

Data Visualization Final Project

Introduction:

The data I choose is 'Suicide Rates Overview 1985 to 2016' from kaggle. This compiled dataset pulled from four other datasets linked by time and place and was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum.

The reason I chose this particular data is that I want to know how number of suicides is distributed throughout the world and generations. Does the distribution match the assumption in my mind? For example, is the place with higher GDP with lower number of suicides? Or is there any difference in number of suicides between generation? Therefore, I hope to use the data to answers these questions through data visualization.

Summary of Data:

There're twelve columns in this dataset, including country, year, sex, age, count of suicides, population, suicide rate (suicides/100k pop), HDI (Human Development Index) for year, gdp_for_year (\$), gdp_per_capita (\$), gdp_for_year (\$) and generation. Each data entry in the dataset refers to a single suicides record.

First, we check the distribution of suicide rate in 1985 (Figure 1). It's a right-skewed distribution for the suicide rate. The most frequent suicide rate in 1985 happens to be with

0~5/100k pop, so does 2016 (Figure 2). We can also find that in 1985, there're more extreme suicide rate between 100~150/100k pop.

Distribution of Suicide Rate in 1985

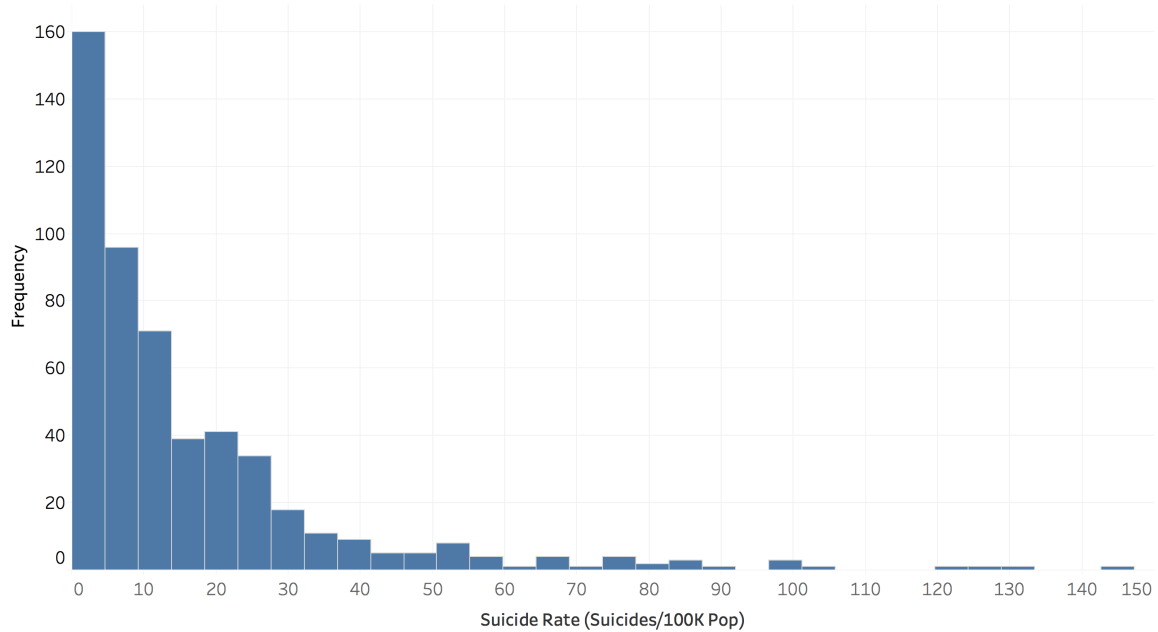


Figure 1: Distribution of Suicide Rate in 1985 (interactive plot)

Distribution of Suicide Rate in 2016

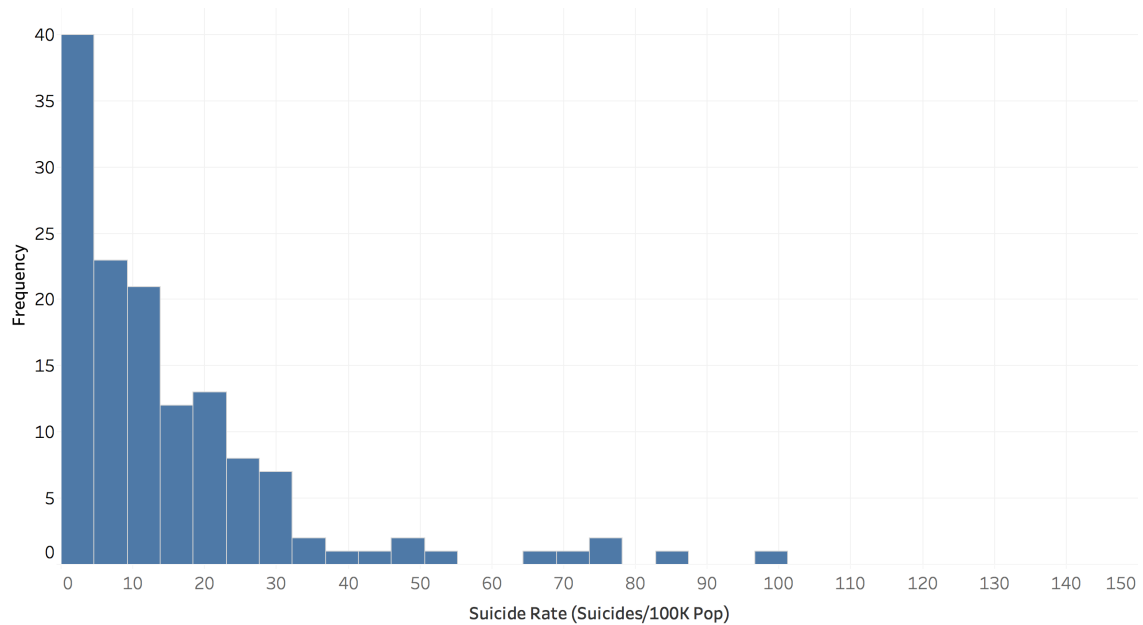


Figure 2: Distribution of Suicide Rate in 2016 (interactive plot)

Now, we look at the information of countries and suicides. I sum up the number of suicides trough 1985 to 2016 and show the top 10 countries with highest total number of suicides (Figure 3). We can tell from Figure 3 that Russia, US and Japan have highest number of suicides and the numbers are much higher than the following countries.

Top 10 Countries with Highest Number of Suicides

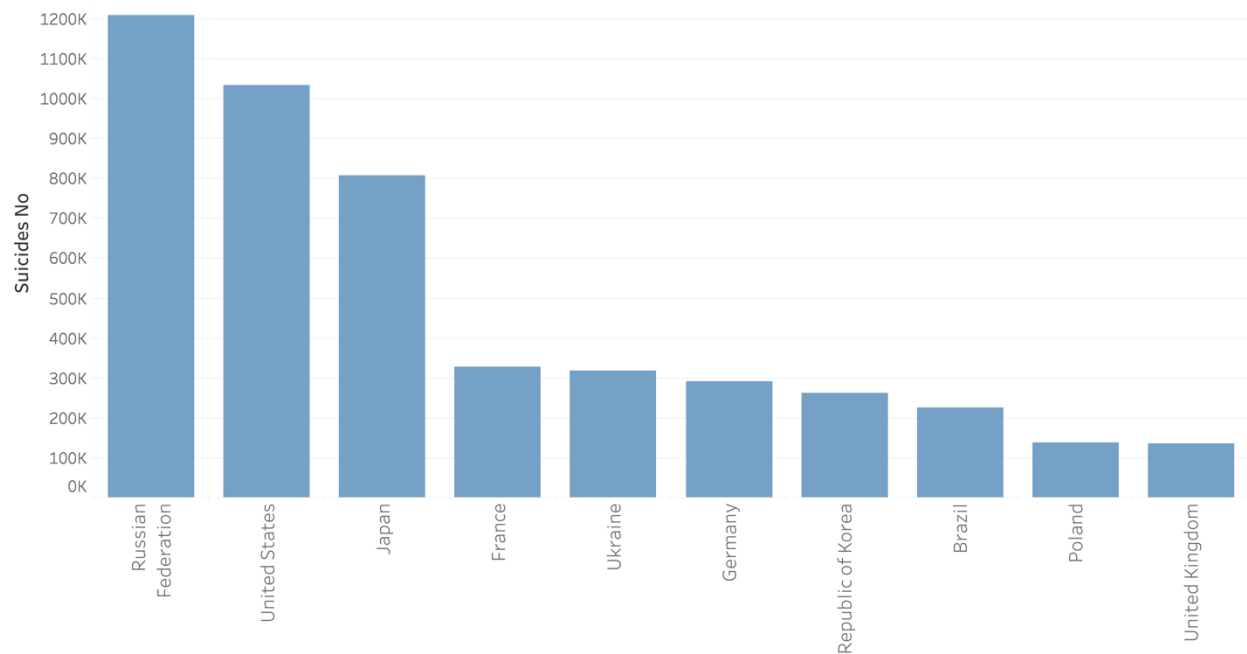


Figure 3: Top 10 countries with highest number of suicides (interactive plot)

For different gender, the distribution of number of suicides are different. We can see from Figure 3, 50% of number of suicides of male is in the range of ~0 to 250. But 50% of number of suicides of female is in the range of ~0 to 75.

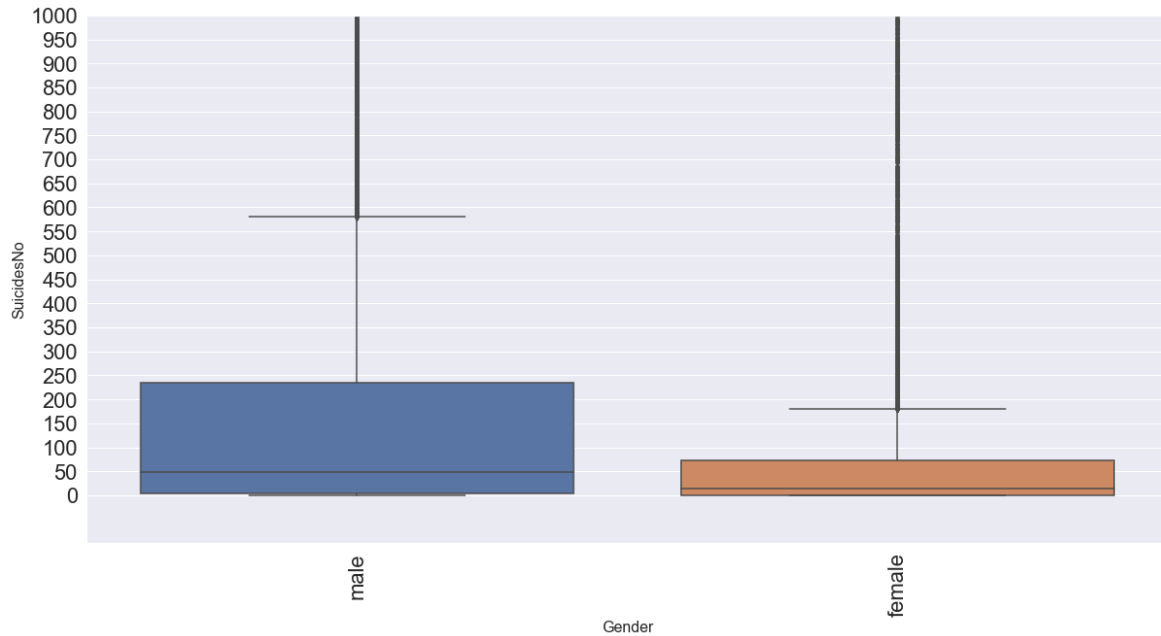


Figure 4: Distribution of Suicide Rate in 2016 (interactive plot)

Figure 5 shows the number of suicides of each year for each generation. Here are the birth years for each generation:

- Gen Z, iGen, or Centennials: Born 1996 – TBD
- Millennials or Gen Y: Born 1977 – 1995
- Generation X: Born 1965 – 1976
- Baby Boomers: Born 1946 – 1964
- Traditionalists or Silent Generation: Born 1945 and before

We can see there are some usual high number of suicides of generation Boomers during 1991 to 2007.

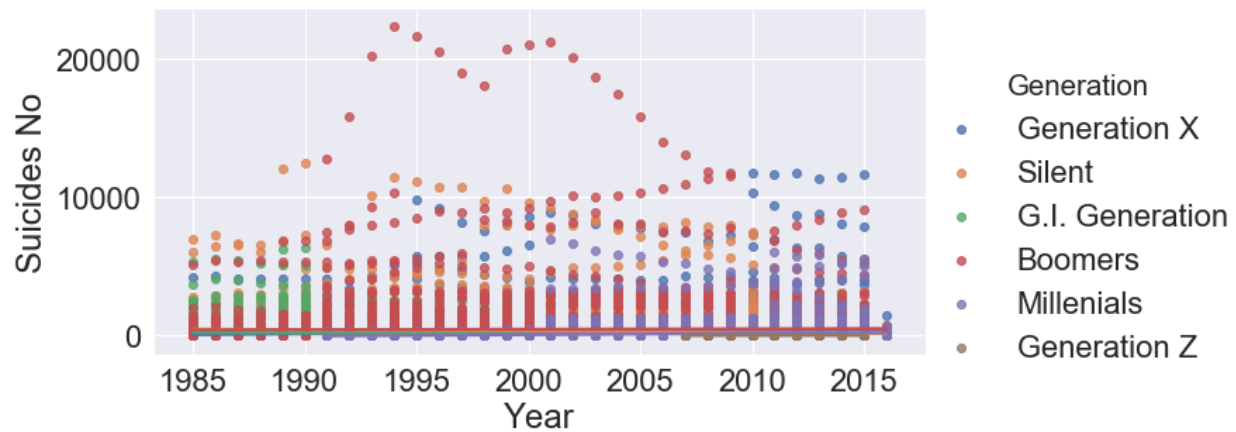


Figure 5: Distribution of Suicide Rate in 2016 (interactive plot)

The bubble plot below shows 2 information, 1) suicides/population of each country and 2) the maximum GDP of those countries. The more reddish the color is, the lower GDP that country has. The larger the size is, the higher suicide rate the country has. For example, if we check United States and Canada, both countries have similar number of suicides, but the GDP/Capita of Canada is lower that United States. There're also some countries with low GDP and low number of suicides, which is impressive.

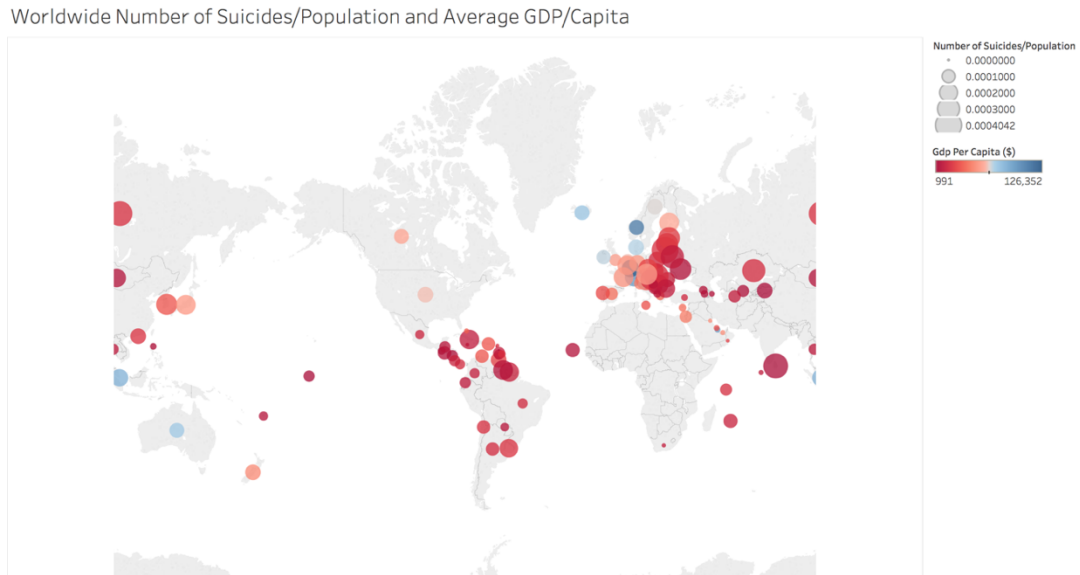


Figure 6: Worldwide Suicide Rate and Average GDP/Capita (interactive plot)

In Figure 7, we can see the Human Development Index (HDI) of those countries, which have records of suicides in the dataset. The HDI is a statistic composite index of life expectancy, education, and per capita income indicators, which are used to rank countries. It's pretty obvious that Russia has the lowest HDI. On the other hand, northern America seems to have higher HDI.

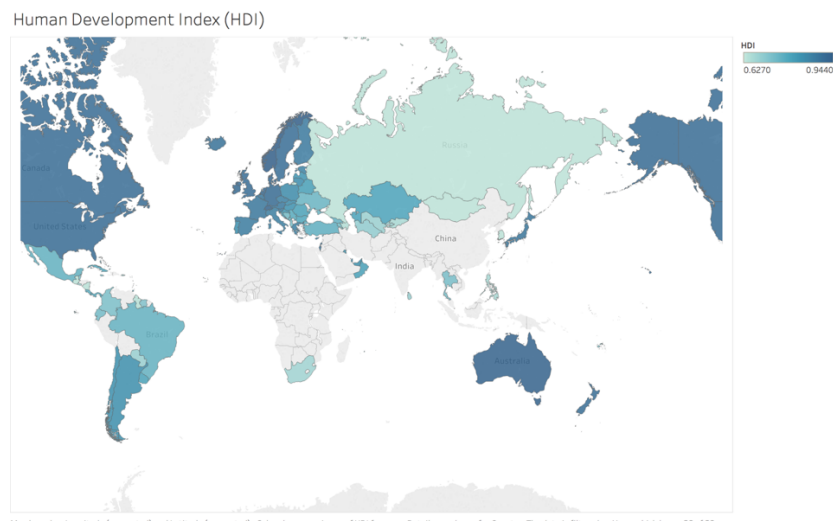


Figure 7: Worldwide Human Development Index (HDI) (interactive plot)

It's hard to create a connection plot basing on this dataset. Per talked to Aleks, she wants us to show our abilities to create a connection plot. Thus, the following connection is just connecting the top 5 countries with highest number of suicides. They are actually across four main continents.



Figure 8: Connection plot of top five countries with highest number of suicides (interactive plot)

Figure 9 is a heatmap showing correlation between 5 columns in the dataset. The darker the color is, the higher correlation it is. Take `GdpPerCapitaMoney` as an example, it has highest correlation with `HDI`. This result makes sense since `HDI` is a statistic composite index of life expectancy, with higher `GDP`, the `HDI` is reasonably higher.

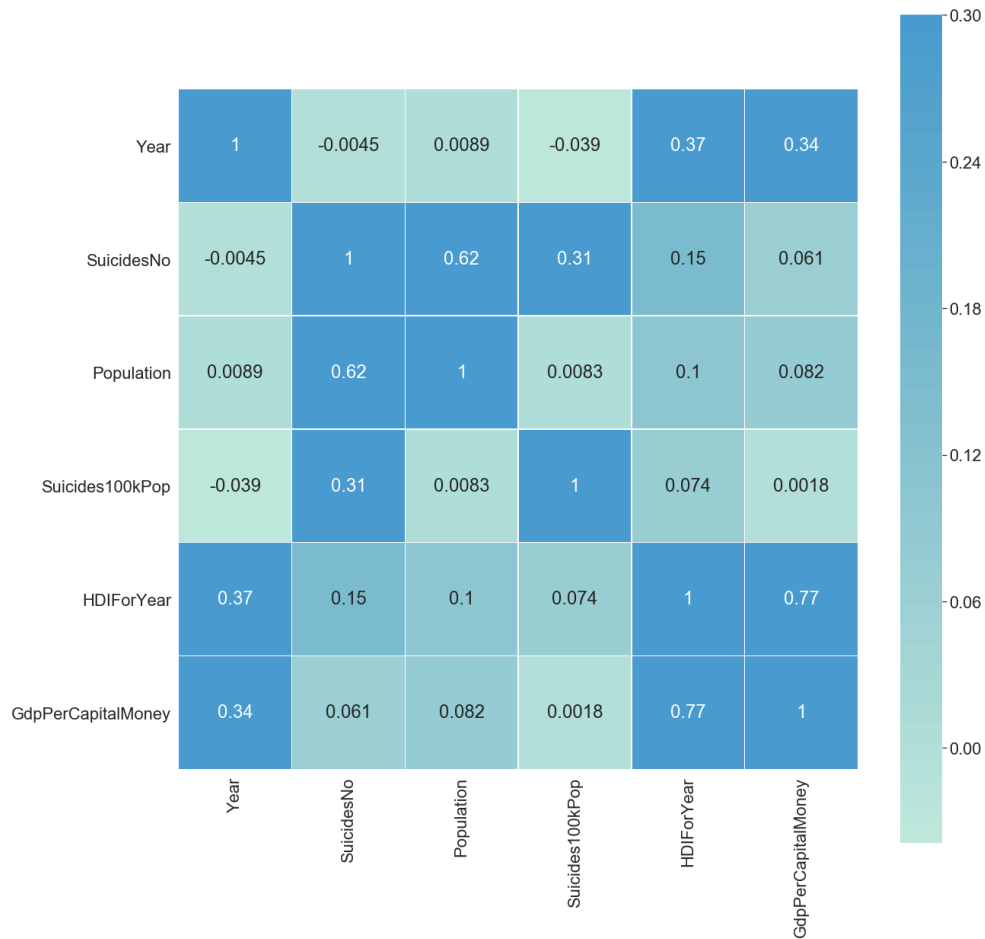


Figure 9: Connection plot of top five countries with highest number of suicides (interactive plot)

Figure 10 shows the stacked area plot of top ten countries with highest number of suicides. The width of each county reveals the number of suicides of that year, the wider the width is, the more suicides happened in that year. Some data are missing in the dataset so that's why there're some discontinuity in the graph. We can tell from the graph that most of the countries keep stable numbers of suicides each year except Russia and South Korea.

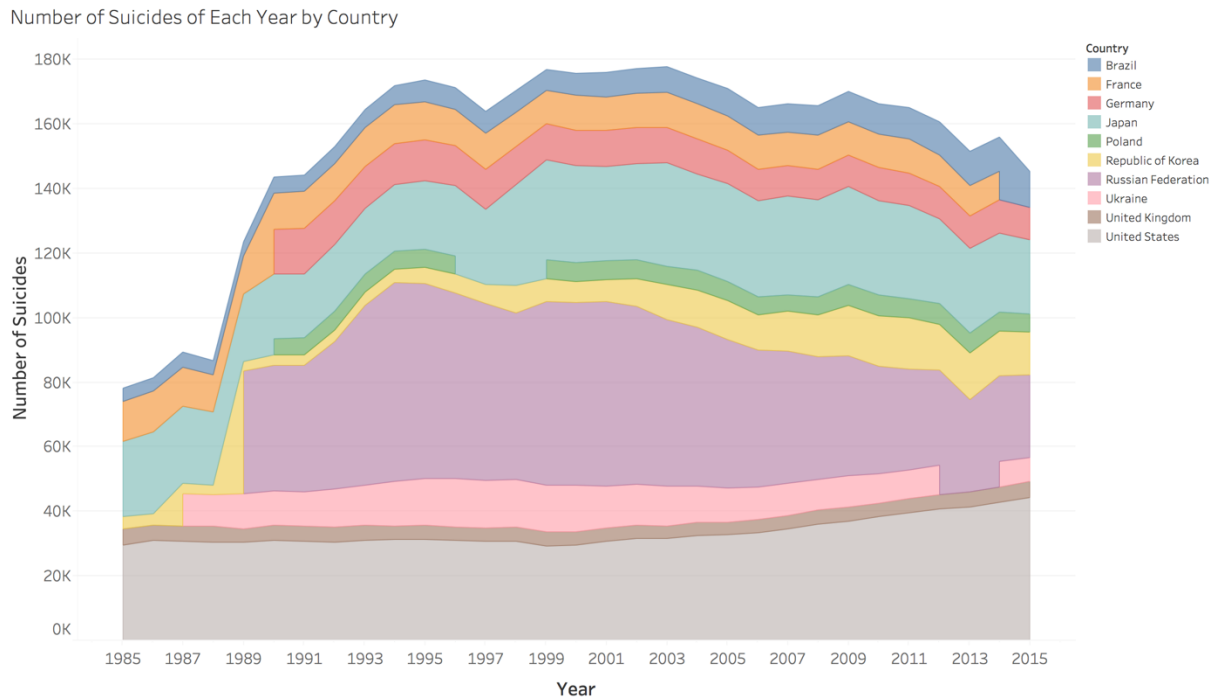


Figure 10: Stacked area plot of top ten countries with highest number of suicides

(interactive plot)

The treemapping below shows the number of suicides of each year, from 1985-2016. Both the area and the color of the box give information about it. The more reddish the color is, the more suicides happened in that year. In terms of the area, the larger the area is, the more suicides happened in that year. Based on this plot, it easy to notice that in the year of 1999, we have the most number of people suicided. There were least number of suicided in 2016, which is a bluest box on bottom right corner. Although the detail is not shown on the box here due to limitation of the area of the box, the detail is shown when you hover over that box. There were 15603 people suicided in 2016.

Number of Suicides each Year

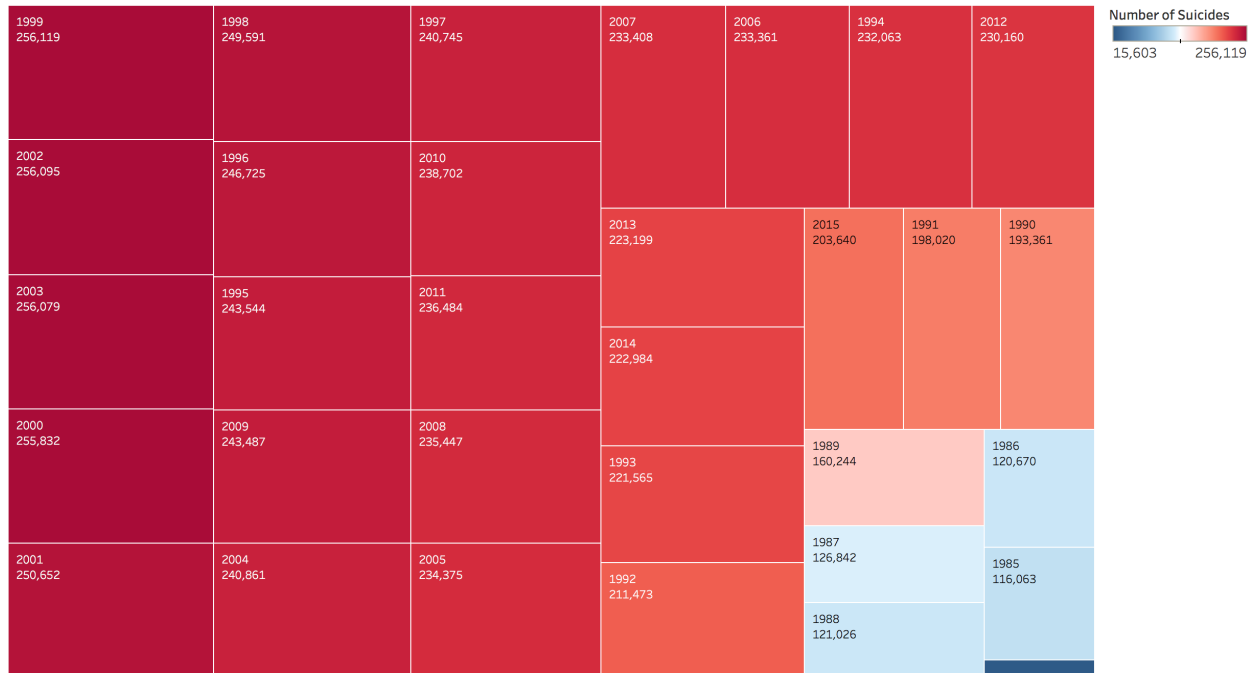


Figure 11: Stacked area plot of top ten countries with highest number of suicides
(interactive plot)

My storyline:

Before working on this suicide dataset, I was imagining that GDP would be a strong indicator of the number of suicides, since GDP is the most commonly used measure of economic activity and serves as a good indicator to track the economic health of a country.

In figure 12, it shows 2 information, 1) the suicide rate (number of suicides/population) of each country and 2) the maximum GDP of those countries. The more reddish the color is, the lower GDP that country has. The larger the size is, the higher suicide rate the country has. Based on the graph, we can conclude that most of the bubble with more reddish color are generally larger. However, when we compare Australia and New Zealand, they have quite different GDP but they have almost the same suicide rate. Also, Countries like Fiji and Kiribati, they both have lower

GDP but they have small suicide rate. This changes my stereotype for the relationship of suicide rate and economics status. There're more reasons that lead to suicides.

Worldwide Number of Suicides/Population and Average GDP/Capita

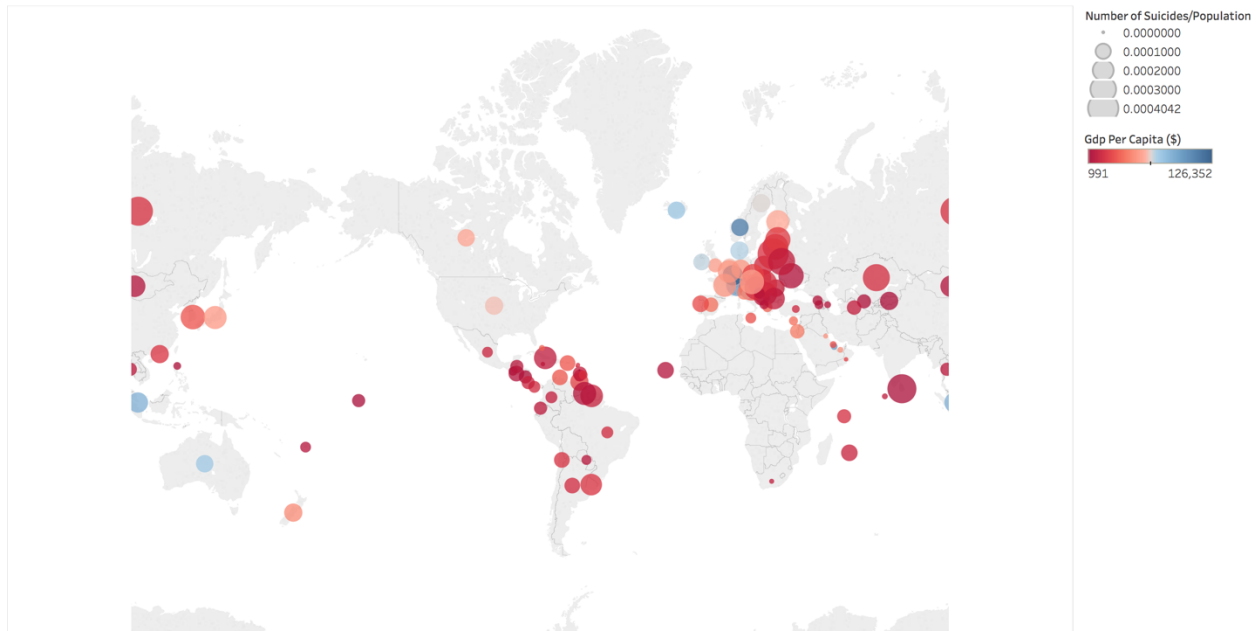


Figure 12: Worldwide Number of Suicides/Population and Average GDP/Capita
(interactive plot)

Results/Summary/Conclusion:

While we have some stereotype of correlation between data, it's better for us to deep dive the data first before any conclusion. Like my assumption in my storyline, I won't notice the special cases in Fiji and Kiribati if I didn't create the bubble plot on the map. Besides, it's important to think about how to process your data while visualizing data. For example, if I choose 'average GDP' as the indicator, the plot will dramatically change. We need to deeply elaborate the data before what we want to present on the plot.

Link to github page with this analysis:

<https://github.com/JoyceMTChang/datavis>