Final Project Report
Haoyu Wang, Ziyu Wang

For the final project, we discovered the variables that affect the suicide rates and did a bit of research to find out the reasons behind those phenomenons. The dataset we mainly focused on is the Suicide Rates Overview 1985 to 2016 from Kaggle. We also utilized the 2015 World Happiness Report dataset from Kaggle as a supplement for getting more insights into the topic. The factors that will affect the suicide rates will be analyzed from four aspects: age and gender, country happiness rank, generation, and GDP.

Age and gender:
From Figure 1, it is easy to tell that people between the ages of 35 to 75 which is also known as middle age are much easier to commit suicide. We can tell that the midlife crisis is not an urban legend. From Figure 1 we can also tell that males always have a higher suicide rate than women. In the age ranges 5 to 14 and over 75 years old, the difference between the suicide number based on gender is not obvious. However, men are 3 to 4 times easier to commit suicide than women for the other age ranges. After researching, the most reasonable explanation is that traditional male gender roles discourage emotional expression and this will make them more likely to get mental illness. Based on the United Nations' report, there is a well-established link between suicide and mental health issues such as depression and alcohol use disorders.

Country happiness rank:
In order to find out how counties' happiness rank affects the suicide rate, we filtered the countries with the top 5 happiness rank (Figure 2) and the countries with the bottom 5 happiness rank (Figure 3) that are included in both of the datasets. The smaller the number means the happier the country is. Then we use the results we got to filter the suicide number per 100000 people in the top (Figure 4) and bottom 5 (Figure 5) countries from 1995 to 2015. The labels on the end of each line mean the happiness rank for that country in 2015. After comparing the two figures, we noticed that for the upper 2/3 countries in the World Happiness Rank, countries with a higher happiness rank generally have a higher suicide rate. One of the reasons causes this phenomenon is the difficulty of collecting data from underdeveloped countries based on the United Nations' report. There are a total of 158 countries in the World Happiness Report dataset, but we can find non of the bottom-ranked 56 countries in the Suicide Rates Overview dataset. From Figure 6 we can tell the countries that have a higher happiness score are mostly developed countries and countries with a lower happiness score are mostly underdeveloped countries. Low level of literacy, the difficulty of access to the people because of the war, and the expensive price of the census are all the reasons for limited data collection in underdeveloped countries.
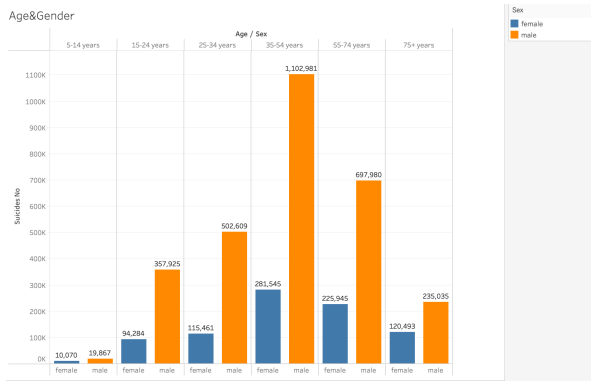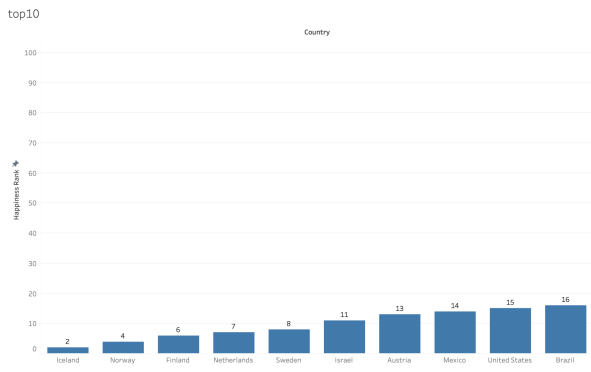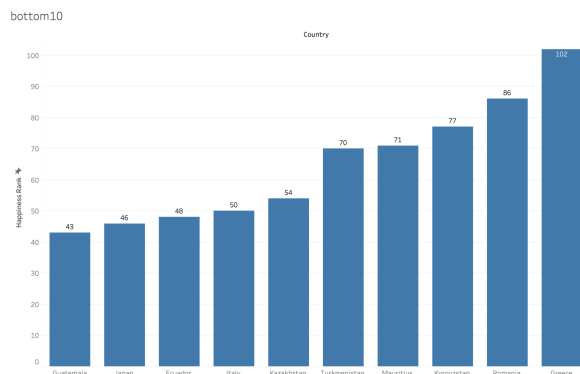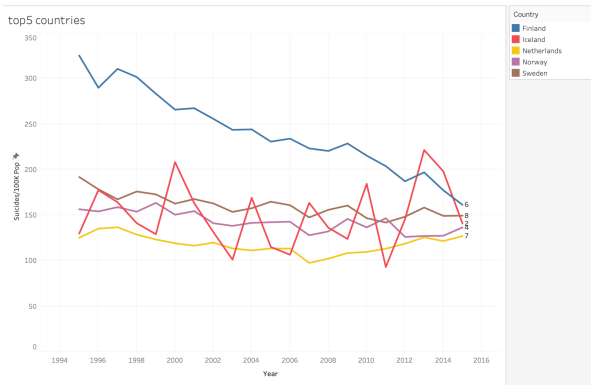
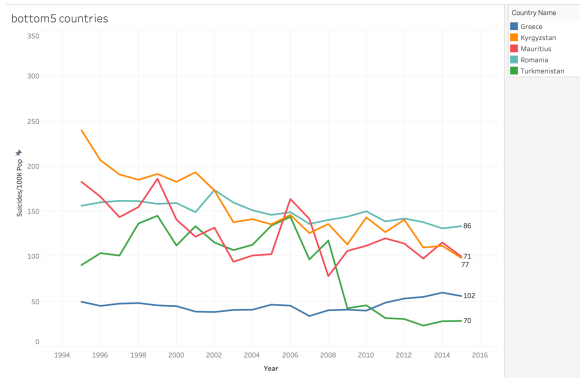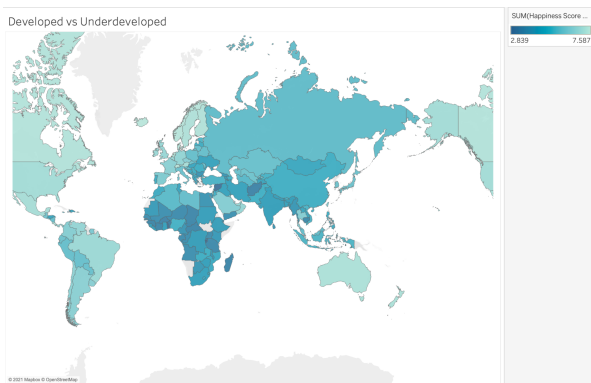Figure 1

Figure 2


Figure 3

Figure 4

Figure 5

Figure 6

Generation:

As an important categorical variable listed in the dataset, it is crucial to find out the suicide patterns over generations.

It is common sense that worldwide events that are severe enough could influence a whole generation. To see the actual relationship between the suicide rate and generation we analyzed the data and it pictured us some important insights. According to the figure below, it is easy to detect that the Silent and the Boomers are the first and second high suicide rate generations. These results seem plausible when we consider the fact that these generations were from the time period of World War II. The Silent period defines people who were born in 1928–1945 while the Boomers generation consisted of individuals who were born in 1946–1964. This provokes us to further analyze the generation with the suicide rate.

Over 50 million people lost their lives due to this war, and thus, many people lost their fathers, mothers, sisters, brothers, and friends. According to some psychological reports, people raised by single parents or in an orphanage have a higher chance to form an unhealthy personality plus a pessimistic social atmosphere. People were left alone and can not tolerate such an extreme mental burden that they no longer wanted to live anymore, and unfortunately might have committed suicide. Not to mention the trauma of the events would remain to haunt them throughout the rest of their lives especially among the soldiers experiencing PTSD and Veterans with disabilities, the picture below shows a severe suicide trending on the soldiers returned from battle compred to other citizens.
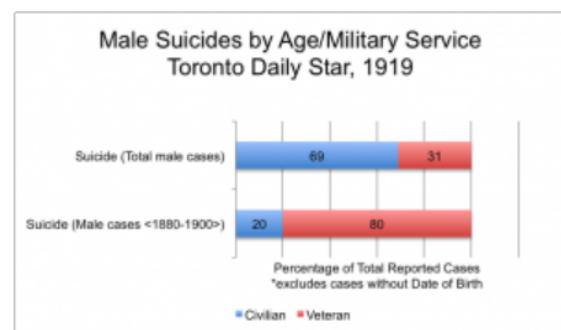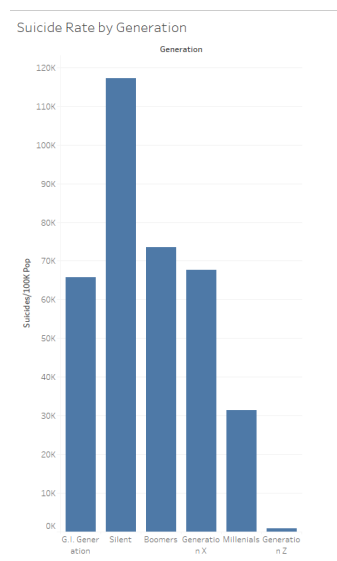

GDP:

As Abraham Maslow mentioned in his theory of the Hierarchy of Needs, physiological needs, food, water, sleep, sex, homeostasis, and excretion, are always the fundamental part of surviving. So the amount of wealth accumulated by a person also has an effect on the suicide rate.

We can see in the visualization below that the suicide rate keep dropping since 1995 while the GDP per Capita has maintained a growth trend. Here, we can infer that the growth in personal wealth has a positive effect on reducing the suicide rate.
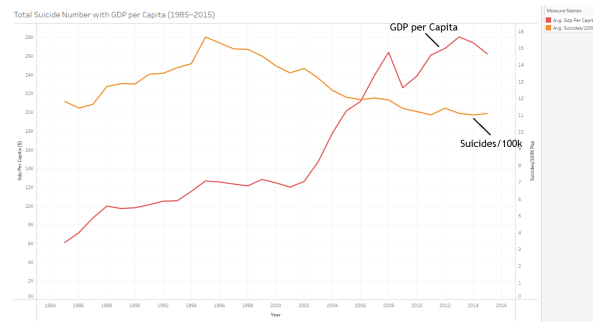

As we dive into the level of countries, different patterns emerged. In the US, as well as in the Republic of Korea, the suicide rate keeps growing while GDP per Capita increases.
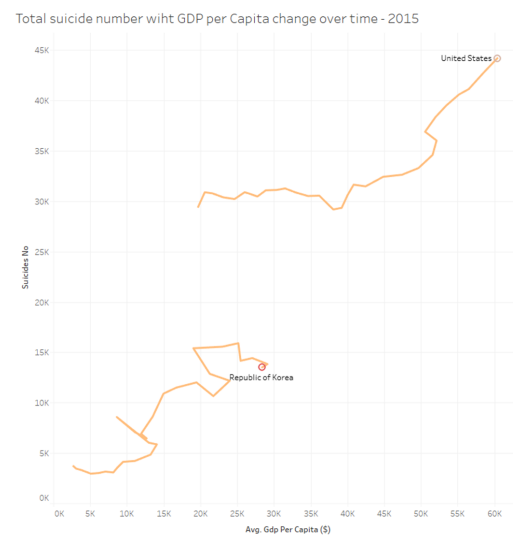
In some countries like the UK, the suicide rate is relevantly the same when the GDP per Capita increases.
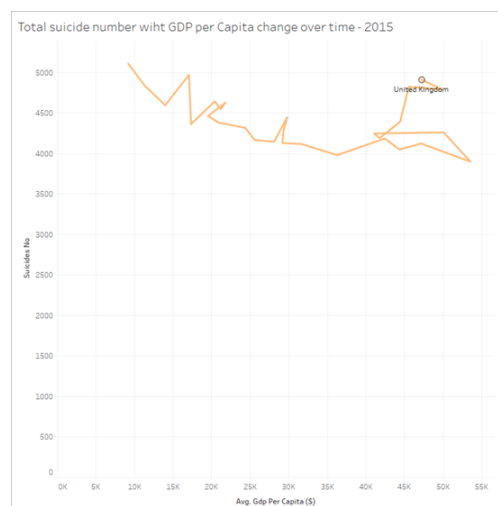
Last but not the least, the are some countries' suicide rates keep dropping with the growth of GDP per Capita, like Russia and Germany.
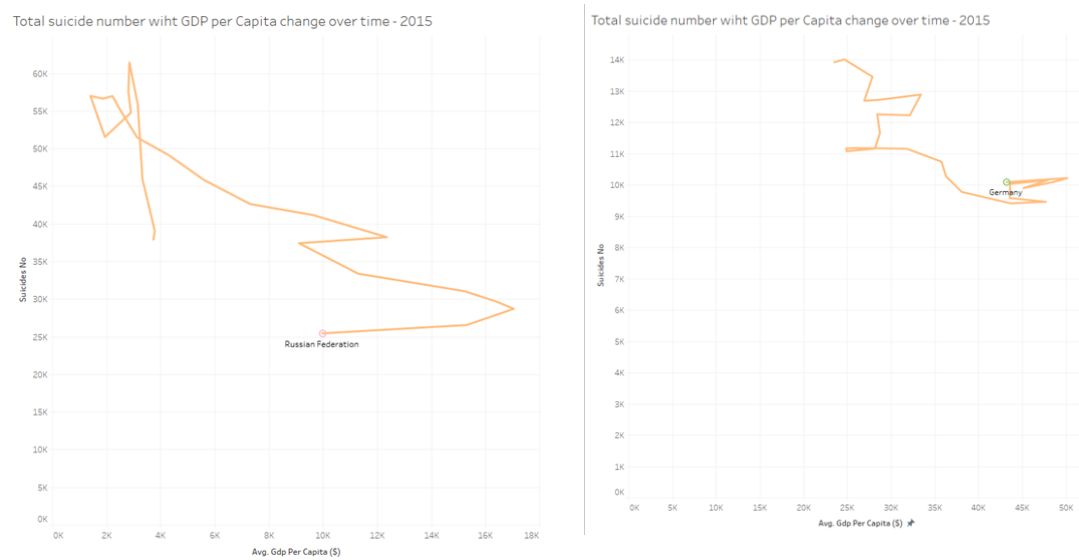
This conclusion yields that GDP per Capita has a correlation with other factors, but we could still use the suicide rate change over different age groups, sex, and time to find clusters. It is reasonable to select the suicide rate and the GDP per capita to reduce the different populations among countries and decrease the level of the dimension of the data frame. The final data frame received by the algorithm is shown in the figure below.
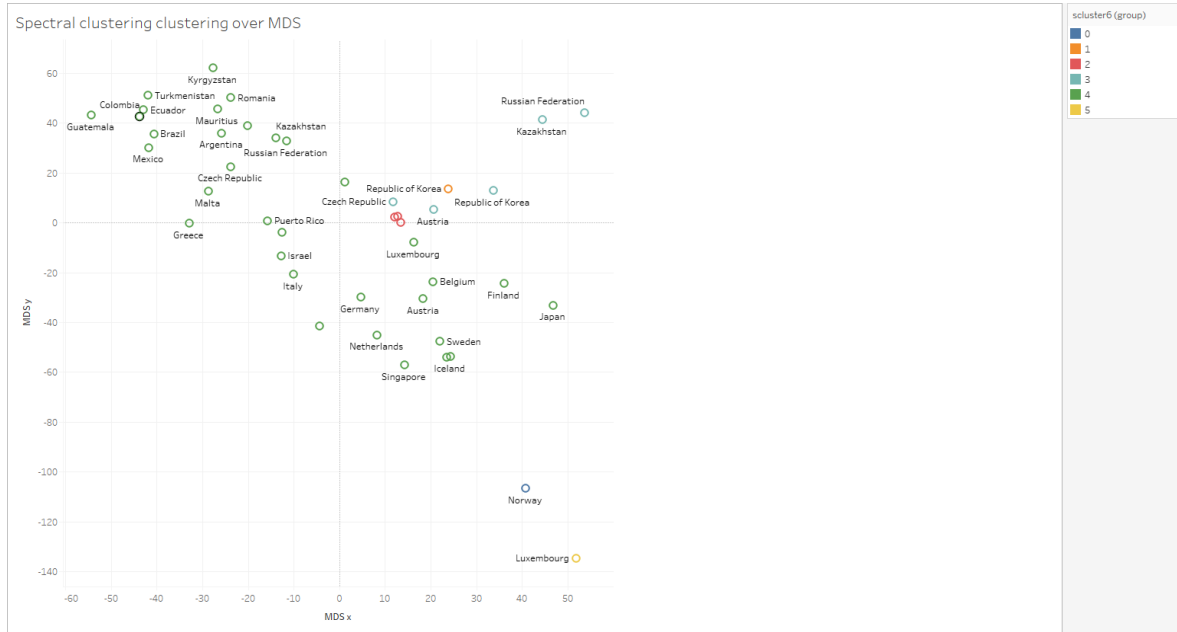

| country | age | sex | suicides/100k pop | | | | | | | | | ... | gdp_per_capita ($) | | | | | | |
			1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	...	2006	2007	2008	2009	2010	2011	2012
Argentina	15-24 years	female	-0.589713	-0.582829	-0.570257	-0.599596	-0.631629	-0.583944	-0.511492	-0.499893	-0.433699	-0.480958	...	-0.923603	-0.897554	-0.865788	-0.876307	-0.814315	-0.756214	-0.738198
		male	-0.300347	-0.294296	-0.326905	-0.232608	-0.184114	0.003350	0.098384	0.142297	0.310898	0.335200	...	-0.923603	-0.897554	-0.865788	-0.876307	-0.814315	-0.756214	-0.738198
	25-34 years	female	-0.626984	-0.647656	-0.605479	-0.683463	-0.700299	-0.649261	-0.630095	-0.587541	-0.601587	-0.564738	...	-0.923603	-0.897554	-0.865788	-0.876307	-0.814315	-0.756214	-0.738198
		male	-0.292065	-0.269853	-0.298621	-0.234297	-0.207392	-0.068725	0.007688	0.067572	0.140284	0.091074	...	-0.923603	-0.897554	-0.865788	-0.876307	-0.814315	-0.756214	-0.738198
	35-54 years	female	-0.578842	-0.528630	-0.545175	-0.600159	-0.618245	-0.554664	-0.523701	-0.521805	-0.490388	-0.520906	...	-0.923603	-0.897554	-0.865788	-0.876307	-0.814315	-0.756214	-0.738198

After using the MDS, TSNE to gain visualizations of the level of similarity of individual cases of a dataset, we use K-mean and Spectral Clustering and set the parameter of a number of clustering to 6. Since the Spectral Clustering algorithm yields an unreasonable distribution of the labels, it is hard to find a reasonable explanation to interpret this kind of phenomenon. We can also see the performance of Spectral Clustering is bad on the next page.

But on the other hand, the K-mean algorithm gives an understandable result. As shown on the visualization, there are basically no overlapping areas and countries that have similar patterns resulting in the same group. And it showed the outlier countries which are Luxembourg and Norway whose suicide rates are significantly greater.
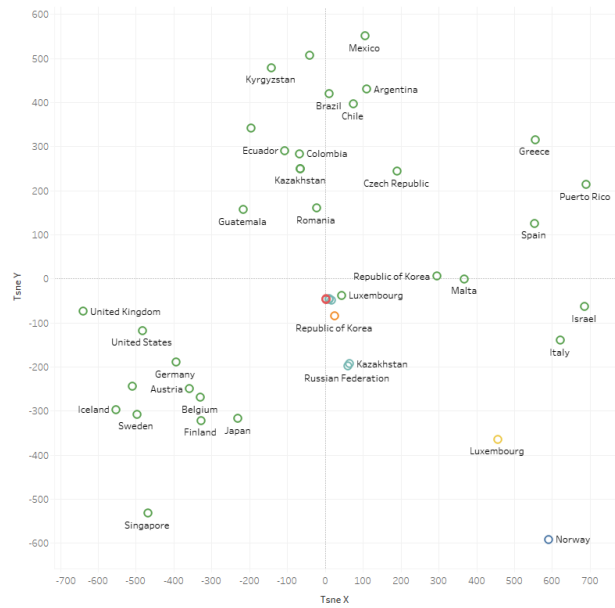
Thus, it is reasonable to conclude that for developing countries like Russia, the living standard has improved, and the people can feel it firsthand. They are happy to get positive feedback when they contribute their time and labor to their work, which is the improvement of their income. For countries like the US and Korea, it is very hard to survive — social anxiety, mortgage, pressure from work, and lacking a proper health care system. These could lead to mental problems over time and increase the chance of people committing suicide. And it is reasonable to form a theory that mental health is the most essential part to prevent suicide.

Spectral clustering clustering over MDS

Spectral clustering clustering over TSNE

scluster6 (group)
- 0
- 1
- 2
- 3
- 4
- 5
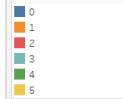
Kmean clustering over MDS
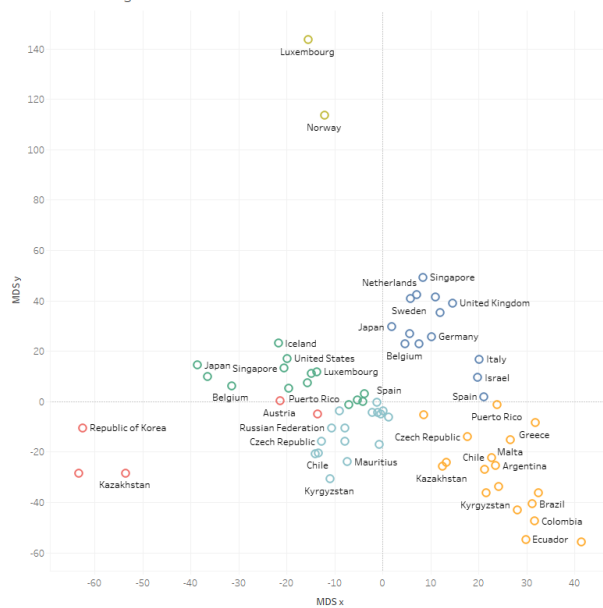
Kmeans6 (group)
- 0
- 1
- 2
- 3
- 4
- 5
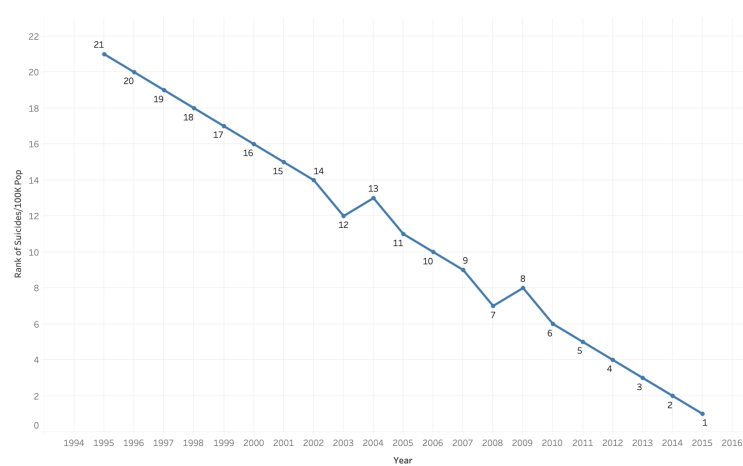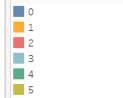
After finishing this project, we were shocked by the fact that there is a trend within a generation and a huge suicide rate. Even though it shows a trend of decreasing as the bump chart displayed above, it is essential for each of us not only to work hard to gain a better quality of life but also keep monitoring our mental condition. This really shows us how the issue of suicide is so relevant in every generation that we look into. No matter how little or more, in every generation there have been a significant number of suicides. This is a harsh reality that we all must realize so that we could work as a society and encourage the government to find solutions to bring an end to such tragic events.