1, should be multiplied by 100% to convert to percentage)
- Flood affected = Total number of people getting affected, injured, or killed in flood during given year (number of people)
- Forest coverage = Percentage of total land area that has been covered with forest during the given year, excluding other wood land which is spanning more than 0.5 hectares, with trees higher than 5 meters and a canopy cover of 5-10 percent (scala: from 0-1, should be multiplied by 100% to convert to percentage)
- Energy production = Refers to forms of primary energy (number of products)
- Total income = Gross domestic product per person, inflation adjusted. (in PPP)

Research Questions

Using the data, there are several questions to be explored and answered

- Which regions have the lowest and highest performance of each indicator?
- Have certain regions of the world been growing in each of the indicators?
- Is there any correlation between selected indicators?

Data Wrangling Process

- Load in the data, check for cleanliness, and then trim and clean the dataset for analysis
- Since all datasets are separated we need to merge into one dataset
- Before merging the data, we also need to change the structure of the dataset
- Previously, the structure is like this

	country	1961	1962	1963	1964	1965	1966
0	Afghanistan	0.577	0.578	0.579	0.580	0.580	0.581
1	Albania	0.450	0.450	0.450	0.449	0.451	0.453
2	Algeria	0.191	0.189	0.187	0.185	0.185	0.185
3	Andorra	0.553	0.553	0.553	0.553	0.553	0.553
4	Angola	0.459	0.459	0.459	0.459	0.459	0.460

- After the data wrangling process, the structure is like this

	country	year	agr_value	fld_value
0	Albania	1990	0.409	0
1	Albania	1991	0.411	0
2	Albania	1992	0.411	35000
3	Albania	1993	0.411	0
4	Albania	1994	0.411	0

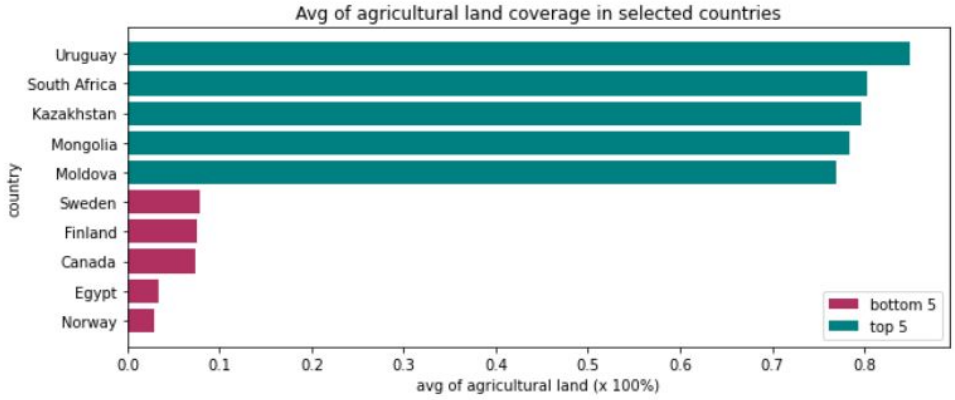
- So I transpose the year into values instead of separate column, and put each indicator value in different columns
- After that I clean the dataset by using inner join on country and year
- The reason why I use inner join because if there is a country that does not exist in one of the indicator, it will cause many null values and will be hard to predict what the values are
- I also change the data type and fill the the missing value with mean
- Some quick description on the data after being cleaned
    - There are unique 12 countries and unique 19 years
    - On average, 43% of land is covered by plants that can be cultivated
    - On average, around 1mio of people are affected by floods. a country in a certain year also lost around 2.4mio of people. The data will be probably skewed
    - 30% of land is covered by forest, on average. the characteristic seems almost the same as agricultural land data
    - Countries in certain years have produced around 80K of energy on average. but at maximum there is data that shows up to 1.9mio of energy production. this is probably an outlier and the data seems skewed
    - On average, GDP per capita is around 14K dollars
- The details can be found in the python script

Exploratory Data Analysis

Q1: Which regions have the lowest and highest performance of each indicator?

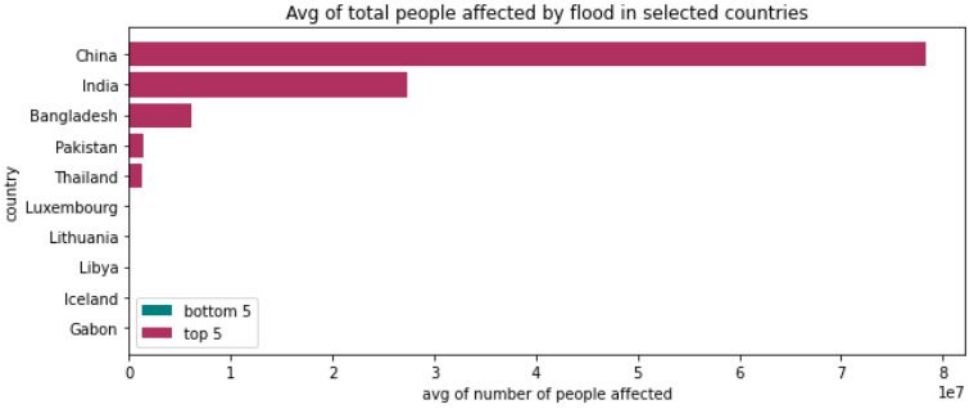
To see the lowest and highest performance of each indicator, I used mean to see one number that describes the 'performance'

Agricultural land coverage



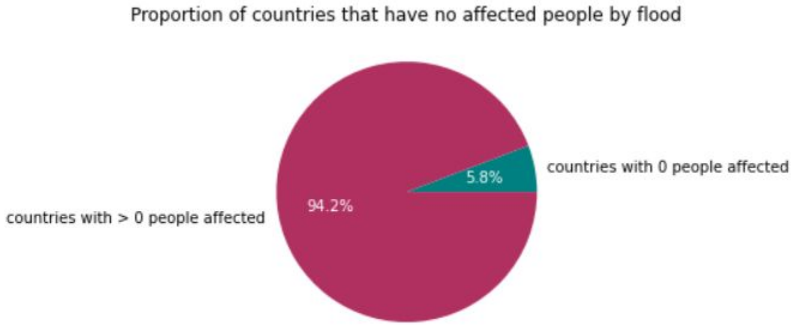
- There are 5 top countries that have the largest agricultural land. Those are **Uruguay, South Africa, Kazakhstan, Mongolia, and Moldova**
- Actually **Kazakhstan and Mongolia are quite close geographically**
- Then, there are 5 countries that have the smallest agricultural land. Those are **Sweden, Finland, Canada, Egypt, and Norway**
- Three of the five countries are **Scandinavian countries**

People affected by flood



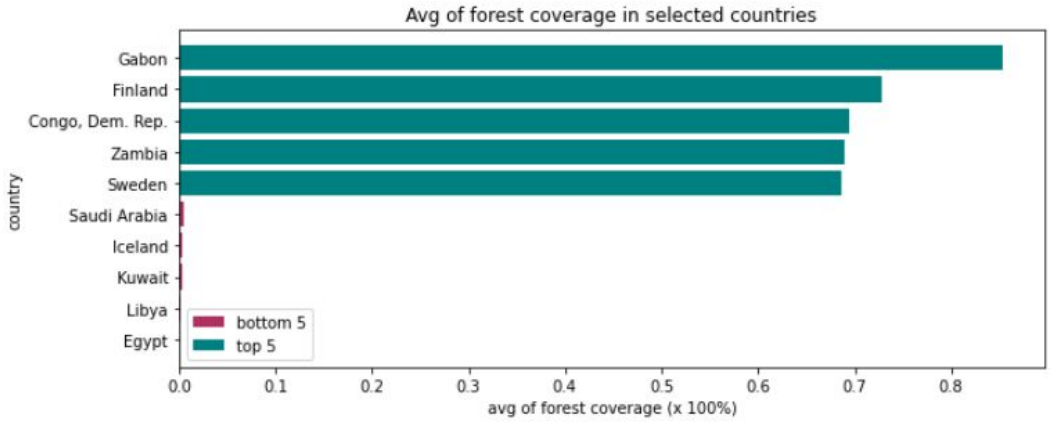
- There are 5 top countries that have the most affected people by flood. Those are **China, India, Bangladesh, Pakistan, and Thailand**

- The difference of number of people affected by flood in China is quite huge compared to India as the second rank
- Then, India and Bangladesh which are in the next ranks, are also close geographically
- 5 countries that have the smallest number of people affected by flood are **Luxemburg, Lithuania, Libya, Iceland, and Gabon**



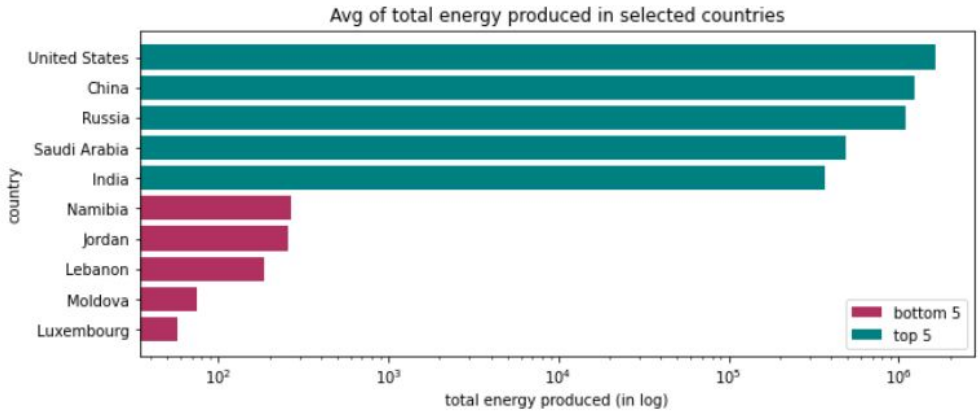
- Other than that, there are **5.8%** of countries that never have victim by flood

Forest coverage



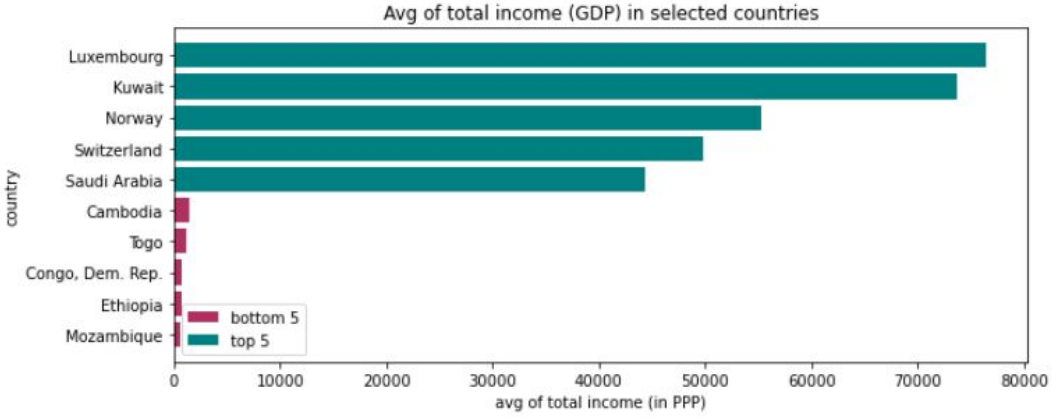
- **Gabon, Finland, Congo, Zambia, and Sweden** have the largest land for forest
- **Saudi Arabia, Iceland, Kuwait, Libya, and Egypt** have the smallest land for forest

Total energy produced



- **US, China, Russia, Saudi Arabia, and India** produced energy more than other countries
- Those top 5 countries are definitely one of the most powerful countries
- Meanwhile, **Namibia, Jordan, Lebanon, Moldova, and Luxembourg** are countries that produced the least energy

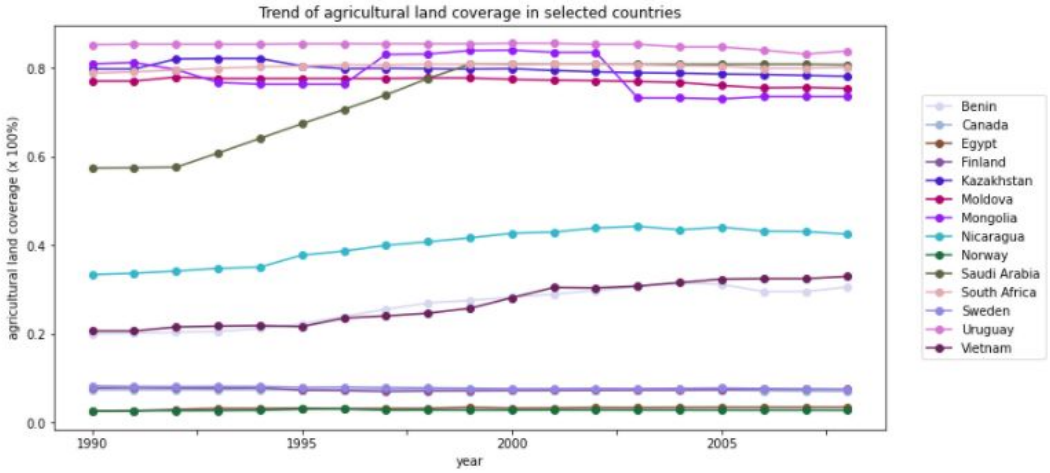
Total income



- **Luxembourg, Kuwait, Norway, Switzerland, and Saudi Arabia** have the highest total income
- **Cambodia, Togo, Congo, Ethiopia, and Mozambique** have the lowest total income
- Most of the lowest income countries are in **Africa** region which we also know that in general, Africa has the lowest income compared to other region in the world

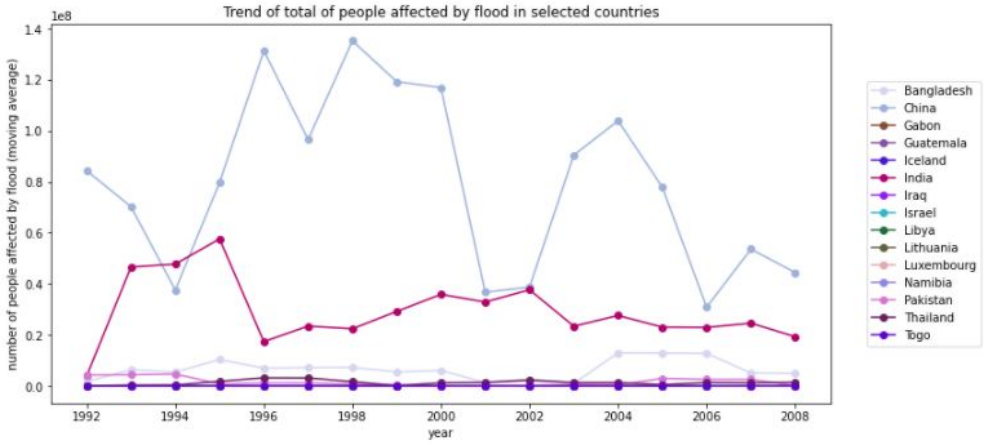
Q2: Have certain regions of the world been growing in each of the indicators?

Agricultural land coverage



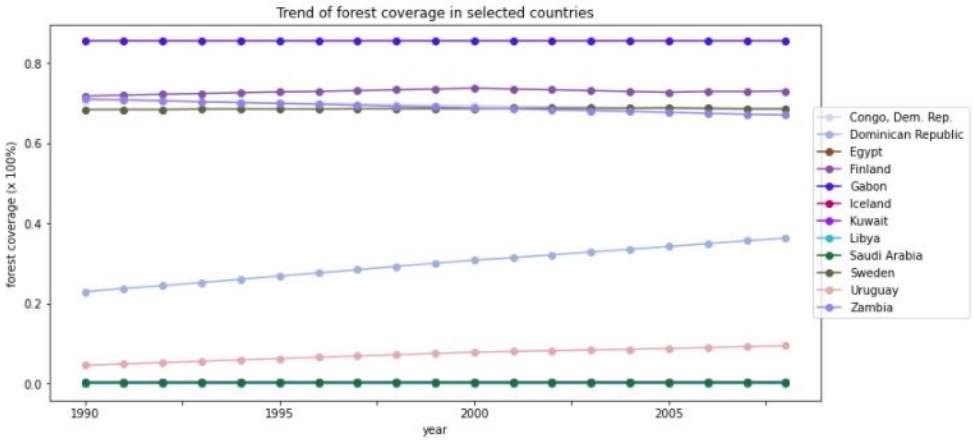
- There is a significant increase for **Saudi Arabia from 1992 to 1999**. After that, the trend is stagnant
- **Nicaragua, Vietnam, and Benin** have also increased gradually
- The other countries that has the highest and lowest mean of agricultural land coverage do not have a significant growth

People affected by flood



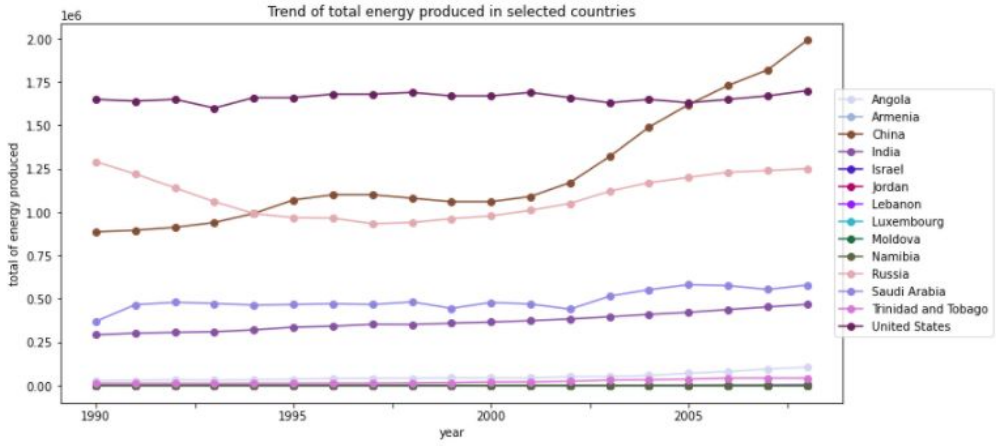
- We can see **China** is still the only country that has a high trend
- Followed by **India** that also has a high trend

Forest coverage



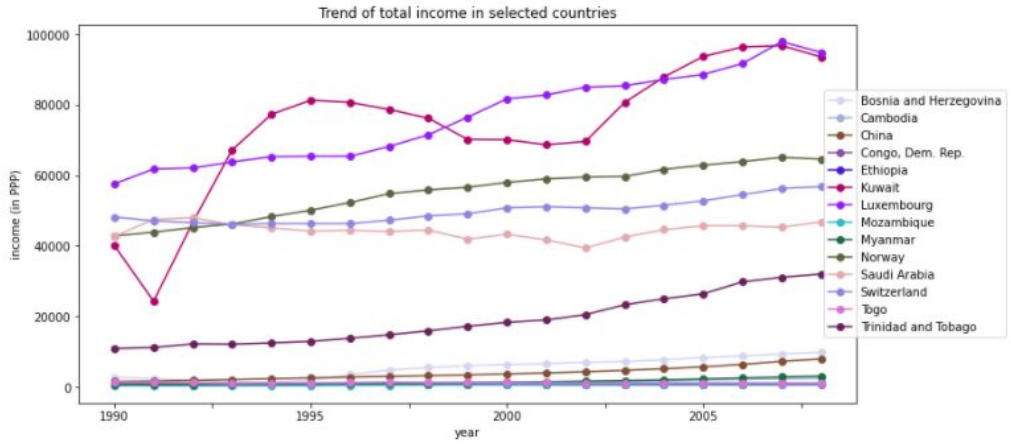
- Dominican Republic **is gradually increasing from time to time**

Total energy produced



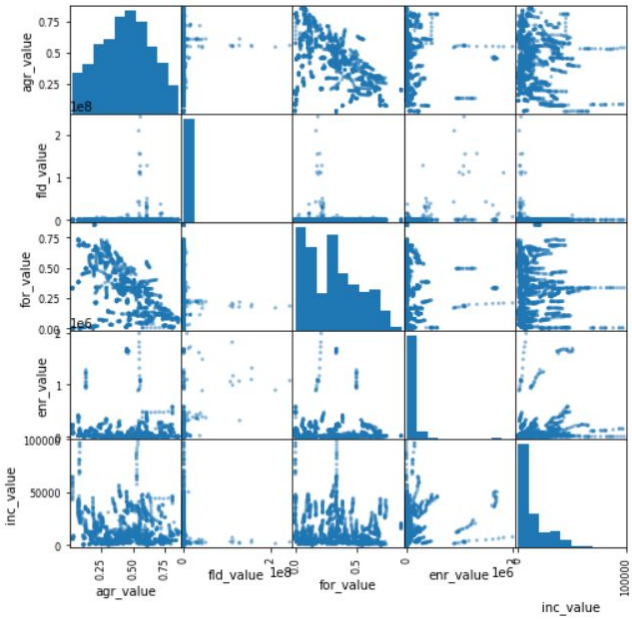
- **China experienced a significant increase from 2000 onwards**, we can see its good growth
- **Russia** ever had a drop from 1990 until 1994. But it started to gradually increase from 2000

Total income



- **Luxembourg** has a **consistent** trend. it's always increasing until now
- **Kuwait** also has a **high** total income, but its trend sometimes ups and downs
- **Norway, Switzerland, and Trinidad & Tobago** started were stagnant at first, but then **gradually increasing** until now
- Meanwhile **Saudi Arabia's total income starts to decrease**

Q3: Is there any correlation between selected indicators?



- Agricultural land coverage and forest coverage have **negative moderate correlation**
- Total number of people affected by flood and total number of energy produced **have positive weak correlation**
- As well as total number of energy produced and total income **have positive weak correlation**

Limitations

- I have mentioned in the report and script that I **use inner join** because if there is a country that does not exist in one of the indicators, it will cause many null values and will be hard to predict what the values are. This actually affects the decrease in the number of records. I believe that I lost some data, but I guess this is the better way than fill the missing value that might cause a **misleading**
- To see a summary or one number of each country in the analysis, I only use mean. This sometimes is nor recommended, but I think I'd be good to capture the outlier
- For the last question, I just calculate the correlation but I don't say any implication since we need a deeper analysis on that
- I don't check the outliers

Notes

- I feel the dataset is a little bit confusing between the other dataset, I don't see any dependent variable. Thus, I had some difficulties setting the research questions. But I still feel this dataset is challenging, so I kept trying
- I also feel that the questions and answers are still to descriptive due to the difficulties of choosing the research questions

Resource

- https://stackoverflow.com/questions/28999287/generate-random-colors-rgb