Exploratory Data Analysis

Ishan Ambike

1/28/2022

Introduction:

In this problem, I will be working on Exploratory Data Analysis. Exploratory Data Analysis (EDA) is an initial investigation to analyze the data. It can be used to discover trends or patterns in the data. It can also be used to spot anomalies and test hypotheses. Statistical calculations and graphical visualizations are used in EDA. For this problem, EDA will be performed on a chosen dataset. I will then present the information found out using EDA in detail.

The Dataset:

The dataset I am going to work on is available on Kaggle. It can be viewed here. The name of the dataset is “Suicide Rates Overview 1985 to 2016”. This dataset gives information about suicides and suicide rates along with socio-economy information from 1985 to 2016.

The Below table describes each column in the dataset.

Column Name	Description
Country	Name of the country
Year	Year in YYYY format
Sex	Male or Female
Age	Mentions the age group
Suicides_no	Number of suicides for the demographic
Population	Population for the demographic
Suicides/100k pop	Number of suicides per 100k population of the demographic
country-year	Country-Year combination
HDI for year	Human Development Index for the year

gdp_for_year ($)	Gross domestic product for the year in USD
gdp_per_capita ($)	Gross domestic product per capita in USD
Generation	Generation of the demographic as per birth

Analysis from EDA:

What information is available in the given dataset?

The dataset contains 27820 rows and 12 columns. Below is the list of columns present in the dataset.

```
Index(['country', 'year', 'sex', 'age', 'suicides_no', 'population',
       'suicides/100k pop', 'country-year', 'HDI for year',
       ' gdp_for_year ($) ', 'gdp_per_capita ($)', 'generation'],
      dtype='object')
```

The dataset basically gives details about suicides committed from 1985 to 2016. It provides demographic information like country, age, generation, and sex. It also gives stats like population, HDI, GDP for each year, and GDP per capita.
Below is a snapshot of a brief description of numerical data present in the dataset.

	year	suicides_no	population	suicides/100k pop	\
count	27820.000000	27820.000000	2.782000e+04	27820.000000	
mean	2001.258375	242.574407	1.844794e+06	12.816097	
std	8.469055	902.047917	3.911779e+06	18.961511	
min	1985.000000	0.000000	2.780000e+02	0.000000	
25%	1995.000000	3.000000	9.749850e+04	0.920000	
50%	2002.000000	25.000000	4.301500e+05	5.990000	
75%	2008.000000	131.000000	1.486143e+06	16.620000	
max	2016.000000	22338.000000	4.380521e+07	224.970000	

	HDI for year	gdp_for_year ($)	gdp_per_capita ($)
count	8364.000000	2.782000e+04	27820.000000
mean	0.776601	4.455810e+11	16866.464414
std	0.093367	1.453610e+12	18887.576472
min	0.483000	4.691962e+07	251.000000
25%	0.713000	8.985353e+09	3447.000000
50%	0.779000	4.811469e+10	9372.000000
75%	0.855000	2.602024e+11	24874.000000
max	0.944000	1.812071e+13	126352.000000

Which data types/categories of data are there in the dataset?

```
country               object
year                   int64
sex                   object
age                   object
suicides_no            int64
population             int64
suicides/100k pop    float64
country-year          object
HDI for year         float64
 gdp_for_year ($)       int64
gdp_per_capita ($)      int64
generation            object
dtype: object
```

The dataset contains integer, float, and string types of data. There are categorical data like country, sex, age, country-year, and generation. The dataset contains numerical data such as year, number of suicides, population, suicides/100k population, HDI for year, GDP for year($), and GDP per capita($).

How many unique values are there in each variable?

The below snapshot shows each variable and the unique values it contains.

```
country                101
year                    32
sex                      2
age                      6
suicides_no           2084
population           25564
suicides/100k pop     5298
country-year          2321
HDI for year           305
 gdp_for_year ($)     2321
gdp_per_capita ($)    2233
generation               6
```

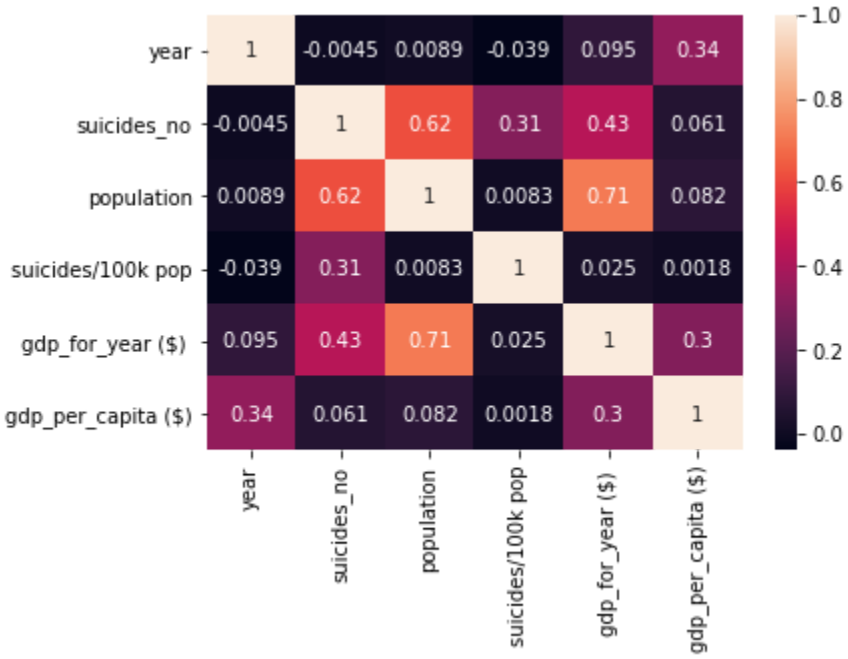
So there is data for 101 countries, 32 years, and 6 generations of people.

Are there missing values in the dataset?

```
country                   0
year                      0
sex                       0
age                       0
suicides_no               0
population                0
suicides/100k pop         0
country-year              0
HDI for year          19456
 gdp_for_year ($)         0
gdp_per_capita ($)        0
generation                0
```

As seen from the data above only HDI for year column has missing values. There are no missing values for other variables.
After calculating, we get that 69.93% (~ 70%) values are missing in HDI for year column. Since the majority of data is missing for this variable, it can be dropped from further analysis.

What is the correlation between the variables?

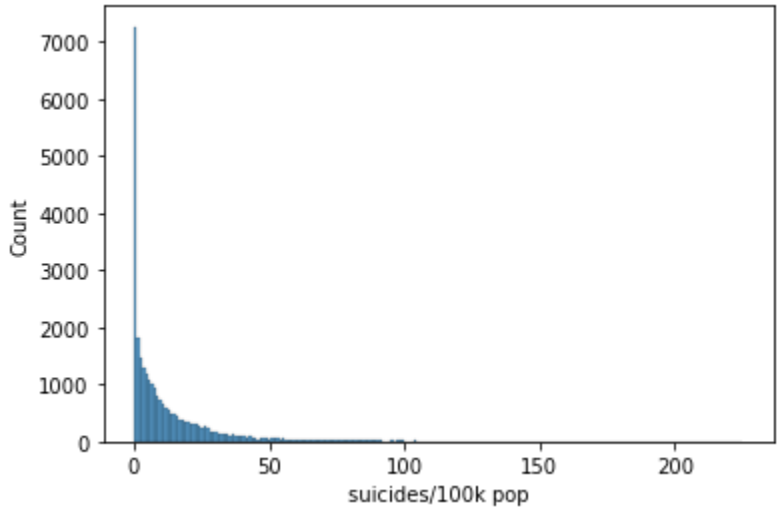


The above heatmap shows the correlation between all the numeric variables. There is a strong correlation between the number of suicides and the population. It is expected as the number of suicides would be more in a country with a higher population. This is also evident from the positive correlation between the number of suicides and suicides/100k population. Year variable has a negative correlation with the number of suicides and suicides/100k population. Which would indicate that as years are passing suicides are decreasing. The year variable has a positive correlation with GDP per capita which indicates GDP per capita is increasing with the years. The Strong correlation between GDP for year and population could be because with more population there is more workforce which improves GDP.

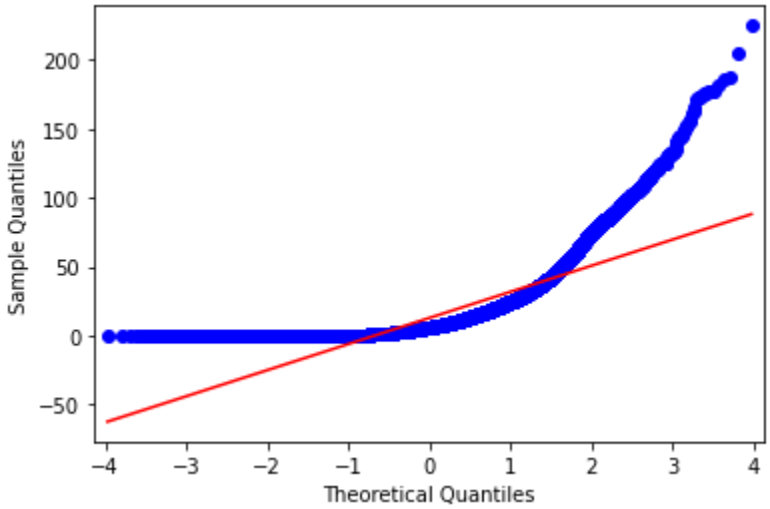
Is suicides/100k population normally distributed?

Let’s see distribution using plots.

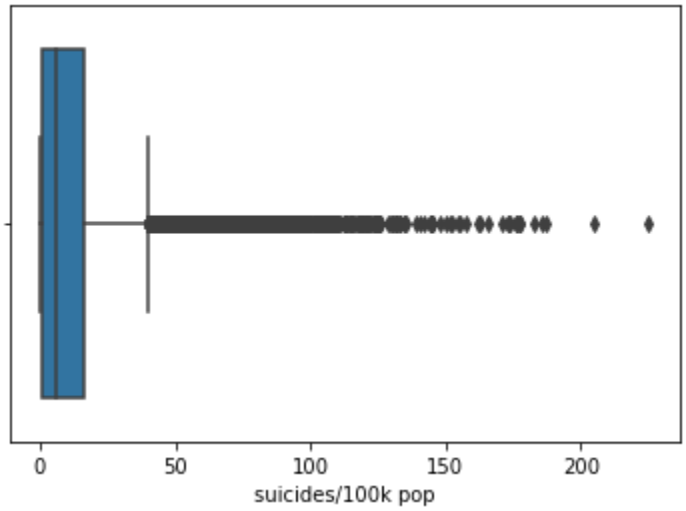
Histogram:



Q-Q Plot:

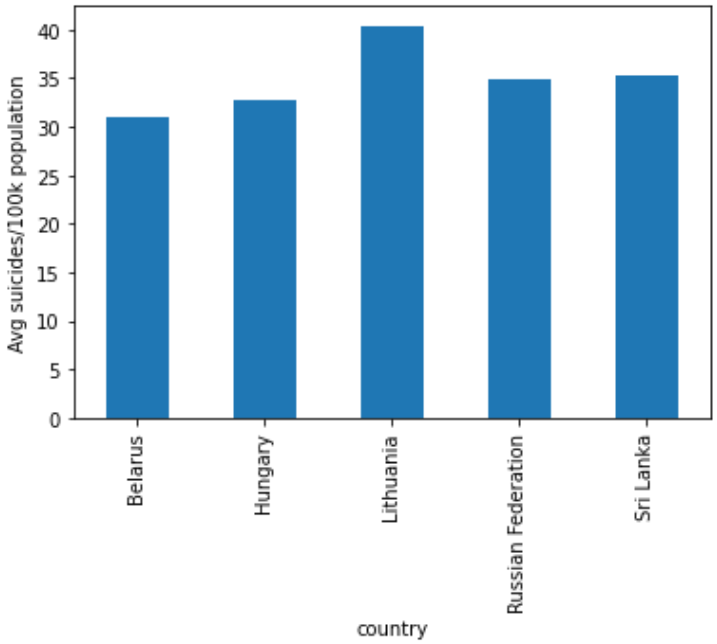


Box Plot:



So from the plots, it is evident that suicides/100k population is not normally distributed. It is right(positively) skewed.
This can also be confirmed by numerical data. Suicides/100k population has Mean (12.81) > Median (5.99) > Mode (0).

Which countries have the highest average suicides/100k population?

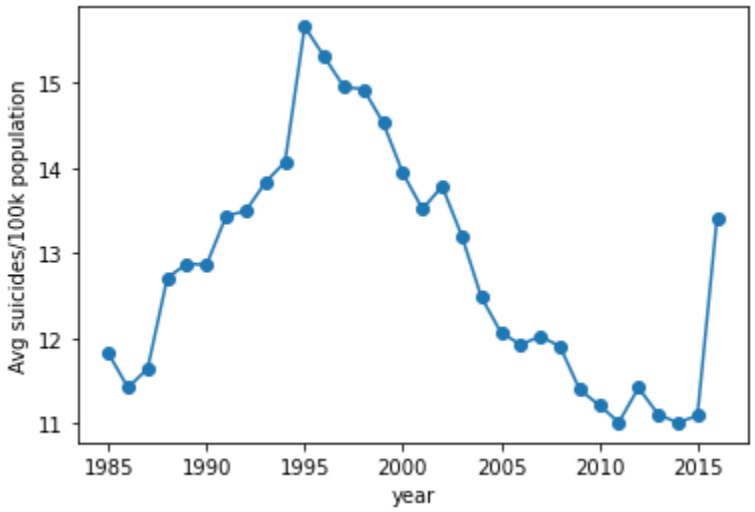


country	
Lithuania	40.415573
Sri Lanka	35.295152
Russian Federation	34.892377
Hungary	32.761516
Belarus	31.075913

Lithuania has the highest average suicides/100k population followed by Sri Lanka, Russian Federation, Hungary, and Belarus respectively.

What is the trend of suicides/100k population over the years?

Below is a time series plot showing average suicides/100k population over the years.

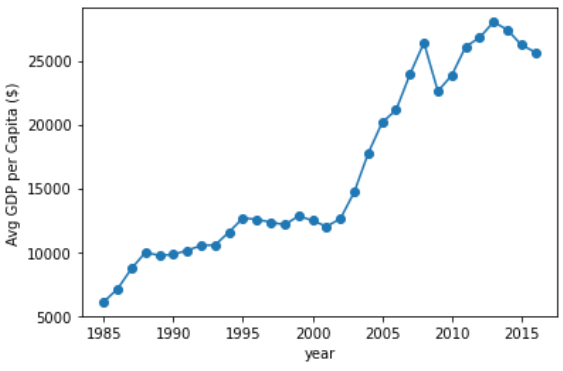


Years with the highest average suicides/100k population:

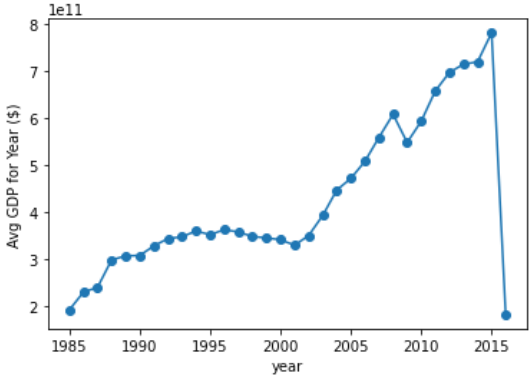
year	
1995	15.662671
1996	15.305422
1997	14.954361
1998	14.926920
1999	14.532038
1994	14.073272
2000	13.941328
1993	13.833705
2002	13.786550
2001	13.519138
1992	13.498564
1991	13.438880
2016	13.421188
2003	13.205019
1989	12.879071
1990	12.862956

1995 is the year with the highest average suicides/100k population. We can observe that period of 1990-2000 had a high suicide rate. Let's see if we can find the reason for this from the data available.
Let's check if the economy is the reason for high suicides during this period. Below are the time series plots for GDP per capita, GDP for year, and population.

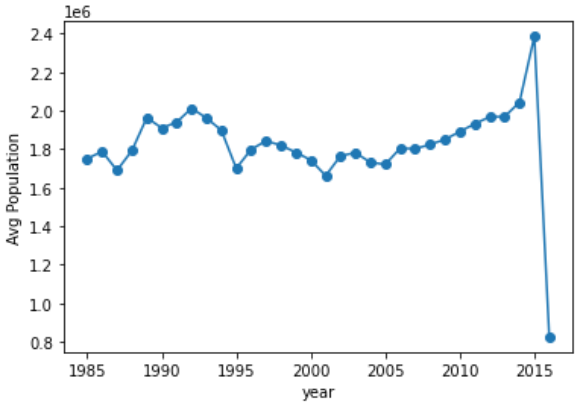
Avg GDP per capita v Year



Avg GDP for year v Year



Avg Population v Year



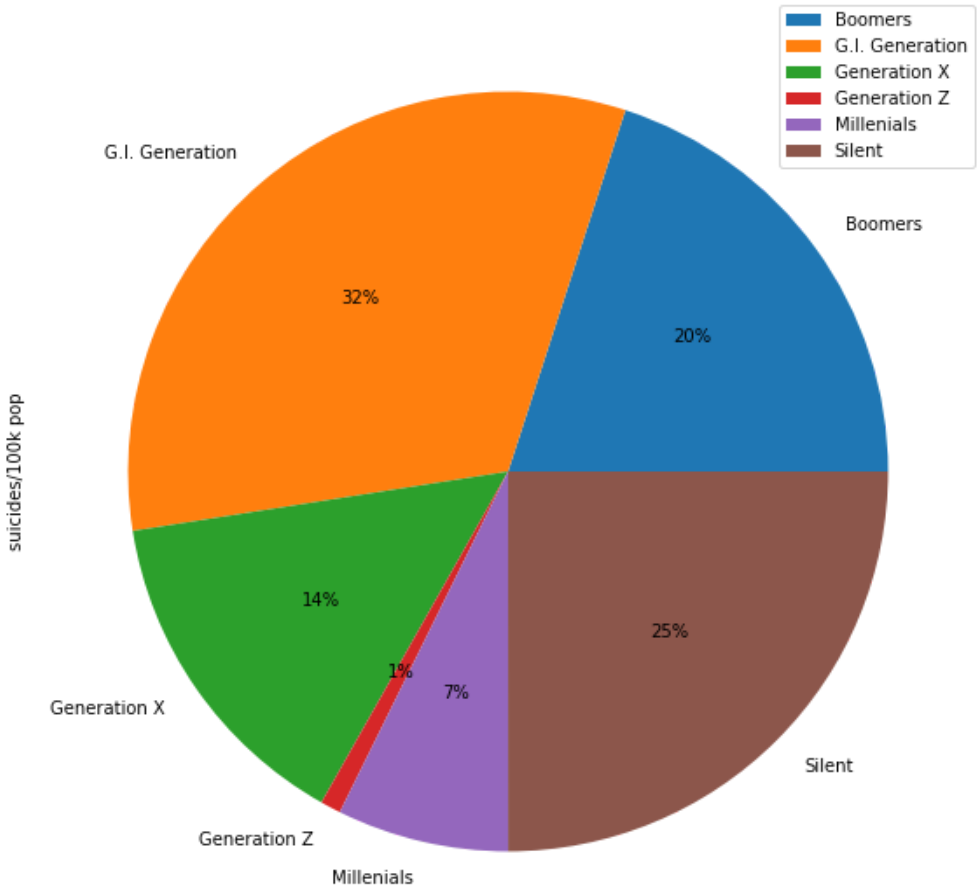
So from these graphs, we can see that GDP per capita and GDP for years 1990-2000 were lower than years post-2000. So economy might be the reason for higher suicide rates during this period. We can also see population dip around the year 1993-1998 which might be due to suicides.

What patterns are seen in the generation variable?

Let's find out the average suicide rate for each generation to answer this question.

Avg Suicides for each generation per 100k population:

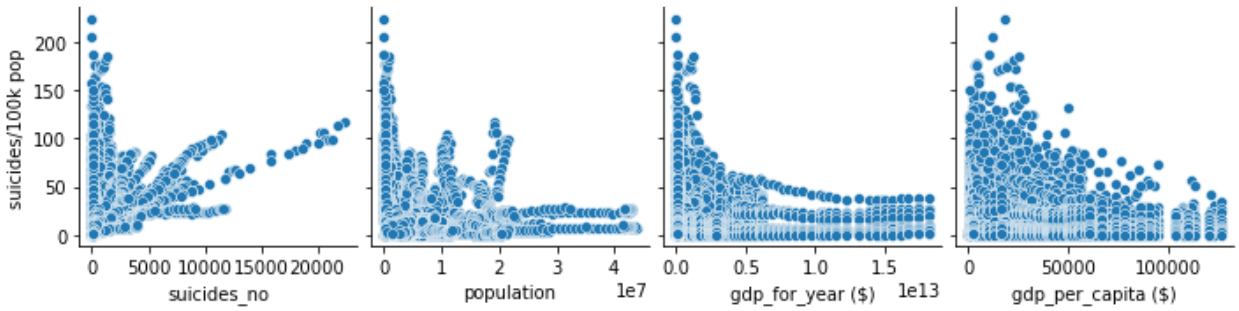
generation	
G.I. Generation	23.946378
Silent	18.418848
Boomers	14.742094
Generation X	10.556874
Millenials	5.383597
Generation Z	0.642299



So from the table and pie chart, we can clearly see that G. I. Generation has the highest proportion in the suicide rate followed by Silent, Boomers, Gen X, and Millenials. Gen Z has the least suicide rate.

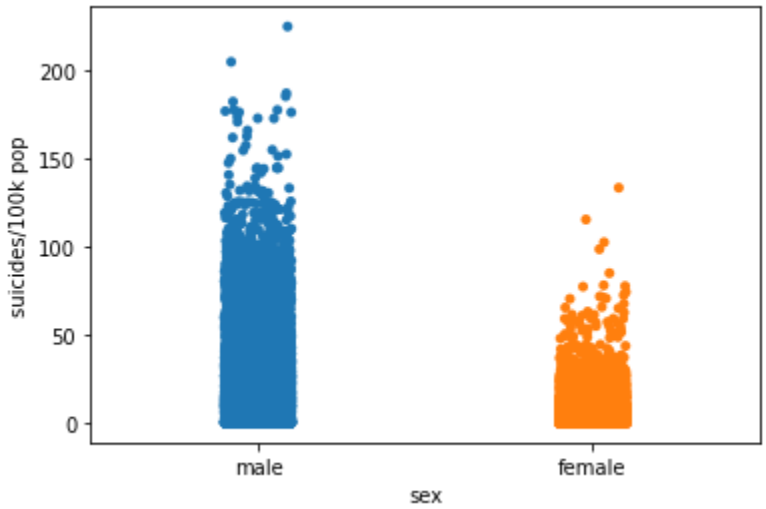
The reasons for higher suicide rates in G. I. Generation could be World War 2 and the great depression. The people of this generation had to participate in WW2 and had to live during the economic depression. The silent generation could have been affected by the after-effects of economic depression and the Korean war. Gen Z has lived in a time with economic and war stability and has the least suicide rate. Hence it seems that wars and the economy have played role in the suicide rate.

How do other numerical variables have a relation with suicides/100k population?

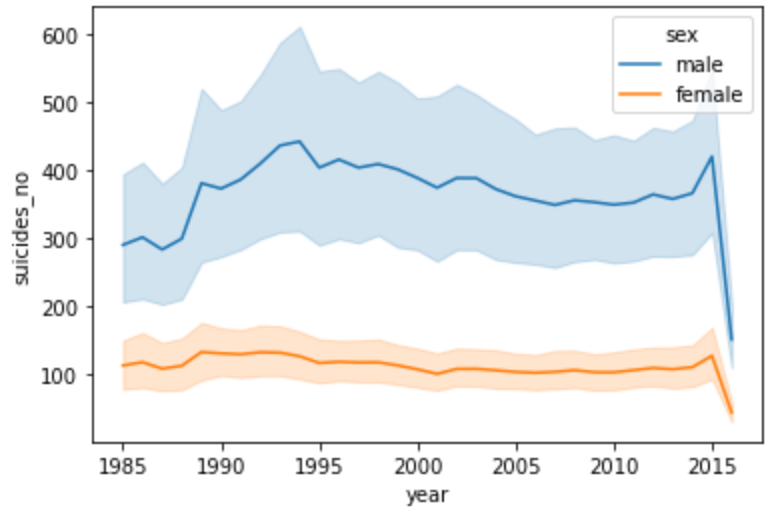


There does not seem to be a pattern between the number of suicides and suicides/100k population. But we can see that for a higher suicide number, the suicides/100k population is also more(increasing). In the second plot we can conclude that as population increases, suicides/100k population decreases. From the third and fourth graphs also we can notice that as GDP for year or GDP per capita increases, suicides/100k population decreases. Hence it seems that better economic conditions result in a lesser suicidal rate.

What pattern is observed in sex?

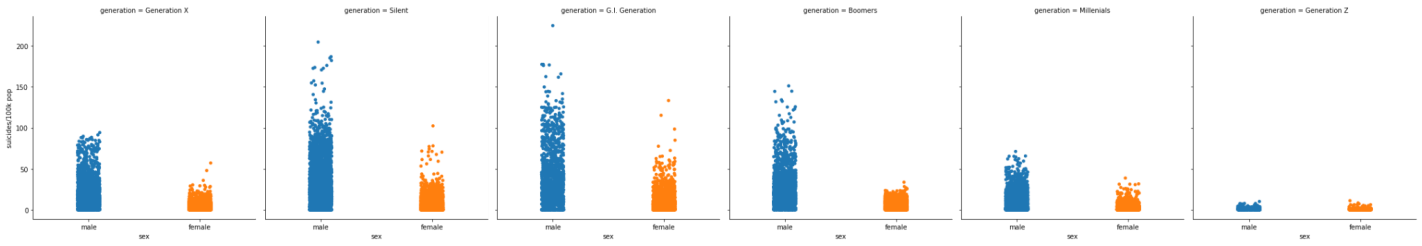


From the above strip plot, we can see that males have a higher average suicides/100k population than females.



The above plot shows the number of suicides for males and females individually from 1985 to 2016. We can see that males have a higher number of suicides throughout these years.

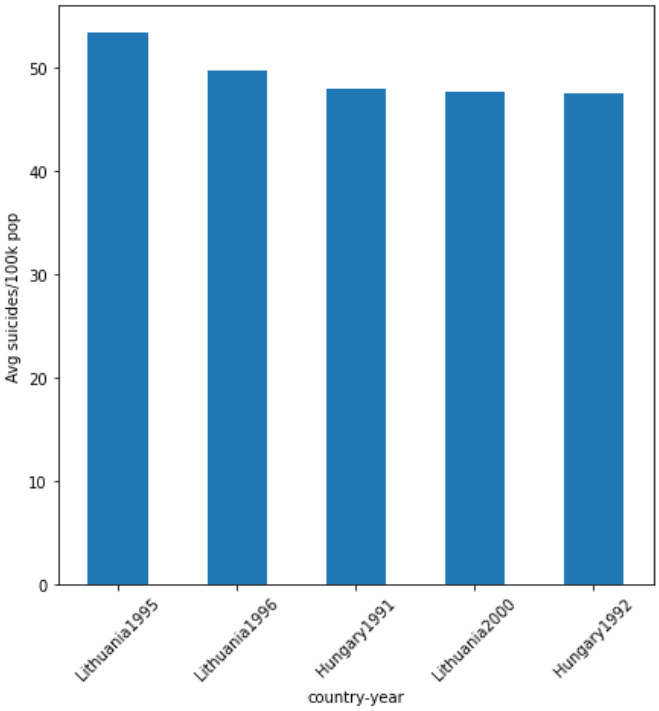
Is the pattern observed in sex consistent across all the generations?



Across all the generations we can see that males have had a higher average suicides/100k population. But in later generations, we see that the difference is reducing. In Gen Z there seems almost no difference. This might be because females also started taking part in wars and jobs. In earlier generations as only males used to work and serve in the military, significantly higher suicides were observed in males.

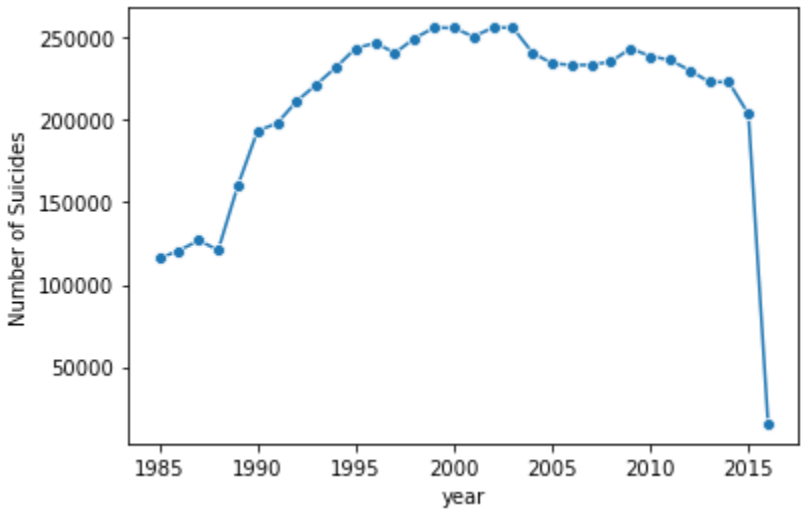
Which country-year has the highest average suicides/100k population?

```
country-year
Lithuania1995    53.275000
Lithuania1996    49.634167
Hungary1991      47.916667
Lithuania2000    47.650000
Hungary1992      47.521667
```



Lithuania in 1995 has the highest suicide rate from 1985 to 2016 in the country-year category. It seems that Lithuania and Hungary had high average suicides/100k population in the 1990 to 2000 period.

Which year has the most number of suicides?



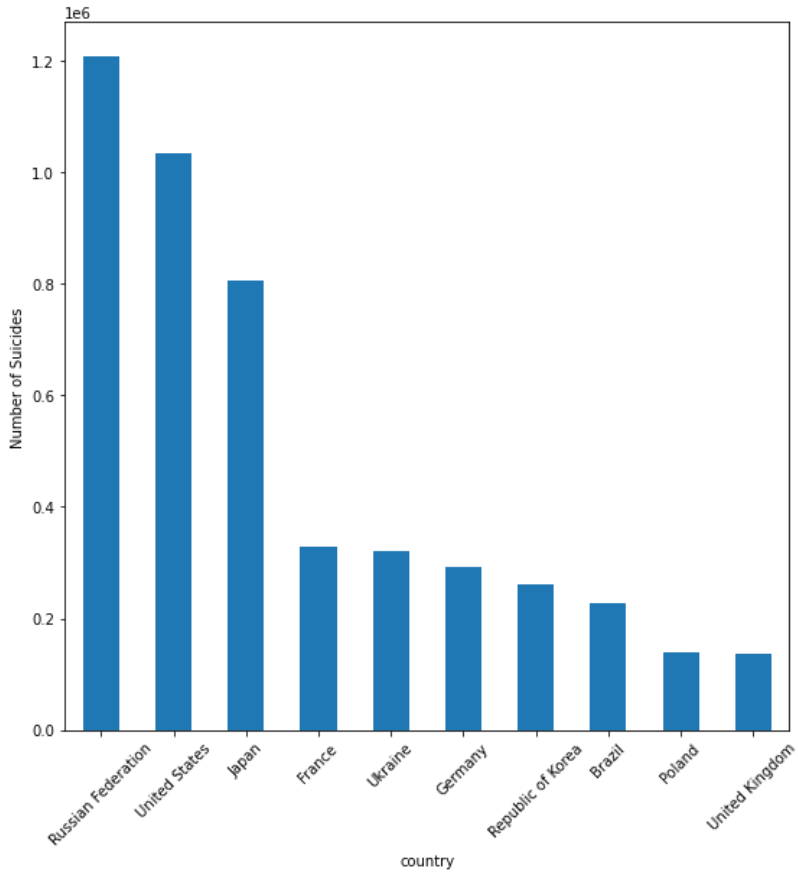
Above is the time series plot for Number of Suicides v Year.

The year 1999 has the most number of suicides with 256,119 suicides. It is followed by 2002, 2003, 2000, 2001, and 1998 respectively. Hence it seems that 1998-2003 had lots of suicides.

Which countries have the highest and least number of suicides?

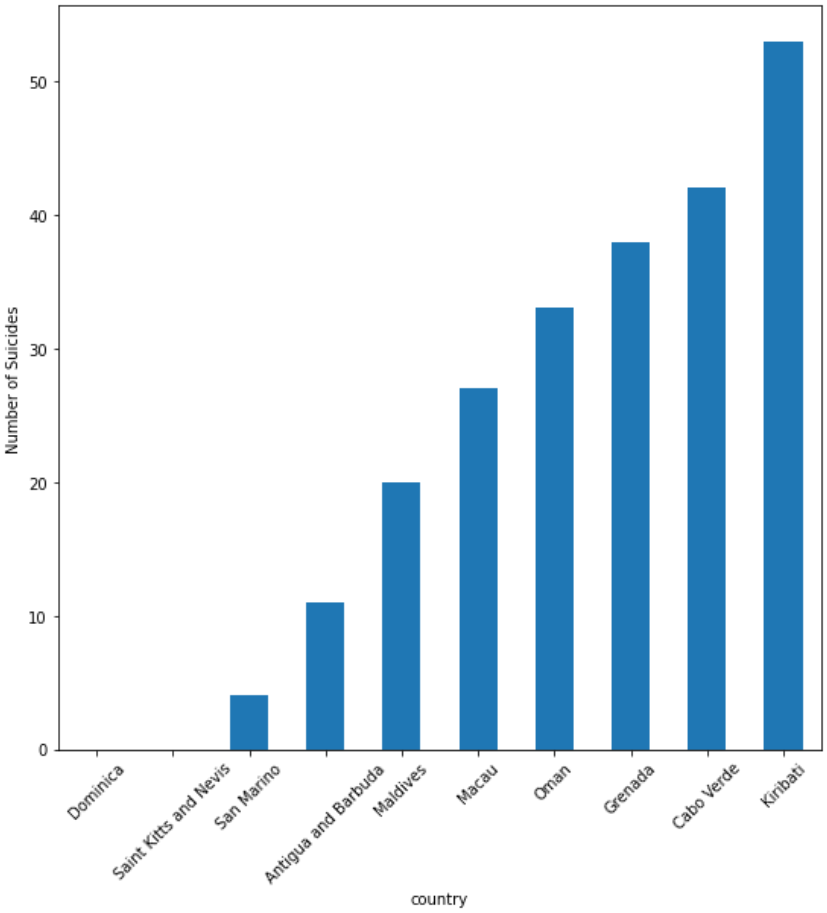
List of countries with the highest suicides:

country	
Russian Federation	1209742
United States	1034013
Japan	806902
France	329127
Ukraine	319950
Germany	291262
Republic of Korea	261730
Brazil	226613
Poland	139098
United Kingdom	136805



List of countries with the least suicides:

country	
Dominica	0
Saint Kitts and Nevis	0
San Marino	4
Antigua and Barbuda	11
Maldives	20
Macau	27
Oman	33
Grenada	38
Cabo Verde	42
Kiribati	53



So we can see there is a big contrast between the number of suicides from the two graphs. There are countries like Russian Federation and the USA with greater than 1 million suicides, while there are also countries with 0 suicides.
The general pattern observed from this is that prominent countries with large populations and big economies have higher suicides. Small countries have fewer suicides. Many of these countries are small island nations.

Conclusion:

Successfully performed EDA on the worldwide suicides dataset. Could Discover some interesting patterns. To analyze the data, various types of visualizations were used. Found information like country and year with most and least suicides. Males have more suicides than females. Suicides might have socio-economic reasons like the economy and wars. Hence some interesting insights were discovered in this EDA activity.

Future Scope:

Further predictive analysis can be done on the dataset to predict the likelihood of suicides. The past number of suicides, economic conditions, gender, and age group can be used for prediction. This can be used for preventive measures. A country/state can figure out the preventive measures it should take to reduce suicides.