Assignment 1 - Introduction to Data Science & AI

Nathan HAUDOT, Hugo MATH

November 9, 2021

1 GDP per capita vs. life expectancy

1.a Python program for scatter plot of GDP per capita vs. life expectancy, assumptions and decisions while selecting & combining data

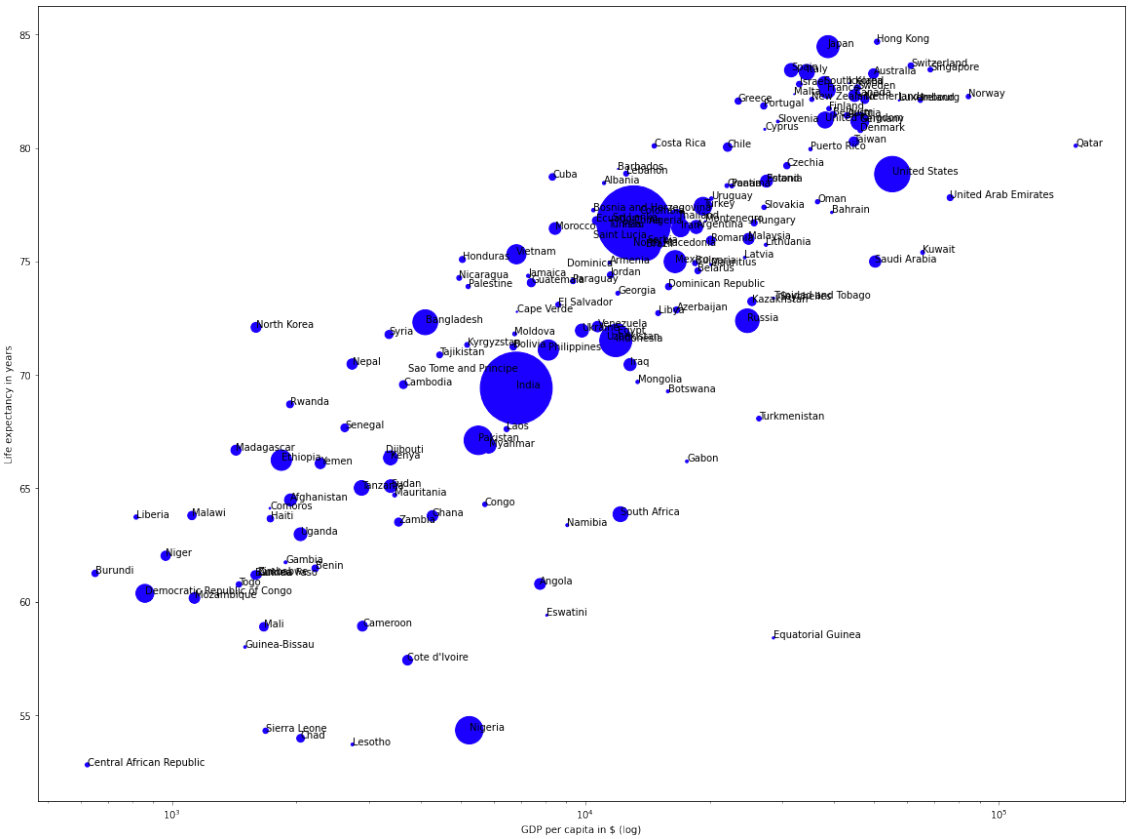

Figure 1: Life expectancy vs. GDP per capita, year 2018

The dataset was retrieved from the Our World In Data website. At first, it was filtered to remove entries that did not contain life expectancy or GDP per capita data. This filtering divided the dataset size by five (from about 60,000 entries to 12,500).

The scatter plot above shows the life expectancy compared to the GDP per capita of 167 countries for the year 2018. The population size is also present in the graph, based on the size of the marker used for each country. For simplicity, the term "life expectancy" will be referred to as "LE" in the following sections.

1.b Relevance & explanation of results

The results seem to be consistent. The more economic activity a country has (reduced to a per capita), the higher LE it provides. However, GDP per capita does not capture income and wealth inequalities within a country's population; one would have to go "in-depth" by correlating other data sources if one wanted to find other factors that improve LE (such as standard of living, medical coverage or lifestyle choices).

1.c Data cleaning for the following questions, which entries have been removed

The entries for the continents and the world have been removed. In the graph, only the entries for the countries will be shown. If we filter the entries for a specific year, we get about 150 entries for as many different countries. For the following questions, we will choose to use the 2018 data, because it is recent and the dataset is substantial.

1.d Which countries have a life expectancy higher than one standard deviation above the mean ?

For 2018, the standard deviation of LE is about 7.7 years, and the mean is 72.6 years. We can see that there is a lot of European countries, 4 in Asia, 2 in Oceania, and 1 in Africa.

List of countries that have a LE higher than one standard deviation above the mean:

Australia	Austria
Belgium	Canada
Cyprus	Denmark
Finland	France
Germany	Greece
Hong Kong	Iceland
Ireland	Israel
Italy	Japan
Luxembourg	Malta
Netherlands	New Zealand
Norway	Portugal
Singapore	Slovenia
South Korea	Spain
Sweden	Switzerland
United Kingdom	

1.e Which countries have high life expectancy but have low GDP?

In our Python code, we merged GDP dataset to LE vs GDP per capita. We can now compare these data.

We can firstly slice our data into countries that have GDP under the mean of total GDP, and LE above the mean of total LE. We get 82 countries including (e.g. Albania with 78 years of LE and $ 15 billion of GDP).

We can reduce the number of countries in this group by applying a filter saying: $LE > mean + one\ standard\ deviation$, and $GDP < mean - deviation/8$. Thus, we get 5 countries: Cyprus, Luxembourg, Malta, Iceland and Slovenia. The most surprising result is Slovenia with a relatively poor GDP but with 81 years of LE.

1.f Does every strong economy (normally indicated by GDP) have high life expectancy?

We can translate strong economy by filtering our data with $GDP > mean + one\ standard\ deviation/8$, and $LE < mean - one\ standard\ deviation/8$. Therefore, we get India with 69 years old of LE and a GDP of $ 2 800 billion.

However, GDP like that corresponds to Europe countries like France for example which have a much better LE (+10 years), which means that GDP is not directly correlated to a better LE.

1.g Related to question f, what would happen if you use GDP per capita as an indicator of strong economy? Results explanation & insights discussion comparing previous questions

We can redo the experiment above with GDP per capita. When applying the same filter we got two countries : Equatorial Guinea and Turkmenistan. The GDP per capita is mind opening because we can directly see that Equatorial Guinea has a LE of 58 years, and a relatively high GDP per capita compared to others.

It is obvious that in high GDP countries, there is a huge disparity in the population and the GDP alone is not enough to understand LE. That is why GDP per capita is more suitable for our analysis. The challenge here is to not confuse causality and correlation.

After removing "NaN" entries and merging our dataset, we get this graphs using a logarithmic scale on the x axis.

2 Other datasets analysis related to happiness, life satisfaction, trust or corruption

2.a Analysis & answered questions with these new datasets, informative visualisations

After downloading the *happiness-cantril-ladder.csv* file from Our World In Data (data indicating life satisfaction for several countries) and *extreme-poverty-headcount-ratio-vs-life-expectancy-at-birth.csv*, we can ask ourselve those questions :

Does happiness is directly correlated to GDP or/and GDP per capita ? Is it also correlated to LE ? If we take Poverty dataset, does it is correlated to GDP per Capita and LE ?

After merging our happiness data, GDP data and removing the "NaN" entries from the resulting data, we can start drawing our figures.

Figure 2: Evolution of the happiness indicator according to GDP and GDP per capita

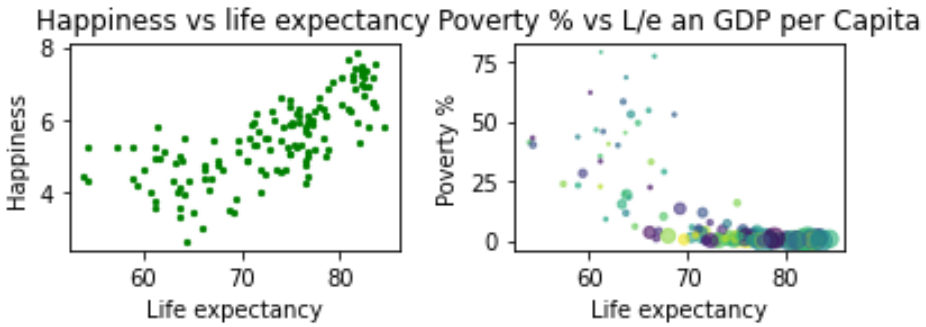

Figure 3: Evolution of the poverty in % according to LE and GDP per capita

2.b Discuss any observations that you make, or insights obtained, from the data visualisations.

According to the figure 2 in 2.a, we can clearly state that GDP per capita seems more relevant when talking about happiness in a country.

In fact, if we draw a best fit line, we can observe that the GDP per capita graph has less "dispersion" and seems to correspond to a more "straight" line than the GDP graphic. **We can now state surely that in most cases, GDP per capita has a positive impact on citizen's happiness**.

Another interesting graph in 2.a is figure 3, showing that LE is not directly correlated to happiness. In fact, we can find countries where the happiness indicator is near the mean but the LE is one standard deviation above the mean (e.g. with Albania which has a score of 5 for life satisfaction or happiness, and 78 of LE).

The graph on the right is also showing GDP per capita represented by the size of the markers. LE in years and poverty in % are represented respectively on the X and Y axis. GDP per capita seems highly correlated to life expectancy and poverty. Indeed, there might be an exponential curve between poverty and LE, and also for the GDP per capita and poverty.