



جامعة عجمان
AJMAN UNIVERSITY

Data Analysis and Visualization

Using R and Rstudio

External Training

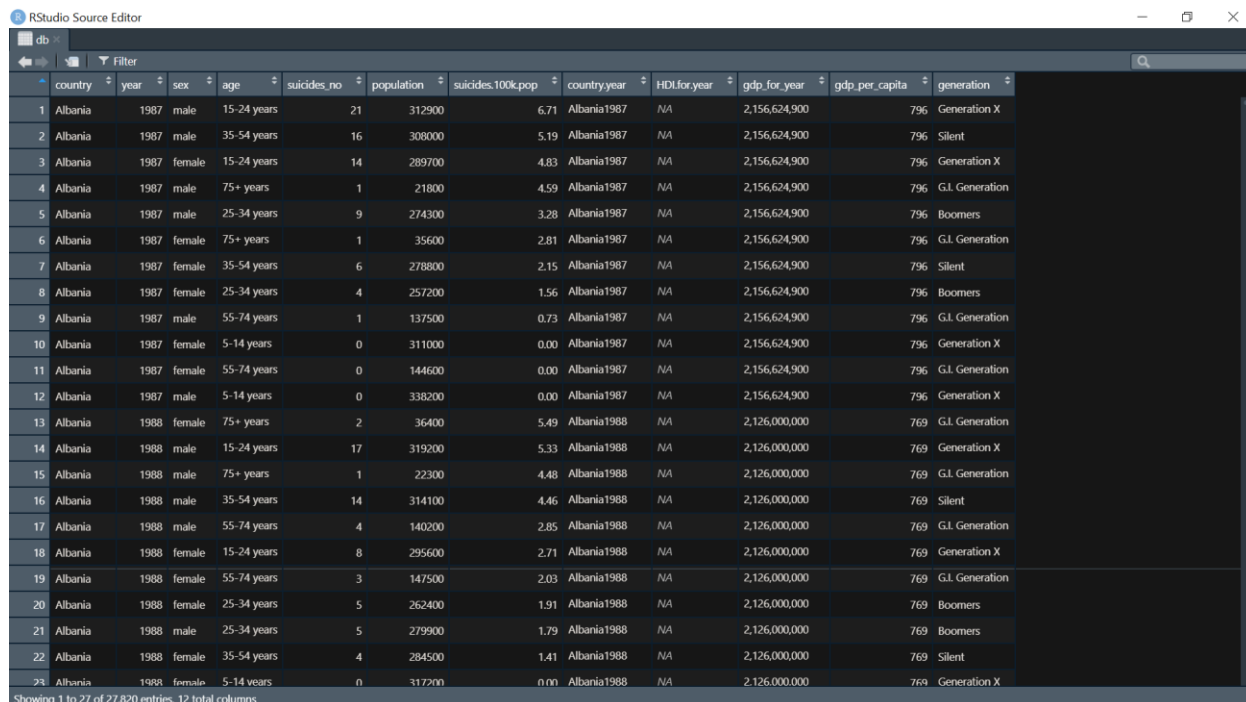
Name : Shaimaa Mahmood Mounir Kouka

ID: 201610434

Summer 2020

1. Choose database and Import it from Kaggle website [1]

The database we choose describes the number of suicides according to various factors, such as age, country and year, and GDP from year 1985 to 2016



	country	year	sex	age	suicides_no	population	suicides.100k.pop	country.year	HDI.for.year	gdp_for_year	gdp_per_capita	generation
1	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NA	2,156,624,900	796	Generation X
2	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NA	2,156,624,900	796	Silent
3	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NA	2,156,624,900	796	Generation X
4	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NA	2,156,624,900	796	G.I. Generation
5	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NA	2,156,624,900	796	Boomers
6	Albania	1987	female	75+ years	1	35600	2.81	Albania1987	NA	2,156,624,900	796	G.I. Generation
7	Albania	1987	female	35-54 years	6	278800	2.15	Albania1987	NA	2,156,624,900	796	Silent
8	Albania	1987	female	25-34 years	4	257200	1.56	Albania1987	NA	2,156,624,900	796	Boomers
9	Albania	1987	male	55-74 years	1	137500	0.73	Albania1987	NA	2,156,624,900	796	G.I. Generation
10	Albania	1987	female	5-14 years	0	311000	0.00	Albania1987	NA	2,156,624,900	796	Generation X
11	Albania	1987	female	55-74 years	0	144600	0.00	Albania1987	NA	2,156,624,900	796	G.I. Generation
12	Albania	1987	male	5-14 years	0	338200	0.00	Albania1987	NA	2,156,624,900	796	Generation X
13	Albania	1988	female	75+ years	2	36400	5.49	Albania1988	NA	2,126,000,000	769	G.I. Generation
14	Albania	1988	male	15-24 years	17	319200	5.33	Albania1988	NA	2,126,000,000	769	Generation X
15	Albania	1988	male	75+ years	1	22300	4.48	Albania1988	NA	2,126,000,000	769	G.I. Generation
16	Albania	1988	male	35-54 years	14	314100	4.46	Albania1988	NA	2,126,000,000	769	Silent
17	Albania	1988	male	55-74 years	4	140200	2.85	Albania1988	NA	2,126,000,000	769	G.I. Generation
18	Albania	1988	female	15-24 years	8	295600	2.71	Albania1988	NA	2,126,000,000	769	Generation X
19	Albania	1988	female	55-74 years	3	147500	2.03	Albania1988	NA	2,126,000,000	769	G.I. Generation
20	Albania	1988	female	25-34 years	5	262400	1.91	Albania1988	NA	2,126,000,000	769	Boomers
21	Albania	1988	male	25-34 years	5	279900	1.79	Albania1988	NA	2,126,000,000	769	Boomers
22	Albania	1988	female	35-54 years	4	284500	1.41	Albania1988	NA	2,126,000,000	769	Silent
23	Albania	1988	female	5-14 years	0	317200	0.00	Albania1988	NA	2,126,000,000	769	Generation X

Showing 1 to 27 of 27,820 entries, 12 total columns

Figure 1 : Database structure ,header, and sample of the data

2. Clean database

When we imported the database we defined the file-encoding description since interpreter of text editor/ web browser interpreted header i>>country Then R interpret it to i..country so to eliminate errors and also in our case, to limit the interpreter representation of character to only utf-8-BPM [3]. After we import the database, we use R functions to get more familiar with the database and to spot any inconsistency or not available data such as unwanted interpretation of headers.

In addition, we added a column to our data base and named it suicides_ratio to represent the number of suicides related to the population of people to better describe the suicide rate considering the population of that nation

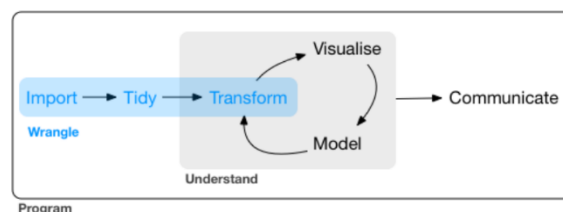


Figure 2: The process of data visualization which consists of import, tidy then manipulate and model the data to draw it in graphs then communicate to public through research papers, or articles

So, now we have 12 headers and the added field: country of type char which we have in our database over 101 different country from around the globe from year 1985 to 2016 and it is of type int. Sex of type char is either male or female, age of type char, suicides_no of type int, population of type int which differ based on the country. Also, suicides.k100pop which represent the number of suicides in population of 1000, country. Year of type char which is merging 2 headers country and years. Then we HDI.for.year of type num, which stands for human development index for year, furthermore, defined by [4] as “statistical tool used to measure a country's overall achievement in its social and economic dimensions. The social and economic dimensions of a country are based on the health of people, their level of education attainment and their standard of living”. Also, gdp_for_year of type char (gross domestic product for a year). GDP is the monetary value of all the finished goods and services produced within a country's borders in a specific time period and includes anything produced by the country's citizens and foreigners within its borders. It is primarily used to assess the health of a country's economy while gdp_per_capita of type int is a measure of a country's economic output that accounts for its number of people. It divides the country's gross domestic product by its total population. That makes it a good measurement of a country's standard of living. It tells you how prosperous a country feels to each of its citizens, and finally generation of type char which is either boomer ,silent, G.I , Millenials ,generation x .

3. data analysis and visualization

3.1 Suicide cases categorized based on gender

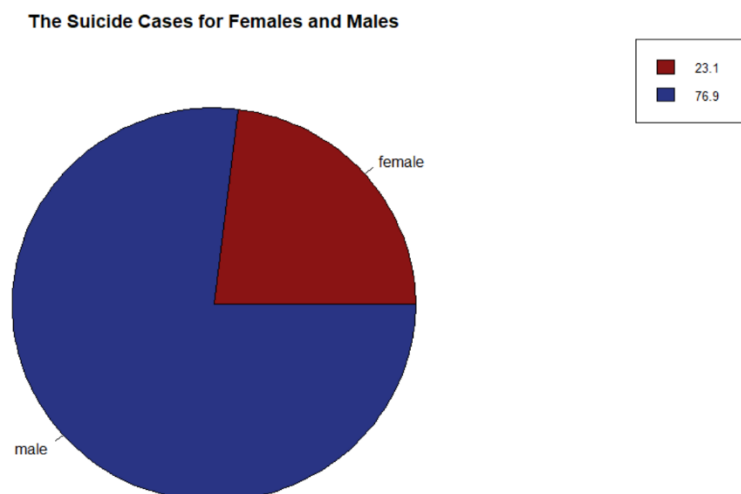


Figure 3: The figure showcases the ratio of women suicides and men suicide over the span of 31 years in various countries around the world

In figure 3, we studied the suicide cases and its relation to gender and its obligations and characteristics, and we found that 23.1 % of the people who committed suicide globally from 1985-2016 were women and 76.9% were men. Since, the majority of people who committed suicide were males we want to furthermore seek the relation between committing suicide over the years and gender of the person who committed suicide.

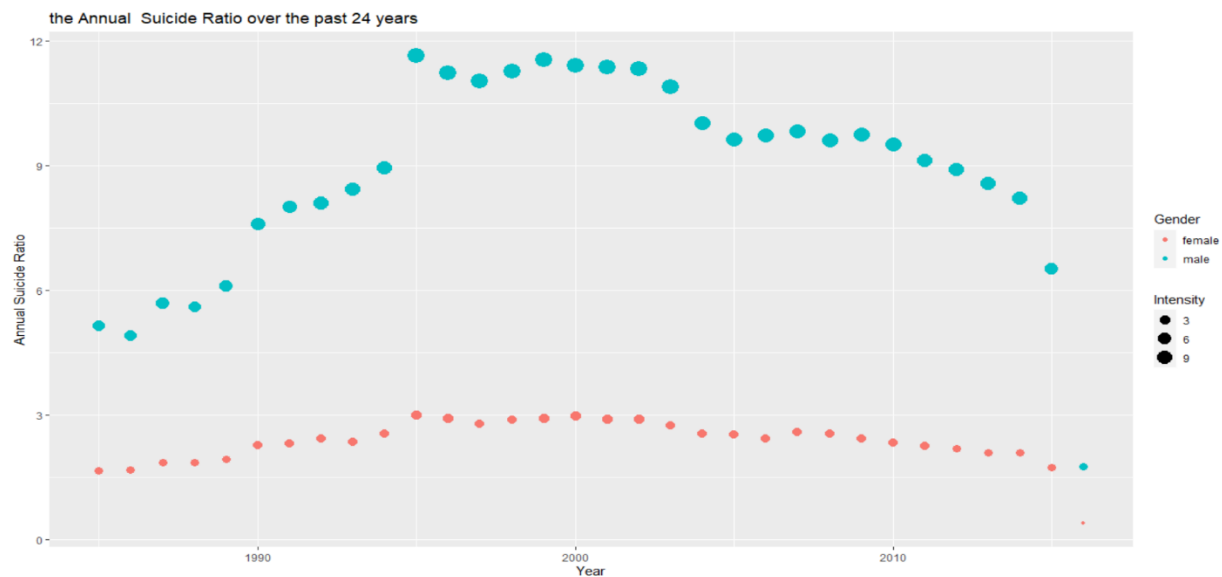


Figure 4: The scattered plot graph illustrates the relation between the gender of people who committed suicide with the number of cases found in the data base

For scattered graph in figure 4, we colored the data related to female with orange and blue for male and displayed the scale of the annual suicide ratio by the size of the dot in the graph. We observed that the annual ratio for female and males confirms the previous conclusion about relation between gender and suicide ratio and the ratio peaked for both genders in 1995 then decreased slightly for males and stayed approximately constant for females. And We can see that points that represents females were changing over time very slightly compared to the changes for the males with remaining higher than females for all years. Finally, in the last year for both genders the ratio shrank compared to year 1985.

3.2 Suicide cases categorized based on Annual GDP per capita

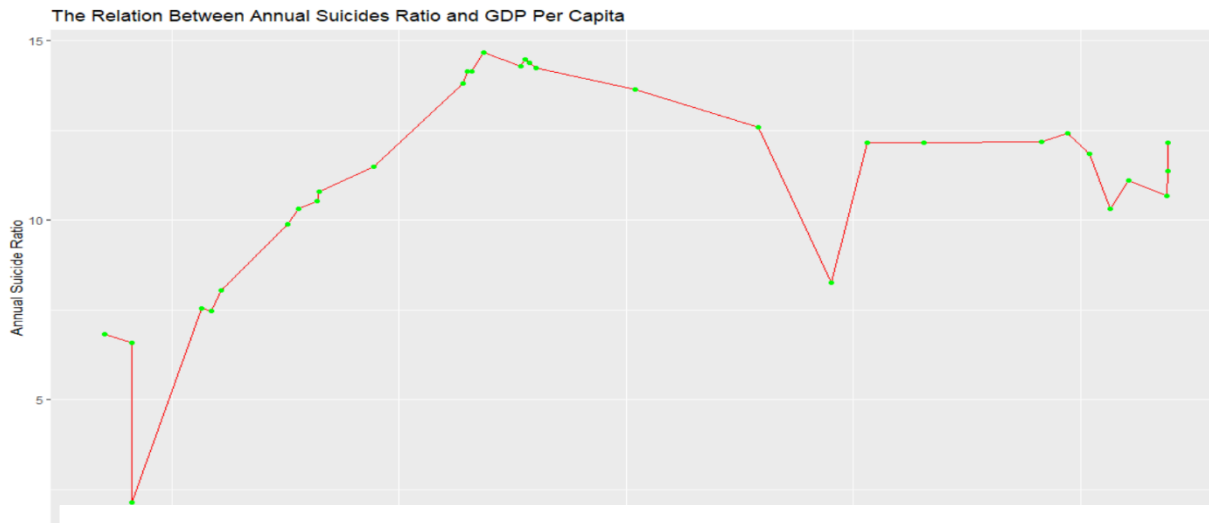


Figure 5: the relation between the Annual gross domestic per capita and Annual suicide rates

The annual gross domestic per capita for all 101 countries and their suicide rates have a proportion relation. Each point represents the global GDP per capita and the global annual suicide rate in a year so, what we discovered, that over the years the GDP increases due to better education system and advanced technologies and as a result suicide ratios increases linearly until it peaks to continuous decreasing. In addition, overall GDP per capita is affecting and the suicide ratio in a non-linear manner. Therefore, we think that probably the events in years are producing this no-linear increase.

3.3 The countries with the highest Suicide rates

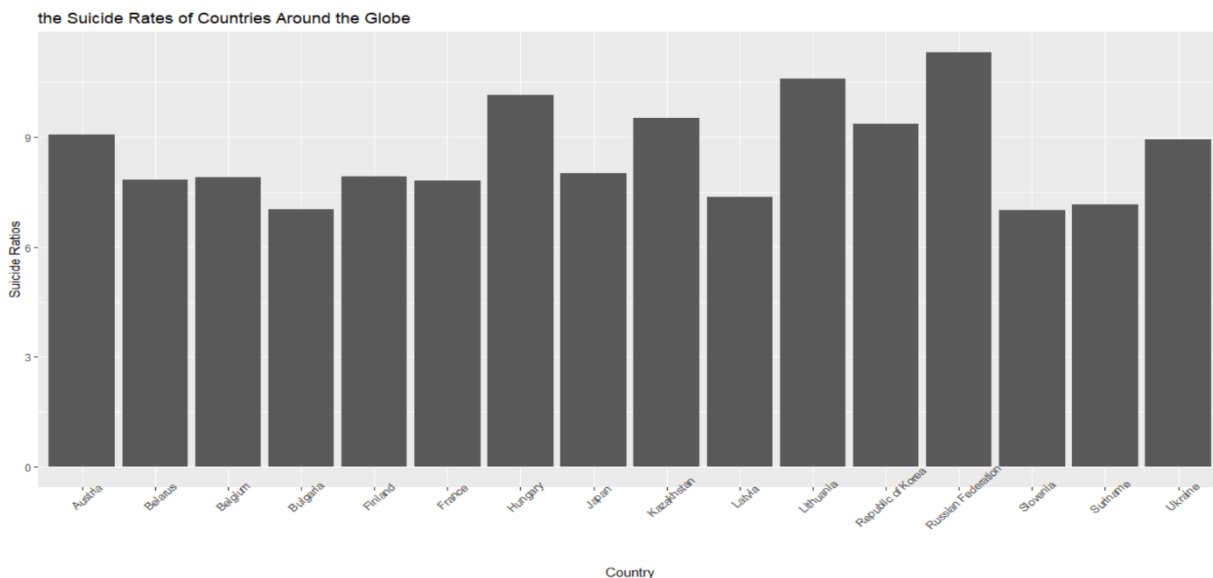


Figure 6: the relation between the country we live in and rate of suicides

Finally , we want to observe the effect of the geography distribution on the suicide rates , so we choose 16 countries that scored the highest rates of suicides between 101 countries and compare between them to find the factor that all these countries have in common.

The country with the highest rate of suicides is Russia, Lithuania , then Hungary with ratio greater than 9. What we found is that the majority of the countries that have high rate of suicide is distributed in Europe and part of Asia.

conclusion

Suicide rate have many factors that contributes in the increase such as geographic distribution, as we concluded the countries with the highest rates are the ones located in Europe and Asia and over the time as the education system and economy is developing thus, the GDP per capita is increasing which does not necessarily means better , or happier life just the growth of economy .So, as economy of countries is expanding the rates are increasing non-linearly. And Finally, the rates of male suicides are always higher than female suicides.

References and resources

[1] Kaggle Webpage , <https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016> , accessed on June 2020

[2]H.Wickham and G.Grolemund,"R for Data Science", <https://r4ds.had.co.nz/explore-intro.html>, O'Reilly Media, 2017

[3] Stackoverflow Webpage <https://stackoverflow.com/questions/24568056/rs-read-csv-prepend-1st-column-name-with-junk-text>, accessed on June 2020

[4]The economics times webpage, <https://economictimes.indiatimes.com/definition/human-developmentindex#:~:text=Definition%3A%20The%20Human%20Development%20Index,its%20social%20and%20economic%20dimensions.&text=Every%20year%20UNDP%20ranks%20countries,released%20in%20their%20annual%20report.> , Accessed on June 2020