Suicide Analysis

Jinal Butani
Department of computer science
University of North Carolina
Charlotte, USA
jbutani@uncc.edu
801077913

Abstract— **The goal of this paper is to do an in-depth analysis of the suicides occurs in all over the world. the suicide ratio is increasing year by year. WHO says that in every 40 sec one human dies by suicide. There should be a suicide prevention program. This survey paper is beneficial for the world health organization for suicide prevention. They can analyze it and make the program according to the city, age group, and sex. My main purpose it that whoever is accessing this for suicide prevention have in-depth information and it should also help them to make an important decision. There are many suicide prevention programs is working on but sometimes they fail because they don't analyze the past data. There are many factors which affect this. They never connect the suicide information with the country's GDP. This project will conduct regression to check the statement "richer the country, higher the suicide rate". I selected this dataset from Kaggle. It covers many countries. This data will allow me to create some visualization in tableau. This will provide insight into human behavior. the overall dashboard will help WHO in suicide prevention. [3]**

Keywords—suicide, suicide prevention, mortality, WHO, GDP per capita, Human Development index

I. Introduction

Suicide is one individual act. Aggregation of total suicide may have a small number for one country.in the longer run large changes can happen and indeed have happened. [1] WHO is working over the past 26 years for this. They improved the mortality rate also. According to me, they have a shortage of such information for many countries. nowadays as per one analysis suicide rate is decreased from 15.3 to 11.6 per 100000 people. Suicide is preventable.

According to WHO, Lithuania and Japan are at the top of the suicide rate. The rates are changing. It can be easily said that mortality rates are shifting from western Europe to Eastern Europe. the most populated counties India and China are also the biggest contributors in this.

This type of behavior creates major mental health problems. In many countries, the suicide rate is higher than the accident death rate. in some countries, some particular age group is more likely to commit suicide. one report says that it becomes the major health concern in many developed and developing countries. in article mental health by WHO, they mentioned that 800k people die every year.

Suicide prevention is difficult when we see the rates in the world, but it is not impossible. if we implement effective interventions on populations and individuals then it can help us to prevent suicide and even suicide attempts. IT is right that WHO also have to work on the prevention of suicide attempts because their report says that every single suicide has an average of 20 suicidal attempts. they also give a connection between suicide and the country by that 79% of the suicide occurs in low-income countries. common methods of suicide are pesticide, hanging, and firearms.

As suicide is a serious health problem, but it is also true that suicides are also prevented over time.[2] for the effective response of the country's prevention strategies is needed. I can say that the main objective is to conduct a deep analysis of the suicide data provided by the WHO. From this survey paper, every presentation organization will have a clear idea that which country needed which kind of policy. The country in which younger people are more likely to commit suicide. should have a policy like treatment for mental problems. The country should have reduced access to the means of suicide like certain pesticide. Next, I have applied various Regression models to analyze the data. We have used Multiple Linear Regression and Regression Tree for the above purpose.

I also connect the GDP and the suicide rate to take the answer to the question that is countries with lower income has a high suicide rate?

II. Dataset

Suicide Rates Overview 1985 to 2016 was collected from the Kaggle. This data was used for making visualization and making knowledge out it for suicide prevention. It will help the WHO for making suicide prevention program for a particular city or generation or sex. Results are derived using various data processing techniques. All the parameter is very crucial regarding the suicide rate in a particular country. The database contains 12 columns: country, year, sex, age group, count of suicides, population, suicide rate, country-year composite key, HDI for a year, GDP_for_year, GDP_per_capita, generation. Detail description of all the parameter follows: [4]

- country: It defines the country name. There are 101 countries from 5 continent (Africa, the Americas, Oceania, Asia, Europe).

- year: It defines the year in which suicide occurs. there are the data from 1985 to 2016.

- sex: It defines the gender of victims. It may be male or female.

- age group: It defines the age group of victims. There is 6 age group in this database. 5-14, 15-24, 25-34, 35-54, 55-74 and 75+

- count of suicides: it is a total suicide number of victims for a particular gender and age group.

- population: It the total population of a particular age group in the country.

- the suicide rate: It shows the suicide rate per 100k population. This column is important because analyzation becomes easy with small numbers.

- country-year: This column is a combination of the country name and the year.

- HDI for a year: HDI stands for Human development index. this is the HDI of the country for that particular year.

- gdp_for_year: Gross domestic product is one of the most common indicators used to track the health of a nation's economy. It includes a number of different factors such as consumption and investment.

- gdp_per_capita: GDP per capita 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.

- generation: In this suicide dataset, you will see different generation values indicating the born year. Roughly speaking, the generation means = 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

Before preprocessing the dataset has a total of 27,821 records. It will be further preprocessed to achieve the final dataset.

A. *Pre-Processing of Data*

The data was thoroughly analyzed, and various functions were used for pre-processing the dataset. The dataset was imported into python data frame. “pandas.read_csv()” was used for importing the data.

Data cleaning

1. There are 7 county which have less than 3 years of data. This is negligible compare to other data, so I removed the 7 countries.

2. In this database, there are data from 1985 to 2016. For particular 2016, some countries don’t have data and other have many missing fields, so I removed the 2016’s data.

3. HDI (human development index) was removed. Most of the countries don’t have the HDI. The ratio is around 2:3.

4. Country-year is irrelevant field as there are separate country and year field present. I removed this unnecessary felid.

5. When we are talking about data around the world. Continent plays very important role. So, added the continent. I divide all the given country in 5 continents (Africa, the Americas, Oceania, Asia, Europe)

6. After adding the continent field, I found that There is not much information about the countries of Africa.

7. I found that generation column has many problems as it overlaps with the age group. This leads to wrong conclusions, so I did not use generation for my any result.

8. I found that suicide rate is not that much useful for my representation, so I removed it.

After cleaning:
- Number of rows: 27,492
- There will be 9 columns: country, year, sex, age, suicide_no, population, gdp_for_year, gdp_per_capita, continent

III. BUSINESS USE

The objective of this project is to prepare dataset which helps WHO for starting the suicide prevention program. There are lots of factors affect when one person is committing suicide. increasing the suicide rate is not good for the country and society. There are many organizations who are working on this. Reasons for suicide also vary from age group, sex, income, etc. They also require different prevention program. nowadays the younger generation most likely to suffer from depression so for them anti-depression programs will help. In some countries, there is gender inequality. due to gender equality, the female becomes more dependent and at the result, she becomes the victim of Domestic violence. in this project, I will work on each parameter separately. my data mining project will classify the output. from all the analysis, we can easily see which country, which sex and which age group needed the most help. The outcome of this project will help WHO or any organization who works for better human life for any decision making

IV. DATA VISUALIZATION

A. *Heat Map: -*

A heat map is a graphical representation of data. In the heat map, charts are represented as colors. Higher or density of the record is more intense color it has. It can be represented in highlighted table or geographical map.

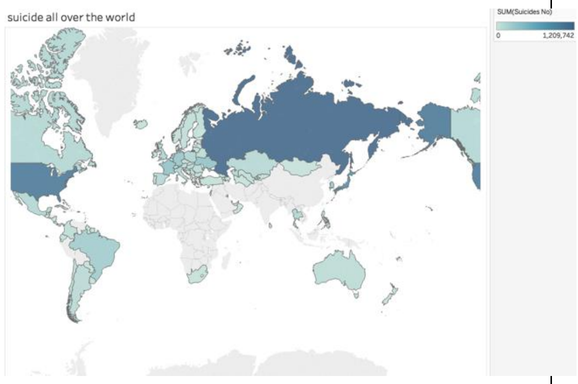

Figure 1: Geographical Heat Map for representation of Total suicide all over the world

figure 1 clearly represent that countries in Europe continent has more suicide cases in past years.

B. Tree Map: -

The tree map displays data in nested rectangles. The dimensions define the structure of the tree map and measures define the size or color of the individual rectangle. The rectangles are easy to visualize as both the size and shade of the color of the rectangle reflect the value of the measure.

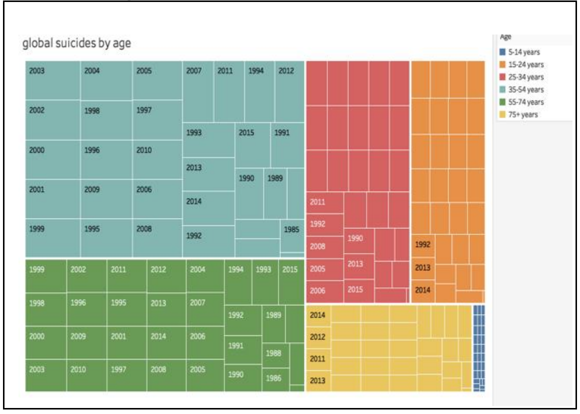

Figure 2: Tree map for representation of total suicide by age group and year

There are 6 age groups. From the above figure one can analyze that people in age group 35-54 are committing suicide more compare to other age group.

C: Bar chart: -

Bar chart is a chart that represents the categorical data in rectangular bars. Height proportional to the values they represented. It can be vertical or horizontal depends on need. In the bar chart, one axis represents all the categories and other represents the value.in most of the cases, categories are discrete.[5]

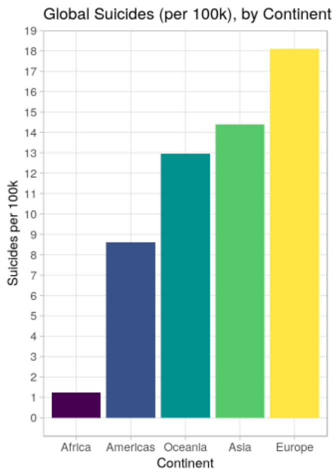

Figure 3: division of total suicide by continent

If I divide all data by the continent in bar chart, I shows that Africa has a smaller number of the suicide rate. This might be leads to bias decision as data not contains all the countries in Africa. As well as data has no record of the most populous countries India and China

D: line graph:

Line Charts are primarily used for showing time-series data. They can be used to show changes in values of quantitative data over a short or a long period of time, e.g. data about changes in chances of admit with change in examination scores.

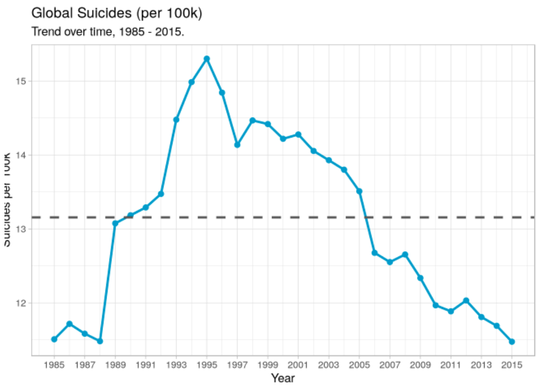

Figure 4: suicide over time 1985-2015

In 1995, there are highest rate of the suicide over the world. Due to development, this rate is decreasing

E: Area Chart:

Area Chart represent the one or more qualities over time. IT is similar to line chart. Area between the X axis and line are filled with colors. Staked Area chart might be good choice when one wants to compare the more than two qualities. Through this one can easily see the difference.

Here I user the year, total suicide number and sex as parameter.

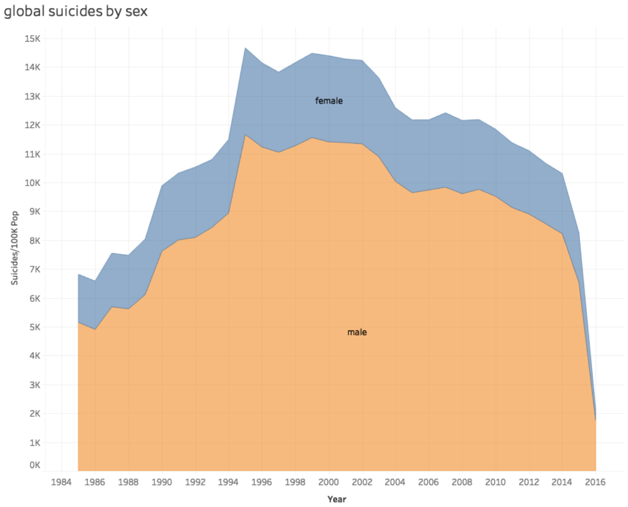

Figure 5: Area chart for representation of suicide rates in men and women over time

Figure 5 shows that suicide rate for men is higher (around 3.5 times more). In 1995 both men and female have high suicide rate. Ratio of man: women suicide rate is increasing. Thought is remained constant in 90’s

V. ALGORITHMS

A. Multiple linear regression

Multiple linear regression is the one form of linear regression analysis. Multiple linear regression is a statistical technique that uses explanatory variables to predict the output of a response variable. Basically, the Main goal of the linear regression is to model the linear relationship between the independent (explanatory) variables and dependent(response) variable. One can identify outlier with the use of Multiple linear regression [10]

C. Regression Tree

A regression tree/decision tree is a process which is built through as binary recursive partitioning. This is an iterative process that splits the data into partitions or branches. This process continues splitting each partition into smaller groups as the method moves up each branch. Since the target variable does not have classes, we fit a regression model to the target variable using each of the independent variables [9]

VI. GDP AND SUICIDE RATE

One question arises when I see the data of GDP in the dataset that how suicide rate and the country’s financial condition are connected? Research shows that its completely depends on the country. In last 25 years, every country’s GDP is linearly increasing. By calculating the correlation between year and GDP, I found that they have linear relationship. The correlation was 0.877. [13]

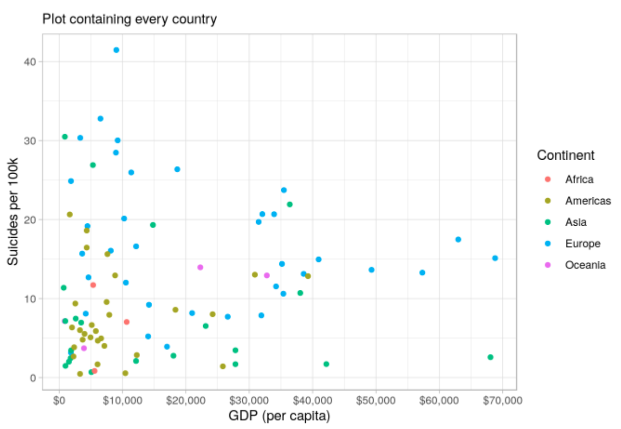

Figure 6: scatter plot of all countries GDP and suicide Rate

In the Figure 6 we can see that I took all the countries and make one scatter plot. I divided the countries into their continent. Purpose for this scatter plot is to see how they are connected.

I removed the outliers. P value of the Model is 0.0288. which is less than 0.05. so, it is proven that GDP is associated with the suicide rate. We cannot completely ignore it.$r^2 = 0.0544$ GDP per capita has very little variance in suicide rate.

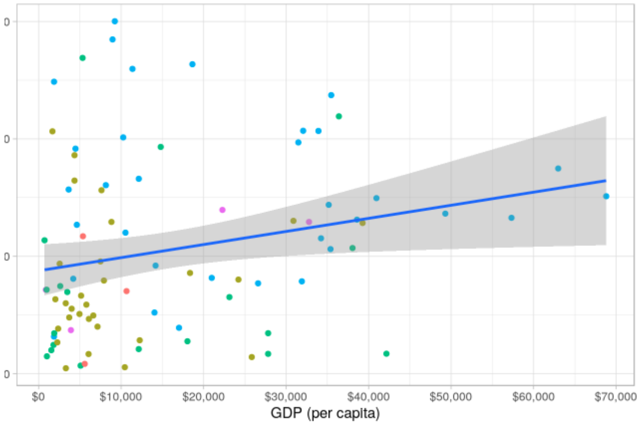

Figure 7:scattle plot with outlier removed

$$\text{Suicides}=8.7718+0.1115*\text{GDP}$$

From the regression line I can say that the increase GDP by $8,967 there was additional one suicide per 100k people per year.

VII.CODE

For scatter plot of all countries GDP and Suicide Rate: [15]

```
country_mean_gdp <- data %>%
  group_by(country, continent) %>%
  summarize(suicide_per_100k =
(sum(as.numeric(suicides_no)) /
sum(as.numeric(population))) * 100000,
            gdp_per_capita = mean(gdp_per_capita))

ggplot(country_mean_gdp, aes(x = gdp_per_capita, y =
suicide_per_100k, col = continent)) +
  geom_point() +

scale_x_continuous(labels=scales::dollar_format(prefix="
$"), breaks = seq(0, 70000, 10000)) +
  labs(title = "Correlation between GDP (per capita) and
Suicides per 100k",
       subtitle = "Plot containing every country",
       x = "GDP (per capita)",
       y = "Suicides per 100k",
       col = "Continent")
```

Correlation between GDP and Suicide Per 100k: [15]

```
ggplot(gdp_suicide_no_outliers, aes(x = gdp_per_capita,
y = suicide_per_100k, col = continent)) +
  geom_point() +
  geom_smooth(method = "lm", aes(group = 1)) +
scale_x_continuous(labels=scales::dollar_format(prefix="
$"), breaks = seq(0, 70000, 10000)) +

 labs(title = "Correlation between GDP (per capita) and
Suicides per 100k",
      subtitle = "Plot with high CooksD countries removed
(5/93 total)",
      x = "GDP (per capita)",
      y = "Suicides per 100k",
      col = "Continent") +
 theme(legend.position = "none")
```

VIII. CONCLUSION

WHO realize that prevention of suicide should be the public health priority. They also provide the suicide report in 2014. so that they increase the awareness of public health. they think that suicides and suicide attempts are major problems globally Due to their efforts the reports are getting better for mortality. WHO claims that mortality rate is decreasing for eastern Europe. The suicide rate is increasing in eastern Asia. [14]

- The suicide rate is decreasing
- Suicide rate have correlation with GDP that every increase in around $8500 suicide is increase by 1 per 100 k
- 44.5% of total suicide occurred between 1996 and 2005
- Male at their middle age in Europe are more likely to commit suicide.

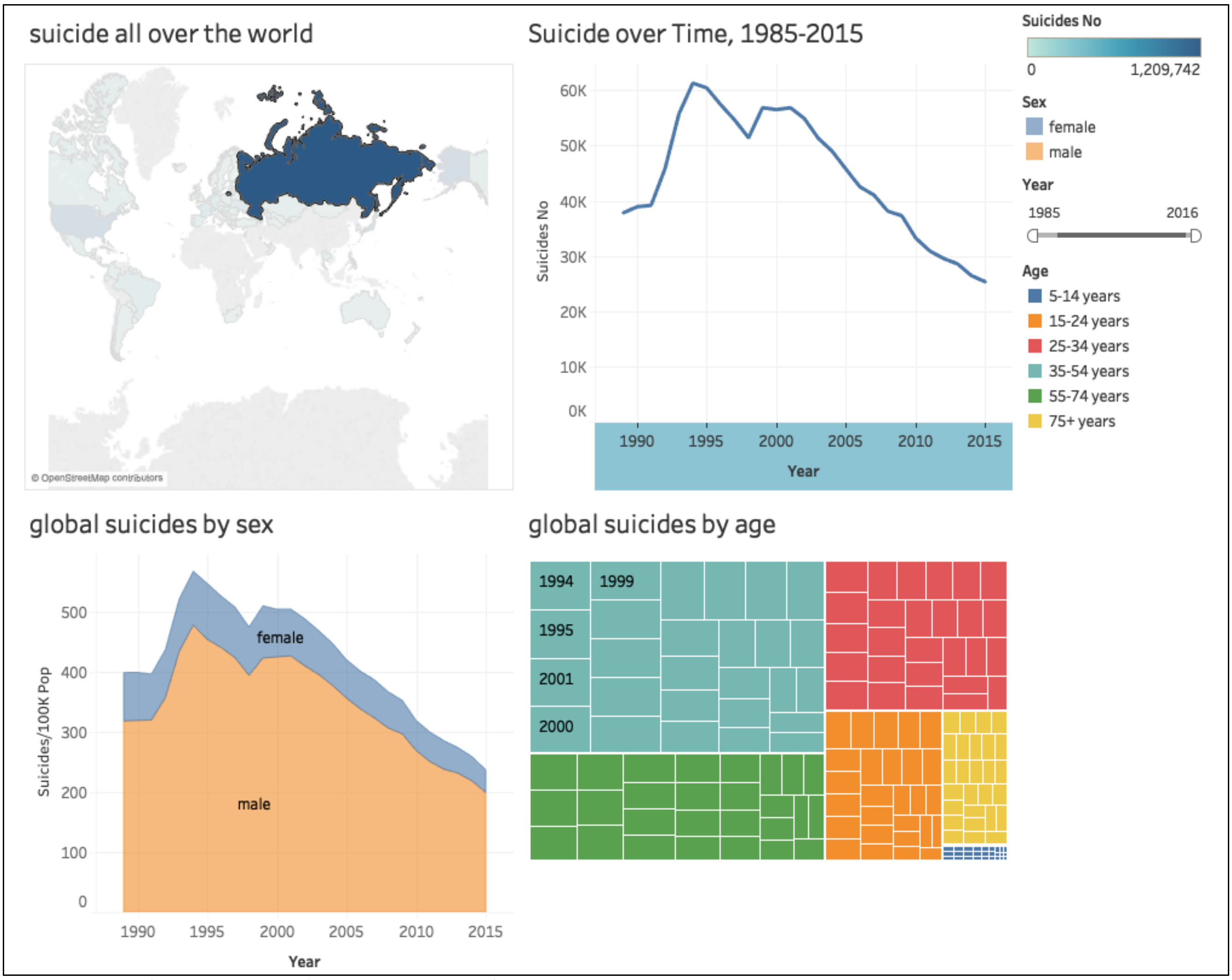

figure 8: workbook for the suicide analysis for continent with highest suicide rate and other data according to that

REFERENCES

[1] World Health Organization *mental health suicide data*. Available: https://www.who.int/mental_health/prevention/suicide/suicideprevent/en/

[2] Schmidtke, A., Weinacker, B., Apter, A., Batt, A., Berman, A., Bille-Brahe, U., ... & Grad, O. (1999). Suicide rates in the world: update. Archives of Suicide Research, 5(1), 81-89..

[3] Värnik, P. (2012). Suicide in the world. International journal of environmental research and public health, 9(3), 760-771..

[4] Kaggle data set. *Suicide rates overview 1985-2016*. Available: https://www.kaggle.com/mohansacharya/graduate-admissions

[5] Sankhe-Savale, S. (2016). Tableau Cookbook–Recipes for Data Visualization. Packt Publishing Ltd.

[6] United Nations Development Program. (2018). Human development index (HDI). Retrieved from http://hdr.undp.org/en/indicators/137506Sankhe-Savale, S. (2016). Tableau Cookbook–Recipes for Data Visualization. Packt Publishing Ltd.

[7] World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016. Retrieved from http://databank.worldbank.org/data/source/world-development-indicators#

[9] I. Frontline Systems. (2019, April 20, 2019). *Regression Trees*. Available: https://www.solver.com/regression-trees

[10] K. J. Preacher, P. J. Curran, and D. J. Bauer, "Computational Tools for Probing Interactions in Multiple Linear Regression, Multilevel Modeling, and Latent Curve Analysis," *Journal of Educational and Behavioral Statistics,* vol. 31, no. 4, pp. 437-448, 2006/12/01 2006.

[11] [Szamil]. (2017). Suicide in the Twenty-First Century [dataset]. Retrieved from https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook

[12] T. Point. (2019, April 23, 2019). *Tableau - Tree Map*. Available: https://www.tutorialspoint.com/tableau/tableau_tree_map.htm

[13] B. Leiva. (2016, April 23, 2019). *Creating Scatter Plots in Tableau*. Available: https://www.thedataschool.co.uk/borja-leiva/creating-scatter-plots-tableau/

[14] World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/

[15] https://www.kaggle.com/lmorgan95/r-suicide-rates-in-depth-stats-insights/report#comparing-the-uk-ireland-america-france-denmark