**DEVELOPED vs DEVELOPING COUNTRIES**

NOTEBOOK: Developed\_vs\_developing.ipynb

In this section, we’ll be exploring the suicide rates between developed and developing nations. The first challenge is to determine what criteria will be used to define a developed vs developing country.

A developed country is a country that has an advanced, mature economy and infrastructure. This is led to believe to a higher standard of living and prosperity for it’s citizens. On the other hand, a developing country is one that has an under-developed economic system and low to mediocre infrastructure, thus providing a lower quality of life to it’s citizens.

There are several metrics to measure how developed a country is: income inequality, GDP per capita, average income. Since we have one of the metrics available to us in the dataset (GDP per capita), we will be using this to classify developed vs developing countries.

There is no concrete rule about a cut-off level for GDP per capita in terms of determining how developed a country is, but a rule of thumb is anything above $20,000 of GDP per capita is considered a developed country. Of course, there are several other factors that go into determining the eligibility, but this is a quick way that we’re going to use for our analysis.

## **ADDING TYPE ATTRIBUTE**

After loading the dataset into a dataframe, we will add a new column called ‘Type’. This attribute tells us whether the country specified in the column is developed or developing. The following line of code was used for adding the column:

This line works almost like an if/else statement where anything above with a GDP per capita above 20,000 is assigned Developed and Developing otherwise.

The first task is to see the suicide rates between the developed and developing countries over the years. In order to do this from the main dataframe, we needed to group the dataframe by year and by country because the data is split up into multiple rows (by age). After the grouping, we can make a dataframe based on the mean of the 'suicides/100k pop' column values. The following code achieves this:

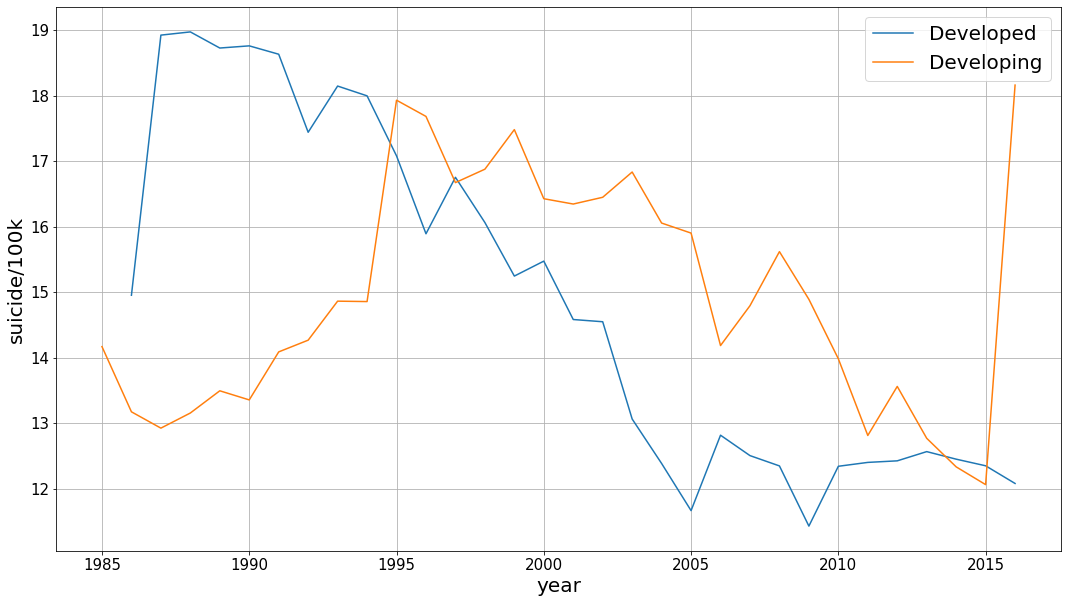
After the grouping, we can plot our dataframe with matplotlib (Figure 1). There are three interesting things to note here. First, right off the bat, the suicide rate for developed countries is way higher than developing countries in the first few years. It is almost twice as much as the developing countries. Secondly, as the time moves forward, the suicide rate for developing countries increases while the suicide rate for developed countries decreases, intercepting at around 1994. Lastly, there’s a sudden spike in 2005 (and a subsequent downturn) followed by a trend where the rates start increasing after about 2009 for developed countries. Additionally, there’s a sharp increase for developing countries after 2015.

Figure - Developed vs Developing nations suicide rates (per 100k)

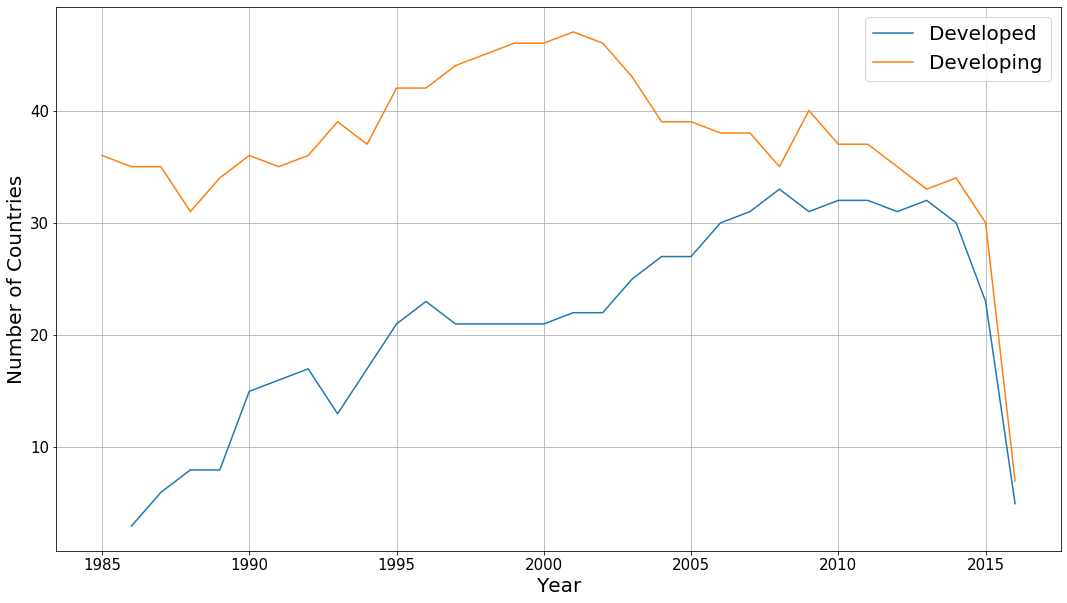
To address the first note, we decided to gather the number of countries for each year that fall into each category. To do this, we grouped the dataframe by year, country and Type and get a count for the number of countries in each type by year. A plot of the count is shown in figure 2.

Figure - Count of countries in developed vs developing group per year

As you can see, the number of countries in developed countries is significantly lower compared to developing countries. It’s very possible that initially (around 1985), a lot of countries didn’t meet the criteria for a developed country, which is why the count for them is so low. The count increases dramatically over time, so a lot more countries are counted as a developed country as time goes on.

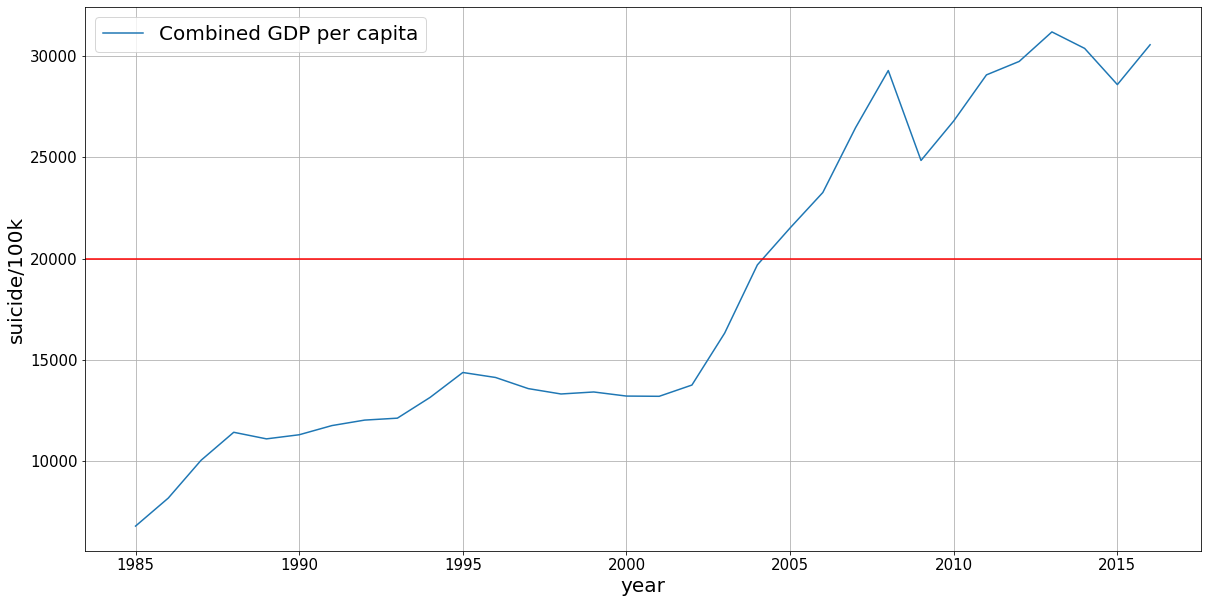
As you can see from the graph in figure 3, the combined GDP per capital has steadily increased over 1985 onwards which means a lot more countries end up meeting the criteria of being developed. (indicated by the horizontal red line).

Figure 3 - GDP per capita of world per year

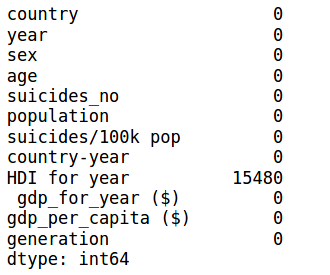
There is also a huge drop in the number of countries after 2015. This shows us that the data for 2016 is faulty as it contains relatively fewer countries compared to previous years. Therefore, we’ll be ignoring this outlier.

As observed earlier, a trend can be noticed after 2008 for developed countries where the suicide rates begin to increase over time, after a period of significant decrease in the 2 decades prior to that. 2008/09 was the beginning of the financial crisis that led to the Great Recession. This time saw the life savings wiped for a lot of people and many developed countries were especially hit hard. It’s very possible this is what led to the increasing suicide rates after 2008 among developed countries.

## **Using the HDI attribute**

Human Development Index (HDI) is a metric used by United National Development Programme to measure the level of human development in each individual country. The index consists of three components: life expectancy, education and per capita income. Scoring high in these components results in a high HDI.

Within our dataset, we have a column called HDI that measures the HDI for each country for each year. Checking for null values for the HDI column, we can see there are quite a lot of empty values.

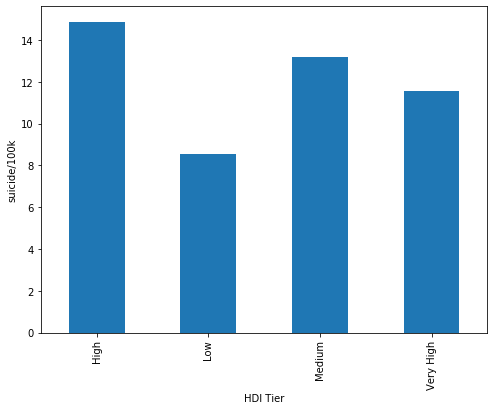


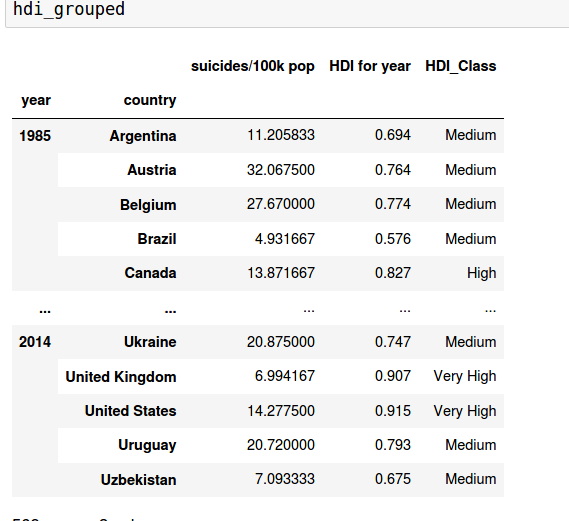
There are 15,480 null rows for the ‘HDI for year’ column, which gives us about 8000 rows to work with. First step is to make a dataframe that doesn’t contain any null values for the ‘HDI for year’ column.

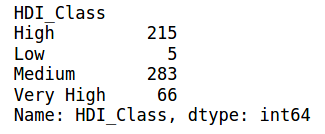
A new dataframe can be created by grouping the HDI dataframe by year and country and including the mean for ‘suicides/100k pop’ and ‘HDI for year’ columns.

Next step is to classify the HDI column into tiers of very high (0.90-1.00), high (0.80-0.89), moderate (0.55-0.79) and low (less than 0.55). The classification can then be done with the classify\_hdi function.

After applying the function on every row, a new column can be made with the new tiers and then the mean suicide rate of each tier can be plotted.

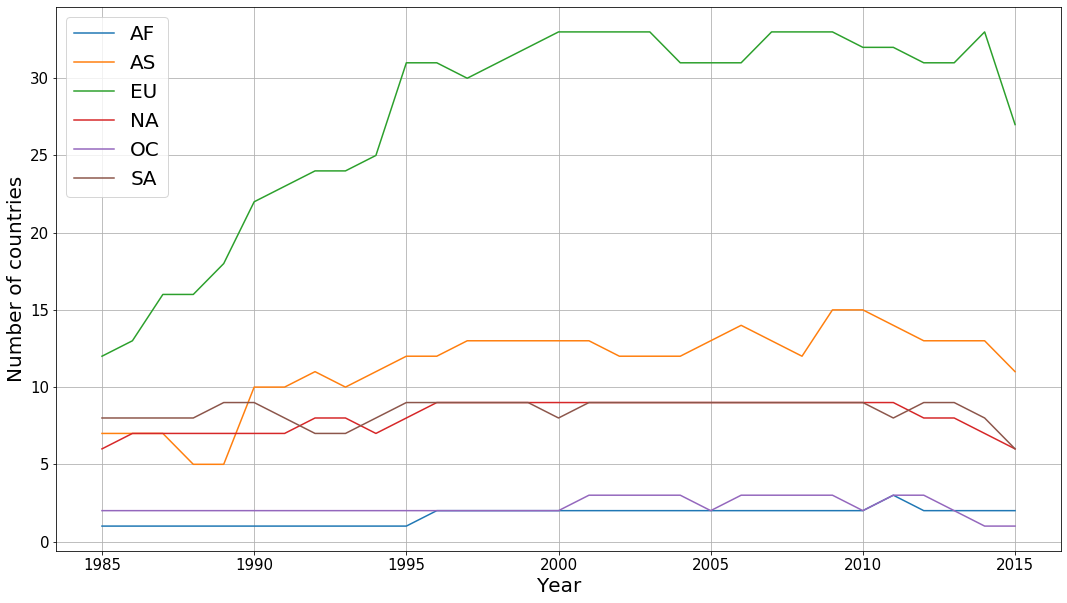




From the plot, there might be a general negative linear relation between HDI and suicide rates (i.e.: the higher the HDI, the lower the rate), but the low HDI tier having the lowest rate disproves that. Upon investigating further for the number of values in each category (see figure below), we can see that the value counts for low category are significantly lower compared to others. Since it has such low numbers, we’ll ignore this category as it can’t be used for any conclusions. This means that the linear relation between HDI and suicide rates is maintained.

## **BREAKDOWN BY CONTINENT**

Since we have the breakdown by country, we can get a breakdown of suicide rate by continent as well. In order to do this, we need to make a new column called ‘continent’ and map the country names to the corresponding continent. In order to do this, we used a python library called pycountry-convert (<https://pypi.org/project/pycountry-convert/>).

This involved a two-step process. The first step is to convert each country to a 2-letter code with the country\_name\_to\_country\_alpha2 method. The next step is to convert the 2-letter country code to a 2-letter continent code with the country\_alpha2\_to\_continent\_code method. This process was defined in the country\_to\_cont function and applied to every column. Figure 4 shows the plot of suicide rates of each continent over all years.

Two things jump out from this plot: The rate for Africa (AF) has a very high variance and a rate of the lowest rate in the later year and the rate for Europe is consistently higher compared to other continents and begins a period of gradual decline after 1995.

Figure 4 - Suicide rates/100k pop for each continent

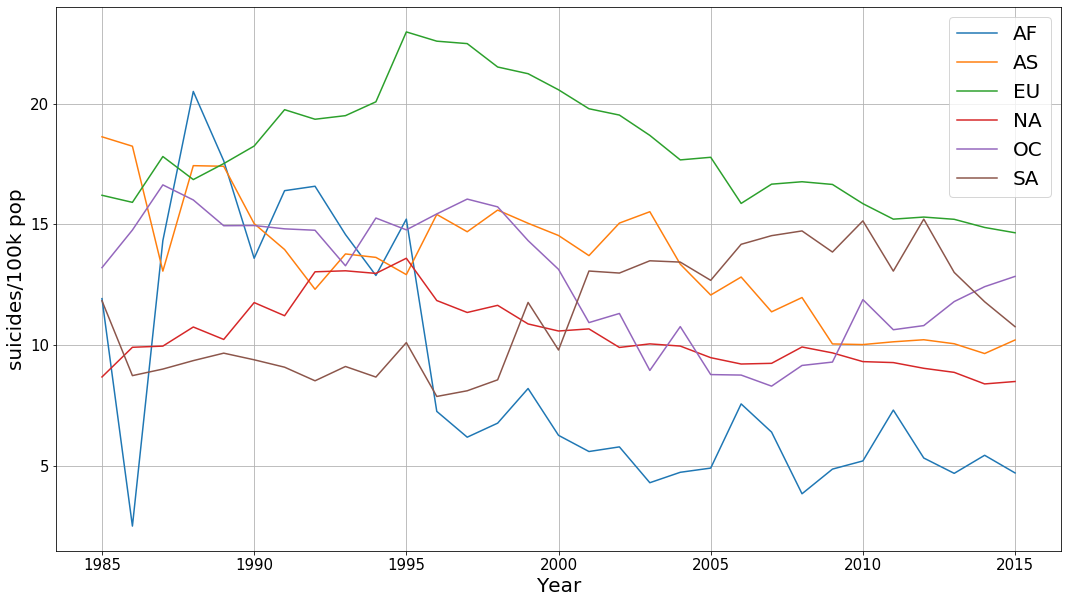
If we plot the number of countries in each year by continent (figure 5), we can see that the number for Africa is very low. This means the variability of suicide rate in each country within Africa has a large effect on the overall variability for Africa. Additionally, Africa has 54 countries and only one to two of them are represented in this dataset. It’s very possible that only countries providing the data are the ones with the lowest suicide rates, resulting in sampling bias.

Figure 5 - Count of countries in each continent per year