SCA 2024 - Sustainability Championship of America 2024

Fabian Hildebrandt
Matriculation Number: 23392857
Email: fabian.hildebrandt@fau.de

***Abstract*—'Drill, baby, drill' is an old republican slogan, that became famous during the recent election campaign of Donald Trump - the future president of the USA. Despite countless extreme weather incidents and the visible effects of the climate change, many American countries decide to further expand non-sustainable energy supplies, deforest the rain forests and cause harm to the nature. The Sustainability Championship of America 2024 (SCA 2024) presents a comparison of national sustainability performance of the Americas in three categories: environmental, economical, and social. The analysis uses carefully selected, comparative indicators that avoid penalizing smaller or economically emerging nations. The report introduces three distinct recognition categories: overall champions demonstrating consistent high performance, rising stars with significant recent improvements, and latecomers struggling with sustainability challenges. The scope of this report is to explain the data sources of the indicators and the data pipelines, which are used to extract, transform and load the data for further analysis.**

I. Question

Which countries are America's current sustainability champions, the rising stars and latecomers?

II. Data Sources

The goal is to have a comprehensive data set for most of the 45 countries of North America, Latin-America and the Caribbean over the last years. Therefore, two common data sources are used.

A. World Bank Group

The *World Bank Group* offers a really comprehensive composed dataset called *World Development Indicators* containing a large variety of environmental, social and economical indicators. The database of the *World Bank* is composed of numerous publications, statistics and evaluations [1]. It offers a graphical user interface and an application programming interface (API) to flexibly select and extract the needed indicators. The data has a tabular structure and can be downloaded in different formats (`CSV`, `XLSX`, `Pandas DataFrame`). The data obtained via the API has a wide format as shown in the table I.

economy	series	YR1850	YR1851
AFG	EN.GHG.CO2.PC.CE.AR5	42	43
	EN.GHG.ALL.PC.CE.AR5	22	23

TABLE I
World Bank Group - Data structure

The data has a double-index (economy and series) and a column for each year starting with 'YR'. The economy index is the country code and the series index is the ID of a specific indicator listed in the data bank of the *World Bank Group*. To obtain a series containing the values over all years, both the economy and series need be provided.

The data is highly reliable, as it originates from governmental sources and research institutes. The World Bank Group has a well-described selection process, which examines the trustworthiness, relevance, quality and reliability of the data. Common data sources for the indicators are OECD, NSD, UN, UNHCR, UNFPA and many more. The data is either annually or quarterly updated. However, the World Bank notes that many countries have incomplete data reporting, making it hard to obtain a complete data set for a certain indicator. Possible limitations regarding correctness, completeness and relevance are transparently described in the metadata of each indicator [2].

The licenses of all indicators are contained in the metadata. Generally, the World Bank has an open data policy using the Creative-Commons Attribution 4.0 international licenses (CC-BY 4.0) for most of the indicators [3]. These licenses allow a free usage. However, an attribution to the original data source is required when using, adapting, or reproducing the data. Changes to the data must be indicated. The license conditions will be met in the following reports by adding a detailed list with all indicators, their origin, performed changes and their licenses in a license file in the project GitHub repository.

B. Our World in Data

Our World in Data is a non-profit research initiative that makes knowledge accessible and understandable by providing datasets, reports and case studies. It was founded by Oxford University researches and follows common academic standards. The datasets originate from various organizations (UN, WHO, World Bank etc.), research articles, governmental data sources and specialized institutes and cover a wide range of topics mostly focused on showing the most challenging problems of humanity [5]. The CO_2 emissions data is available in a `CSV`-file in a GitHub repository and has the format, which is demonstrated in the table II. The format is called wide

country	year	iso_code	indicator_1	indicator_2
Afghanistan	1850	AFG	42	22
Afghanistan	1851	AFG	43	23

TABLE II
Our World in Data - Data structure

format, in which each column is an indicator and each row can be referenced by specifying the country and year or iso_code and year.

All visualizations, the data, and code produced by Our World in Data are open access under the Creative Commons BY license (as the World Bank Data). The usage is possible as long as the source is provided and the authors are credited. The data produced by third parties made available by Our World in Data is subject to the license terms from the original third-party authors [5]. The sources are already part of the metadata and will be cited and displayed in the future reports and presentations and the GitHub repository.

III. Data Pipeline

A generic Python pipeline class is designed to extract, transform and load the data (ETL). This way, further data sources can be added with low efforts. Two child classes of this pipeline class are implemented for the two data sources. In addition, a configuration file can be used to change the API settings, add further indicators, change the required metadata attributes and the included countries. This configurability ensures quick adaptations in the later data analysis, which might be needed, if the data doesn't meet the quality requirements. The following figure 1 shows the general pipeline structure for a single data source.

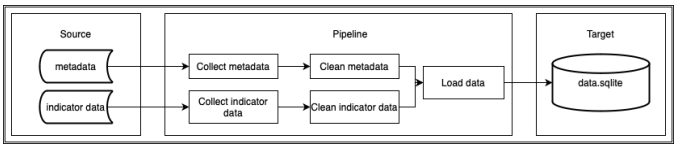

Fig. 1. Structure of the designed pipeline

The extraction step is different for each data source. Therefore, the extract and preprocessing methods are specified for each child class separately. The transformation is split into a data source-specific step and a generic step. The data source-specific step deals with transforming the data format from the data source to a generic long format as shown in the table III.

country	year	indicator	value
AFG	1850	indicator_1	42
AFG	1851	indicator_1	43
AFG	1850	indicator_2	22
AFG	1850	indicator_2	23

TABLE III
GENERIC LONG DATA FORMAT

During the generic transformation step of the indicator data (called `check_and_clean`), duplicate and Not a Number (NaN) values are removed and two data quality indicators are calculated for each indicator.

1) NaN ratio: Ratio of NaN values for the indicator.
2) Country NaN ratio: Ratio of countries, for which no data is available for the indicator.

The resulting data of each step of the pipeline can be reviewed using the `debug` mode of the pipeline. After the data transformation, the preprocessed metadata and indicator data are loaded to an `SQLite` database (using `sqlite3`) with the tables *metadata* and *indicator_data*. This method is called `load_data`.

A. World Bank Group

As described, the design goal of the pipeline is to be flexible to later changes in the required indicators. Therefore, the *World Bank Group* Indicators API is used, which offers access to nearly 16,000 time series indicators over the last 50 years (and more) [6]. The open source python library WBGAPI [7] is used to access the API of the World Bank. It is licensed under the MIT License, which allows a full usage of the software without restrictions. It has intuitive methods to select multiple indicators, countries and time periods in a single request. Therefore, it is a suitable tool to access the *World Bank Group* indicators.

The preparation step is to collect all required indicators. Therefore, the search function of the World Bank Open Data is used. By entering keywords, it is possible to identify suitable indicators. These indicators have a unique id (e. g. NY.GDP.MKTP.CD for GDP (current US$)), which are then added to the pipeline configuration file. The first data extraction step collects the indicator metadata using the `wb.series.metadata.get` method. It was quite challenging to convert the returned `metadata` object to a `dictionary`, as the usage of the method is really sparsely documented. The resulting `dictionary` is converted to a `Pandas Dataframe`. After the extraction, the metadata `DataFrame` is cleaned. Only the required columns defined in the configuration (e. g. ID, indicator name, license type, unit, definition) are kept and the rest is dropped. It is important to store the license information for the later analysis, to give accreditation to all data sources, which are used (as required with the CC-BY 4.x licenses). This function is already implemented and will be executed as soon as the scope of indicators used for the analysis is fixed. Afterwards, the indicator data is read out using the `wb.data.DataFrame` method, which requires a list of the series (indicators), economies (countries) and a specified time range (start year, end year/ number of recent years) and returns a `Pandas DataFrame`. The data is then preprocessed. The first step is to reset the indexes (country, indicator) and then convert the data from the wide-format to the long format using the `Pandas pd.melt` method. Afterwards, the columns are renamed to match the defined format (country, indicator, year, value). Then, the generic `check_and_clean` method is used to clean the data and check the data quality. Afterwards, the data is loaded to an `SQLite` database using the generic `load_data` method.

B. Our World in Data

The pipeline for the second data source is quite similar. However, compared to the first data source, the scope of the available indicators is fixed. The indicator data and metadata are available in a GitHub repository as `CSV`-files [8]. The `CSV`-files are extracted using the `requests` library and then

converted to `Pandas DataFrames`. The cleaning of the metadata is almost similar to the first data source and only necessary columns (indicator name, description, unit and data source) are kept. The preprocessing of the indicator data is slightly different. First, the configured indicators are kept and the rest of the data is discarded. Then, the columns are renamed to match the uniform naming. Afterwards, the data is filtered, such that it only contains the American countries and is in the configured date range. Afterwards, the data is converted to the standardized long format. Finally, the `check_and_clean` method and the `load_data` method are used to finish the pipeline by loading the data in the `SQLite` database.

IV. Results and Limitations

The final data format of the pipeline is `SQLite`, which is more performant compared to reading text-based files like `CSV` or `XLSX` sheets. The standardized long format of the resulting data comes with a few downsides and many advantages compared to the wide format. The downsides are, that it is hardly human-readable, occupies more memory/ storage space and it might result in more complex queries. However, the data size is in the range of a few MB so the required storage space doesn't matter. However, the long format is more flexible for data cleaning and for tasks related to data aggregation, visualization and adding new indicators. Specifically during the quality check, it is important to remove NaN values, which is easier in the long format. A potential issue that might occur is the large ratio of NaN values or missing countries. On average, 33.3 % of the countries didn't report data across all indicators and more than 30 % of the values are NaN values. This is a serious limitation. The goal is that the SAC 2024 is a comprehensive and fair comparison. Therefore, it is essential to have a dataset that covers the most recent years for the most countries. As reported by the two data sources, it is hard to obtain complete data, because many countries don't report certain indicators. Data imputation would distort the results due to the long reporting cycles (annually/ quarterly). As a result, it might be necessary to switch the indicators during analysis to allow a fair competition. This is possible with a low effort due to the large variety of available indicators in the *World Bank Group DataBank*, thhe configurability of the pipeline and the quick feedback regarding the data quality.

References

[1] World Bank Group (2024) - "World Development Indicators" Published online at DataBank. Retrieved from: 'https://databank.worldbank.org/source/world-development-indicators/' [Online Database]

[2] Neil Fantom and Tariq Khokhar (2014) - "How we do Open Data: #1 - choosing development indicators" Published online at World Bank Blogs. Retrieved from: 'https://blogs.worldbank.org/en/opendata/how-we-do-open-data-choosing-development-indicators' [Online Resource]

[3] World Bank Group (2024) - "Public Licenses" Published online at World Bank Data Catalog. Retrieved from: 'https://datacatalog.worldbank.org/public-licenses#cc-by' [Online Resource]

[4] Hannah Ritchie, Pablo Rosado and Max Roser (2023) - "CO_2 and Greenhouse Gas Emissions" Published online at OurWorldinData.org. Retrieved from: 'https://ourworldindata.org/co2-and-greenhouse-gas-emissions' [Online Resource]

[5] Our World in Data (2024) - "FAQs- Our World in Data" Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/faqs#citing-work-produced-by-third-parties-and-made-available-by-our-world-in-data' [Online Resource]

[6] World Bank Group (2024) - "About the Indicators API Documentation" Retrieved from: 'https://datahelpdesk.worldbank.org/knowledgebase/articles/889392-about-the-indicators-api-documentation' [Online Documentation]

[7] Tim Herzog (2022) - "WBGAPI: A Simple Python Interface for the World Bank's Data API" Retrieved from: 'https://pypi.org/project/wbgapi/' [Software]

[8] Our World in Data (2024) - "Data on CO2 and Greenhouse Gas Emissions" Published online at GitHub. Retrieved from: 'https://github.com/owid/co2-data/tree/master' [Online Repository]