Exploratory Analysis of World Suicide Statistics

by Blake List, Nina Sun, Jaibir Batth and Shun Li

20 October 2018

1. Introduction

1.1 Background

Suicide is one of the most tragic and complex topics of discussion in this day and age. It is a terrible cause of death and one that affects not only families and friends, but also the wider community as a whole. The purpose of this analysis is to shed light on certain statistics concerning suicide and potential factors related to these, with the intention of creating open and informed conversation surrounding this very sensitive issue.

According to estimates from the World Health Organisation (WHO), every year, over 800,000 people die from suicide. This translates to a suicide rate of around 11.5 per 100,000 people – a figure equivalent to someone dying of suicide every 40 seconds. In New Zealand alone, it was found that between June 2017 and July 2018, 668 Kiwis died by suicide - the highest number of suicides since records began in New Zealand. Every suicide is a tragedy, yet suicides are preventable with timely, well-supported interventions. While it is not possible to determine the precise motives or the cause of suicide, one theme is present as a recurring risk factor - mental health. This report looks to use data wrangling techniques in R and Julia to compare world suicide statistics between males and females, one of the most explanatory heterogeneous forms of division, while also determining any correlation between suicide rates and world happiness indicators, like social support, and economic measures, like Gross Domestic Product (GDP) and health expenditure per capita.

1.2 Targets of interest

Throughout the scope of this project, our aim was to answer three key questions regarding world suicide statistics:

- How do male and females suicide rates differ by country per 100,000 population?
- How are world suicide statistics correlated with world happiness report indicators such as social support?
- Is there a significant trend that can be recognized from the data we have collected?
- How does a country’s public health expenditure relate to its suicide statistics?

The various datasets collected allowed us to answer these questions with an appropriate level of certainty that the assumptions and conclusions were correct. In addition, we formed several final relational data frames in the form of CSV’s that could be used by others to conduct future research.

1.3 Dataset sources

http://apps.who.int/gho/data/node.main.MHSUICIDE
https://www.gapminder.org/data/
https://s3.amazonaws.com/happiness-report/2018/WHR2018Chapter2OnlineData.xls
https://en.wikipedia.org/wiki/List_of_countries_by_total_health_expenditure_per_capita
https://en.wikipedia.org/wiki/List_of_countries_by_suicide_rate

https://en.wikipedia.org/wiki/World_Happiness_Report#2016_World_Happiness_Report
https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita

2. Method

2.1 Dataset collection and pre-processing

The analysis of suicide statistics was centered primarily around the dataset containing suicide rate estimates per 100,000 population by country for the years 2000-2016, sourced from the Global Health Observatory data repository from the World Health Organization. This featured estimates for females, males and both sexes and was obtained in the form of a CSV. In addition, we obtained a more comprehensive dataset of age-adjusted world suicide statistics per 100,000 population by country dating from 1950 to 2016 from the Gapminder data repository, also in the form of a CSV. This dataset, however, did not contain information pertaining to males and females of each country.

2.2 Relational datasets combination

Since the variables in the dataset that we collected from the World Health Organization are not enough abundant to help generate enough information related with this topic .So, we search other data source from the World Happiness Report to get more different variables.

The data was collected from the World Happiness Report and featured information concerning a country's life ladder - where one would consider themselves on a scale of 0 (worst possible life) to 10 (best possible life), social support - the perception and actuality that one feels they have a social support network, log GDP per capita, and healthy life expectancy at birth, among other variables. This data was obtained in the form of a tidy Excel spreadsheet.

Due to there were gaps between the dataset that collected from World Happiness Report and the dataset that picked from the World Health Organization, we need to combine these two dataset by the shared common variables like countries that both of them have.

However, since the incomprehensiveness of official statistics, some columns exist lacking of information, which were needed to be mutated; In additions, applying the knowledge that we have learnt from the class(changing the table type), sometimes , it is a bit difficult to convert the wide relational dataset into a long table. Even we have done the job, *NA* value will be created.

2.3 Scraping data from webpages

Lastly, due to the difficulty of finding datasets regarding the amount of funding spent on public health by countries per year, it was necessary to scrape the data from a valid source - in this case Wikipedia. Although Wikipedia allows for database dumping through its Pip package, it was important to be able to demonstrate knowledge and ability of scraping data from the web. As this data was to be combined with other sources, the year 2016 was selected.

3. Results

3.1 Achievements

Throughout the project, we believe we were able to adequately answer the three key questions posed at the beginning of the report. Evidence to support this comes from the visualisations of data that follows. As gender is a very important separator of other demographics like age and race, one of our main successes during the project was comparing the suicide statistics between males and females.

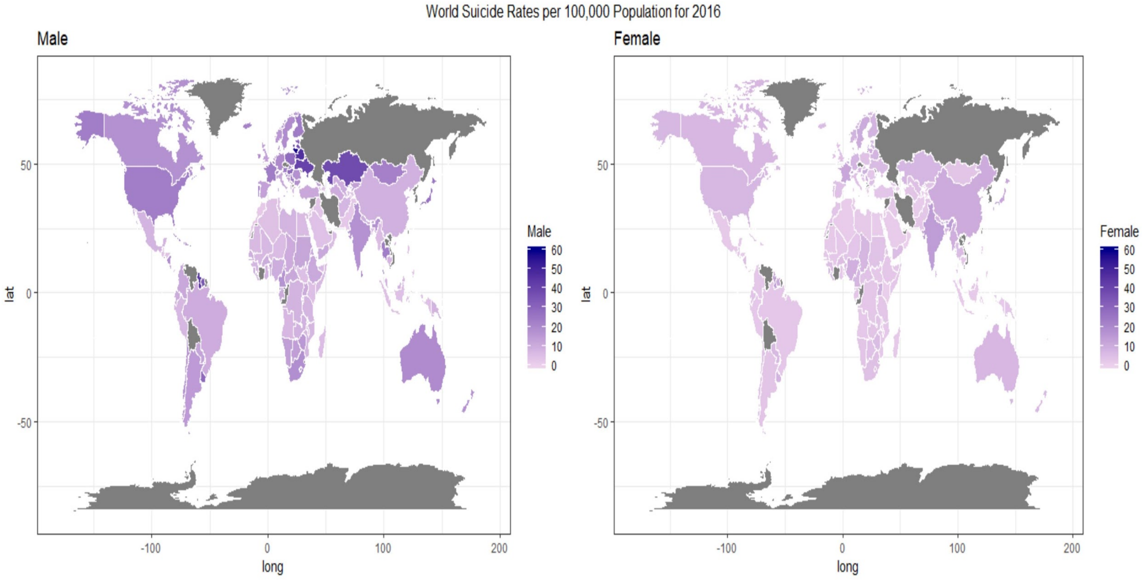

Figure #1: World suicide rates per 100,000 population for males and females for the year

The above figure uses the maps package from R to demonstrate the world suicide rates per 100,000 population by gender. We can see that the graph for males is much darker overall which shows that males, on average, have higher suicide rates per 100,000 population that females. It is also worth mentioning that total suicide rates are higher in higher income countries like the U.S, Australia, New Zealand, France, and China.

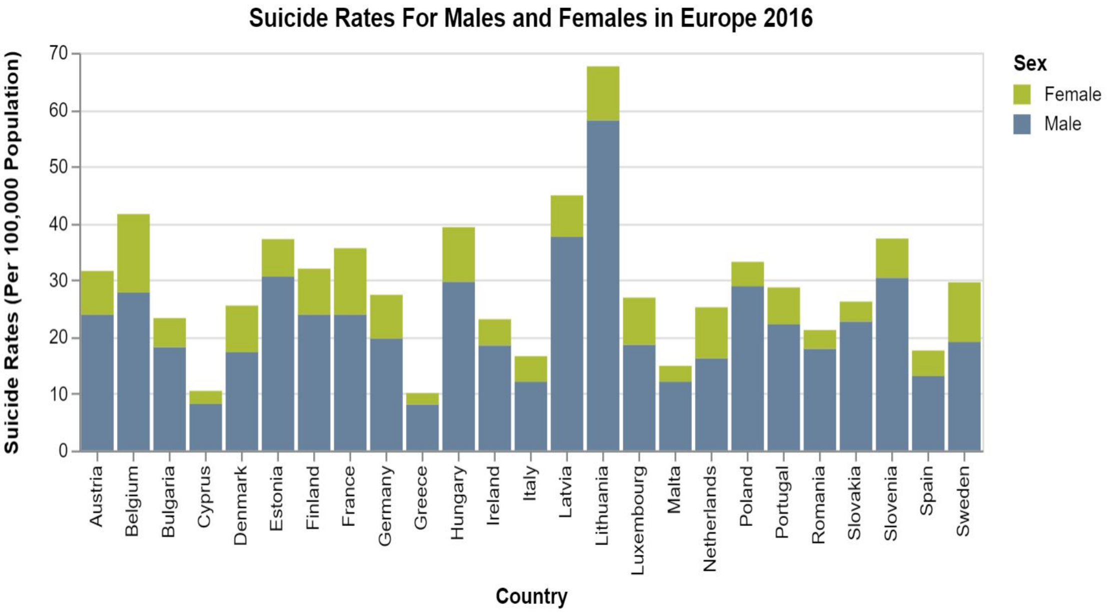

Figure #2: Suicide Rates by gender in European Union Countries for the year 2016.

The above plot was made using the VegaLite package in Julia. It follows on with the same point as *Figure # 1* that males, on average, have higher suicide rates per 100,000 population than females. In this case, the observation is demonstrated using data for the European Union countries. We can see that females contribute very little to the total suicide rate per 100,000 population compared to males. In addition, Lithuania has the highest total suicide rate whereas Cyprus and Greece have the lowest total as well as the lowest female and male suicide rates for the countries in Europe.

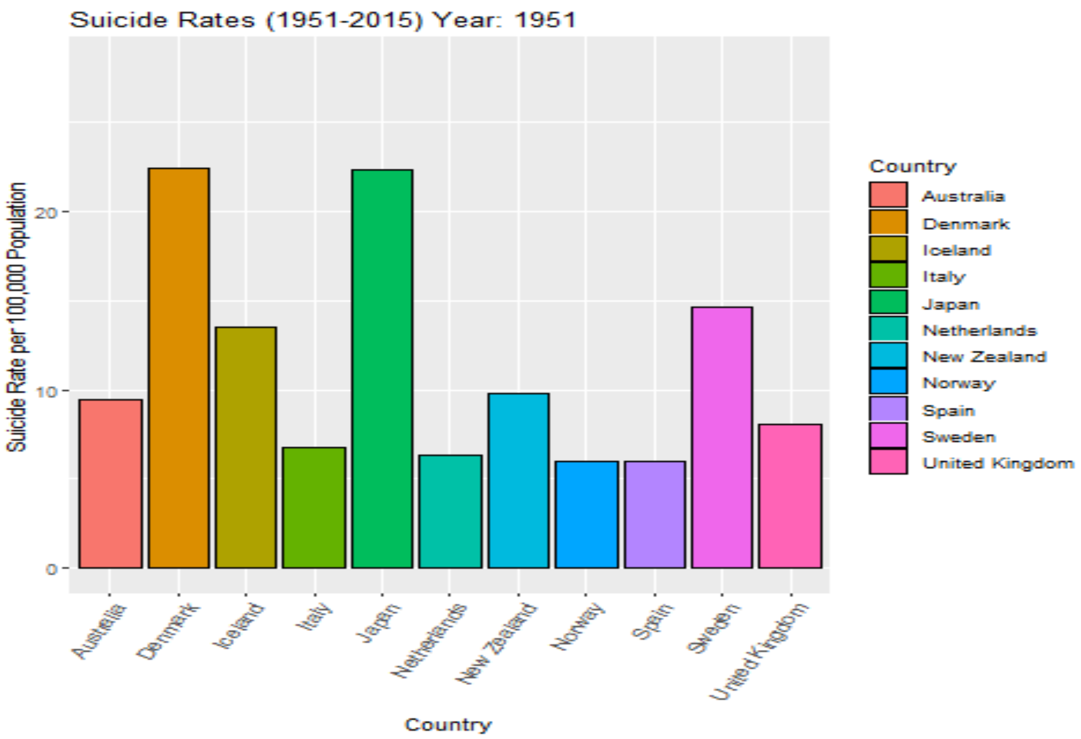

Figure #3: Animated bar plot of suicide rates per 100,000 population between 1951 and 2015

The above plot utilizes the gganimate package to transform between suicide rates per 100,000 population for the given countries between 1951 and 2015. As the animation progresses, we see an overall decrease in the total suicide rates for each country from the late 1980's and early 1990's. From the beginning of the data during the 1950's, Denmark and Japan have some of the highest suicide rates. Italy and Spain have consistently low rates of suicide per 100,000 population for the period of 1951 to 2015.

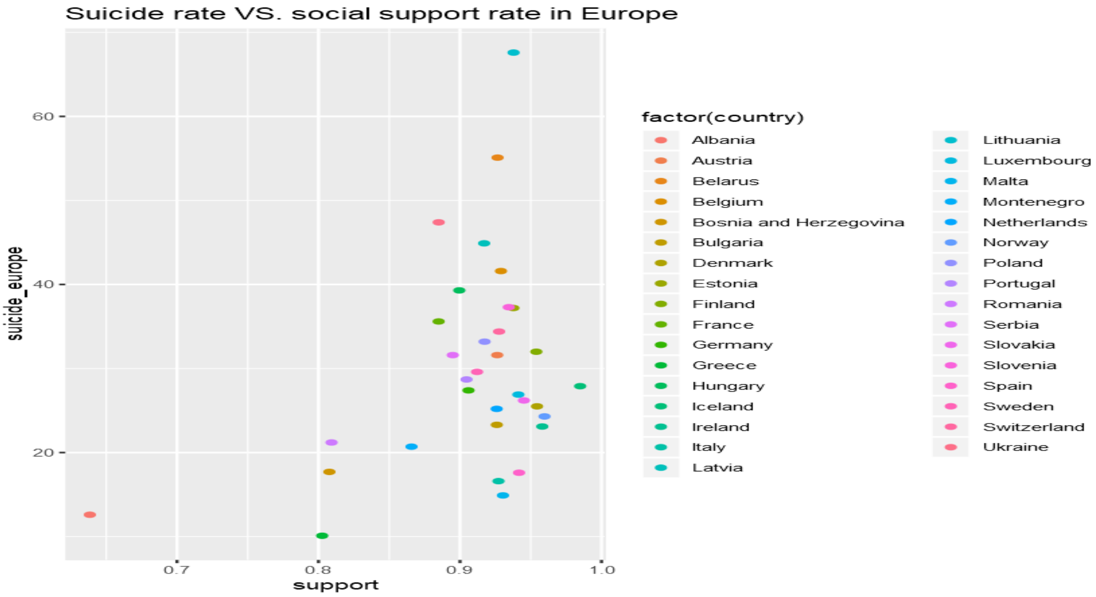

Figure #4: Scatter plot for Social support rate and suicide rate in Europe countries in 20

The above plot shows a correlation between suicide rate and social support rate in Europe countries in 2016. The correlation between suicide rate and social support rate is clearly presented here. Though couple of outliers existed in the data, generally speaking, this plot presented a positive correlation between suicide rate and social support rate. Also there is enough data for Europe countries compared with Oceania country dataset.

3.2 Failures

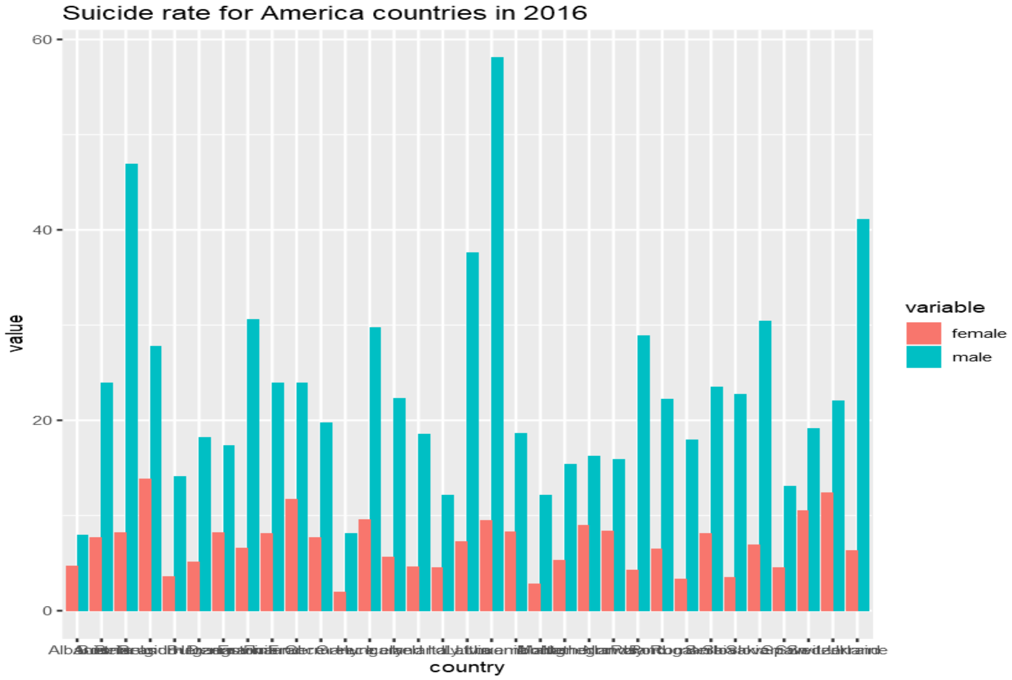

Figure #5: Bar chart for suicide rate for female and male in Americas countries in 2016

The above plot showed a comparison between female and male suicide rate in Americas countries in 2016. Some bar plots were generated but because of the number of rows for the countries, the country name on x-axis is not clearly presented. Since there are only a small number of data in each datatable, dropping the some rows is not a good choice. This could make the country names on x-axis clear but will result in a loss of information.

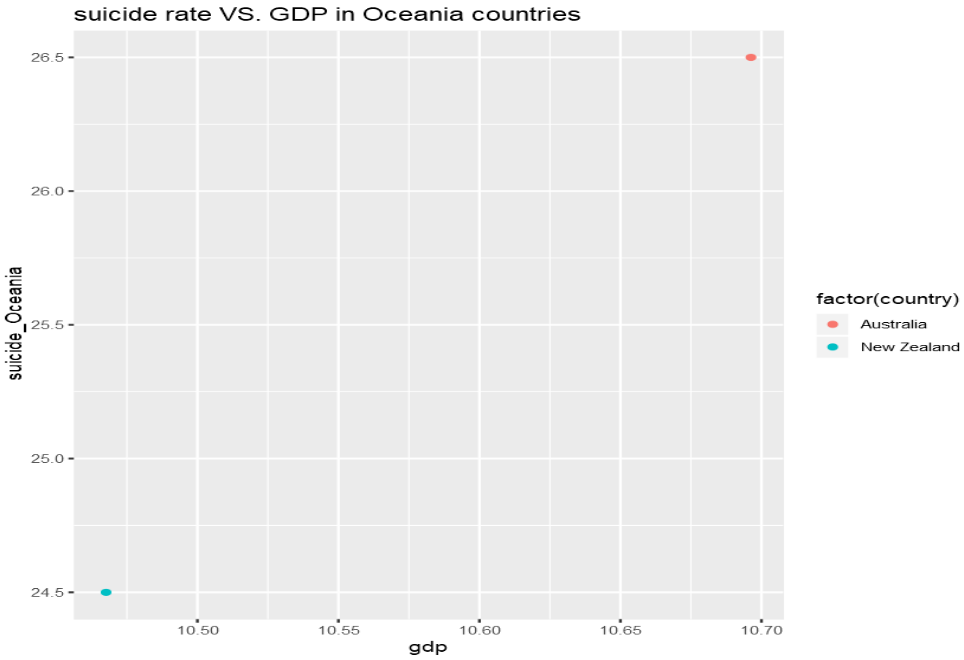

Figure #6: Scatter plot for GDP and suicide rate in Oceania countries in 2016

The above plot shows is supposed to show a correlation between GDP and suicide rate in Oceania countries in 2016. But there are only two countries in the Oceania dataset, which are New Zealand and Australia. Due to a lack of data here, the correlation between GDP and suicide rate in Oceania countries can not be analyzed because there are only two points in the plot.

4. Discussion

4.1 Difficulties

As is the case with most data analysis, the lack of meaningful data can lead to constrained results. Throughout our project, we found many datasets that could have been useful, had they not contained so many missing values. This often resulted in data being thrown away as it was not sufficient enough to achieve a significant outcome. In addition, when merging dataframes based on certain columns or variables, any non-intersecting information was always removed. In order to get a meaningful relational dataframe that could be analysed and visualised, we were often left with minimal data.

The use of new packages, such as gganimate and maps, brought forward some roadblocks as we were treading through new territory. Usually, these packages had few examples of uses and so trialing by error was often the only solution. Furthermore, Julia's package updates and installs became an issue which stalled us for a lot of time. Individual library files needed to be manually deleted then rebuilt to fix the issue.

4.2 Future work

In future, we would like to collect more data from various sources that could be combined to eradicate any missing values from the dataframes. This would mean that more meaningful analysis could be performed and data could be visualised more clearly leading to higher certainty about any conclusions that were made. It would also be of interest to compare cross-country figures and potentially the suicide estimates of Eastern and Western continents. This could provide some insight into the factors and characteristics indicative of suicidal behaviour between different societies, demographics and ages.

Furthermore, we would like to explore the effect of burdening diseases and disabilities on suicide rate. It cannot be assumed that mental health and depression are the primary factors that lead to an individual's choice to commit suicide, so collecting more information about the quality of one's life and their health could potentially produce some indication as to what is causing such tragic decisions. Lastly, the ability to work closely with governments and councils regarding suicide in their communities may allow for more open debate into ways in which these statistics can be reduced, particularly in New Zealand.

5. Conclusion

From the bar charts, it appears that males tend to have a higher suicide rate than females in almost all the countries.

Regarding the correlation between suicide rate and Log GDP, the scatter plots for different continents did not present a clear correlation between the two. All the points in the plot seems randomly distributed.

For the correlation between suicide rate and social support rate, the scatter plots for different continents all generally follow a similar pattern with a positive correlation but the correlation is still not strong. Therefore it is not clear whether a higher social support rate is one of the factors to determine suicide rate.

As for the correlation between suicide rate and healthy life expectancy rate, there is also not a clear correlation between the two. Similar as the scatter plot for suicide rate and Log GDP, the points here are also randomly distributed and did not follow any linearity.

References

Lee, L., Roser, M., & Ortiz-Ospina, E. (2018). "Suicide". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/suicide' [Online Resource]

Helliwell, J., Layard, R., & Sachs, J. (2018). World Happiness Report 2018, New York: Sustainable Development Solutions Network.