**Name: P.Varshith Reddy**

**Reg. No.: 12017306**

**Section: K20MD**

**Roll No.: RK20MDB69**

**Suicide Rates Overview 1985 to 2016.**

# Description

In this analysis I will be focusing at investigating factors that could influence suicide rates among different user cohorts and countries.

The dataset includes info on suicide rates among different cohorts together with the country, region and country economical status. 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.

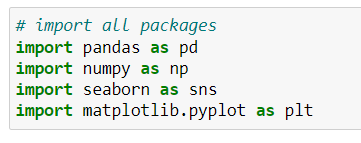
# Dataset Structure

#### The dataset contains below variables:

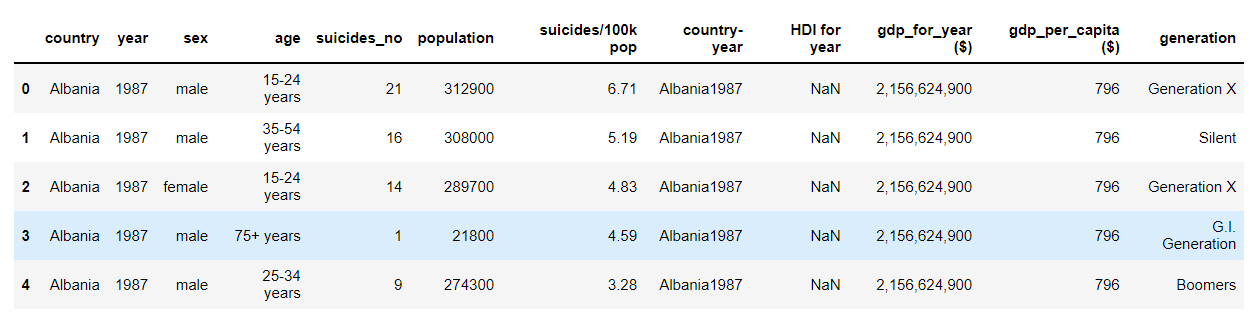
* country - country to where the statistics belong.
* year - year from where the statistics was taken.
* sex - cohort sex.
* age - cohort age group.
* suicides\_no - the total number of sucides in the country for the given cohort.
* population - the size of the cohort for the given year.
* suicides/100k pop - suicides per 100k for this particular cohort.
* contry-year - country-year pair.
* HDI for year - Human Development index for the given year.
* gdp\_for\_year()−𝐶𝑜𝑢𝑛𝑡𝑟𝑦𝐺𝐷𝑃.
* gdppercapita() - Country GDP per household.
* generation - social generation to which the current generation belong.

**Import Data and Required Packages**

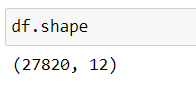
Importing Pandas, Numpy, Matplotlib and Seaborn



**Show Top 5 Records**



**Shape of the dataset**

****

## **Exploratory Data Analysis**:

Creating a new variable, suicide rate, that contains the percentage of suicides for the countries, this might help a little bit in dealing with countries with extreme populations.

I’ll be using ggplot to generate a few visualizations in order to get a better understanding of the underlying patterns in the data set. Answering my research question, suicides among the different age groups.

Very notable spike right at the beginning of the histogram for the youngest age group, this corresponds to the fact that lots of countries have low suicide rates among the youth.

Notably, the 75+ age group appears to have a slightly more significant tail as compared to the other age groups, indicating the fact that some countries have high suicide rates when it comes to the older age groups.

## **Relationship between Suicides and GDP per Capita:**

When it comes to inspecting suicides globally, a key question that comes to mind is whether third-world countries tend to have more suicides, corresponding to a lower standards of living, or if first-world countries also have significant numbers of suicides.

I am looking at the average suicides for each country to see the relationship with low standards of living.

The United States and Japan have the highest average number of suicides, in stark contrast to their high standards of living.

#### **What features in the dataset will help support the investigation?**

The primary focus for the analysis will be around universal suicide ratio of the number of suicides per 100k population.

This metric is taken as a centralised object of analysis as it is independent from population size and therefore will be more objective in representing general trends.

# Data cleaning:

# 

### **Insights:**

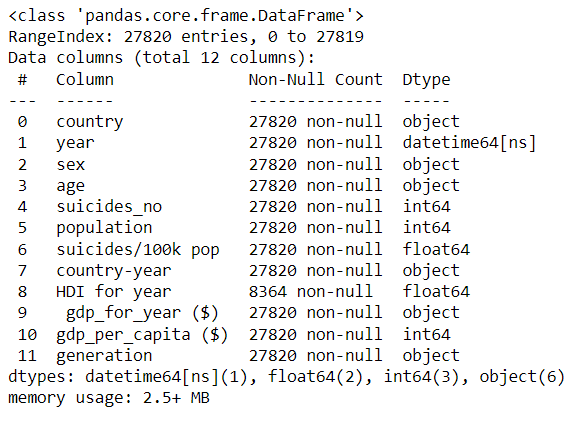
### From above we can see HDI for year having null values.

# Null value Percentage:



From above we can see 69% of null values are present in that column so we cannot drop those values. It effects the distribution of respective column.

## **Converting year column into datetime format:**

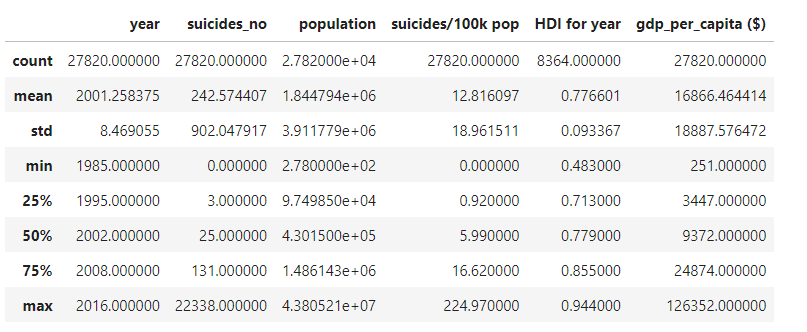


### **Insights:**

Year column converted into datetime format so we can easily get insights from that column.

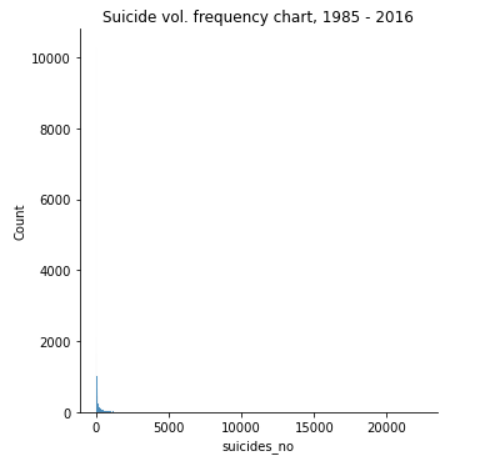
# Univariate Exploration:

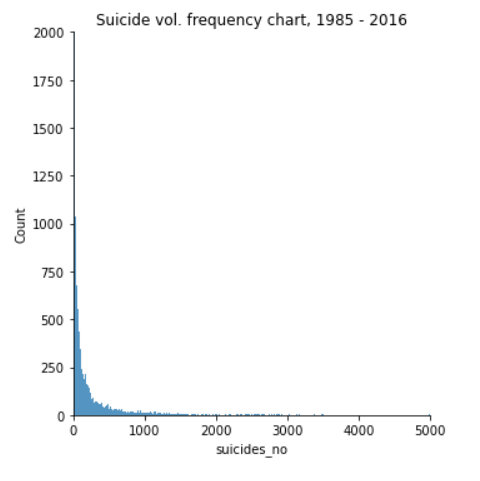
Let's havea look at the overall structure of the dataset



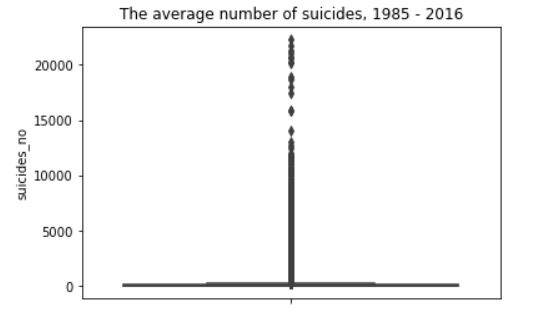
In total, there's a range of observations for 101 country for the period of 31 years, which is really impressive.

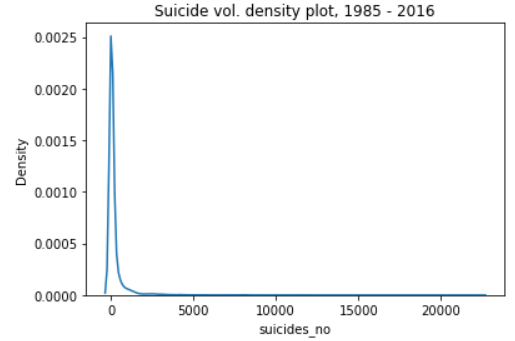
From here if we look at max and min numbers, we can already spot that suicides number has values close to zero, as well as its maximum value of 22,338. Let's investigate this value a bit closer.

****

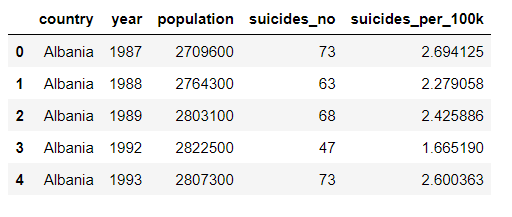
The plot immediately shows that the majority of suicides are sitting somewhere within the range of approx. 2,000. Let's have a closer look at values < 5000 and below 2000 to confirm. ****

Let's look at boxplots and density distribution

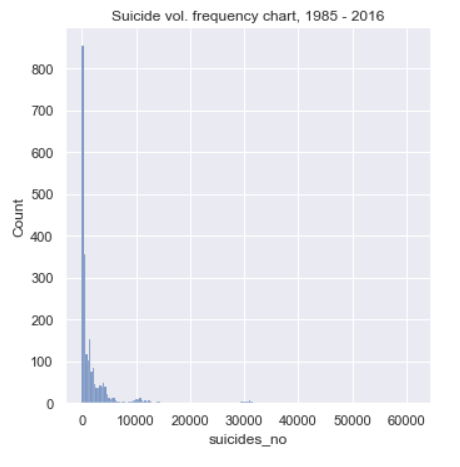


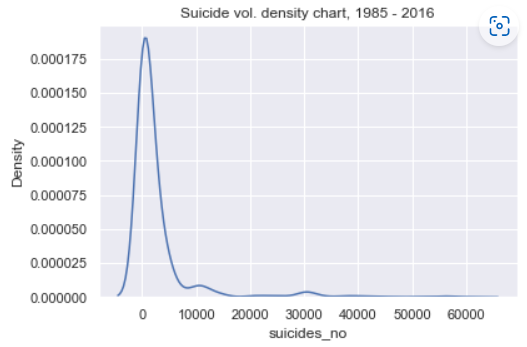


The given data has very long tail on the number of suicides preliminary driven by high granularity of the dataset (data is given per country per generation). Let's see if re-aggregating the data by country and year would give us more meaningful result.

****

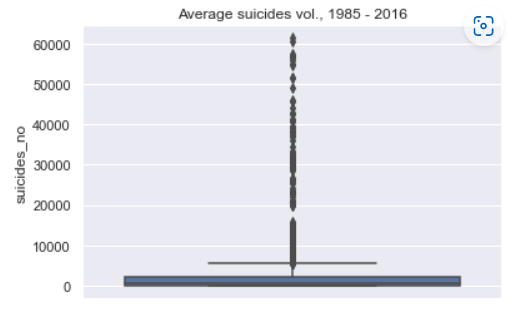
Let's revisit the metric after the transformation to see if more meaningful picture can be discovered

****

****

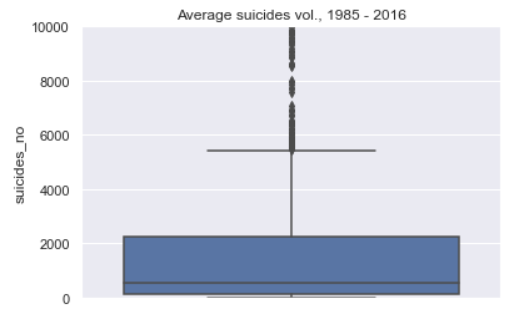
The above chart makes way more sense with the majority of annual suicides per country sitting somewhere within the range below 10,000 cases. However, there is a long tail of suicides that lasts up until 60,000 cases.

Another interesting thing to notice are two peaks within the range of 10k-15k annual suicides and an area within the range of 30k-35k.

****

The above chart doesn't give much sense due to many outliers sitting above ~8k suicides per year, which we can investigate later.

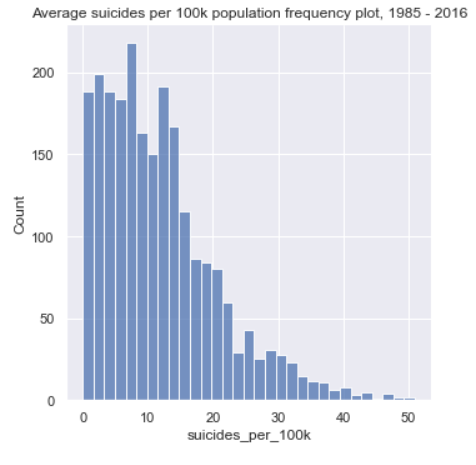
Let's look at the boxplot body a bit closer to understand the average world figure a bit better.

****

Above picture is way more optimistic.

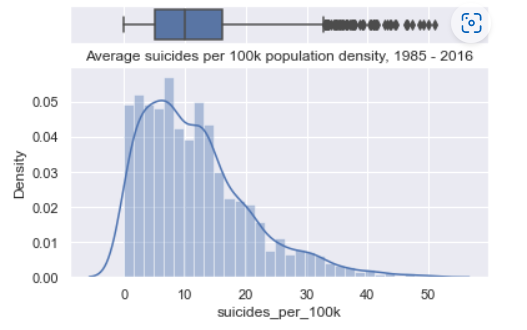
Despite the number of outliers, the average aggregated picture is reassuring. 50% of all suicide numbers within the given number of countries and date ranges will be sitting below 1,000 people, and 75% of the whole dataset is below 2000, with the majority of 101 countries across 31 year of observations is sitting below 6,000 suicides per annum.

Let's investigate the suicide per 100k population before jumping into outliers investigation to identify if any of them needs to be removed from the dataset.

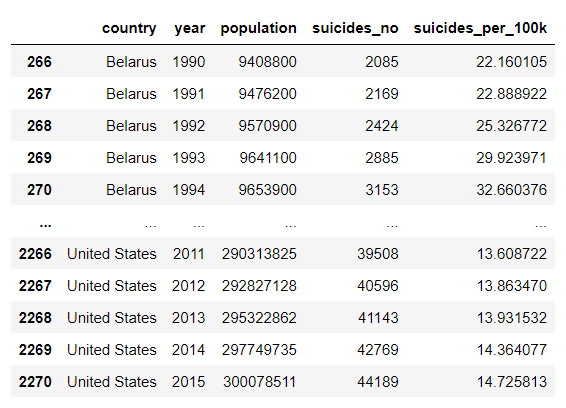


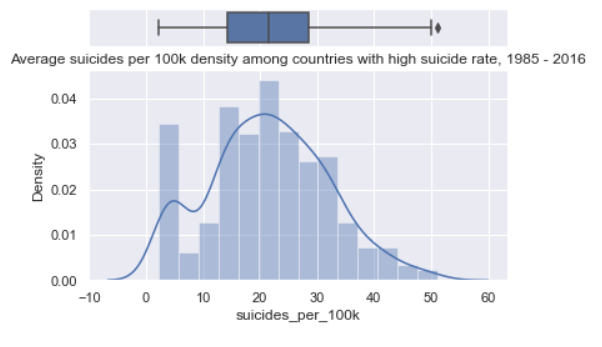
We can see that world average across 31 year would be sitting somewhere within 10 suicides per 100k population. Overall, the dataset is heavily skewed to the right having non-normal distribution.

Together, after the investigation we can categorize outliers as countries having > 6000 suicides per year or over 31 suicide cases per 100k population. Let's review those in details.



**filtering data with high suicide rates to investigate a little bit further**



Let's investigate data distribution for countries with high suicide rates.

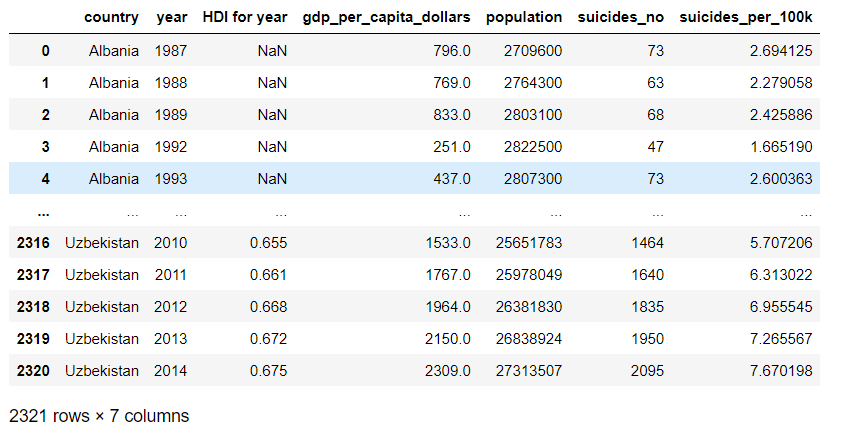
### Discuss the distribution(s) of your variable(s) of interest. Were there any unusual points? Did you need to perform any transformations?

On countries with high suicide rates - we ended up having more normally distributed data while still having one outlier being Lithuania.

According to Wiki, high suicide rate is a significant social issue in the country due to its high rate reaching its peak in 1995 year.

In my further analysis I am going to categorize countries as low/moderate/high/extreme suicide rates and will try define if the population, Human Development Index, or GDP per capita have a strong influence on country suicide rates.

**Final wrangling before bivariate exploration**

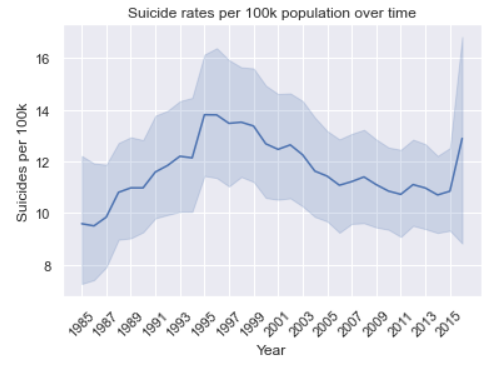


The dataset now is categorised based on suicides per 100k. Let's jump to bivariate exploration.

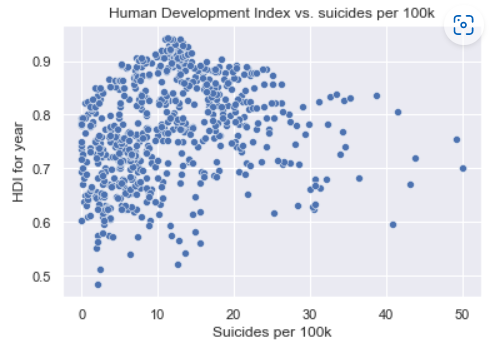
# Bivariate Exploration:

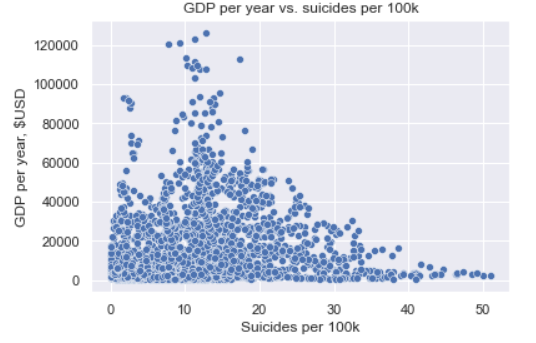
As we can see from here, the average worldwide suicide rates been climbing up constantly to reach peak between 1195 to 1997, followed by a gradual decrease in rates until 2015. However, after they spiked again in 2016!

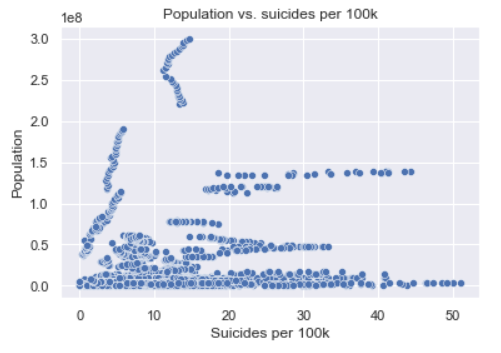
look at how mean suicide rates been changing over time



Let's investigate potential dependencies / correlations between suicide rates and other factors including population, country GDP, GDP per capita and human development index wherever it is possible.







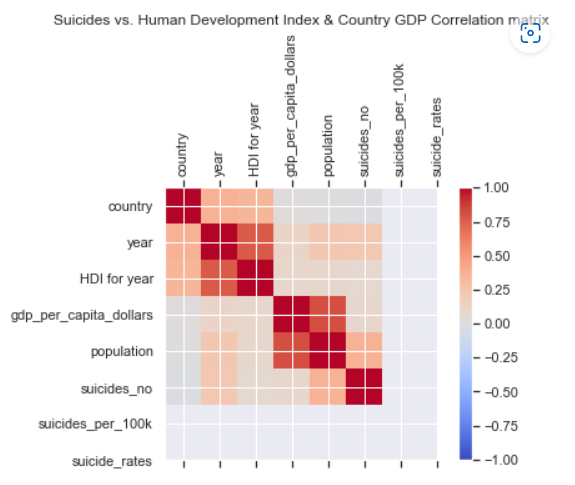
# Multivariate Exploration:

Based on the correlation matrix, there is nothing we can strongly associate with the plot.

A few mild/strong positive correlation points were discovered based between time variable and human development index, as well as GDP growth (good news). However, it is interesting to observe only a slight/weak correlation between gdp per capita and country gdp. That means that it is unlikely that if the country is producing more GDP, its households will get much richer.

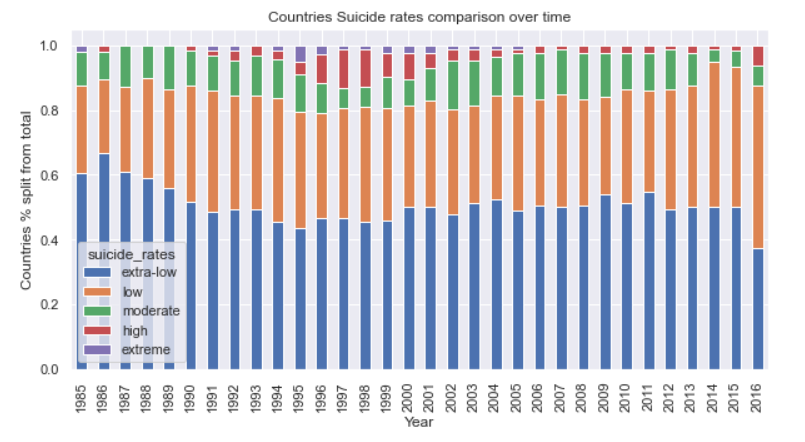
Let's have a look into what's going on within the ratio of countries with different suicide rates over time.

**Correlation matrix**

****

We can see from the chart, that period 1996-2001 there was way higher number of countries with high suicide rates. As the world moves on, more countries are shifting from extreme cohorts towards lower cohort.

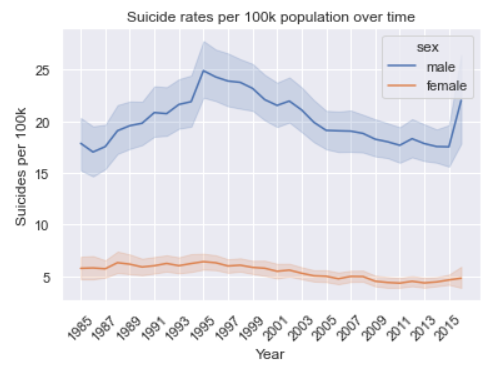
However, in the meantime, extra-low segment shows almost no improvement at all, what can set good targets in the future.

****

In this section, I want to put more focus to understand the relationship between countries with different suicide rates and key life characteristics, as well as to understand if high suicide rates are driven by a particular group/generation of people.

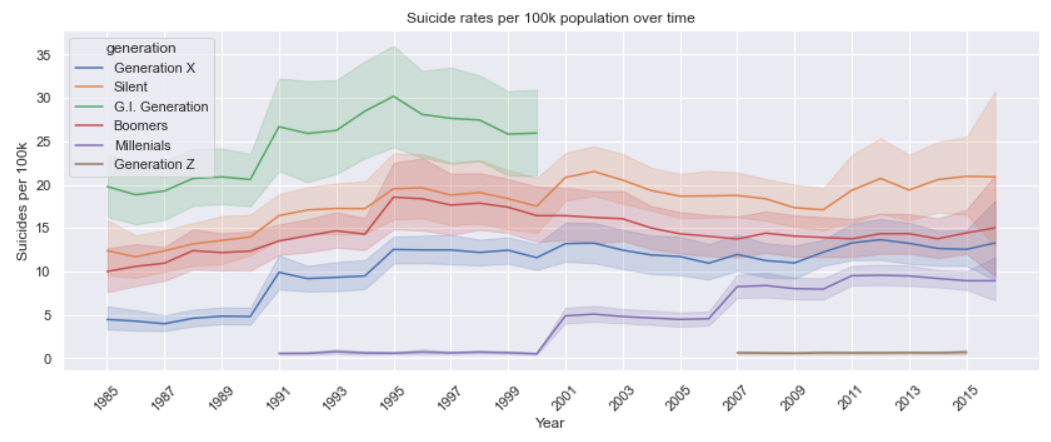
### **Men tend to commit suicide more often, than women.**

First of all, males tend to be more vulnerable group than females, as their suicide rates per 100k population is many times higher than females. That pattern is observed among all age groups, and observed consistently over the course of the given time period from 1985 to 2015.



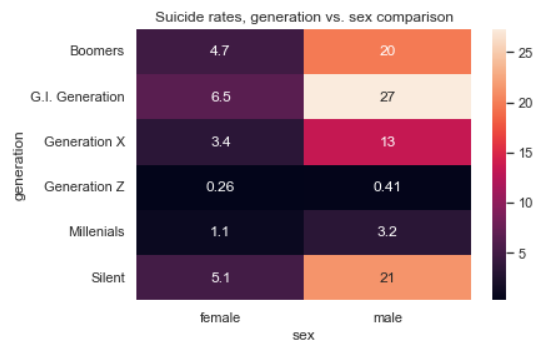
## **WWII Generation reported highest suicide rates than everyone else**

Based on the chart, G.I. Generation as well as Silent one tend to have the highest suicide rates per history.



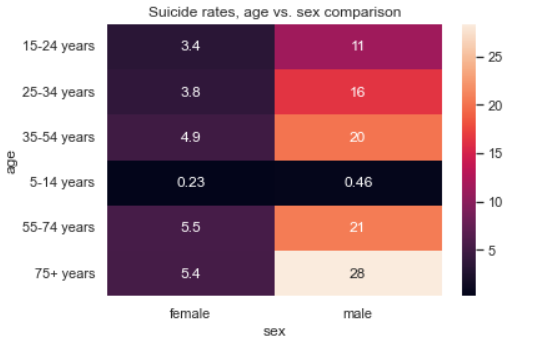
## **WWII Generation reported highest suicide rates than everyone else**

Based on the chart, the trend is similar across both males and females.

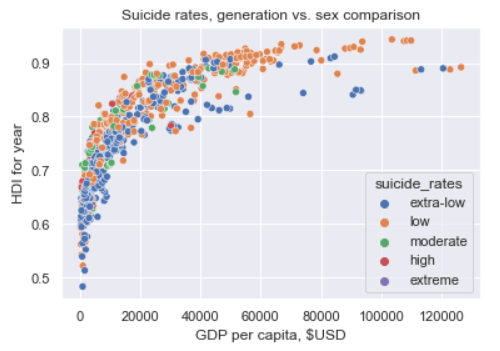


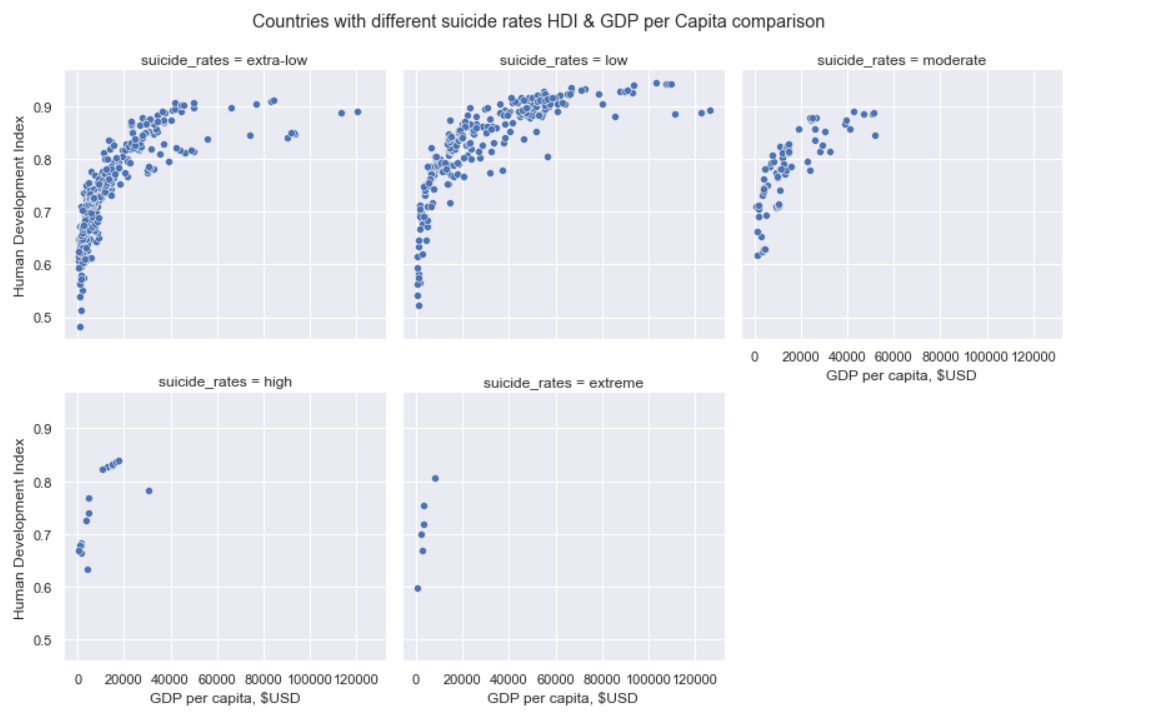
#### **The older the person, the higher the suicide rate is.**

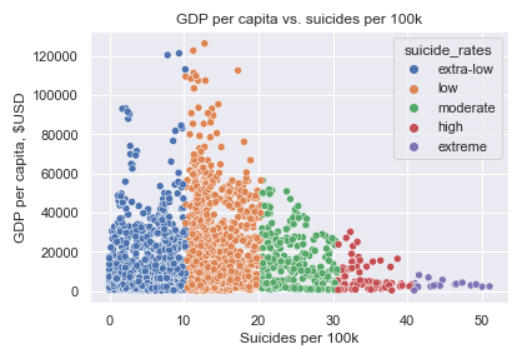
Another interesting discovery was done based on the age. The older the group, the higher suicide rate per 100k population is. This trend is consistent across sex, and probably partially driven by decrease of the size of the group, as well as increased number of suicides.

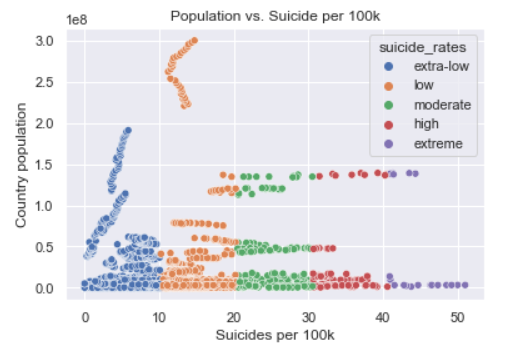
It is alarming considering that planet population is getting older in general.

**Scatter Plot:**

****

****

****

****

## **Limitation:**

All countries are not included in the dataset, i.e there are missing countries and cannot analyze them.

**Conclusion:**

1- Suicide are higher among men than women.  
2-Higher standards of living do not necessarily indicate lower suicide rates.  
3-Suicide numbers were decreasing up till 2010, after which they started rising again.  
4-Middle-aged people are more likely to commit suicides globally.  
5-The year, sex, and population are strong indicators when used to predict the number of suicides for a country.   
6-There is a weak positive relationship between a countries GDP (per capita) and suicide rate.

## **Reference:**

<https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>

United Nations Development Program. (2018). Human development index (HDI). Retrieved from <http://hdr.undp.org/en/indicators/137506>

https://github.com/Varshithreddy33/Suicide-Rates-Overview