Kseniya Husak
Final Project Report

Motivation

Mental health is getting more attention now than ever. With increasing figures concerning widespread depression and anxiety, many in our society are concerned with the impact of worsening mental health on suicide rates. As someone who lost two friends to suicide, it is a problem that I care deeply about. So, when I searched data on suicide trends on Kaggle, I found a dataset that allowed me to engage with this topic in a systematic way. The aim of my project was to analyze suicide trends to search for some patterns and possibly correlations across the socio-economic spectrum and to visualize suicide trends across countries and time.

Research Questions

1. How do suicide rates and totals differ across countries, generations, and gender?
2. How do suicide rates and totals change by year?
3. Is there a statistically significant relationship between suicide rates and economic prosperity?
4. Are there overarching common patterns when it comes to high suicide rates? And do suicide victims share similar attributes? (i.e. are majority of them men of a certain age regardless of country or does each country differ from one another?)

Data

Data Source: Kaggle
Data URL: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
Dataset Name: "*Suicide Rates Overview 1985 to 2016*"
Data Format: CSV
Data Variables:

Variable Name	**Variable Type**	**Description**
Country	string	Country
Year	Integer	Year
Sex	String	Male or Female
Age	String	Age group
Suicides_no	Float	Count of suicides
Population	Integer	Total population
Suicides/100k pop	Float	Rate of suicides per 100k people
HDI for year	Float	Human Development Index
gdp_per_capita	Integer	GDP per capita in $
generation	string	Generation

Total Records: 27,820
Time Period Covered: 1985 to 2016

Methods Question 1

(a) how did you manipulate the data to prepare it for analysis?

One of the biggest challenges of working with the dataset is the fact that it is a timeseries – each country has observations for each year. So, it's a very long dataset. In addition, there is further breakdown of observations based on demographic variables (sex, age and generation). In order to analyze trends across countries, I relied heavily on grouping the data (by sex, age, and generation) and consequently collapsing the dataset from long to wide as much as possible. I also filtered my data when needed (top 10 vs. bottom 10 etc.). Luckily, all data columns were formatted well so I didn't have to do any manipulations to actual entries. I wanted to analyze both the total suicides as well as rate/100k of population across countries. The data manipulation was similar for both, save for the variable in question (I aggregated totals with sum and rate with mean function).

(b) How did you handle missing, incomplete, or noisy data?

When I looked for missing data, it was found exclusively in the 'HDI for year' column (a total of 19,456 missing). That's a huge chunk considering that the raw dataset has 27,820 records. Normally, I would drop the rows with missing data but in this case propping so many records would have been unwise. So, for this question specifically I ignored the HDI column entirely as it did not impact the analysis I had to perform. I also wanted to make sure that I had complete data for all of the years. After visualizing my data by total suicide numbers, I saw a sharp drop-off in the year 2016, which prompted a suspicion that my data for that year was incomplete. I investigated this pattern a bit further by grouping the data by year and counting numbers of unique countries for each year. As suspected, there were only 16 countries observed in 2016 – significantly fewer than previous years. Since my analysis would have been impacted by this missing data, I simply dropped all observations from 2016.

(c) How did you perform data analysis in code, i.e. Briefly describe the workflow of your source code?

To obtain the cumulative totals by country, I grouped my data by country and summed the suicide numbers, sorted in descending order. I then used a seaborn distribution plot to visualize the distribution of total suicides across countries. I also used the same grouped/sorted dataframe to plot total suicide number per country with a bar chart and individual country names. Next I created a bar plot of the top ten and bottom ten countries in terms of suicide totals (relying on. tail(10) and .head(10) to filter observations). Similarly, I grouped my data by 'generation' and 'sex' (separately) and used an aggregate function to get the cumulative total of suicide numbers for each category (sum). I then visualized the totals for each variable with a bar graph. In addition, I used a bar plot to visualize suicide totals by age and by sex to so how age categories might differ between the two sexes. I repeated parts of this analysis to explore the rate of suicide per 100k of population, grouping my data in a similar fashion and relying on bar plots to visualize top/bottom ten countries. Last but not least, I got some descriptive statistics for the variables of interest with .describe( ).

(d) What challenges did you encounter and how did you solve them?

The biggest challenge was to group data correctly and get numbers that were conducive to my analysis. I found the aggregating options to be helpful (doing mean for rates and sum for totals) as well as only using columns that I needed for my analysis, thus simplifying the data into a smaller subset. I also initially wanted to do my analysis by country and by year together, but after some trial and error I thought it would be clearer and more effective to just focus on countries at first so as to guide the rest of my analysis and present easy-to-digest visualizations.

Methods Question 2

(a) how did you manipulate the data to prepare it for analysis?

For this question, I had to group my data by year since I was interested in analyzing trends across time in addition to trends across countries. I was able to create a grouped table with the average rate of suicide/100k of population and sum of total suicides per year across all countries. As states previously, my data filtered out any observations from year 2016 as those were incomplete and would introduce bias. This helped conduct high-level trend analysis. But I also wanted to look at trends across time for various generations and age groups. To get my data in the right format I relied on pivot tables to summarize total suicides by generation/sex across years.

(b) *H*ow did you handle missing, incomplete, or noisy data?

Although I didn't have any noisy data, some countries, like Russian Federation, did not have any observations for years prior to about 1989. I could have dropped all of the data for those years for all countries, as technically it's missing/incomplete data, but I chose to keep all observations and have some 'gaps' in my final visualizations. Although somewhat incomplete, I thought it was important to keep as much of my data as possible to get a fuller picture. And generally, I think even with incomplete data we can infer overall trends better with the full dataset as opposed to the abridged complete one.

(c) How did you perform data analysis in code, i.e. Briefly describe the workflow of your source code?

I grouped my clean dataset (with .groupby( )) by country and year, summing all other variables (I only needed "suicides_no" for my analysis). I also create a list of top and bottom 10 countries based on my previous analysis to serve as a filter (I used .isin( ) to filter my data). Afterwards, I used seaborn boxplot to visualize respective distributions of suicide totals.
To visualize the overall trend of suicides across years without country-specific breakdown, I again relied on .groupby( ) year and summed suicide numbers. I then used this new dataframe to do a line plot with Matplotlib.
To perform my analysis of trends across time by generation and sex, I relied on pivot tables as opposed to groupby. For generation, I used 'year' as the index, 'generation' as columns and 'suicide_no' as values aggregated as sum. For sex, I used 'sex' as columns. I then used matplotlib to plot the trends.
Lastly, I wanted to gain further insight into trends across time for the top 10 countries in terms of total suicide cases. I used the pivot table method again (index='year', columns='country',

values=’suicides_no’ aggr=sum or ‘suicide_rate/100k pop aggr=mean). Again, I relied on matplotlib to visualize the trneds.

(d) What challenges did you encounter and how did you solve them?

The biggest challenge was to group my data correctly. I initially used groupby for all of the tables I needed but I couldn’t get the right columns to be able to plot trends by sex/generation/country. What worked for me however was using pivot tables since they simplified my data and I was able to have each categorical variable in a column with sum/mean values of the variable of interest. I also spent quite a bit of time trying to make my visualizations legible. Using online documentation for Matplotlib was super helpful.

Methods Question 3

(a) how did you manipulate the data to prepare it for analysis?

There are three variables that I needed for this analysis – ‘suicides/100k pop’, ‘gdp_per_captia’ and ‘HDI for year’. I didn’t have to do much data manipulation. I relied on .groupby( ) to obtain average values of variables of interest per country.

(b) How did you handle missing, incomplete, or noisy data?

Based on the analysis for the previous question, I was aware that my data was incomplete (missing values for suicide rate, HDI for year for a number of countries). However, since my statistical analysis relied on aggregate data averaged out across years, I ended up keeping the full data for one of the regressions (OLS on GDP/Cap) and dropping all rows with missing HDI/year and suicides/100k values for a different OLS. In the latter, my observations went down from 100 to 90, which is still a representative sample. I would have liked to have more accurate/complete data, but I am not worried about the accuracy of the models.

(c) How did you perform data analysis in code, i.e. Briefly describe the workflow of your source code?

After grouping my data, I ran OLS first on complete dataset (without dropping NaNs), regressing ‘GDP/cap’ on ‘suicides/100k population’. I then used the OLS on data devoid of rows with NaNs, regressing both ‘gdp_per_capita’’ and “HDI for year’ on ‘suicides/100k pop’.

(d) What challenges did you encounter and how did you solve them?

The biggest challenge was trying to decide whether or not it’s worth dropping rows with incomplete data. To solve the problem, I did some trial and error and performed OLS on both datasets to see how many observations I was giving up and whether the results were radically different. I arrived at the conclusion that it’s fine to drop some data.

Methods Question 4

(a) How did you manipulate the data to prepare it for analysis?

First, I dropped columns that I didn't need for the clustering analysis ('HDI per year', 'country-year', 'gdp-current-year'). Then I converted all of the categorical socio-demographic variables (sex, age group, generation) to dummies. I then calculated the nominal values for total suicide cases for each category and stored results in a new column. For example, to figure out how many males died in a specific country in a specific year, I took the 'suicides_no' value * 1 (binary for male) and saved the results in a new column 'male'. I repeated the process for females, each generation and each age group. I then dropped the columns with binary variables leaving only columns with numeric values, which is what I needed to do clustering. I then performed a final grouping of this data by country. After clustering, I created three separate tables for each cluster with mean values for each variable, and concocted all three into one table to make comparisons easier.

(b) How did you handle missing, incomplete, or noisy data?

As previously, I came to a conclusion that more data is better than incomplete data especially given the fact that all of it was getting 'collapsed' by country, aggregating numeric values. So, I kept all rows with missing suicide rates, etc. but dipped the column with the greatest number of missing values, namely HDI, as it would cut my data by more than 70% if I were to actually drop the rows. It also didn't make sense to fill in the missing values with zeros as this would bias the aggregated values (mean).
To generate the dummies, I relied on both 'get_dummies' method (for generation and age) and . map method (where male=1 and female=2). I wanted to play around with the two options to see which one was best. I then used python algebra to calculate the number of suicide victims in each level of my three categorical variables and stored results in a newly created column.

(c) How did you perform data analysis in code, i.e. Briefly describe the workflow of your source code?

After analyzing the distribution of suicides, I decided on using three clusters with the Euclidian distance method. I used AgglomerativeClustering to fit the model and predict cluster membership of individual observations. I then used model.labels_ to label each observation with respect to clusters. I then plotted a dendrogram and finally used .describe( ) to look at attributes of each cluster. I focused on mean values, queried the mean table by cluster to create a table of results per cluster and then used .concat to put everything together.

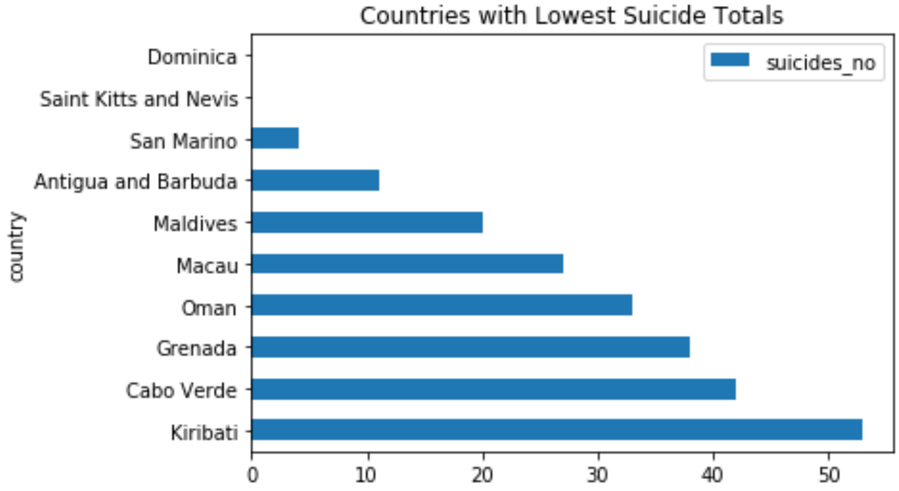
(d) What challenges did you encounter and how did you solve them?

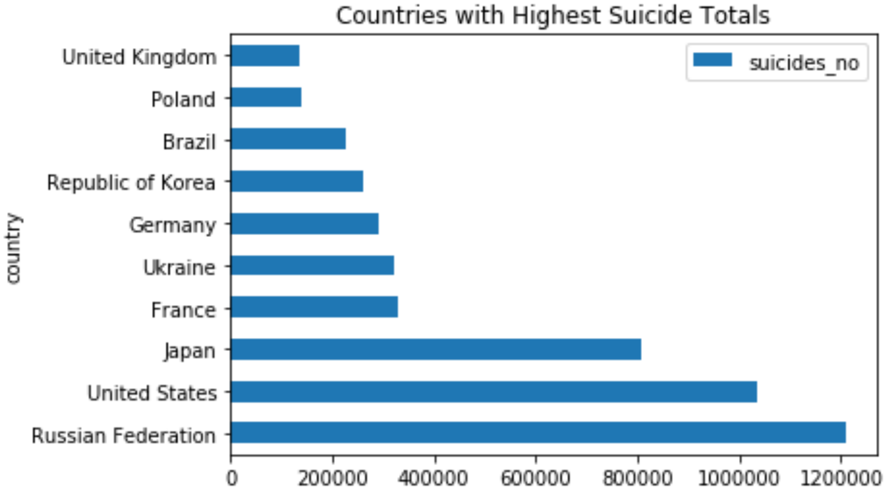
The most challenging part of this question was converting categorical socio-economic variables stored as strings into numeric format that would actually make sense. It took me a while to do this as I initially used groupby method but couldn't get the actual values. I solved the problem by converting them into dummies and manually calculating totals.

Analysis and Results

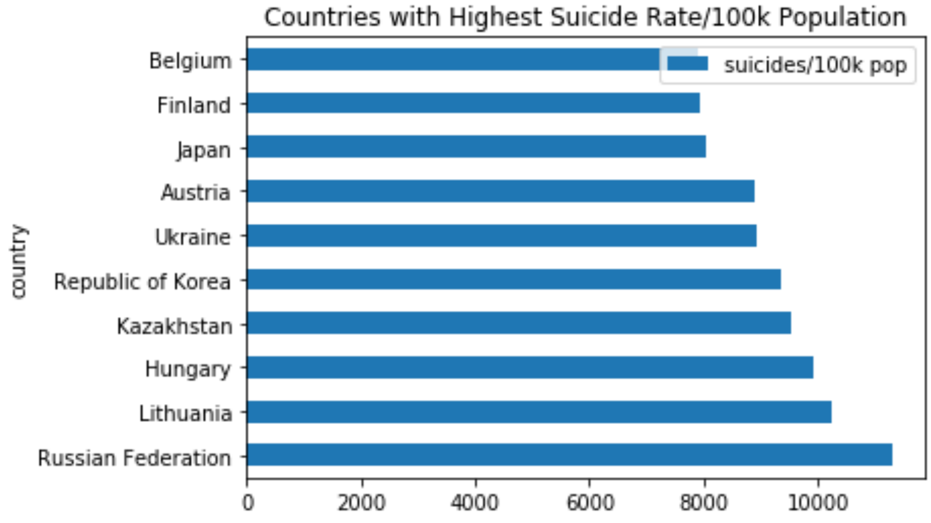
Question 1

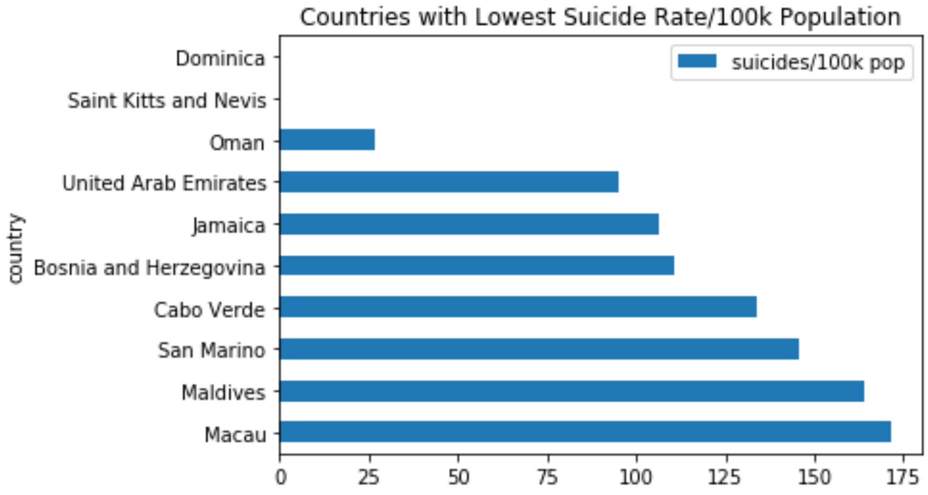
The distribution of total suicides is actually highly skewed to the right, with majority of countries in the dataset having fewer than 100,000 cases over the period from 1985-2015. There are, however, outliers and the range of total suicides is wide – from 0 to around 1,200,000. It was surprising and more importantly encouraging to actually see that a number of countries in the dataset have no recorded/very low number of suicides. Although, it is possible that this is simply due to non-reporting or poor data collection. Of the top ten countries with highest levels of suicide in this time period, US came second, whereas Russian Federation first. United Kingdom, Poland, Brazil, Republic of Korea, Germany, Ukraine, France and Japan are also in this group. Since this analysis reflects the total number of suicides, I would expect higher numbers for countries with bigger populations. It is not surprising then to see US, Brazil, and Russia on the list. Seeing Poland and Ukraine, both of which have small populations, is a bit surprising and definitely prompts questions as to why the two countries have such high numbers (my intuition tells me that there is undeniable connection with the economic downturn following the fall of the Soviet Union). Countries with fewest overall suicides are Dominica, Saint Kitts and Nevis, San Marino, Antigua and Barbuda, Maldives, Macau, Oman, Grenada, Cabo Verde and Kiribati (all small nations and all, with a few exceptions, in warm climates, actually). The range for suicide totals is from 0-50.



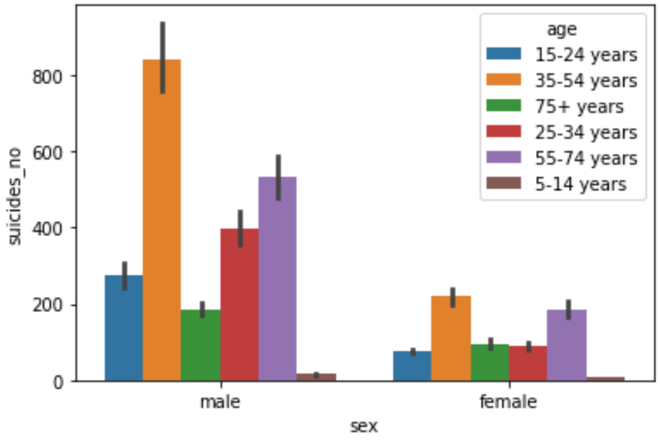


However, totals can only give so much insight into the rate at which people take their own lives as more populous nations might have higher overall numbers but lower rates. When I visualized the rate of suicide per 100K of Population, US actually dropped off the top 10 list, while countries with small populations emerged (Hungary, Lithuania, Belgium).



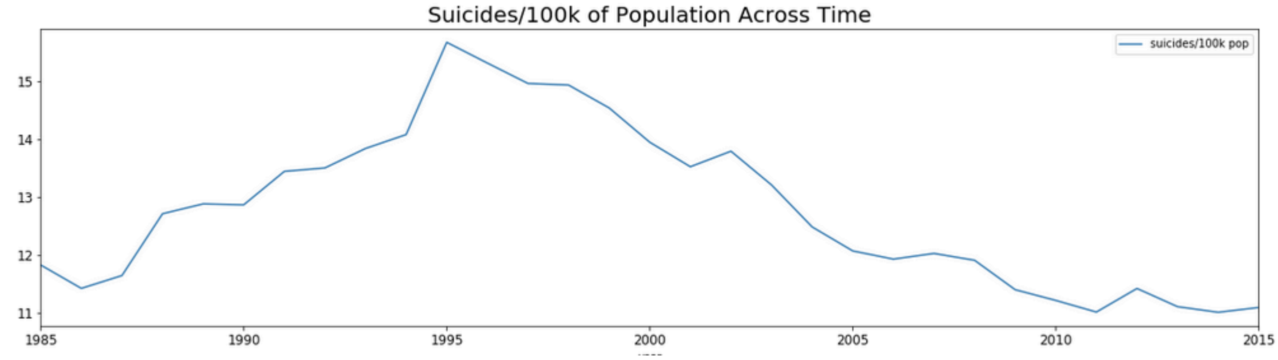


In terms of sex, age group and generation, a disproportionate number of men commit suicide overall compared to women, with the older population (35 and up) making up more than half for each sex. For both sexes, more 35-54 year-olds have committed suicide than 55-74 year-olds.

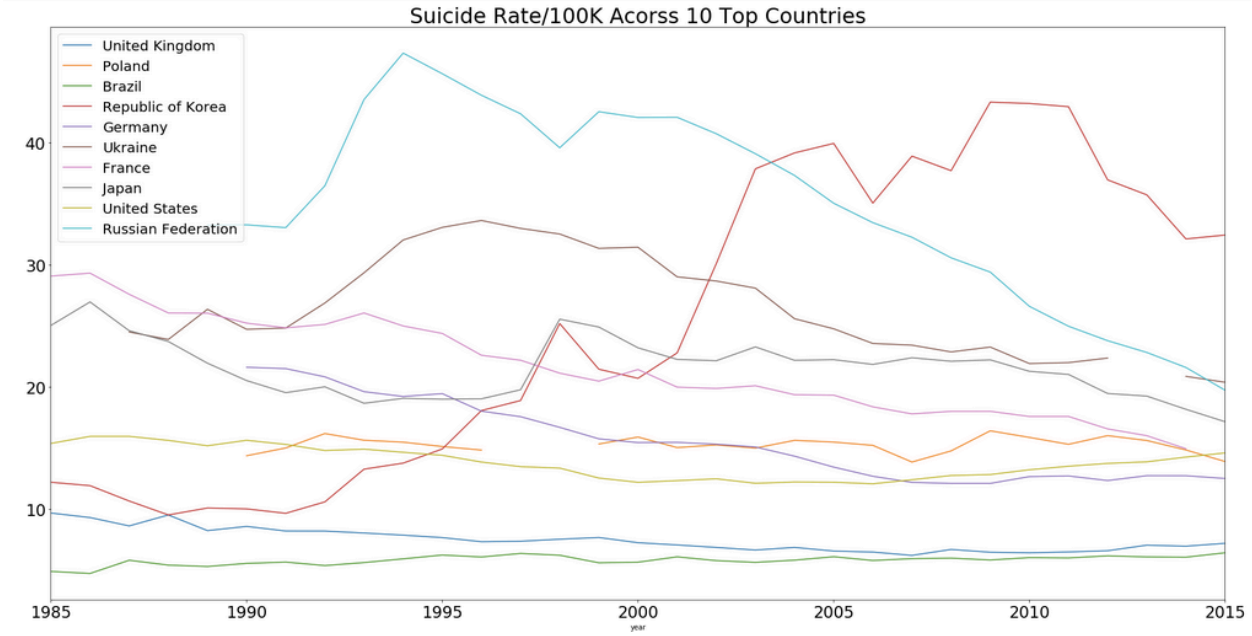


Question 2

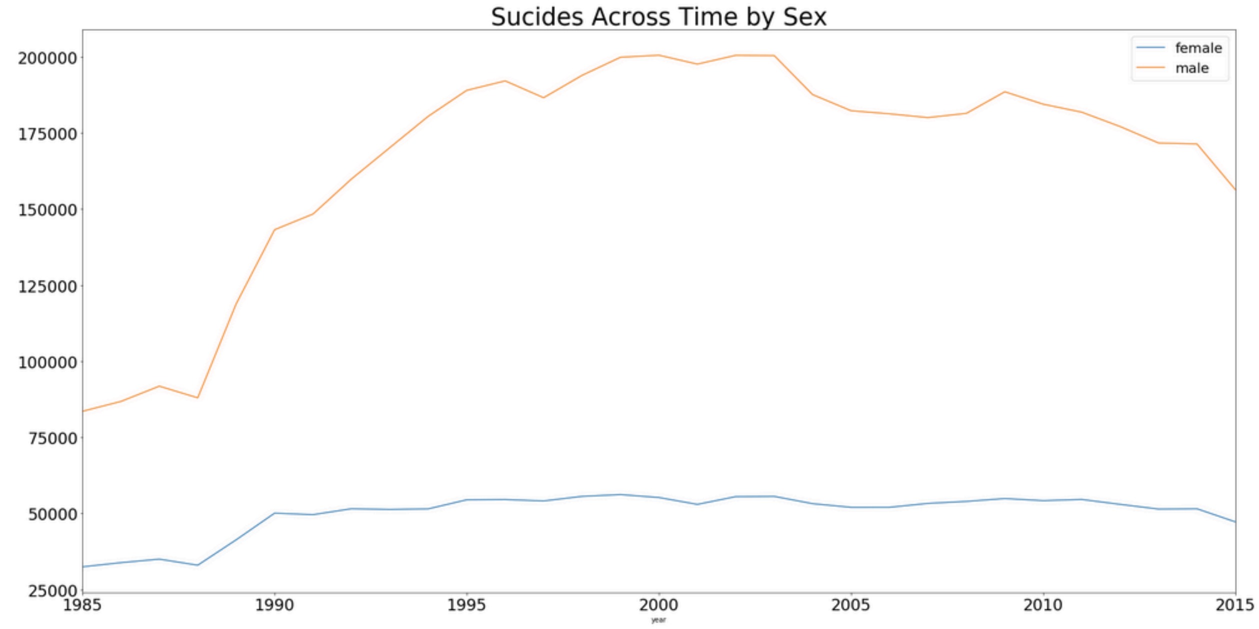
In spite of my assumptions, the rate of suicide (measure per 100k of population) has actually been declining overall starting in 1995, with levels in 2015 lower than in 1985.



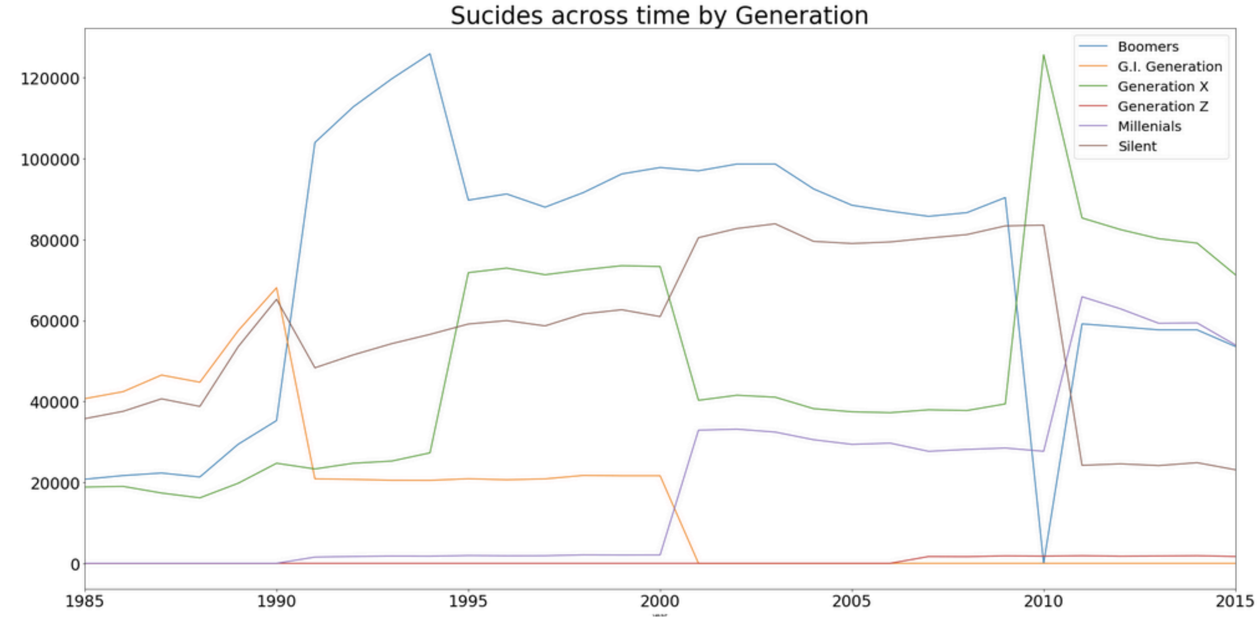
In looking more closely at the top ten countries (with highest suicide totals), it seems that this decreasing trend still holds for the most part. The rates have decreased sharply for the Russian Federation and France. They have remained relatively flat overall for a number of countries. A notable exception is Republic of Korea, where the rates have gone up overall since 1985. The distribution of total suicides across these ten countries over time also varied greatly with Russian Federation having the largest one.



Suicide rates across time by gender, however, reveal that the rate has increased disproportionately for men, peaking around 2004 before declining somewhat. For women, the rates have remained comparatively flat although there was a slight increase overall.

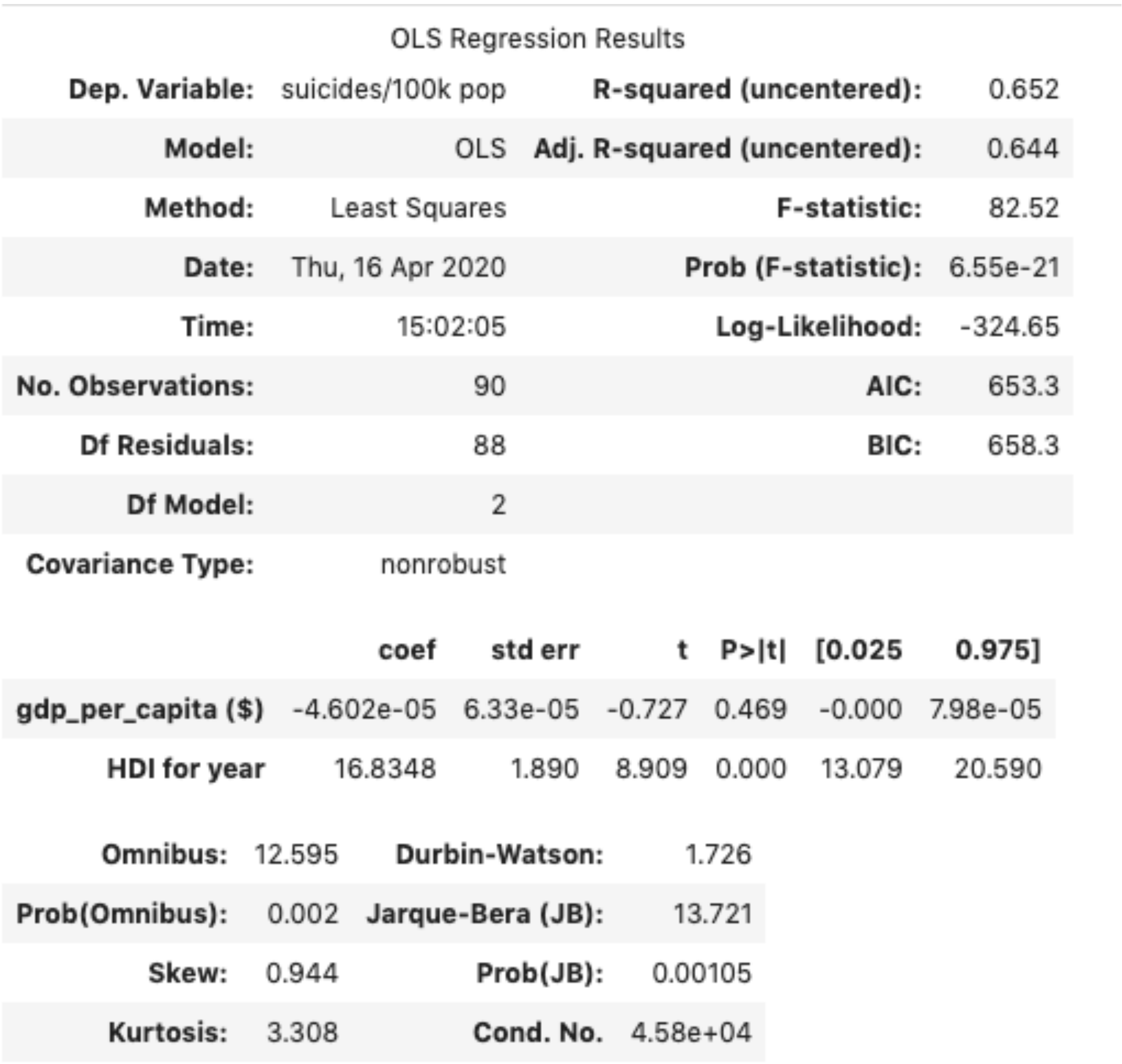


In terms of generations, there is a lot of ups and downs for Boomers and Generation X, although the rates are higher for both in 2015 than in 1985. Millennial suicides are on the rise.



Question 3

After regressing the suicide rate variable on HDI and GDP/capita, I determined that GDP/capita is negatively correlated with suicide rate although the coefficient failed to reach statistical significance at the 5% level of significance. Surprisingly, HDI was positively correlated with the rate of suicide, with on average 1-point increase in HDI/year being associated with 16.8 increase in suicide rate. This was a statistically significant coefficient at the 1% level of significance. This was not the outcome I anticipated, since I assumed that higher human development index (and therefore higher life expectancy, education, and per capita income indicators) would translate into greater economic security/prosperity/healthcare and possibly higher life satisfaction. But the model outcome is pointing in the opposite direction – the higher the human development the higher suicide rates are. This does make me question the validity of the model. For one, HDI might already include GDP/Capita which might be messing with the results. And two, there are so many variables not measured in my data that might give more accurate results (i.e. it's very likely this model is biased).



OLS Regression Results

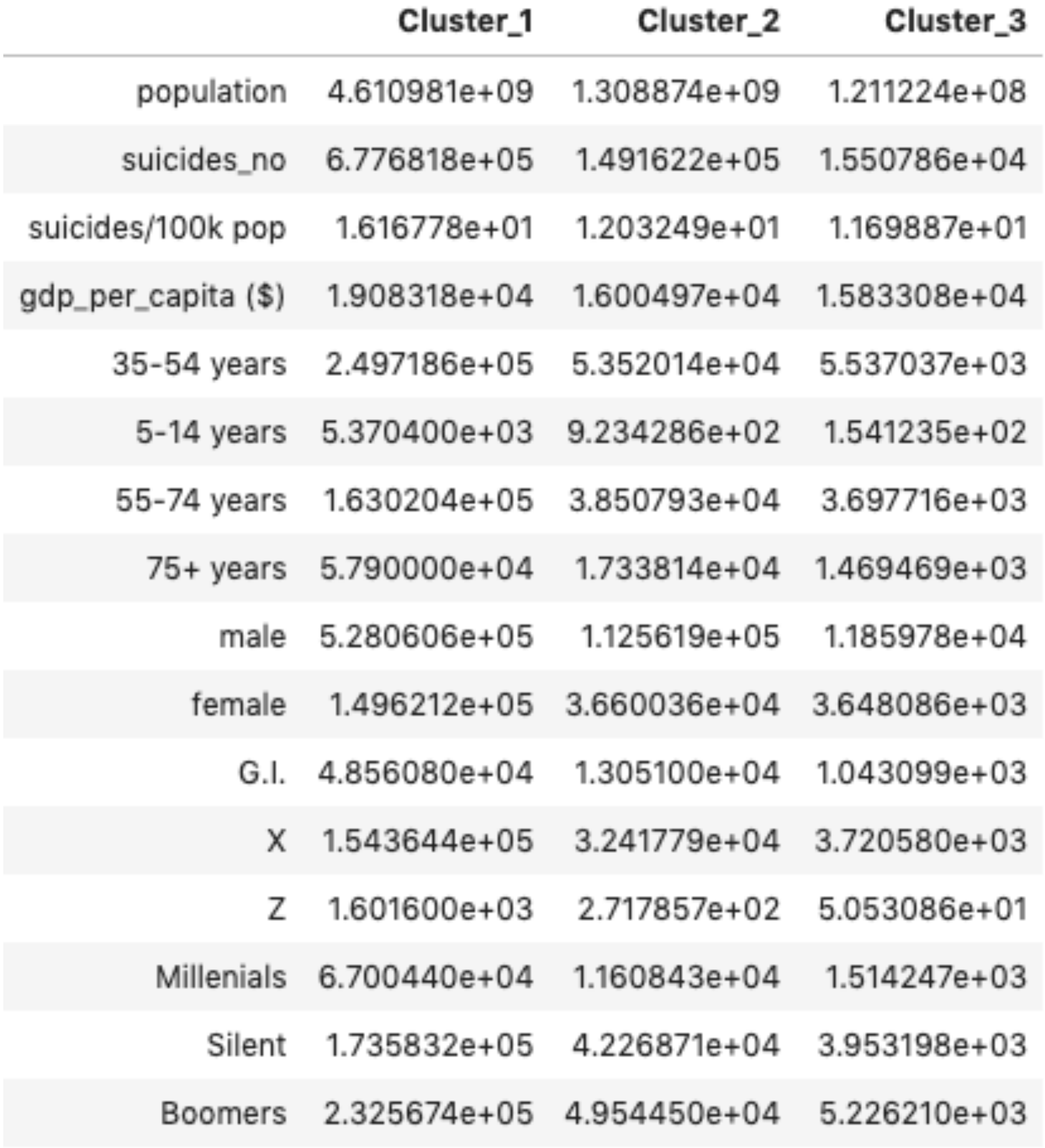
Dep. Variable:	suicides/100k pop	**R-squared (uncentered):**	0.652
Model:	OLS	**Adj. R-squared (uncentered):**	0.644
Method:	Least Squares	**F-statistic:**	82.52
Date:	Thu, 16 Apr 2020	**Prob (F-statistic):**	6.55e-21
Time:	15:02:05	**Log-Likelihood:**	-324.65
No. Observations:	90	**AIC:**	653.3
Df Residuals:	88	**BIC:**	658.3
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
gdp_per_capita ($)	-4.602e-05	6.33e-05	-0.727	0.469	-0.000	7.98e-05
HDI for year	16.8348	1.890	8.909	0.000	13.079	20.590

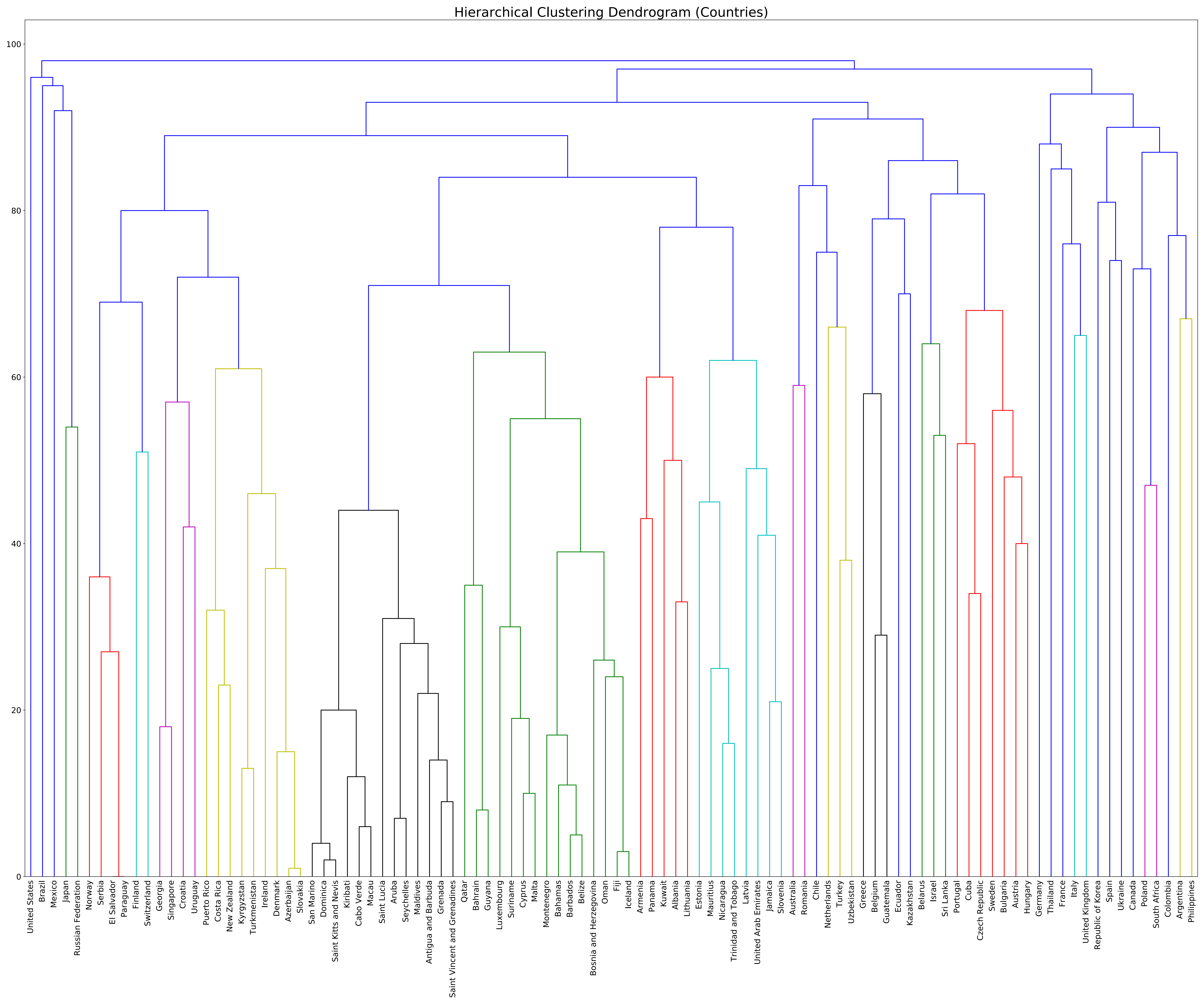
Omnibus:	12.595	**Durbin-Watson:**	1.726
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	13.721
Skew:	0.944	**Prob(JB):**	0.00105
Kurtosis:	3.308	**Cond. No.**	4.58e+04

Question 4

After clustering the countries into three clusters, I obtained both the dendrogram (see last page) and the descriptive statistics of each cluster. Based on the dendrogram, US, Brazil and Mexico were closest to one another (but didn't get grouped really). Republic of Korea, Ukraine, Canada and Spain were grouped together as well. Generally, though, I think it's more informative to look at the descriptive statistics of each cluster than the dendrogram. The first cluster (0 in the output, but labeled 1 in the table below) has only 5 countries, compared with the second (14 clusters) and the third (81 countries). It also has the highest average population, the greatest number of suicide totals, highest suicide rate, and the greatest GDP/Capita. Brazil, Japan, Mexico, Russian Federation and US make up this cluster, which is somewhat in-line with the previous analysis. The following countries are in the second cluster: Argentina, Canada, Colombia, France, Germany, Italy, Philippines, Poland, Republic of Korea, South Africa, Spain, Thailand, Ukraine, and United Kingdom. The rest of the countries in the dataset are in the third cluster. The third cluster has the lowest average values for non-demographic variables. Across all three clusters, the greatest average number of suicides is in the 35-54 age group, with highest numbers in the first cluster. Boomers generation is the most represented in across all three clusters. Overall, I think the results are in-tune with the analysis performed earlier.



	Cluster_1	Cluster_2	Cluster_3
population	4.610981e+09	1.308874e+09	1.211224e+08
suicides_no	6.776818e+05	1.491622e+05	1.550786e+04
suicides/100k pop	1.616778e+01	1.203249e+01	1.169887e+01
gdp_per_capita ($)	1.908318e+04	1.600497e+04	1.583308e+04
35-54 years	2.497186e+05	5.352014e+04	5.537037e+03
5-14 years	5.370400e+03	9.234286e+02	1.541235e+02
55-74 years	1.630204e+05	3.850793e+04	3.697716e+03
75+ years	5.790000e+04	1.733814e+04	1.469469e+03
male	5.280606e+05	1.125619e+05	1.185978e+04
female	1.496212e+05	3.660036e+04	3.648086e+03
G.I.	4.856080e+04	1.305100e+04	1.043099e+03
X	1.543644e+05	3.241779e+04	3.720580e+03
Z	1.601600e+03	2.717857e+02	5.053086e+01
Millenials	6.700440e+04	1.160843e+04	1.514247e+03
Silent	1.735832e+05	4.226871e+04	3.953198e+03
Boomers	2.325674e+05	4.954450e+04	5.226210e+03


Hierarchical Clustering Dendrogram (Countries)
100
80
60
40
20
0
United States
Brazil
Mexico
Japan
Russian Federation
Norway
Serbia
El Salvador
Paraguay
Finland
Switzerland
Georgia
Singapore
Croatia
Uruguay
Puerto Rico
Costa Rica
New Zealand
Kyrgyzstan
Turkmenistan
Ireland
Denmark
Azerbaijan
Slovakia
San Marino
Dominica
Saint Kitts and Nevis
Kiribati
Cabo Verde
Macau
Saint Lucia
Aruba
Seychelles
Maldives
Antigua and Barbuda
Grenada
Saint Vincent and Grenadines
Qatar
Bahrain
Guyana
Luxembourg
Suriname
Cyprus
Malta
Montenegro
Bahamas
Barbados
Belize
Bosnia and Herzegovina
Oman
Fiji
Iceland
Armenia
Panama
Kuwait
Albania
Lithuania
Estonia
Mauritius
Nicaragua
Trinidad and Tobago
Latvia
United Arab Emirates
Jamaica
Slovenia
Australia
Romania
Chile
Netherlands
Turkey
Uzbekistan
Greece
Belgium
Guatemala
Ecuador
Kazakhstan
Belarus
Israel
Sri Lanka
Portugal
Cuba
Czech Republic
Sweden
Bulgaria
Austria
Hungary
Germany
Thailand
France
Italy
United Kingdom
Republic of Korea
Spain
Ukraine
Canada
Poland
South Africa
Colombia
Argentina
Philippines