# STATISTICS FOR DATA SCIENCE: MINI-PROJECT
# SUICIDE RATE OVERVIEW


## Team Details:

1. Michelle Mary Varghese - PES1UG19CS271
2. Navya Eedula – PES1UG19CS293
3. Parimala V- PES1UG19CS324
4. Pradnya Sanjeev - PES1UG19CS331



## Abstract:

According to the World Health Organisation, approximately one million people commit suicide every year. This translates to a global mortality rate of 0.00016 (per 1,00,000 people) or 1 death every 40 seconds. Due to the COVID-19 pandemic, there's been an increase in suicide rates due to many factors like unemployment, domestic violence, stress, anxiety, etc. Hence, we thought that it would be relevant to discuss this issue through our project.
In our project, we selected an unclean dataset of suicide rates across 40 countries, made appropriate modifications by imputing values into the HDI column and removing those countries that had a large percentage of missing values. We then plotted various graphs in order to extract meaningful insights from our data. We then standardised and normalised the variables of our data in order to make it more consistent and readable. This not only helped us find suitable variables for comparison, but also proved to be effective to conduct hypothesis testing for our problem statement. Finally, we found the relationship between all the variables in our dataset, in order to comprehend what factors affect suicide rates the most.


## Introduction:

Through our data collection and visualisation, we aimed to analyse suicide rates over the years, amongst males and females of all age groups. According to WHO, suicide rates are increasing amongst 15-24 year olds, and is currently, the second leading cause of death amongst them. As students who lie, more or else in that age group, we made a conscious decision to research about the various factors that lead to suicides in the country with the aim of becoming socially aware citizens and play a small part in creating awareness amongst our peers and family. Therefore, we tried to study the trends and factors affecting suicide rates across various countries, analyse them and draw meaningful insights.

The two variables, HDI for year and GDP for year, and their correlation to mean suicides, are the key factors our project is based on. These two parameters, are not only important to analyse development and progress of a country, but also talk a great deal about the suicides trends in that particular country. The **Human Development Index** (**HDI**) is a measure of a country's overall development in socio-economic dimensions, which include health of people, their level of

education attainment and their standard of living. The **Gross Development Product (GDP)** is the standard measure of the value of the production of goods and services in a country during a certain in a year. It gives a perspective of the economic growth of a country.

## Data Set:

Our original dataset consists of data about suicide rates across 101 countries from the years 1985-2016. It has 12 columns – country name, year, sex, age, number of suicides, population, suicides per 100k, HDI for year, GDP for year, GDP per capita and generation. It has about 4% NULL values.

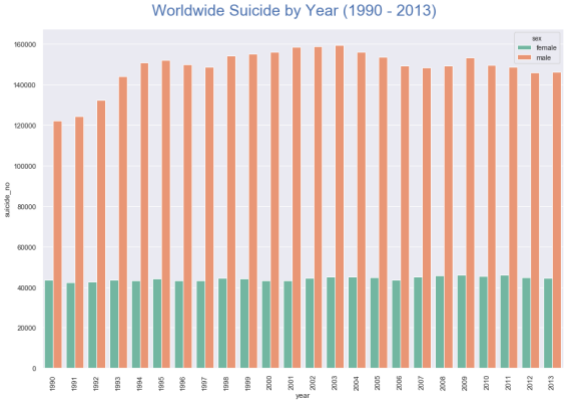
## Data Cleaning:

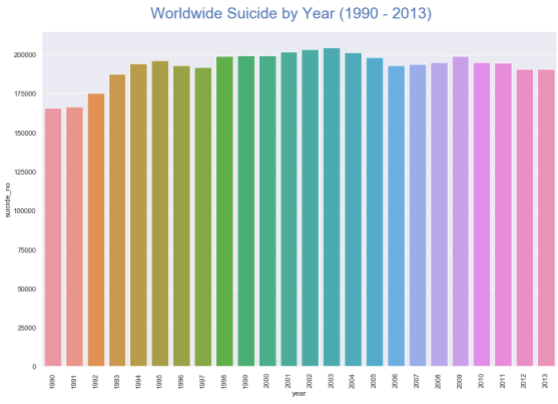
The dataset for 'Suicide Rates Overview' had all its NULL values concentrated in the HDI for year column. Hence, we had to clean it to make the data valid and complete in order to perform accurate analysis.
After cleaning, our dataset consists of data from 40 countries (with 11521 rows). We decided to impute values into the HDI for year column from the official UNDP website, by merging the .csv files. After that, we found that many years still had missing values for man countries. Hence we decided to remove those rows, in order to make our data consistent. Finally, we converted the year column into categorical data.

## Exploratory Data Analysis:

Through our analysis, we have made the following insights:

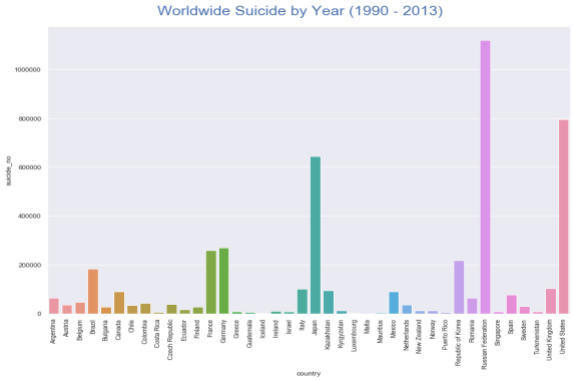
1. The suicide rates amongst males is higher than females, in fact more than double that of females.
2. The year 2003 has had the most number of suicides, followed by 2002 and 2009.

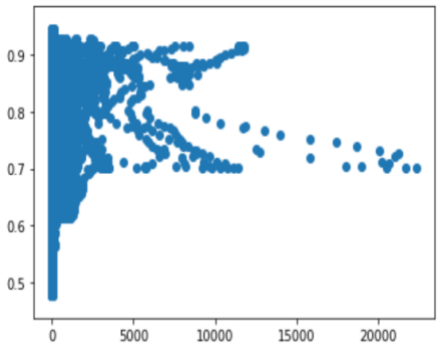



3. We can conclude that Russia has the most number of suicides, followed by Japan. According to our dataset, the Scandinavian countries are proven to be the ‘happiest’ countries with least number of suicide rates.

4. As suicide rate is plotted along the X-axis and the HDI for year is plotted along the Y-axis, we can conclude that they are negatively correlated, i.e.., as HDI increases, the number of suicides reduces.

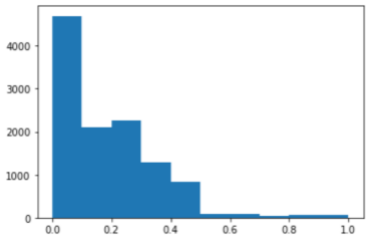


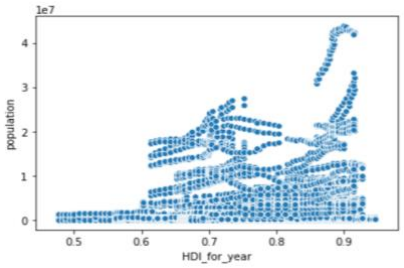



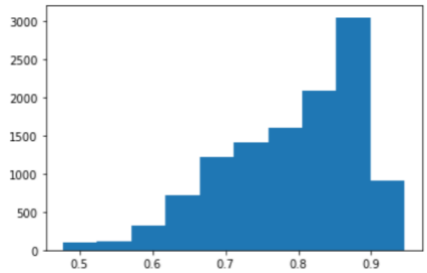

5. We can arrive at the conclusion that generally, countries with a large population have a larger Human Development Index.

6. The histogram is plotted for GDP per capita for all the countries. The histogram is right-skewed. The skew factor is 1.472. There are more than 4000 values that have values that have GDP (gross domestic product) per capita between 0 to 20,000.

7. The histogram is plotted for HDI for year. The histogram is left skewed. Most countries have HDI for year between 0.8 and 0.9

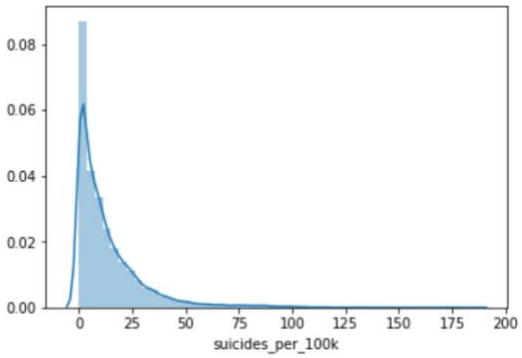


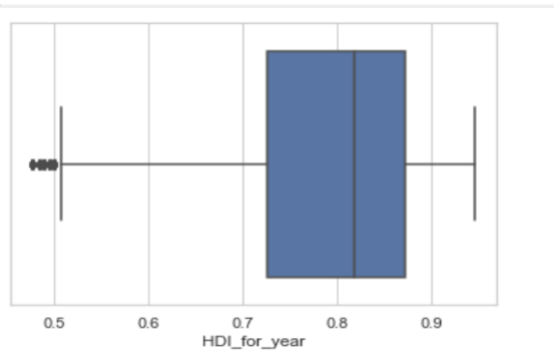




8. The histogram is plotted for suicides per 100k. The highest number of suicides lies between 0 to 25 (per 100k).

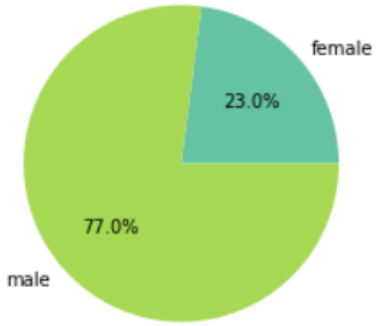
9. HDI for a year has a minimum value around 0.5 and a maximum value around 0.95 The first quartile line lies around 0.73 while the third quartile line lies around 0.88. The median for our data is approximately 0.82. There are a few outliers which indicate that a few countries have HDI less than 0.5.

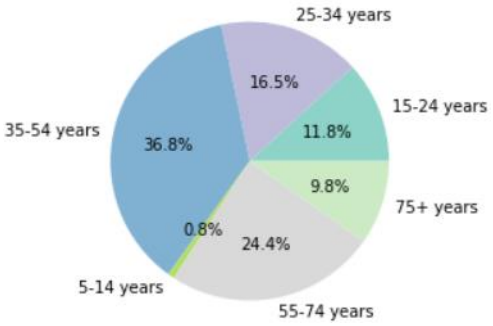



10. Suicide rates amongst males is greater than females.

11. Suicide rates are maximum in the age group of 35-45 years, followed by 55-74 years.




<u>Standardisation and Normalisation:</u>

Converting the data, such that the mean is equal to 0, and the standard deviation is equal to 1 is called standardisation.

In [9]:
```
#Standardisation - Scale down distribution with the parameters of standard normal distribution
# Mean = 0; Variance/SD = 1
Standardised_Data = X_Data.apply(lambda x: (x-x.mean(axis=0))/(x.std(axis=0)))
Standardised_Data
```

Out[9]:

| | suicide_no | population | suicides_per_100k | HDI_for_year | gdp_for_year | gdp_per_capita |
|---|---|---|---|---|---|---|
| 0 | -0.138703 | -0.488249 | 2.386403 | -0.914818 | -0.346866 | -0.875045 |
| 1 | 0.079274 | -0.149756 | 0.577621 | -0.914818 | -0.346866 | -0.875045 |
| 2 | 0.029519 | 0.144200 | -0.082947 | -0.914818 | -0.346866 | -0.875045 |
| 3 | -0.269015 | -0.442529 | -0.235120 | -0.914818 | -0.346866 | -0.875045 |
| 4 | -0.173453 | -0.116437 | -0.325616 | -0.914818 | -0.346866 | -0.875045 |
| ... | ... | ... | ... | ... | ... | ... |
| 11515 | 0.697665 | 3.561455 | -0.428794 | 1.300355 | 8.046216 | 1.651521 |
| 11516 | 0.452836 | 3.708713 | -0.523326 | 1.300355 | 8.046216 | 1.651521 |
| 11517 | 0.023200 | 1.695324 | -0.565980 | 1.300355 | 8.046216 | 1.651521 |
| 11518 | -0.118169 | 3.616864 | -0.713542 | 1.300355 | 8.046216 | 1.651521 |
| 11519 | -0.204254 | 3.434629 | -0.741786 | 1.300355 | 8.046216 | 1.651521 |

11520 rows × 6 columns

Converting all the data, such that it lies in the range of 0 to 1 is called normalisation.

```
In [7]: #Normalisation - All data will lie within 0 and 1
        #Normalising data using lambda function
        Normalised_Data = X_Data.apply(lambda x: (x-x.min(axis=0))/((x.max(axis=0))-x.min(axis=0)))
        Normalised_Data
```

Out[7]:

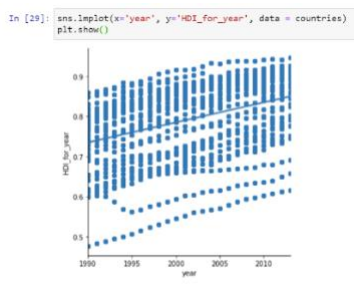
| | suicide_no | population | suicides_per_100k | HDI_for_year | gdp_for_year | gdp_per_capita |
|---|---|---|---|---|---|---|
| 0 | 0.010117 | 0.009278 | 0.296542 | 0.490405 | 0.008394 | 0.037309 |
| 1 | 0.022473 | 0.048477 | 0.127259 | 0.490405 | 0.008394 | 0.037309 |
| 2 | 0.019653 | 0.082517 | 0.065437 | 0.490405 | 0.008394 | 0.037309 |
| 3 | 0.002731 | 0.014573 | 0.051195 | 0.490405 | 0.008394 | 0.037309 |
| 4 | 0.008148 | 0.052335 | 0.042725 | 0.490405 | 0.008394 | 0.037309 |
| ... | ... | ... | ... | ... | ... | ... |
| 11515 | 0.057525 | 0.478242 | 0.033069 | 0.931770 | 1.000000 | 0.459245 |
| 11516 | 0.043648 | 0.495295 | 0.024222 | 0.931770 | 1.000000 | 0.459245 |
| 11517 | 0.019294 | 0.262141 | 0.020230 | 0.931770 | 1.000000 | 0.459245 |
| 11518 | 0.011281 | 0.484659 | 0.006420 | 0.931770 | 1.000000 | 0.459245 |
| 11519 | 0.006402 | 0.463556 | 0.003776 | 0.931770 | 1.000000 | 0.459245 |

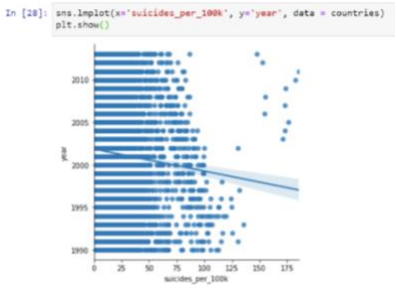
11520 rows × 6 columns

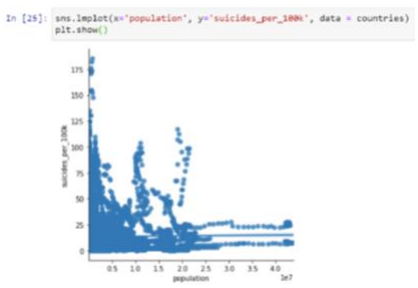
Standardisation and normalisation are done in order to ensure uniformity, consistency and avoid variability amongst the different variables being compared, due to differences in ranges or units.

Correlation:

i) Year and HDI for year are positively correlated.
ii) Year and suicides per 100k are negatively correlated.
iii) Population of the country and suicides per 100k not correlated.




## Hypothesis Testing:

As a part of our hypothesis testing, we decided to conduct two student's T test in order to obtain valuable insights from our dataset.
In our first research hypothesis, we decided to check to if the mean suicides per 100k in 2003 would be lesser than the population mean suicides per 100k for a set of countries whose HDI was greater than 0.801325. The null and alternate hypothesis are as follows:
*#H0:The total suicides_per_100k is* $\mu$ *>= 2.368484*
*#H1:The total suicides_per_100k is* $\mu$ *< 2.368484*

```
tst, pval=stats.ttest_1samp(test_sample["suicides_per_100k_log"],
                  2.368484)
print("tst",tst)
print("pvalue",pval/2)
if pval/2 < 0.05:    # alpha value is 0.05 or 5%
   print(" we are rejecting null hypothesis")
else:
  print("we cannot reject null hypothesis")
```

```
tst 1.2687148265813202
pvalue 0.11536665042348582
we cannot reject null hypothesis
```

On calculations, the p value was found to be 0.115, which is greater than our significance level (0.05). Hence, we fail to reject the null hypothesis and therefore, H0 is plausible. We then concluded that there is not sufficient evidence to support the claim that the total suicides per 100k will be lesser than 12.588647 if the HDI is greater than 0.801325.

As a part of our second research hypothesis, we conducted a similar hypothesis test, but for GDP for year. A set of countries whose GDP was greater than 21346.625 wanted to prove that their mean suicides per 100k in 2003 would be lesser than the population mean suicides_per_100k.
*#H0: The total suicides_per_100k is $\mu$ >=2.368484*
*#H1: The total suicides_per_100k is $\mu$ < 2.368484*

```
tset, pval=stats.ttest_1samp(test_sample_gdb["suicides_per_100k_log"],
                  2.368484)
print("tst",tset)
print(pval/2)
if pval/2 < 0.05:    # alpha value is 0.05 or 5%
   print(" we are rejecting null hypothesis")
else:
  print("we cannot reject null hypothesis")
```

```
tst 2.159887632709311
0.026857397818945045
 we are rejecting null hypothesis
```

On calculations, the p value was found to be 0.027 which is clearly lesser than the significance level. Hence, we reject the null hypothesis and accept the alternate hypothesis. We then concluded that, there is some evidence that total_suicides_per_100k will be lesser than 12.588647 if the GDP is greater than 21346.625.

## Result and Conclusion:

After finishing our analysis, we have drawn many meaningful insights from our dataset. Suicides were maximum in the age group of 35-54, in the years 1990-2013. And this statistic was uniform through all the forty countries. Suicides were significantly greater amongst males than amongst females. Amongst the 40 countries analysed, suicide rates were found to be highest in Russia. On further research we found that this was mainly due to two reasons: economic crisis and heavy alcohol consumption. Thankfully, as of recent years, suicide rates have been on decline in Russia, alongside decrease in alcohol consumption throughout the country.

In our hypothesis testing, we found concluded that
a) There is not sufficient evidence to support the claim that the suicides per 100k, will be lesser than the mean suicides per 100k,  if the HDI is greater than 0.801325.
b) There is some evidence that total_suicides_per_100k will be lesser than the mean suicides per 100k if the GDP is greater than 21346.625.
On performing correlation analysis, we proved that the variables, GDP per capita and HDI for year, and year and HDI for year are positively correlated. This means that as the years go by, the HDI of all countries worldwide are improving hand in hand with economic growth. This in turn means suicide rates are on the decline. As long as the population of a country are well educated and are in good health and happiness, it is possible to minimise the deaths caused due to suicides.