# DATA MANAGEMENT PROJECT REPORT

(Project Semester: August-December 2020)

## *WORLD SUICIDE RATE ANALYSIS 1985-2016*


Submitted by

S Nandhini

Registration No. 11804841

Programme and Section: B.Tech (Computer Science),

Course Code: INT 217

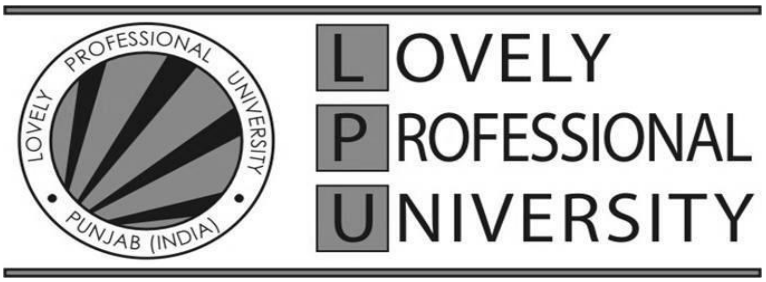
Under the Guidance of

**Savleen Kaur, 18306**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

# CERTIFICATE

This is to certify that S NANDHINI bearing Registration no. 11803529 has completed INT 217 project titled, **"WORLD SUICIDE RATE ANALYSIS (1985-2016)"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Savleen Kaur, 18306**

School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 18 December, 2020

# DECLARATION

I, S Nandhini student of Bachelor in Technology under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 18 December 2020

Registration No. 11804841 SNandhini

# ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my teacher Ms. Savleen Kaur who gave me the golden opportunity to do this wonderful project of analysis of the data of a superstore namely “WORLD SUICIDE RATE ANALYSIS: 1985-2016” which also helped me in doing a lot of research and I came to know about so many new things. I am thankful to them. Secondly, I would also like to thank my friends and faculties who helped me a lot in finalizing this project objectives within the limited time frame.

# Table of Content

# INTRODUCTION

**Data Analysis** is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

In this project WORLD SUICIDE RATE ALALYSIS OVER THE YEARS 1985-2016, the data does not provide information about religion/culture or law across countries which could be factors to suicide rate. (In many countries, suicidal behaviour is condemned by the society or is even against the law for religious/cultural reasons; also, physician-assisted-suicide might be a contributor to suicide rates in countries where it is legal.) Given the scope of this dataset, the project focuses on determine whether gender, age or GDP has an impact on suicide rate and how they impact suicide rate.

WORLD SUICIDE RATE ANALYSIS OVER THE YEAR 1985-2016: Data contains the following fields:

The dataset for this project includes 12 variables: country, year, sex, age group, suicide number, population, suicide rate, HDI, GDP per year, GDP per capita and generation. Data is recorded for a total of 101 counties from year 1985 to 2016.

# SCOPE OF ANALYSIS

I wanted to analyze the SUICIDE data collected from all over the world (1985-2016 year). Death by suicide is an extremely complex issues that causes pain to hundreds of thousands of people every year around the world. The main objective of this data entry is to contribute to an informed, open debate about the ways to prevent suicide by understanding it in a deeper level. The project aims to find and explain some relationship between suicide rate and gender, age group/generation or GDP across countries. I cleaned, analyzed and presented it in the form of a dashboard for it to be more understandable for users and to get more meaningful insights.

The Objectives:

# SOURCE OF DATASET

The data is being taken from the Kaggle.

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

**About Kaggle**: Kaggle is an Airbnb for Data Scientists – this is where they spend their nights and weekends. It's a crowd-sourced platform to attract, nurture, train and challenge data scientists from all around the world to solve data science, machine learning and predictive analytics problems. It has over 536,000 active members from 194 countries and it receives close to 150,000 submissions per month. Started from Melbourne, Australia Kaggle moved to Silicon Valley in 2011, raised some 11 million dollars from the likes of Hal Varian (Chief Economist at Google), Max Levchin (PayPal), Index and Khosla Ventures and then ultimately been acquired by the Google in March of 2017. Kaggle is the number one stop for data science enthusiasts all around the world who compete for prizes and boost their Kaggle rankings. There are only 94 Kaggle Grandmasters in the world to this date.

# ETL PROCESS

In computing, extract, transform, load (ETL) is a process in database usage to prepare data for analysis, especially in data warehousing. Data extraction involves extracting data from homogeneous or heterogeneous sources, while data transformation processes data by transforming them into a proper storage format/structure for the purposes of querying and analysis; finally, data loading describes the insertion of data into the final target database such as an operational data store, a data mart, or a data warehouse. A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions.

Precisely, ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. ETL stands for Extract, Transform and Load.

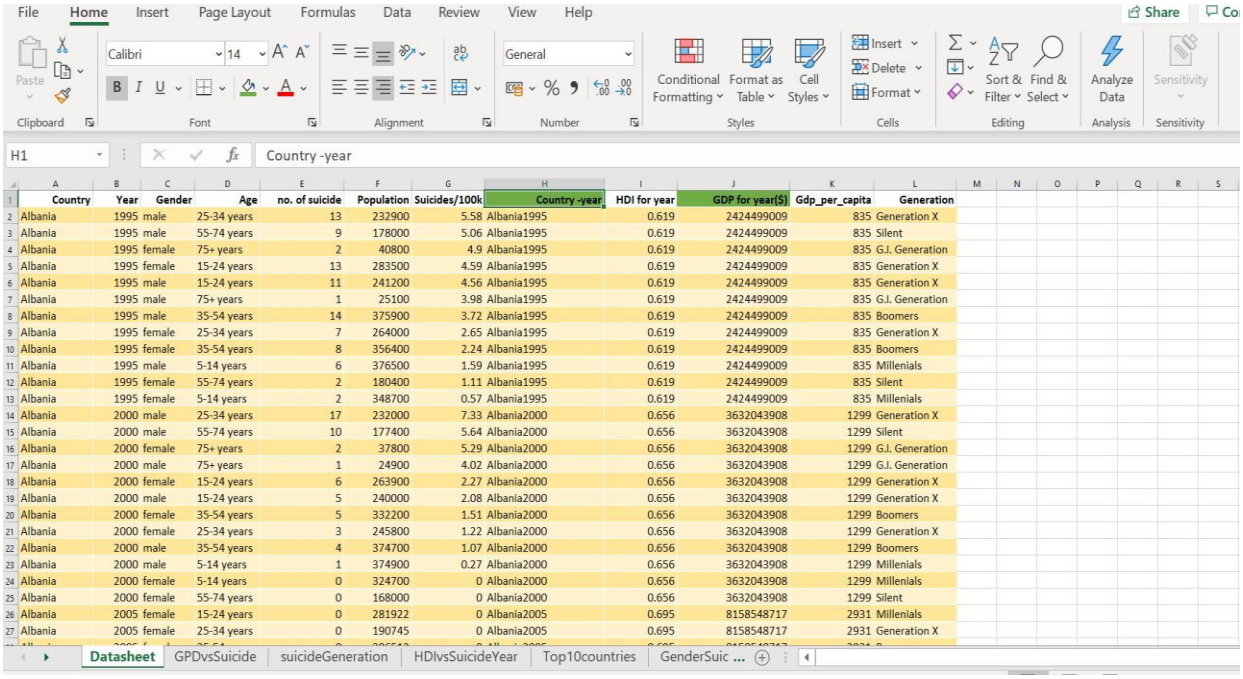
Through the process of ETL, we are going to clean the dataset and bring all the entities to their proper data format.

**Step 1: Removing the blank cells from the dataset.**

For this, select the whole dataset. Go to Find and Select in the Home tab of excel. Select Go to Special from the drop-down menu and then tick the blank option. All the blank cells will be selected. Then go to Delete option in the home tab again and select Delete Rows from the drop-down menu. This will remove any rows with blank cells.

**Step 2: Removing columns which are not properly defined or not crucial to our analysis.**

For this we will columns which are redundant like the column with just the index numbers. For this we will select that particular column and then go to delete option in the home tag and then select Delete Columns from the drop-down menu.




| Country | Year | Gender | Age | no. of suicide | Population | Suicides/100k | Country -year | HDI for year | GDP for year($) | Gdp_per_capita | Generation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Albania | 1995 | male | 25-34 years | 13 | 232900 | 5.58 | Albania1995 | 0.619 | 2424499009 | 835 | Generation X |
| Albania | 1995 | male | 55-74 years | 9 | 178000 | 5.06 | Albania1995 | 0.619 | 2424499009 | 835 | Silent |
| Albania | 1995 | female | 75+ years | 2 | 40800 | 4.9 | Albania1995 | 0.619 | 2424499009 | 835 | G.I. Generation |
| Albania | 1995 | female | 15-24 years | 13 | 283500 | 4.59 | Albania1995 | 0.619 | 2424499009 | 835 | Generation X |
| Albania | 1995 | male | 15-24 years | 11 | 241200 | 4.56 | Albania1995 | 0.619 | 2424499009 | 835 | Generation X |
| Albania | 1995 | male | 75+ years | 1 | 25100 | 3.98 | Albania1995 | 0.619 | 2424499009 | 835 | G.I. Generation |
| Albania | 1995 | male | 35-54 years | 14 | 375900 | 3.72 | Albania1995 | 0.619 | 2424499009 | 835 | Boomers |
| Albania | 1995 | female | 25-34 years | 7 | 264000 | 2.65 | Albania1995 | 0.619 | 2424499009 | 835 | Generation X |
| Albania | 1995 | female | 35-54 years | 8 | 356400 | 2.24 | Albania1995 | 0.619 | 2424499009 | 835 | Boomers |
| Albania | 1995 | male | 5-14 years | 6 | 376500 | 1.59 | Albania1995 | 0.619 | 2424499009 | 835 | Millenials |
| Albania | 1995 | female | 55-74 years | 2 | 180400 | 1.11 | Albania1995 | 0.619 | 2424499009 | 835 | Silent |
| Albania | 1995 | female | 5-14 years | 2 | 348700 | 0.57 | Albania1995 | 0.619 | 2424499009 | 835 | Millenials |
| Albania | 2000 | male | 25-34 years | 17 | 232000 | 7.33 | Albania2000 | 0.656 | 3632043908 | 1299 | Generation X |
| Albania | 2000 | male | 55-74 years | 10 | 177400 | 5.64 | Albania2000 | 0.656 | 3632043908 | 1299 | Silent |
| Albania | 2000 | female | 75+ years | 2 | 37800 | 5.29 | Albania2000 | 0.656 | 3632043908 | 1299 | G.I. Generation |
| Albania | 2000 | male | 75+ years | 1 | 24900 | 4.02 | Albania2000 | 0.656 | 3632043908 | 1299 | G.I. Generation |
| Albania | 2000 | female | 15-24 years | 6 | 263900 | 2.27 | Albania2000 | 0.656 | 3632043908 | 1299 | Generation X |
| Albania | 2000 | male | 15-24 years | 5 | 240000 | 2.08 | Albania2000 | 0.656 | 3632043908 | 1299 | Generation X |
| Albania | 2000 | female | 35-54 years | 5 | 332200 | 1.51 | Albania2000 | 0.656 | 3632043908 | 1299 | Boomers |
| Albania | 2000 | female | 25-34 years | 3 | 245800 | 1.22 | Albania2000 | 0.656 | 3632043908 | 1299 | Generation X |
| Albania | 2000 | male | 35-54 years | 4 | 374700 | 1.07 | Albania2000 | 0.656 | 3632043908 | 1299 | Boomers |
| Albania | 2000 | male | 5-14 years | 1 | 374900 | 0.27 | Albania2000 | 0.656 | 3632043908 | 1299 | Millenials |
| Albania | 2000 | female | 5-14 years | 0 | 324700 | 0 | Albania2000 | 0.656 | 3632043908 | 1299 | Millenials |
| Albania | 2000 | female | 55-74 years | 0 | 168000 | 0 | Albania2000 | 0.656 | 3632043908 | 1299 | Silent |
| Albania | 2005 | female | 15-24 years | 0 | 281922 | 0 | Albania2005 | 0.695 | 8158548717 | 2931 | Millenials |
| Albania | 2005 | female | 25-34 years | 0 | 190745 | 0 | Albania2005 | 0.695 | 8158548717 | 2931 | Generation X |



**Step 3: Giving proper and appropriate column names.**

The dataset does not have proper columns so our next step would be to giver proper column names to the columns wherever required.

**Step 4: Excluding the NULL values from the data.**

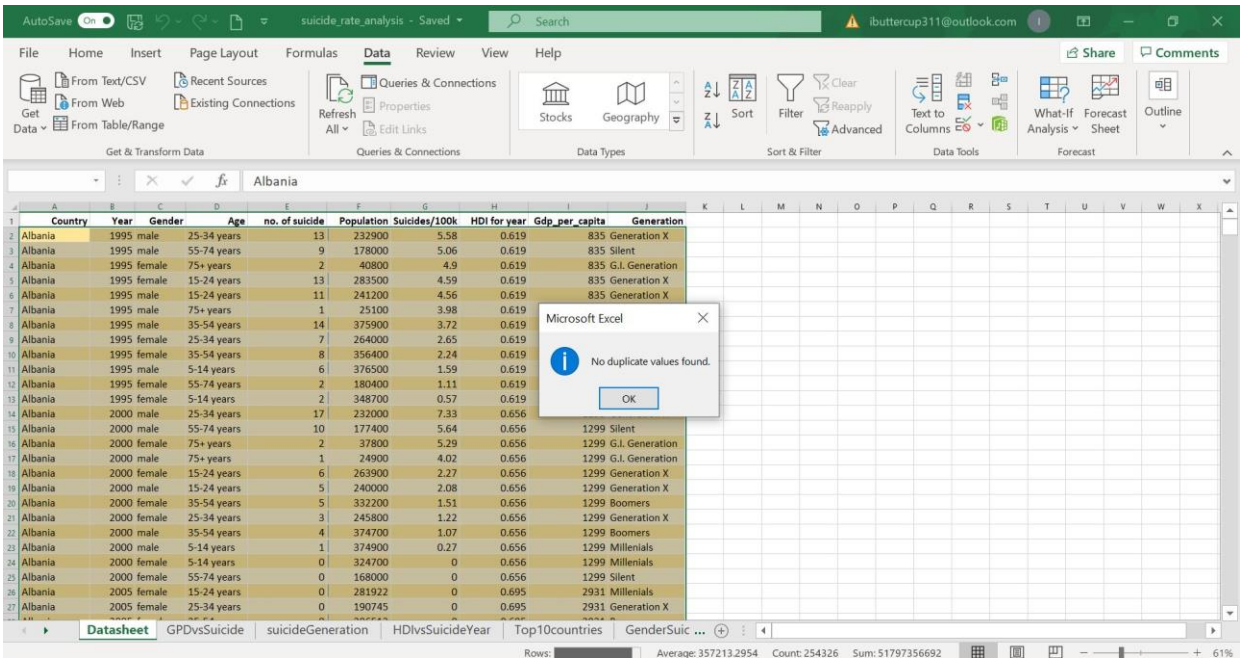
We'll be using Tableau prep for this work as it'll make the work simple and faster because we might not know how many null values could be there in this huge data set. Tableau helps us doing one step cleaning with ease.

### Step 5: Improvising Proper Data Formatting

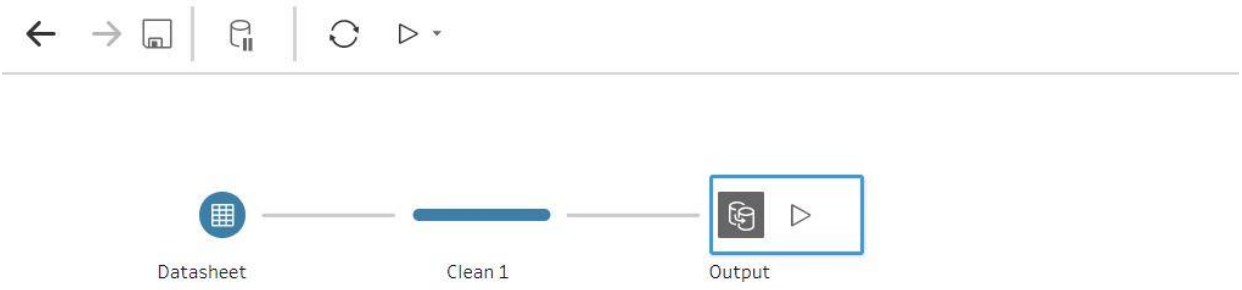
Without proper Data Formatting, proper analysis will not take place. So, we will bring down certain columns to their proper format. For example, the dates should be in the date format and price and sales should be in currency format for better results.

### Step 6: Removing Duplicate Values if Any

It might be possible that our data may be containing duplicate values which may hinder in precise analysis. So, our last task in ETL will be removing duplicate values and making our data perfect for analysis.



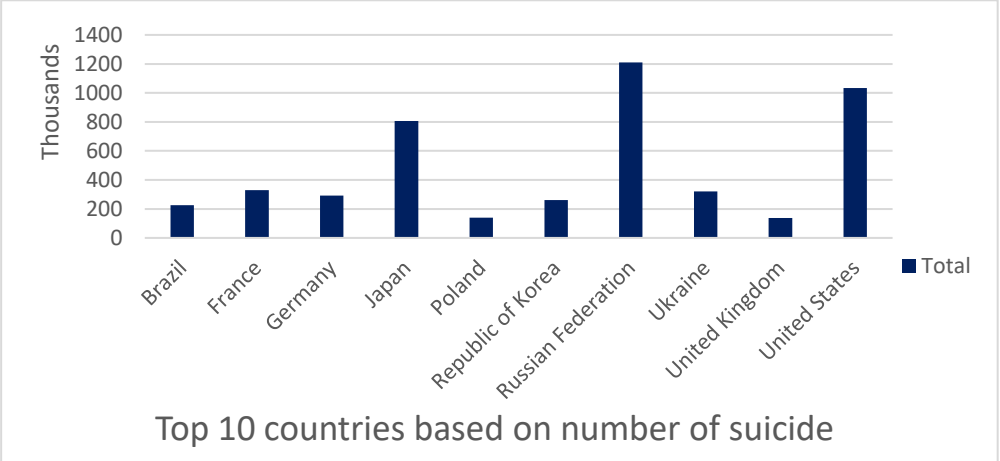
**Cleaned Data after ETL process in Tableau prep:**



# ANALYSIS OF DATASET:

1. **Trend in Suicides over the years in different countries:**

   Analyzing the number of suicides in each country over the country. As it will be congested to display all the values, only the top 10 countries are displayed, where as the sum of suicides in other country can be checked using slicer.

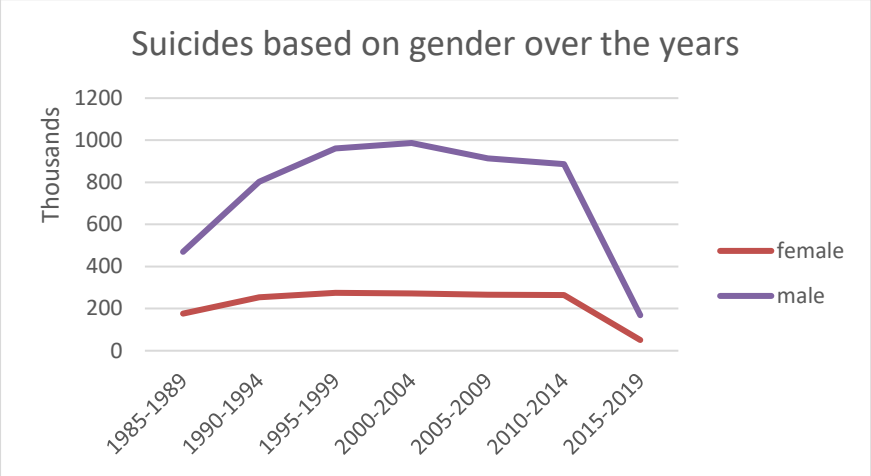



| Countries | Sum of suicides_no |
|---|---|
| Brazil | 226613 |
| France | 329127 |
| Germany | 291262 |
| Japan | 806902 |
| Poland | 139098 |
| Republic of Korea | 261730 |
| Russian Federation | 1209742 |
| Ukraine | 319950 |
| United Kingdom | 136805 |
| United States | 1034013 |
| **Grand Total** | **4755242** |

**Observation:** From the above graph, **Russian Federation** has the highest number of suicides over the years 1985-2016

2. **Trend in Suicides over the years by Gender:**

Analyzing the number of suicides in each country over the years (grouped by 5 years) based on **gender**. In general, men are more likely to become alcoholic due to stress or use lethal way to end their lives, whereas women have strong religious orientations it is expected to see higher number of suicide cases in the higher gender population case.

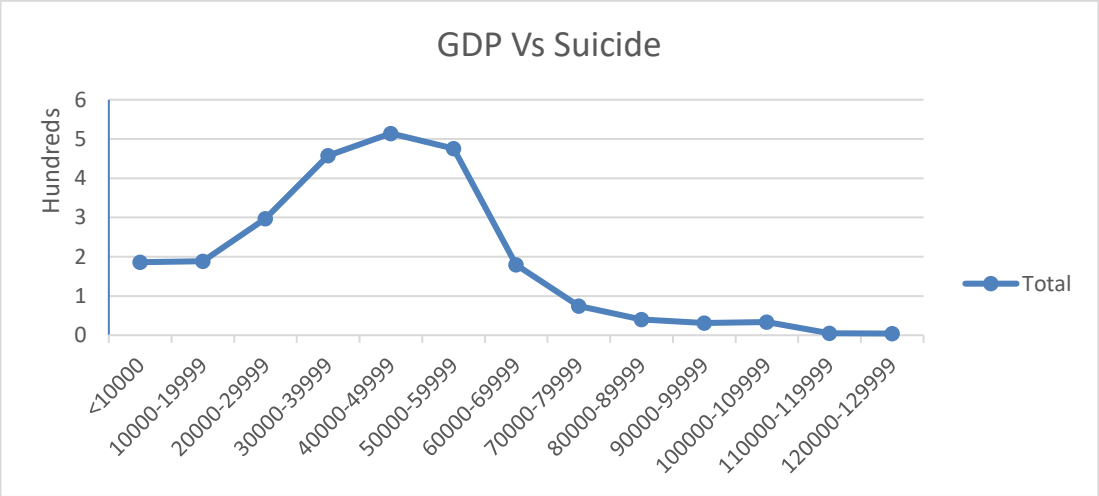



| Sum of suicides_no | Column Labels | | |
|---|---|---|---|
| Year | female | male | Grand Total |
| 1985-1989 | 175713 | 469132 | 644845 |
| 1990-1994 | 254170 | 802312 | 1056482 |
| 1995-1999 | 275059 | 961665 | 1236724 |
| 2000-2004 | 272661 | 986858 | 1259519 |
| 2005-2009 | 266291 | 913787 | 1180078 |
| 2010-2014 | 264864 | 886665 | 1151529 |
| 2015-2019 | 50752 | 168491 | 219243 |
| **Grand Total** | **1559510** | **5188910** | **6748420** |

Observation: Over the years, suicide rate is higher among the men by 2-3 times that of women. Especially in the years, 1995-2014 the number of men suicide was 5 times higher than that of men. From above it can be seen, suicide **rate is higher among the male population.**

3. **How GDP (gross domestic product) affects the suicide rate:**

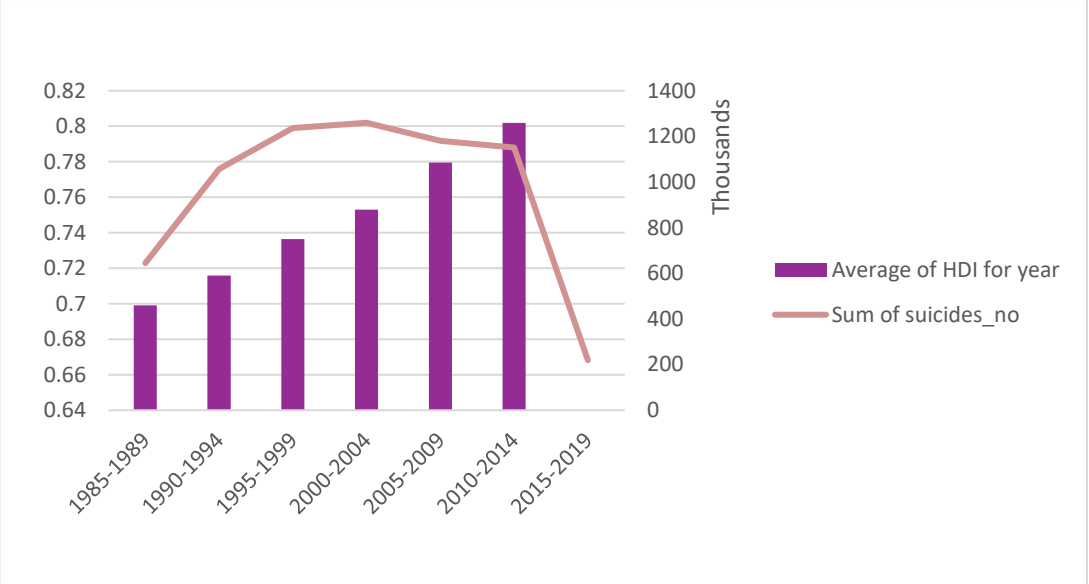
The relationship between GDP per capita and suicide may follow an inverted U-shaped curve, with suicide trends declining after peaking at a certain threshold of economic development. Thus, although at low GDP levels, increases in GDP are associated with increases in suicide rates, once a given threshold (which depends upon various factors like social, economic and cultural differences) of economic development is reached, further increases in GDP do not correlate with further increases in suicide rates




| GDP per capita | Average of suicides_no |
|---|---|
| <10000 | 186 |
| 10000-19999 | 188 |
| 20000-29999 | 297 |
| 30000-39999 | 458 |
| 40000-49999 | 514 |
| 50000-59999 | 476 |
| 60000-69999 | 180 |
| 70000-79999 | 74 |
| 80000-89999 | 40 |
| 90000-99999 | 31 |
| 100000-109999 | 34 |
| 110000-119999 | 5 |
| 120000-129999 | 4 |
| **Grand Total** | **243** |

**Observation:** GDP per capita strongly correlates to suicide rates worldwide, and the direction and magnitude of the correlation differs between developing and developed countries

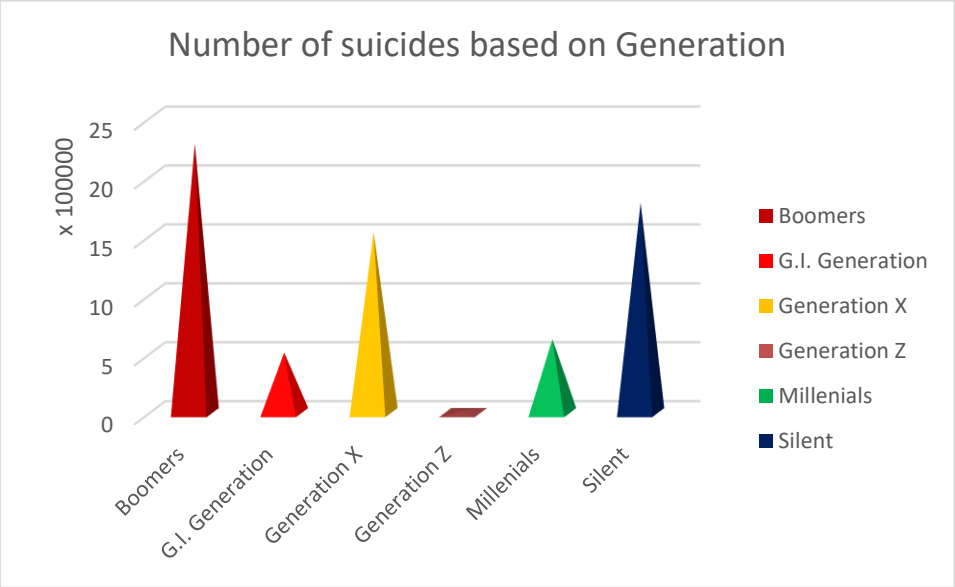
4. **How HDI (Human development index) is affected by the suicide rate:**




| Row Labels | Average of HDI for year | Sum of suicides_no |
|---|---|---|
| 1985-1989 | 0.699162162 | 644845 |
| 1990-1994 | 0.7158 | 1056482 |
| 1995-1999 | 0.736428571 | 1236724 |
| 2000-2004 | 0.752960526 | 1259519 |
| 2005-2009 | 0.779434211 | 1180078 |
| 2010-2014 | 0.801962025 | 1151529 |
| 2015-2019 | | 219243 |
| **Grand Total** | **0.776601148** | **6748420** |

**Observation:** The suicide rates increased with increasing levels of HDI. Especially in developed countries, where HDI is high the suicide rate is higher. It could be because of the very big financial difference among the poor and the rich.

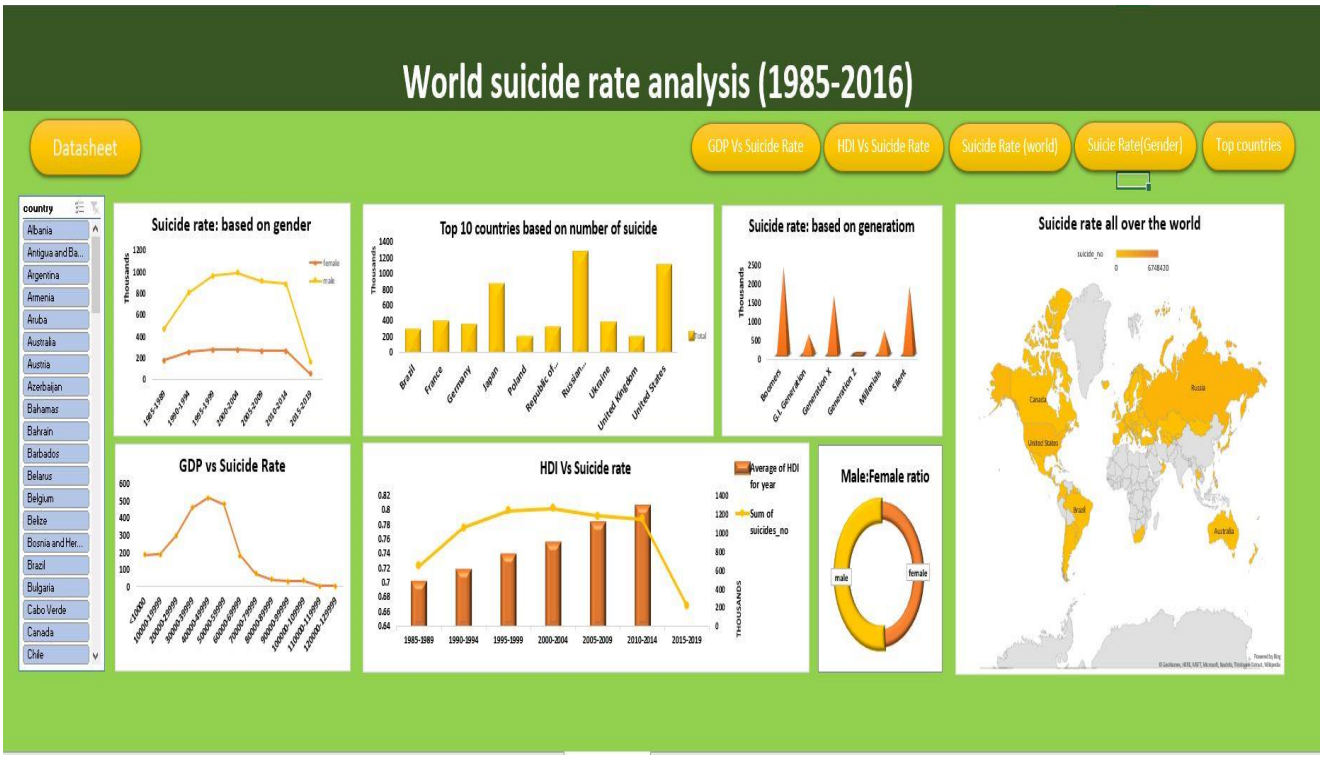
5. **Average number of suicides of each generation in different countries**




- Boomers : 1946 to 1964
- G.I.Generation: 1901 to 1927
- Generation Z: 1997 to 2012
- Silent: 1928 to 1945
- Generation X: 1965 to 1980
- Millenials: 1981 to 1996

**Observation:** The number of suicides is higher in the **Boomers** (post war years), followed by Silent generation where people were affected a lot, both mentally and financially due to war which led to depression, having fewer children and many such effects.

# DASHBOARD:


World suicide rate analysis (1985-2016)
Datasheet
GDP Vs Suicide Rate
HDI Vs Suicide Rate
Suicide Rate (world)
Suicie Rate(Gender)
Top countries
country
Albania
Antigua and Ba...
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Barbados
Belarus
Belgium
Belize
Bosnia and Her...
Brazil
Bulgaria
Cabo Verde
Canada
Chile
Suicide rate: based on gender
Thousands
female
male
1985-1989
1990-1994
1995-1999
2000-2004
2005-2009
2010-2014
2015-2019
Top 10 countries based on number of suicide
Thousands
Brazil
France
Germany
Japan
Poland
Republic_of_...
Russian...
Ukraine
United Kingdom
United States
total
Suicide rate: based on generatiom
Thousands
Boomers
G.I. Generation
Generation X
Generation Z
Millenials
Silent
Suicide rate all over the world
suicide_no
0
6748430
Russia
Canada
United States
Brazil
Australia
GDP vs Suicide Rate
<10000
10000-19999
20000-29999
30000-39999
40000-49999
50000-59999
60000-69999
70000-79999
80000-89999
90000-99999
100000-109999
110000-119999
120000-125999
HDI Vs Suicide rate
Average of HDI for year
Sum of suicides_no
THOUSANDS
1985-1989
1990-1994
1995-1999
2000-2004
2005-2009
2010-2014
2015-2019
Male:Female ratio
male
female

# References


- YouTube
- EasyExcel