Suicide Rate Analysis Across Regions Between 1985-2015

RETAKE - Programming and Data Analysis for Business (BINTV2006E) - Home assignment - written product (UC)

Exam number: S129049

Program: BSc International Business in Asia

Date of submission: Feb 28th 2022

Number of characters: 19.487

Number of pages: 10

Abstract

Understanding the influencing factors of suicides can help governments identify which part of the population should they pay more attention to and what suicide prevention measures should be introduced. This paper aims to shed light on the differences between the suicide numbers within different regions, while analysing the number of suicides based on Age, Generation, GDP and Country. Thus, answer the question, *To what extent do suicide rates differ between regions?*

An Explorative Data Analysis have been carried out by visualising the data and carrying out a hierarchical cluster analysis. Visualising the data helped identify the key patterns amongst the different variables, and cluster analysis helped identify five groups which were then analysed on their unique attributes.

Based on the results high suicide numbers were found in countries in Europe, with Russia having the highest number of suicides; amongst males; in the age group 35 until 54 and within Boomers and Generation X. It was also seen that there may be a connection to GDP and the number of suicides as the global suicide rate decreased while the global yearly GDP increased.

The paper aims to serve as a basis for further research in this area and suggest to take into consideration various other social and economic factors to determine the best strategy possible for suicide prevention.

Keywords:

Suicide Rate, Exploratory Data Analysis, Hierarchical Clustering, Global Suicide Rate, Suicide Numbers

Introduction

Suicide is the act of a person ending their lives on purpose. Suicidal behaviour accounts for a major problem in regards to mental health and on a social level. It takes an effort from both primary and secondary health care settings to deal with it. In some countries, the number of suicides are higher than the number of deaths in traffic accidents. In many countries the suicide is one of the main leading causes of death for younger age groups, however it is also an increasing problem amongst the elderly (Schmidtke et al., 2007, p. 81).

There is one person every 40 seconds who pass away due to suicide which adds up to 800.000 people yearly. Suicide is a complex illness which can be influenced by personal, social, psychological, cultural, biological and environmental factors. It is very important to have context to understand the reason of it. It is important for societies and governments to recognise it in the early stages while it is treatable as it has numerous consequences for the people left behind (World Health Organization, 2014, p. 11). According to Liu (2009, p. 204) suicide rates differ across the different countries in the world. He argues that social factors due to geography have an influencing and important role in suicide.

The aim for this paper is to analyse and answer the question "*To what extent do suicide rates differ between regions?*". It aims to understand the reason behind the different suicide rates and how it varies taking into consideration the different metrics the dataset contains such as Age, yearly GDP, GDP per capita, Generation, Country, Region and Subregion. The paper takes the Suicide Dataset from Kaggle.com which has information of the afore-mentioned information in the period 1985-2016. The findings can have implication on the different factors societies have to take into consideration in order to prevent suicide within their country.

Data Analytics Methods

The paper carries out an Exploratory Data Analysis, the steps for this are the following according to Wickham (2017):

1. A person has to generate questions about the data they will be working with.
2. Answers need to be searched through visualising, transforming and modelling the data.
3. The findings can be used later to either refine the question or generate new questions.

This way of analysing data does not have a formal process or a strict set of rules, it is more of a "state of mind". EDA is and important part of analysing data as it investigates the quality of data and analyses if it meets the necessary expectations during the research process by asking questions. The goal in EDA is to understand the data someone is working with. The two useful questions to ask during EDA is about the variation and covariation between the variables. EDA is

rather a creative process, in which by asking questions to dive deep in the data it increases the chances of making a discovery (Wickham, 2017).

The paper also uses Hierarchical Clustering which does not require a particular choice of the number of clusters (K). Clustering is a set of techniques to find subgroups, or clusters in a data set. When clustering the observations in a chosen set of data are analysed to distribute them into distinct groups, making the observation in a group similar to each other and different towards other groups. The first step is to define what it means for observations to be similar or different, this has to be done according to the data. Clustering is an unsupervised method as we are trying to distinct clusters based on the data set. The paper have chosen Hierarchical clustering as this method is used when we are unsure of the amount of clusters we want. This method creates a dendrogram, which allows us to view the obtained clustering for each number of clusters at once. The paper uses the bottom-up or in other words, agglomerative clustering which means the dendrogram is built up from the observations and is combining clusters as going up. The Euclidean distance is used to define dissimilarity amongst the pair of observations (James et al., 2021).

Dataset Description

The paper took the "*Suicide Rates Overview 1985 to 2016*[1]" and the "*Country Mapping - ISO, Continent, Region*[2] " datasets from Kaggle (2022) and the "*ISO-3166-Countries-with-Regional-Codes*[3]" dataset from Github (2022). The *Suicide Rates Overview* dataset contains 27820 observations and 12 different variables. The variables are the country's name; years ranging from 1985 to 2016; gender; age; the number of suicides; population; suicides per 100.000 population; combination of the country's name and the according year; HDI for year, which is the Human Development Index for the according year; GDP per year; GDP per capita and generations.

Both the *Country Mapping* and *ISO-3166-Countries-with-Regional-Codes* dataset contains 249 observations and 11 columns. It was important to merge these two datasets together as some countries names´ are presented in a different way in the two datasets, thus this allows the best combination to the Suicide Rate dataset. The variables are the name of the country; alpha-2; alpha-3; the country code; the ISO 3166-2; the region; the subregion; the intermediate region; the region code; the subregions code and the intermediate region code.

Data Wrangling and Visualization

[1] https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

[2] https://www.kaggle.com/andradaolteanu/country-mapping-iso-continent-region

[3] https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/blob/master/all/all.csv

Most of the information that is needed for this analysis derives from the *Suicide Rate Overview* dataset. This has been combined with the *Country Mapping* and the *Countries with regional codes* datasets in order to analyse the suicide numbers within the different regions and subregions. The two datasets containing the regions and subregions have been reduced to three columns which have been used in this analysis, these are the country's name - to be able to match the columns -, the regions and the subregions. Firstly the two continents datasets were combined using `rbind()` into one dataset, then a function, called `mymerge()` has been created, allowing two datasets to be combined in line with the country name.

```
mymerge <-
  function (x, y) {
    masterdata <- merge (
      x,
      y,
      by.x = c("country"),
      by.y = c("country_name"),
      all.x = TRUE
    )
    return(masterdata)
  }
```

Due to combining the two datasets with the country's name and its regions and subregion there is a duplicate for each line in this phase. However this is cleaned later on with the `distinct()` function. There are 20.824 rows that contain missing values in the HDI, region and subregion columns. The data is checked for irregularities, the 2016 year is excluded as it is incomplete for the countries, furthermore, countries with less than three years of data are also excluded. One column that contained purely the combination of the Country's name and the Year and another one containing HDI have been removed as it was not necessary for the analysis. The missing values from the region are removed. The data frame is tidied up by cleaning the nominal factors, setting age and generation as ordered factors. A value which contains the global suicide rate over the period of time for visualisation is created. The finalised data looks the following. It has 26.820 rows and 12 columns. Out of these columns 4 are characteristic variables, 6 are numerical and 2 are ordered factors.

```
## tibble [26,820 × 12] (S3: tbl_df/tbl/data.frame)
##  $ country       : chr [1:26820] "Albania" "Albania" "Albania" "Albania" ...
##  $ year          : num [1:26820] 1995 1995 1995 1999 1999 ...
##  $ sex           : chr [1:26820] "male" "male" "male" "male" ...
##  $ age           : Ord.factor w/ 6 levels "5-14"<"15-24"<..: 6 4 2 4 3 5 2 3 5 4 ...
##  $ suicides_no   : num [1:26820] 1 14 11 31 19 14 19 13 6 5 ...
##  $ population    : num [1:26820] 25100 375900 241200 391100 242300 ...
##  $ suicides_100k : num [1:26820] 3.98 3.72 4.56 7.93 7.84 7.56 6.4 4.7 3.18 1.34 ...
##  $ GDP_year      : num [1:26820] 2.42e+09 2.42e+09 2.42e+09 3.41e+09 3.41e+09 ...
##  $ GDP_per_capita: num [1:26820] 835 835 835 1127 1127 ...
##  $ generation    : Ord.factor w/ 6 levels "G.I. Generation"<..: 1 3 4 3 4 2 4 4 2 3 ...
##  $ region        : chr [1:26820] "Europe" "Europe" "Europe" "Europe" ...
##  $ subregion     : chr [1:26820] "Southern Europe" "Southern Europe" "Southern Europe" "Southern Europe" ...
```

The numerical variables include the year, the number of suicides, the population, the number of suicides per 100.000 people, the yearly GDP and the GDP per capita. The range of these

variables are presented below.

```
      year          suicides_no        population        suicides_100k       GDP_year          GDP_per_capita
 Min.   :1985   Min.   :    0.0   Min.   :     278   Min.   :  0.00   Min.   :4.692e+07   Min.   :   251
 1st Qu.:1995   1st Qu.:    3.0   1st Qu.:  102588   1st Qu.:  0.99   1st Qu.:9.399e+09   1st Qu.:  3397
 Median :2002   Median :   25.0   Median :  435925   Median :  6.01   Median :4.811e+10   Median :  9387
 Mean   :2001   Mean   :  241.3   Mean   : 1857394   Mean   : 12.78   Mean   :4.517e+11   Mean   : 16959
 3rd Qu.:2008   3rd Qu.:  129.0   3rd Qu.: 1456899   3rd Qu.: 16.60   3rd Qu.:2.579e+11   3rd Qu.: 25191
 Max.   :2015   Max.   :22338.0   Max.   :43805214   Max.   :224.97   Max.   :1.812e+13   Max.   :126352
```

Except for the year which is skewed to the left more, each of the other variables are skewed to the right, meaning the mean here is larger than the median. For the ordered factor variables, age has 4.470 variables for all levels. The variables for Generation differ, it is presented below.

```
table(rate$age)
```

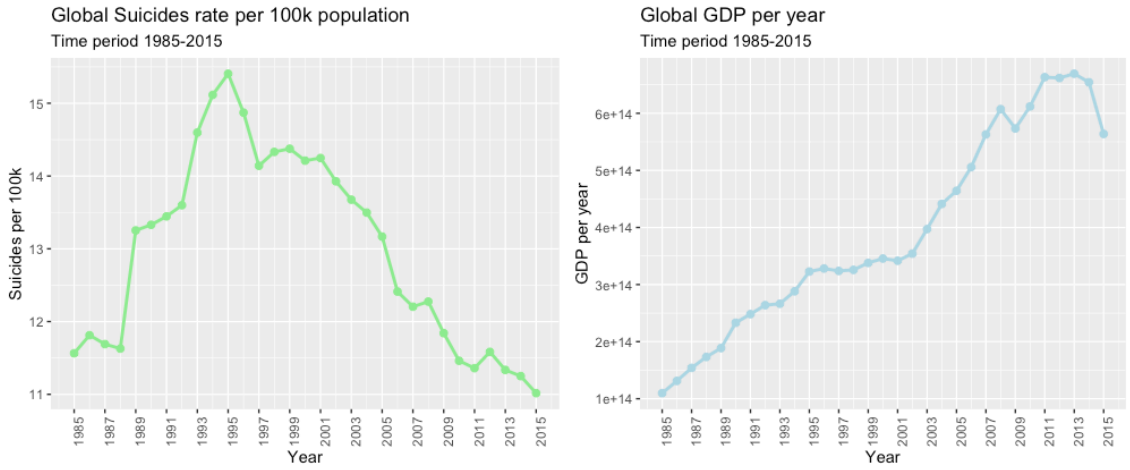
```
##
##  5-14 15-24 25-34 35-54 55-74   75+
##  4470  4470  4470  4470  4470  4470
```

```
table(rate$generation)
```

```
##
## G.I. Generation          Silent         Boomers    Generation X      Millenials
##            2654            6146            4810            6178            5610
##    Generation Z
##            1422
```

Explorative Data Analysis

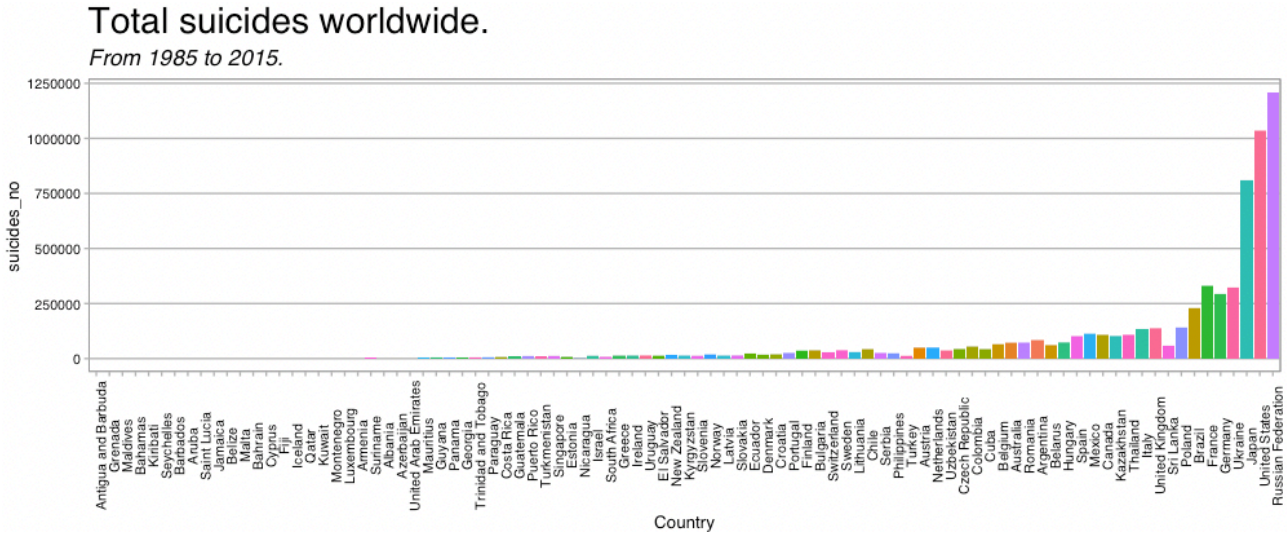
The data visualisation have been done with the ggplot package. The first diagram shows the global average suicide rate and the global GDP per year from 1985-2015. The suicide rate is calculated per 100.000 of population. It is interesting to see that there is a peak in 1995, however the global suicide rate seem to decline afterwards reaching its lowest point in 2015. It seems the suicide rate decreases as the global yearly GDP increases, however there is a drop in the GDP in 2015 which do not affect the rate of suicide.



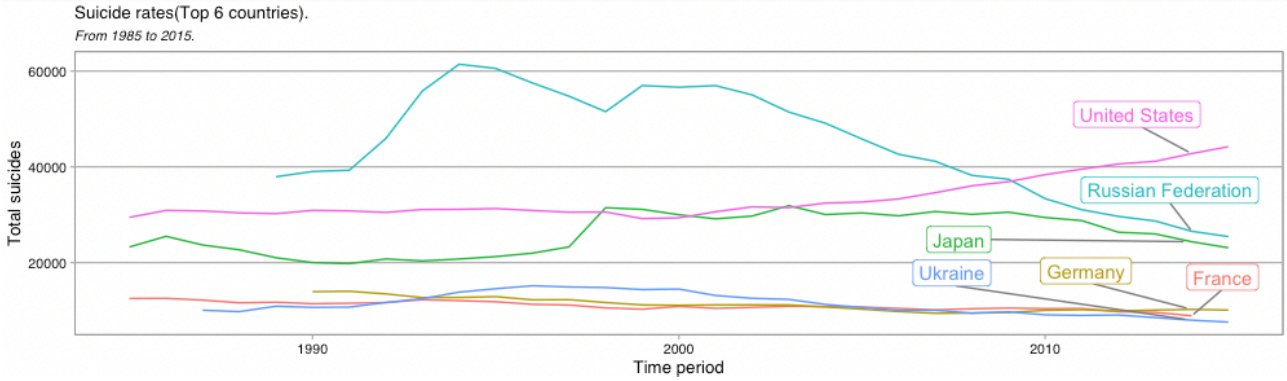
According to the WHO (1999) the peak can be due to the end of the USSR and that some of its former Republics started to report individually, therefore increasing the global rate.

The analysis further looks at the regions and their suicide numbers over 1985-2015. Europe has the highest rate of suicides over this time period followed by Asia, Oceania and Americas, while Africa has the lowest rate of suicide. However Liu (2009) states this could be due to that African countries do not collect morality data routinely or they do not send this to international organisations.

The countries number of suicide is analysed specifically as well. It can be seen that the Russian Federation has the largest number of suicides followed by the United States, Japan, Ukraine, Germany and France. The lowest suicide number is within Antigua and Barbuda, Grenada, the Maldives, the Bahamas and Kiribati, mainly smaller countries and islands.



Analysing the top 6 countries with the most suicides the lines do not seem to have a similar pattern except for Ukraine, Germany and France, where they tend to have a similar line where the suicide numbers seem fairly constant. By 2015 the United States have surpassed the Russian Federation in terms of suicide numbers, however up until around 2010 the Russia Federation had the higher number of suicides over the years. The number of suicide seem to fluctuate for countries over the years. It seems to have an increasing pattern for the United States. According

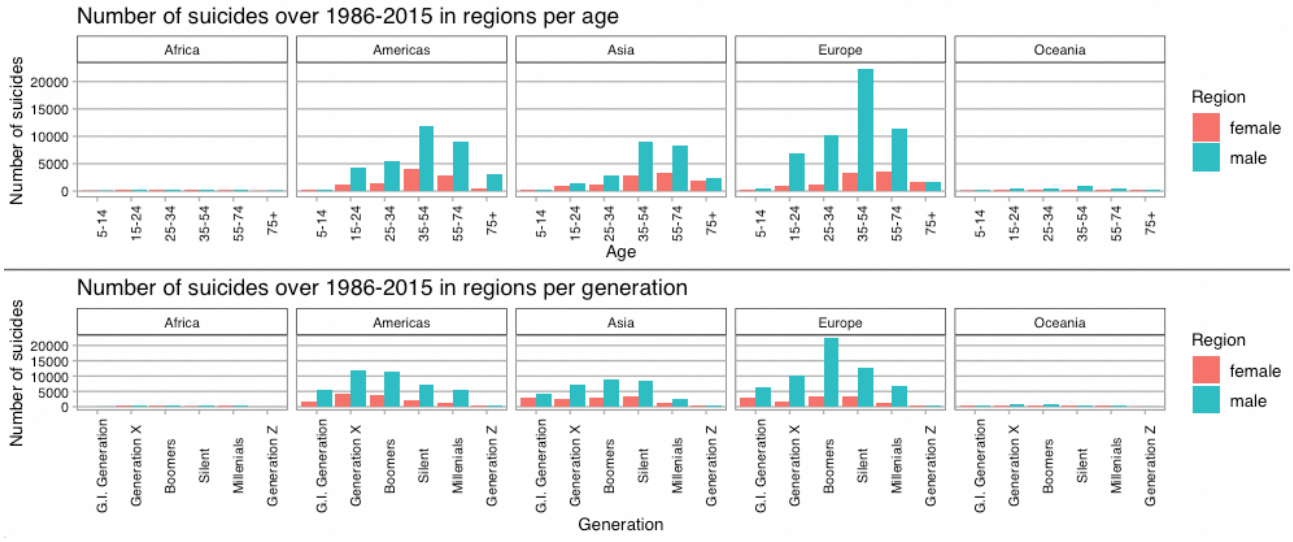


to Centers for Disease Control and Prevention (Stone et al., 2018), suicide rates in the US have

risen 30%. This can be due to mental health conditions, social and economic problems, access to lethal means such as alcohol, substances and firearms for people at risk and poor coping and problem solving skills. They add that suicide is often related to mental health conditions, however according to CDC´s study (Stone et al., 2018) half of suicide decedents do not have a registered mental condition, this could be due to the limited and expensive resources for mental health support.

According to Walsh (2003) the peak for Russia around 1995 is due to the social and economic problems that people failed to adapt to their current environment since the fall of the Soviet Union. The highest-risk groups are men unable to find career. Alcoholism is also an issue and is connected to a higher number of suicides.

The different age and generation groups are also analysed within the regions and between the two genders, male and female. In all regions people between 35-54 have the highest number for committing suicide. In Europe Boomers, which includes people born between 1946-1964, have committed the most suicide over the years as well as in Asia, and in the Americas region it is the Generation X, which includes people born within 1965-1980. Due to the increasing proportion of elderly in developing and developed countries people within the ages where they are able to work and pay taxes it committing suicide can be a major public health concern (WHO, 2014).



It can be seen that males significantly tend to commit suicides more. According to WHO (2014, p. 20) there are several reasons for this unbalancing data. It can be due to differences in dealing with stress and conflict that is socially acceptable for men and women, the availability of and preference for different suicide, patterns of alcohol consumption and differences between care-seeking rates for mental disorders between the two genders.

Cluster Analysis

To further analyse the data the Hierarchical Cluster Analysis method was used to identify clusters. A subset named as *rate_sg* has been created from the rate dataset containing solely the numerical variables. Missing values were omitted if any once again. A dendrogram have been plotted which was then set to have five clusters. Since within the dataset we have used five regions, the cluster analysis was set to look for five clusters. This have been visualised with the Cluster plot.


```
table(rate_clusters, rate$region)
```

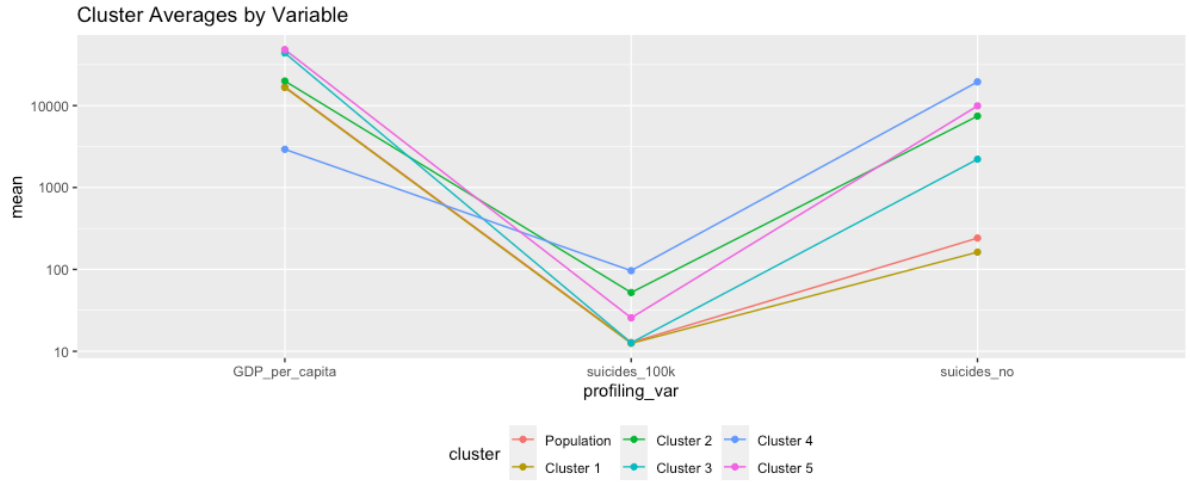
```
##
## rate_clusters Africa Americas  Asia Europe Oceania
##             1    828     8554  4841  11184     972
##             2      0       24    50     70       0
##             3      0      251     5      0       0
##             4      0        0     0     14       0
##             5      0       27     0      0       0
```

All of the five clusters have later on been analysed by creating a plot of them to see the differences within the groups. These differences have been analysed on the country, age, generation, gender and GDP per capita level.

1. *Cluster one* contains almost all of the countries from all regions, however it has most countries from the European region. It contains all ages from ranging from 5 to over 75 years olds. It contains the G.I Generation, Silent Generation, Boomers, Generation X, Millenials and Generation Z. It contains both female and male observations. It contains data where the GDP per capita is ranging up to 125.000 USD. This is the biggest cluster.
2. *Cluster two* contains data for Russia, Japan and the United States. Its contains ages ranging from 15 to 54 years olds. It contains data for the G.I Generation, the Silent Generation, Boomers and Generation X and Millenials. It only has one gender, male. The GDP per capita here for countries ranges until 50.000.

3. *Cluster three* contains data for Japan and the United states, the age group here is from 5 to 75 and more years olds. It contains observations for the G.I Generation, Silent Generation, Boomers, Generation X, Millenials and Generation Z. Both genders are represented in the cluster and its GDP per capita ranges until 60.000 USD.
4. *Cluster four* contains solely the Russian federation, ages from 35-54, the Boomer Generation, males and GDP per capita until 5.000 USD.
5. *Cluster five* contains the United States, ages ranging from 35-74, which includes the Silent Generation, Boomers and Generation X. It contains data for males solely and the GDP per capita here ranges until 60.000 USD.

When looking at the means of the different clusters a new row is introduced which contains the summarisation of the population's average. The data is then represented with a snake plot, however GDP per year is taken out for an enhanced view which gives us the following diagram.



The clusters seem to follow a similar pattern, however their means across the variables are rather different. Cluster three, population and cluster one seems to have the lowest mean for suicide rates followed by cluster five, cluster two and cluster four, which means that in general the highest number of suicide can be found in the Russian federation between the ages of 35 and 54, specifically in the Boomer Generation, when the GDP per capita was solely 5.000 USD and within males.

Results

The information obtained from the data's visualisation through different diagrams and clustering shows that the highest number of suicides can be found in countries within Europe, more specifically in Russia. Most of the suicides occurs amongst males. Boomers and Generation X has the highest number of suicide in regions. Suicide happens within the age group from 35 until

54 years olds. Lower amount of GDP per capita and per year is usually in connection with a higher rate of suicide, however not in all cases, such as the United States.

Discussion & Future Work

The analysis showed us that while there are similarities between the different regions and their suicide rates compared to their GDP per capita and yearly GDP, it was possible to identify five different clusters. It could be seen that the first cluster contained most of the data but in the other ones Japan, the United States and Russia had values mostly. From the visual analysis it could be seen that these three countries have the highest number of suicides across the 1985-2015 period. While for Russia it has been declining since the 2000s, and for Japan it has been sort of constant, for the United States the suicide rates have been increasing and around 2010 it had more suicides than Russia.

The paper provides a base for further research which could take into consideration the different social an economic factors amongst regions and countries. It would be interesting and beneficial to look at a country's happiness index, its quality of healthcare and education system. It can also serve as guideline for preventing suicide by putting more focus on men and ensuring they get the help they need. Furthermore, to focus on people between 35-54 years.

According to WHO (2014, p. 48) knowledge about behaviour related to suicide has been increasing. Risk and protective factors have been identified and it has been found out that cultural variability can be a suicide factor as well. There are now suicide prevention strategies on a national level. Information about suicide and its prevention has also been increasing. Self-help groups have been established and telephone counselling is provided when needed through trained volunteers.

Limitations of the dataset/work

Limitations for the dataset in general lies within the missing values. The hierarchical clustering has disadvantages as well, such as the difficulty of decision making as it has a great impact, which is why it takes time to look for the most useful solution. Validating a cluster can also be difficult as there are a number of techniques, however there is no consensus on a single best approach. Assigning each observation a cluster might not be appropriate at all times, due to outliers the clusters may be heavily distorted. Due to the dataset being quite big it would be interesting to break it down in the future to more subsets and analyse them separately. The clustering had challenges due to the big amount of data, however this research wanted to analyse the dataset as a whole and to look for the regions and the differences or similarities between them specifically.

Bibliography

Schmidtke, Weinacker, B., Apter, A., Batt, A., Berman, A., Bille-brahe, U., Botsis, A., Leo, D. D., Doneux, A., Goldney, R., Grad, O., Haring, C., Hawton, K., Hjelmeland, H., Kelleher, M., Kerkhof, A., Leenaars, A., Lönnqvist, J., Michel, K., … Wasserman, D. (1999). Suicide rates in the world: Update. *Archives of Suicide Research*, 5(1), 81–89. https://doi.org/10.1080/13811119908258318

WHO., Saxena, S., Krug, E. G., & Chestnov, O. (2014). *Preventing Suicide A Global Imperative. World Health Organization*.

Liu. (2009). Suicide Rates in the World: 1950-2004. *Suicide & Life-Threatening Behavior*, 39(2), 204–213. https://doi.org/10.1521/suli.2009.39.2.204

Wickham, & Grolemund, G. (2017). *R for data science* . O'Reilly Media.

James, Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to statistical learning : with applications in R* (2nd ed.). Springer.

WHO. (1999). *Figures & Facts About Suicide*. World Health Organization.

Stone, Simon, T. R., Fowler, K. A., Kegler, S. R., Yuan, K., Holland, K. M., Ivey-Stephenson, A. Z., & Crosby, A. E. (2018). Vital Signs: Trends in State Suicide Rates — United States, 1999–2016 and Circumstances Contributing to Suicide — 27 States, 2015. MMWR. *Morbidity and Mortality Weekly Report*, 67(22), 617–624. https://doi.org/10.15585/mmwr.mm6722a1

Walsh, Nick Paton. (2003, July 9). *Russia's suicide rate doubles*. The Guardian. https://www.theguardian.com/world/2003/jul/09/russia.nickpatonwalsh

Appendices

Appendix A