POLITÉCNICA

REPORT:
SUICIDE RATES OVERVIEW 1985 TO 2016

Statistical Analysis HW 1.1

Junhui Liang, Miguel Pérez, Ángel Igareta

October 20, 2019

1 Dataset Description

The dataset used in this assignment was retrieved from a Kaggle project named *"Suicide Rates Overview 1985 to 2016"* [2]. It is a mix of four datasets linked by time and place which contains relations between socio-economic information and suicide rates by year and country. The main motive of publishing this dataset is to help prevent suicides in the global scope.

This dataset is composed by the following variables:

- **Country:** Name of the country annotated. It contains 101 different ones.
- **Year:** Year of the annotation. It ranges from 1985 to 2016.
- **Sex:** Sex of the annotated population (male or female).
- **Age:** Age group of the annotated population. There are 6 different groups: [5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, 75+ years].
- **Suicides number (suicides_no):** Number of suicides of the annotated population.
- **Population:** Number of inhabitants of the country in the annotation.
- **Suicide Rate per 100k population (suicides/100k pop):** Suicide rate (number of suicides divided by population) multiplied by 100k in order to be more intelligible.
- **Country-Year:** Variables country and year joined by a dash.
- **HDI for year:** Human Development Index per year, which indicates a composite index of life expectancy, education, and per capita income indicators.
- **GDP for year (gdp_for_year):** Gross Domestic Product for year, which indicates the total market value of all finished goods produced within a country´s borders in a year.
- **GDP per capita (gdp_per_capita):** GDP for year divided by the population.
- **Generation:** Demographic cohort of the annotated population. There are 6 different groups: [Generation Z, Millenials, Generation X, Boomers, Silent, G.I. Generation]

The dataset contains 12 columns and 27.821 rows. After a careful observation, we concluded the data is not homogeneous, there are some countries that only have one year annotated. Nonetheless, the requirement is not needed in our research questions, because they are focused on a global scope, not in comparing only two countries. The only requirement would be that in every year annotated per country, the split in ages and sex would be the same, and it is met.

In respect of null values, there were only observed in the column HDI for year, which is not used in our questions (due most of the values are null), so there was no need for handling those null values.

2 Research Questions

2.1 Suicide Rate and GPD

First, it will be studied if suicides are proportional to Gross Domestic Product per country and if there is a relation between the GPD per capita and the suicide number, in order to research if there are more suicides in poor countries than in rich ones.

2.2 Suicides in Countries over the years

Next, we will try to study how suicide rate has changed over the years, if there has been a trend increasing or decreasing suicide rates on a global scope.

2.3 Suicides during European debt crisis

It would be interesting to research if there was a relation between the European debt crisis and the suicide rate focusing on the European countries. We will try to investigate if there was an increase in suicide number and in which countries.

2.4 Suicides, Gender and Age

Here, we will try to investigate if suicide rate is higher depending on the gender or the age group.

2.5 Suicides in low populated age groups

Finally, it would be interesting to study if there is a relation between suicides number and population, being able to answer: Is suicide more likely to occur in countries with lower population?

3 Findings

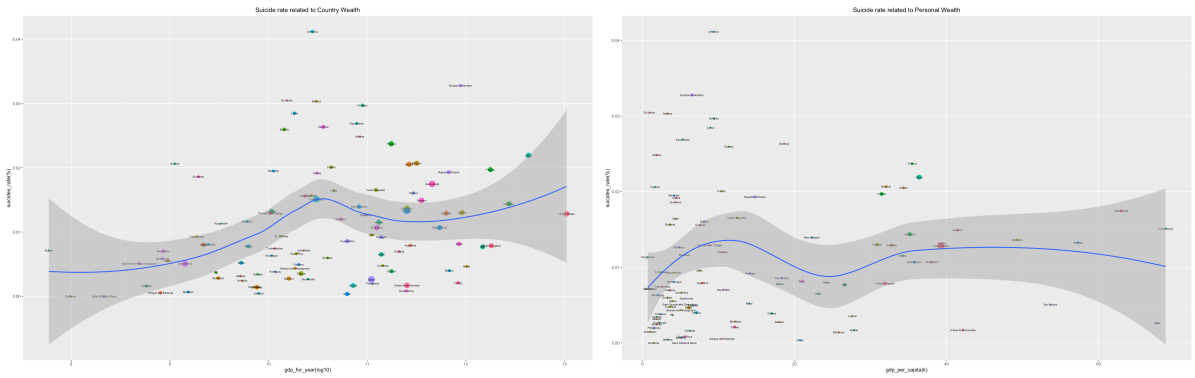
3.1 Suicide Rate and GPD

3.1.1 Data preparation and manipulation

Firstly, in order to obtain the desired variables, the averaged suicide rate and GDP over the years, we used group_by twice to filter the relevant columns and remove some irrelevant factors. The first time was to aggregate the number of suicide and population over ages and genders per year, while the second one is averaged by the years.

3.1.2 Findings

After that, we selected a scatter plot to describe the trend and relationship between suicide rate and GDP per country. Besides, gdp_for_year, representing country wealth, adopted "log" function with the purpose of intuitive observation (Figure 1a), while gdp_per_capita (1b) is gdp_for_year divided by population, representing personal wealth.



(a) (b)

3.1.3 Conclusion

As the graph below showed, using fitting curve to simulate the trend, we can come to the conclusion that suicide rate is concentrated on the specific range of GDP and not obviously linearly related to GDP at some extent due to the dispersed distribution. Comparing both figures, we can see an interesting discovery that high suicide rate focus on the middle level of country wealth, yet concentrate on the low level of personal wealth at the same time. We can conclude that people living in medium wealth country with low income have high tendency to suicide.

3.2 Suicides in Countries over the years

3.2.1 Data preparation and manipulation

The main purpose of this section is to find a trend along the years for the suicide rates. As overall life conditions tend to improve, we may think that suicides decrease as well. However this is a biased perspective, since suicides may happen even in a wealthy society, so we must approach this issue carefully.

In the figure 2, we have plotted the evolution along the years for the suicide rates, which is the result of dividing the total suicides registered in the world, and the entire global population. We must emphasize that this is a simplified graph, since it does not show the distribution of countries, ages nor genders, since we just added all suicides and population data along an entire year in all countries and divided the sum. Note as well there is less data in the last recorded year (2016).

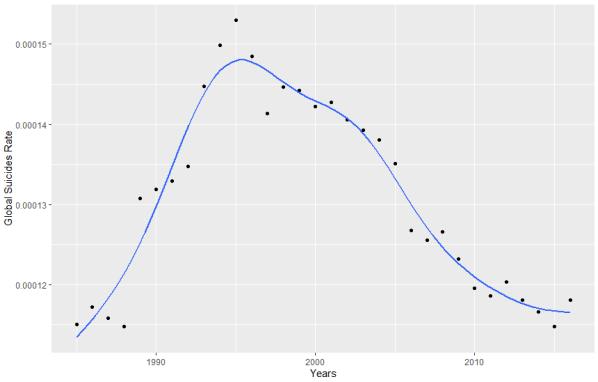


Figure 2

3.2.2 Findings

As we can see, there was a raising trend in suicides until 1995, from where it started decreasing. Besides, we can notice that at the pace of the suicides decrease is stagnating. Curiously, we can see that the 2008 crisis has not affected to the global suicides, but it may have more local impact in suicides rather than in the global perspective. We will find out more about it later on in the section. 3.3.

Other issue we must address is the overall suicide distribution in one country along all years. In the graph 3, we can see an extract for some countries about the suicide rates mean along the years.

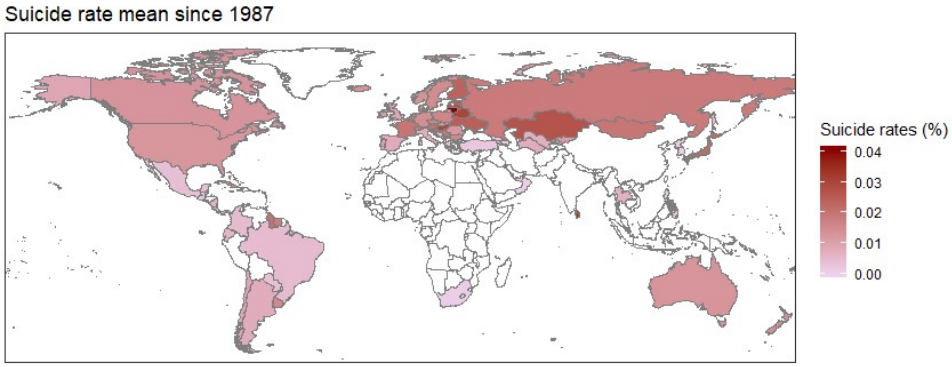


Figure 3

3.3 Suicides during European Crisis

3.3.1 Data preparation and manipulation

In order to find a relation between our data and the European debt crisis, a brief research had to be done. The multi-year debt crisis has been taking place in the European Union since the end of 2009 [1]. As most of the countries were able to exit their bailout programs during the end of 2014, we will focus on examining the European countries suicide data in the range of years [2008, 2014].

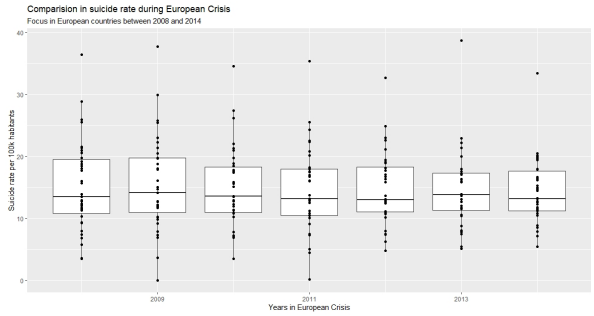
As the variables 'age' and 'sex' are not relevant to answer this question, the data was summarized by adding the suicide number of each of those groups and dividing by the total population of the country. The summary result was the suicide rate per country and per year, in order to research its evolution.

3.3.2 Findings

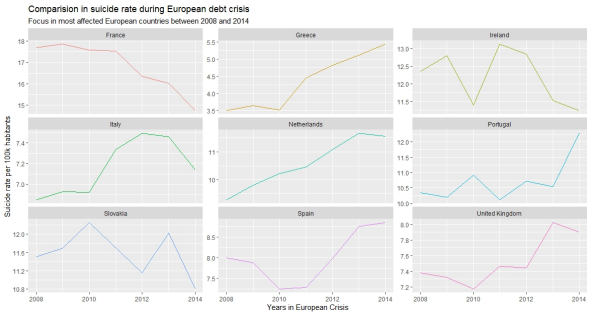
The first approach consisted on representing the statistical data of the suicide rate in all the European countries, in order to find out if there was a significant increment in this variable over the years. Through the figure 4a, it can be seen that its median does not increase notably so there is no increment.

Although there is no major increment in the suicide rate median in the European countries set, it is evident that some countries were affected more than others in the debt crisis. After studying which were

those countries, we could filter our dataset. In order to research if there was an increment in the suicide rate over the years, we should compare each country in its own suicide rate scale, as represented at 4b.



(a)



(b)

3.3.3 Conclusion

It can be seen that the suicide rate in some of the most affected countries significantly increased in the years of the European Crisis, as it is the case of Greece, Italy, Netherlands, Portugal, Spain and United Kingdom. On top of that, in the plot it can be observed the total increment of the suicide rate, being almost a 2% in some countries such as Greece and Portugal. Hence, we can conclude there is a relation between the suicide rate in the European countries and the European debt crisis period.

3.4 Suicides, Gender and Age

3.4.1 Data preparation and manipulation

In order to prepare the data to tackle this question, as we seek to find a relation between suicide rate, gender and age, the year variable is not needed, so it could be summarized by averaging the population and suicides number over the years per country. However, averaging these variables is a way of approximating the data, it is better to maintain the original values per year.

3.4.2 Findings

The best way to find a relation between the suicide rate, gender and age would be through a box-plot where it can be seen the suicide rate statistics per age group and per sex, as presented in the figure 5. Each dot of the box plot would represent a suicide rate in an age group, sex, country and year.

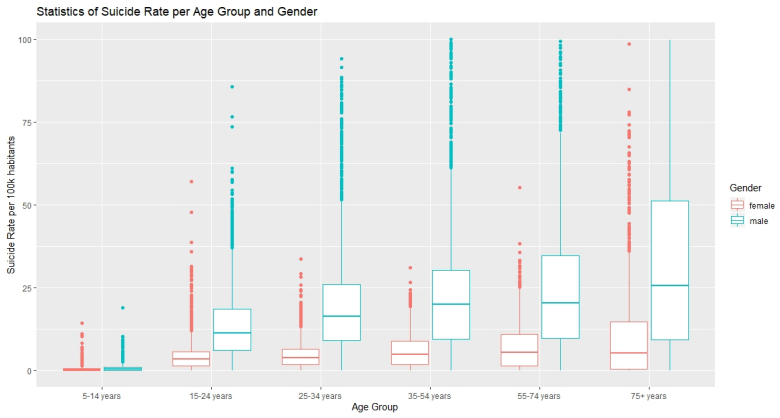


Figure 5

3.4.3 Conclusion

As conclusion for this research question, it can be seen that male gender median is higher than female one in all the groups, so it can be demonstrated that men tend to suicide more than women.

On the other side, when the focus is on the age groups, it can be appreciated that age is directly proportional with the suicide rate, being the oldest group the one with the highest suicide rate of all the groups.

3.5 Suicides in low populated age groups

3.5.1 Data preparation and manipulation

To address this question, we selected every country and in each one we studied all population clusters along the years. This way we had 6 groups by ages: (5-14, 15-24, 25-34, 35-54, 55-75, +75) years. All data about suicides and population were added together, except country belonging (fact which separates clusters), averaging those clusters over all the years to see their general behavior.

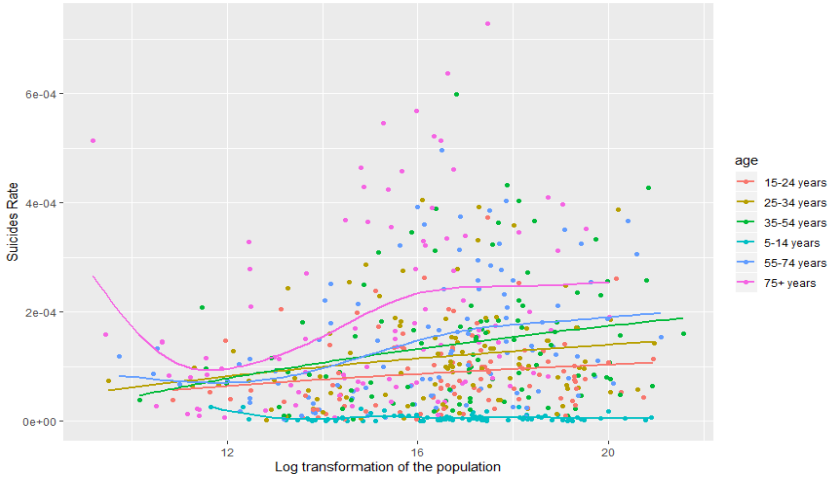


Figure 6

3.5.2 Conclusion

As it can be seen through the figure 6, suicide is more likely to happen in big populated cluster areas (except for some '+75 years' and '5-14 years' clusters, in which suicide is surprisingly higher) and its often proportional to age.

4 Data Analysis Plan

The purpose of this study is to find some viable empirical correlations and dependencies. Hence, we would use supervised learning algorithms such as decision trees or other classification models in order to **predict** rates of suicides given some specific future demographic evolution of a subject of study.

On top of that, some more sophisticated curve fitting and regression plots could be used in order to find the relationships and correlations between variables, trends and making predictions in the future. The main variable interesting to predict would be the suicide rate, a quantitative continuous variable.

Through the use of these machine learning algorithms, we could focus on predicting the suicide rate of a person given its gender, age and the country she or he lives in. Furthermore, the use of convolutional neural networks to solve these problems is also interesting, especially for those difficult to fit with low degree fitting function, but it might require more examples over the years and data input.

References

[1] Bird, B., and Kenton, W. European Sovereign Debt Crisis. https://www.investopedia.com/terms/e/european-sovereign-debt-crisis.asp*https://www.investopedia.com/terms/e/european-sovereign-debt-crisis.asp*.

[2] Rusty. Suicide Rates Overview 1985 to 2016. https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016*https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016*.