Suicide Causes and Prevention
Haoyu Wang, Ziyu Wang

A) Project and datasets summary

During the pandemic, millions of lives have been taken by the virus around the world and countless families fell into pieces because of the loss of loved ones. At the time that people are trying so hard to fight against death, an estimated 703000 people loss their precious life by committing suicide worldwide each year based on the American Foundation for Suicide Prevention. It makes us wonder about the reason behind this phenomenon. For the final project, we are planning to discover the correlation between suicide and various variables and contribute our findings to prevent suicide.

The dataset we will use is Suicide Rates Overview 1985 to 2016 from Kaggle (https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016). This dataset was created to identify signals linked to higher suicide rates in various cohorts throughout the world, across the socioeconomic range. We will first generate several visualizations to find out the suicide number changing by year and country. Next, we will combine the findings with the generation to analyze the reasons behind the changes. The suicide rates may also differ between sexes and age groups so we will also dig into these aspects. The economic situation is a big part of life, thus the discovery of the relationship between country GDP and the suicide rate is essential.

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	796	Generation X
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	796	Silent
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	796	Generation X
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	796	G.I. Generation
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	796	Boomers

To get more insights, we would like to utilize the World Happiness Report from Kaggle (https://www.kaggle.com/unsdsn/world-happiness) to check the interrelation of the happiness index and suicide rate based on countries. Since the year range of the cleaned Suicide Rates Overview dataset is 1995 to 2015, we will only use the 2015 World Happiness data.

	Country	Region	Happiness Rank	Happiness Score	Standard Error	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
0	Switzerland	Western Europe	1	7.587	0.03411	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678	2.51738
1	Iceland	Western Europe	2	7.561	0.04884	1.30232	1.40223	0.94784	0.62877	0.14145	0.43630	2.70201
2	Denmark	Western Europe	3	7.527	0.03328	1.32548	1.36058	0.87464	0.64938	0.48357	0.34139	2.49204
3	Norway	Western Europe	4	7.522	0.03880	1.45900	1.33095	0.88521	0.66973	0.36503	0.34699	2.46531
4	Canada	North America	5	7.427	0.03553	1.32629	1.32261	0.90563	0.63297	0.32957	0.45811	2.45176

B) EDA on the selected datasets

1. Datasets investigation

Multiple datasets were explored during the process. The first topic we were planning to do is about air pollution. Variables that affect air pollution are stored in different datasets. After data cleaning and merging, we realized there is not much year intersection between these datasets which means the actual available data is too little. The lack of tuples makes the data visualization and analysis unrepresentative. Our second choice is how government policies affect the COVID-19 infection number. Many datasets about this topic can be merged for gaining more insights. However, the policy description is inconsistent for all countries, and it makes us hard to evaluate them under the same criteria. The third dataset we found is World Trending YouTube Video Statistics. The difficulty of using this dataset is that the video titles, channel titles, and video tags are written in different languages. This becomes a barrier for us to understand and interpret the data.

2. **Datasets composition**

The metadata of the Suicide Rates Overview 1985 to 2016 dataset are from four other datasets stated below:

- Human Development Index dataset from the United Nations
  http://hdr.undp.org/en/indicators/137506
- GDP by Country dataset from the World Bank
  http://databank.worldbank.org/data/source/world-development-indicators#
- Suicide Rates dataset from the World Health Organization
  https://www.who.int/data/gho/data/themes/mental-health/suicide-rates
- Suicide Prevention dataset from the World Health Organization
  https://www.who.int/health-topics/suicide#tab=tab_1

The metadata of the World Happiness Report dataset is from the United Nations Sustainable Development Solution Network - https://worldhappiness.report/

3. **Variable types and definition**

For the Suicide Rates Overview 1985 to 2016 dataset, there are 12 variables and 27820 tuples. Categorical data are country, sex, age, country-year, and generation. Numerical data are year, suicides number (suicides_no), population, suicides per 100000 population (suicides/100k pop), Human Development Index of the year (HDI for the year), Gross Domestic Product for the year (gdp_for_year), and Gross Domestic Product per capita (gdp_per_capita).

In this dataset, the variable “country” is the name of the country, “year” is the year for the tuple, “sex” is the gender of the age group, “age” is the age group, “suicides_no” is the number of suicide people in the specific age group of the particular gender, “population” is the population in the specific age group of the particular gender, “suicides/100k pop” is the suicide rate per 100000 people, “country-year” is the name of the country combines the year, “HDI for year” is a composite indicator between 0 and 1 where 1 is the aspirational target. It measures the average performance in three key aspects of human development: a long and healthy life, knowledge, and a decent standard of living. “gdp_for year” is the Gross Domestic Product in the unit of U.S.

dollar, “gdp_per_capita” is the GDP per person based on the country, and “generation” is the generation the year belongs to.

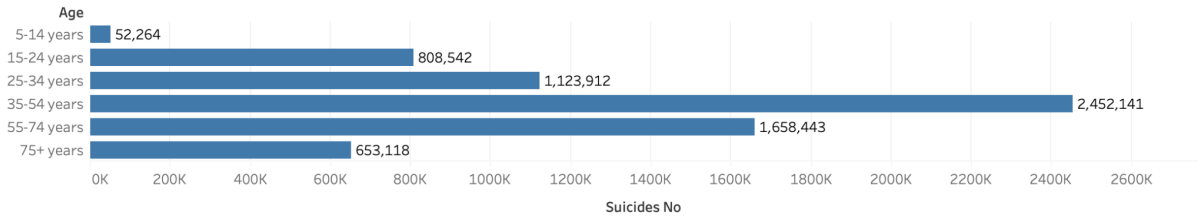
For the World Happiness Report dataset, there are 10 variables and 158 tuples. Since this dataset is supplementary to the topic, we will not use all the variables in the dataset. For the variables we are going to use in the project, the categorical data are country and region, and the numerical data are happiness rank and happiness score. In this dataset, the variable “country” is the name of the country, “region” is the region the country belongs to, “happiness rank” is the rank of the country based on the happiness score, and “happiness score” is a rating between 0 and 10 where 10 is the happiest.

In general, the variables in both datasets have the correct types, the size of the dataset is manageable, and the speed of the software is sufficient.

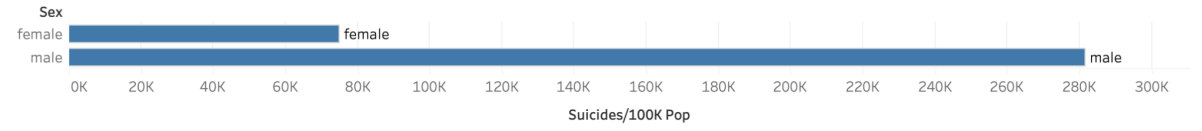
4. Exploratory data analysis

A) Graphs

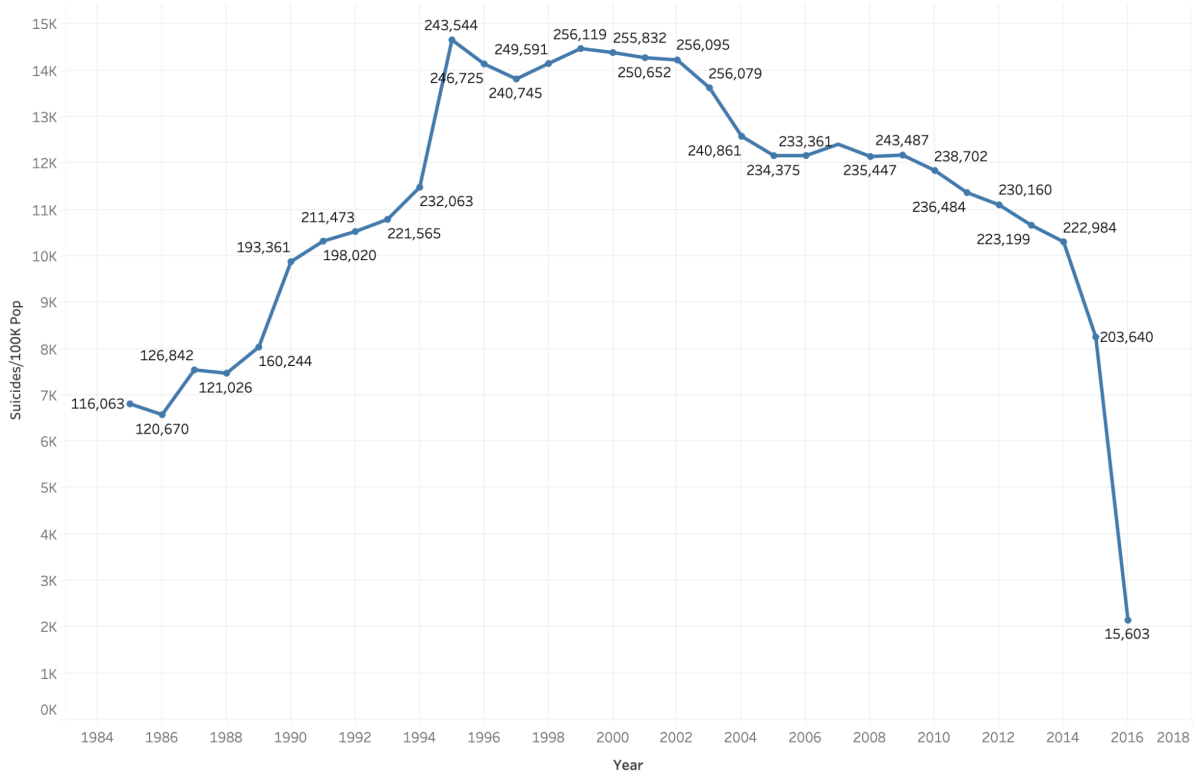
For the Suicide Rates Overview dataset, we generated several graphs to check whether there are correlations between the variables. The first graph is the correlation between the suicide number and age. The correlation is obvious. The people who are in the 75+ ages range are most easily to commit suicide.



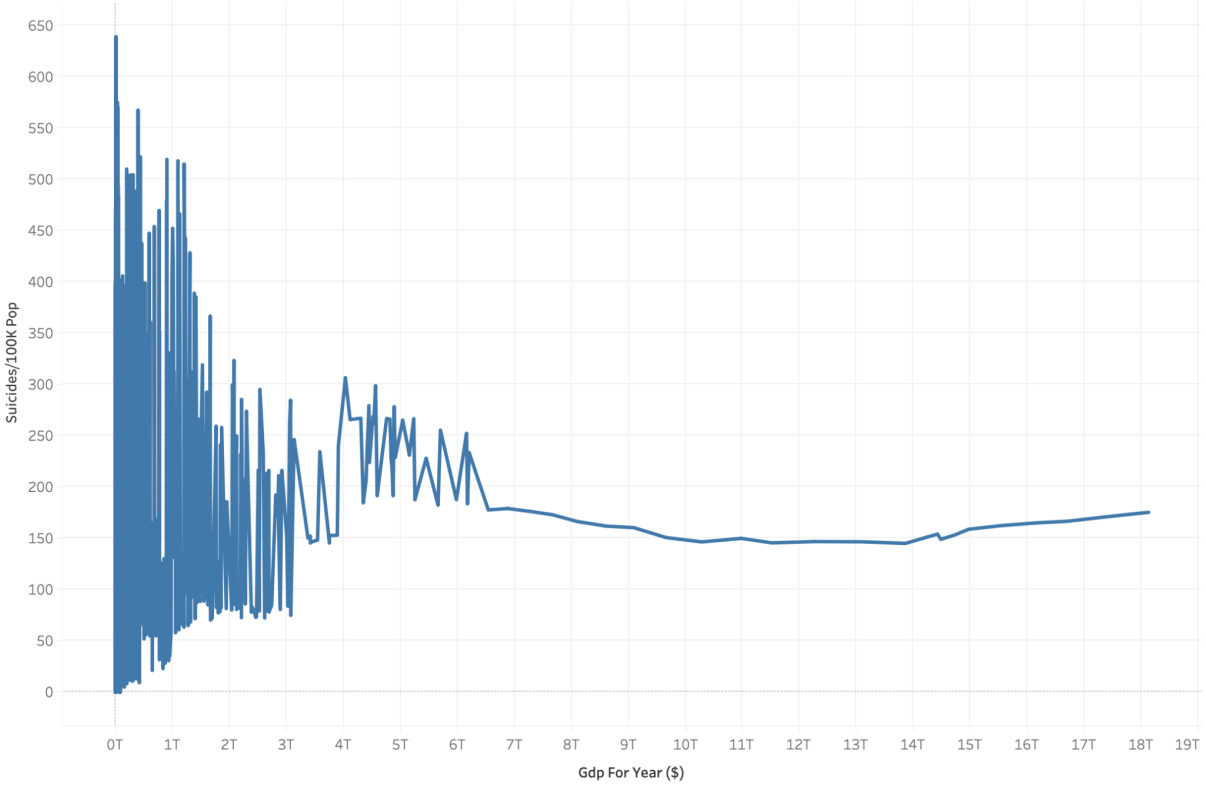
The second graph is the correlation between the suicide number per 100000 people and gender. The correlation is obvious. It is much more easier for males to commit suicide than women.



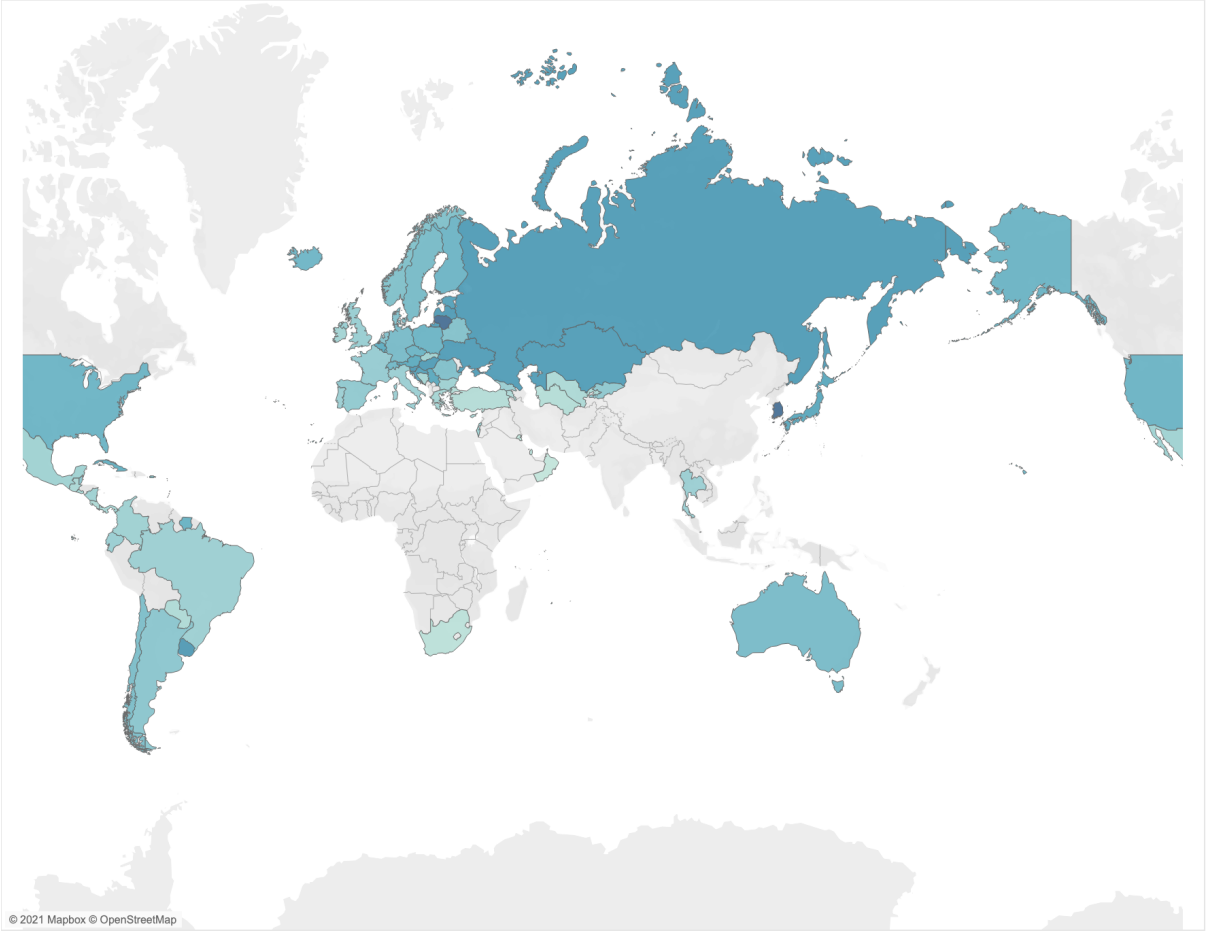
The third graph is the correlation between the suicide number per 100000 people and year. The correlation is obvious. The number of people committing suicide kept increasing from 1988 to 2004 and 2006 to 2010. There might be some problems with the 2016 year’s data because of the huge drop in suicide numbers.



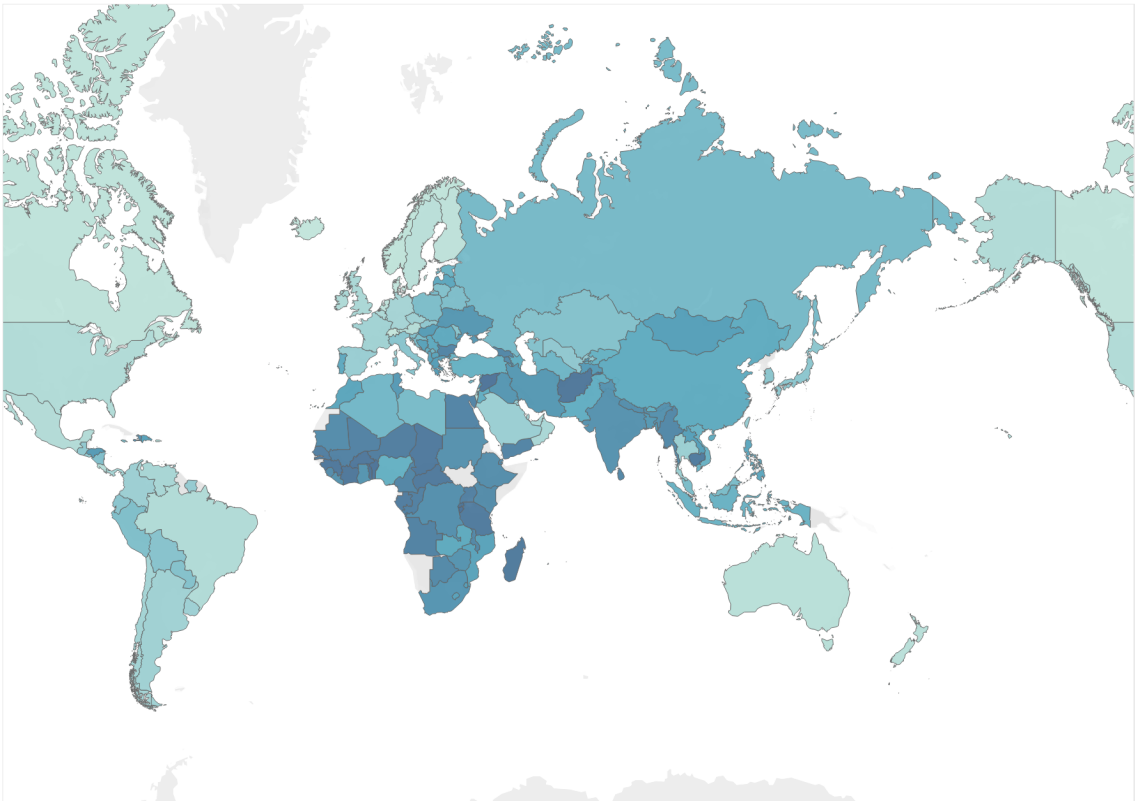
The fourth graph is the correlation between the suicide number per 100000 people and GDP for years. The curve becomes smoother after 7T but the relationship between the suicide number per 100000 people and GDP for years is not obvious. It would be best if we can find out the reason that cause the strong changes before the 7T to increase the data accuracy.



The fifth graph is the correlation between the suicide number per 100000 people and the country in 2015. The correlation is obvious. The deeper the color is means the more people committed suicide in 2015.



For the World Happiness Report dataset, we generated a data visualization to show the happiness rank on the map. The country with a deeper color means the rank is lower. Also, from the chart, we can tell that most of the countries in the world are covered in the dataset.



B) Numerical variables

For the Suicide Rates Overview dataset, there are 6 numerical variables. After checking the variable summary, three variables should be paid more attention to during the null value and outlier detection process. The variable that contains many null values is HDI for the year. The total count of HDI for the year is much less than the other variables which means there should be many null values in that record. The variables that may contain outliers are suicide number and GDP per capita. The minimum suicide number is 0 and the maximum number is 22338. If the corresponding suicide rate for the maximum suicide number is too large, then there must be outliers in that record. The minimum and maximum numbers for the GPD per capita illustrate the large wealth gap. Analyzing these numbers with the standard deviation of the record, we can find the potential outliers.

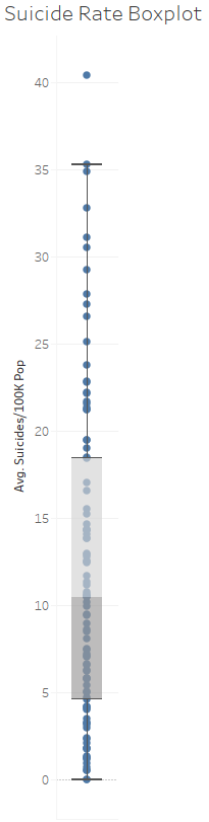
	year	suicides_no	population	suicides/100k pop	HDI for year	gdp_per_capita ($)
count	27820.000000	27820.000000	2.782000e+04	27820.000000	8364.000000	27820.000000
mean	2001.258375	242.574407	1.844794e+06	12.816097	0.776601	16866.464414
std	8.469055	902.047917	3.911779e+06	18.961511	0.093367	18887.576472
min	1985.000000	0.000000	2.780000e+02	0.000000	0.483000	251.000000
25%	1995.000000	3.000000	9.749850e+04	0.920000	0.713000	3447.000000
50%	2002.000000	25.000000	4.301500e+05	5.990000	0.779000	9372.000000
75%	2008.000000	131.000000	1.486143e+06	16.620000	0.855000	24874.000000
max	2016.000000	22338.000000	4.380521e+07	224.970000	0.944000	126352.000000

For the World Happiness Report dataset, there are two numerical variables we are going to use for the project. There seems no obvious problem with the data based on the summary of these two variables.

	Happiness Rank	Happiness Score
count	158.000000	158.000000
mean	79.493671	5.375734
std	45.754363	1.145010
min	1.000000	2.839000
25%	40.250000	4.526000
50%	79.500000	5.232500
75%	118.750000	6.243750
max	158.000000	7.587000

C) Null values and outliers

Situations varied from country to country for this section. There are definitely outliers in the record of ‘GDP’, ‘GDP per Capita’, ‘suicides_no’, and ‘suicides/100k pop’ due to the different sizes and different levels of development of the country. The boxplot below shows the existence of an outlier in the suicide rate. And we can see that there are some points near 0. Two kinds of situations can explain these points. First is the points themselves equal to 0. Second, it could be caused by a lack of years of record, such as there are zero records for Dominica in the dataset.



For NULL values, there is a huge amount of “HDI for year” missing from the dataset. The reason behind this is a concept introduced by the 2010 Human Development Report. The concept divided the target country into 4 different ranks indicating different levels of human development. There are a few records of 2016 in our dataset in which case most of the countries don’t have any records.

In categorical variables, ‘age’ is missing some of the records as the frequency table shows below. After using ‘country’ to break down the aggregate we can find out that Mongolia is missing the records of the age group from 5 years old to 14 years old. Also for the column ‘sex’ and ‘generation’ there is no missing data, for ‘generation’ there are different frequencies due to different years of record among countries.

Age	
5-14 years	4,610
15-24 years	4,642
25-34 years	4,642
35-54 years	4,642
55-74 years	4,642
75+ years	4,642

Sex	
female	13,910
male	13,910

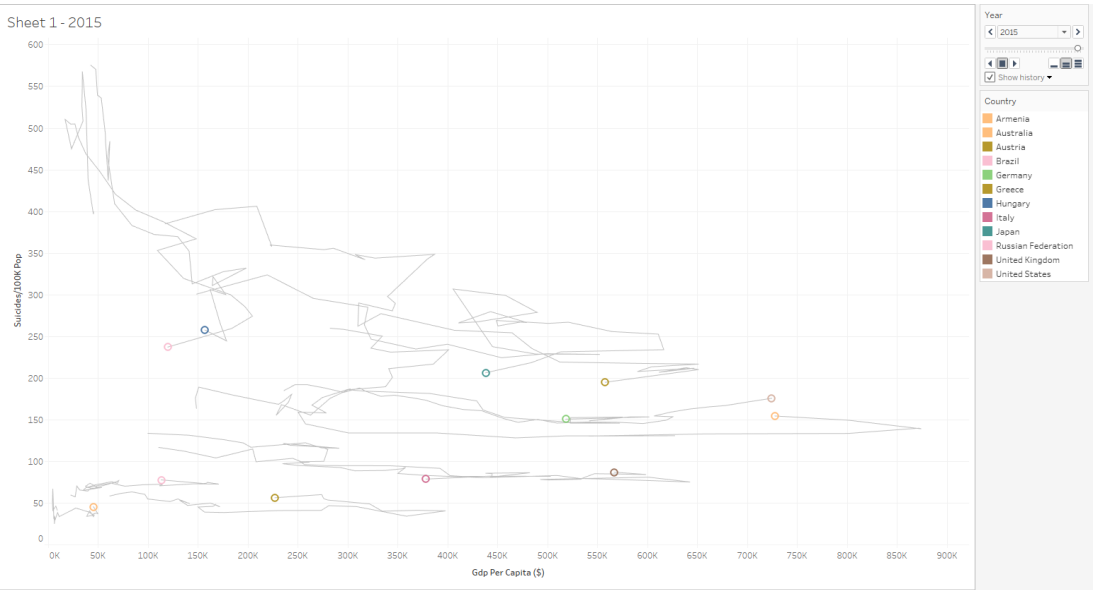
Generation	
Boomers	4,990
G.I. Generation	2,744
Generation X	6,408
Generation Z	1,470
Millenials	5,844
Silent	6,364

Last but not the least, for the variable 'year', there are several countries only have only a few years of records such as Cabo Verde, Macau, Turkey, Dominica

Country	
Albania	264
Antigua and Barbuda	324
Argentina	372
Armenia	298
Aruba	168
Australia	360
Austria	382
Azerbaijan	192
Bahamas	276
Bahrain	252
Barbados	300
Belarus	252
Belgium	372
Belize	336
Bosnia and Herzegovina	24
Brazil	372
Bulgaria	360
Cabo Verde	12
Canada	348
Chile	372
Colombia	372
Costa Rica	360
Croatia	262

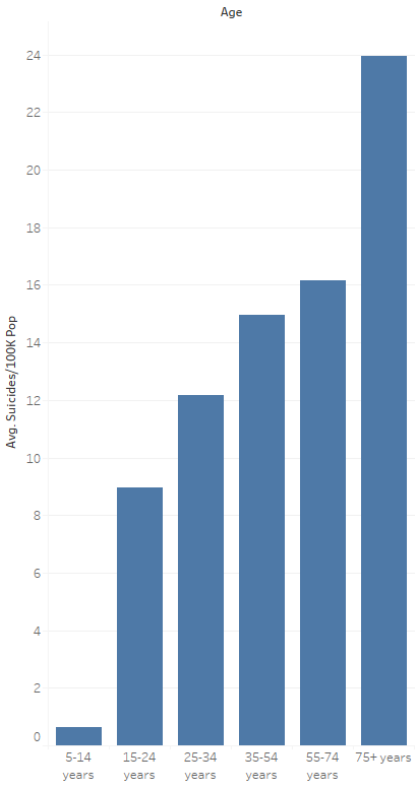
5. Data correlation

The variable we should be focusing on is Suicides/100K population due to the different sizes of the population. And the variables that could influence the Suicides/100K population could be sex, age, generation, GDP per Capita, etc.
We are using the connected Scatterplot utilize Suicides/100K population over GDP per Capita that to examine the correlation. The polt show an uncertain pattern that some country's number of Suicides/100K population does drop as the growth of the GDP per Capita, some country's suicide rates over 100K people have a very little change as the growth of GDP per Capita, some country's suicide rate and GDP per Capita keep growing at the same time. These different patterns could be displayed by the clustering algorithm in Python after breaking the original dataset into a low-dimensional dataset. Each cluster could be used to describe a unique pattern.



And some of the categorical variables do have an inner relationship with the suicide rate such as age, after plotting the bar chart shows a huge trend that elder people are more likely to suicide.

Last but not the least, we can see a clear view by plotting Suicides/100K population by different age groups and gender. And we can see that there is an almost 3 times more male kill themselves than female



	Sex	
Age	female	male
5-14 years	1,065	1,793
15-24 years	10,045	31,487
25-34 years	10,614	45,957
35-54 years	13,732	55,654
55-74 years	16,534	58,461
75+ years	23,024	88,177

Generation	
Boomers	4,990
G.I. Generation	2,744
Generation X	6,408
Generation Z	1,470
Millenials	5,844
Silent	6,364

6. **Data cleaning**

For the World Happiness Report dataset, their records are ordered so there is no need to do extra cleaning.

For the Suicide Rates Overview dataset, the first thing to do is delete the column of Human Development Index - this concept is first shown in 2000 which leads to 70% of missing data in the dataset. Secondly, deleting the records of several records of 2016 won't help us since most of the data are from 1985-2015 and most of the countries do not have the record of 2016. Third, delete the countries that have very few records such as Cabo Verde, Macau, Turkey, Dominica.

Last but not the least, we should create a subset of the original dataset that has a reasonable intersection of the time period which most of the countries hold. After doing the EDA of the dataset we found that it is reasonable to use the constraint of 1995<=year<=2015 to subdivide the dataset. Notice that we do allow the countries to miss some years of records because the dataset is used to visualize the trend of the data. Also, it is not proper to miss too many years to see the trend so it is necessary to set a limitation that all the countries in the subset should have at least 10 years of record. And for the dataset we used to do the clustering, it is necessary to make the countries have the same period of time to run the algorithm. Thus, we filter the country which contains exactly 10 years of data to form a subset.

7. **Merging datasets**

Though we will be using two datasets in this project, we are not going to merge the Suicide Rates Overview dataset with the World Happiness Report dataset. The year range for the Suicide Rates dataset is from 1995 to 2015, but the year range for World Happiness Report dataset is only 2015. Also, there are a few differences between the countries included in these two datasets. If we use tools such as Python to merge two datasets with different labels, then too much work will be conducted. Instead of merging the datasets, we will use the results from the World Happiness dataset to filter the Suicide Rates dataset for final results.

With the development of the research, we found mental health is one of the biggest factors that affect suicide rates. After searching, we found a mental disease dataset on World in Data website. We thought about merging this dataset with the Suicide Rates dataset, but the differences between the categorical variables make this thought difficult to be achieved. In the Suicide Rates Overview dataset, all records are broken down by gender and age while the Mental Disease dataset is recorded by year and country. It is hard and erroneous to combine age and gender to merge the dataset because this will lose detail of the pattern about gender and age.

8. **Storytelling**

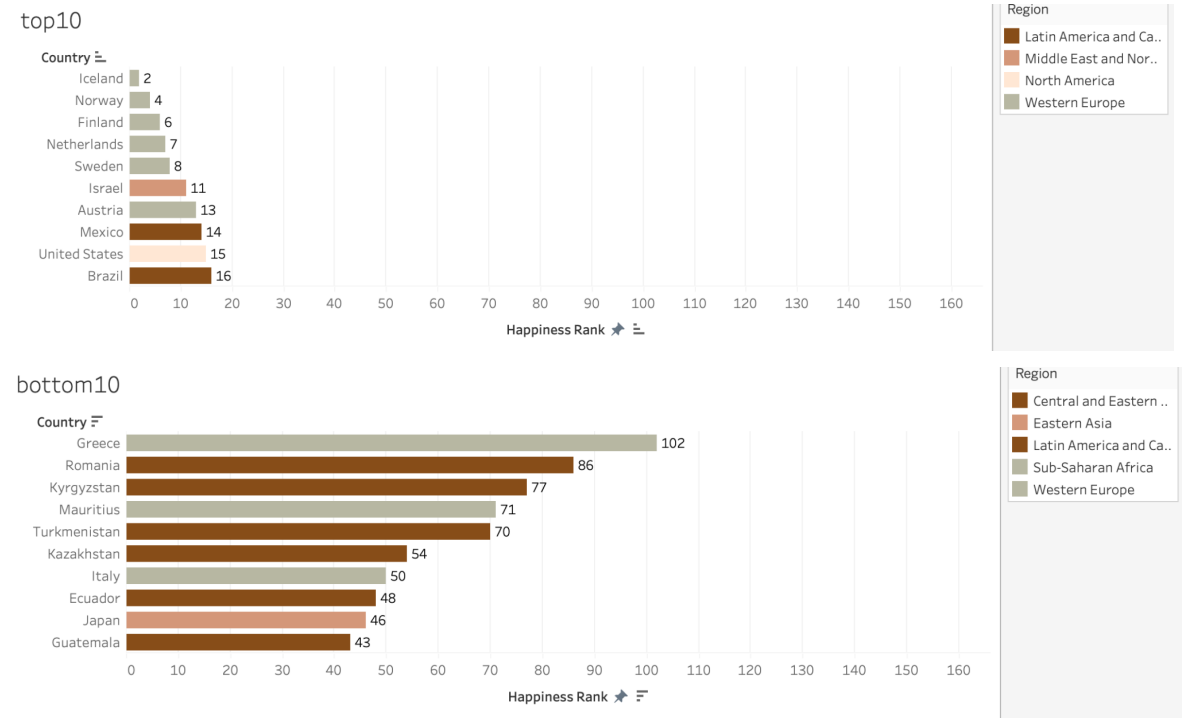
We are going to tell the story of how mental health influences on suicide rate and worldwide events influence the generation trend of suicide rate. The most important part of our storytelling is about GDP. As mentioned above, there is not a

general pattern that shows a relationship between GDP and suicide rate. With further research, we found there are actually several patterns for the GDP and suicide rate relationship. So it is meaningful to use the clusters to distinguish the different patterns, and based on that we could dive into a different group of countries to find what drives the growth of the suicide rate.

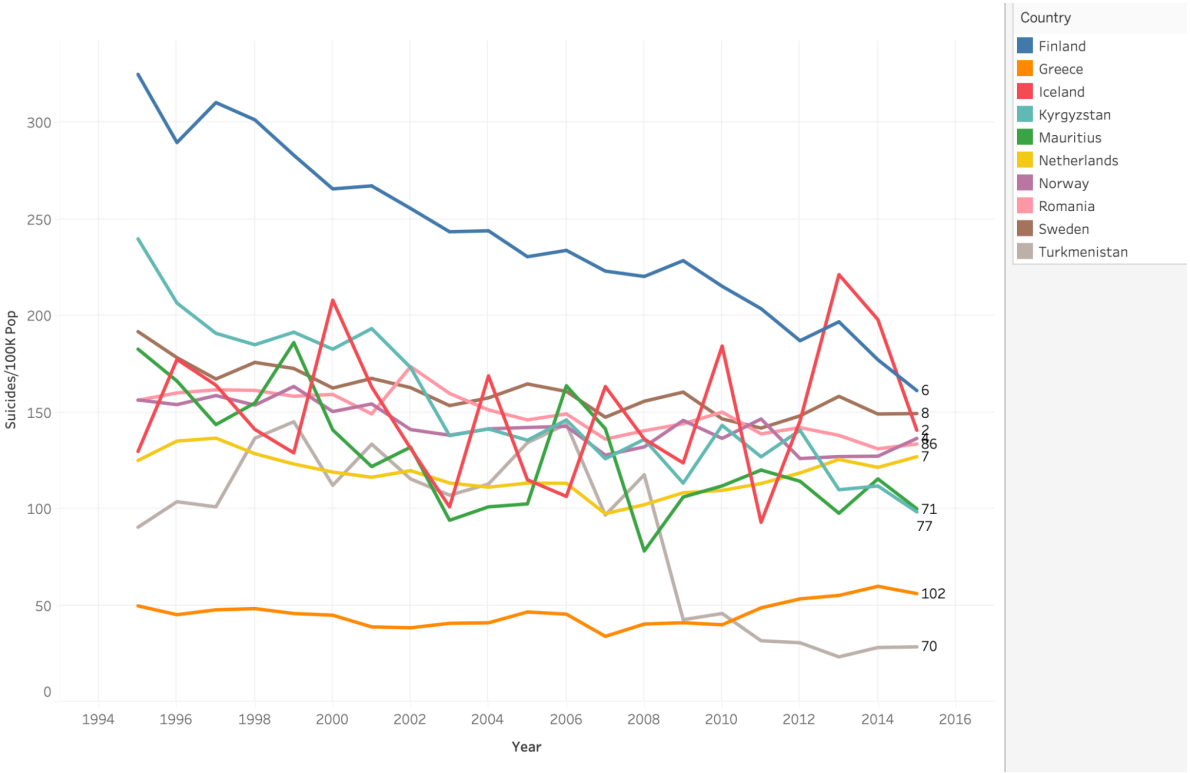
C) Visualization plan

The visualizations that we are planning to use are tables, line charts, bar charts, pie charts, bullet graphs, and heat maps. High Dimensional Visualization and clustering will also be used to help the research. Below are prototype visualizations for the project.

During the research process, we found 32 countries are in both the Suicide Rates Overview and World Happiness Record datasets. We utilized bar charts to show the top and bottom 10 countries which are in both of the datasets. The labels on the end of the bars represent the country's Happiness Rank worldwide. The lower the number is means the people in that country feel happier. The color of the bar represents the region the specific country belongs to.



Then we filtered the top and bottom 5 ranked countries from the results we got above to see the overall trend of the relationship between Suicide number per 100000 people and World Happiness Rank. The labels on the end of the lines are the country's happiness rank worldwide. The lower the number is means the people in that country feel happier. From the graph below, it is obvious that the countries with a higher Happiness Rank in 2015 have higher suicide rates from 1995 to 2015 except Norway. Next, we will find out the reason behind this.


Country
Finland
Greece
Iceland
Kyrgyzstan
Mauritius
Netherlands
Norway
Romania
Sweden
Turkmenistan
Suicides/100K Pop
300
250
200
150
100
50
0
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
Year
6
8
2
86
7
71
77
102
70