Exploratory Analysis of World Suicide Statistics by Blake List, Nina Sun, Jaibir Batth and Shun Li

Introduction

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 statistics, 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 causes 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.

Dataset summary

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. Furthermore, 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 an Excel spreadsheet. 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.

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

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.

Method

Dataset Collection and Preprocessing Techniques

We first began the wrangling component of this project by loading the suicide rate and world happiness report datasets into RStudio. The data was imported as either a CSV or XLSX file type so readr was used. The condition of the data was ordered but not tidy, however, reaching a tidy state was achievable. Columns were labeled as exceedingly long strings with spaces which needed to be changed for the purposes of filtering and visualising data, especially in Julia, which struggles to deal with column names with spaces. The tidyverse package in R allowed for mass import of all libraries that would be needed for filtering, mutating, gathering, spreading and piping data.

After the column names were changed to a more easily referenced format, summary statistics about each dataframe were output. The dataframes were then filtered to include any interesting information and to exclude variables which contained many missing values or would not be of importance. The datasets were then converted to wide and long formats using the gather and spread functions in R, and the melt and unstack functions in Julia. Once the dataframes were in an appropriate format, simple plots (bar and point) were made using ggplot to determine any visible relationship between variables.

Once individual dataframes, such as the suicide rates per 100,000 population and world happiness report statistics, were in a format suitable for visualisation, they were then inner joined on intersecting countries and years. For visualisation in R, ggplot, ggmaps, worldmap and gganimate were used. The plotting in Julia was performed using the VegaLite package.

Scraping Data

The foremost work to be done for scraping the data from a web page is to find an appropriate source which has the relevant data and allows us to scrape data from their website. This initial stage task was the most challenging, wrangling the untidy data being at the second spot as far as the challenges are concerned. We started off by finding the web pages that had some relevant data related to the suicidal rates for different countries. We decided to analyse the trends for the year 2016, so the data for 2016 was the most relevant, in our case.

The second step was to read the “Terms and Conditions” of the website that we were planning to scrape, to ensure they allowed their data to be scraped. We got the data related to suicidal rates and other attributes like Health Expenditure, Happiness Score and GDP of the countries on Wikipedia and some other websites that had no issues with their data being scraped.

The “rvest” package of R makes it convenient to scrape the data from a web page either by using the CSS selectors or the XPath of components of a web page. We were able to fetch this data in the form of HTML tables. We then converted these tables into dataframes, as we need to perform further operations on this data. The irrelevant columns were omitted from these dataframes to keep it simple and concise, and the names of the columns were changed to make them more readable. Another challenge we faced here were the figures mixed with special symbols or some Unicode characters in some of the dataframes. gsub() proved to be helpful in this case.

Finally, we were ready with the relevant data to be visualised.We used the functions of the ggplot package which proved to be handy to visualise the data through some bar charts and scatter plots. We were able to produce some inferences and conclusions based on the scraping that we did, which are included in the conclusion section in this report.

Results

Achievements

Throughout the project, we believe we were able to adequately answer the three key questions posed at the beginning of the report, or give reason for a lack of support. 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.

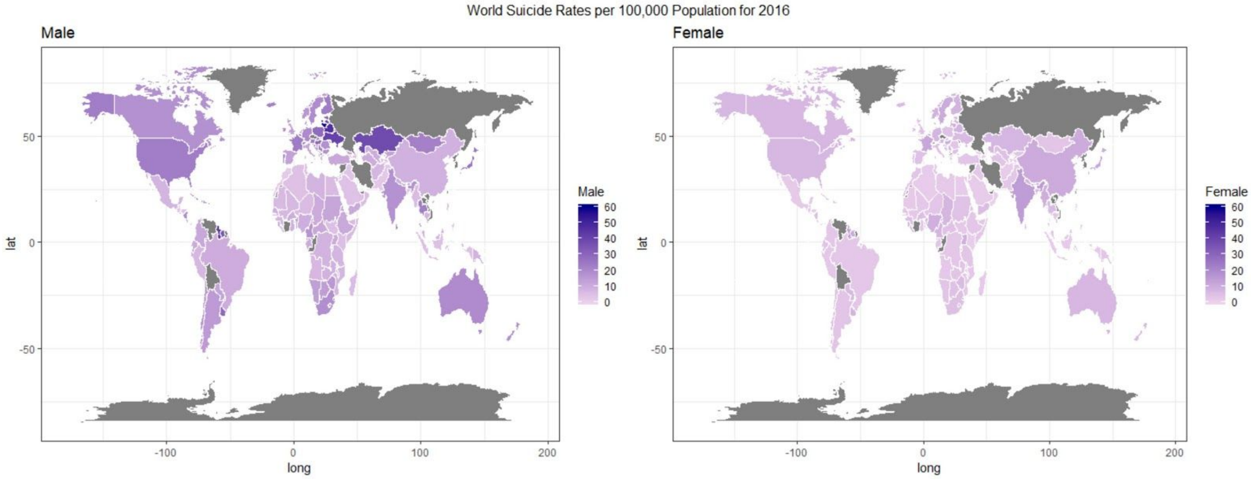

Figure 1: World suicide rates per 100,000 population for males and females for the year 2016

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.

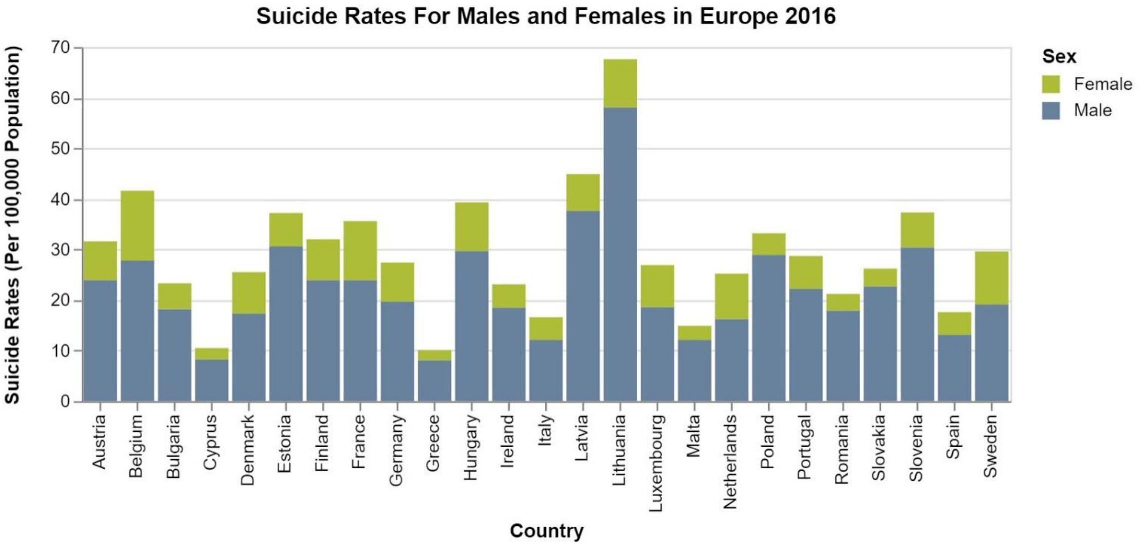

Figure #2: Suicide Rates by gender in European Union Countries for the year 2016.

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.

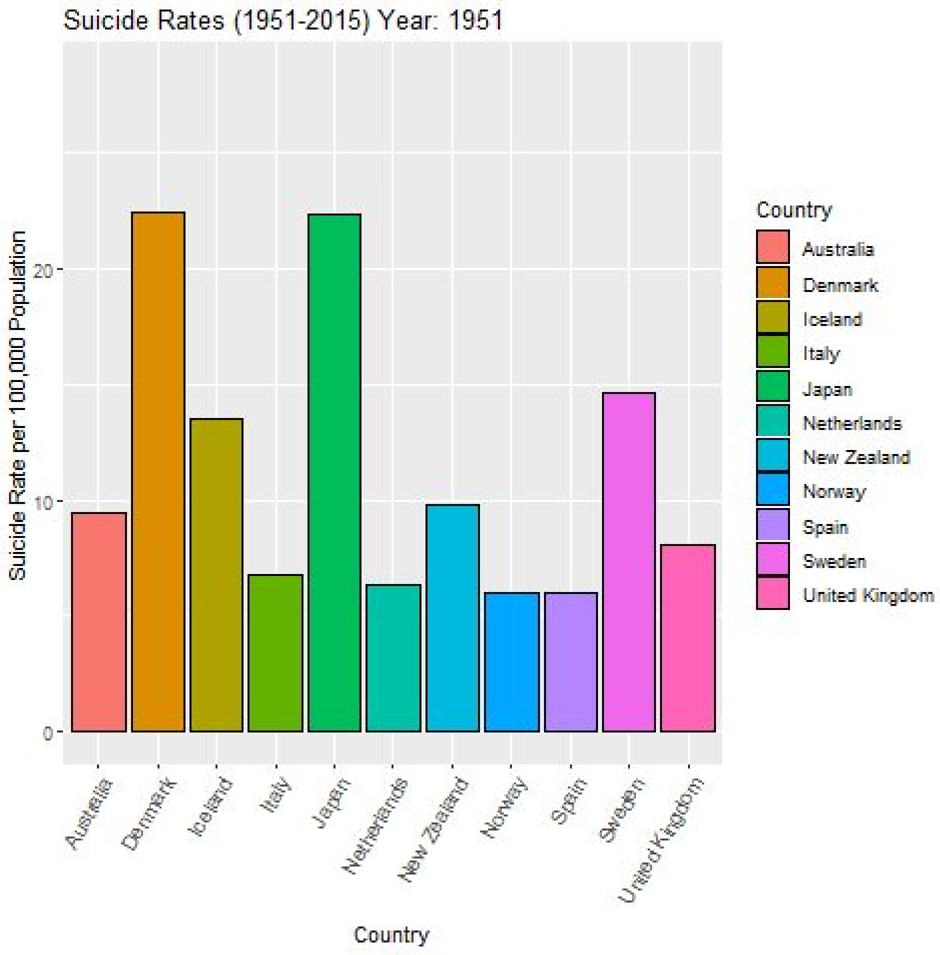

Figure #3: Animated bar plot of suicide rates per 100,000 population between 1951 and 2015

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.

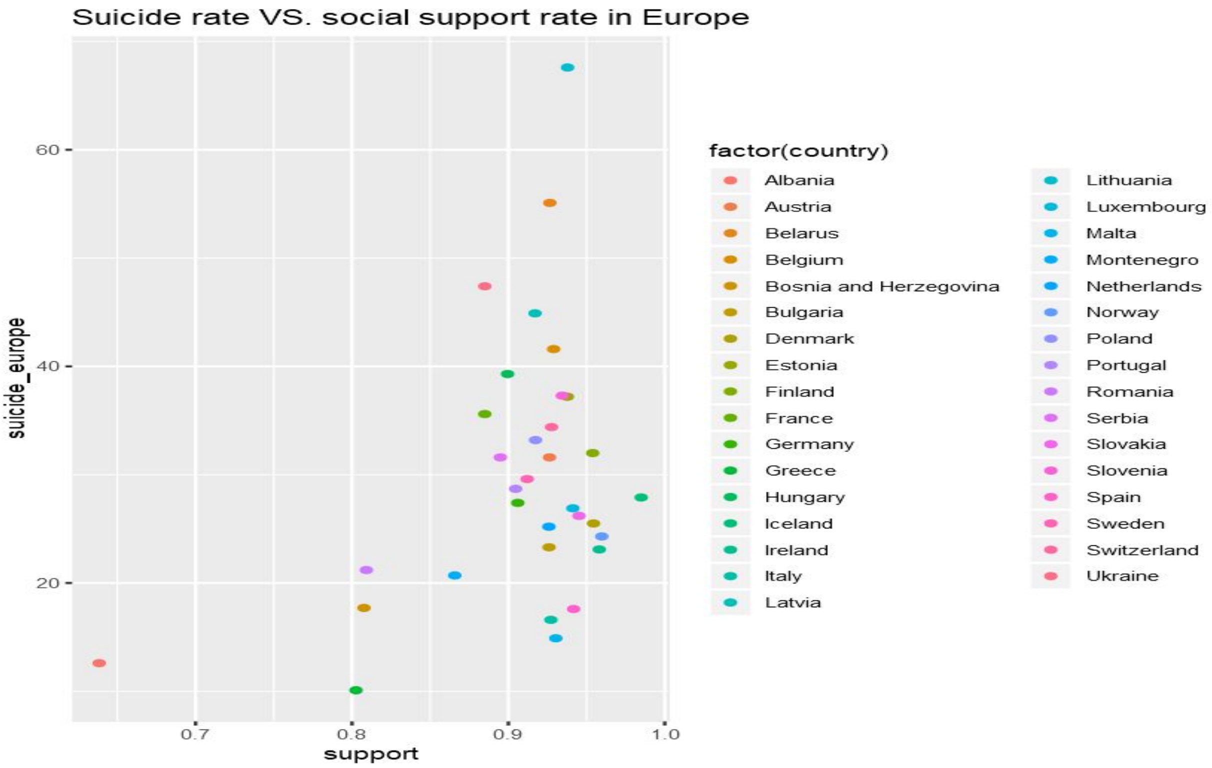

Figure #4: Scatter plot for Social support rate and suicide rate in Europe countries in 2016.

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 a couple of outliers exist in the data, generally speaking, this plot presents a positive correlation between suicide rate and social support rate. Also, there is enough data for European countries compared with the Oceania country dataset.

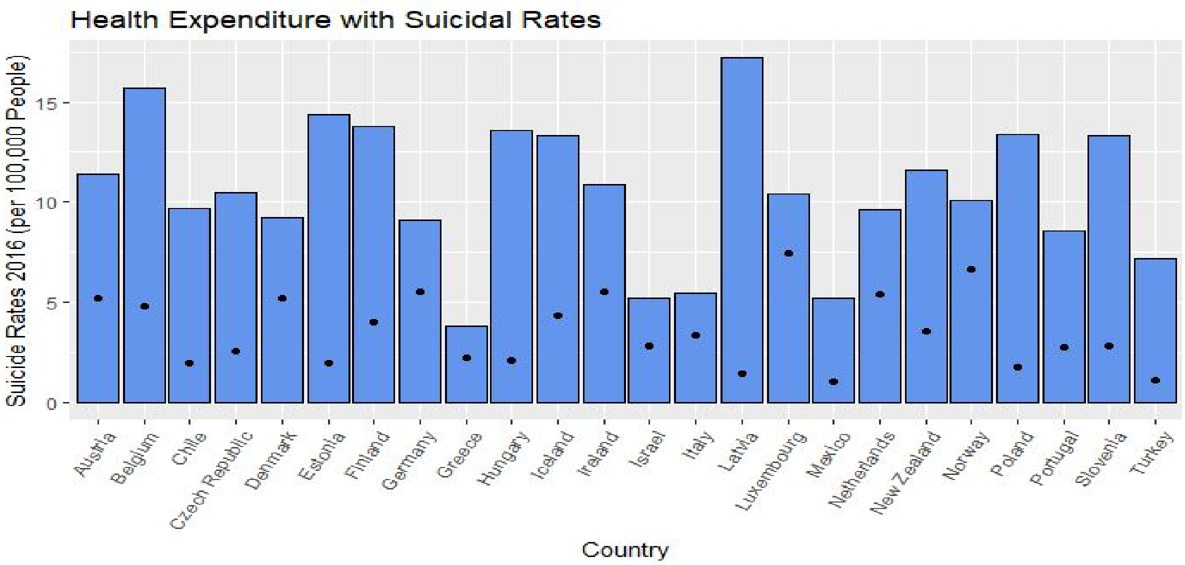

Figure #5: Health expenditure verse suicide rates by country for 2016.

The above graph shows the Suicide rates (per 100,000 people) for different countries along with the health expenditure shown as points on the bar chart. The graph comes out to be as expected for some countries, but for some of them it is bit different from the expectations. Like for Latvia, it can be seen that the expenditure on Healthcare is too low and therefore the suicidal rates are high. But for some other countries, like Luxembourg, even if the health expenditure is high, the suicidal rates are still high.

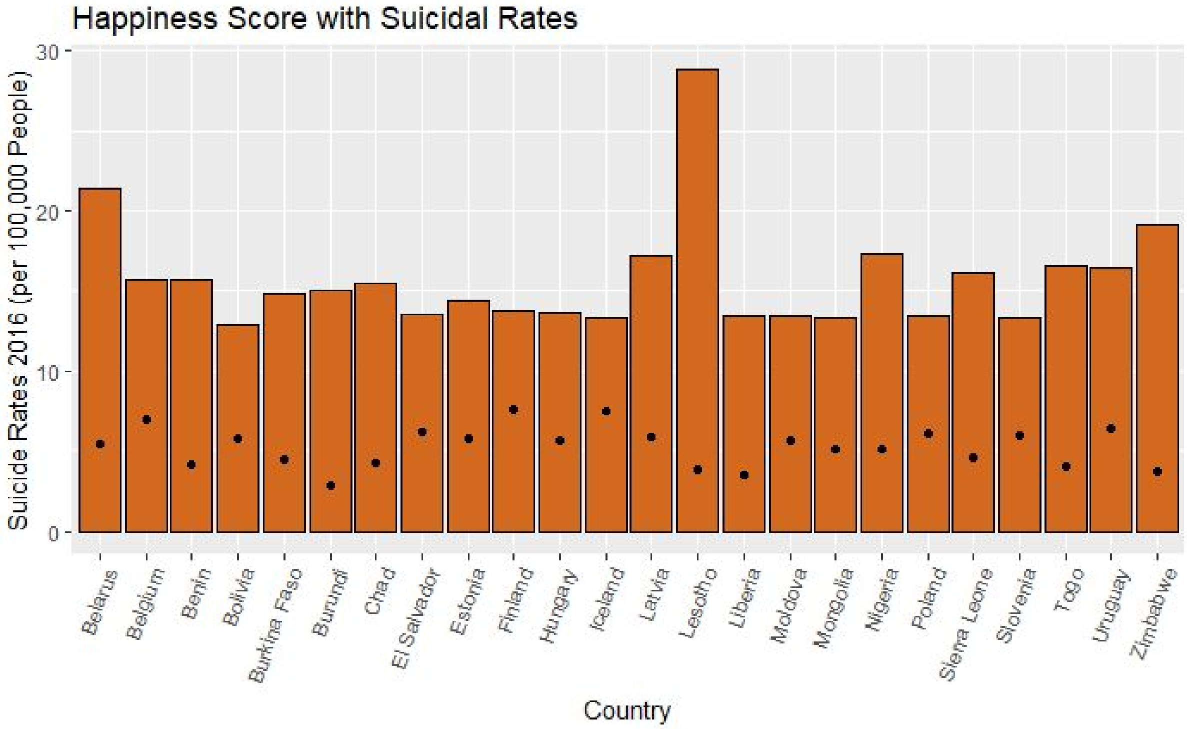

Figure #6: Happiness score verse suicide rates by country for 2016.

The above graph shows the suicide rates (per 100,000 people) for different countries along with the Happiness Index shown as points on the bar chart. The results here are closer to what was expected than it was in the plot of suicidal rate with health expenditure. Like for Burundi, if the Happiness Index is low, the suicide rate is high and that was what was expected. However, for a country like Finland, even if the happiness Score is high, the suicidal rates still comes out to be high. Thus, there might be some other issues other than the happiness score that need to be considered to reach some conclusion about the number of suicides.

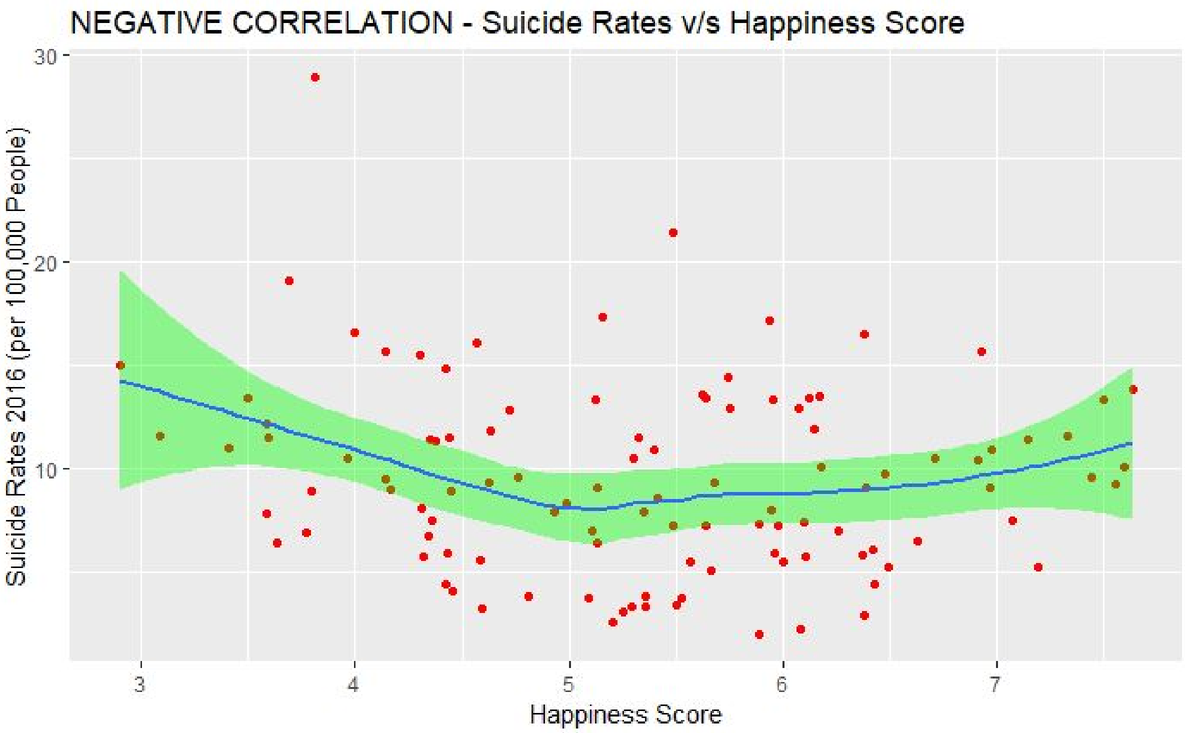

Figure #7: Correlation between suicide rates and happiness score for 2016.

We cannot consider the correlation of the suicidal rate and happiness score as significant. However, it comes out to be negative which was expected as suicidal rate is inversely proportional to the happiness score, and the inclination of the model line should be negative.

Failures

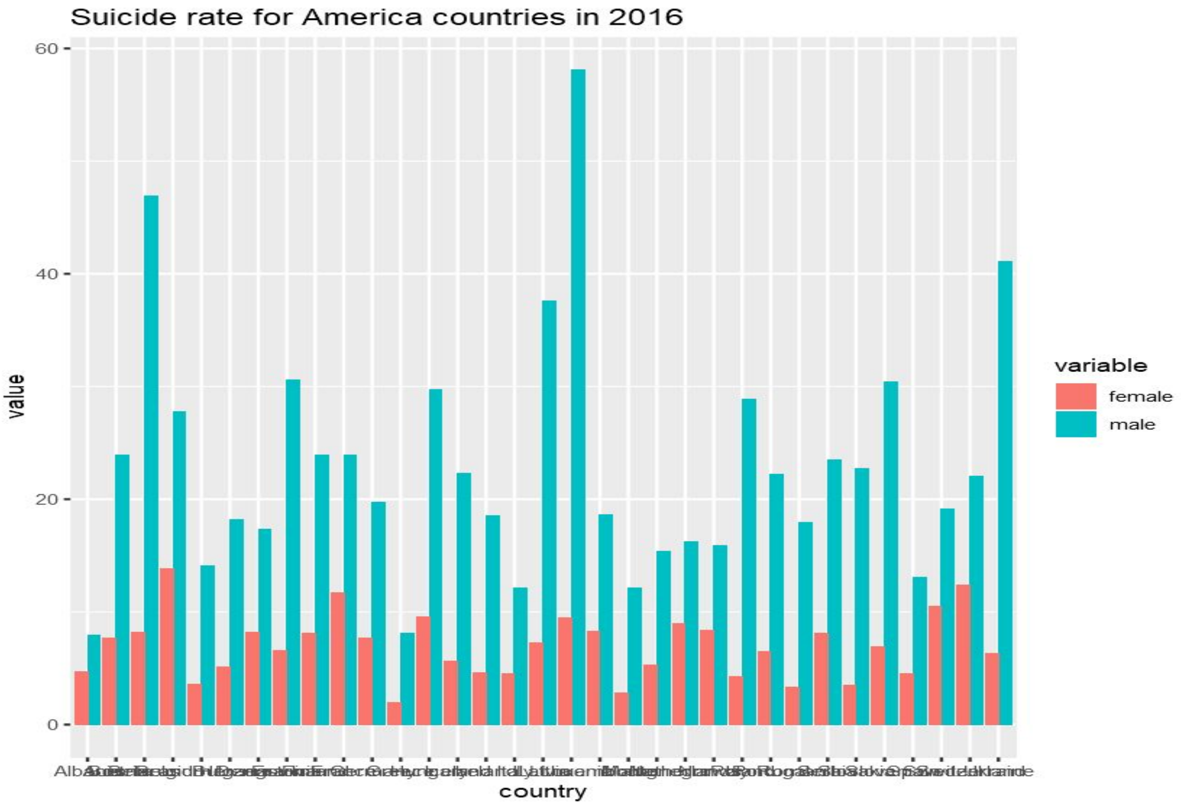

Figure #5: Bar chart for suicide rate for female and male in Americas countries in 2016.

The above plot shows 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 the x-axis is not clearly presented. Since there are only a small number of data in each dataframe, dropping 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. We have also realized the bar chart is not a good way to show linearity between variables so decided to use a scatter plot.

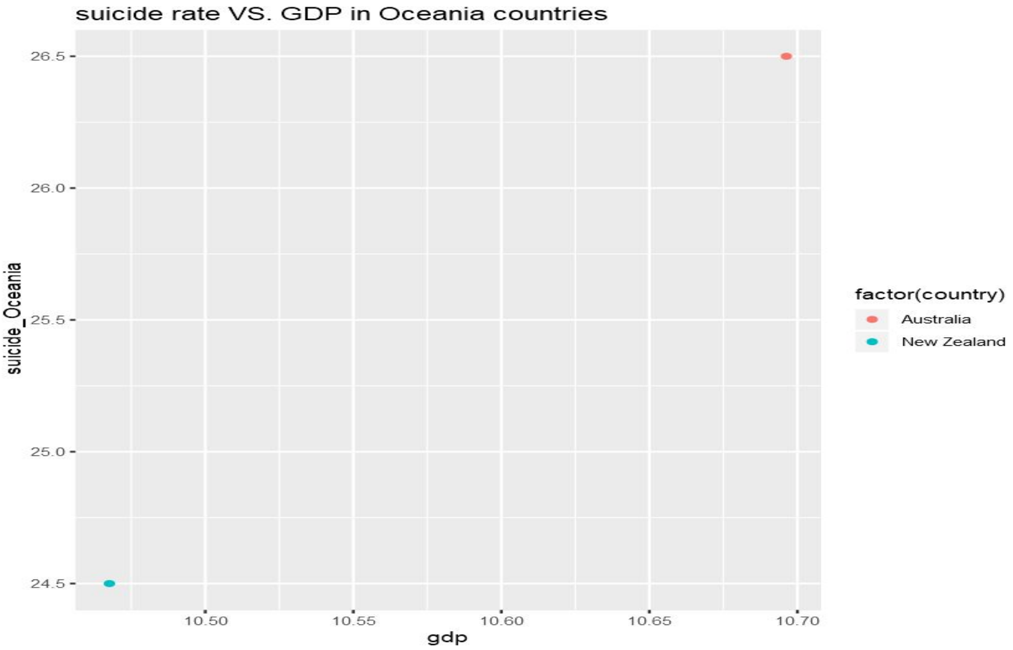

Figure #5: Scatter plot for GDP and suicide rate in Oceania countries in 2016.

The above plot 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, New Zealand and Australia. Due to a lack of data here, the correlation between GDP and suicide rate in Oceania countries can not be analysed because there are only two points in the plot.

Discussion

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.

Future Work

In future, we would like to collect more data from various sources that could be combined to 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.

Conclusion

This report demonstrates that, on average, males tended to have a much higher suicide rate than females in almost all 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 seemed randomly distributed. For the correlation between suicide rate and social support rate, the scatter plots for different continents all generally followed a similar pattern with a weak positive correlation. Therefore it is not clear whether a higher social support rate is one of the factors contributing to suicide rate. As for the correlation between suicide rate and healthy life expectancy rate, there is not a clear correlation between the two. Like the scatter plot for suicide rate and Log GDP, the points here are also randomly distributed and did not follow any linearity.

More data would need to be added to confirm any relationship between the world happiness report indicators and the suicide rates per 100,000 population for males and females. Without conclusive evidence, we cannot claim that the overall ‘happiness’ of a country implies a change in suicide statistics. However, more public health and mental health expenditure by governments could lead to a drop in suicide, especially in higher socioeconomic countries. In summary, more awareness of mental health and depression and open discussion into the causes of suicide is needed to make a significant impact to decrease the tragic occurrence of these deaths, both at the national level and within communities.

References

GHO | By category | Suicide rate estimates, crude - Estimates by country. (2018). Retrieved from http://apps.who.int/gho/data/node.main.MHSUICIDE

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

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

List of countries by GDP (nominal) per capita. (2018). Retrieved from https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita

List of countries by suicide rate. (2018). Retrieved from https://en.wikipedia.org/wiki/List_of_countries_by_suicide_rate

List of countries by total health expenditure per capita. (2018). Retrieved from https://en.wikipedia.org/wiki/List_of_countries_by_total_health_expenditure_per_capita
World Happiness Report. (2018). Retrieved from https://en.wikipedia.org/wiki/World_Happiness_Report#2016_World_Happiness_Report