Suicide Rates 1985 - 2016

# Summary

Based on descriptive and predictive analyses, it is recommended that all countries and municipalities immediately implement suicide prevention programs. The rates of suicide are greatest among older men, so it is recommended that this demographic be focused on. Across all demographics, the stigma of mental health needs to be addressed throughout the world. The apparent direct relationship between gun prevalence, internet accessibility, and suicide rates must also be addressed to reduce the rates of suicide.

Nash Delcamp

Keith Kunz

Allison Roth

Table of Contents

[**Summary**](#_vtpa15lyrwdm)1

[**Specification**](#_hrjio33o1zch)3

[**Observation**](#_g2rolh7ufwjh)3

[**Analysis**](#_wr5bfjl1btyi)5

[K-Means Clustering](#_1durryxovp87) 5

[Multiple Linear Regression](#_yixh8ki2a53n) 9

[Random Forest](#_aksnrmnjurgf) 10

[**Recommendation**](#_h3cw54xmhidm)11

[**References**](#_xv8szpa7keba)11

# Specification

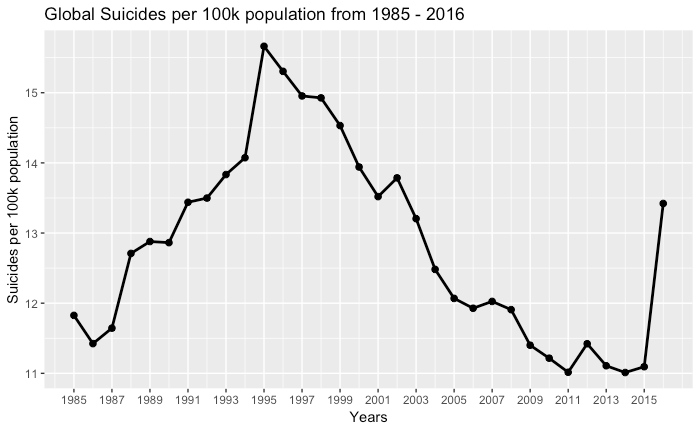
Suicide is a global public health issue. Many of us know a friend, family member, or acquaintance who has attempted or committed suicide. According to the World Health Organization, every year approximately 800,000 people commit suicide, while even more attempt suicide¹. Every suicide has a lasting effect on the families left behind and their communities.

Suicide can be prevented, but countries need to make suicide prevention a priority. To date, only 38 countries report having a national strategy for suicide prevention². These strategies vary from treatment and support services to further research on suicide prevention. The goal of this analysis is to take a deep dive into the World Health Organization’s Suicide Rate Overview from 1985 - 2016, and determine what factors contribute to the suicide rate to better understand what drives the number of suicides, and potentially, how to reduce the unnecessary deaths.

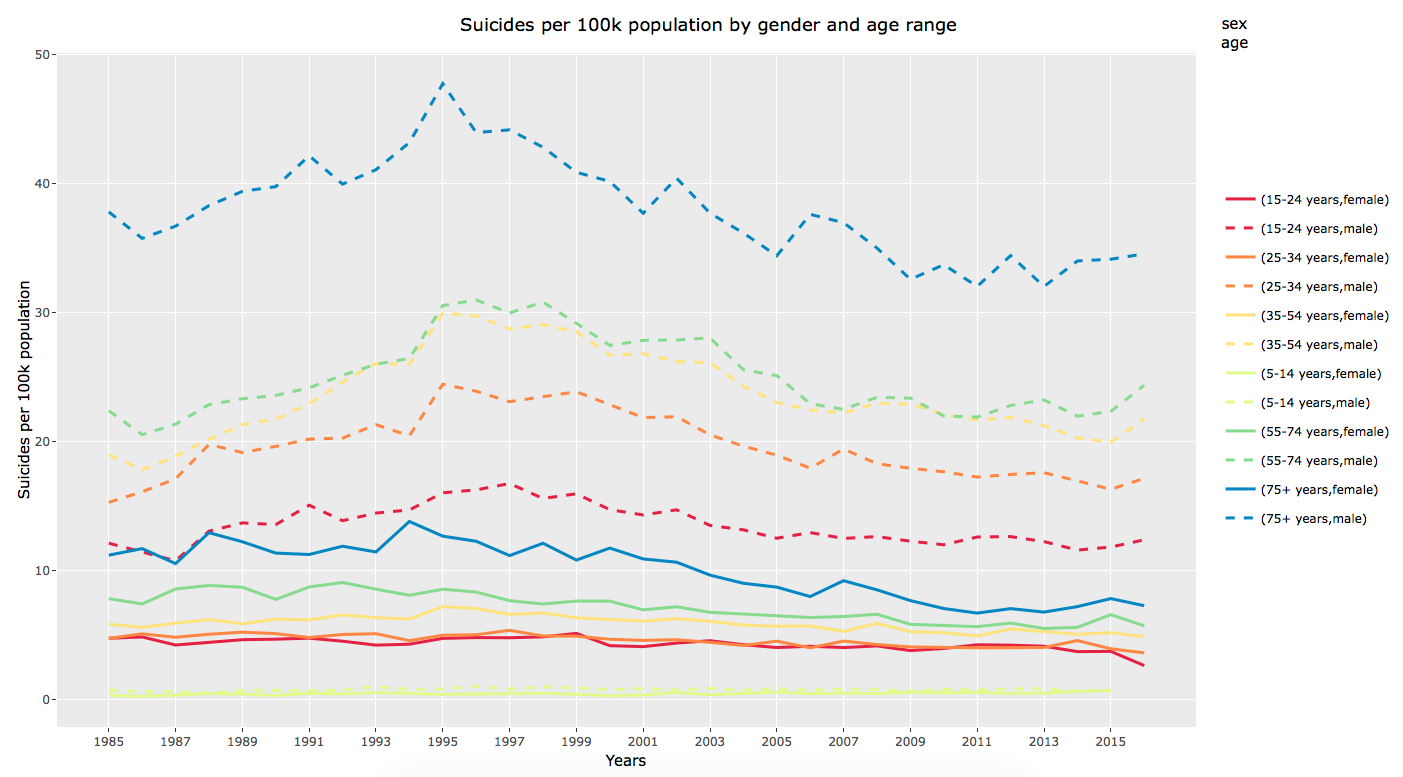
Many factors were considered that may contribute to suicide rates, including access to guns, social media, computer access, universal healthcare, unemployment rate, access to mental health care, and veteran status. After further investigation, we decided to focus on access to guns and internet usage. It was hypothesized that these metrics would be directly related to suicide rates. These external data were obtained from the Guns Ownership dataset from Wikipedia³ and internet usage data from Wikipedia⁵ as well.

# Observation

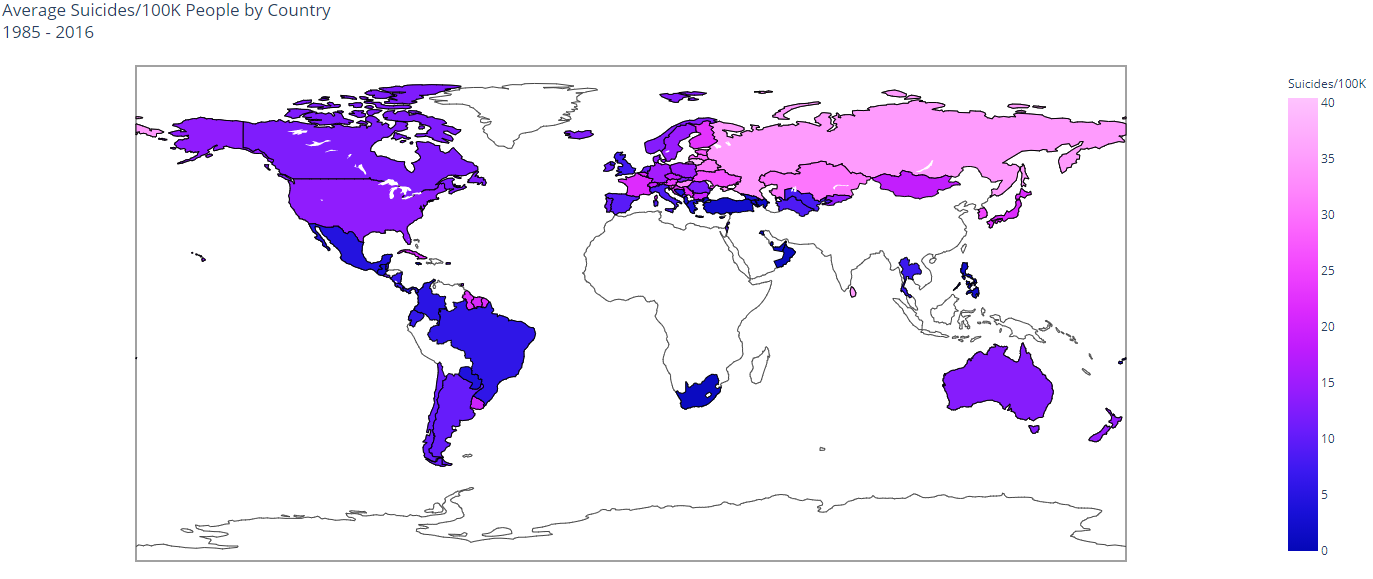
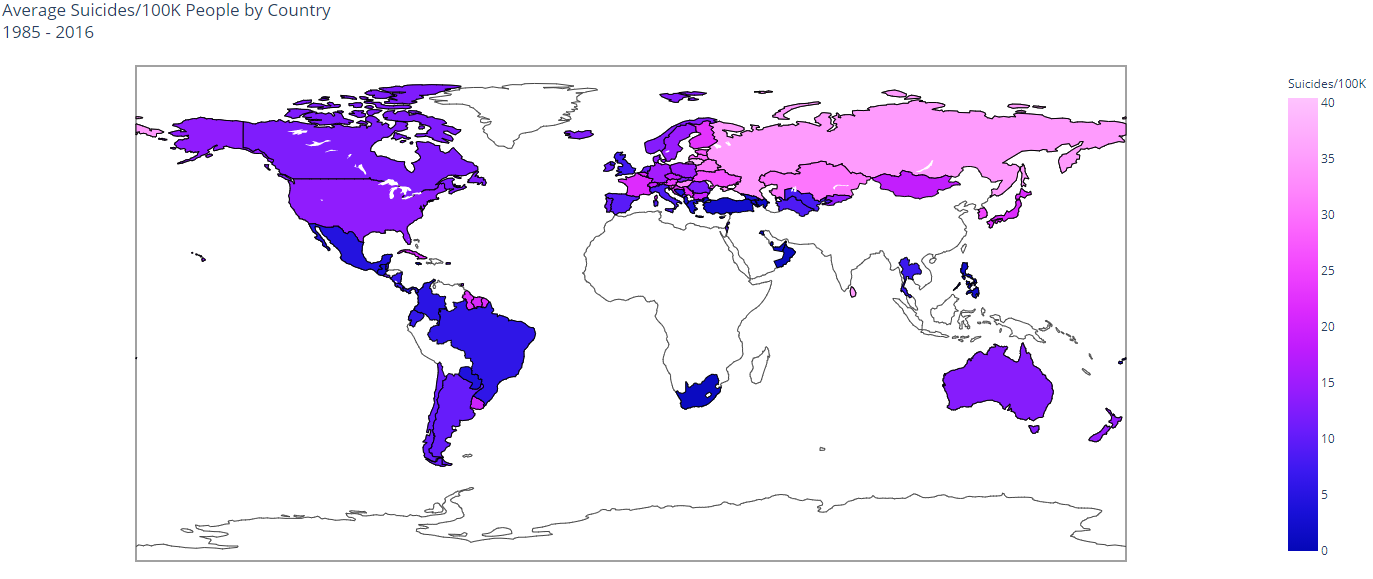
Through exploratory analysis, it is concluded that globally, the number of suicides has been decreasing since the mid-1990s after steadily increasing the decade prior. From the figure below, it is clear that the global average suicide rate is steadily decreasing. It was at a peak in 1995, where an average of 15.66 people committed suicide per 100 thousand people.



Further, there is a noticeable difference in suicide rates when grouping by gender and age. The figure below shows the suicide rates of males (dotted line) and females (solid line). For both genders, the suicide rates for 5-14 year olds are the same and consistent. Interestingly, the subsequent suicide rates by age are the same for both male and female (15-24 years, 25-34 years, 35-54 years, 55-74 years, 75+ years). Additionally, even the lowest suicide for males (not considering ages 5-14) is greater than the greatest suicide rate for females. This graph continues to support the trend of suicide rates decreasing.



The countries with the greatest suicides per 100 thousand people overall were identified as Lithuania (40.4), Sri Lanka (35.3), Russia (34.9), and Hungary (32.8). The average suicide rate for each country is presented in the map below. Darker blue represents a low suicide rate while light pink represents a high suicide rate. This map also illustrates that data are missing for the majority of Africa and Southern Asia.



# Analysis

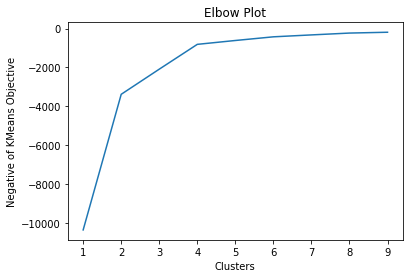
For this analysis, the World Health Organization’s Suicide Rate’s from 1985 to 2016 dataset was combined with Wikipedia’s gun ownership data set and Internet Usage dataset. 101 countries are included in the data over a range of 26 years (1985 - 2016).

The suicides, guns, and internet usage datasets were all cleaned and merged by country. Many variables representing monetary values or other numeric values contained dollar signs and commas, so all variables were stripped of dollar signs, commas, hyphens, and special characters. Country codes were merged from the Plotly Master Dataset collection⁴ and missing codes were manually determined and input. Some countries with different names in the three source datasets were converted to match the suicides dataset.

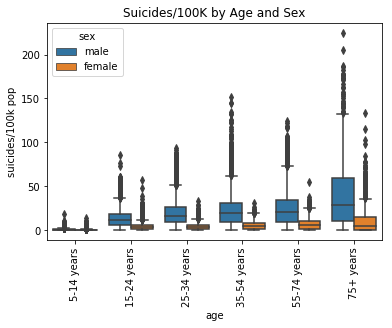
## K-Means Clustering

Two different K-Means models were built to better understand the relationships between suicide rates and other factors. The K-Means method is an unsupervised learning algorithm which minimizes the distance between the samples and cluster centers. This type of model is very beneficial for discovering patterns in the data.

The first K-Means model included all records from the dataset, i.e., all countries, years, sex, and age groupings. This model included the response variable - suicides per 100 thousand people, sex, and age group. Since the K-Means algorithm requires all variables to be numeric, the latter two variables were represented in numeric forms. Sex was represented as a binary variable - 0 represented male while 1 represented female. Age group was represented by the minimum of the age group range provided. These three variables were the only variables included because based on previous attempts, the inclusion of sex and age group would determine both the appropriate number and assignment of clusters. Therefore, other variables were not considered in this model. The number of clusters was determined by building models with clusters ranging from 1 to 9. Each model was scored based on the sum of the distance between samples and cluster centers. Based on the elbow plot below, 4 clusters is the optimal number of clusters for this model.



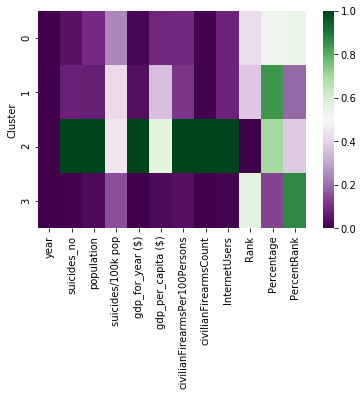
These clusters appear to simply assign females age 5 to 54 to the Young Female cluster, females age 55 - 75+ to the Elder Female cluster, males age 5 to 54 to the Young Male cluster, and males age 55 - 75+ to the Elder Male cluster. That said, there are extreme differences between the suicide rates of these four clusters. The boxplot and table below illustrate the clear difference between these four clusters, primarily that elderly males have a suicide rate twice as high as young males, nearly four times as high as elderly females, and over eight times as high as young females.



|  |  |  |
| --- | --- | --- |
| **Cluster** | **Mean Population** | **Mean Suicides/100K** |
| **Young Female** | 2,137,515 | 3.83 |
| **Elder Female** | 1,392,135 | 8.52 |
| **Young Male** | 2,162,833 | 14.55 |
| **Elder Male** | 1,078,035 | 31.59 |

The second K-Means model was built using only records from 2015 without the response variable - suicides per 100 thousand people. These data were also normalized on a min/max scale and aggregated such that each country belonged to only one record. Suicides per 100 thousand people was recalculated to adjust for unequal weighting from aggregation.

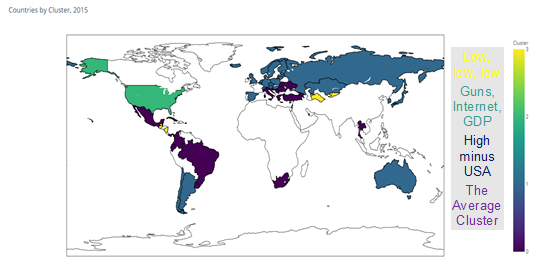
Since the goal of these clusters was to find clusters that represent distinct groups with respect to suicide rate, cluster models based on all combinations of all numeric variables that are not directly related to suicide rate were compared based on the standard deviation of the mean of suicide rate. The combinations also included iterations for 2, 3, and 4 clusters. The combination that yielded the greatest standard deviation was built with 4 clusters and variables civilianFirearmsCount and Percentage (percentage of population with access to the internet). This was a positive result, because civilianFirearmsCount came from the guns dataset while Percentage came from the internet usage dataset. A heatmap of numeric values by cluster is presented below which shows a distinct difference in suicides per 100 thousand people, GDP, gun presence, and internet usage in cluster 2.



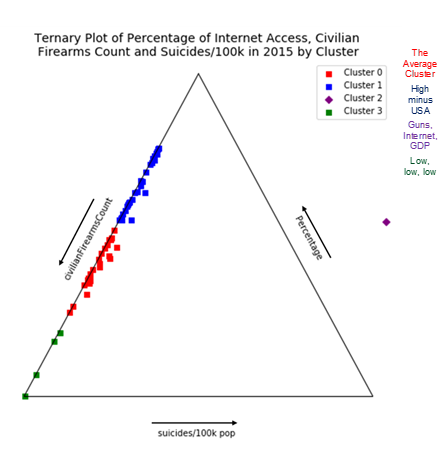
Cluster 2 was determined to contain US, only. The striking difference in total number of guns, guns per 100 people, suicide rate, GDP metrics, and internet rank between the US and the average of all other countries are portrayed further in the table below. The cluster most like the US was called High minus USA. This cluster has a mean suicide rate very close to that of the US, the next highest mean GDP, the next highest number of guns per person, and the next highest internet accessibility. The four countries included in the low, low, low cluster have low suicide rates, very low GDP, few guns, and little access to the internet. The countries in The Average Cluster are near the 2015 average for all metrics.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster** | **Suicides/ 100K** | **Pop** | **Suicides** | **GDP** | **GDP Per  Capita** | **Guns/ 100 pop** | **Guns** | **Internet Rank** | **Percent  Rank** | **Countries** |
| **The Average Cluster** | 7.9 | 30 mil | 1,918 | $300 bil | $11,037 | 12 | 4 mil | 83 | 99 | 24 |
| **High minus USA** | 13.6 | 20 mil | 3,396 | $700 bil | $38,960 | 14 | 2 mil | 73 | 37 | 33 |
| **Guns, Internet, GDP** | **14.7** | 300 mil | 44,189 | $20 tril | $60,387 | **121** | **400 mil** | **3** | 70 | US,  only |
| **Low, low, low** | 5.0 | 7.5 mil | 340 | $30 bil | $3,847 | 5 | 65K | 109 | 157 | 4 |

The map below is colored by cluster. This helps to identify the countries clustered together as well as countries missing from the dataset used for this model; note that not all these countries are missing for all years.



To conclude this clustering model, the below ternary plot demonstrates the separation between the US and the countries in all other clusters while also showing that the variables clustered by are appropriate in modeling suicide rate.

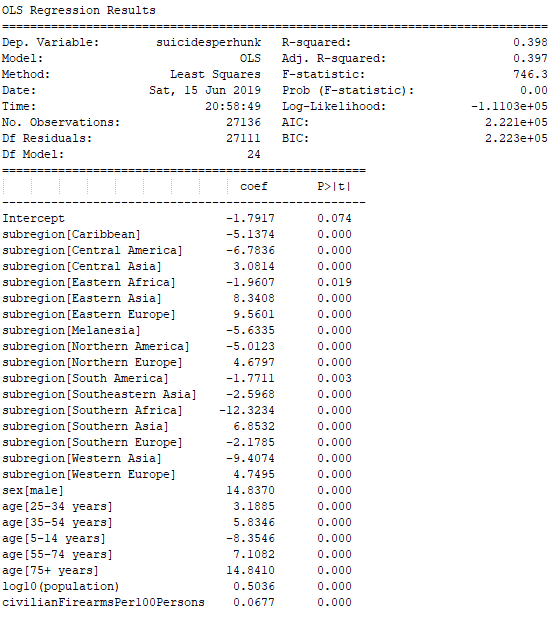


## Multiple Linear Regression

Using multiple linear regression to determine the largest factors for determining suicide rates was accomplished by testing many combinations of predictor variables. Variables that were not significant were iteratively removed, starting from the variable with the largest p-value. Only 5 variables remained in the final model for determining suicide rates. Those variables include the subregion, sex, age, the log of the population, and civilian firearms per 100 persons. While the R-squared of 0.397 indicates that the data do not represent even half of the variability in the suicide rate data there are some interesting outcomes.

Many trends that are seen in other analyses are also seen here. For example, males add 14.8 to the rate over females and the eldest population (75+) adds 14.8 as well over the population aged 15-24. Both are the largest factors in the model. Also, as population and the civilian firearms ownership rate increases the suicide rate increases modestly. Interestingly, South Africa and Western Asia have the largest negative effect on suicide rates while Eastern Europe and Eastern Asia add the most to the suicide rates as far as subregions are concerned.

Below is the output of the final model:



## Random Forest

The random forest method is an ensemble method that creates many random, uncorrelated decision trees to make continuous or discrete predictions. Since decision trees can have high variance depending on training and testing data, using many of these trees improves the accuracy and reliability of this method. The Random Forest method also only uses a subset of all the features in each decision tree, making the decision trees uncorrelated.

The data were split into train and test sets and were stratified based on the tertile assignment of suicides per 100 thousand people. The train set was used to determine the best model through cross validation. Model performance was measured based on Mean Squared Error. The base MSE of this model would occur when each prediction is simply the mean of the train response, 12.84. The base MSE is therefore estimated at MSE = 365.61.The maximum error when predicting the mean in the test set is 212.13 for Aruba in 1995 where there were 224.97 suicides per 100 thousand people.

Since there were many variables to choose from, all combinations of variables were looped through in to determine the best subset of variables for the random forest model. Additionally, each of these models was tested using a grid search with three-fold cross validation testing all combinations of max depth of 3, 7, 12, 16, and 24 with four different max features methods. These methods were auto, log2, 0.15, and None. Max features = 0.15 represents 15% of the number of features used; auto represents the square root of the number of features; log2 represents log2 of the number of features; and None represents the number of features. This was a very time intensive process since there were 20 models built on 510 combinations of variables. The testing of all models took a total of 6 hours.

Without including sex and age, the most successful combination of variables was year, population, civilianFirearmsPer100Persons, civilianFirearmsCount, InternetUsers, Rank, Percentage. This model had a max depth of 16 and max features of auto. The MSE of this model was 137.5, which is 2.65 times lower than the base MSE. The maximum absolute residual from these predictions is 192.69 for Seychelles in 2006 where there were 204.92 suicides per 100 thousand people but only 12.23 were predicted. In this observation, there were 2 suicides with a population of only 976.

This model was improved tremendously with the addition of features sex and age. The optimal max depth became 24 while the max features became log2. With these 9 variables, the random forest model achieved a MSE of only 48.46. The most inaccurate prediction for this model had an absolute residual of 209.46 where there were 2 suicides with a population of 889.

The records with very low populations appear to result in the greatest residuals. This makes sense because an increase of only one or two suicides has a much greater impact on these records.

# Recommendation

All countries should immediately implement a national strategy for suicide prevention. This strategy should include the spreading of mental health awareness, especially in men over 55 years old. Countries with high internet accessibility and/or a high number of guns per person should perform studies to determine why the suicide rate appears to increase proportionately to these metrics. These factors were shown to be related to suicide rate through visualizations, unsupervised learning methods, and supervised learning methods.

# References

1. National Suicide Prevention Strategies: Progress, Examples and Indicators - World Health Organization - [https://apps.who.int/iris/bitstream/handle/10665/279765/9789241515016-eng.pdf](https://apps.who.int/iris/bitstream/handle/10665/279765/9789241515016-eng.pdf?ua=1)
2. Suicide Key Facts - World Health Organization - <https://www.who.int/en/news-room/fact-sheets/detail/suicide>
3. Wikipedia Gun Ownership - <https://en.wikipedia.org/wiki/Gun_ownership>
4. Plotly World GDP with Codes Dataset - <https://raw.githubusercontent.com/plotly/datasets/master/2014_world_gdp_with_codes.csv>
5. Wikipedia Internet Usage - <https://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users>