2021

Suicide Rate 1985 - 2016

BAN140ZAA FINAL PROJECT

PROFESOR: OMAR ALTRAD

Leslie Stefany Lopez Espinoza ID 133803205

Punn Prateek Singh ID 146980180

Prathap Reddy Padigapati ID 128942208

Aysegul Turk ID 128176203

Contents

Introduction

Personal Objective

In the visualization project, we have chosen the Suicide Data Set for the years 1885 to 2016. We are working for a not-for-profit Mental Health Service Company, whose goal is to identify the priority suicide risk groups worldwide in order to decide in which country and demographic the project should focus on.

The main objective of this project is to visualize and find a relatable pattern in the data to get a better understanding of it and use the findings to answer our business question. We will focus on the suicide rate (Suicide/100K population) to deliver our analysis for this project and every finding will be shown related to this metric.

Intended Outcomes

Our analysis is expected to find the most affected by suicide per country, age, gender & generation. We will review suicide rates and the independent variables over time as well as make comparisons between different countries, age groups, gender, and GDP rates, etc. to get a clear picture of the factors that are actually associated with suicide globally.

Description of the needs of the intended audience

It is our goal to show great explanatory data visuals so that everyone in the audience gets the actual idea behind our work. Our classmates would like to understand the reasoning behind our recommendation to the Not-for-profit Mental Health Company that we are working for.

Foreseeable Challenges

The data set has to be prepared before importing to SAS for data cleaning because some variable names are not compliant with SAS syntax. The biggest problem that we see in the raw data set is the incorrect description of variable generations.

There are some other challenges in the project as well such as missing values in the variable HDI and data redundancy therefore we have used mean values to make accurate visualizations.

Description of the dataset

The data set selected is the suicide data set, which has 27820 observations and 12 features, out of which five are numerical and the rest are categorical.

The target variable is the suicide number “suicide_no”. The independent or predictor variables are country, year, sex, age, population, country_year, HDIforyear, gdp_for_year, and generations. The data set also has 2 derived variables, suicide_100pop and gdp_per_Capita, which are calculated by dividing suicide_no and GDP_for_year by population.

The following is the description of each variable and its type in the data set, which shows gdp_for_year and HDIforyear as char type, the latter with length 1.

#	Variable	Type	Len	Format	Informat
1	country	Char	7	$7.	$7.
2	year	Num	8	BEST12.	BEST32.
3	sex	Char	6	$6.	$6.
4	age	Char	11	$11.	$11.
5	suicides_no	Num	8	BEST12.	BEST32.
6	population	Num	8	BEST12.	BEST32.
7	suicides_100k pop	Num	8	BEST12.	BEST32.
8	country_year	Char	11	$11.	$11.
9	HDIforyear	Char	1	$1.	$1.
10	gdp_for_year	Char	15	$15.	$15.
11	gdp_per_capita	Num	8	BEST12.	BEST32.
12	generation	Char	15	$15.	$15.

We also can see a listing of the first 10 observations where we can see a significant number of missing values for HDIforyear and inconsistencies bet. age and generation.

Listing the First 10 observations of Data Set Suicide Rates 1984-2016

country	year	sex	age	suicides_no	population	suicides_100k pop	country_year	HDIforyear	gdp_for_year	gdp_per_capita	generatio
Albania	1987	male	15-24 years	21	312900	6.71	Albania1987		2,156,624,900	796	Generatior
Albania	1987	male	35-54 years	16	308000	5.19	Albania1987		2,156,624,900	796	Silent
Albania	1987	female	15-24 years	14	289700	4.83	Albania1987		2,156,624,900	796	Generatior
Albania	1987	male	75+ years	1	21800	4.59	Albania1987		2,156,624,900	796	G.I. Gener
Albania	1987	male	25-34 years	9	274300	3.28	Albania1987		2,156,624,900	796	Boomers
Albania	1987	female	75+ years	1	35600	2.81	Albania1987		2,156,624,900	796	G.I. Gener
Albania	1987	female	35-54 years	6	278800	2.15	Albania1987		2,156,624,900	796	Silent
Albania	1987	female	25-34 years	4	257200	1.56	Albania1987		2,156,624,900	796	Boomers
Albania	1987	male	55-74 years	1	137500	0.73	Albania1987		2,156,624,900	796	G.I. Gener
Albania	1987	female	5-14 years	0	311000	0	Albania1987		2,156,624,900	796	Generatior

Statistics for Numerical Variables

When running Proc. Means on SAS to create a statistics table for numerical variables, we can see that there are no missing values in the numerical variables.

The mean for suicide_no is 242.57 over the last 30 years and the maximum suicide number is 22338. This very high std deviation, suggests the presence of outliers. Therefore, we will use suicides_100kpop to work with a more smooth metric that considers also the country population. We have the same situation with GDP_per_capita, which spreads between $251 and $126352. However, because GDP has a high variability per country due to size and other socio-economic realities, we will continue working with it as it is.

Descriptive Statistics

The MEANS Procedure

Variable	N	N Miss	Mean	Std Dev	Median	Minimum	Maximum
year	27820	0	2001.26	8.47	2002.00	1985.00	2016.00
suicides_no	27820	0	242.57	902.05	25.00	0.00	22338.00
population	27820	0	1844793.62	3911779.44	430150.00	278.00	43805214.00
suicides_100k pop	27820	0	12.82	18.96	5.99	0.00	224.97
gdp_per_capita	27820	0	16866.46	18887.58	9372.00	251.00	126352.00

Statistics for Categorical Variables

We ran Proc Freq to find out the balance of each categorical value and found the following:

- Human Development Index (HDI) has around 19456 missing values and length is rounded to 1 by SAS in the frequency table.
- Age groups are equally distributed with frequency 4642 and Gender is also equally distributed with frequency 13910. It means the dataset is balanced in terms of age groups and Gender.
- Given the errors in Generation values, GenerationX has the highest frequency with a 6406 count.

Challenges

While doing the descriptive statistics, we found some challenges like data inconsistencies and missing values in some of the features. Therefore, we performed the following cleaning tasks:

Generation values were mapped with the correct age group to remove the inconsistencies.

HDIforyear has 19456 missing values. Therefore, we are not considering it appropriate for analysis.

GDP_for_year and HDI were changed to numeric type in Tableau, as were Char type in the original dataset.

Exploratory analysis of data

In order to find out whether there is any relationship between the dependent variable suicide_no and gdp_per_capita or gdp_for_year, we generated scatter plot matrixes on SAS, which can be seen in Appendix 1. In this report we are showing the matrix of scatterplots of variables suicide_no, suicide_100kpop, and gdp_per_capita only in Figure 1 as follows:

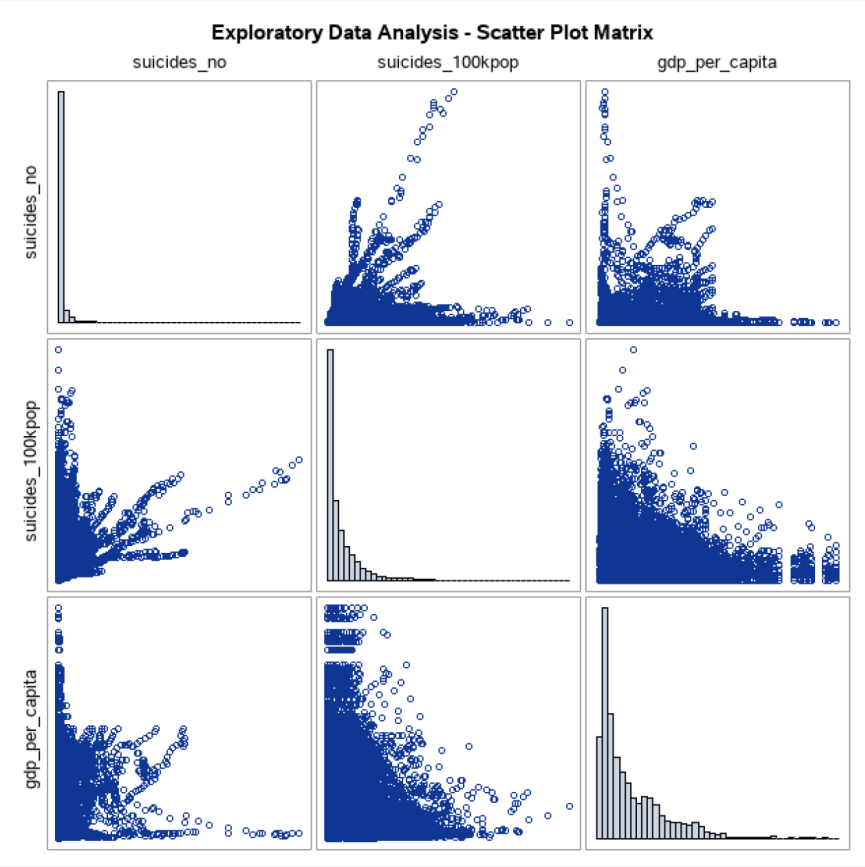

Figure 1

As seen in figure 1, we can deduct the following:

- There is an association between GDP_per_Capita and Suicide_no or Suicide_100pop
- The variables seem to follow a logarithmic distribution (non-linear)

Therefore, as mentioned earlier in this report, we will focus on analyzing Suicide_100pop as a derived variable from Suicide_no because it provides a better fit for analysis as shown in the matrix (less noise in data).

Besides, in order to verify the relationship between GDP and suicide rates, we created a visualization on Tableau showing the scatterplot of gdp_for_year and suicide_100kpop, which you can see below in Figure 2. The size of the circles represents the magnitude of the rate. The bigger the circle the higher the suicide_100Kpop rate.

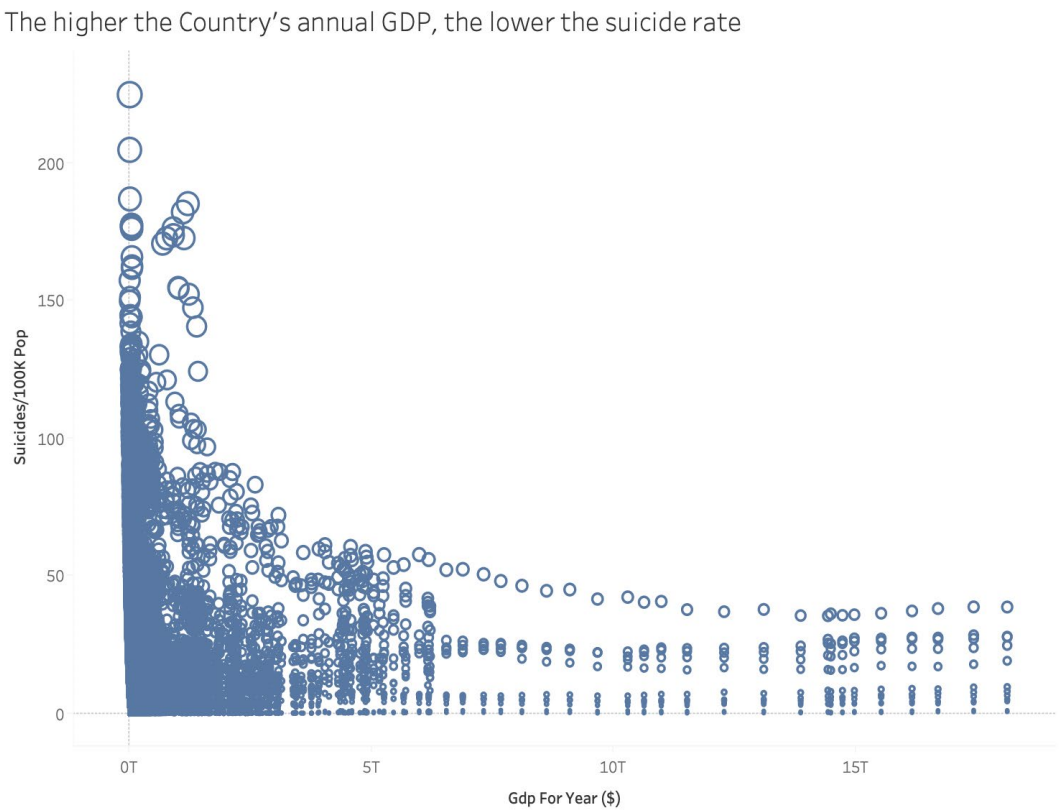

Figure 2

Here, we can clearly see that annual GDP and suicide rates are associated inversely. This means that as annual GDP increases, suicide rates decrease and vice versa. As we will see soon in our project, countries with high annual GDP recorded low suicide rates.

In order to see this association over 30 years, we ran a time series visualization for variables gdp_per_capita and suicide rate, as seen in Figure 3 as follows:

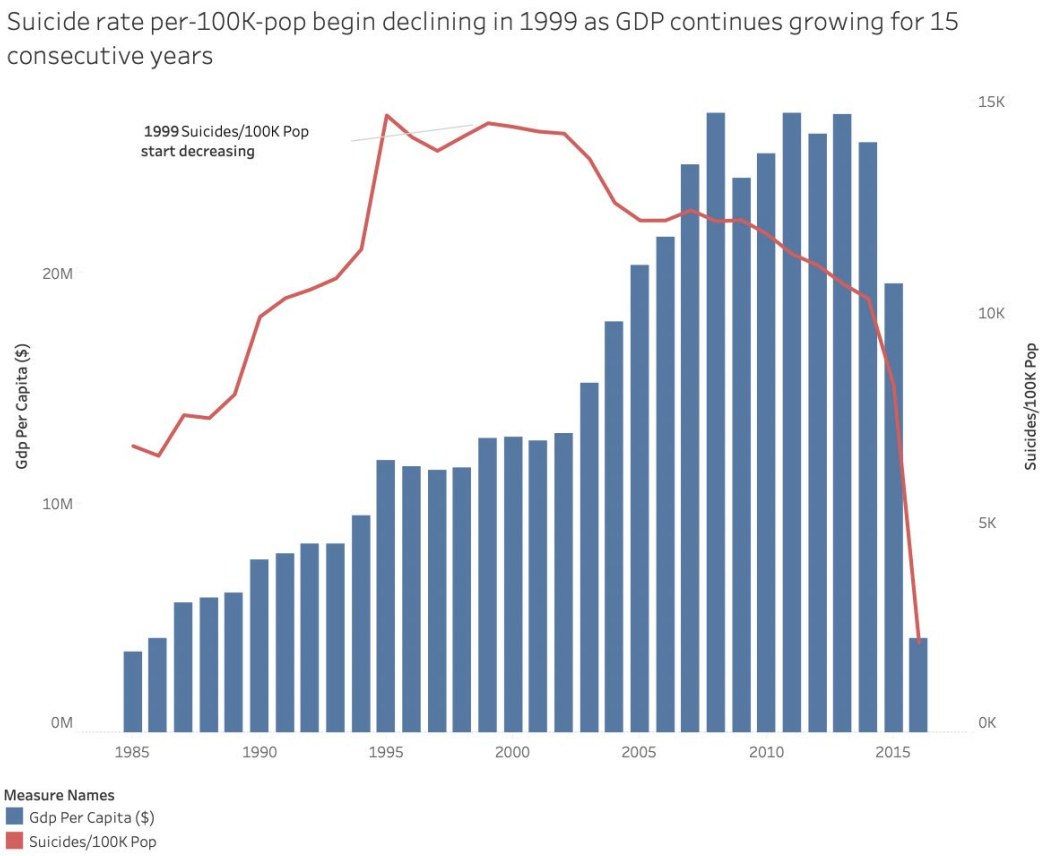

Figure 3

With Gdp_per_capita represented in blue bars and suicide rates as a red trend line, we see that in the year 1999, after 15 years of consecutive GDP growth, the suicide rates started declining.

Is it important to mention that there is a clear inconsistency in the year 2016, which shows the presence of a data outlier, which we did not remove when cleaning the dataset. Therefore, we are aware that these data points may appear in our visualizations.

In conclusion, we can say that there is an actual association between GDP and suicide rates. We will continue analyzing other independent variables like gender, generation/age, and country next.

From analyzing the frequency of suicide rates per gender and generation, as shown in Figure 4 below, we found out that, globally between 1985 and 2016, there was a **clear disparity in suicide rate between genders**, being the male suicide rate at least 4 times the female rate.

The generation with a higher suicide rate worldwide was the G.I. generation (the greatest generation), which is the generation that had more than 75 years in 1984. This **generation of people fought in World War II, which helps explain one of the reasons for suicide**. Trauma caused by War is associated with PTSD, which could cause suicide according to the U.S. Department of Veteran Affairs and other mental health institutions.

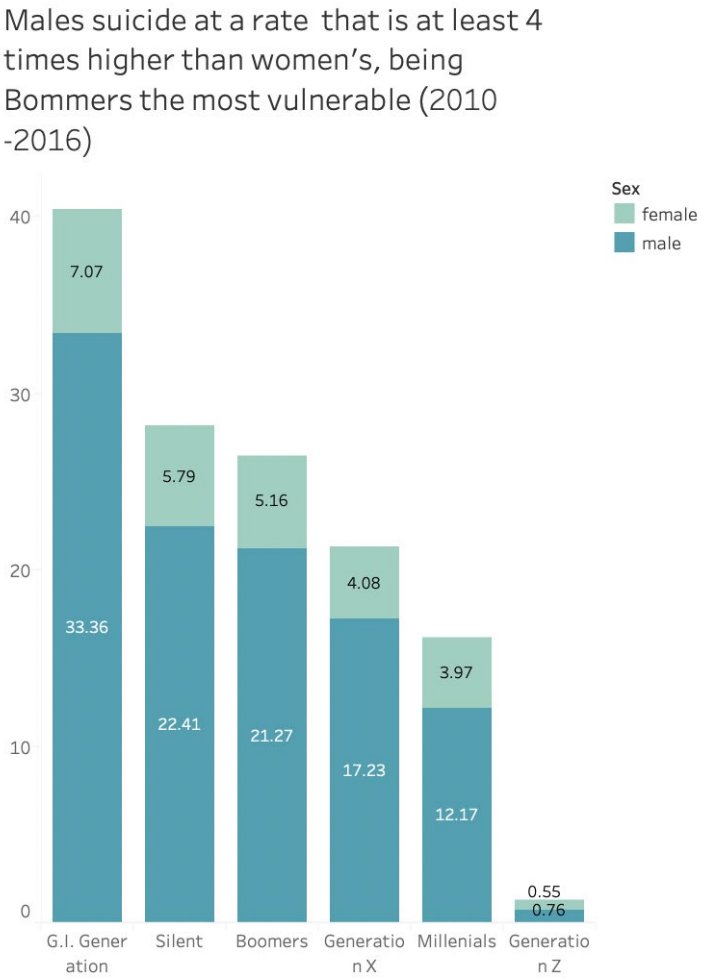

Figure 4

All these visualizations are consolidated in the dashboard "Exploratory Analytics".

Where does our mental health project should focus on to prevent suicide?

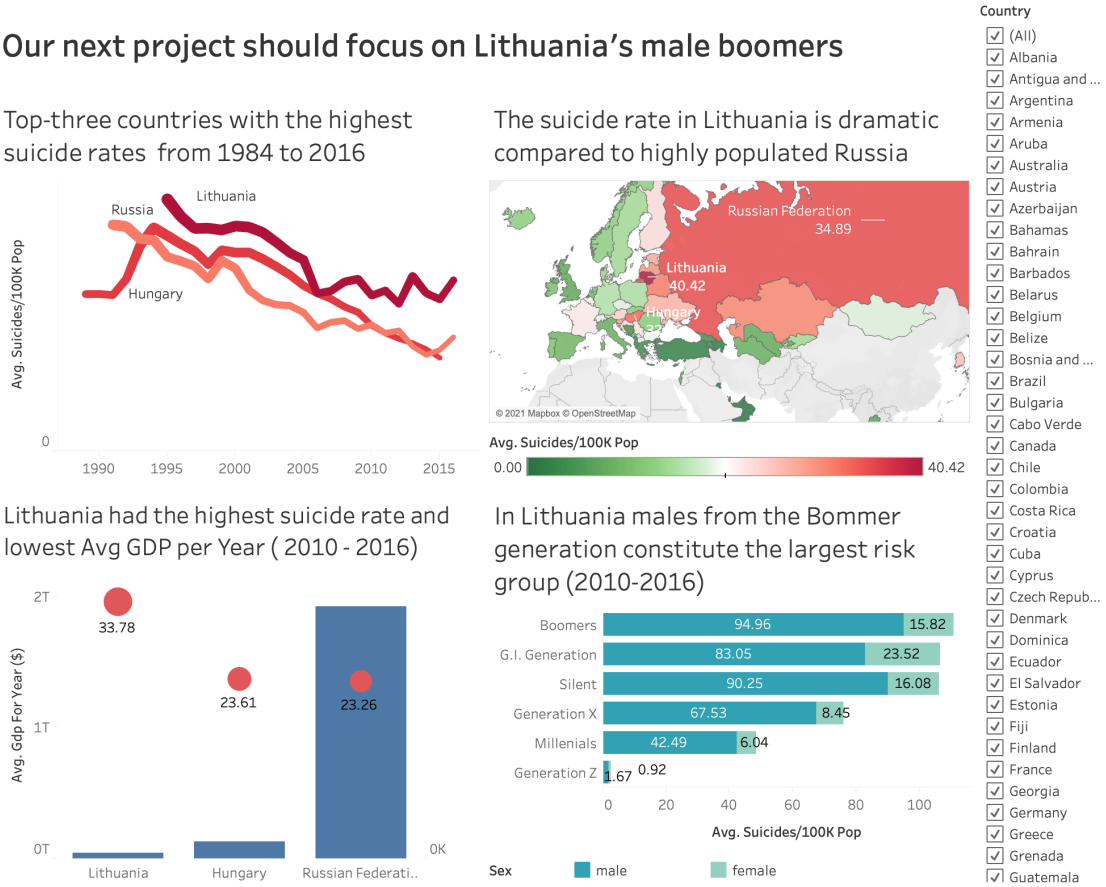

Figure 5

In order to find out which country and risk groups our mental health project should focus on, we did a deeper analysis of suicide rates per country, gender, and generation. The dashboard shown above summarizes such analysis step by step.

Next, we are going to narrate our progression of findings showing the story that we built on Tableau in figures 6, 7, 8, and 9.

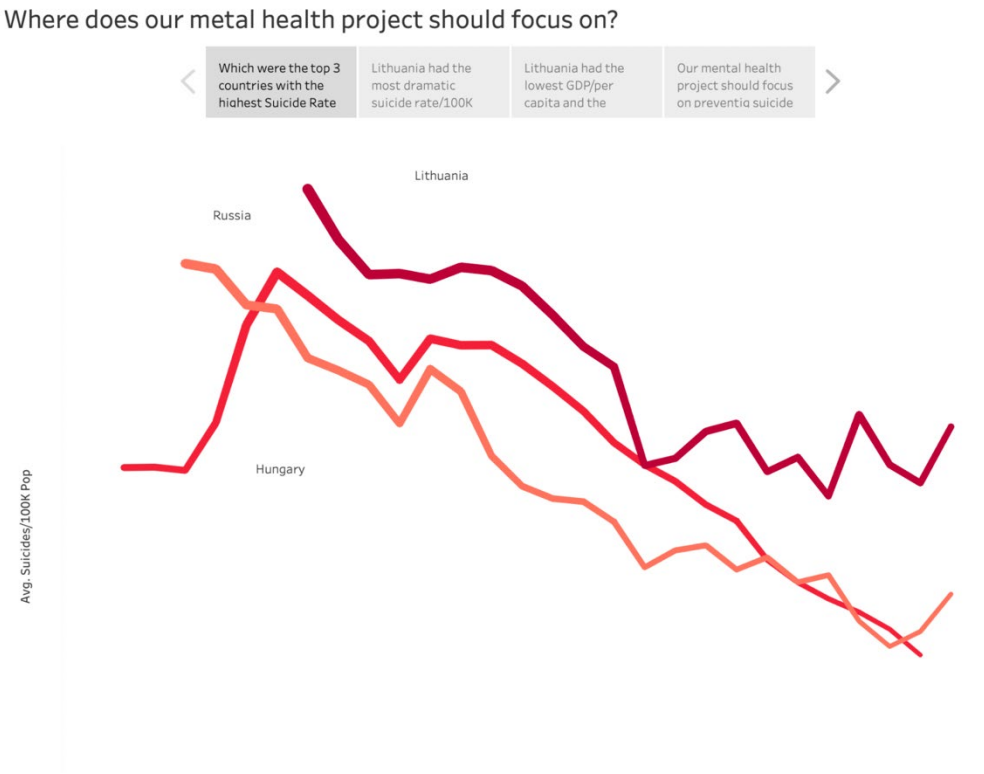

Figure 6

We started with visualizing the top 3 countries which have the highest AVG number of suicides per 100k population in the past 30 years. The result came with Hungary, Russia & Lithuania. This means Europe and Northern Asia are the continents most affected by suicide.

Lithuania came as most affected. Although, we can see here that suicide rates are reducing, Lithuania is still struggling and has been on the top in the 30 years registered in the data set. Russian Federation was the second most affected, followed by Hungary. The intensity of the color red shows the intensity of the suicide problem.

Once we found which were the top countries affected by suicide, we wanted to see the magnitude of the problem better based on the size of each country and its population. Therefore, we created a heat map, where the most affected countries are colored with red and the least affected were colored with green. All countries are colored from the most intense red to the lightest green depending on their average suicide rate. See Figure 7.

Our story and our dashboard allow us to filter each country if needed to see them separately for comparison.

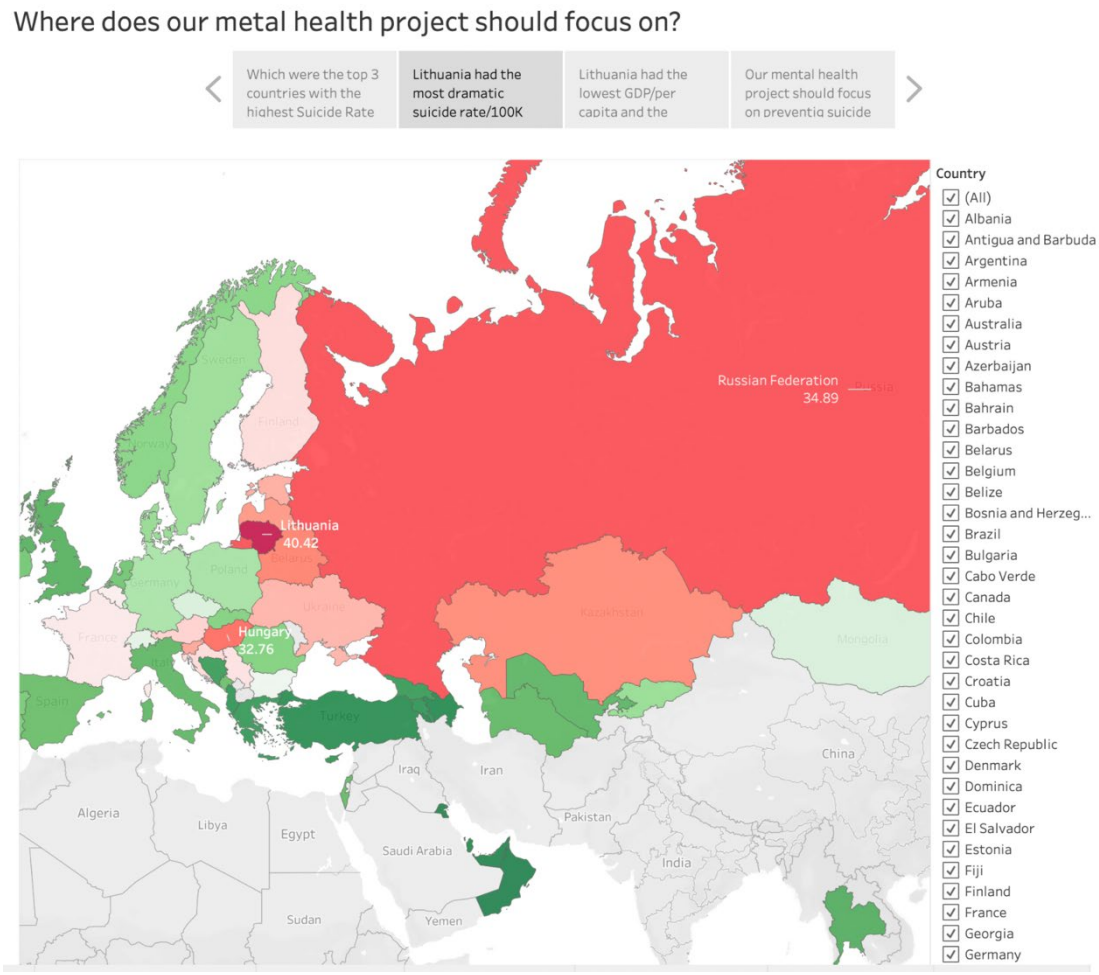

Figure 7

Here, we see that out of all three countries Lithuania has the most dramatic suicide rate by 100k of population. Even though, Russia has the highest number of suicides (due to its large territory and population), Lithuania has the highest rate / 100K of population because of the small size of its nation.

At this point, we are deciding to focus on Lithuania to conduct our mental health project to prevent suicide.

In order, to check that GDP has an association with suicide rates in Lithuania we analyzed the AVG suicide rates for each of the top countries versus their AVG gdp_for_year from years 2010 to 2016. This finding will also allow the mental health company to understand how chronic the suicide problem would be depending on the economic outlook of the country in the next years.

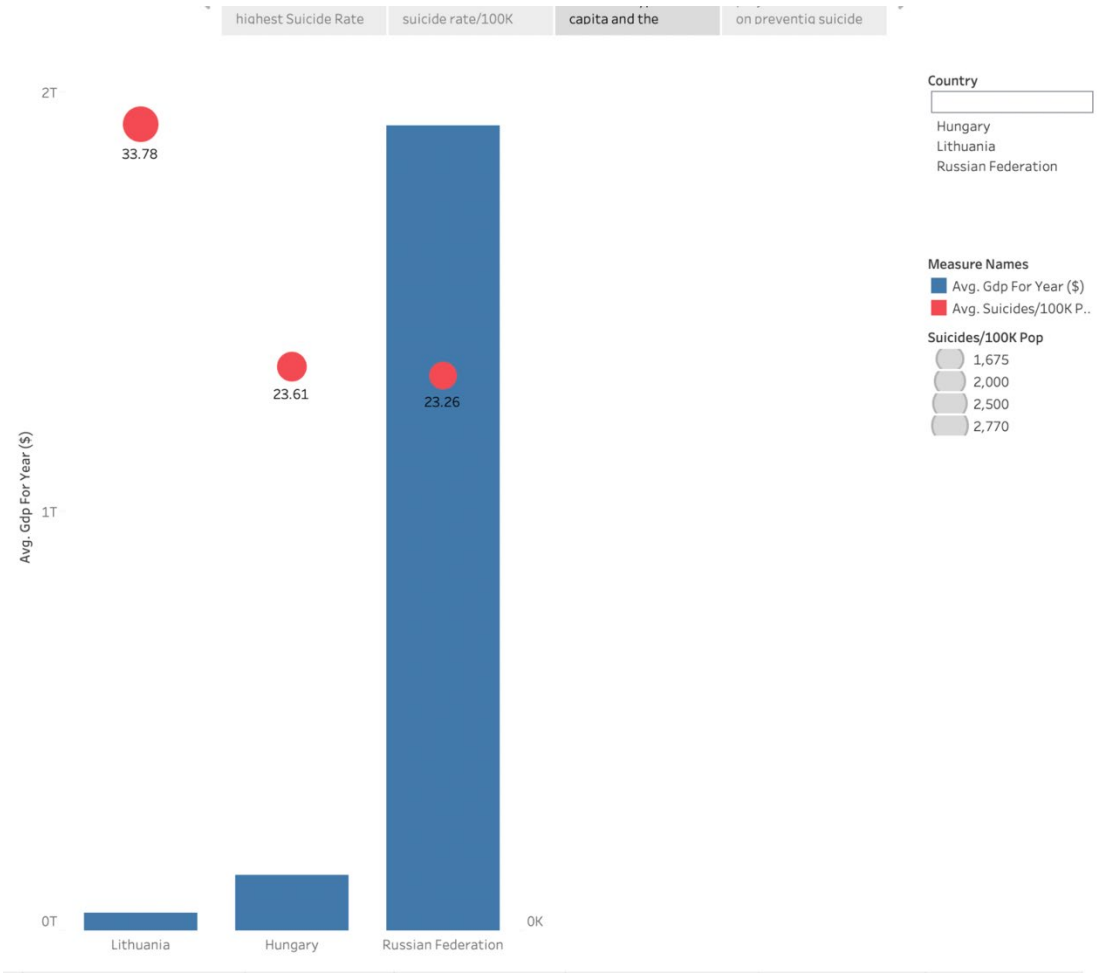

Figure 8

In Figure 7, we can see that Lithuania has the lowest AVG annual GDP and the highest suicide rate.

As a result, of these 3 pieces of analysis, we are deciding that our project will focus on Lithuania to prevent suicide. Then, we broke down the suicides in Lithuania by gender and generation in order to find the highest-risk groups that our not-for-profit should serve. Figure 8 shows our findings.

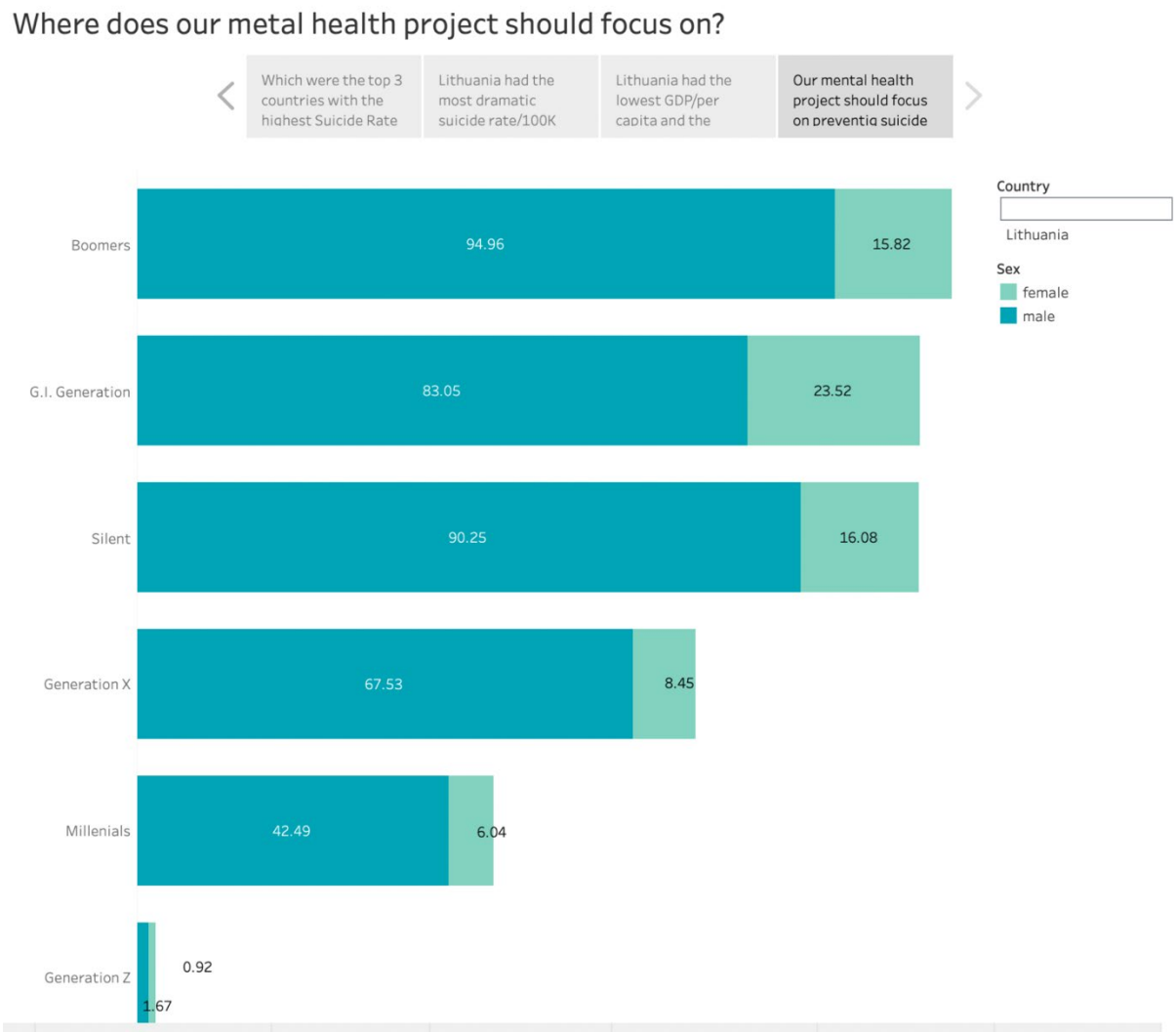

Figure 9

In this visualization, we can clearly see that the boomer generation, which belongs to the age group of 35-54 years by the year 1985, is the most affected age group by suicide (110 rates). The Greatest and Silent generations follow with a quite similar rate (106).

On the other hand, we can see that in all generations, the majority of suicides are made by male individuals, being 86% of the total suicides for this generation.

Conclusions and Recommendations

As a result of our analysis, we recommend that our mental health not-for-profit mental health company should focus on preventing suicide in Lithuania, specifically in the boomer population of males (35-54 y-o).

In our suicide prevention work, gdp_per_capita should be considered as a major factor associated with suicide.

References

Suicide Rate Overview 1985 -2015, www.Kaggle.com

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

The U.S. Department of Veteran Affairs, National Center for PTSD,

https://www.ptsd.va.gov/understand/common/common_veterans.asp

Appendix 1: SAS code for Data Cleaning