COM-480 Data Visualization
Project - Milestone 1

Team UNSDG_viz
Sophia Ly, Lorenzo Rovati, Julia Wälti

April 7th 2023

1 Dataset

The data-set [1] used for this project is an aggregation of many United Nations Sustainable Development Group (UNSDG) indexes, created with the goal to assess the sustainability of a country's development. The author took the data from the "International Bank for Reconstruction and Development" and the "United Nation" database. The quality of the data-set is therefore quantified as good. It consists of two features describing time, two features describing location, one feature about the economic development level and 15 features describing the country's state with focus on sustainability. Over half of the features contain many missing values. Due to this and because the features do not fully represent the sustainable development goals of the UN, more data was added. The goal was to find some indicators that represent social, environmental, and economic sustainability.

For completion, data from "Our World in Data" is considered. Environmental indexes such as land degradation and biodiversity is difficult to compare spatially since the natural condition differs. Therefore, for environmental sustainability CO2 emissions and share of the population exposed to air pollution levels above WHO was used. For representing social sustainability, the share of the population with access to safely managed drinking water or undernourished, education rate and literacy rate was chosen. Lastly, the continuous GDP was added since the economic development level is a categorical feature. The data-source is reliable, therefore the obtained data is also assessed with a good quality.

The data-sets cover different spatial and time domain:

- UNSDG: 2002 - 2021, 259 countries
- CO2 emissions: 1750 - 2021, 269 countries
- Literacy rate: 1970 - 2021, 178 countries
- GDP : 1960 - 2020, 225 countries
- Education rate: 1970 - 2050 (predicted), 120 countries

Therefore, some prepossessing needs to be done: One needs to take care that the ID used for linking the different data-set is homogeneous. The features with too many missing values should either be discarded or a small sub-sample should be extracted. One has to look for possible outliers and check if they are within reasonable range or due to mistakes in the data recording.

2 Problematic

Many developing countries have experienced economic growth in the last 15 years. This is linked with a higher consumption of goods and energy. It is predicted that the levels of energy use will significantly increase to power further economic and social development [2].

In order to reach the goals of the Paris agreement, action has to be taken fast to mitigate floods, storms and wildfires [3]. Therefore, the goal of this project is to visualize the development of a country with respect to environmental sustainability regarding climate change driving CO2 emission. The aim is to show how features representing sustainable development are linked with the level of CO2 emission and how its inter-correlation looks like. With visualization through time and space it will be possible to understand the data in a more intuitive manner and to find strategies on how one can decouple CO2 emissions from economic growth as well as on how to ensure a sustainable development.

3 Exploratory Data Analysis

The data-sets are merged following two procedures: an inner and a left merge (the left side being the UNSDG data-set). The inner merge only selects records with matching values in both tables, while the left merge fills missing values with NaNs.

Using the 75^{th} percentile and maximum value, one can see outliers in most features. Since the number of outliers is small, the values exceeding the 75^{th} percentile plus 1 standard deviation are converted to NaNs. The percentage of NaNs in both data-sets is analyzed and can be found in figure 1. The inner join data-set has fewer NaNs and contains information of 50 countries. The left join data-set has more NaNs and contains information of 259 countries. Since the number of countries in the inner join data-set is much smaller, the left joined data-set will be used in the future. The handling of NaNs will depend on the type of visualization.

The distribution of NaNs per row and per country is checked and shown in figure 2. There are 24 features in total, including 4 that that contain no NaNs: country, region, date and year. All rows contain at least 4 NaN values. Some countries have more NaNs compared to others. The distribution of NaN per row and per country is similar. Therefore, one could assume that NaNs mostly stem from the same countries. All countries with more than 18 NaNs per row on average are dropped from the data-set.

The distribution of economic development level is shown in figure 3. One can see that it does not change between the first and last year of the data-set. This indicates that it is determined for the whole time-series. If a higher time resolution is needed, the GDP feature can be used instead.

Some mean distributions for important features can be found in figures 4 and 5. The correlation between all features is shown in figure 6. Some result behaves as expected, such as the negative correlation between literacy rate and rate of no education or the positive correlation between CO2 emissions from fuel consumption and fossil fuel subsides consumption and production. There are also some unexpected results such as the negative correlation between renewable energy share on the total energy consumption to the proportion of population relying on clean fuels.

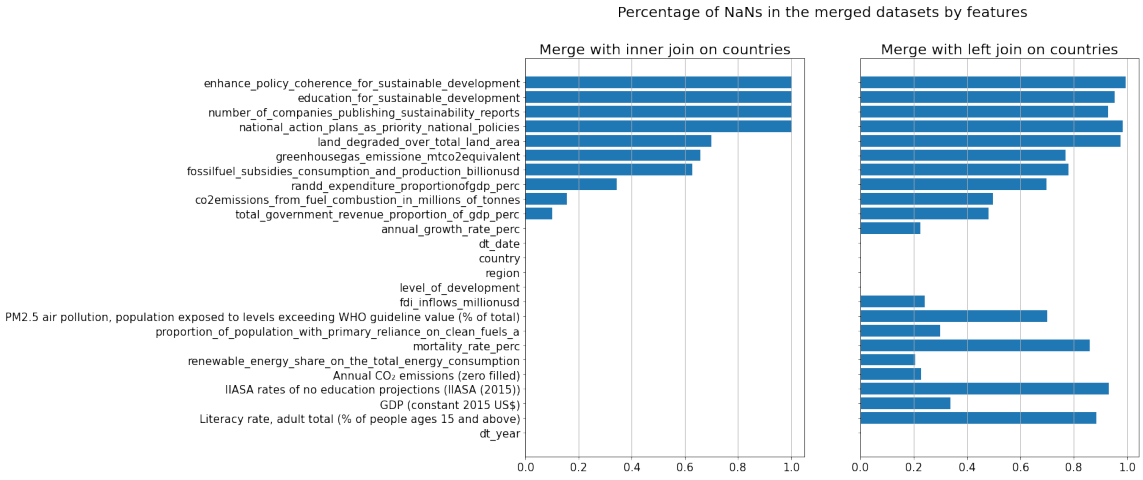

Figure 1

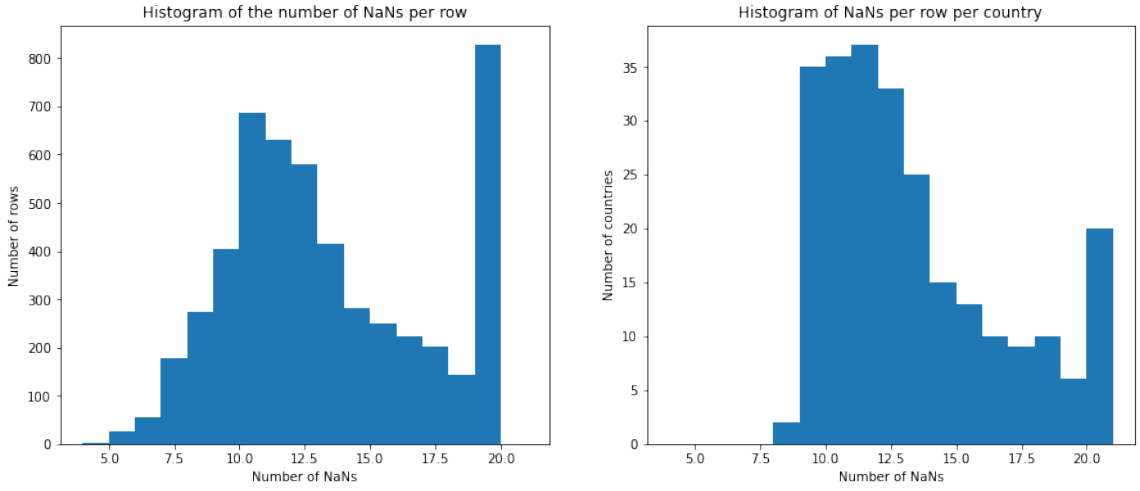

Figure 2: for information, 24 features in total

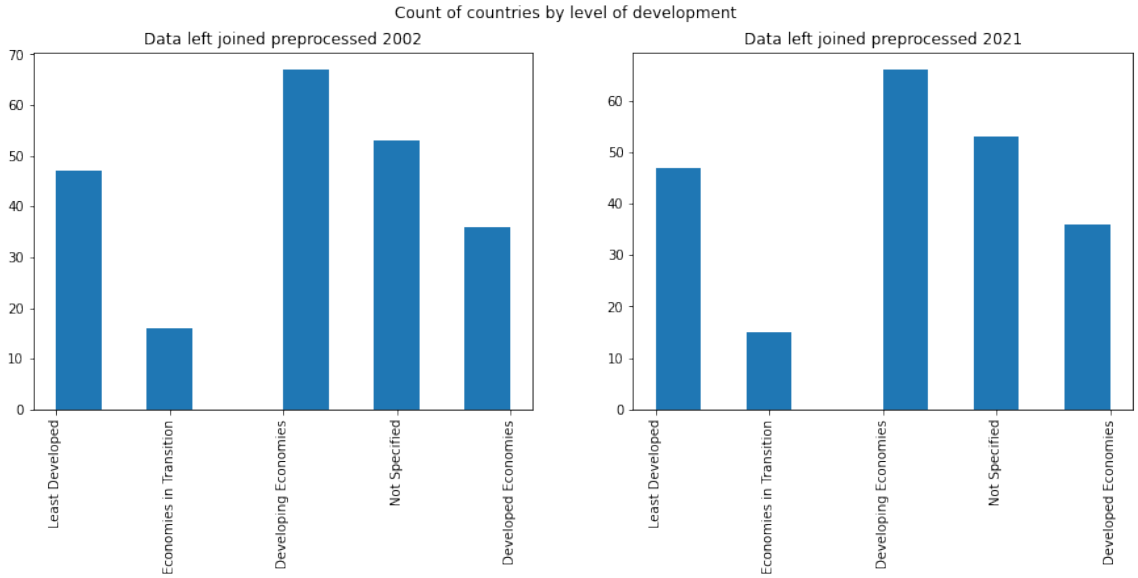

Figure 3

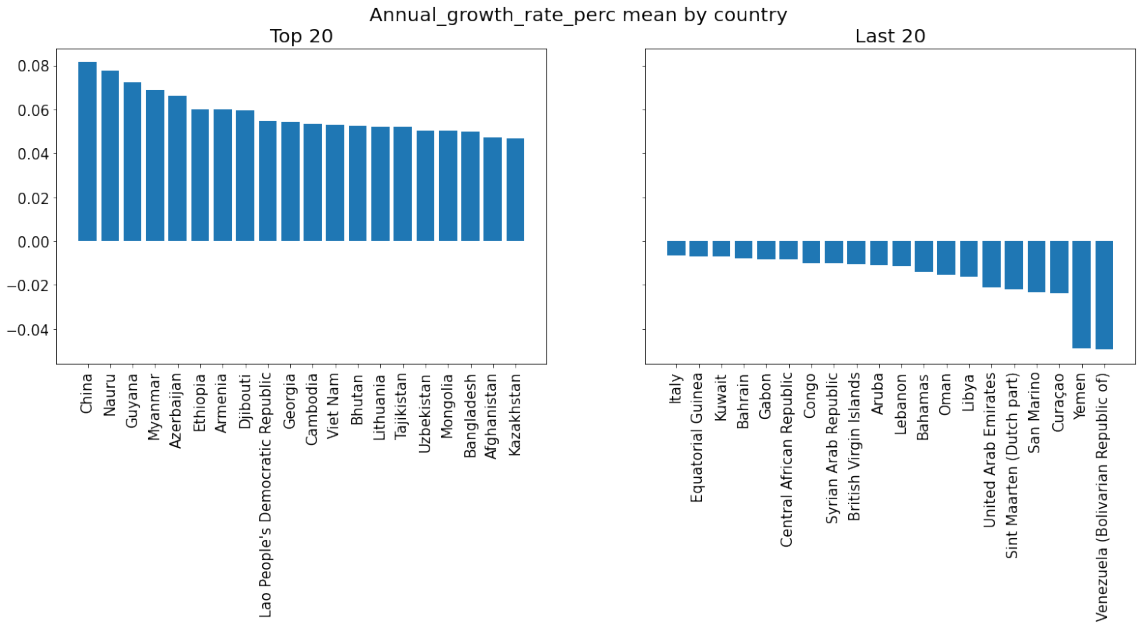

Figure 4

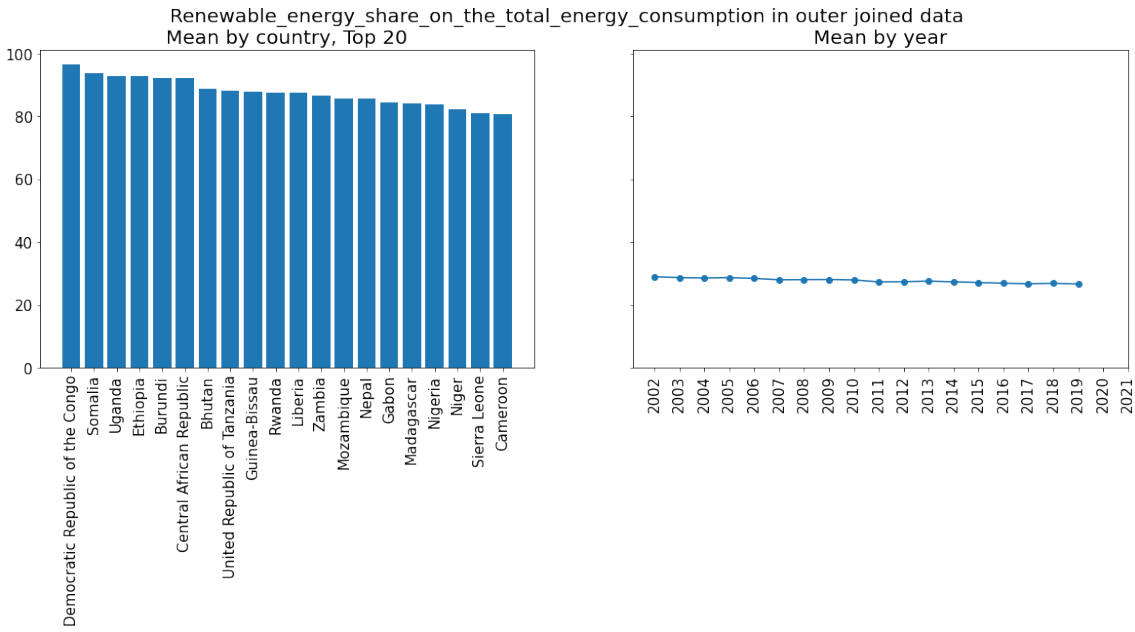
Renewable_energy_share_on_the_total_energy_consumption in outer joined data
Mean by country, Top 20
Mean by year
100
80
60
40
20
0
Democratic Republic of the Congo
Somalia
Uganda
Ethiopia
Burundi
Central African Republic
Bhutan
United Republic of Tanzania
Guinea-Bissau
Rwanda
Liberia
Zambia
Mozambique
Nepal
Gabon
Madagascar
Nigeria
Niger
Sierra Leone
Cameroon
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

Figure 5

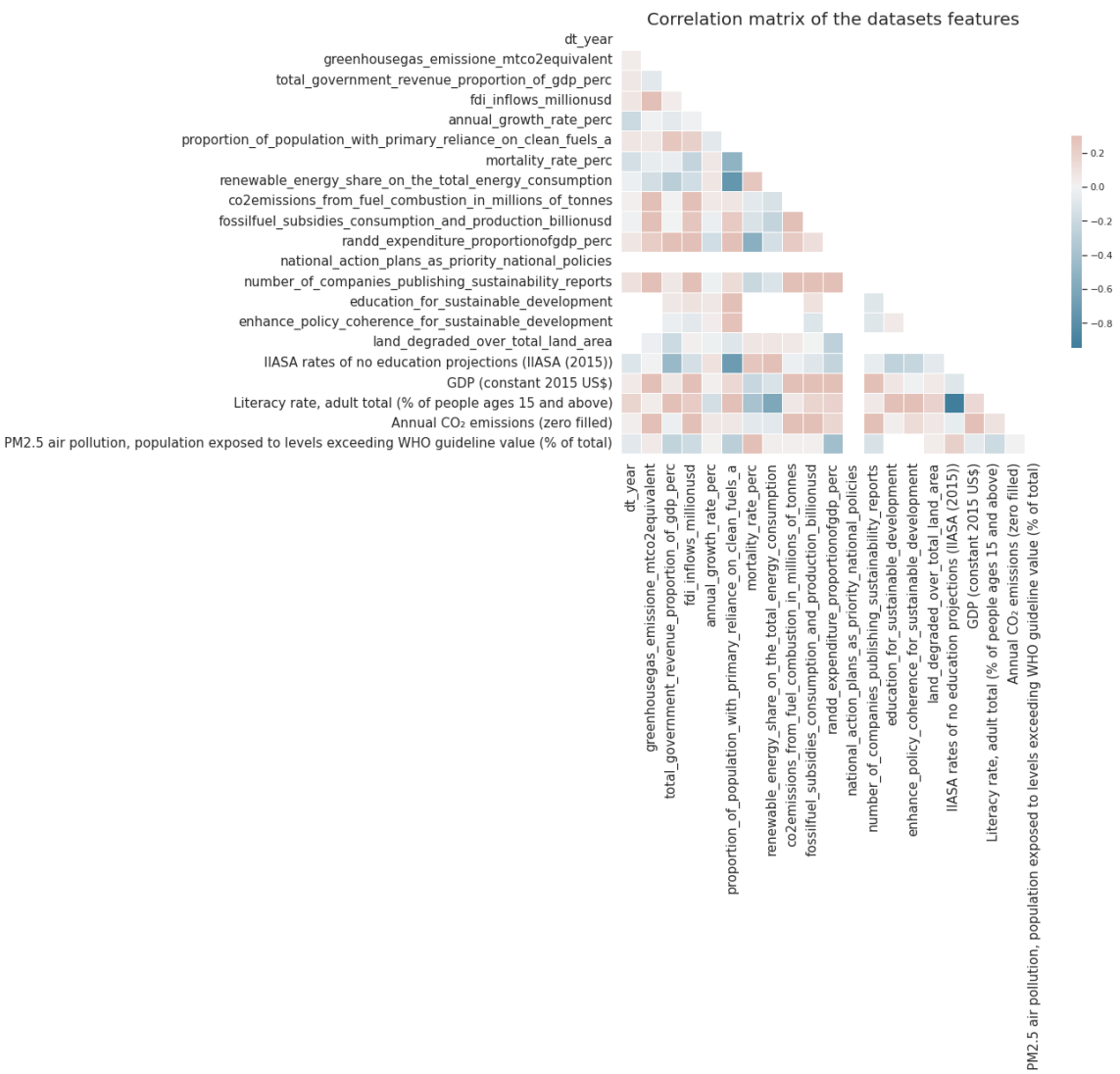
Correlation matrix of the datasets features
dt_year
greenhousegas_emissione_mtco2equivalent
total_government_revenue_proportion_of_gdp_perc
fdi_inflows_millionusd
annual_growth_rate_perc
proportion_of_population_with_primary_reliance_on_clean_fuels_a
mortality_rate_perc
renewable_energy_share_on_the_total_energy_consumption
co2emissions_from_fuel_combustion_in_millions_of_tonnes
fossilfuel_subsidies_consumption_and_production_billionusd
randd_expenditure_proportionofgdp_perc
national_action_plans_as_priority_national_policies
number_of_companies_publishing_sustainability_reports
education_for_sustainable_development
enhance_policy_coherence_for_sustainable_development
land_degraded_over_total_land_area
IIASA rates of no education projections (IIASA (2015))
GDP (constant 2015 US$)
Literacy rate, adult total (% of people ages 15 and above)
Annual CO₂ emissions (zero filled)
PM2.5 air pollution, population exposed to levels exceeding WHO guideline value (% of total)
0.2
0.0
−0.2
−0.4
−0.6
−0.8

Figure 6

4 Related work

The data-set [1] concerning the United Nations Sustainable Development Group indexes, was used to explore the importance of features among themselves. In order to do this, models were fit to each feature based on the remaining features. The goal of this project on the other hand is to visualize a countries development with respect to CO2 emissions and sustainable development.

Most studies focus on the relationship between economy and CO2 emissions. While economic growth is the main target of many countries, the sustainable aspect regarding the social standpoint should not be forgotten. This projects therefore aims to visualize all three sections. With this, exemplary countries could be detected that have achieved a sustainable development program, taking all sections into account. By making the visualization interactive, the user can explore the data freely and understanding of the data can be improved by linking multiple visualizations together.

Sources of inspirations for this project are the visualizations of education from the open world data website[4]. They visualized a time-series in form of a heat-map representing the countries shape. In order to show the relationship between CO2 emission and the other features a partial dependence and individual conditional expectation plot can be done for each feature. An example can be seen on the scikit website[5]. Since the features in this data-set are dependent from each other, a diagonal correlation matrix can help to understand the data better. An example is shown on the Seaborn website[6].

References

[1] V. GIATTI. (2022) Onu sustainability of countries development. [Online]. Available: https://www.kaggle.com/datasets/vittoriogiatti/unsdg-united-nations-sustainable-development-group

[2] P. Benoit, "Energy and development in a changing world: A framework for the 21st century," *Center on Global Energy Policy*, 2019.

[3] H.-O. Pörtner, D. C. Roberts, H. Adams, C. Adler, P. Aldunce, E. Ali, R. A. Begum, R. Betts, R. B. Kerr, R. Biesbroek *et al.*, *Climate change 2022: Impacts, adaptation and vulnerability*. IPCC Geneva, Switzerland:, 2022.

[4] M. Roser and E. Ortiz-Ospina, "Global education," *Our World in Data*, 2016, https://ourworldindata.org/global-education.

[5] Scikit-learn: Partial dependence and individual conditional expectation plots. [Online]. Available: https://scikit-learn.org/stable/auto_examples/inspection/plot_partial_dependence.html

[6] seaborn: Plotting a diagonal correlation matrix. [Online]. Available: https://seaborn.pydata.org/examples/many_pairwise_correlations.html