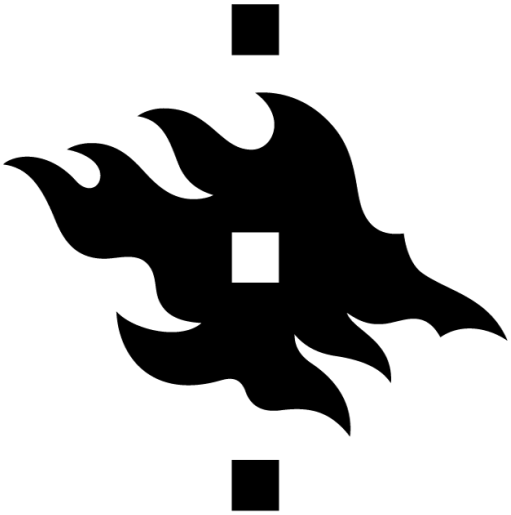

Design an interactive visualization

Course project
Interactive Data Visualization

Enrico Buratto

University of Helsinki

Faculty of Science

Academic Year 2021-2022

Contents

1 Introduction **2**
1.1 The problem 2

2 Data **2**
2.1 Chosen dataset 3
2.2 Data pre-processing 3
2.3 Data usage 4

3 Task abstraction **4**
3.1 Iterations 5

4 Design rationale **5**
4.1 Visualization 6
4.2 Interaction 7

5 Lessons learned **8**

Appendices **9**

A Used variables **9**

B Screenshots from the visualization **10**

1 Introduction

The following document describes the work done for the Interactive Data Visualization course project. This project started in the beginning of April 2022, and went through several iterations as the course progressed; all the steps have already been mentioned in the four learning diaries, but this document describes them in a detailed fashion.

This report is composed as follows:

- The next subsection describes in detail what is the scope of the project; this includes the problems that need to be solved through designing the interactive visualization;
- §2 describes the used dataset and the procedures performed to make it usable with the visualization software;
- §3 describes the task abstraction, *i.e.* the various visualizations that were implemented to address the problem. It also reports the iterations I went through during the development;
- §4 discusses about the design rationale of the visual representation and interaction, *i.e.* the design choices that have been made in order to have a coherent and aesthetically pleasing visualization;
- §5 describes the lessons learned from this project.

Different tools have been used to achieve the final visualization; the most important are:

- Pandas library[1] for data loading and preprocessing;
- Jupyter notebooks[2] for quick data visualization during development;
- Tableau[3] for the final interactive visualizations.

1.1 The problem

The idea of this project is to study the global trends in energy consumption and electricity generations between 1900 and 2019; even though, as we will see in the next sections, the dataset contains information until 2021, the last two years are incomplete, hence they are excluded by this study.

In order to study these trends, as specified below, different visualizations had to be implemented, the most important of them being the map visualizations. The final scopes of the visualizations are then:

- To offer an interactive way to access data for each country and each year, offering useful tools like filters in order to being able of studying different aspects of the problem;
- To offer a more static, though still interactive, way to visualize the general trends through the years;
- To have information quickly available, without the need for complicated data analysis from the user side.

2 Data

Before starting with the task abstraction and the implementation of the visualizations, an introduction on the data should be done; the next subsections describe where the data come from, how it has been processed and how it is used for the scope of this project.

2.1 Chosen dataset

The data I used is *Data on Energy by Our World in Data*[4]; this dataset is, as the authors report, a collection of the most important metrics maintained by Our World in Data[5], which is an influential online publication with a focus on large global problems. This dataset tries to keep track of many different energy-related factors in the world; in order to do so, Our World in Data combines different sources and put all the data together in a coherent manner.

The data reported in the dataset belongs to three categories:

- **Data on energy consumption**: this data regards, as the name suggests, information on energy consumption of each country from 1900 to 2021. In this conceptual section we can find data about consumption from different energy sources, *e.g.* coal, oil, gas, wind, sun, nuclear. We can also find information about the change in consumption through the years;
- **Data on electricity generation**: this data regards, as for the energy consumption, information on energy generation; again, the collected data regards all the countries of the world between 1900 and 2021, and here we can find information about electricity generation from different energy sources;
- **Other variables**: this data is a set of different useful variables, including information on year, country, gdp (gross domestic product) and population of the countries.

This dataset is composed by two different files: a *codebook* and the dataset itself. The former contains the description of every variable of the latter, and it has been very useful for data preparation.

The data was available in three different formats: `CSV`, `XLSX` and `JSON`. Since, as reported in the below section, I used the python's Pandas library for pre-processing, I chose the first because of better compatibility; however, the final result was an excel file in order for it to be compatible and ready to use with Tableau.

2.2 Data pre-processing

When dealing with the dataset, two main problems had to be taken into account:

- The amount of variables is huge: the dataset is, in fact, composed by 125 unique variables, some of them being useless for the scope of the project and some other being redundant;
- The data is incomplete: for certain countries, in fact, useful data starts from 1965 circa; this brought the problem of dealing with missing data, and it therefore needed to be solved.

The biggest effort has been to select the right variables to work with; this took several iterations, since it was not clear which data to represent in the first instance. A complete list of the used variables is reported in Appendix A; note that:

- The ISO code of the countries has been removed since Tableau is able to recognize the countries in an automated fashion from their name, and was therefore useless;
- The change of energy consumption and generation through the years has been removed since these were redundant information: the change can be inferred, in fact, quite easily from the remaining variables.

After having chosen the variables to use, there was now the problem of dealing with missing data. In order to address this problem, I decided to perform some data pre-processing using

python's Pandas library; thus, I loaded the CSV file into a Jupyter Notebook and I started working on the data. In the first iteration, I decided to fill all the empty values with zeros; however, using this data led to quite incomplete map and bar charts visualizations. Therefore, I decided to adopt a different strategy to fill the data: I filled the empty values with medians and averages of the column values. This way, the data was now interesting to analyze and represent with Tableau.

2.3 Data usage

After having performed the first pre-processing, the data was loaded into Tableau as an Excel file (`.xlsx`). As already reported, the first iteration was unsuccessful due to missing data, hence the second iteration introducing the median and the mean of the variables to fill the empty values. In the second iteration, however, I reputed the data to be interesting enough to be represented; therefore, I started from here for the subsequent work, which is described in the next sections.

3 Task abstraction

In order to study the trends in energy consumption and electricity generation, different visualizations could be chosen; for instance, simple line charts could be used to display energy consumption in function of the years. However, this could result in a cluttered visualization due to the high amount of variables.

After some careful thinking, then, I first decided to implement four visualizations, two for type of data to display. To be more precise, these are:

- A map visualization containing information about electricity generation. The standard data visualized when the visualization is first open is the total electricity generation for each country from 1900 to 2019. The user, then, can interact with this data in various way:
  - Zoom on the map and/or on single countries;
  - Hover with the mouse on the countries to visualize all the pertinent data for the selected time span;
  - Click on a country to isolate it and visualize all the statistics for the selected time span;
  - Filter the data with regards to different variables.

  The variables the user can use as a filter are the following:
  - Population;
  - GDP;
  - Types of source: biofuel, coal, gas, hydro, nuclear, oil, solar and wind;
  - Year.
- An interactive bar chart representing almost the same energy generation data but in a different fashion: the user can select the country to study and then visualize the energy consumption from different sources through the years; this last parameter can be modified using a simple slider as a filter. The standard data visualized when the visualization is

first open is the trend from 1900 to 2019 of a random country. The user can then interact with it in various ways:

  - Hover on the single bars that make the stacked bars to visualize the pertinent data;
  - Click on one or multiple parts of bars to isolate them and visualize all the statistics for that;
  - Filter the data by year.

- A map visualization equivalent to the first one, but with information regarding the electricity consumption;
- An interactive bar chart equivalent to the first one, but with information regarding the electricity consumption.

As reported in the learning diaries and in the below section, however, I had some more time to work on the project; therefore, I decided to implement two more visualizations of the same type. These are two bar chart races[6], and their main scope is to visualize interactively, almost like a video, the top countries for energy generation and energy consumption through the years. The user can interact with them by just pressing the play button and then visualize the bars expanding and contracting, or selecting a specific year to know which was the country with the highest electricity production or consumption.

All these visualizations can be viewed and used at the link provided[7]; however, for completeness, some screenshots are provided at Appendix B.

3.1 Iterations

As already mentioned, I implemented several iterations in order to have a complete and pleasing visualization. These are the following:

1. In the first iteration, I downloaded the data and prepared it for Tableau. Then I loaded the data in the latter and I started making the first map visualizations;
2. As already stated in §2, the first data pre-processing was kind of unsuccessful due to the missing data; therefore, in the second iteration I performed some other pre-processing and I restarted with the map visualizations;
3. In the third iteration I finished the barebone map visualizations and I started making them more meaningful and aesthetically pleasing; this design rationale is described in §4;
4. In the fourth iteration I implemented the two stacked bar charts, again starting with a barebone version and then improving them from a meaningfulness and aesthetic perspective;
5. In the fifth iteration I created the final Tableau story to display as a coherent visualization, adding also the title page and the two bar chart races;
6. A sixth iteration was performed to improve again the visualizations, following again the same principles.

4 Design rationale

As already said, during the development of the project I went through several iterations. However, for convenience only the final results are reported in this section, although a comparison

between different possible implementations was performed. This chapter is divided into two subsections: visualization and interaction; the former is about the visualization design principles that have been followed, while the latter explains the interaction choices that have been taken.

4.1 Visualization

We can say that the design rationale behind the visualizations I implemented is given by two principles: the L.A.T.C.H. principle[8] and Tufte's design principles[9].

The former is a way to categorize the type of visualizations into five macro-areas: Location, Alphabet, Time, Category, Hierarchy. The implemented visualizations can be framed in different areas: the map visualizations belong to the Location type of visualization, while the bar charts to the Time type. Also the order of the visualizations on the Tableau story (the main deliverable) is not casual, even if it is just a small detail: the visualizations are, in fact, close by category: the two maps, the two stacked bar charts, the two bar races.

Coming to Tufte's principles, which are the most crucial in this discussion, there are several aspects that had been taken into consideration while developing the visualizations; these are listed in the next paragraphs.

Principle of graphical integrity

Graphical integrity is an umbrella term for several different aspects. In the developed visualizations, graphical integrity is given by the following features:

- In the map visualizations, the **background** is always the same. They also respect the **area ratios** (as most as possible, since it is not possible to perfectly represent a sphere on a plane) and the standard maps are the same but with different colors; this is done to underline that they display different data (one for energy consumption, the other for energy production) but in the same way;
- **Labeling** is extensively used to defeat ambiguity; this is specially true for the filters and for the measure names in the stacked bar charts. For the latter, I made the precise decision to keep the measure names along with the hover in order to not lose the reference of the different types of fuel;
- The **area** of the values for the stacked bar charts and for the bar chart races is proportional to the quantities represented; since the areas are rectangular, there was no need to introduce a **lie factor**;
- The graphics does not quote data out of context: everything is precisely labeled in order to let the users understand immediately what they are watching at.

Data-ink ratio and chart junk

The data-ink principle says that the the ratio between the *ink* (*i.e.* the text, images, etc) which represent data and the total *ink* of the visualization should be maximized; that is, useless information should be avoided. Chart junk principle, similarly, suggests to not add useless elements to the charts.

Even though, as reported in the above section, some representation is redundant (*e.g.* the measure names in the stacked bar charts), no extra *ink* has been used for the implemented visualizations: they have been built in a minimalist way, yet reporting all the needed data. The

only exception to these laws could be the title page; however, this page is not meant to convey information and can be easily skipped, so it was deemed not to be a problem.

Data density

The data density principle states that the ratio between the number of entries and the area of the data graphic should be maximized, along with the size of the data matrix itself. This principle has been applied mainly for the development of the map visualizations: when the user puts on hover the mouse on a country on the map, all the available data is displayed. Note that, in order to avoid possible confusion and to make the interaction more intuitive, the different energy sources are listed alphabetically.

A final note to make is the **proximity**: it is a good practice to put similar information close, in order to reduce the moving time inside the visualization. This have been made with the information of the map visualizations and with the filters on both the map and the stacked bar charts.

4.2 Interaction

Coming to the interaction choices, we can frame them into the different actions the used can perform in order to explore the data inside the visualization. These actions are reported in the next paragraphs; in addition to this, a last paragraph is reserved to the information seeking mantra, which has been crucial to the development.

Select

The action of selecting is to mark something as interesting and then to explore further combining this with other actions. In the visualizations, the user can select mainly two things: the countries on the map and a portion of bar in the stacked bar charts. With the former, the user can then *isolate* the visualization to the selected country only, and can then play with the filters or play the year animation.; this applies as well with a multiple selection of countries. With the latter, the user can isolate a single energy source, and play with the filters as well, or just isolate a single value to read the data.

Abstract/elaborate

The action of abstracting is the ability to show less details, while elaborating is the ability to show more details. In the visualizations, this is possible both on the map and on the stacked bar visualizations. In the former, the user can hover with the mouse on a country to visualize all the pertinent data for the selected time span; in the latter, the user can hover with the mouse to have a precise number instead of the size of the bar only.

Filter

The action of filtering in the ability to show something conditionally; this is probably the most important action the user can perform on the developed visualizations. As already mentioned above in this document, on the map visualizations the user can filter the data on different features:

- Population;
- GDP;

- Year;
- Different energy sources: biofuel, coal, gas, hydro, nuclear, oil, solar and wind.

All these filters, except for the year one, are implemented with "at most sliders": the user can select a span between the minimum in the dataset and a chosen maximum.

On the stacked bar charts, the user can filter the data on two features: year and country. In this case, only the year is a slider, while the country filter is a dropdown menu.

Information seeking mantra

The information seeking mantra can be summarized as "*Overview first, zoom and filter, details on demand*". As already mentioned, every time the users open a page on the Tableau story they will find a standard visualization. This displays only a discrete amount of data; the user can then interact with filters and, on the maps, with zooming in and out, and display further details by hovering and clicking with the mouse.

5 Lessons learned

In conclusion, I can say that this project gave me some more comprehension on interactive data visualizations. The most important aspect of it is that it helped in translating the theoretical concepts we saw during lectures in practical knowledge. To be more specific, I learned:

- How to effectively represent multi-dimensional data: since the dataset contained a high amount of useful variables, I had to reason on how to represent them in an uncluttered and simple to access way;
- How to apply Tufte's design principles to visualizations: this was also in the scope of the second course assignment, but this project had a different order of size and, therefore, it has been more difficult to apply this concepts and guidelines to it;
- Different types of actions and how to implement them when it comes to interactive visualizations; of great importance also the information seeking mantra, which guided me during the development.

From a more materialistic perspective, I've also learned how to work with Tableau which is, if I have to give a subjective opinion, a strong and powerful tool with sometimes too much complexity: in an expert's hands, it can be used for a huge amount of tasks, but for a beginners it can be tough. However, now that I used it for a higher-size project, I can say I am more familiar and confident with it. For instance, it took me quite much time to implement the first map visualization because I was still uncomfortable with the different parts of the software; the last visualizations, on the other hand, took less time to be completed since I had learned how to move inside it.

Appendices

A Used variables

column	description
country	Geographic location
year	Year of observation
population	Total population
gdp	Total real gross domestic product, inflation-adjusted
biofuel_consumption	Primary energy consumption from biofuels, measured in terawatt-hours
coal_consumption	Primary energy consumption from coal, measured in terawatt-hours
gas_consumption	Primary energy consumption from gas, measured in terawatt-hours
nuclear_consumption	Primary energy consumption from nuclear power, measured in terawatt-hours
oil_consumption	Primary energy consumption from oil, measured in terawatt-hours
other_renewable_consumption	Primary energy consumption from other renewables, measured in terawatt-hours
hydro_consumption	Primary energy consumption from hydropower, measured in terawatt-hours
wind_consumption	Primary energy consumption from wind, measured in terawatt-hours
solar_consumption	Primary energy consumption from solar, measured in terawatt-hours
renewables_consumption	Primary energy consumption from renewables, measured in terawatt-hours
primary_energy_consumption	Primary energy consumption, measured in terawatt-hours
fossil_fuel_consumption	Fossil fuel consumption, measured in terawatt-hours. This is the sum of primary energy from coal, oil and gas.
electricity_generation	Electricity generation, measured in terawatt-hours
biofuel_electricity	Electricity generation from biofuels, measured in terawatt-hours
coal_electricity	Electricity generation from coal, measured in terawatt-hours
gas_electricity	Electricity generation from gas, measured in terawatt-hours
nuclear_electricity	Electricity generation from nuclear power, measured in terawatt-hours
oil_electricity	Electricity generation from oil, measured in terawatt-hours
other_renewable_electricity	Electricity generation from other renewable sources including biofuels, measured in terawatt-hours
hydro_electricity	Electricity generation from hydropower, measured in terawatt-hours
wind_electricity	Electricity generation from wind, measured in terawatt-hours
solar_electricity	Electricity generation from solar, measured in terawatt-hours

column	description
renewables_electricity	Electricity generation from renewables, measured in terawatt-hours
fossil_electricity	Electricity generation from fossil fuels, measured in terawatt-hours. This is the sum of electricity generation from coal, oil and gas.
other_renewable_exc_biofuel_electricity	Electricity generation from other renewable sources excluding biofuels, measured in terawatt-hours

Table 1: Used variables from the dataset.

B Screenshots from the visualization

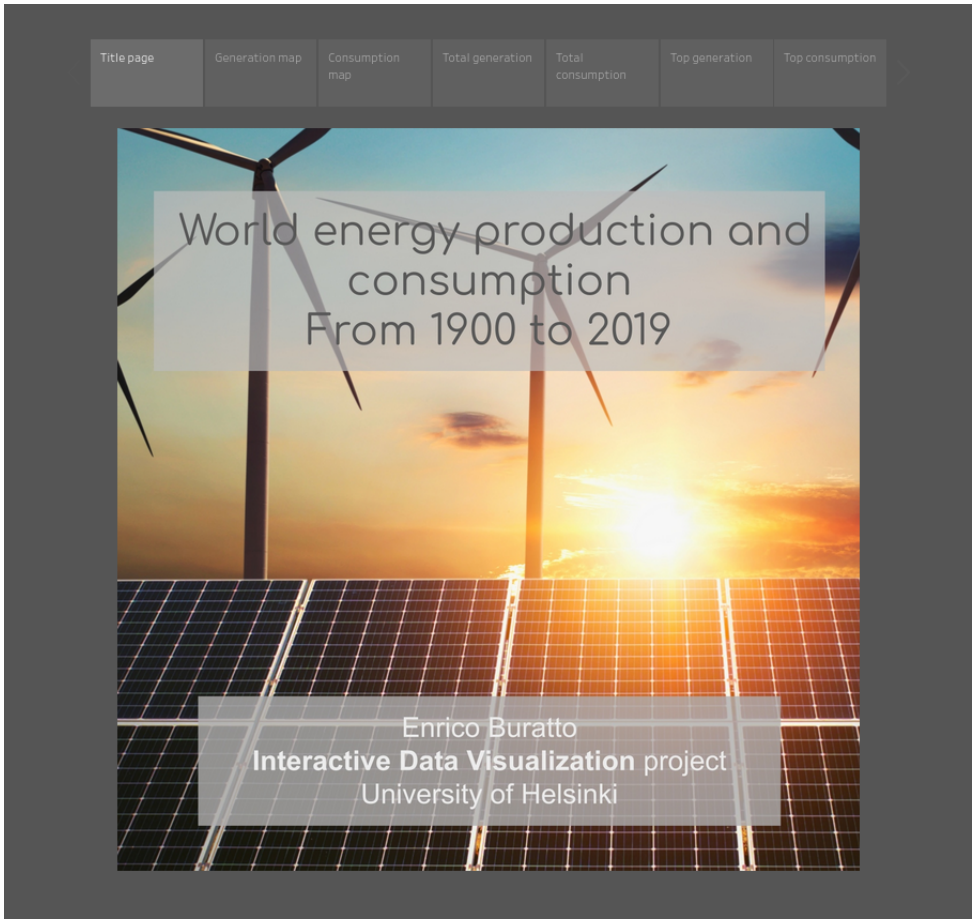

Figure 1: Title page of the visualization.

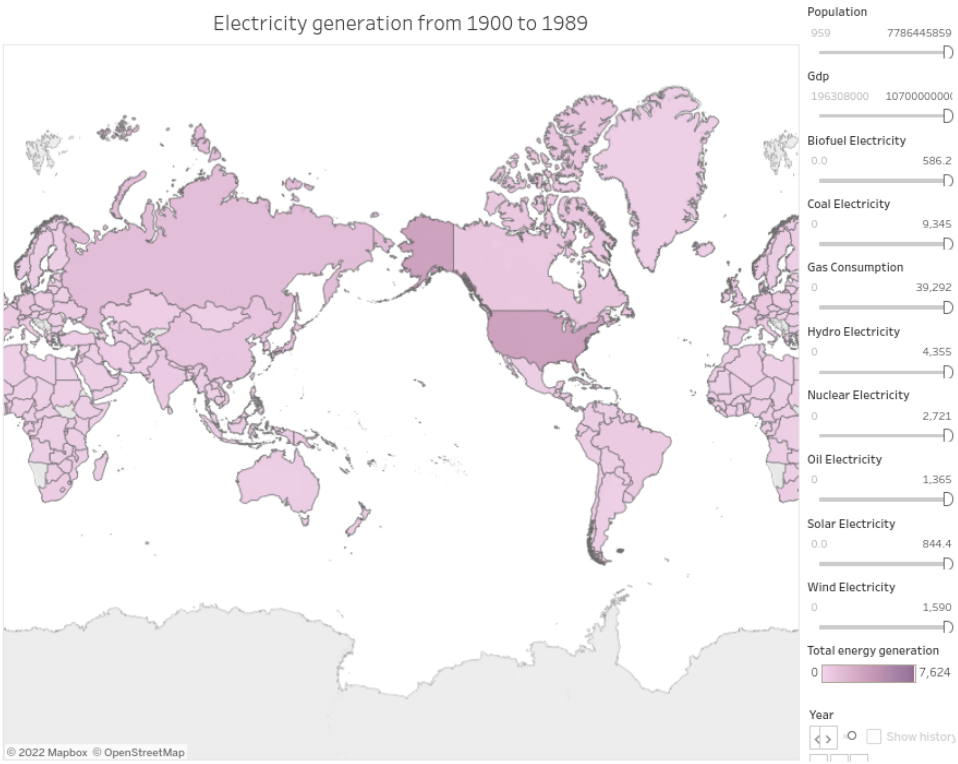

Figure 2: Interactive map visualization.

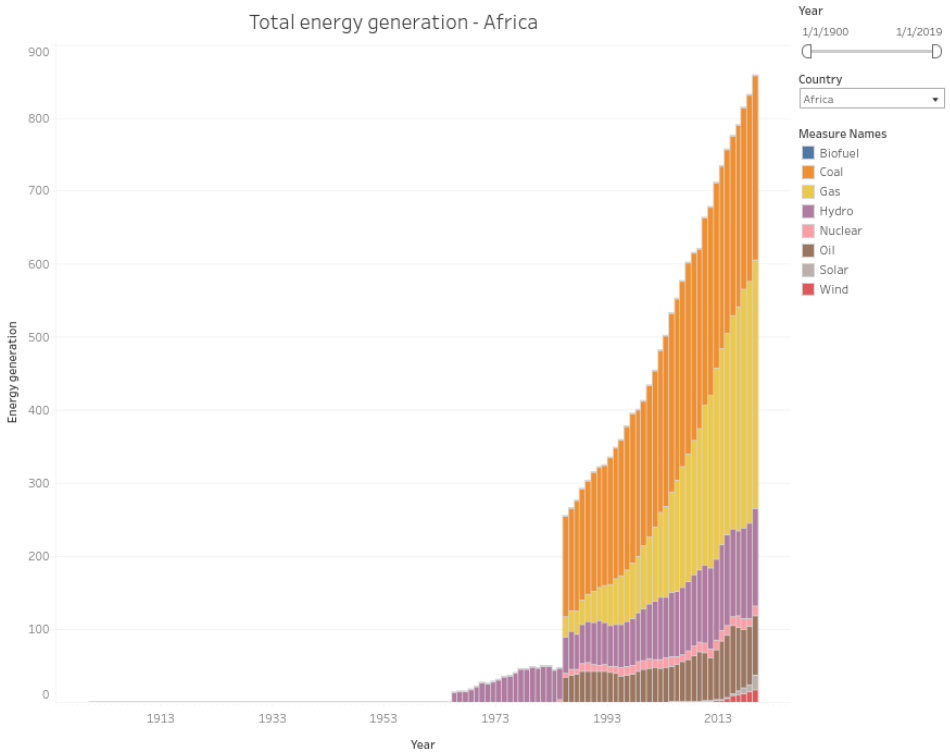

Figure 3: Interactive stacked bar visualization.

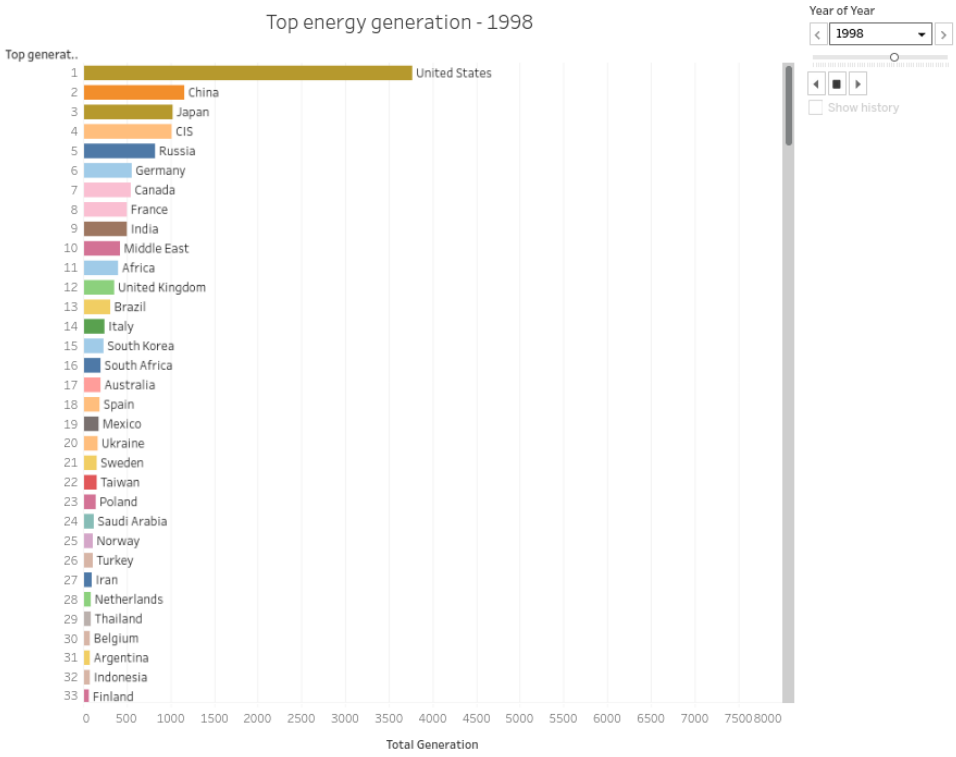

Figure 4: Bar race visualization.

References

[1] Pandas library for python, https://pandas.pydata.org/.

[2] Jupyter Notebooks, https://jupyter.org/.

[3] The Tableau visualization software, https://www.tableau.com/.

[4] Data on Energy by Our World in Data, retrieved from https://github.com/owid/energy-data.

[5] Our World in Data website, https://ourworldindata.org/.

[6] A Beginner's Guide to Build an Animated Tableau Bar Chart Race in 6 Minutes, retrieved from https://jnyh.medium.com/a-beginners-guide-to-build-an-animated-tableau-bar-chart-race-in-6-minutes-998b0087d30e.

[7] Final visualization, Enrico Buratto, https://public.tableau.com/app/profile/enrico2496/viz/IDVproject_16515785081680/Finalstory?publish=yes.

[8] Wurman, R. S., 2000, *Latch: The Five Ultimate Hatracks*, from *The information anxiety*, retrieved from https://www.informit.com/articles/article.aspx?p=130881.

[9] Tufte, E. R., 1942 (2001), *The Visual Display of Quantitative Information*, Cheshire.