Data Visualization Assignment- R

Name: Abhishek Dubey

Student ID: D20123718

Full time: MSc Data Science 2020-21

Class: TU59 Full Time

Date: 15/12/2020

Analysis of Suicide Trends Globally

Title: Analysis of Suicide Trends Globally

Introduction:

Suicide is the act of causing your own death intentionally, there are so many reasons for it. We are not discussing the reasons here but we will discuss how is this trend growing in numbers.

As per the sources 8,28,000 cases reported in 2015 in the world, which is almost 7,12,000 more cases from 1990.

This makes suicide 10th leading cause of death worldwide.

From 1985 to 2016 we will see suicide rates with respect to every country. And further we will see the relation of rich countries with the suicide cases.

My Goal

My goal is to only provide awareness to people about these cases from the visualization I prepared. so that we can safe more human being.

STOP SUICIDE

Problem:

Suicide is the major problem since long back, all governments are taking steps to control it. Now in this analysis we will see the global trend of suicide cases from 1985 to 2016.

We are using global suicide data per country from World Health Organization. And also, we are using country per capita data from World Bank to analyze which countries per capita shows what relations to suicide cases?

Which Rich countries shows maximum suicide rate till date? This will lead to many more questions like why in rich countries suicide rates are so high?

What is the global trend of suicide cases each year from 1985 to 2015?

Description of intended audience:

Governments use these visualizations and can check the stand of their country in global cases. compare suicide cases between countries so that government can implement awareness program to overcome these cases.

Health Ministry of every country can visualize these insights and can implement consultation program organized by different countries to reduce this rate. Providing psychologist one to one consultation to the people who feel stress. According to the study 99% cases of suicide are because of some kind of stress. And lots of governments are organizing different consultations program so that people can feel open and these programs shows best results.

Research Students can use this study and visualizations to organize some specific kind of technique to overcome these cases.

General population should check these visualizations so that they scan help their friends, relatives, family person by providing extra support and care.

Dataset

I am using 2 datasets from public resources

Dataset 1:

Contains data related to suicide cases by country every year from 1985 to 2016

This data set taken from World Health Organization

Reference:

World Health Organization. (2018). Suicide prevention.

Retrieved from http://www.who.int/mental_health/suicide-prevention/en/

	country	year	sex	age	suicides_no	population	suicides/100k pop
1	Albania	1987	male	15-24 years	21	312900	6.71
2	Albania	1987	male	35-54 years	16	308000	5.19
3	Albania	1987	female	15-24 years	14	289700	4.83
4	Albania	1987	male	75+ years	1	21800	4.59
5	Albania	1987	male	25-34 years	9	274300	3.28
6	Albania	1987	female	75+ years	1	35600	2.81
7	Albania	1987	female	35-54 years	6	278800	2.15
8	Albania	1987	female	25-34 years	4	257200	1.56
9	Albania	1987	male	55-74 years	1	137500	0.73
10	Albania	1987	female	5-14 years	0	311000	0.00

Dataset 2:

Contains data related to per capita income of country from 1985 to 2016

This data will be helpful in analyzing which countries are very rich and which are poor.

This dataset taken from World Bank

Reference:

World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016.

Retrieved from http://databank.worldbank.org/data/source/world-development-indicators#

	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
1	Albania1987	NA	2156624900	796	Generation X
2	Albania1987	NA	2156624900	796	Silent
3	Albania1987	NA	2156624900	796	Generation X
4	Albania1987	NA	2156624900	796	G.I. Generation
5	Albania1987	NA	2156624900	796	Boomers
6	Albania1987	NA	2156624900	796	G.I. Generation
7	Albania1987	NA	2156624900	796	Silent
8	Albania1987	NA	2156624900	796	Boomers
9	Albania1987	NA	2156624900	796	G.I. Generation
10	Albania1987	NA	2156624900	796	Generation X

PRE- PROCESSING, CLEANING AND WRANGLING OF DATA

1. Both datasets are uploaded in my GitHub so that we can use them directly and anywhere. So, no need to import from computer as data is already in cloud.

   Country_df = Data related to country with per capita income from 1985-2016

   Country_df<- read_csv("https://raw.githubusercontent.com/Abhidubey96/Analysis-of-Suicide-Trends-Globally-in-R/main/Country%20Data.csv")

   Suicide_df = Data related to suicide cases by each country from 1985-2016

   Suicide_df <- read_csv("https://raw.githubusercontent.com/Abhidubey96/Analysis-of-Suicide-Trends-Globally-in-R/main/Suicide%20Data%20by%20Country.csv")

2. After importing them in R Studio we will merge them in final data frame.

   df <- cbind(Suicide_df, Country_df)

	country	year	sex	age	suicides_no	population	gdp_per_year	gdp_per_capita	generation	continent
1	Albania	1987	male	15-24 years	21	312900	2156624900	796	Generation X	Europe
2	Albania	1987	male	35-54 years	16	308000	2156624900	796	Silent	Europe
3	Albania	1987	female	15-24 years	14	289700	2156624900	796	Generation X	Europe
4	Albania	1987	male	75+ years	1	21800	2156624900	796	G.I. Generation	Europe
5	Albania	1987	male	25-34 years	9	274300	2156624900	796	Boomers	Europe
6	Albania	1987	female	75+ years	1	35600	2156624900	796	G.I. Generation	Europe
7	Albania	1987	female	35-54 years	6	278800	2156624900	796	Silent	Europe
8	Albania	1987	female	25-34 years	4	257200	2156624900	796	Boomers	Europe
9	Albania	1987	male	55-74 years	1	137500	2156624900	796	G.I. Generation	Europe
10	Albania	1987	female	5-14 years	0	311000	2156624900	796	Generation X	Europe

3. Removing variable “HDI per Year” from data frame as variable is showing 75% missing values. No need of this variable

4. Removing variable “suicides/100k pop” from data frame as field calculated is wrong. We will recalculate this later and use it in our visualization

5. Renaming some variables for our statistics like “gdp_per_year”, “gdp_per_capita”, “country-year”

```
# Removing unnecessary columns
# Renaming the variable names
df <- df %>%
  select(-c(`HDI for year`, `suicides/100k pop`)) %>%
  rename(gdp_per_year = `gdp_for_year ($)`,
         gdp_per_capita = `gdp_per_capita ($)`,
         country_year = `country-year`) %>%
  as.data.frame()

View(df)
```

```
'data.frame':   27820 obs. of  10 variables:
 $ country       : chr  "Albania" "Albania" "Albania" "Albania" ...
 $ year          : num  1987 1987 1987 1987 1987 ...
 $ sex           : chr  "male" "male" "female" "male" ...
 $ age           : chr  "15-24 years" "35-54 years" "15-24 years" "75+ years" ...
 $ suicides_no   : num  21 16 14 1 9 1 6 4 1 0 ...
 $ population    : num  312900 308000 289700 21800 274300 ...
 $ country_year  : chr  "Albania1987" "Albania1987" "Albania1987" "Albania1987" ...
 $ gdp_per_year  : num  2.16e+09 2.16e+09 2.16e+09 2.16e+09 2.16e+09 ...
 $ gdp_per_capita: num  796 796 796 796 796 796 796 796 796 796 ...
 $ generation    : chr  "Generation X" "Silent" "Generation X" "G.I. Generation" ...
```

Extensive Cleaning in Data:

6. As we can see in our data frame that every year should have 12 observations. As per 6 age groups and 2 genders. So, we need 12 for every year. But 2016 year shows less entries. So, we are not taking it further, we will remove 2016 year as data is very less.

```
df <- df %>%
  filter(year != 2016) %>%
  select(-country_year)
```

7. In our final data frame, we need continent data so we will use **country code library** to impute the continent field

```
# Making Continent variable by using country code library
df$continent <- countrycode(sourcevar = df[, "country"],
                            origin = "country.name",
                            destination = "continent")
```

8. Some variables have wrong datatypes so we will correct the datatype for age to be **Ordinal using factor function**
9. Similarly, variable generation has wrong data type so we will correct it with **Ordinal datatype using factor function**

** for code please check .Rmd File attached

```
df$age <- factor(df$age, ordered = T,
                 levels = c("5-14 years",
                            "15-24 years",
                            "25-34 years",
                            "35-54 years",
                            "55-74 years",
                            "75+ years"))

df$generation <- factor(df$generation,
                        ordered = T,
                        levels = c("G.I. Generation",
                                   "Silent",
                                   "Boomers",
                                   "Generation X",
                                   "Millenials",
                                   "Generation Z"))
```

10. Calculating global suicide rates over time from 1985 to 2015

*Global <- (sum(as.numeric(df$suicides_no)) / sum(as.numeric(df$population))) * 100000*

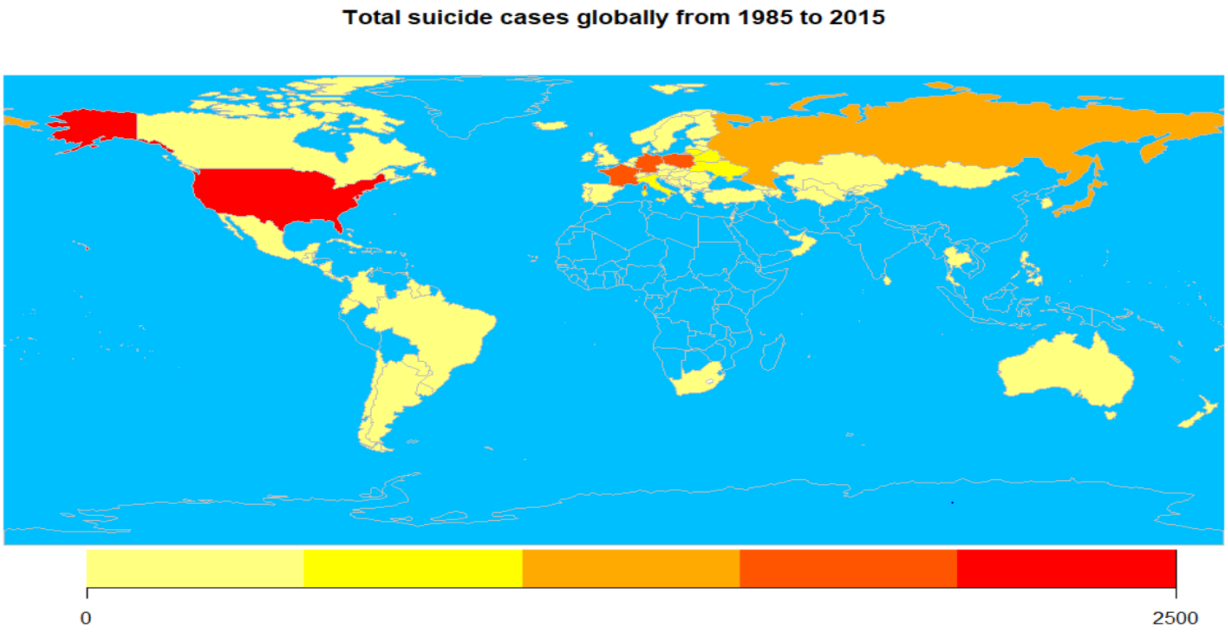
VISUALIZATION 1

What will be the World heat map according to the suicide per 100k population for each country?

Also, which countries and continents shows higher cases?

To achieve this goal first we make normal world map with total suicide cases.

Iteration 1




```
countrydata <- joinCountryData2Map(df, joinCode = "NAME", nameJoinColumn = "country")

mapCountryData(countrydata,
nameColumnToPlot="suicides_no",
colourPalette = "heat",
mapTitle="Total suicide cases globally from 1985 to 2015",
mapRegion = "world",
oceanCol="deepskyblue",
catMethod = "pretty")
```

This Graph Develops the confusion about the countries and their suicide cases.

Countries which are not present in dataset shows no colour, which makes hard for audience to analyze.

Also, we need to calculate the suicide cases per 100k population.

Iteration 2:

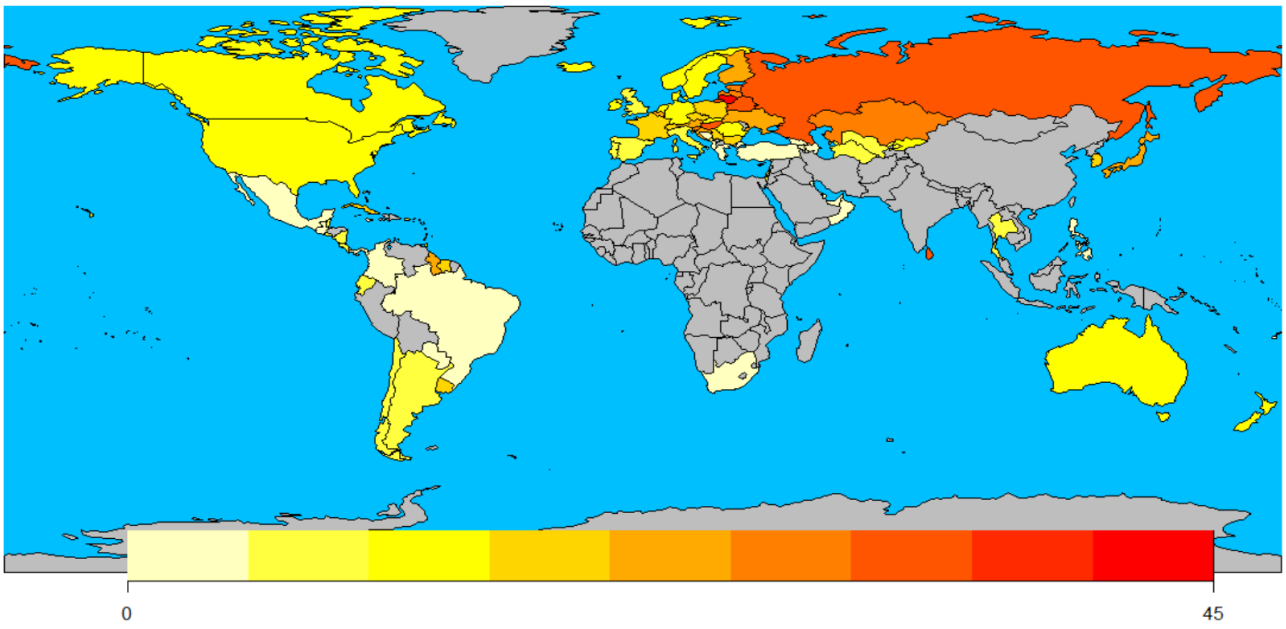
Calculating suicide per 100k population

*Variable "suicide_per_100k" = (sum(as.numeric(suicides_no)) / sum(as.numeric(population))) * 100000*

And giving missing countries to gray colour

missingCountryCol="gray"

Total suicide cases globally as per (suicide per 100k population). Shows overall cases by country from 1985 to 2015



Map shows countries in colour from white to dark red according to the suicide cases per 100k population of the country.

Scale is there from 0 to 45.

Being 0 be lowest in suicide cases per 100k population

Being 45 means highest in suicide cases per 100k population

**Some of the countries are gray means their data is missing, they are not part of this analysis.

Some countries from ASIA and AFRICA doesn't have sufficient data for analysis.

Blue colour shows the ocean.

Insights from Map:

1. Russia and Lithuania show highest number of suicide cases per 100k population from 1985 to 2015. Margin is very close to 45 index.
2. Because of insufficient data from some continents, we can say majority cases are coming from Europe. But is we have more data then this statement will be wrong.

Code for Map

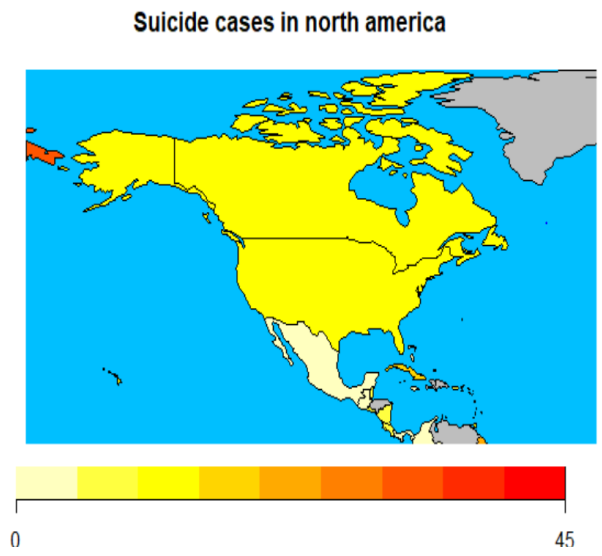
```
region <- df %>%
  group_by(country) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) /
sum(as.numeric(population))) * 100000)

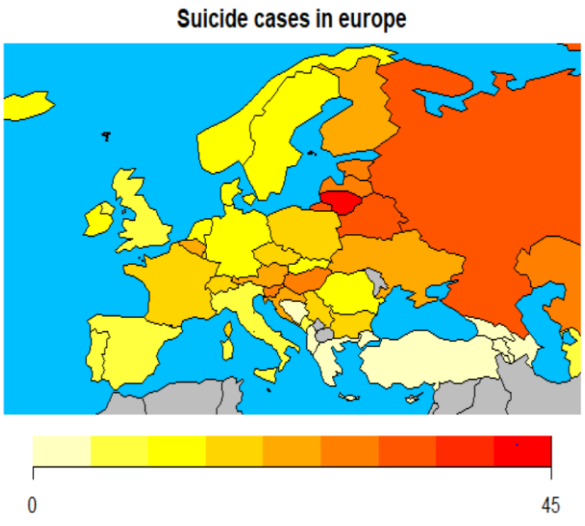
countrydata <- joinCountryData2Map(region, joinCode = "NAME", nameJoinColumn =
"country")

par(mar=c(0, 0, 0, 0)) # margins


mapCountryData(countrydata,
nameColumnToPlot="suicide_per_100k",
colourPalette = "heat",
mapTitle="Total suicide cases globally as per (suicide per 100k population). Shows
overall cases by country from 1985 to 2015",
mapRegion = "world",
oceanCol="deepskyblue",
missingCountryCol="gray",
borderCol = "black",
catMethod = "pretty")
```

As we can see most of the cases are coming from Europe and north America
Visualizing suicide per 100k population cases in that region
MapRegion = “north America” MapRegion = “europe”





VISUALIZATION 2:

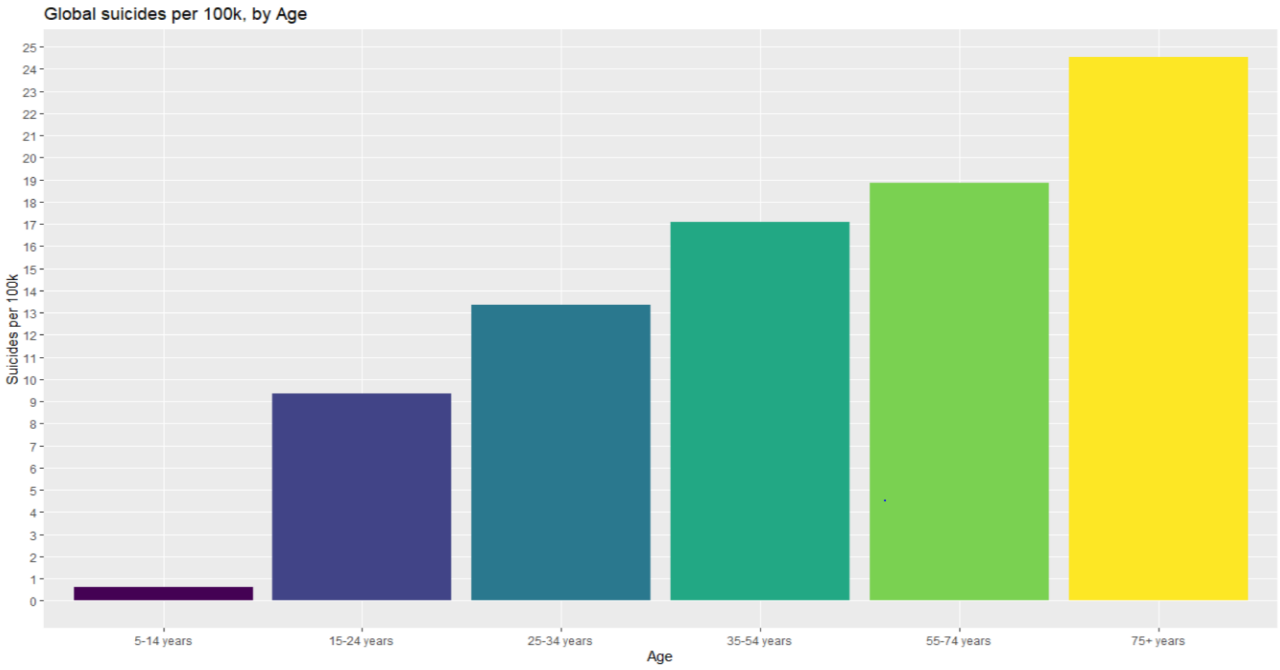
What is the relation between age and generation over years from 1985 to 2015 suicide cases per 100k population?

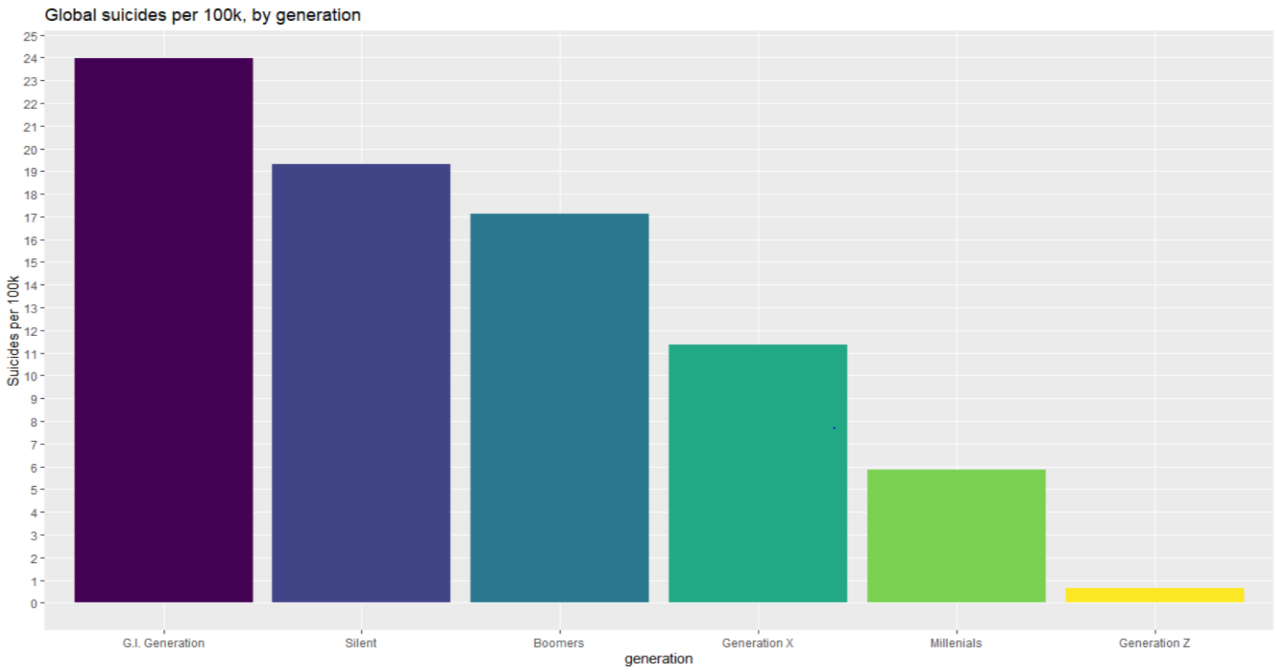
Is it True that new generation suicide cases globally are growing rapidly?

To Answer above question, we will follow below iterations

Iteration 1

Analyzing age variable with suicide cases per 100k population. To identify which age group shows more cases

```
visualization for age varibale
```{r}
df %>%
  group_by(age) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) /
sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = age, y = suicide_per_100k, fill = age)) +
  geom_bar(stat = "identity") +
  labs(title = "Global suicides per 100k, by Age",
       x = "Age",
       y = "Suicides per 100k") +
  theme(legend.position = "none") +
  scale_y_continuous(breaks = seq(0, 30, 1), minor_breaks = F)
```
```



This shows age group 75+ years having the greatest number of suicide cases. And age group 5-14 years shows a smaller number of cases.

Now we have relation between age and suicide pe 100k population globally

Let's find out the relation between generation and suicide cases

Iteration 2:

Let's Analyze the relation between generation with suicide per 100k population

```
visualization of generation varibale
```{r}
df %>%
  group_by(generation) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) /
sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = generation, y = suicide_per_100k, fill = generation)) +
  geom_bar(stat = "identity") +
  labs(title = "Global suicides per 100k, by generation",
       x = "generation",
       y = "Suicides per 100k") +
  theme(legend.position = "none") +
  scale_y_continuous(breaks = seq(0, 30, 1), minor_breaks = F)
```
```



For Knowledge purpose generation groups belongs to:

G I Gen: Born 1996 – TBD

Gen Z: Current

Millennials: Born 1977 – 1995

Generation X: Born 1965 – 1976

Boomers: Born 1946 – 1964

Silent: Born 1945 and before

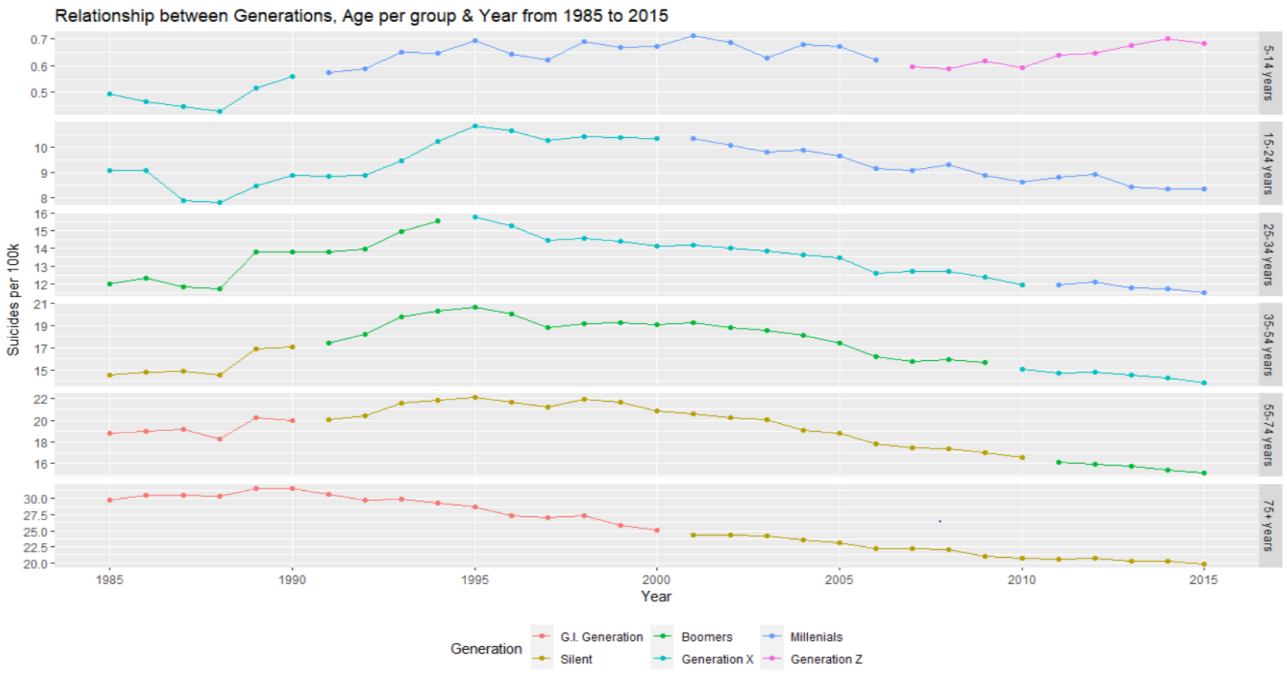
Insights:

- Mostly cases are from people who born between 1996 to TBD
- Least number of cases by Generation Z, but there is a flaw in dataset Generation Z also comes under GI Generation

Let's visualize it by combining both age and generation on the basis of year 1985 to 2015.

Iteration 3:

The analysis shows the answer of 1st question the relation between all.



Insights from visualization 2

- This shows 75+ age group is decreasing the number of cases; however, age group 5-14 years shows recently hike in suicide cases.
- Generation Millennials: Born between 1977 and 1995 shows the greatest number of suicide cases in all over the world per 100k population

Answer of 2nd question in Visualization 2

- **Overall, all age groups are going down and Generation Z which are current generation is slightly going up in suicide cases in 2015 rising from 2007.**

VISUALIZATION 3:

What is the global trend of suicide cases per 100k population from 1985 to 2015?

Is there any relation between countries having high gdp per capita income showing a greater number of suicide cases?

To Analyse this, we will perform following iterations

Basically, countries which have high per capita income as per world bank data from 1985 – 2015

Iceland
Luxembourg
Norway
Switzerland
United States

Iteration 1

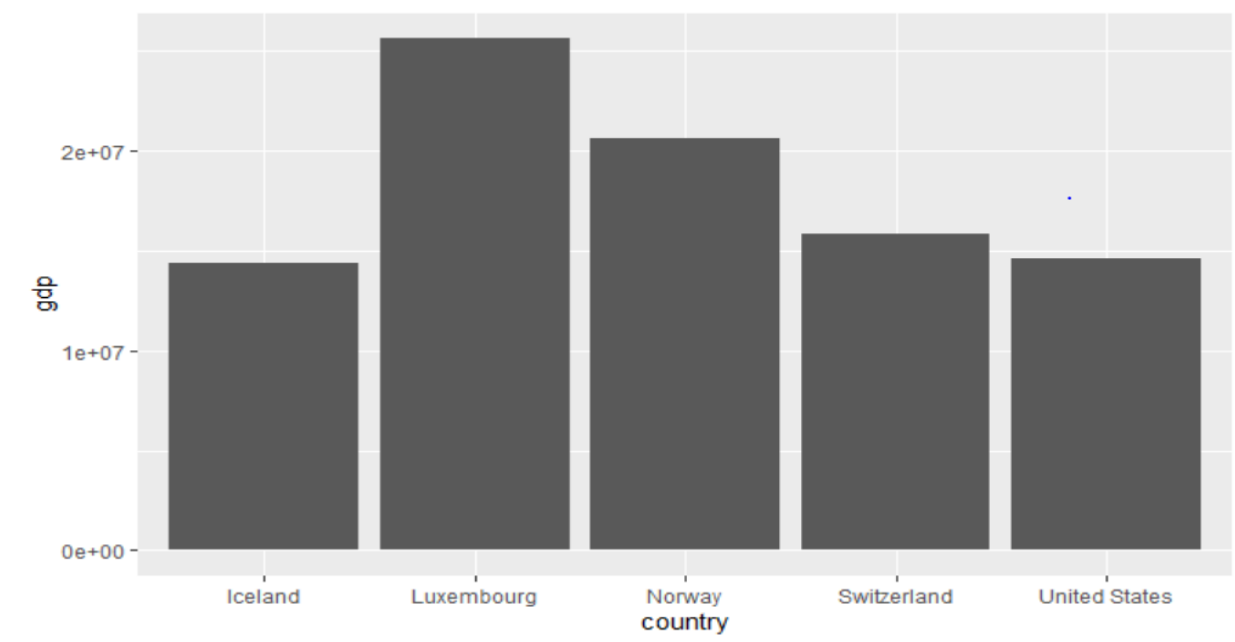
Finding 5 countries with high gdp per capita income

Analyzing countries with high gdp per capita income

```{r}
library(sqldf)

gdp_df <-sqldf("select country, sum(gdp_per_capita) as gdp from df group by country")
gdp_df

# finding 5 countries with max gdp per capita income

gdp_df1 <-sqldf("select country, gdp from gdp_df where gdp > 14300000")
gdp_df1

```

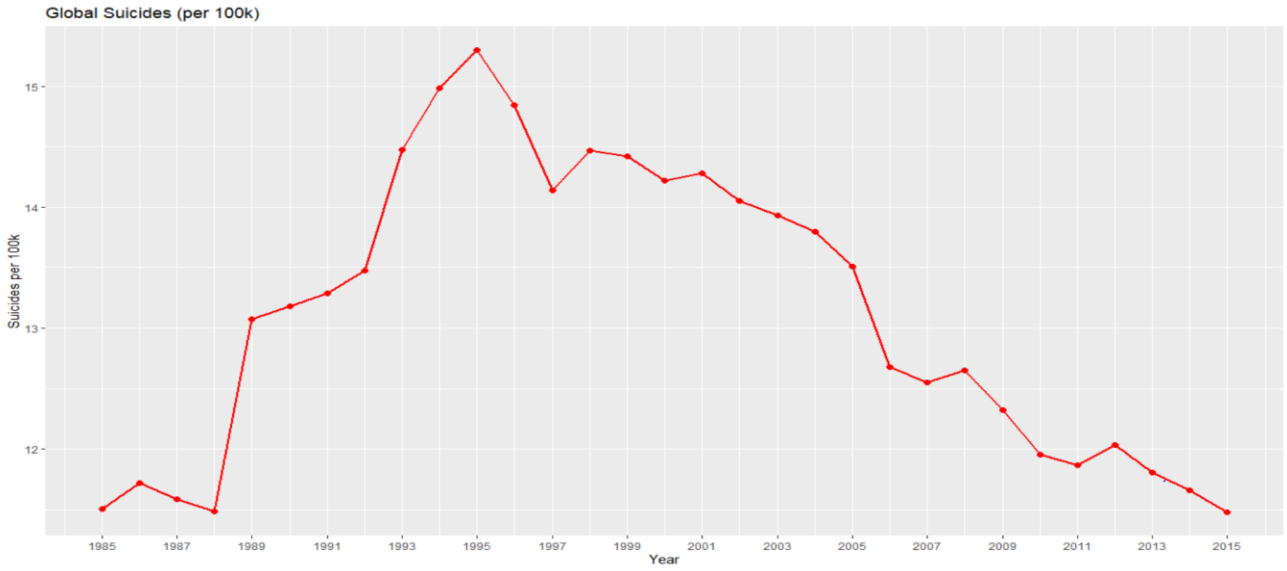
country <chr>	gdp <dbl>
Iceland	14355876
Luxembourg	25593000
Norway	20635056
Switzerland	15871404
United States	14608296

```{r}
ggplot(data=gdp_df1, aes(x= country, y=gdp)) +
geom_bar(stat="identity")
```



Iteration 2

Analysing global trend of suicide cases over time from 1985 to 2015



Insights:

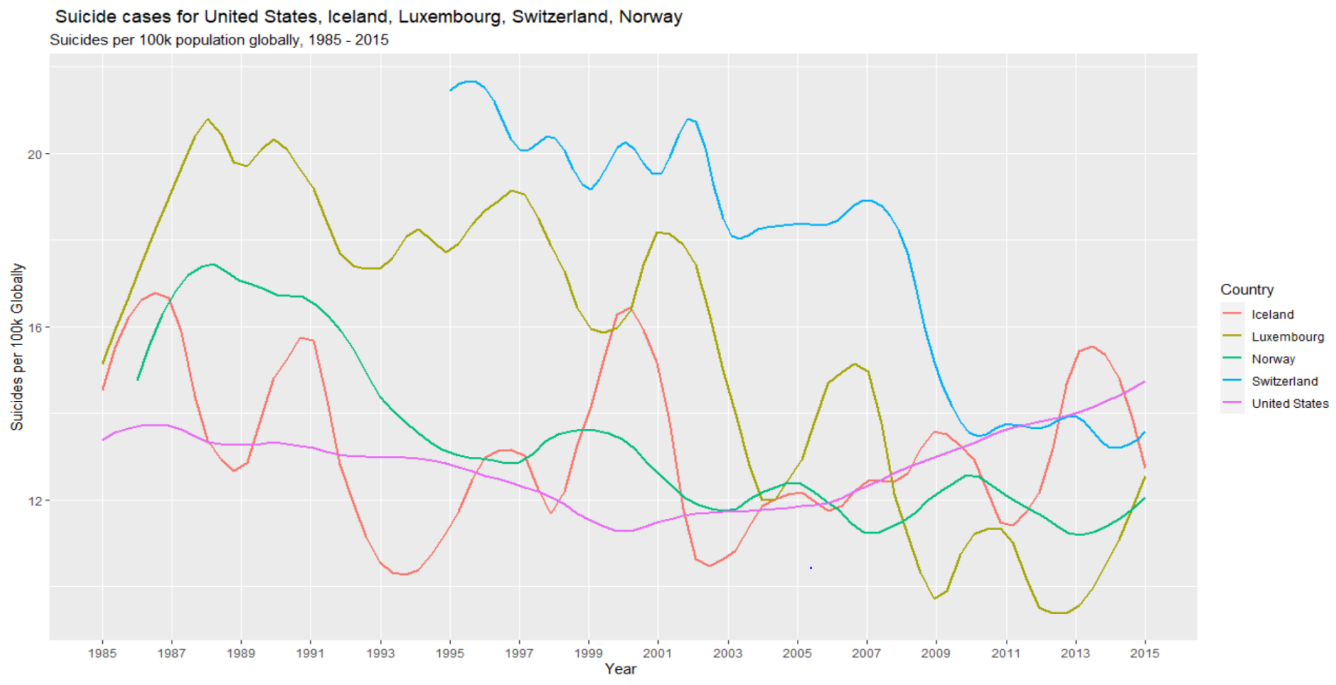
- World suicide cases are going down, which is very good
- Suicide cases has peak on 1995.

Iteration 3

Final Visualization:

Only specific to high GDP per capita income countries with suicide cases per 100k population over a period from 1985 to 2015

```
df_specific <- df %>%
  filter(country %in% c("Iceland",
                        "Luxembourg",
                        "Norway",
                        "Switzerland",
                        "United States"))

df_specific %>%
  group_by(country, year) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) /
sum(as.numeric(population))) * 100000) %>%
  ggplot(aes(x = year, y = suicide_per_100k, col = country)) +
  geom_smooth(se = F, span = 0.2) +
  scale_x_continuous(breaks = seq(1985, 2015, 2), minor_breaks = F) +
  labs(title = " Suicide cases for United States, Iceland, Luxembourg, Switzerland,
Norway",
       subtitle = "Suicides per 100k population globally, 1985 - 2015",.
       x = "Year ",
       y = "Suicides per 100k Globally",
       col = "Country")
```



Insights:

Starting from 1985 all developed nations with high GDP showing suicide cases per 100k population to be nearly 15. Data for Norway comes under consideration in 1995 as it is showing very large number of suicide cases. But there is significant drop in suicide cases over a period of time.
There are so many online articles on Norway suicide cases.
Refer:
https://tidsskriftet.no/en/2019/08/kronikk/why-suicide-rate-not-declining-norway

Study shows:
Most people who did suicides are not in psychiatric treatment and societal conditions are a main reason why the suicide rate is not falling.
As we discussed in introduction that 99% world wide reported is because of stress and now governments are taking action by providing one to one consultation with psychiatrist.

Similarly, condition showing in Luxembourg in 19's Luxembourg shows large number of suicide cases, but by implanting consultations programs in various cities at free of cost. Country shows some improvement in 2009 but there is a significant rise showing from 2013-2015

Case with Iceland is very fluctuating, with small number of population Iceland shows still more suicide cases and similar case with Switzerland.

United States of America one of the developed nation having GDP 20.54 trillion USD reported in 2018. With highly skilled population despite shows significant number of increase suicide cases from 2000.

As covered in Article by BBC
Refer: https://www.bbc.com/news/world-us-canada-44416727

In 17 years, cases rise to 30%, means 16 out of 100k population taking their own life. There are so many reasons for suicide.

There is one thing reported by Prof Cerel
"Our mental health systems are just really struggling across the country," Prof Cerel says. "In terms of training mental health professionals, we're not doing a great job."

** As the purpose of making these visualization and report is only provide awareness to the people**

STOP SUICIDE