Giada Gabriele - 235799

Università della Calabria - Master Degree in Artificial Intelligence and Computer Science

Data Analytics (Data Warehouse and Visualization) course - Essay

Introduction

The dataset we are going to analyze is focused on **global** climate change, in particular on the exposure to the climate in some economics, social and geographical areas.

Sources

The dataset was found via Kaggle and it comes from *The World Bank*'s Data Catalog.
Here you can find the Excel file that we are going to examine as our starting dataset.

Goal

We want to analyze climate change to see exclusively how it evolves, so the values that we care most are the **fields of application** and the **values through the years**.

Data Preparation

First we have to rename in the original dataset some cells that have dots - **..** - as NULL values in such a way that Tableau Prep Builder recognizes them correctly as NULL values.

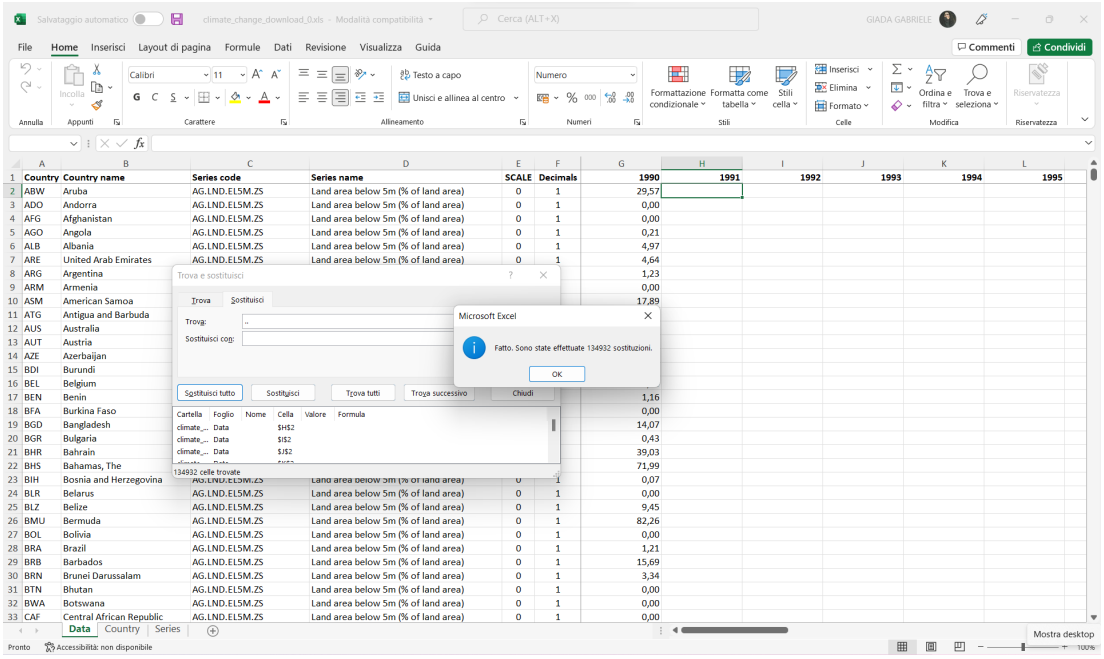

Figure 1.

We are going to analyze if there are some attributes that are not important based purely on their meaning.

We start from the **Series** sheet. The columns are:

- Series code
- Series name
- Scale
- Decimals
- Order
- Topic
- Definition
- Source

Scale, *Decimals* and *Order*, are values referred to the syntax and the place in the database, they are useless for the analysis that we are going to do, so we drop them.

Then we analyze the **Country** sheet. The columns are:

- Country code
- Country name
- Capital city
- Region
- Income group
- Lending category

In this sheet there is no cleaning to do.

Last, we analyze the **Data** sheet, the most important. The columns are:

- Country code
- Country name
- Series code
- Series name
- SCALE
- Decimals
- 1990
- 1991
- 1992
- 1993
- 1994
- 1995
- 1996
- 1997
- 1998
- 1999
- 2000
- 2001
- 2002
- 2003

- 2004
- 2005
- 2006
- 2007
- 2008
- 2009
- 2010
- 2011

Like in the previous reasoning, we drop *SCALE* and *Decimals* attributes because they don't bring any contribution to the analysis. Then we notice that to join with other sheets we just need the code of the *Country* and the code of the *Series*, so we drop the useless duplicates like the *name*s.

Data Selection and Cleaning with Tableau Prep Builder

At this point we can exploit Tableau Prep Builder (TPB). We upload the excel file on TPB and we are going to analyze the **Data** sheet (because the other two are mostly descriptive), in particular, the *Series name* column. By selecting all the NULL bars (as shown in the figure below) we can discover the total percentage of each attribute's NULL values.

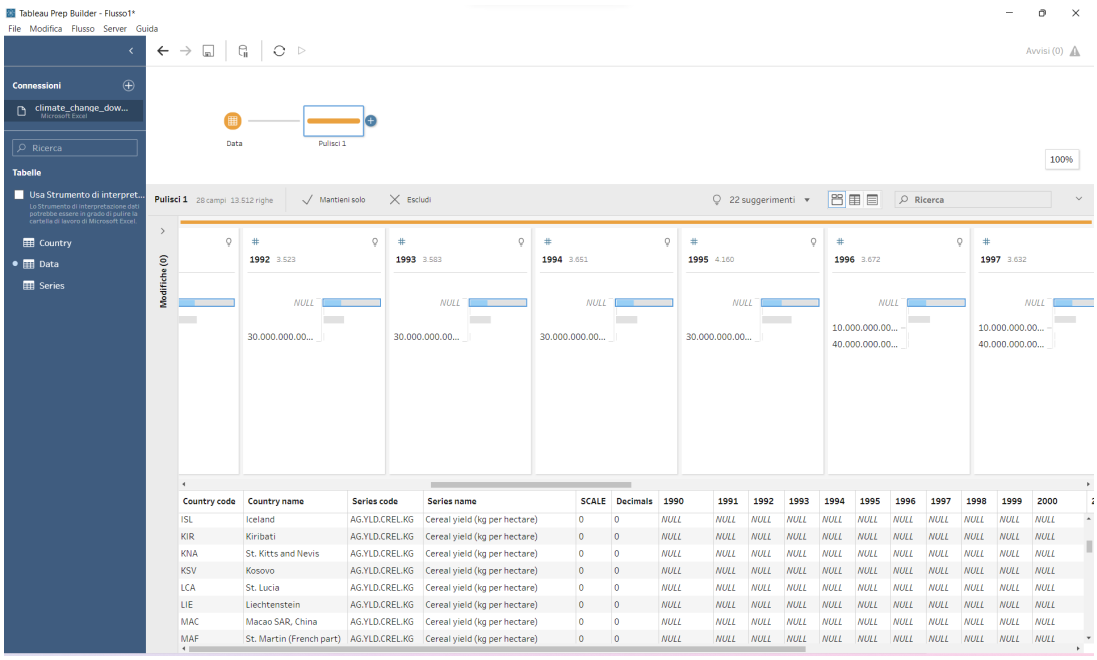

Figure 2.

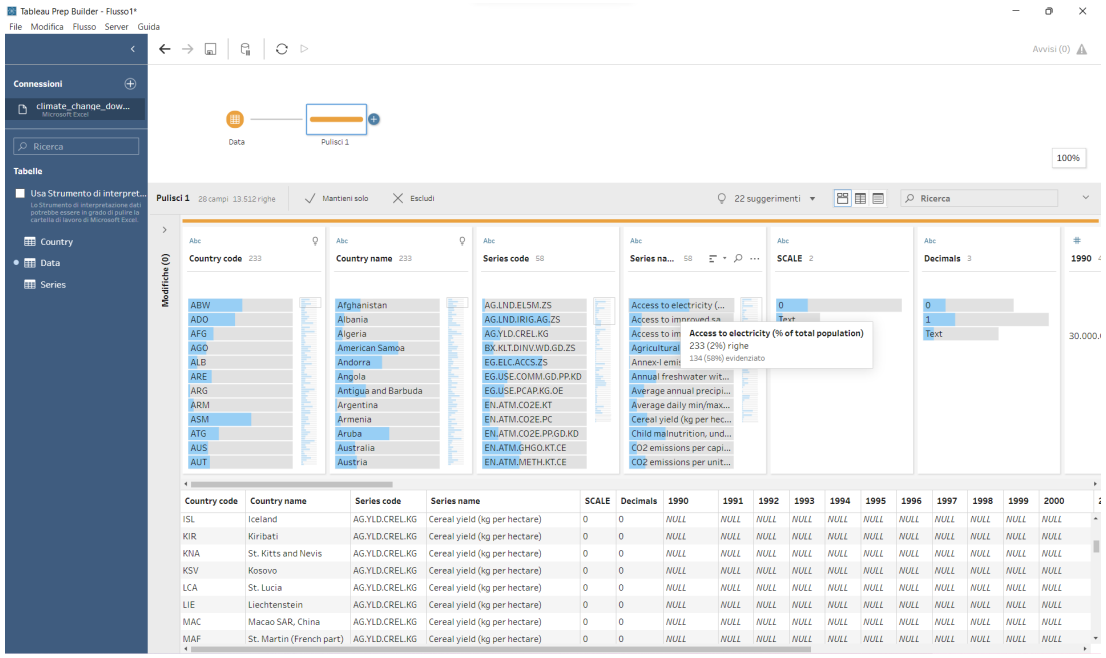

Figure 3.

Now we decide to drop all the attributes that have more than 20% of NULL values, so we are going to exclude from the future analysis:

- Access to electricity (% of total population)
- Agricultural land under irrigation (% of total ag. land)
- Annual freshwater withdrawals (% of internal resources)
- Child malnutrition, underweight (% of under age 5)
- Disaster risk reduction progress score (1-5 scale; 5=best)
- Droughts, floods, extreme temps (% pop. avg. 1990-2009)
- Ease of doing business (ranking 1-183; 1=best)
- Energy use per capita (kilograms of oil equivalent)
- Energy use per units of GDP (kg oil eq./$1,000 of 2005 PPP $)
- GHG net emissions/removals by LUCF (MtCO2e)
- Hosted Clean Development Mechanism (CDM) projects
- Invest. in energy w/ private participation ($)
- Invest. in telecoms w/ private participation ($)
- Invest. in transport w/ private participation ($)
- Invest. in water/sanit. w/ private participation ($)
- Issued Certified Emission Reductions (CERs) from CDM (thousands)

- Malaria incidence rate (per 100,000 people)
- Methane (CH4) emissions, total (KtCO2e)
- NAMA submission
- Nitrous oxide (N2O) emissions, total (KtCO2e)
- Other GHG emissions, total (KtCO2e)
- Population in urban agglomerations >1million (%)
- Population living below $1.25 a day (% of total)
- Public sector mgmt & institutions avg. (1-6 scale; 6=best)
- Renewable energy target

After making this choice now we made choices less technical related to the initial Goal.

We have some Series that are average of other years, so to not make confusion we decide to drop these rows:

- Average annual precipitation (1961-1990, mm)
- Average daily min/max temperature (1961-1990, Celsius)
- Projected annual precipitation change (2045-2065, mm)
- Projected annual temperature change (2045-2065, Celsius)

Then we have other *Series* that are textual values, not useful in a column when we expect numbers to make a graph. We drop also these rows:

- Annex-I emissions reduction target
- Hosted Joint Implementation (JI) projects
- Issued Emission Reduction Units (ERUs) from JI (thousands)
- Latest UNFCCC national communication
- NAPA submission
- Projected change in annual cool days/cold nights
- Projected change in annual hot days/warm nights

At this this point we only have these *Series*:

- Access to improved sanitation (% of total pop.)
- Access to improved water source (% of total pop.)
- Cereal yield (kg per hectare)
- CO2 emissions per capita (metric tons)
- CO2 emissions per units of GDP (kg/$1,000 of 2005 PPP $)

- CO2 emissions, total (KtCO2)
- Foreign direct investment, net inflows (% of GDP) *
- GDP ($)
- GNI per capita (Atlas $)
- Land area below 5m (% of land area) *
- Nationally terrestrial protected areas (% of total land area)
- Nurses and midwives (per 1,000 people) *
- Paved roads (% of total roads) *
- Physicians (per 1,000 people) *
- Population
- Population below 5m (% of total) *
- Population growth (annual %)
- Primary completion rate, total (% of relevant age group) *
- Ratio of girls to boys in primary & secondary school (%) *
- Under-five mortality rate (per 1,000)
- Urban population
- Urban population growth (annual %)

The rows denoted by * will be dropped for personal choice. These *Series* are near others already present or far away from the ideal analysis that we are going to do. So our final dataset will have these *Series*:

- Access to improved sanitation (% of total pop.)
- Access to improved water source (% of total pop.)
- Cereal yield (kg per hectare)
- CO2 emissions per capita (metric tons)
- CO2 emissions per units of GDP (kg/$1,000 of 2005 PPP $)
- CO2 emissions, total (KtCO2)
- GDP ($)
- GNI per capita (Atlas $)
- Nationally terrestrial protected areas (% of total land area)
- Population
- Population growth (annual %)
- Under-five mortality rate (per 1,000)
- Urban population

- Urban population growth (annual %)

Before starting with the Design phase we are going to analyze the new, smaller, dataset on TPB. We notice that with this choices the last two columns of the years' values are more than 50% NULL:

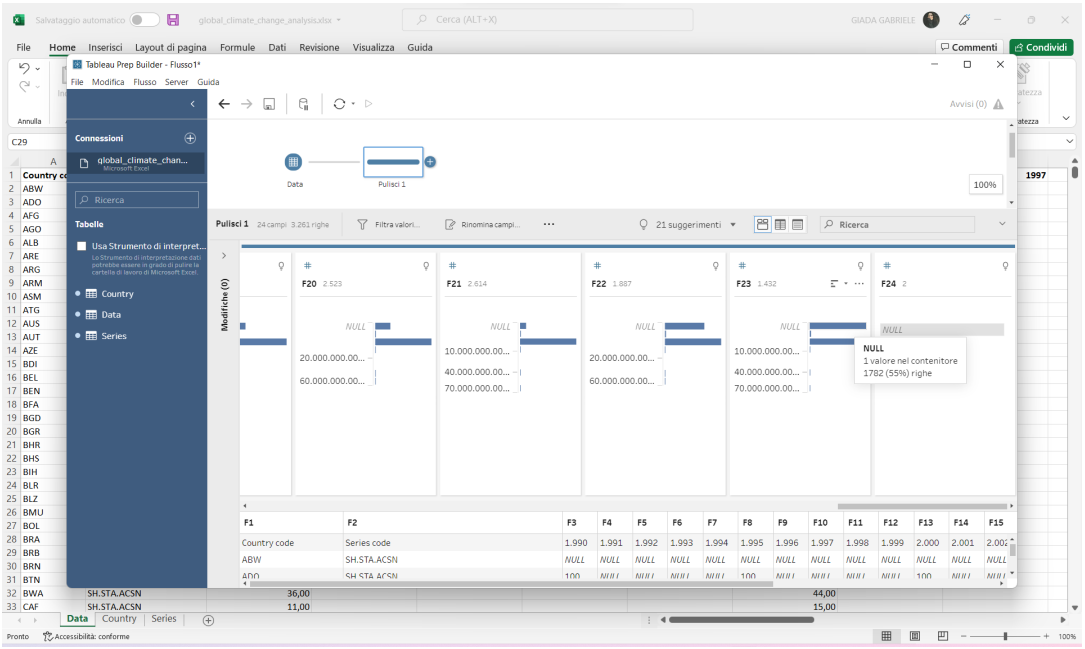

Figure 4.

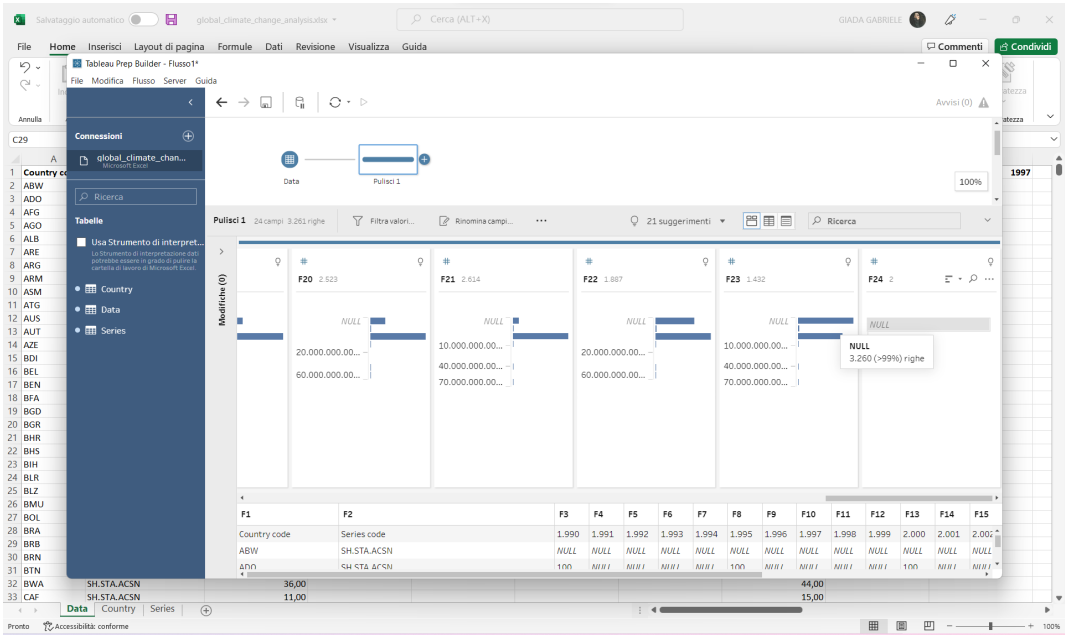

Figure 5.

So we drop them. Lastly, we have to modify the columns' values. We need the **year**, that now is the name of each column, as an **attribute** *(date type)* to improve in the modeling phase the graphical part, so the best way to reach our goal is to pivot the *Data* table.

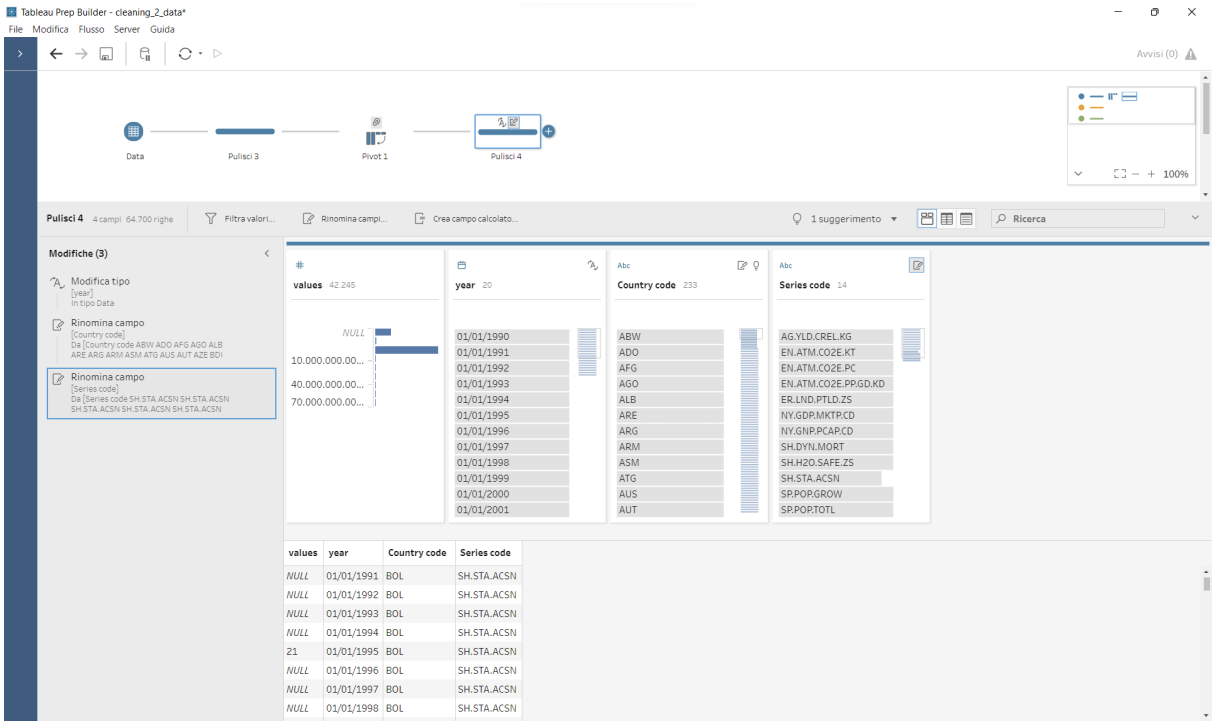

Figure 6.

Design

Relational schema

Data(Country code*, Series code*, year, values)

Country(Country code, Country name, Capital city, Region, Income group, Lending category)

Series(Series name, Topic, Definition, Source)

Attributes Tree

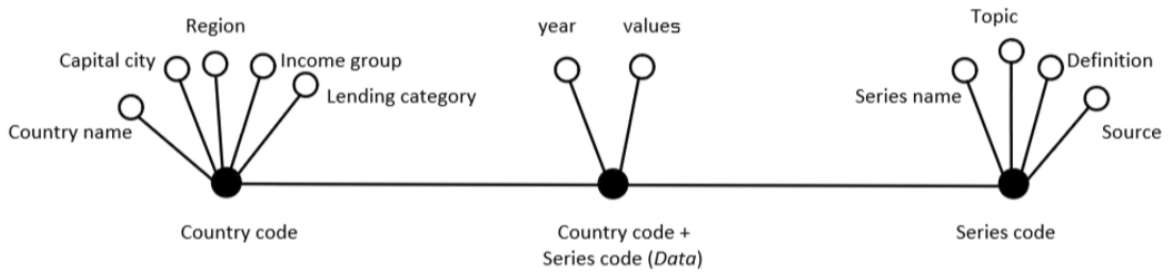

Figure 7.

Editing Attributes Tree

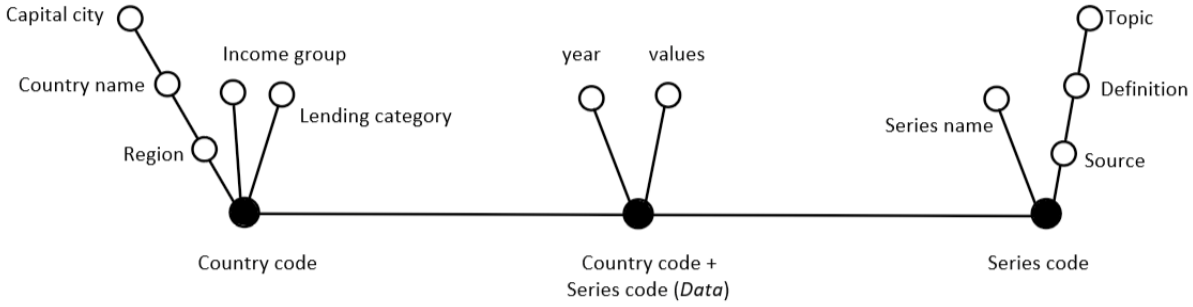

Figure 8.

Fact Schema

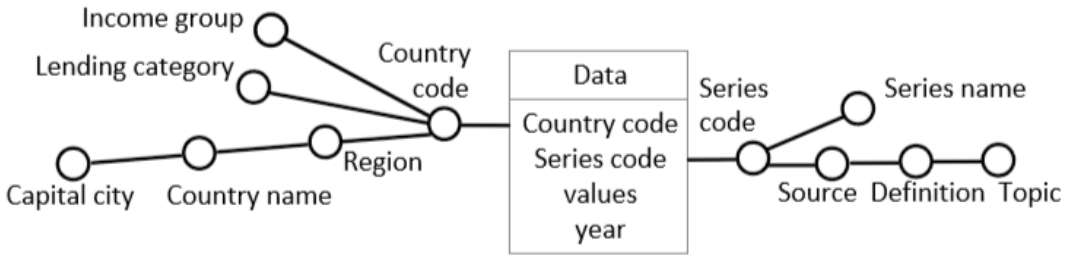

Figure 9.

Star Schema

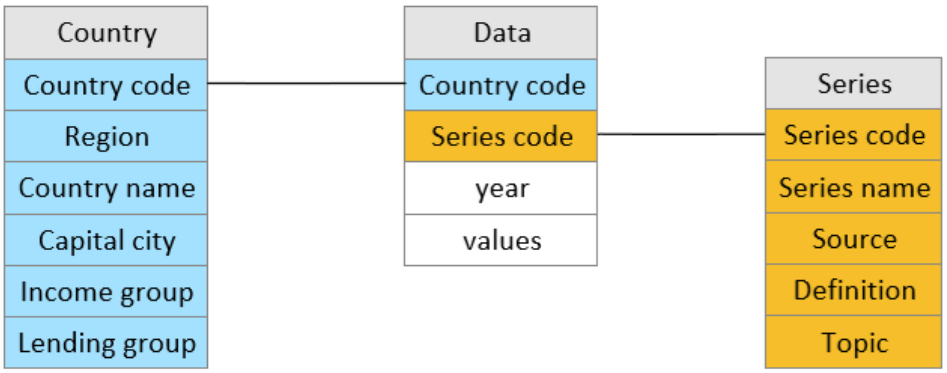

Figure 10.

Analysis sheets and Dashboards with Tableau

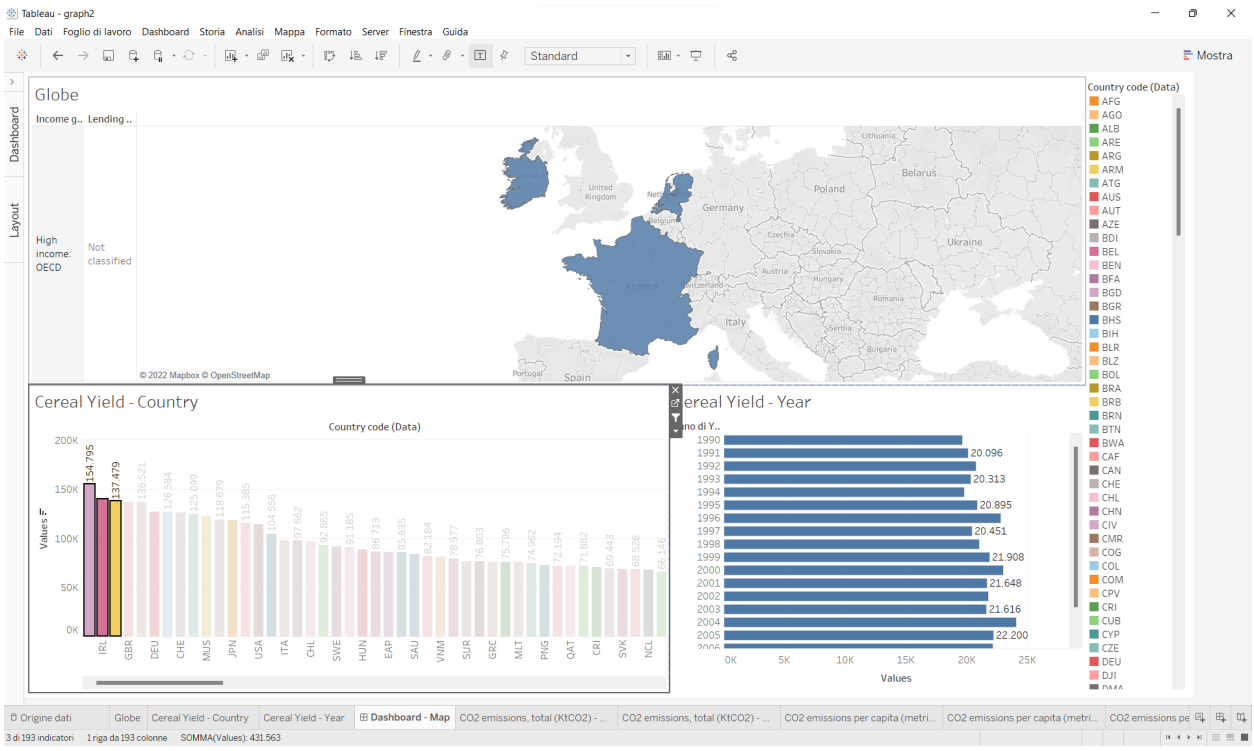

Figure 11.

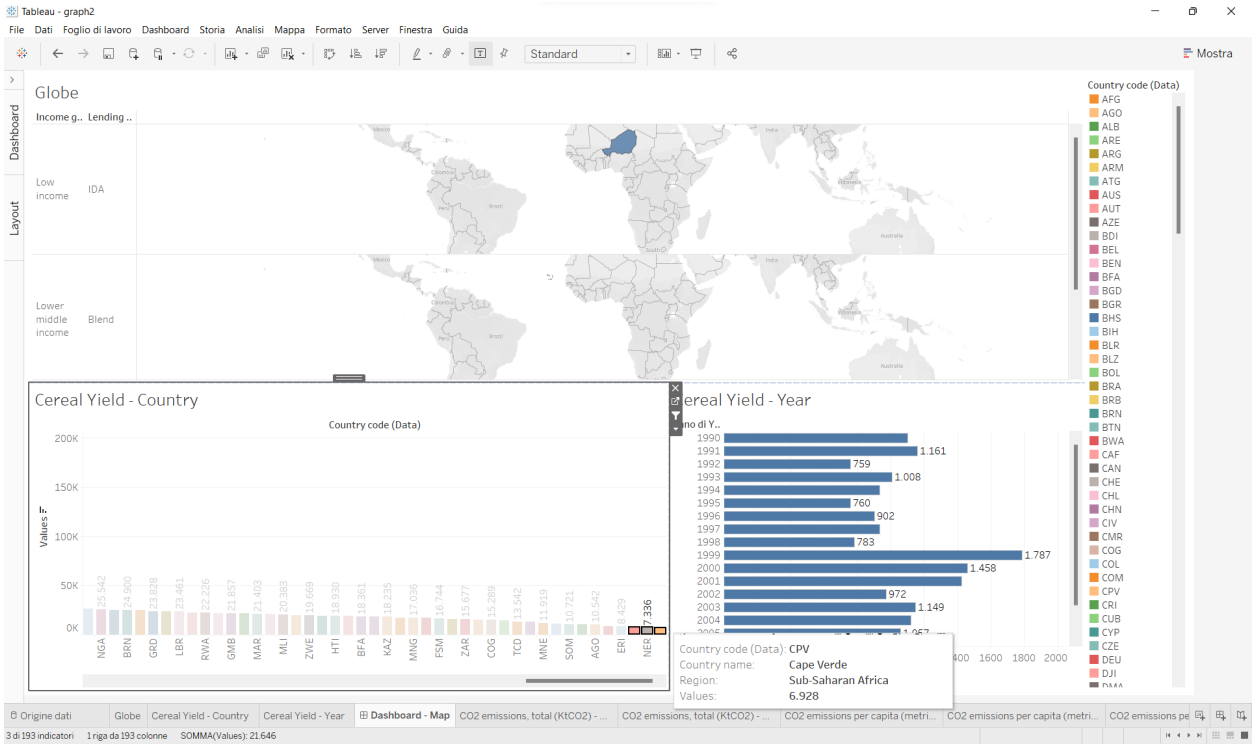

Figure 12.

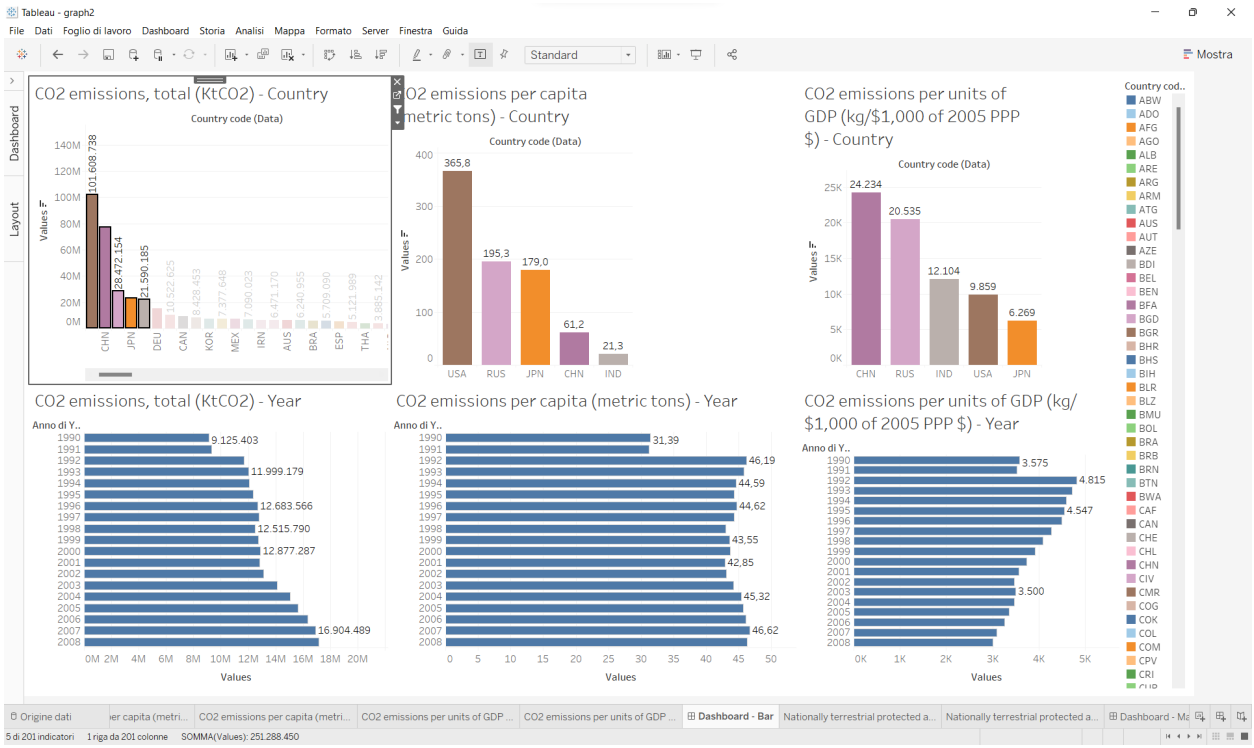

Figure 13.

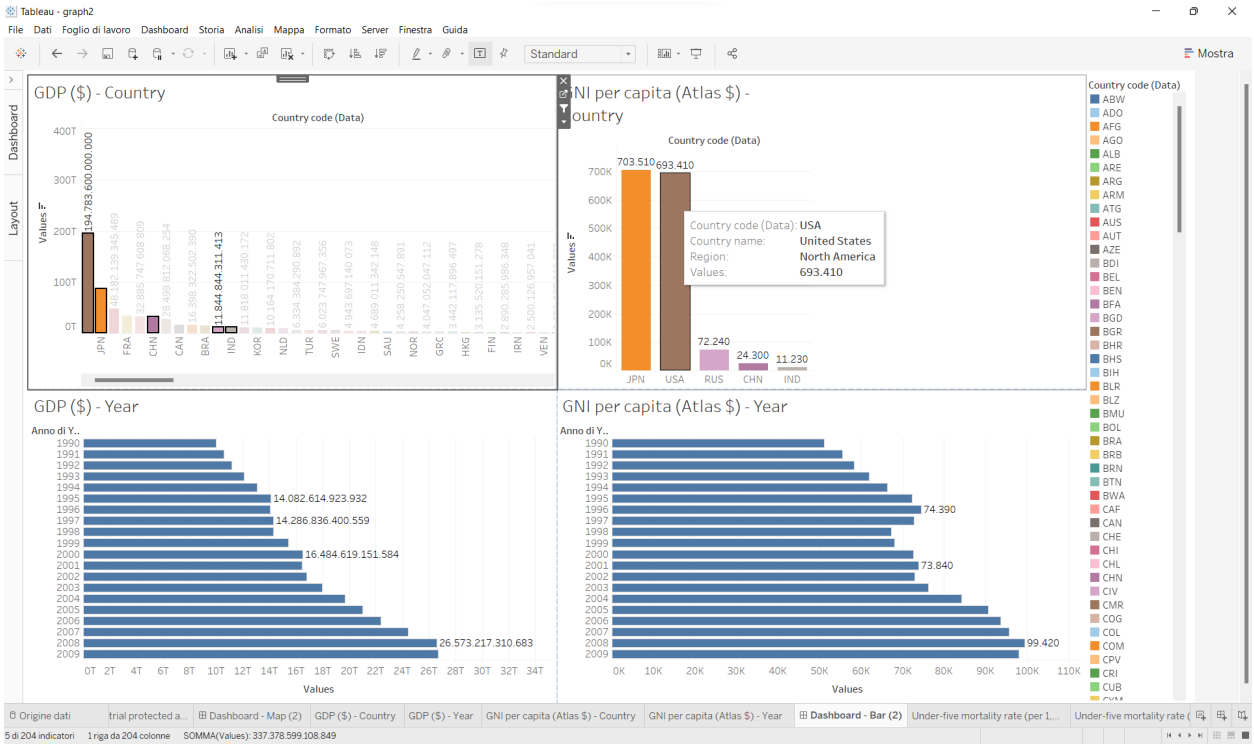

Figure 14.

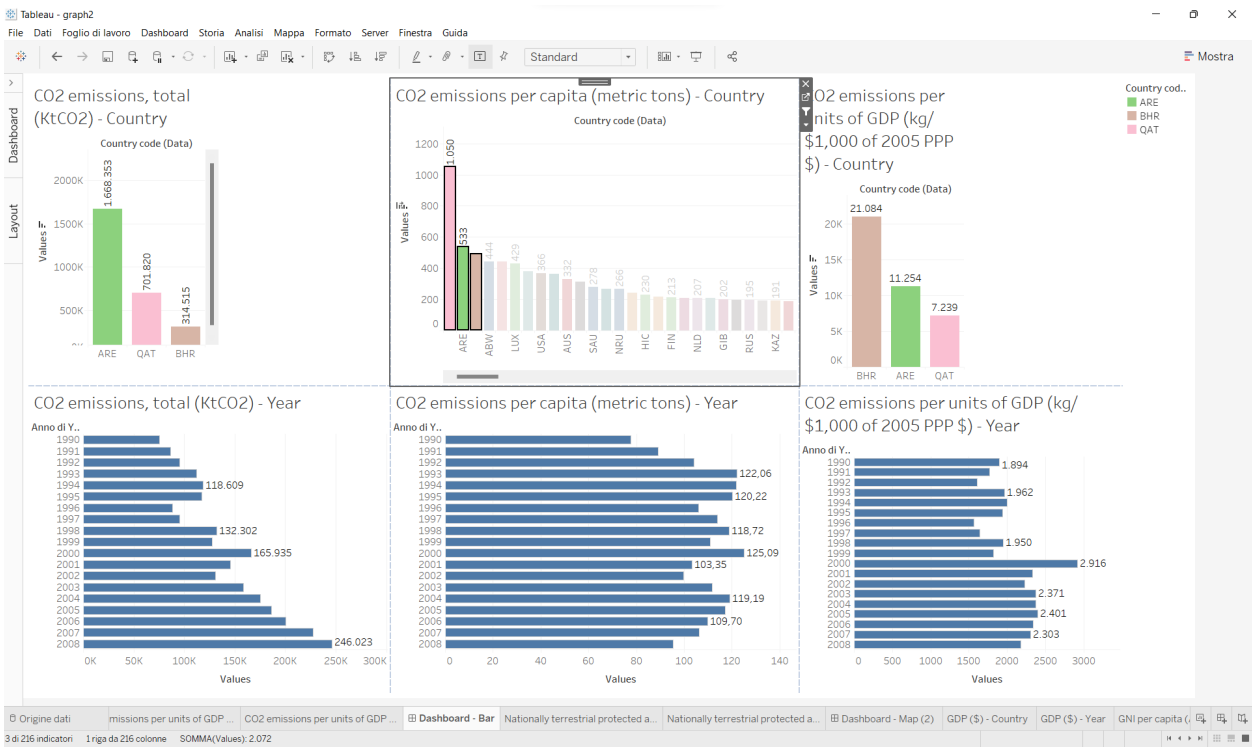

Figure 15.

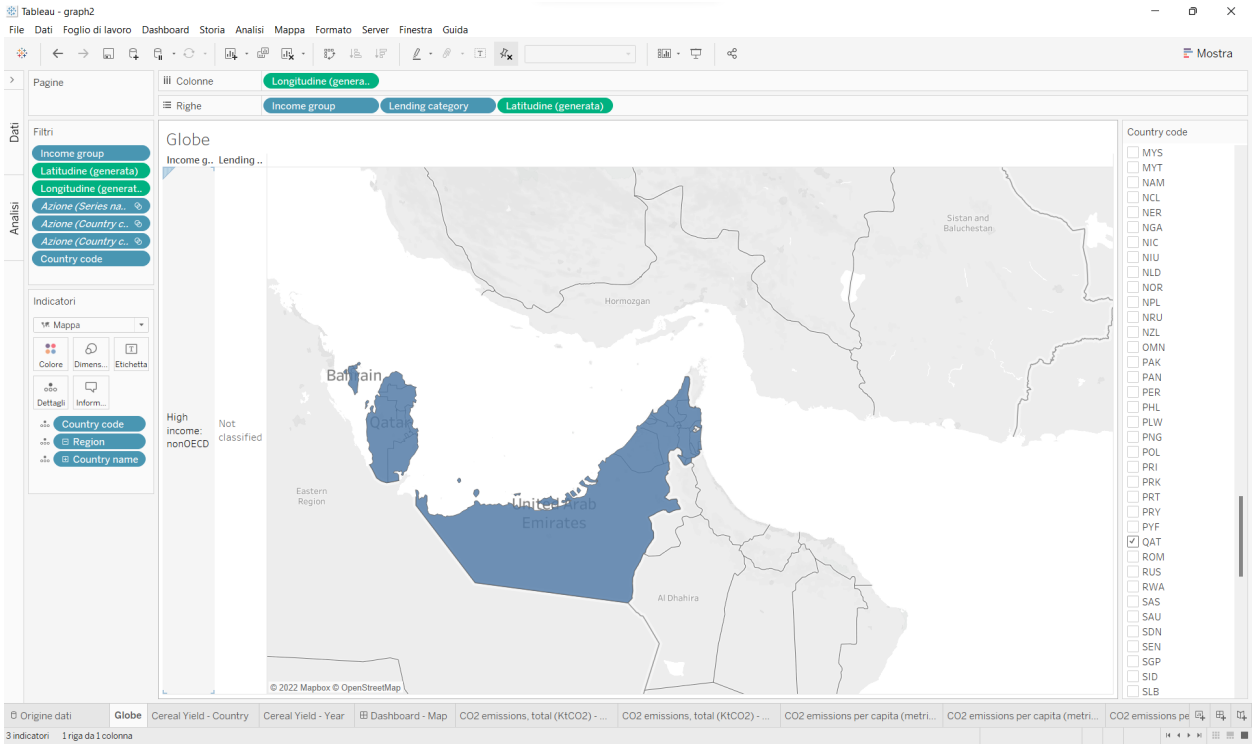

Figure 16.

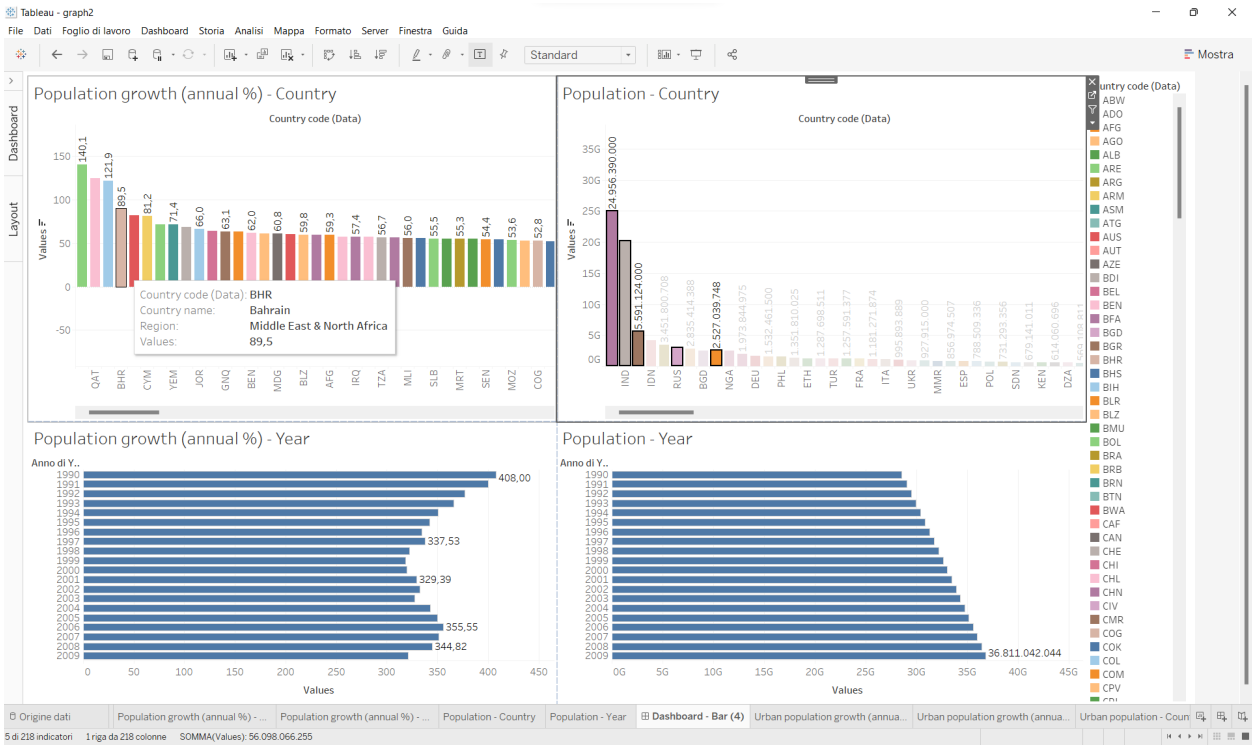

Figure 17.

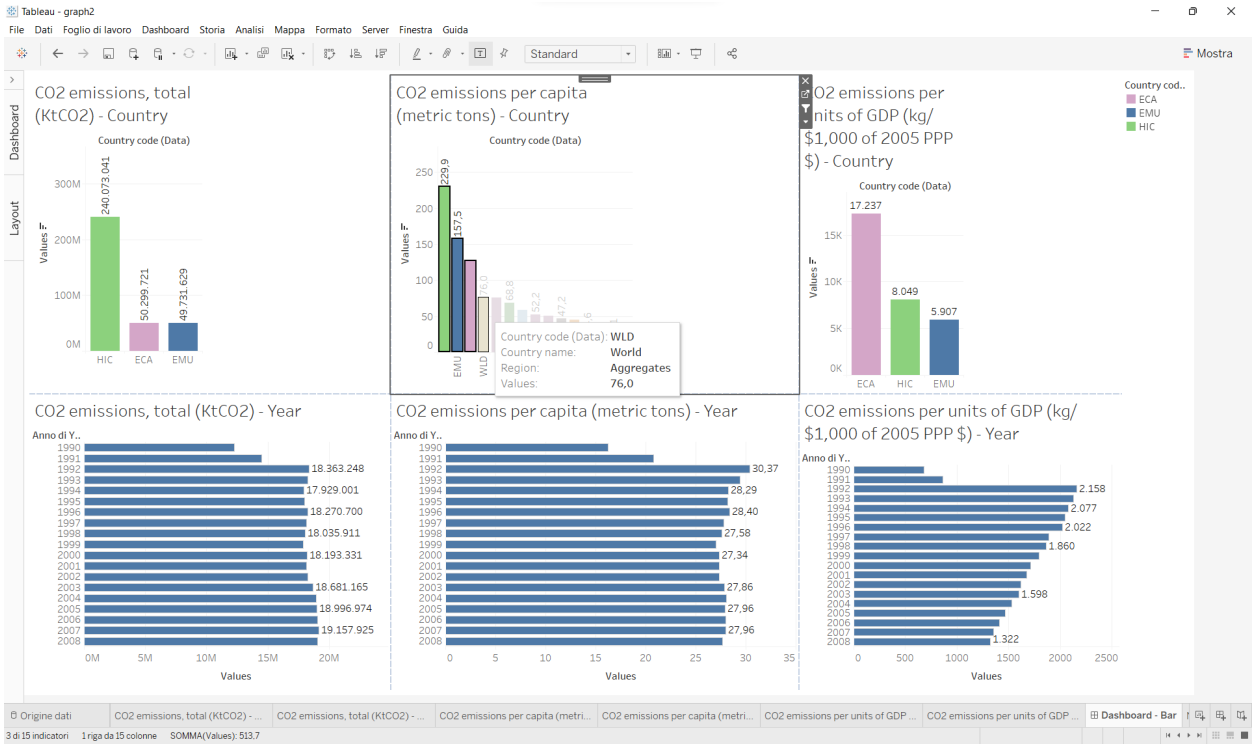

Figure 18.

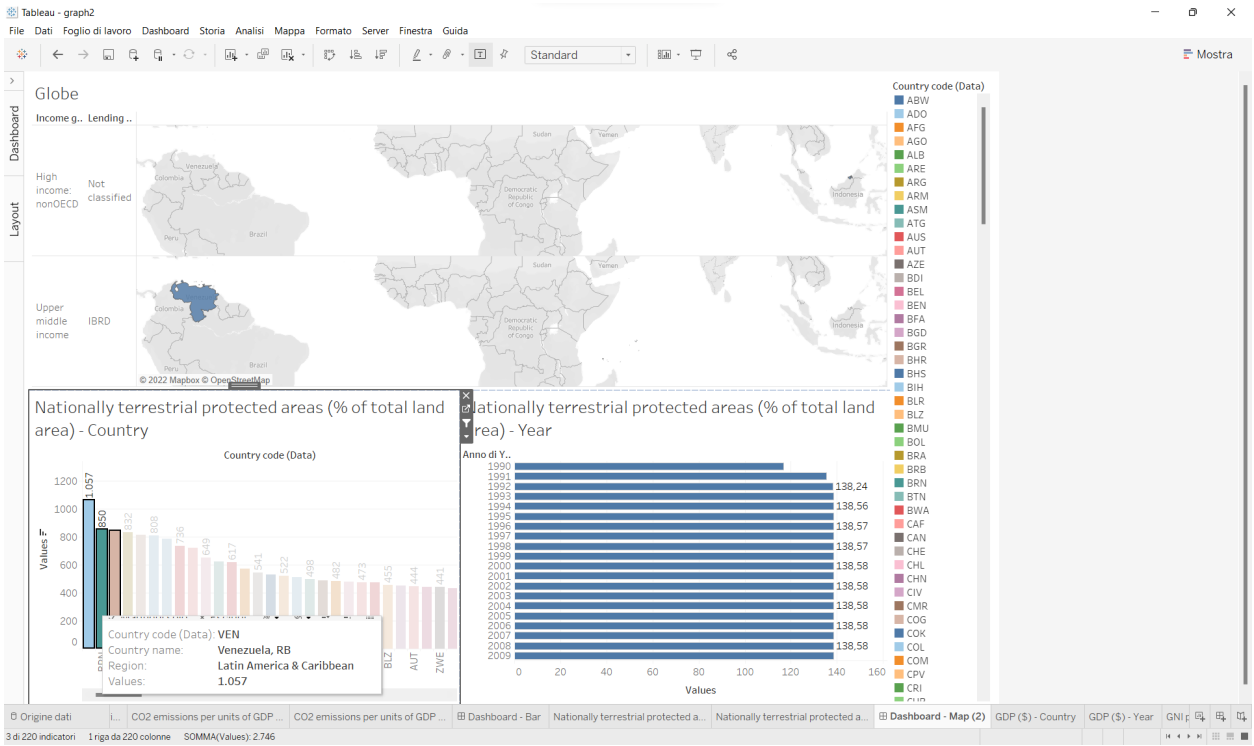

Figure 19.

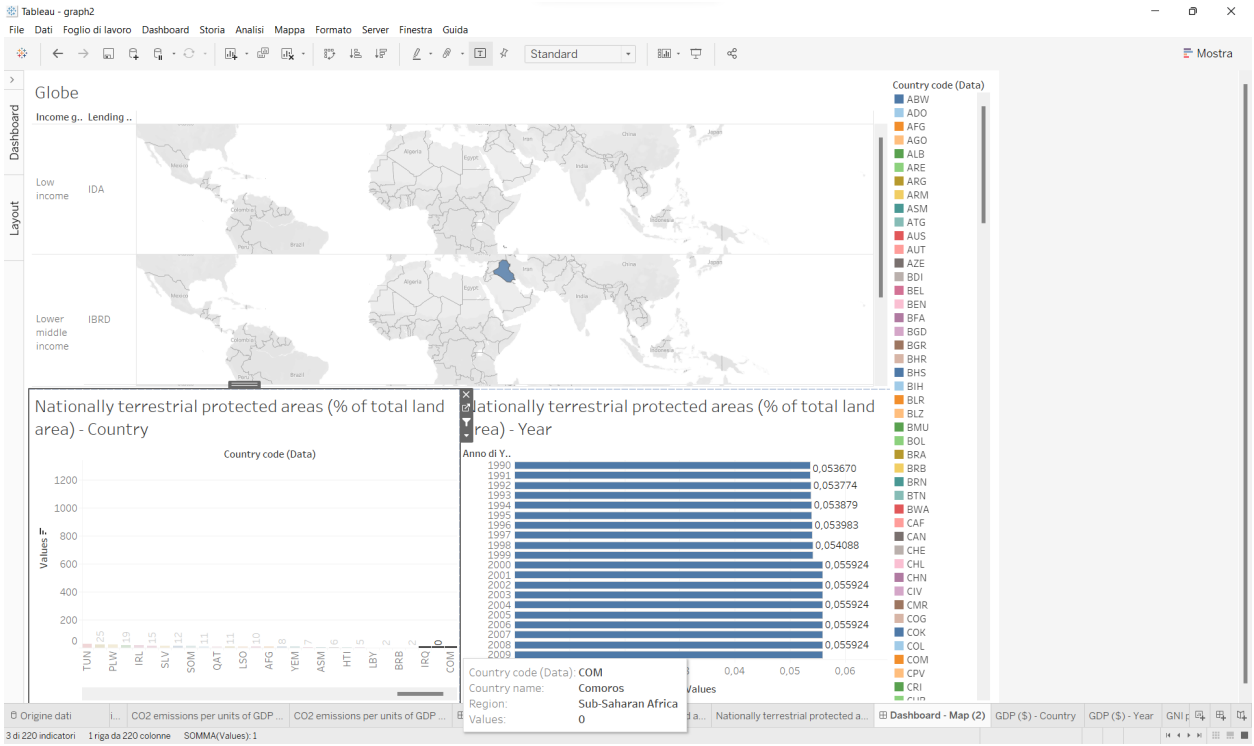

Figure 20.

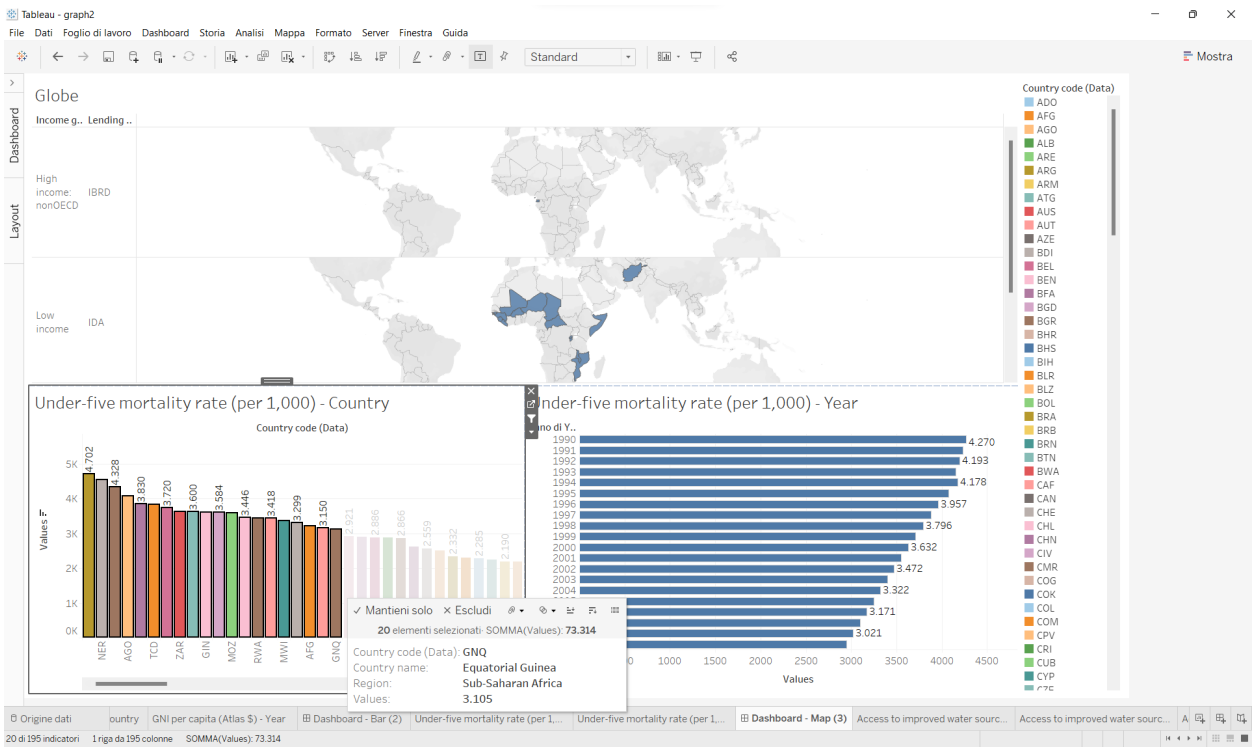

Figure 21.

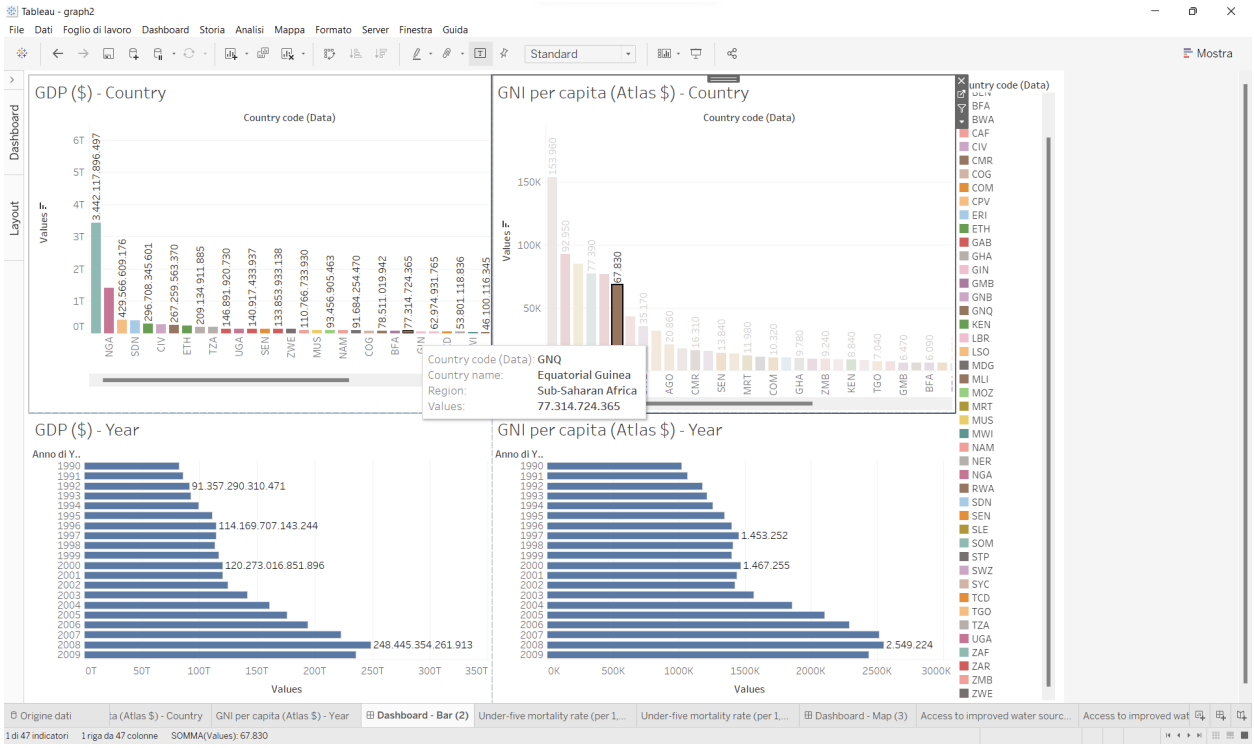

Figure 22.

Conclusion

Notes:

1. *here you can find an article to understand how World Bank classifies countries*
2. *legend to understand how we are going to describe the final analysis:*
    - ❖ *Series name*
        - ➢ *Topic - according to our dataset (Series sheet)*
            - ■ *observation*

- ❖ Cereal Yield
    - ➢ Resilience
        - ■ Looking at Figure 11 we can notice that the top 3 for cereal yield are all Countries classified as *High income* and all in the European area (one of the strongest economies). Instead, looking at Figure 12, the last 3 are at most in the *Lower middle* classification. So we can conclude that there is a strict correlation between production and income, it's easier for a Country with a higher income to overcome difficulties in the agricultural system.

- CO2
  - GHG emissions and energy use
    - The top 5 for CO2 emissions (Figure 13) are the Countries with the highest *GDP*, as we can see in Figure 14. Also for *CO2 emissions per capita* (Figure 15) the top 3 are all Countries classified as *High income* (Figure 16). So, also here there is a correlation between income, GDP, basically a Country's economy, but also with the number of *population* and *growth population*, visible in Figure 17. Moreover, we can notice from Figure 18 that *HIC* (high income), *ECA* (Europe & Central Asia) and *EMU* (Euro area) have higher *CO2 emissions per capita* than the *WLD* parameter that refers to *WORLD*, so European and Asian Countries with *High income* emit more than global average.

- Nationally terrestrial protected areas (% of total land area)
  - Exposure to impacts
    - Looking at Figure 19 we can notice that the top 3 Countries with protected areas are all classified **at least** as *Upper middle income*, while for the last 3, as we can see in Figure 20, all the Countries are **at most** *Lower middle income*. Once again, of course, there is a strict correlation between economy and the ability to respond to Country needs.

- Under-five mortality rate (per 1,000)
  - Exposure to impacts
    - In Figure 21 we highlighted the "top" Countries for *under-five mortality rate*, and we noticed that to find a Country with an *High Income* we have to reach position 20, *Equatorial Guinea (GNQ)*, which is part of a Region not so rich, as we can see in Figure 22; only in the *Sub-Saharan Africa* area *GNQ* is not even in top 3 for *GDP per capita* and far away the top 20 in the *GDP* ranking.

We can deduce that, for any *topic* related to climate change, the richest Countries (from an economic point of view) are the ones that have more impact, like in the CO2 graphs, and at the same time they have the possibility and the economic capacity to face up climate impacts. The poorest countries are the ones that suffer more and have less operating space.