Exploratory Data Analysis on suicide data of past 32 years

*Abstract*— Suicide is a global phenomenon and threat to the social wellbeing which effects not only the victim but also the people around them. According to WHO close to 800,000 people die due to suicide every single year. General study in USA shows that everyday approximately 123 American die by suicide. In America alone 44,965 people take their lives according to save.org. More alarmingly WHO also said suicide is one of the top 10 reasons of death worldwide and second leading cause of death among the people of age 15-29 [1]. Most of these cases could have been prevented and is still possible to prevent if enough steps can be taken after understanding what causes these phenomenon. Finding patterns in these data will open up more spots to shed light on. The more we understand the data, it is more likely that we will find a way to reduce the number in a great deal.

Keywords—EDA, suicide, data analysis, data visualization, data science.

# Introduction

Suicide is a worldwide concern and one of the major causes of death. Each suicide can be compared to a pebble in a pond. It has a ripple effect just like that wave in the pond. It affects the family and society. When a well-known celebrity takes his or her own life, it has an effect on the psychology of the people around the world. If we can analyze the deaths and can figure out which people are most prone to suicide, it will be easier to take actions accordingly and it might be possible to reduce the number of deaths significantly.

We have used a dataset which has 32 years of suicide data of 100 different countries. This dataset gives us an insight of the whole world scenario. This dataset is publicly available on a very popular website called Kaggle.com [2]. This dataset has been analyzed thoroughly to find out patterns. These patterns will help us to study those deaths with more meaning and insights. Several library were used to analyze the data-set. For example seaborn, matplotlib for visualization and pandas, numpy to interact with the data and os to iterate the paths of data-set. Some patterns were extracted that will help us understand different aspects of those suicide cases.

# Releted Work

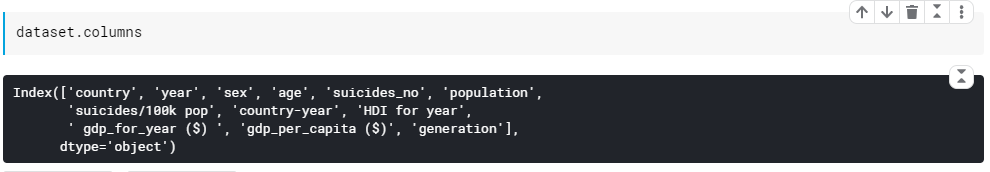
Not much works have been done using Exploratory Data analysis technique on suicidal case study. Closest work found which was authored by Antonio Rodriguez Andres who analyzed income inequality and unemployment of 15 European countries data from 1970 to 1998 and found out that economic growth, fertility rate and alcohol consumption seem to have a significant impact on male and female suicide rates [3].

# Dataset

The dataset was taken from Kaggle website. It has data of 100 countries suicide counts from 1985 to 2016. The file contains 12 rows and 27,820 samples. Suicide numbers of different countries were counted for 32 years. There are columns for five categories age groups. Other columns are sex, population, suicides/100k pop which is the number of suicides occurred per 100,000 population, HDI for year, gdp\_for\_year ($), gdp\_per\_capita ($), generation of population. Other column country-year was dropped. HDI column was also dropped as it had a lots of null values.

# Exploratory data analysis

This work has been done using Jupiter Notebook. After reading the dataset from csv to pandas data-frame, whole data-frame was grouped and sorted using the year column. Let’s look at the columns before moving any further:



We will first look at the suicide rates combining all the 100 countries over the 32 year time.

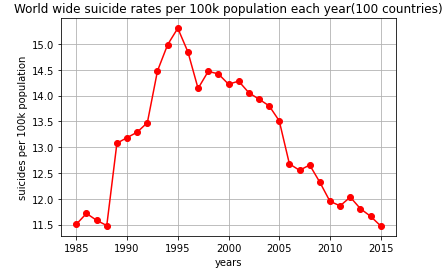


Fig 1: Suicide rate

This is very evident from the graph above (fig 1) that in 1995 there was huge numbers of suicide cases all around the world. Now let’s see the suicide graph in total number of suicides.

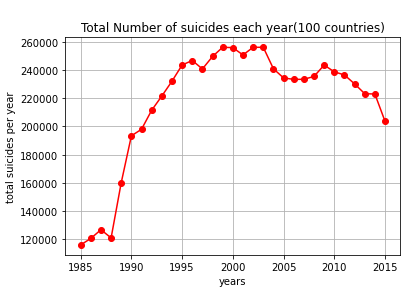


Fig 2: Total suicide numbers

This graph suggests number of people committed suicide was very high from 1995 to 2004. Then the number began to lower down. More standard parameter to realizing the suicide cases is the suicide rate. In 1995, the suicide rate was peak and the number was also very high at that year. Then after 1995 the numbers started to go down gradually. But which countries were helping this curve to climb up to that peak? When the countries will be found then we can analyze deeper what factors were contributing to the conditions of that area, or which age groups were prone to suicide and what was these countries GDP. Because often economic crisis can induce the attempts of suicides. Let’s look at the countries suicide cases in numbers and rates over the 32 year span.

Country wise data visualization:

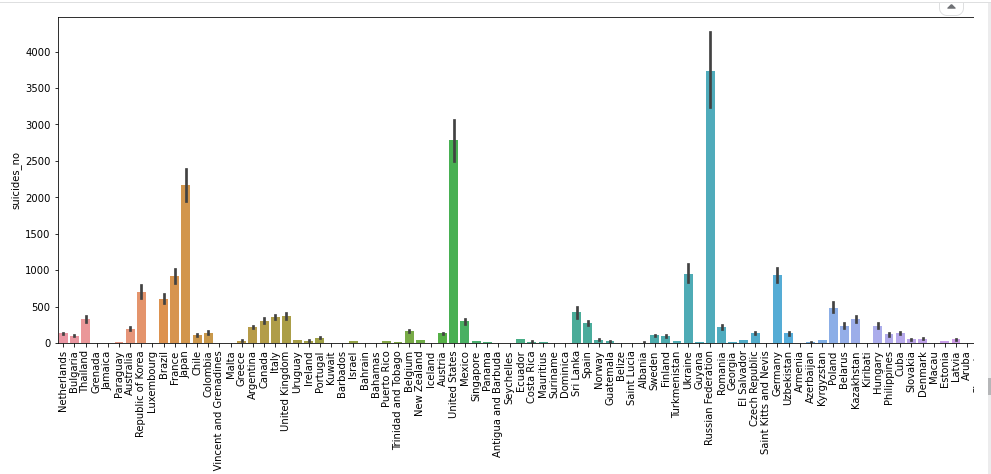


Fig 3: Suicide number in different countries

This is the total number of suicide cases among different countries. Top three countries are – Russian Federation, United States and Japan. Number of cases in France, Germany and Ukraine are also very high. But these countries have lots of populations compared to the other countries. So it doesn’t show us the real picture. So, let’s visualize the suicide cases among the countries by per 100 thousand people. This is a proper metric to understand the situation in different countries.

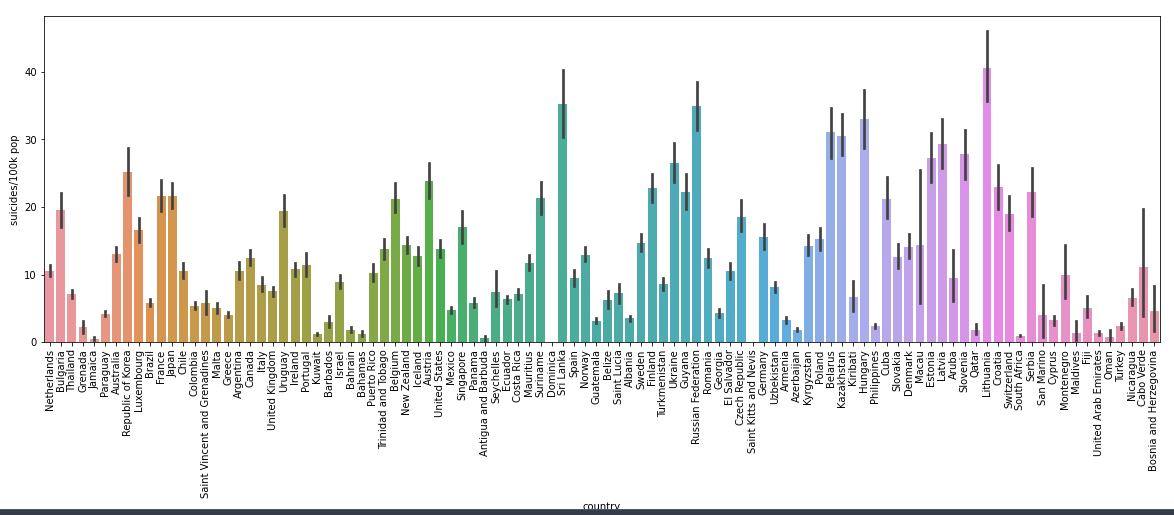


Fig 4: Suicide rates in different countries

This shows much better and proper situation of the countries. This is showing the cases among per 100 thousand people. Now it shows that Lithuania has the highest suicide rate among all other countries. Other countries with high suicide rates are Sri Lanka, Russian Federation, Belarus and Hungary. Now let’s focus in year 1995 as this the year with highest suicide rate. We have shown in the graphs below, top 5 countries among those 100 in the dataset, which has highest number of suicides and highest suicide rates:

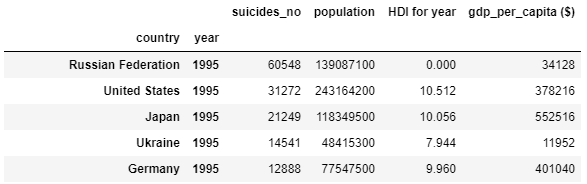


Fig 5.1: Top five countries according to suicide numbers

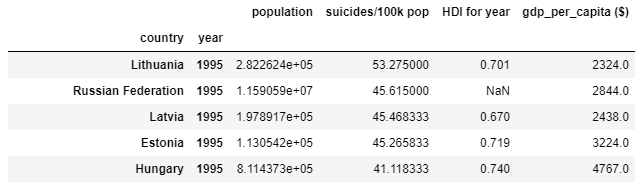


Fig 5.2: Top five countries according to suicide rates

It is very clear from the ranking of suicide rates (fig 5.2), Lithuania was in the first position. This country suffered a great deal with suicide in the year 1995. Though it started to decrease afterwards. History tells us about the tragedy of sudden increase of suicidal rates in Lithuania and Sri Lanka. Lithuania’s high suicide rate is thought to be the result of poor social and economic condition. This is a powerful indication to look into and further research. But the scariest picture shown by both the graphs are in Russian Federation. Despite of huge population, it is number 2 in the suicide rates (fig 5.2) ranking in 1995. And number 1 in total number of suicides (fig 5.1) in the total suicide number ranking list. More than 60.5 thousand people committed suicide and the rate was 45.615. These are areas where focus of researches need to be projected. That is only when we shall understand the factors that escalate the number of suicides. Only then it is possible to take necessary steps to reduce the death toll from suicide.

Here are some countries suicide rate graph along with GDP graph to look for patterns amongst low GDP and suicide rate:

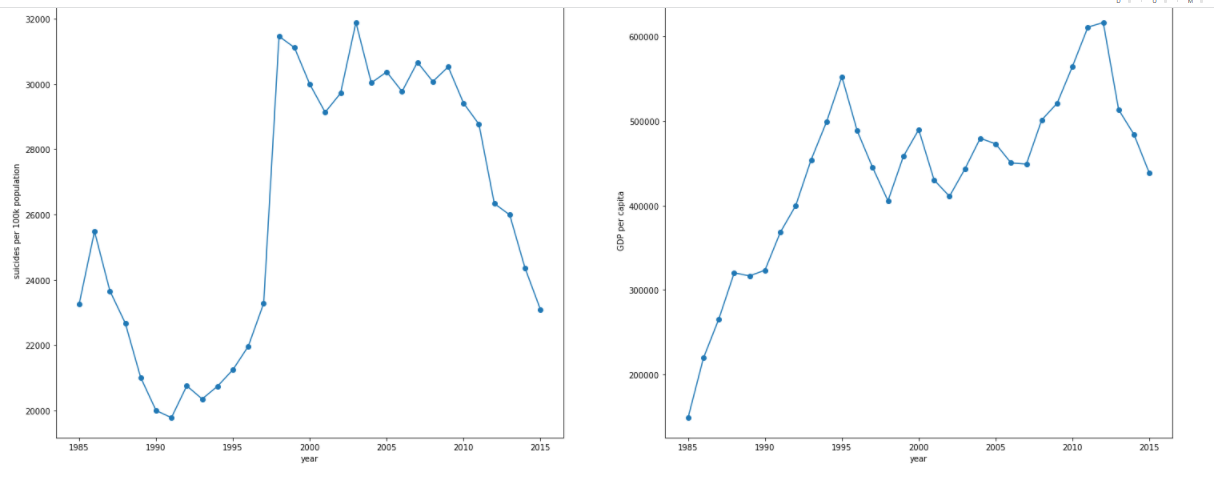


Fig 6.1: Japan

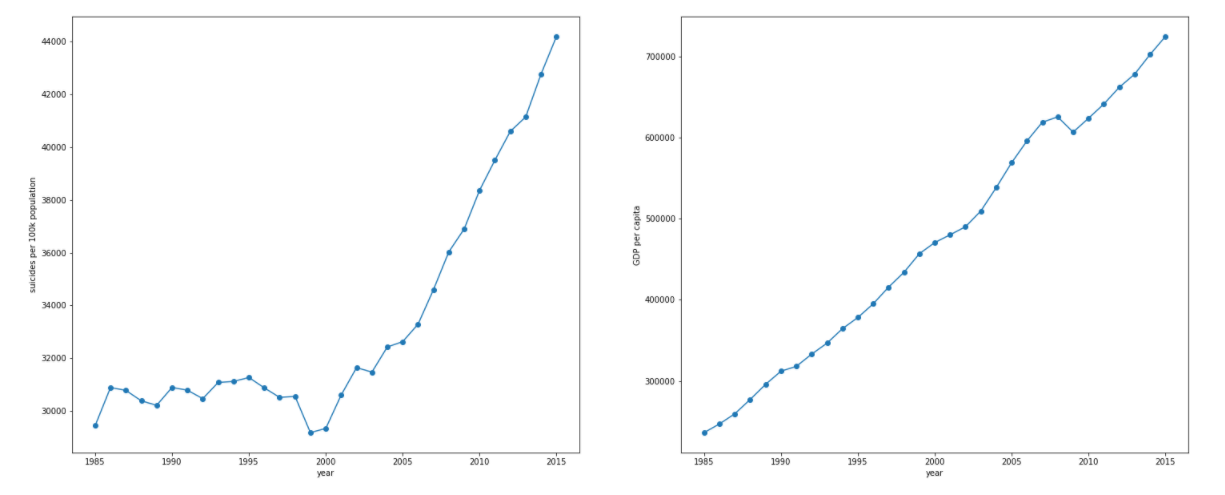


Fig 6.2: US

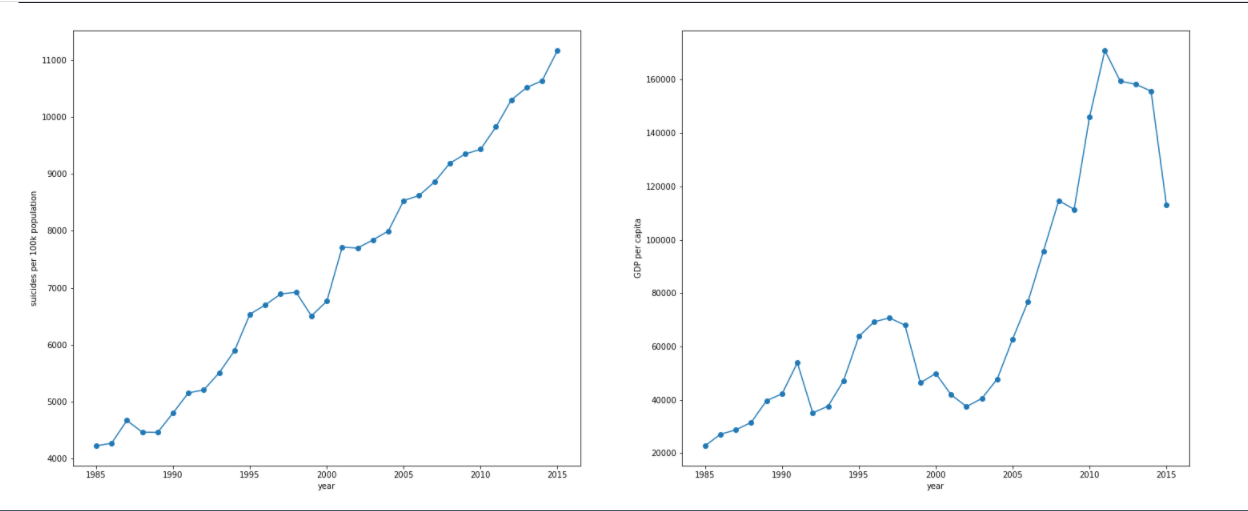


Fig 6.3: Brazil

In a glace it looks like theres strong casual effect between rise in suicidal rate and rise in GDP. Which says if GDP rises, suicidal rate increases too. But this in a common mistake when mining meaning from different datasets. This is like the popular proof that storks delivers babies. Robert Matthews showed very strong correlation and p-value of 0.008 in his paper that storks are the one who delivers the human babies [4]. But in all those cases there are other factors that causing both the variable to have a rise or fall relationship in tha graph. So, in these cases we have to be careful to choose the meaning of that correlation.

Now let us visualize which gender has more suicidal tendency than the other.

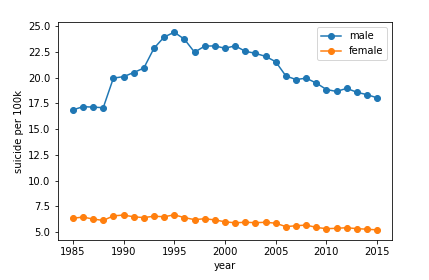


Fig 7: Suicide rates in male and female

So it clearly shows men have very high suicide rate compared to that of women have. Often it is as high as 5 times than the women’s rate of suicide. We have a column of different age 6 different age groups. So it will be interesting to know how much suicide rate differs when it comes to the age.

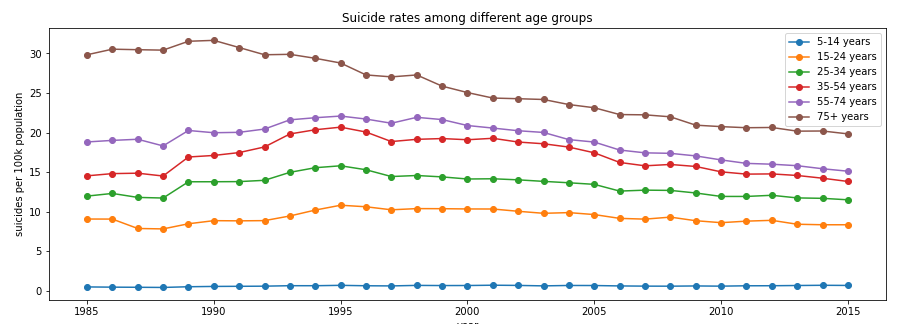


Fig 8.1: Suicide rate in different ages around years

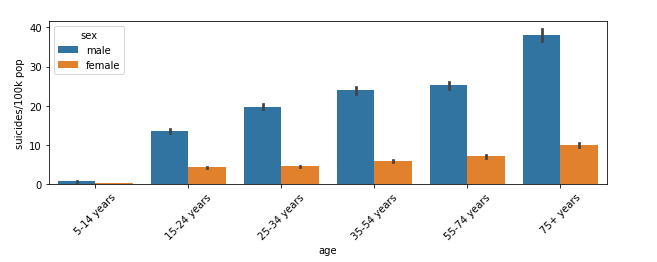


Fig 8.2: Suicide rate in different ages in male & female

So this is very evident that people of age 75+ have the highest suicide rates. That rate is very concerning as they are the people who already served the world economy and reached their retirement. The factors behind their high suicide rate need to be investigatd. Is it the loneliness that grasp them and takes to the pitfall of depression? Or is it there something else. These researches can create the possibilites to reduce their suicide rate. And the graph has a interesting pattern too. It looks like the more people age the more the suicide rate. Let’s look at the total number of suicides in different age groups:

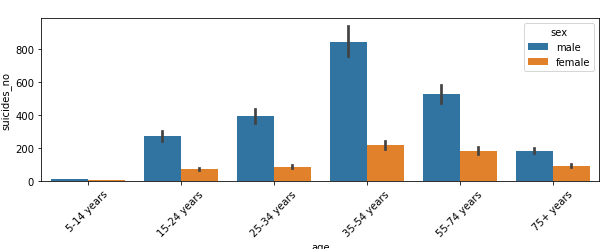


Fig 9: Total suicide numbers in male and female

As we look at the number it is very concerning as the 35-54 aged people have most number of suicides. This is the age when a lots of responisibity lies among them. Childrens, economic responsibilities, loans and lots of other factors can create pressurized situations among this age group of people. So this can be a reason of the high tendency of their committing suicide. This group needs more attention and more researches.

##### Conclusion

From this dataset we have gathered a lot of insights about the suicidal tendency among different people of different age and places around the years. When this is evident that a group of people in a certain place and time have higher tendency of committing suicide we can focus on them more and have research on them more intensively. This is the whole intension of this Exploratory Data Analysis. We can now focus more on the evident group people to find out more about these cases.

##### References

1. Study conducted by World Health Organization(WHO) in 2016: [http://who.int/news-room/fact-sheets/detail/the-top-10-causes-of-deat](who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death)h
2. Kaggle.com, ‘Suicide rates overview 1985 to 2016’ available: <http://www.kaggle.com/> [2018]
3. Antonio Rodiguez Andres, “Income unequality, unemployement and suicide: a panel data analysis of 15 European countries.” 21-Aug-2006, Doi: <https://doi.org/10.1080/0003684042000295304>
4. Robert Metthews, “Storks Deliver Babies”