# Suicide Rate Data Analysis

**Date: July 24, 2019**

**By Team-5:**

**Vaishali Siddeshwar, Gaurav Rattan Singla**
**,Jie Ren, Javier Rodriguez, Duane Robinson,**
**Liangliang Shi, Wing Ho &Chandni Sehgal,**

## Table of Contents



## Introduction

Every individual is responsible for their actions and outcomes. However, the society and the socio-economic environment can have a huge impact on an individuals life. Suicide is global and according to World Health Organization, "Close to 800 000 people die due to suicide every year, which is one person every 40 seconds" (WHO, 2019). In the report below, Suicide Rate data has been explored to understand what variables act as dominant factors and what is the correlation between these variables, that forces people to commit suicide.

## Objective

The dataset is explored to understand what significance each variable could potentially have on an individual's life. It is further explored to see if any of these variables have any correlation with each other. For ex: GDP which represents the economy of a country and its residents; does GDP have any correlation with number of suicides? In other words, if the GDP of a country goes up, the number of suicides should go down and vice-versa. Furthermore, the significance of gender variable and its correlation with number of suicides is also explored.

The data was obtained from Kaggle, which is further combined from four other datasets, all linked by time and place as a common variable among them. In addition, another dataset about temperatures is combined, which was also obtained from Kaggle.

## Data Preparation

The suicide rate dataset has 12 variables, containing Country, Year, Sex, Age, Suicides No, Population, Suicides per 100k Population, country-year, HDI for year, GDP for year, GDP per capita and Generation. The variable Country is represented as String, Year is in range from 1985 to 2016, Age is divided into five intervals ranging from years of 5-14, 15-24, 25-34, 35-54, 55-74 and 75+. HDI is Human Development index indicating life expectancy, income and education of the country for the given year.

In addition to the above suicide rate dataset, it is further combined with temperature dataset. The temperature dataset contains four variables, Date in YYYY-MM-DD format ranging from 1743-11-01 to 2013-09-01 represented for each country. The common variables between suicide rate dataset and temperature dataset were country and year. The year in temperature dataset had to be cleaned up to match the year range and format of Suicide Rate dataset. Hence at this point, the total number of variables considered for this report comes down to 13. The variables were categorized into categorical and numerical features. Since categorical features are continuous, the variables Country, Year, Sex, Age and Generation were categorized as categorical features, whereas Population size, Number of Suicides, Suicides per 100k population, GDP for year, GDP per capita and HDI for year are categorized as numerical features for the analysis.

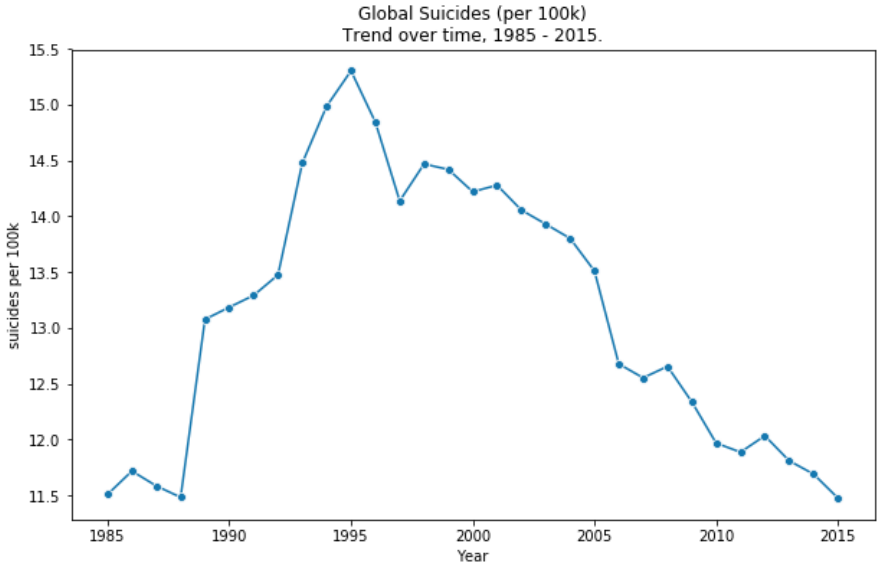
The Suicide Rate contains 27820 rows x 12 columns, out of which 19456 (~70%) of rows are missing data for HDI variable. It was decided to drop data with NaN values for this column because populating it with arbitrary or 0 value would have caused bias and heavily skew the data. Hence, HDI with only non-empty data was used for this analysis. Country-Year was also a redundant column, which was not needed for our analysis. Hence, this column was entirely

dropped as well. The format of Year variable was also changed, in order to smooth the plotting. Also, the data is available from 1985 to 2016, however only few countries had data available for 2016. Hence, the data for the year of 2016 was dropped in order to keep it consistent. And if a country had less than 5 years of data available, those countries were dropped too. At this point, the total number of rows and columns left for analysis was 27,492 rows and 11 columns. In short, after data cleaning and preparation, only 1.18% of data was lost, which in turn indicates that dataset obtained had good quality of information for the rest of the analysis.

## Analysis

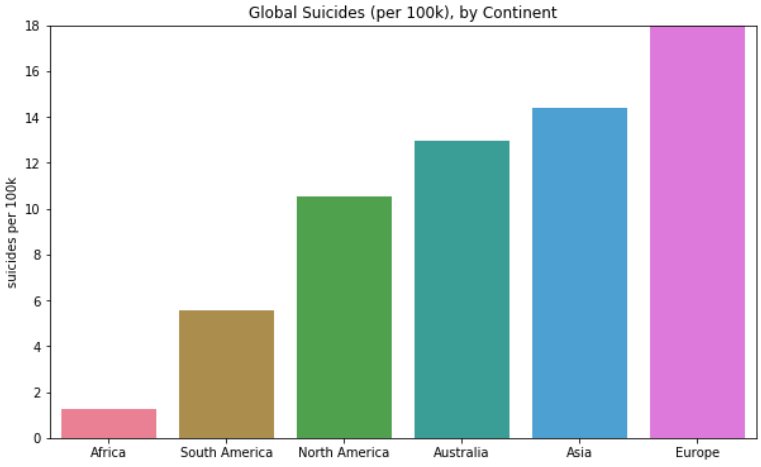
### Global Trends:

To begin with, one of the biggest challenges is to determine whether the death of a person in a country is recorded with how much accuracy. In other words, it is hard to determine if the death of an individual that was recorded in the country was actually suicide or a natural death. This uncertainty about this is especially high in developing or poor countries where the system to record such information may or may not be present. Hence, it is first decided to get the overall bigger picture about the trend of suicide rate across the world from the year of 1985 until 2015. In order to do this, the average of suicides_per_100k and group by Year variables were utilized to achieve the overall global trend shown in the plot attached in appendix (Global Trend of Suicide Rate, 1). Some of the insights that could be further drawn from this graph were the fact that the peak rate of suicides in the world was in year 1995, with the rate of 15.3 deaths per 100k population. The trend after 90's tend to be decreasing steadily at the rate of approximately 25% whereas the insights before 90's seem to be relatively weak as there is not enough data available. Only 5 data points were available for the trend before 90's.



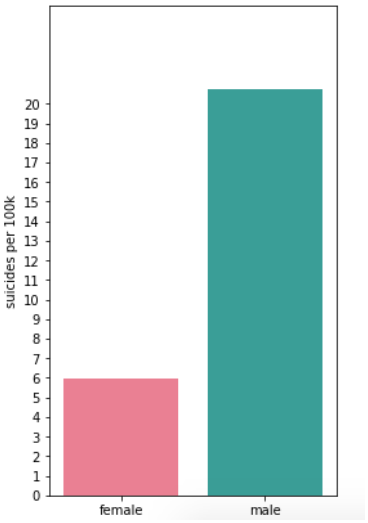
## Global Trend by Continent:

To break it down further from global level trend, we introduced a new column called Continent. The name of countries were parsed using a Python library called country_alpha2_to_continent_code, which returned back continent each country belongs to. Again, the variable of suicides_per_100k and Continent were used to analyze the trend. It turned out, Europe has highest number of suicides rate in total.



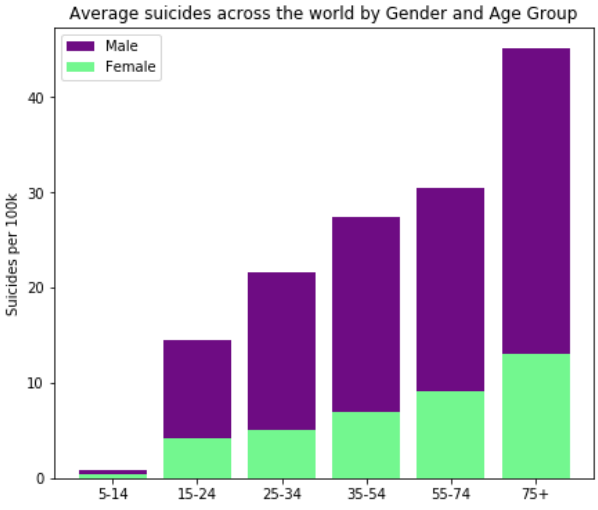

On the bright side, the trend of the suicide rate in Europe has steadily decreased by ~40% alongside Africa and Australia, however the rate is still increasing at an alarming rate in Asia, South America and North America continents.

## Global Trend by Gender:




It was also interesting to see the number of suicides committed by men was significantly higher than women. The global rate of suicide among men is 3.5 times more higher than women. According to a report by BBC, the gender gap is because, "women tend to have higher rates of depression diagnoses. Women also are even more likely than men to attempt suicide. In the US for example, adult women in the US reported a suicide attempt 1.2 times as often as men. But male suicide methods are often more violent, making them more likely to be completed before anyone can intervene" (Schumacher, March.18, 2019). Women in general are considered as more expressive whereas if men try to express their emotions, society tends to generalize them as weak. Instead of hearing them out, men often get told to 'man up'. Such judgements forces them to take irreversible actions and hence it explains, why suicide rate is higher among men than women.

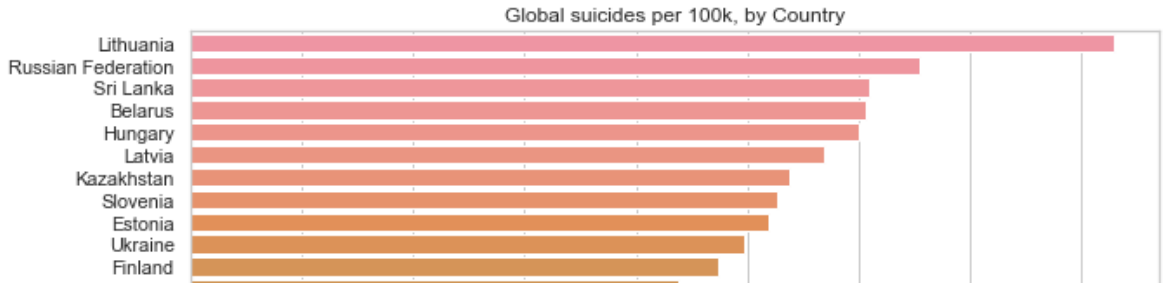
## Global Trend by Gender & Age:




By analyzing the suicide rate based on individual's Age, there was a positive correlation where the likelihood of suicide increases with Age. In fact, the age group of people with age greater than 75 has also dropped by more than 50% since 1990. We took this further by looking at the Gender (a categorical variable) grouped by Age (a numerical variable) and suicides_per_100k. our insights were further consolidated and verified that suicide rate is higher for men than women and men across all age groups have higher suicide rate than women.

## Global Trends by Country:

Since the analysis of global trend showed decline in suicide rate but per continent showed the rate declining only in Europe, Australia and Africa but not the other others, it made sense to deep dive into suicide rate by per country. Furthermore, there is large overrepresentation of European countries with high suicide rates and very few countries representing low suicide rates. Hence, it made sense to look at suicide trends by country instead of continent. For suicide rate in each country, the suicides_per_100k variable was analyzed against grouping by Country. The results showed the country Lithuania, followed by Russia, with Lithuania having the highest suicide rate with greater than 41 suicides per 100k each year. Since Lithuania and Russia is part of Europe, it supports the analysis why Europe as continent has highest total number of suicides among all other continents.




Further analysis were performed to see what are the suicide rate trends like within each country. In other words, the top 5 countries with steepest increase and decrease of suicide rates over the time period from 1985 to 2015 were found. The five countries with steepest increase in suicide rate were Cyprus, Suriname, Guyana, Korea and Montenegro with Korea and Guyana being the highest concern. The top five countries with sharp decline in suicide rate were Kiribati, Estonia (from 43.5 in 1995 to 15.7 in 2015), Latvia, Lithuania and Hungary. It is an interesting point to note that even though Lithuania as a Country probably had highest number of suicides at

some points between 1985 to 2015 but the country has managed to be one of the top 5 countries in decline of the suicide rate. One of the many other things that could cause bias is not accounting the immigration and emigration of population in these countries. The definition of death by natural cause vs death by suicide and arrangement of recording the mortality could be significantly different too.

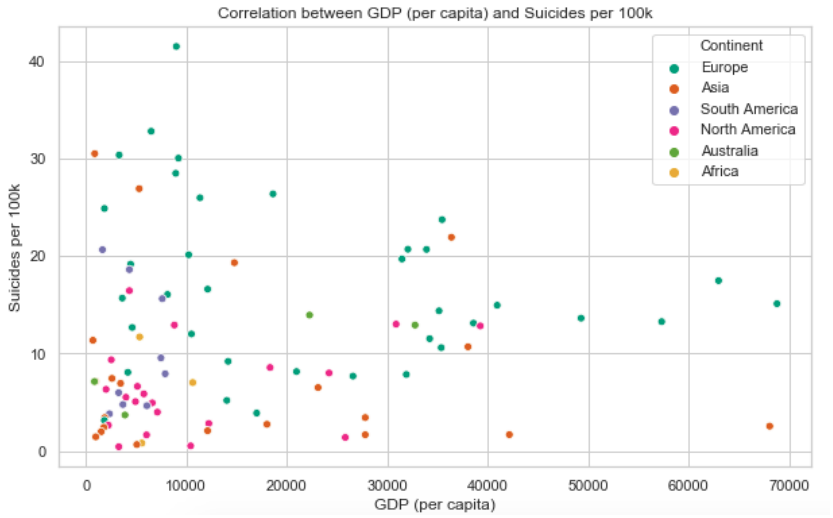
## Suicide rate and GDP

### Relationship between Suicide Rate and GDP

Money plays an important factor in an individual's life, society, economy and the country. GDP stands for Gross Domestic Product, indicating the gross income a country has for the year. Hence, it was viable to see if GDP per capita over the Years has any impact on suicide rate. A Pearson correlation calculation was performed, which is a measure of the linear correlation between two variables, valuing between -1 to +1, where correlation value of -1 indicates negative linear correlation, 0 indicates no linear correlation and +1 indicates total positive linear correlation. In our case, the two variables of interest are GDP and Year and the mean correlation calculated to 0.878. This showed a strong positive linear relationship between the Year and GDP and hence was a good indicator to explore the relationship between two variables to find out "does suicide rate in a country increase as time progresses?"

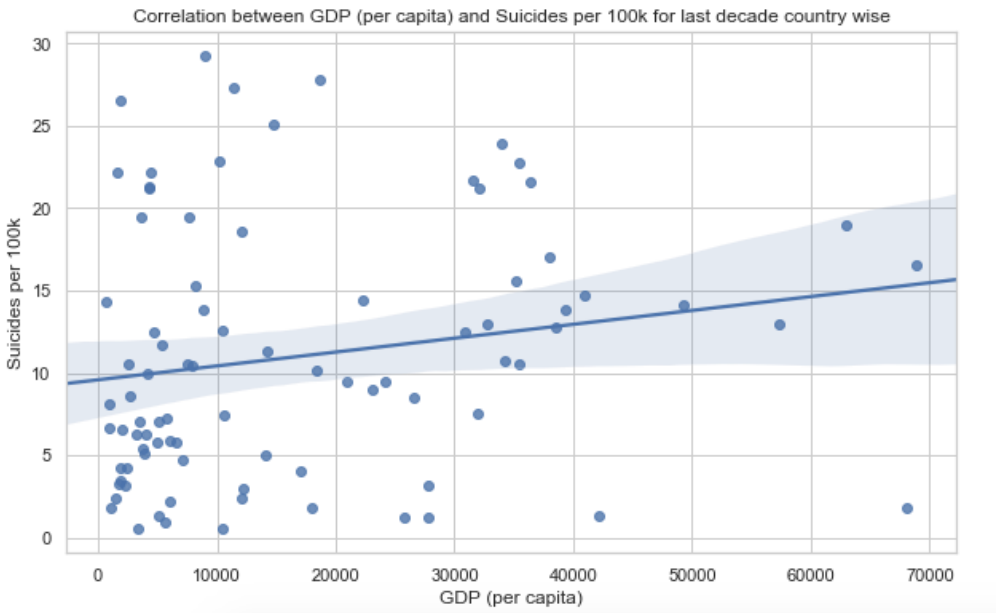
### Correlation between Suicide Rate and GDP of a country grouped by continent

It is well known that GDP and Year of a country has a strong positive correlation, as over the period of time, the GDP of country increases and with Pearson correlation test, we have confirmed GDP and suicide rate has strong positive correlation. Hence for this analysis, a mean of GDP for each Country across the years was calculated and was analyzed against suicides_per_100k.




There are quite a few high leverage & residual countries that could have a significant impact on the fit of regression line (e.g. Lithuania, top left). Therefore, we excluded those

outliers with the condition of greater than 30 suicides per 100k on y-axis and plotted the graph.




The model used was OLS Regression analysis and p-value is 0.034 which is way less than 0.05 which indicates that it is safe to reject the hypothesis that GDP (per capita) of a country has no association with suicide rate per 100k variable. But the r-squared of the model is 0.0544, so impact of GDP (per capita) on suicide rate is small. In other words, we can conclude that there is a positive linear relationship between GDP (per capita) and suicides per 100k but that relationship is weak. In short, rich countries are associated with higher rates of suicides (for ex: USA) but the relationship is weak.

## Suicide Rate during recession

The Recession during 2007-2008 has been the second most financial disaster since the Great Depression during the 1930's. hence, it was vital to see if tough time of this economic period had any impact on suicide rate in country across the world or not.

The variables used to perform this analysis an aggregation of mean of GDP (per capita), sum of population and sum of number of suicides, followed with a group by of Country and Year between the year of 2005 and 2008 inclusive.

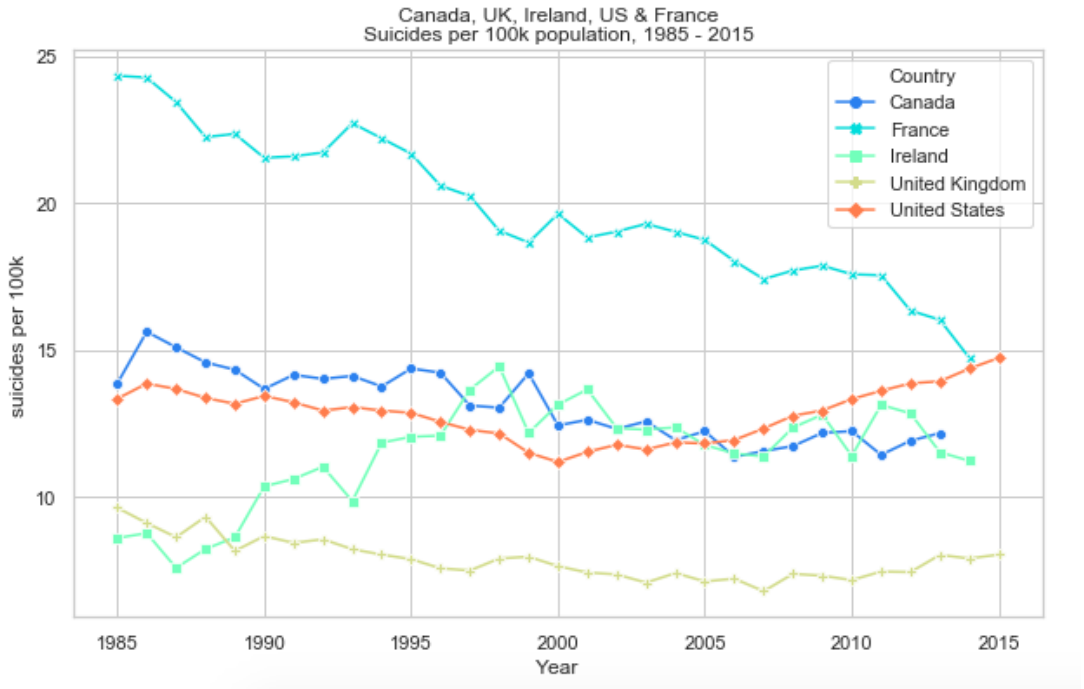
The above analysis made 2 strong conclusions, first only 26 out of 90 countries showed significant positive trend in suicide rate during the recession. Almost every country in the world was impacted by recession but only 26 countries had their population affected by its GDP (per capita).

"Correlation between GDP of Countries and their suicide rate during Years 2005-2008 =-0.10534690279838384"

The second finding was that the suicide rate turned out to be negative. This implies that suicide rates went down with the year, however the correlation is weak. The weak correlation combined with negative rate signs that data is not strong enough conclude that suicide rate during recession went down. There could be various reasons for this; for one being that the amount of data was not enough to fit the Linear Regression, hence the years chosen were 2005-2008, instead of peak recession years of 2007-2008.

## Trends in Developed Countries

Since the information above wasn't enough to reach a concrete conclusion, we decided to narrow down our search scope to five well developed countries where we know they have good systems to record mortality rate and reasons. Hence, the five countries chosen for this study were Canada, America, UK, France and Ireland. the suicides_per_100k and Years between 1985 to 2015 were the variables used to gather the overall trend analysis.

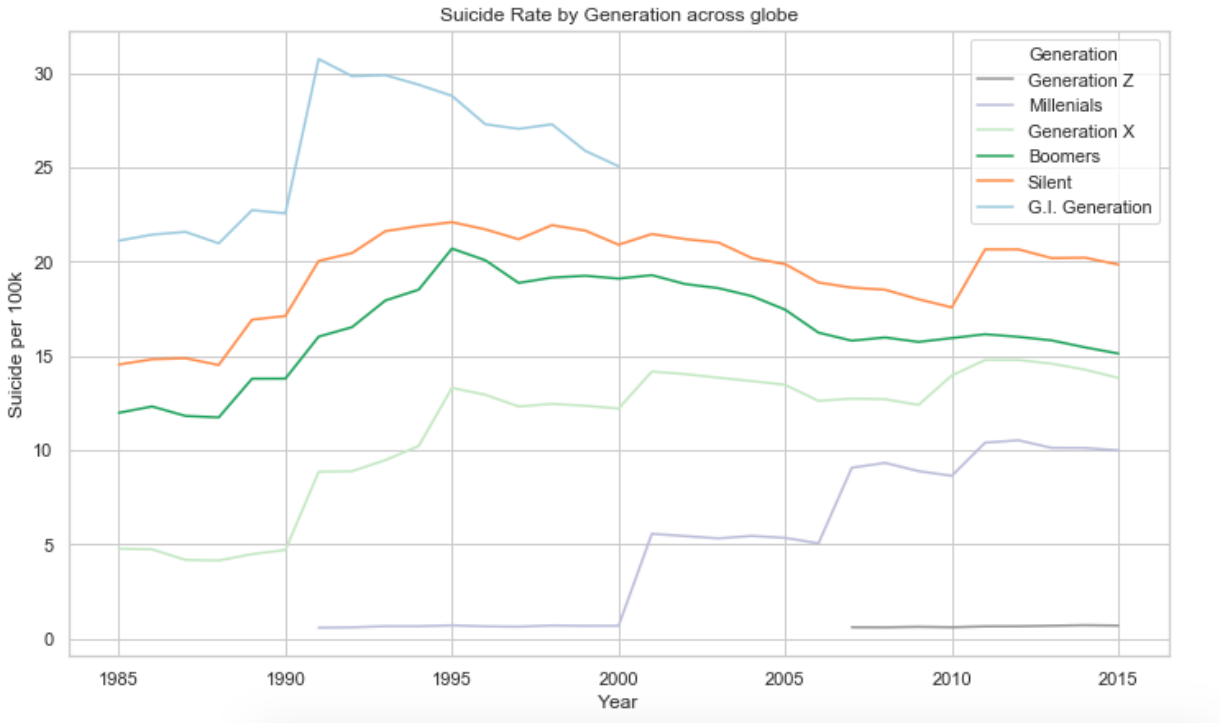



The analysis found to be that UK suicide rate has been consistently lowest since 1990, and has remained fairly static since approximately 1995. On the contrary, France historically had

the highest rate, but is now roughly equivalent to America which has the most concerning upward trend, linearly increasing by ~1/3 since 2000. Canada is showing a downward trend, where its suicide rate is consistently lowering been since 1995 and is at its lowest rate since 1985.

## Suicide Rate and Generation

The generation variable in Suicide Rate dataset is divided into 5 intervals, with age 5-14, to the age of above 75+. The general hypothesis according to society is that teenagers are impulsive, hence they are more prone to commit suicide. Moreover, its viable to find out which generation is more susceptible to suicides and also analyze the trend over the years across the world. The variables of Year and Generation were grouped by to achieve the following analysis:

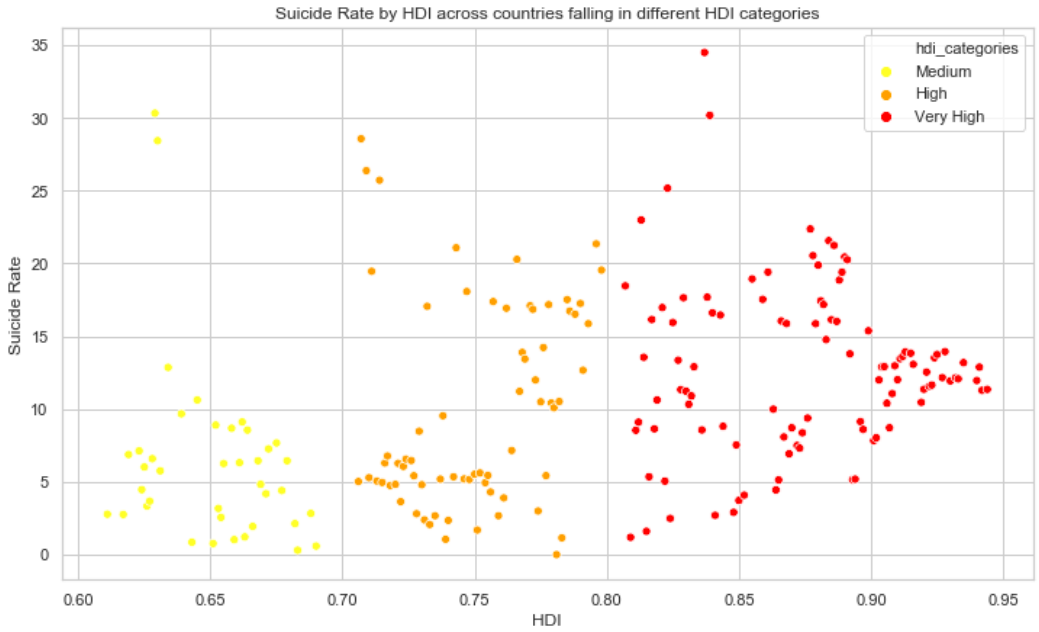



From the graph above, the Boomers generation (born 1946-1964), Silent (born 1921-1945) and Generation X (born 1965-1976) has higher tendencies of committing suicide than the rest. The suicide rate for Generation Z (born 1996 onwards) has plateaued since 2006 but we cannot be assertive about this investigation due to lack of data. Moreover, the suicide rate for Millennials (born 1977 - 1995) is increasing at an alarming rate, while to the contrary, the suicide rate for GI Generation (born 1920 or before) appears to be decreasing after 1991 but, cannot be assertive due to lack of data.

## Suicide Rate and HDI

The Human Development Index (HDI) is a measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. A country scores a higher HDI when the above indicated parameters of human development is higher. As HDI for year column contains missing values to around 70%, the data is filtered for recent years i.e., 2006 to 2016 and those countries with no HDI were removed. Countries were divided based on the HDI categories - Very High, High, Medium and Low representing Human Development Index categories.

| | hdi | suicide_rate |
|---|---|---|
| **hdi** | 1.000000 | 0.344055 |
| **suicide_rate** | 0.344055 | 1.000000 |

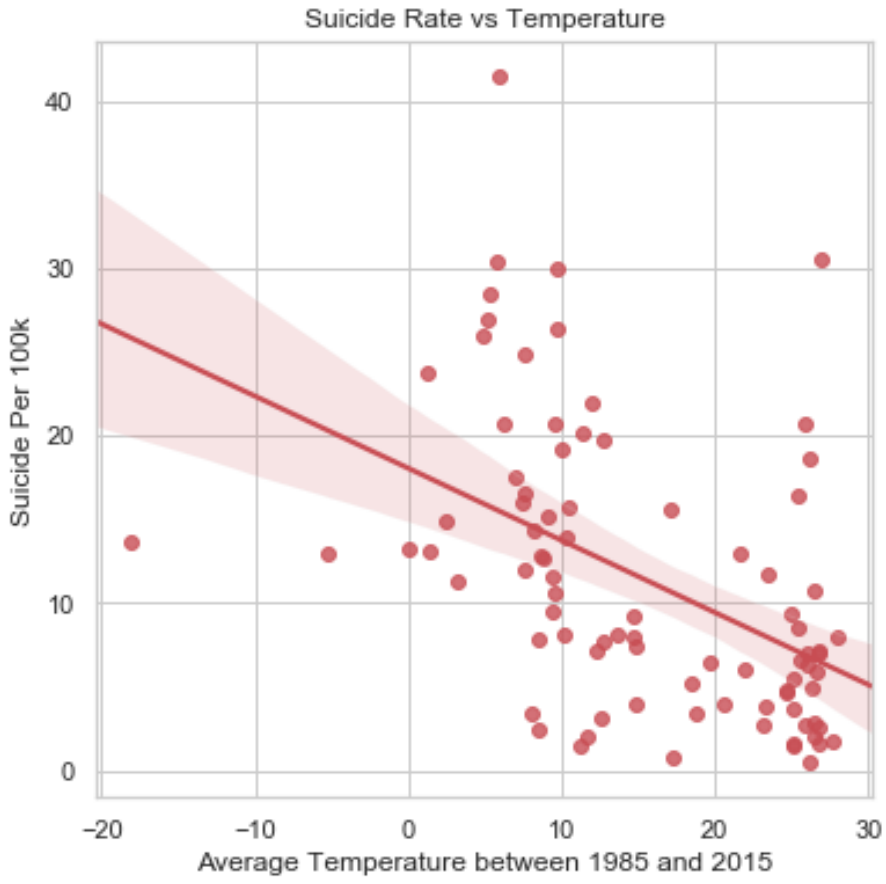



The analysis showed a correlation of 0.43, which indicates there exists a weak positive correlation between suicide rates and HDI variables, implying that the suicide rates are high among countries with high human development index.

## Suicide Rate and Temperature

Climate and temperature of a place has great impact on human body and their moods. Researchers performed a study at Stanford and stated, "By comparing historical temperature and suicide data, researchers found a strong correlation between warm weather and increased suicides. They estimate climate change could lead to suicide rate increases across the U.S. and Mexico." (Horton, 2018). If temperature of a place is too hot, it could cause dehydration and in poor or developing countries, many young kids and elderly citizens die due to heat strokes. On the other hand, if a country has located on earth where winters are too long, that is less amount of sunshine reaches the surface, those people are very highly prone to depression moods and hence higher chances of committing suicide. One can also argue that sun exposure is related to melatonin, which in turn is related to sleep quality and sleep quality is related to mental health.

The temperature dataset was filtered to keep only Year, Country and Temperature fields. This dataset was joined with Suicide Rate dataset with Year and Country as being the common fields or in other words, treated as foreign keys among the two. The temperatures were averaged for each Country across the years and plotted against suicide_per_100k to derive the following analysis.




The correlation between the average temperature of a Country and its suicide rate turned out to be negative with value of -0.47. this implies that countries with warm temperatures has lower changes of suicide rate. Indeed the data shows temperature plays a crucial role in suicide rate of the country, however temperature as an isolated variable does not provide sufficient to conclude the entire hypothesis. If the dataset had other variables such as air quality, pollution levels, etc., it could have solidify the results and analysis to content.

## Conclusion

The data gathered from Suicide Rate and Temperatures dataset had good quality of information which helped find some key insights. To begin with, the global trend analysis

showed the suicide rate are declining across all countries. Through this analysis, it was also found out that on average, the chances of an individual committing suicide increases as the age of a person increases. This was evident when analysis were performed by continent and results were positive for America, Asia and Europe but not for Africa. It makes sense as an individual steps into the real world, their challenges increases. This lead us to think and believe the economic situation of a person would be one of the key factors impacting the suicide rate. This was validated by performing Pearson correlation calculation between GDP (per capita) and suicide_per_100k of a country. The correlation turned out to be 0.87, which shows there is a positive correlation between the two variables that is increase in GDP would have some sort of impact on suicide rate of the country. The hypothesis was that a higher GDP would result in lower suicide rate; however to the contrary it was a *weak* positive relationship between the GDP (per capita) of a country and its suicide rate. It was found that rich countries also have higher rates of suicides for example USA. This forced us to narrow the scope of analysis and focus during the Great Recession period which occurred in 2008. To our surprise, the suicide rate turned out to be negative when we expected it to be positive with the assumption that in recession, a lot of people might have gone bankrupt, lost their home or job due to low economy and hence, the suicide rate must have gone up. This lead us to conclude that either the quality or quantity of data was not up to par for this analysis.

The data was also explored to determine the discrepancy in suicide rate between male and female gender. The suicide rate for men is approximately 3.5 times higher than women. This could be also due to overrepresentation of men in suicide deaths globally at both continent and country level analysis. It was also found out the highest suicide rate every recorded in demographic for a year was 225 deaths per 100k population among males, which occurred in the year of 1995. Nonetheless, the variable of HDI which represents human development index indicating the quality of life in country for the age group as very high, high, good, low or very low was also explored in this analysis. The HDI values were available only from 2006-2016, hence 70% of its data was missing values, and these values were dropped instead of filling them with nulls to avoid bias. Thus, this variable is although the latest, but it is not considered as the most relying indicator for suicide rate analysis.

And lastly, the temperature of a country for the given year was also considered to determine its impact on suicide rate. The correlation between the suicide rate and the average temperature of a Country was a negative relation, with -0.47, which in turn implies that countries with warm temperatures has lower changes of suicide rate. However, a standalone temperature variable did not appears to be sufficient enough to conclude for something as strong and significant as suicide rate trend. If the data was available along with other features such as air quality, water quality, etc. it could have been a reliable factor.

All in all, the data available from the 90 countries, 2/3 of these countries show an explicit linear downward trend over time. This paints a positive hopeful picture of our society for the future. In the 20$^{th}$ century, our society faced two world wars and the most drastic economic depression. Hence it made sense why youngsters in 1990's faced severe aftermath of these catastrophic situations and were forced to commit suicide. It is safe to say their pain-points weren't ignored otherwise we would not have seen the decline in suicide rate around the world

that we see now. There is a lot more awareness of mental health issues across the world and mental health facilities providing such services evidently has positive impact around us.

## Appendix

1. Suicide Rates Overview 1985 to 2016, Retrieved from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016)

United Nations Development Program. (2018). Human development index (HDI). Retrieved from http://hdr.undp.org/en/indicators/137506

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#

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

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

2. Climate Change: Earth Surface Temperature Data. Retrieved from https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data#GlobalLandTemperaturesByCountry.csv

3. Schumacher, Helene (2019, March 19). Why more men than women die by suicide. Retrieved from http://www.bbc.com/future/story/20190313-why-more-men-kill-themselves-than-women

4. Horton, Michelle (2018, July 23). Stanford researchers find warming temperatures could increase suicide rates across the U.S. and Mexico. Retrieved from https://news.stanford.edu/2018/07/23/warming-temperatures-linked-increased-suicide-rates/