# Visualization on Suicide rates

# A Data Analysis

**Introduction:**

An extremely complex issue that causes pain to hundreds of thousands of people every year around the world is death by suicide. Globally, 800,000 people die every year due to suicide, which when calculated is one person every 40 seconds. One of the leading causes of death that occur throughout the lifespan is a global phenomenon. Sixty percent of deaths from firearms result from suicide, which is particularly high in the US. There are many indications that there may have been more than 20 others attempting suicide for each adult who died by suicide. To prevent suicide and suicide attempts, evidence-based and effective interventions are implemented at the population, sub-population, and individual levels. Through this project, I wanted to analyze this very anomaly and explore whether there is a relationship between the cause behind these suicides and the suicide rates in the country with the available data on this topic.

**Related Work:**

Several individuals have been working on this analysis, several papers were published, and journals have also been released. In one of such studies, Barclay D. Johnson attempts to clarify the cause of suicide rates. The two social variables, integration and regulation jointly determine rates of suicide. An extreme condition of integration or of a regulation or by some combination of extreme conditions is caused at a high rate. Another study on the impact of the geographic characteristics on suicide rates by Mayumi Oka, Takafumi Kubota, and Keita Yamauchi. They calculated the standardized mortality ratio from suicide statistics of 3318 municipalities from 1972 to 2002. To find the relation between topographic and climatic variables and suicide rates, they used Correlation analysis, generalized additive model, and multi-regression analysis. They used the Geographic Information System to visualize these relations.

As the causes of suicide are complex, pinpointing the reasons that suicide rates rise or fall, made it challenging in part. Some of the risk factors for suicides include health factors such as serious mental illness, depression, and health conditions, environmental factors, and historical factors. Experts point out that though the climbing rate is cause for concern, they do not tell the whole story. In other words, there is no single, easy solution for suicides nor an obvious culprit for an increase in the suicide to turn the trend around. In these studies, they take each community into consideration for the suicide rates, but for our visualization, we are going to analyze and find out the related facts that cause an increase in the suicide rate.

**METHODS:**

**Dataset Description:**

Out of the data sets I chose Kaggle’s dataset as the dataset allowed me to explore and build models in a web-based data-science environment. The dataset which has been chosen for this project contains information pulled from four other datasets that are linked by time and place. It was built to find signals correlated with increased suicide rates across the socioeconomic spectrum, among different cohorts globally. The dataset provides useful information such as the year in which the suicide occurred, the population for each place. The dataset has information from the year 1985 to the year 2016. Apart from this, the dataset gives us the sex and age of the population in a specific country. We compare the results from male to female from the sex, for an age ranging from 5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, and 75+ years. The number of suicides is calculated for each place based on the information and then the suicides are calculated from 100k population. Additionally, the dataset gives extra information such as country-year, HDI for "year", "gdp\_for\_year", "gdp\_per\_capita", and "generation".

The dataset divides the generation where all the people born and living at about the same time into 6 sub-groups. The individuals born between 1928 and 1945 are considered as Silent Generation. Millenials known as the Generation Y are born in between 1981 and 1996 (a widely accepting defining range). Generation Z uses the mid-to-late 1990s as starting birth years and the early 2010s as ending birth years. Roughly from the early 1960s to the late 1970s are considered as Generation X. The G.I. Generation, also known as the greatest generation and the World War II generation are generally born between 1901 and 1927. The generation most often defined for the individuals born between 1946 and 1964 are considered Boomers.

**Data Cleaning:**

There are different techniques that are possible to clean the data out. First, the countries which do not give enough information to present the suicide rates are removed. A total of 8 countries was removed which is less than or equal to 3 years of total data. Then the years that were missing information have been removed. The year 2016 data has been removed (this extremely less data containing countries often had missing information). One of the columns, HDI, though it additionally gave us the extra information, as two-thirds of the column's data is missing it has been removed.

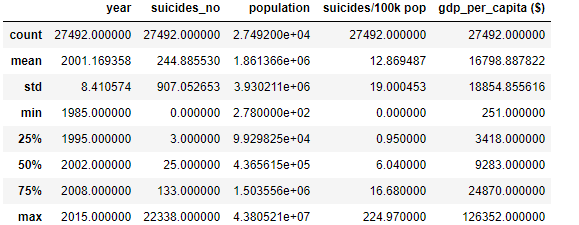
To perform data filtering and manipulation pandas package has been used for this dataset. For this dataset, a fundamental task was adding continent to the dataset using the python package, pyocountry\_convert. The pycountry, extension of the python package, provides conversion functions. This package provides conversion functions between ISO country names, country-names and continent names using the country data which is derived from Wikipedia. The pycountry-convert module is built on python 3 and has dependencies on several python modules that are available within the python package Index PyPI.

**Exploratory Data Analysis (EDA):**

The critical process of performing initial investigations on data to discover patterns, to spot anomalies, to check assumptions, and to test the hypothesis with the help of summary statistics and graphical representation is referred to as exploratory data analysis. Data Scientists work with EDA of the several processes, after acquiring data. This helps in understanding the data by visualizing it with several plots for investigating its characteristics. This technique not only allows data, scientists to know the spread of the information, but provides insights to help devise a plan for the projects. The most widely used languages for data analysis is python programming. This programming language has its own advantages and disadvantages of carrying out different processes of analysis. The data investigation through python is often carried out with matplotlib and Seaborn. Matplotlib requires several changes for appealing plots which are cumbersome.

**Statistical Analysis:**

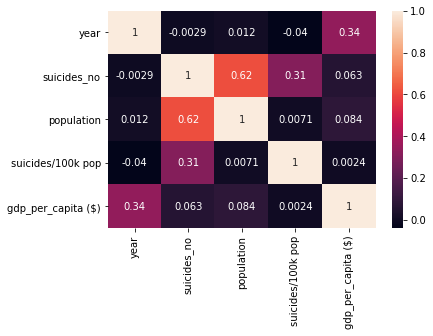
When it comes to building complex analysis pipelines that mix with statistics that include text mining, image analysis, or control of the physical experiment, python is an invaluable asset. Statistical Analysis is about analyzing and summarizing data. One of the main approaches is the quantitative approach. This approach describes and summarizes the data numerically. The statistics about the dataset has been explored in figure 1 below. The dataset gives us information about a total of 27492 counts from the years 1985 to 2015. The maximum number of suicides count is 22338 starting from 0 which can be derived from all the 27492 rows of information. The population in the year 1985 escalated from 2.78 to 4.38 in the year 2015. Through the assessment from the number of suicides and the population figures, it can be observed that the suicide rate for 100k population has ascended from 0 to 224.97 (0.95 to 25% of the population, 6.04 to 50% of the population, and 16.68 to 75% of the population, with a maximum of 224.97 of suicides for the population). It can be observed that the GDP per capita proliferated from 251$ to 126352$ (the mean of which is around 16800 and a standard deviation of around 20,000).



**Figure 1: Statistical description**

**Heatmap:**

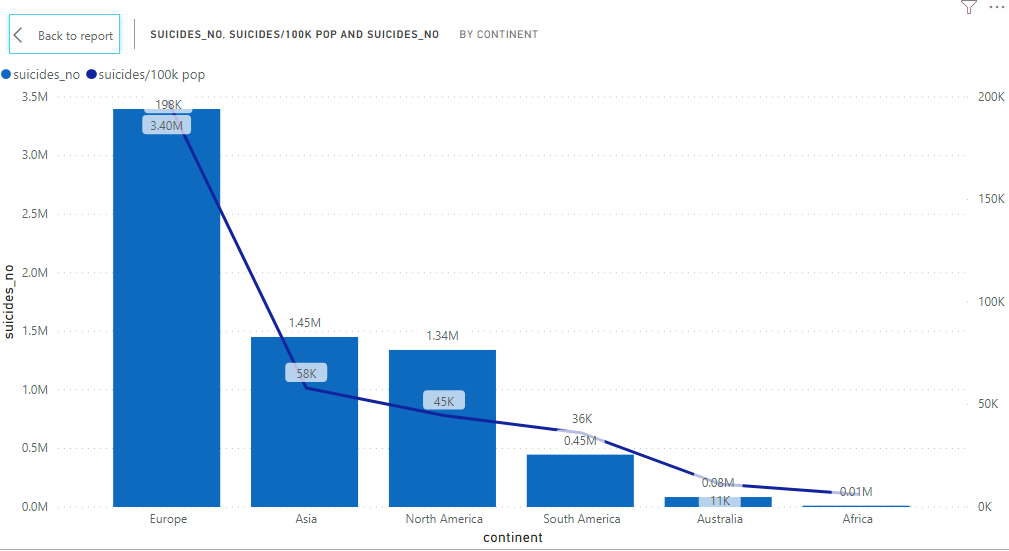
A heatmap uses color in order to communicate the values to a reader. When you have a large volume of data, this is a great tool to assist the audience towards the areas that matter the most which are demonstrated in the below figure 2.



**Figure 2: Correlation Heat map**

While considering the inverse nature of the relationship between “year”, and “suicides\_no” as well as “year”, and “suicides/100k pop”, they share a strong negative correlation. Even though the “population and suicides/100k pop” and “population and suicides\_no” have a strong positive correlation, “population and suicides\_no” have a higher correlation in comparison. The total number of sucides (suicides\_no) slightly decreases as the GDP of the country (gdp\_per\_capita) increases and as the correlation value of the fields “gdp\_per\_capita” and “suicides\_no” is closer to 0.0 (0.063). Through this heatmap, “year, population, suicides\_no, and gdp\_per\_capita” can be considered as the columns that matter most for further analysis.

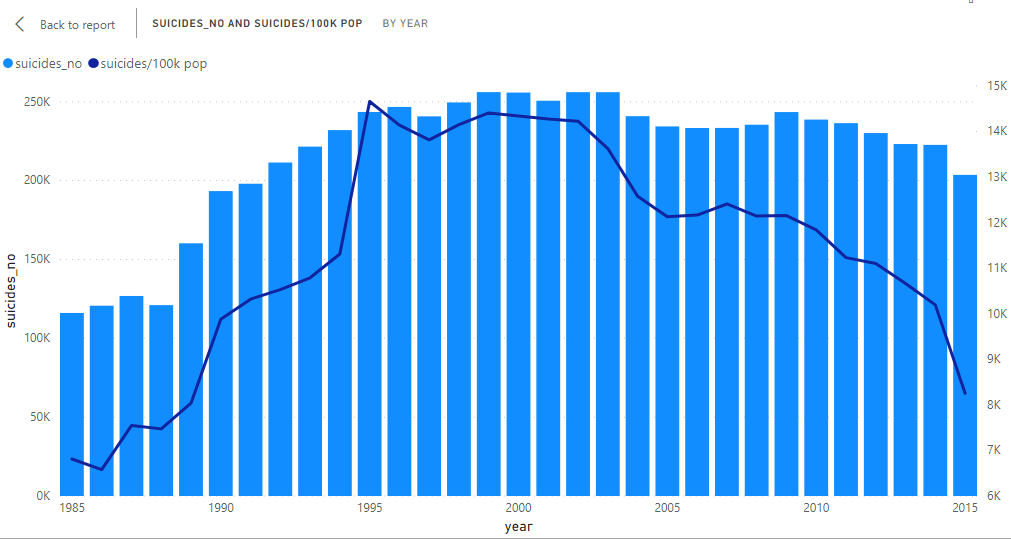
For further analysis we considering 93 countries and 6 continents. From the following visualization in figure 3, it can be observed that the Europeans have the highest overall of the “suicides\_no and suicides/100k\_pop” and the African continent has the least.



**Figure 3: Visualisation showing number of suicides per continent**

And through further analysis, it is observed that, the country Russian Federation has the highest number of suicides overall the 93 countries and the Antigua and Barbuda country has the least.

Over the period of 30 years (1985-2015), the number of suicides for the years 1999-2003 have the highest suicide rate and through the late 1980s and the early 1990s, it is the lowest. When observed from the figure 4 below, the rates from 2015 are now returning to their early 90s rates.



**Figure 4: Visualisation showing number of suicides per continent**

The dataset has information for 7M population that committed suicides from which 5M are men and 2M are women, approximately. Therefore, while considering the rate of suicide globally, the rate of suicide for men has been 2.5X higher for men approximately.

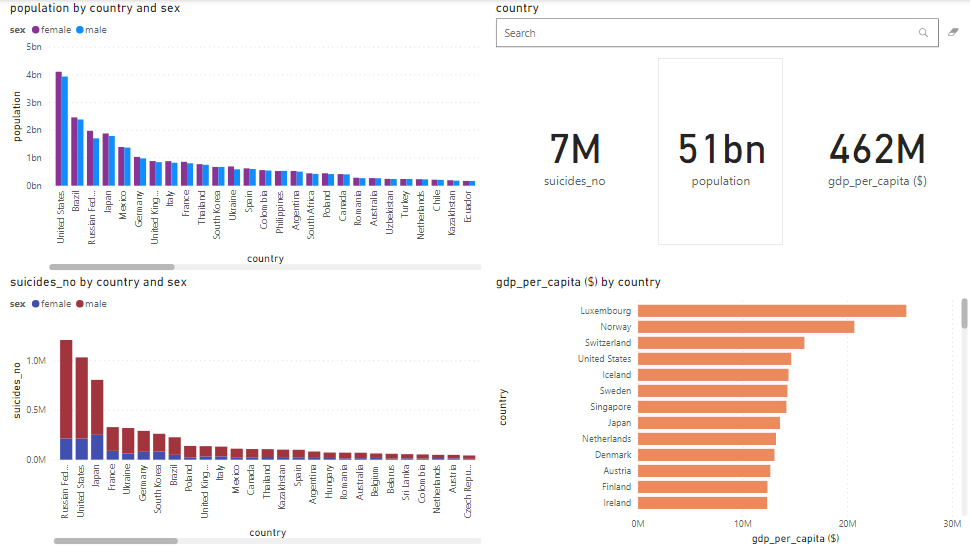
**Hypothesis:**

Till now we have seen a few observations in EDA, the following are my hypotheses for this dataset. Initially, the population and the total number of suicides are positively correlated: if the population is higher then the number of suicides are higher and the GDP (“gdp\_per\_capita”) and the total number of suicides (“suicides\_no”) are negatively correlated: the higher the GDP, lower the number of suicides. Next, the rate of suicide is highest in men through age. Lastly, we hypothesize that the older generation has the highest suicide rate and the newer generation has the lowest.

**Results:**

I came across a wide range of tools to obtain the results by creating the visualizations. One of the great tools to use for data analysis and discovering important insights is Power BI. The reason behind choosing this tool is that it helps create and share interactive data visualizations apart from creating communicative visualizations.

The first visualization from the figure 5 shows us the relation between “population”, “suicides\_no”, and “gdp\_per\_capita” according to the hypothesis. A Clustered column chart has been used to build the visualization between the population in each country and the contribution of males and females among the population. The visualization for the country and their respective “suicides\_no” has been shown using a Stacked column chart. The demonstration of visualization between the country and their respective GDP has been done using a Clustered bar chart. Through the report below, it can be observed from the visualization “*population by country and sex*” that the United States, Brazil, followed by Russian Federation have the highest population and Aruba Kiribati has the least population and it has been observed that most of the countries have a higher female population. The visualization “*suicides\_no by country and sex*” gives us the information that the countries, Russian Federation, United States, followed by Japan has the highest number of suicides and the countries Antigua and Barbuda, followed by the Maldives has the least number of suicides. Considering the top three countries for the highest number of suicides, (Russian Federation, United States, Japan) which are among the top five countries for the highest population, (United States, Brazil, Russian Federation, Japan, Mexico) and countries with least number of suicides tend to have the least population. This gives us the result that the visualization report supports our assumptions: if the country has the highest population then the suicide rates are high. Similarly, from the visualization “*gdp\_per\_capita ($) by country*,” the top three countries (Luxembourg, Norway, Switzerland) with the highest GDP have the least population and least number of suicides. On contrary to the assumption, the GDP is high, for the countries, the United States, and Japan, which has the highest population and the highest number of suicide rates.

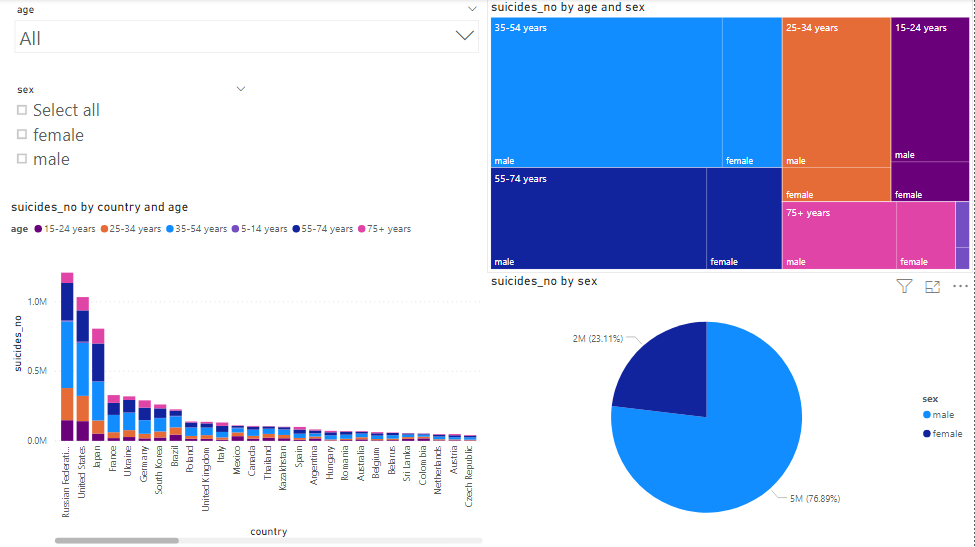


**Figure 5: Visualisation report for hypothesis 1**

The interaction in this report allows us to click on any country and search with the country name to filter that specific country statistics. A visual card has been used to display each statistics (suicides\_no, population, and gdp\_per\_capita) for the visualization and the column “sex” has been plotted as a legend for the fields “population” and “suicides\_no” and to differentiate male from female sex, colors have been assigned to them.

For the total scenario, considering the top few countries with high GDP has less population and the least number of suicides. This gives us the result that, the number of suicides (suicides\_no) is positively correlated with population and both are negatively correlated with GDP (gdp\_per\_capita).

The second visualization report has been demonstrated in figure 6, to show the total number of suicides in the country and the age of that specific country’s population. And the suicide rate for men and women has been shown through age. A Stacked column chart has been used to build the visualization (“*suicides\_no by country and age*”) between each country and the total number of suicides in that specific country and the column “age” has been plotted in legend. Treemap has been used to demonstrate the visualization of the age groups and their respective male and female count by the total number of suicides committed. Through the report below, it can be observed from the visualization “*suicides\_no by age and sex*”, that the age groups 35-54 years, followed by 55-74 years and 25-34 years has the highest number of suicides.



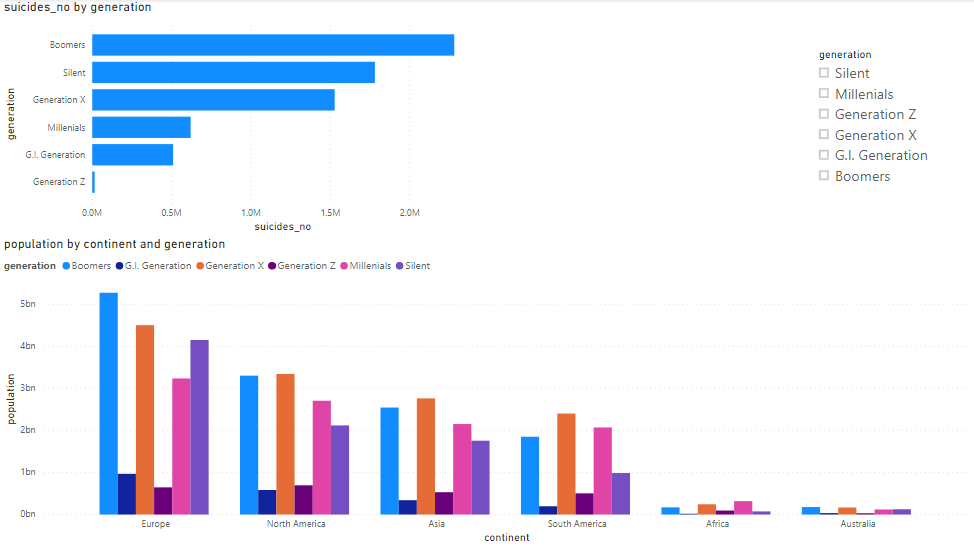
**Figure 6: Visualisation report for hypothesis 2**

From this visualization, it can be observed that, when considering the country's population and age groups, men have the highest suicide rate. A Pie Chart (*suicides\_no by sex*) has been plotted to support this, which shows the male population has the highest suicide rate.

The interaction in this report has been done using Slicer, which allows us to filter the age groups through a drop-down list. The filtering of the report by sex (male and female) can be done using radio buttons that also use Slicer for the implementation. Coming to the total scenario of the report.

The final visualization report has been demonstrated in figure 7 to show the total number of suicides in each generation and to show the population of each generation in each continent. A Clustered column chart has been used to build the visualization between each continent and the total population in that specific continent and the column “generation” has been plotted in legend. Through the report below, it can be observed from the visualization “*population by continent and generation*”, that compared to any continent the generation, Boomers, Generation X, Silent has the highest population. A Stacked bar chart has been used to build the visualization to show the number of suicides in each generation. Here in this visualization “*suicides\_no by generation*”, it can be observed that the number of suicides is highest in Boomers, followed by Silent, Generation X, Millennials, G.I. Generation, and Generation Z. From these visualizations, it can be observed that the generation that has the highest population has the highest suicide rate which supports our first hypothesis. A Slicer has been used to filter the generations using a list.

For the final report on this visualization, it can be shown that through G.I. Generation is the oldest among all generations, it does not have the highest number of suicides which contradicts our hypothesis. But it also has been observed that the newer generations have less number of suicides relatively.



**Figure 7: Visualisation report for hypothesis 3**

**Discussion:**

This project uses custom visuals to create an interactive visualization of multivariate data. The primary visualization in the report has been built for our first hypothesis i.e, the country has the highest population if the suicide rates are high. The visualizations carried out gives us the information that the top three countries with the highest GDP have the least population and the least number of suicides. Contrary to this hypothesis, the top few countries with high GDP have a high population and the highest number of suicides. But in a substantial scenario, when we consider the top 30-40 countries with high GDP, the population and the suicide rates are low. This gives us the result that, the total number of suicides is positively correlated with population and both are negatively correlated with the GDP. Another report has been made through building a visualization to show the total number of suicides in the country and the age of that specific country’s population differentiated suicide rate for males and females through age. The observation that has been made is that higher age groups have the highest number of suicide rates. A conclusion has been made supporting our hypothesis that for the country's population, men for the higher age groups have the highest suicide rate. The final analysis gives us the information that, the generation that has the highest population has the highest suicide rate. Here, it can be observed that this statement supports our first hypothesis. Through this analysis, the oldest generation among all does not have the highest number of suicides which contradicts our final hypothesis, though the newer generations have fewer suicides relatively.

The dataset I have used gives us information only until 2015. It would have allowed us to contextualize the findings better with more recent events if we had the data until 2020 or at least 2019. It would have been helpful if all the countries were considered in the dataset that did not have a large jump or variation in time. The small shifts, in this case, could be due to the individual changes in few countries which would have made it easier to point out.

First, many gaps in our knowledge remain, although there has been a welcome focus on suicide prevention interventions. We hope that the suggestions reported will encourage new strategies and opportunities for interdisciplinary research. In the end, we are hopeful that the developments and the field’s assurance to defeat the distinguished difficulties will consolidate to spare more lives over the globe.

**Furture Work:**

We have summarized the key risk factors for suicide and given a comprehensive, data-driven report on the suicides. This expanded data is collected to address suicide prevention using advanced and updated strategies. To prevent suicides, the interventions should take place from the universal level to the individual level. Here, the one-size-fits-all approach will not come in handy. While determining the measures of suicide prevention, it is necessary to take the characteristics of each community and everyone into consideration.

In this study, we focused on the causes of the increase in suicide rates and the source behind these suicides. There is no research on other factors such as the measures against committing suicides such as the factors that influence the mental health of the individuals (such as deterioration of social network, quality of medical services, or decreasing birthrate).

To date, the prevention factors have not been studied extensively or rigorously as the risk factors. Hopefully, identifying and understanding, prevention factors will be equally as important as researching the risk factors for suicide.

**References:**

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

Oka, M., Kubota, T., Tsubaki, H., & Yamauchi, K. (2015). Analysis of impact of geographic characteristics on suicide rate and visualization of result with Geographic Information System. Psychiatry and Clinical Neurosciences, 69(6), 375-382. <https://doi.org/10.1111/pcn.12254>

Johnson, Barclay D. “Durkheim's One Cause of Suicide.” *American Sociological Review*, vol. 30, no. 6, 1965, pp. 875–886. *JSTOR*, [www.jstor.org/stable/2090966. Accessed 24 Apr. 2020](http://www.jstor.org/stable/2090966.%20Accessed%2024%20Apr.%202020).

AFSP (American Foundation for Suicide Prevention). 2001. Reporting on Suicide: Recommen dations for the Media . [Online]. Available: [http://www​.afsp.org/index-1.htm](http://www.afsp.org/index-1.htm) [accessed December 5, 2001].