Global Suicide Rate Analysis

By:

Hussain Karamali, Ernest Ordu, Jana Taha, Xiaoran Liu, James Loveitt

Introduction

Suicide is a global phenomenon that occurs throughout the lifespan of individuals. Currently, the World Health Organization (2020) estimates that there are 800 000 deaths annually. Putting that into perspective, this means one person commits suicide every 40 seconds. With such high numbers, questions related to demographics and trends come to mind. A better understanding of the causes may help prevent suicides and provide aid to people in need. There are many notable trends related to suicide rate. The general consensus is that men have a higher suicide rate, vis-à-vis women. There is a higher suicide rate among certain countries such as Japan, Korea, and selected Northern European countries. That suicide rates often increase during times of economic recessions.

Our objective is to use a comprehensive dataset obtained from Kaggle on suicide numbers to show the above trends exist and explore the data. Additionally, we wanted to see if there are any trends not well known that we could find.

In addition to performing analysis on this suicide-related data, we were hoping to build a machine learning model that can help predict future rates of suicide. This modeling type could be just as crucial as analyzing past data in the fight to prevent suicide.

Lastly, we feel that it is imperative to point out that this analysis is not exhaustive, and other rational analyses could be performed.

Data Preparation

We chose a dataset (figure 1) from Kaggle that included suicide rates from 101 countries worldwide, spanning 30 years (table 1.). The dataset was created from four different sources linked by time and place; and was built to socio-economic spectrum (Kaggle, 2020). In addition to suicide numbers by country, it includes; corresponding gross domestic products (GDPs) pulled from (WB, 2020), human development index (HDI) gotten from (UNDP, 2020), the suicide rates for these countries retrieved from (WHO, 2018) and the sex and age of those who passed away.

Country	101
Year	32
Sex	2
Age	6
Suicides_no	2084
Population	25564
Suicides/100k pop	5298
Country-year	2321
HDI for year	305

Gdp_for_year ($)	2321
Gdp_per_capita ($)	2233
Generation	6

Table 1. Unique values for each column. Shows the number of countries and years and age groups the data encompasses

The core dataset we used required minimal preparation. Most columns were complete and had no missing values. The only column missing data was HDI. Of the 27,820 rows in the dataset, only 8,364 of the HDI column had entries. There was no way to easily impute this information into the rows with missing data as this information is not available for many countries for specific periods. As such, this column was omitted from our analysis.

The "year" column contains observations that spanned the period 1985 to 2016. And we have two gender groups and six age groups. We should have 12 Rows for every country per year. (6 age groups by the 2 gender groups). This was the case for all the years, except for 2016. Data was missing for some countries, so this year did not meet the 12 Rows of data /country/year style of the preceding years (1985-2015). only a few countries had entries for 2016, so we decided to drop 2016 from our analysis.

The GDP column became an object data type when it was imported into Jupyter, which meant that it was essential to type-cast it to an integer/float data type before it could be used for any analysis. If this were not done, we would not be able to get the measure of central tendency on this column (i.e., mean and median)

For some aspects of the analysis, supplementary data was added to the dataset to streamline the analysis. An example of this would include adding the continent field as another way of categorizing the country. Given that continent is a relatively straightforward addition, a simple dataset was taken from Datahub.io, modified, and merged to the core dataset.

In an ideal world, the dataset would provide the exact age of the individuals who passed away. This scenario would allow anyone utilizing this data to do a more complex analysis or the ages into groups of their choosing. Despite this, the age column is still very useful and can be used for various insight angles.

Analysis

To understand the prevalence of suicide rates and numbers within different demographics, we broke our data down into dependent and independent variables (table 2). After this, we looked at several combinations of independent variables that could influence the dependent variables. These combinations were analyzed to see their impact on suicide rates/numbers. Only those analyses that yielded significant results will be summarized in this report.

Independent	Dependent
Country	Suicide numbers
Sex	GDP/capita ($)
Age	Suicides/100k pop
Year	Population
Generation	

Table 2. Dependent & independent variables.

Trend over time

Now let us look at the global suicide rate over time.

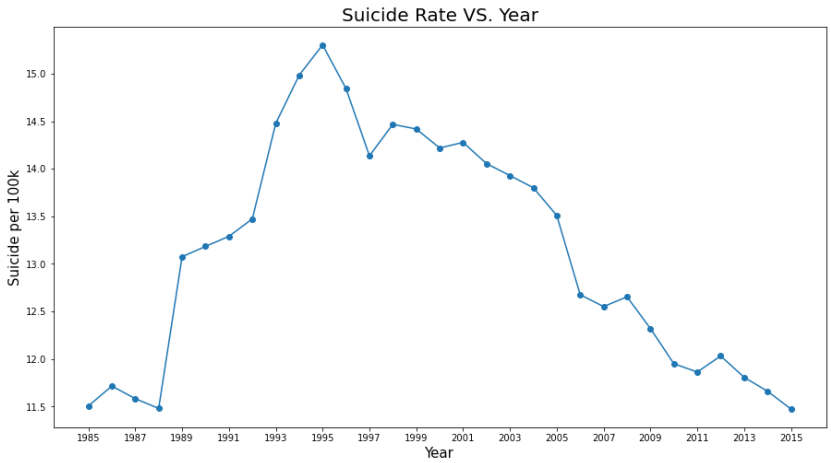

Figure 2.Global Suicide Rate vs. Year

For the y-axis, we calculated the suicide rate by calculating the number of deaths per 100k of the population for each year. We did so by grouping the dataset by year, then taking the sum of the suicide number for each year and dividing it by the sum of the population for that year. We then multiplied that ratio by 100,000.

From the graph above, we see that the suicide rate was increasing from 1985 into the mid-90s. The suicide rate reached its peak at around 15.5 death per 100k in 1995. The rate then started to decrease steadily, and we could see that the rates are now returning to their pre 90s rate.

Country

As mentioned earlier, the continent that a country is a part of was merged into the dataset to facilitate the countries' grouping into regions. Each row was joined to the secondary dataset on the column country, which allowed the continent to be easily added. The six continents that are present in the dataset are Asia, Europe, Africa, Oceania, North America, and South America.

Once this was added, a simple "group by" function was performed to calculate the average suicide rate per 100k for each continent.

Continent	Suicides/100k Population
Africa	7.58
Asia	13.87
Europe	16.31
North America	7.53
Oceania	11.56
South America	11.43

Table 3. The number of suicides/per 100k population by continent.

Table 3 shows the suicide rate in Europe and Asia is notably higher than in the remaining continents. Also, the suicide rate in North America is very low. This insight is somewhat surprising as North America and Europe are very similar, and we would expect the suicide rate to be relatively close.

Country & Year

Once "year" was also brought into this analysis, you can see that each continent has its unique trend related to suicide over the 30 years. Some countries, such as South America, saw increases from the initial few years of data, while others, such as Asia, saw a slow and steady decline.

We thought it valuable to perform an autocorrelation plot for a select continent to see if the time period significantly influenced the rate of suicide.

Below (figure 3) is the autocorrelation for Europe for the full-time period of the dataset.

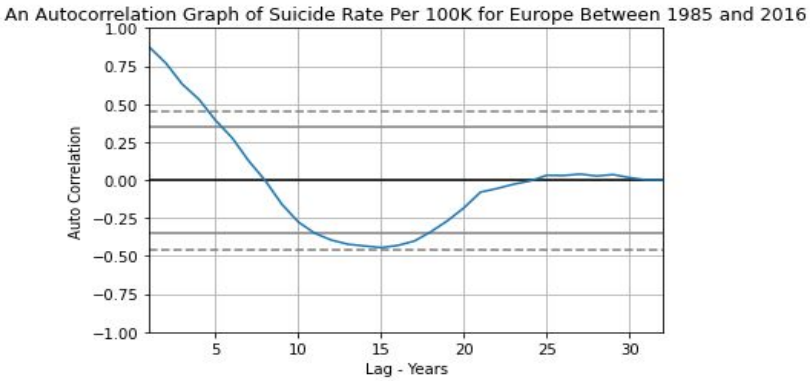

Figure 3. Autocorrelation of Europe's yearly suicide rate

This graph (figure 3) is quite insightful and shows that for the European continent, "year" may have a significant influence on the suicide rate, especially in the early and mid-period. However,

for the latter part of this dataset, it has little to no influence and is close to 0. While not pictured, the South American continent produced a very similar Autocorrelation to Europe.

The below (figure 4) is the autocorrelation of the North American continent.

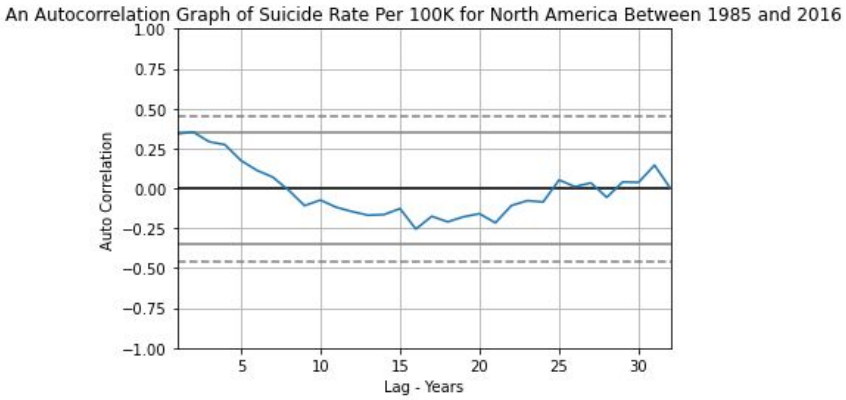

Figure 4. Autocorrelation of North America's yearly suicide rate

North America Correlogram (figure 4) is distinctly different from the European Autocorrelation plot as autocorrelation for all lags falls inside the significance bounds. This shows that year has minimal influence over North America's suicide rate for the given time period. It also means that there is no clear trend in suicide rates over the last few decades.

As mentioned earlier, this is surprising given that many would think that North America and Europe would have similar trends given their close cultural ties.

Country	Year	Suicide Number
Russian Federation	1994	61420
United States	2015	44289
Japan	2003	31881
Republic of Korea	2011	15906
Ukraine	1996	15160

Table 4. Top 5 countries with the highest suicide number and its year

Inspecting more details on suicide numbers and suicide rates in specific countries and years, Table 4 shows us the top five countries with the highest suicide number and its year. This does not show a geographical distribution trend but may be related to political or economic events.

Country	Year	Suicides/100k population
Lithuania	1995	53.27
Hungary	1991	47.91

Russian Federation	1994	47.30
Sri Lanka	1985	46.56
Latvia	1995	45.47

Table 5. Top 5 countries with the highest suicide rate and its year

If we look at the top suicide rates countries and their years, we can find that the top five countries are Lithuania, Hungary, Russian Federation, Sri Lanka, and Latvia (table 5). To be noted, Estonia ranked sixth after Latvia, which the highest suicide rate is 45.27 per 100,000 population in the year 1995. Three of the top six countries are Baltic countries, and two are East European countries, with the highest suicide rate years from 1991 to 1995. Therefore, it indicates that there might be some geographical and geopolitical links between these top suicide rate countries. From 1991 to 1995, the most significant geopolitical event that happened to these countries was the USSR collapse. Russia's suicide rate increases from about 1991 (figure 6), when the USSR collapsed (figure 5), and peaked in 1994. Compared with the data given by Varnik and Wasserman (1992, p.77), the change from 1988 to 1994 is stunning: from 24.3/100k to 47.3/100k.

Table 1. Suicide rates per 100,000 inhabitants in the republics of the former USSR

	1984	1985	1986	1987	1988	% change 1984–1988
Whole USSR	29.6	24.5	18.9	19.1	19.4	−34.5
Republics						
Russia	37.9	31.0	23.0	23.2	24.3	−35.9
Lithuania	35.8	33.7	25.1	28.7	26.3	−26.5
Latvia	33.7	29.0	24.9	23.0	22.5	−33.2
Estonia	32.4	31.5	27.3	25.3	24.3	−25.0
Byelorussia	29.6	23.1	17.7	19.0	18.4	−37.9
Ukraine	26.5	22.3	18.5	19.6	19.0	−28.3
Kazakhstan	26.1	22.2	16.5	16.2	16.8	−35.6
Moldavia	23.3	20.7	18.8	17.1	17.0	−27.0
Kirgizia	14.8	11.6	9.2	11.3	11.2	−24.3
Uzbekistan	8.8	8.2	7.6	6.9	6.3	−28.4
Turkmenistan	8.6	7.5	8.8	8.0	7.7	−10.5
Tadzhikistan	6.2	5.9	5.3	4.4	4.0	−35.5
Georgia	4.8	4.6	4.5	4.3	4.3	−10.4
Azerbaijan	4.5	3.7	3.5	3.7	3.3	−26.7
Armenia	1.9	2.2	1.8	2.4	1.8	−5.3

Figure 5. (Värnik & Wasserman, 1992, p.77)

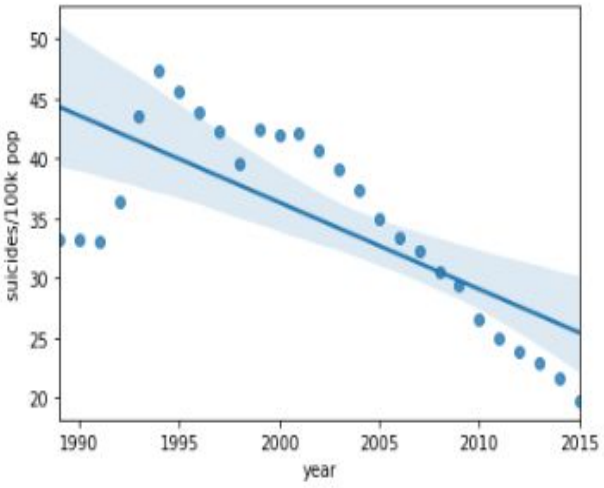

Figure 6. Russian Federation's suicide rates

This also validates European countries' autocorrelation plot results (figure 3), showing a mildly positive autocorrelation of suicide rate small lags values. However, calculating the correlation between suicide rates and political events is beyond the scope of this project. Our objective is merely to point out that more factors affect the suicide rates than what can be deduced from our dataset.

Age (Generation)

Age groups were classified into the following categories for the analysis.

G.I Generation	Pre-1928
Traditionalists/Silent Generation	1928-1946
Baby Boomers	1946-1964
Gen X	1965-1976
Gen Y/Millennials	1977-1995
Gen Z/iGen/Centennials	1995-2010

Table 6. Generation new classification by years

	Suicide Numbers					Suicides/100k Population				
Generation	**Min**	**Max**	**Mean**	**Count**	**Sum**	**Min**	**Max**	**Mean**	**Count**	**Sum**
Boomers	0	22,338	457.81	4,990	2,284,498	0	151.33	14.74	4,990	73,563
G.I. Generation	0	6,401	185.86	2,744	510,009	0	224.97	23.95	2,744	65,708
Generation X	0	11,767	239.20	6,408	1,532,804	0	94.28	10.56	6,408	67,648
Generation Z	0	277	10.82	1,470	15,906	0	11.02	0.64	1,470	944
Millennials	0	6,945	106.68	5,844	623,459	0	71.17	5.38	5,844	31,461
Silent	0	12,517	279.97	6,364	17,817,44	0	204.92	18.42	6,364	117,217

Table 7. Grouping by generation (age groups)

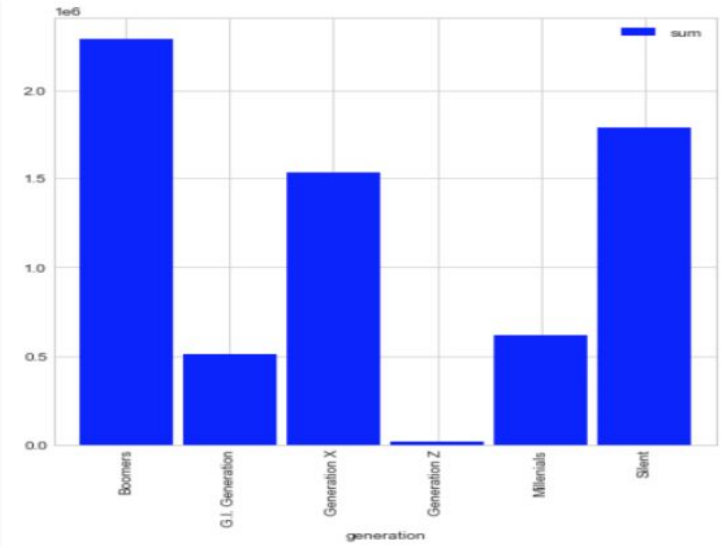

Figure 7. Suicide rates of the age groups

The analysis of different age groups shows that the boomers have a significantly higher suicide rate in general. However, G.I Generation have the highest average suicide rate per 100k of the

population (table 7). The bar chart in figure 7 shows the capriciousness in the suicide rates for the different age groups.

Age and Year

The analysis shows that the boomers saw the single highest ever suicide rate for an age group in 1994. A summary of the top 3 suicide number/group recorded in any age group in a year is shown in table 8. However, it is interesting to note that this group was a large part of the world's population.

Generation	Year	Suicide Numbers
Boomers	1994	125,932
Generation X	2010	125,681
Boomers	1993	119,714

Table 8. Top 3 highest suicide rate in a group recorded in a single year.

If we carry out the same analysis on a group/population basis, we get very different results. In this case, the silent group has the top 5 ever suicide rate/population in this data set. Table 9 shows the top five years on record in which a single group saw the highest suicide/population, and it is shocking to see that the silent group was associated with all 5 top years. This is because the silent group makes up only about 5% of the population compared to the boomers than account for nearly 30%.

Generation	Year	Total Suicides/100k Pop
Silent	2002	7,403
	2001	7,334
	2003	7061
	2004	6495
	2007	6456

Table 9. Age group/year/100k world's population

Age and Country

When sorting the data by average suicide rate and suicide number, it yields two different results. The top five countries with the highest average suicide rate have mostly G.I. generation as their top age groups (table 10), whereas the top five countries with the highest suicide number have mostly Boomers generation as their top groups (table 11). The top 5 countries with the highest suicide number by age group resemble the results by year, which indicates that certain countries have higher suicide numbers than others.

Country	Age group	Avg Suicides/100k population
Hungary	G.I.Generation	82.74
Republic of Korea	G.I.Generation	69.26
Serbia	G.I.Generation	66.12
Slovenia	G.I.Generation	62.57
Lithuania	Boomers	62.29

Table 10. Top 5 countries with the highest average suicide rate by age group

Country	Age group	Suicide number
Russian Federation	Boomers	479140
United States	Boomers	380917
Japan	Boomers	278679
Ukraine	Boomers	124721
France	Boomers	123510

Table 11. Top 5 countries with the highest suicide number by age group

Sex

The "Sex" variable alone was analyzed to see the differences in suicide numbers/rates within the two categories. As shown below (table 12), the counts for both males and females are equal. However, the mean, standard deviation, and max tell us that globally males are more likely to commit suicide. Based on the total suicides per 100k population (table 13), males are 3.7 times more likely to commit suicide.

Sex	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Female	13 830.0	5.39	7.37	0.0	0.41	3.16	7.41	133.42
Male	13 830.0	20.23	23.60	0.0	2.38	13.52	27.36	224.97

Table 12. Global suicide per 100K population summary data

Sex	**Total Suicides/100K population**
Female	74,629
Male	279,767

Table 13. Total global suicides per 100k population.

Figure 8 shows the suicide/100k population spread by sex. The median for males is four times higher than for females. The max is ~2 times higher for males.

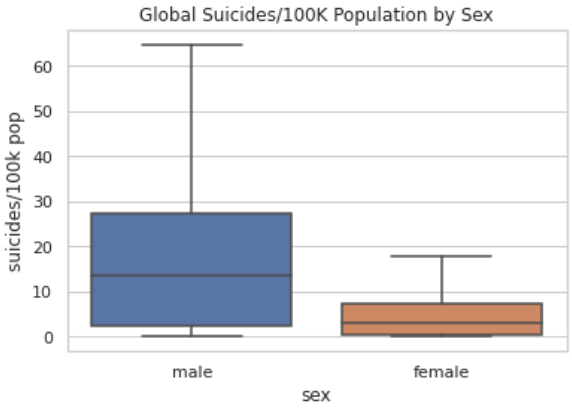

Figure 8. Global suicides per 100k/population by sex.

Country and Sex

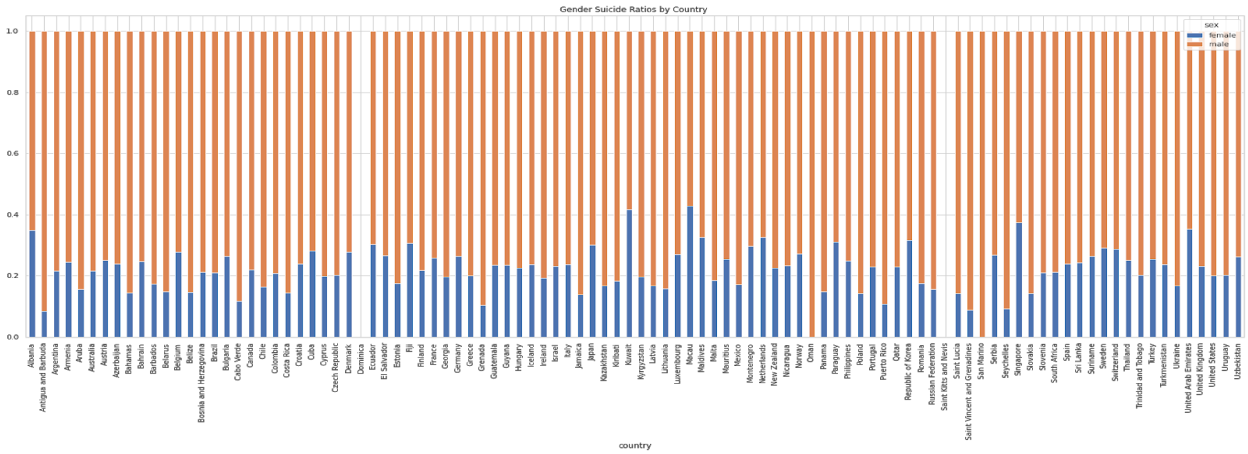

Figure 9. Gender suicide ratios by country.

As was noted above, globally, we see that males are more likely to commit suicide than females. Also, across countries, we see a variation in terms of suicides per 100k population. Due to these findings, we wanted to see if the global gender suicide differences were consistent across all countries, i.e., are males more likely to commit suicide in each country. Figure 9 shows the percentage of suicides for each gender across all countries. Based on Figure 9, we see that global gender suicide differences do hold up in most cases. However, there are differences in countries like Kuwait and Macau, where the split per 100k/population seems more even. These findings are significant as it indicates that the same suicide prevention approach cannot be taken in these countries as others.

Sex and Year

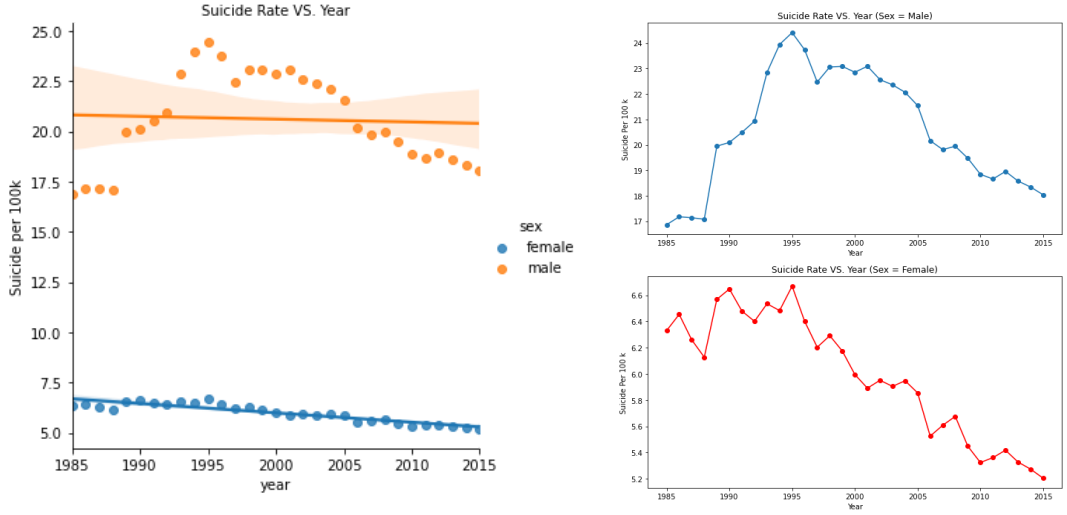

Figure 10. Bi-Variable analysis of gender and year.

From the graphs above, we can see that the rate of suicide for men was always significantly higher than that of the females. We also noticed that both males and females' suicide rates peaked in 1995 and then declined subsequently. A decreasing linear trend can be seen between suicide rate and time for females, as we can see that their suicide rate in the 80s was higher than it is now. On the other hand, we don't see the same linear trend for males. The male suicide rate increased from 1985 until it reached its peak in the mid-90s 90s, then started to decrease steadily and is now returning to its pre 90s rate.

Age & Sex

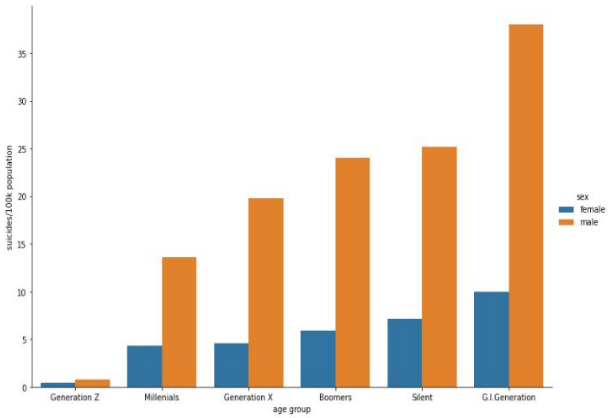

Figure 11. Suicide rate by age groups and gender barplot.

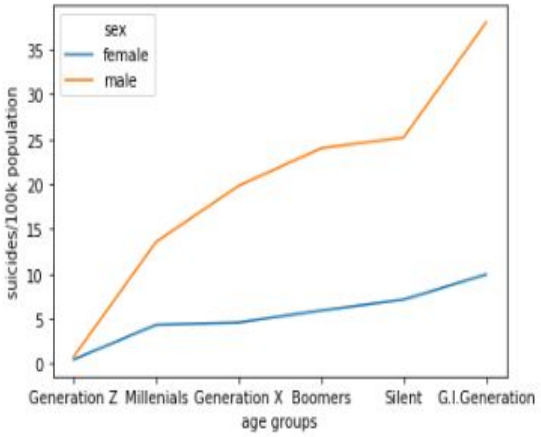

Figure 12. Suicide rate trend by age groups and gender.

Among age groups, there are huge differences between genders. (figure 11) In all age groups, males tend to commit suicide more than females; the difference is large in most groups except for Generation Z. It also shows a growth of suicide rate from the younger generation to the older generation(figure 12.), and male's slopes are significantly steeper than females'.

Sex, Age & Year

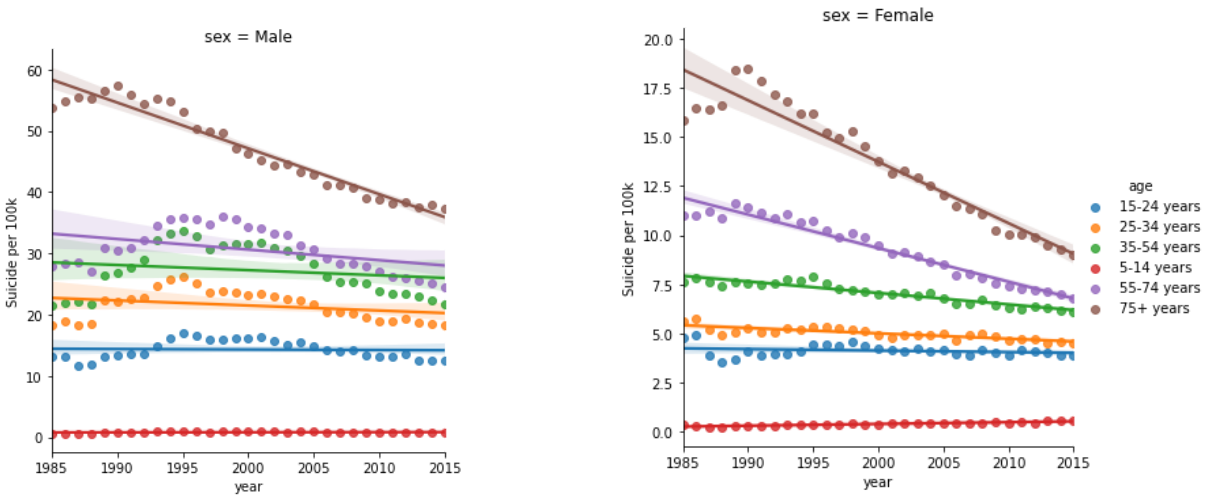

Figure 13. Male's and female's suicide rate by year grouped by age

The overall trends of male and female suicide rates by time are quite similarly decreasing. The female silent generation's suicide rates fall off more significantly than males. Male generation Z, millennials, generation X, and boomers' suicide rates fluctuate more than their female counterparts, especially around 1994 to 2005.

Machine Learning Module

We used supervised machine learning to predict future suicide rates. The model was built using the regression model in "sklearn" and utilizing the Polynomial Feature In 'sklearn.preprocessing'. we attempted different degrees of freedom in the polynomial feature during the analysis. The best result was obtained for n=3. Hence the model was built with the 3rddegree of freedom.

The objective was to build a model using regression analysis to predict future yearly suicide rates and verify that it works. The process of verification was done in two steps. Firstly, we built the model using all the datasets to see how it fits the data, which produced a very acceptable fit[1] (figure 14). After this, we proceeded to verify the model's validity by splitting the data into test and train sets. Using the fit obtained from the train set to predict the test set. The results were very satisfying with an R^2>0.9 (figure 15). This undoubtedly verified the strength of the ML model.

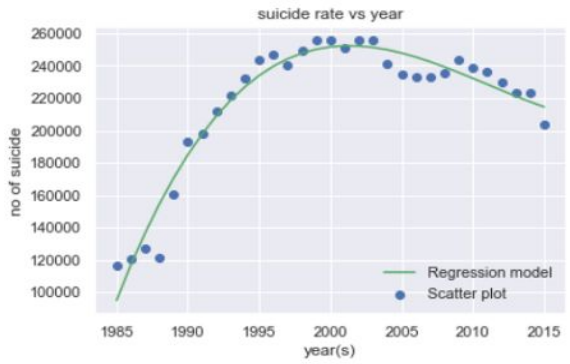

Figure 14. Testing the model: Regression model and scatter plot

Conclusion

By delving deep into the data about global suicides, we have learned about the prevalence of suicides across different demographics and various factors that influence the causes.

This analysis shows that "year" as a variable did not have much impact on suicide rates. However, it does show that global suicide rates have declined to the late 1980's levels. The trend overall shows low correlation across continents, where in some cases, short term impact can be determined but not long term. However, as shown in our "age" and "year" analysis, the year can serve as a marker of high and low suicide rates. Using it as a marker, one can do additional analysis into events surrounding the years to determine the high/low suicide numbers/rates. Additionally, other factors, e.g., geopolitical events, may impact the suicide rates, which can be seen from our "country" and "year" analysis.

Our analysis shows that categorical variables help provide a better breakdown and understanding of the prevalence of global suicide. Overall, we see that Baby Boomers have high suicide numbers as they constitute a vast majority of the global population. Looking at suicide rate, we see that the Silent Generation has the most tendency to commit suicide. We also see a trend in the "age" groups graphs (figure 13) that validated this theory (older people tend to commit suicide more than the younger generations). From analysis based on "gender" variables, we see that males are more likely to commit suicide than females. Finally, based on continental groupings, countries in Asia have the highest suicide rates among all continents.

The categorical analysis helps understand which groups are more prone to commit suicide and can help organizations (e.g., NGOs) and political leaders better distribute their resources on suicide prevention initiatives. As such, our analysis provides similar information as current studies and literature and suggests that males and the older generation are at a higher risk of suicides, especially in Asian and European countries.

The linear regression module is found effective when predicting future suicide rates. We have a result with an $R^2 > 0.9$, which is satisfactory.

References

Kaggle. (2020, 12 10). *Suicide Rates Overview 1985 to 2016.* Retrieved from Kaggle.com: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

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

UNDP. (2020, 12). *United Nations Development Program.* Retrieved from Human development index (HDI) Report: http://hdr.undp.org/en/indicators/137506

Värnik, A., & Wasserman, D. (1992). Suicides in the former Soviet republics.*Acta Psychiatrica Scandinavica*, 86: 76-78. https://doi.org/10.1111/j.1600-0447.1992.tb03230.x

Walsh, N. (2003). *Russia's suicide rate doubles.* Moscow: The Guardian.

WB. (2020, 12). *World Bank.* Retrieved from World development indicators: GDP (current US$) by country:1985 to 2016: http://databank.worldbank.org/data/source/world-development-indicators#

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

WHO. (2020). *Mental Health and Substance Use.* Retrieved from World Health Organization.: Mental Health and Substance Use (who.int)

Appendices

Figure 1. A summary of the different columns of the dataset using the .head () method

Out[29]:

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2.156625e+09	796	Generation X
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2.156625e+09	796	Silent
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2.156625e+09	796	Generation X
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2.156625e+09	796	G.I. Generation
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2.156625e+09	796	Boomers

Figure 15. R^2 Value of the ML model

```
In [50]: #Second verification method
         poly=PolynomialFeatures(degree=3)
         x_train_poly,x_test_poly=poly.fit_transform(x_train),poly.fit_transform(x_test)
         poly.fit(x_train_poly,y_train)
         poly.fit(x_test_poly,y_test)
         Model=LinearRegression()
         Model=Model.fit(x_train_poly,y_train)
         print(Model.score(x_test_poly,y_test))
         r2_score(y_test,Model.predict(x_test_poly))

         0.9669084212629442

Out[50]: 0.9669084212629442
```

1. We dropped 2016 data from the ML analysis because that entry was an outlier (too small a value), probably because data collection was still in progress for the 2016 year when Kaggle put together the data in 2018.