Statistical data analysis

Suicide understanding and prevention

A general assignment for data scientists

Fall 2019/2020

Introduction

You are provided with a suicide worldwide overview collected over approximately 30 years. The overview was built to identify signals correlated to increased suicide rates among different cohorts/subgroups. The cohorts can be defined across countries, continents, genders, age groups and the socio-economic conditions. In the beginning, you are supposed to visualize the main relationships and trends. Your main task is to learn from the dataset and to understand suicide risks, for example, to identify the subgroups with an increased suicide rate. Also, you should construct, test and evaluate a predictive model that gives the total number of suicides in the given subgroup and time period.

The task can be found at: Kaggle. There are some solutions to the task there yet. You can benefit from them namely in the exploratory analysis. In further efforts, you are supposed to provide your own models and statistical tests that you learned within the SAN course.

1 Data description

You are given a data frame that comprehends 27,660 records, each record contains 10 variables. The file is *sucide.RData*, you can easily import it into R with *load(file="suicide.RData")*, it loads into *data*. A csv file is provided too (the same content, less convenient to deal with in R). The individual variables are as follows:

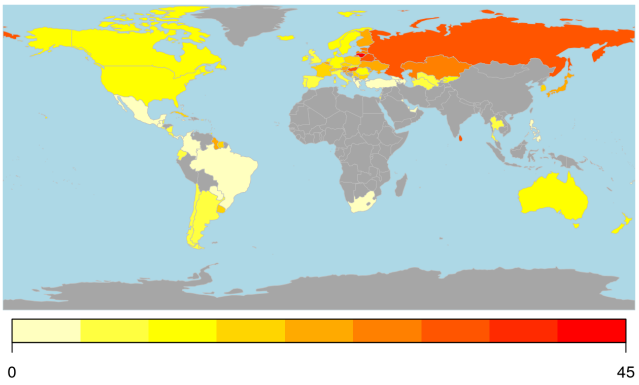
- country, year, sex, age – the meaning is straightforward,
- suicides_no – the total number of suicides in the given group (a country-year-sex-age level, see previous vars),
- population – population at a country-year-sex-age level, if you grouped the data by country and year and summed up the population, you should get the population of each country over time,

- gdp_for_year – Gross Domestic Product at the year of person's suicide within the country, GDP = the total monetary value of all final goods and services produced (and sold on the market) within a country during a period of time (typically 1 year), GDP is the most commonly used measure of economic activity,
- gdp_per_capita – Gross Domestic Product per capita at the year of person's suicide within the country,
- generation – indicates the period of birth, G.I. Generation – born between 1901-1927, Silent – born between 1925-1942, Boomers – born between 1946-1964, Generation X – born between 1960-1980, Millennials – born between 1980-early 2000, Generation Z – born between mid-1990 - 2000s,
- continent – in which continent the country lies.

2 Tasks

The main aim of this assignment is to demonstrate your ability to apply the methods that you learned during the course. The assignment can be decomposed into the following well-defined tasks:

1. **Exploratory analysis of the suicide dataset**. Load the dataset, visualize the main relationships and trends, preprocess the data, carry out dimensionality reduction and clustering. The main goal is to see whether there are regularities in suicide rates across countries, continents, periods of times, etc. You are supposed to conclude this subtask with a brief description of suicide patterns.



A geographical heat map of the suicide rates (taken from Kaggle kernel Suicide Rates (in-depth) – Stats & Insights.)

2. **Hypotheses testing.** The outcome of the previous step should be a set of candidate hypotheses about suicide patterns. In this task you are supposed

to formally test them. An example hypothesis could be: H_{null}: the overall suicide rate across continents does not change, H_a: there is at least a pair of continents whose overall suicide rates differ. You are supposed to propose and test at least 3 hypotheses.

3. **Suicide predictive model.** Create a model that predicts the suicide rate based on the remaining variables. Propose a meaningful train/test scenario. Evaluate and compare the performance of the models, employ cross-validation to get an unbiased estimate. Utilize feature selection to simplify the models and improve their performance. Use at least one method that we touched in the course, but you can compare with other methods too (neural networks, gradient boosting trees, etc.).

4. **Discussion of the results.** Provide a verbal summary of your results (approx. 10 sentences). Compare with an external suicide resource, an example could be: the WHO summary. Propose future work or potentially interesting tasks you could not solve with the given dataset.

3 Submission and evaluation

Submit your solution to the upload system. Submit only the rmarkdown file named DStask$YOURFELUSERNAME.Rmd. This file should be considered as a report containing a definition of the task, description of your implementation details, graphical outcomes and your **detailed answers** to the required tasks.

You can obtain up to 15 points for this assignment, 4 points per each of the first three subtasks plus 3 points for the final discussion. In the first three subtasks, approximately 70% will be given for the concept of the solution (selection of statistical methods and their correct application, depth of the solution), 20% for the answers that summarize your solution in the individual subtasks (interpretation of your results and explanation of their practical impact in natural language), 10% for formal issues (clarity of the code, comments, readability of the rmarkdown as a whole).