PREDICTIVE ANALYTICS AND DATA MINING

Project Name: SUICIDES ANALYSIS

Professor:

Axel Ullern

Project Team Members:

Kouadio Yao Innocent Junior
Thanh Tung Trinh
Sanjeet Maisnam
Shubham Rana
Fabrice Prince

EPITA : École Pour l'Informatique et les Techniques Avancées - 01/06/2020

Table Contents

Introduction

According to WHO (World Health Organization), every year close to 800 000 people take their own life and there are many more people who attempt suicide. Every suicide is a tragedy that affects families, communities and entire countries and has long-lasting effects on the people left behind. Suicide occurs throughout the lifespan and was the second leading cause of death among 15-29-year-olds globally in 2016.

Suicide is a serious public health problem; however, suicides are preventable with timely, evidence-based, and often low-cost interventions. For national responses to be effective, a comprehensive multisectoral suicide prevention strategy is needed. This is the reason why we choose this topic for the learning project of Predictive Analytics and Data Mining.

Dataset 1

Our team has chosen a **Suicide Rates Overview 1985 to 2016** from the website Kaggle.com. This Dataset is based on the real case study from the United Nations Development Program. (2018). Human development index (HDI), World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016, Suicide in the Twenty-First Century and World Health Organization. (2018). Suicide prevention.

```
##   ï..country            year          sex               age
##  Length:27820       Min.   :1985   Length:27820       Length:27820
##  Class :character   1st Qu.:1995   Class :character   Class :character
##  Mode  :character   Median :2002   Mode  :character   Mode  :character
##                     Mean   :2001
##                     3rd Qu.:2008
##                     Max.   :2016
##
##   suicides_no        population       suicides.100k.pop country.year
##  Min.   :    0.0   Min.   :     278   Min.   :  0.00    Length:27820
##  1st Qu.:    3.0   1st Qu.:   97498   1st Qu.:  0.92    Class :character
##  Median :   25.0   Median :  430150   Median :  5.99    Mode  :character
##  Mean   :  242.6   Mean   : 1844794   Mean   : 12.82
##  3rd Qu.:  131.0   3rd Qu.: 1486143   3rd Qu.: 16.62
##  Max.   :22338.0   Max.   :43805214   Max.   :224.97
##
##   HDI.for.year   gdp_for_year....   gdp_per_capita....  generation
##  Min.   :0.483   Length:27820       Min.   :   251     Length:27820
##  1st Qu.:0.713   Class :character   1st Qu.:  3447     Class :character
##  Median :0.779   Mode  :character   Median :  9372     Mode  :character
##  Mean   :0.777                      Mean   : 16866
##  3rd Qu.:0.855                      3rd Qu.: 24874
##  Max.   :0.944                      Max.   :126352
##  NA's   :19456
```

Content

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.

References

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

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#

[Szamil]. (2017). Suicide in the Twenty-First Century [dataset]. Retrieved from https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook

World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/

Inspiration
Suicide Prevention.

Dataset 2

The second dataset has been selected after analysis of the first dataset, A question has arisen in us to know why people committed suicides , due to that doubt we have download the second dataset of **Suicides in India** from Kaggle which give us some information about causes that might be a factor of commit suicides. This data set contains yearly suicide detail of all the states/u.t of India by various parameters from 2001 to 2012.

```
##     State                Year       Type_code            Type
##  Length:237519      Min.   :2001   Length:237519      Length:237519
##  Class :character   1st Qu.:2004   Class :character   Class :character
##  Mode  :character   Median :2007   Mode  :character   Mode  :character
##                     Mean   :2007
##                     3rd Qu.:2010
##                     Max.   :2012
##     Gender           Age_group             Total
##  Length:237519      Length:237519      Min.   :    0.00
##  Class :character   Class :character   1st Qu.:    0.00
##  Mode  :character   Mode  :character   Median :    0.00
##                                        Mean   :   55.03
##                                        3rd Qu.:    6.00
##                                        Max.   :63343.00
```

Content

Time Period: 2001 - 2012
Granularity: Yearly
Location: States and U. T’s of India

Parameters:

a) Suicide causes
b) Education status
c) By means adopted
d) Professional profile
e) Social status

Acknowledgements

National Crime Records Bureau (NCRB), Govt of India has shared this dataset under Govt. Open Data License - India.

NCRB has also shared the historical data on their website

Exploratory Data Analysis in R - Dataset 1

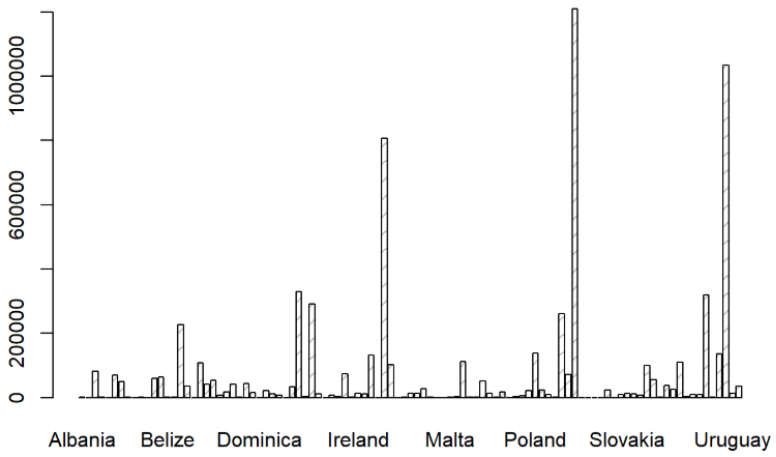
From this dataset, we will see that suicide does not just occur in high-income countries but is a global phenomenon in all regions of the world. In fact, majority of global suicides occurred in low- and middle-income countries.

Importing the dataset & library

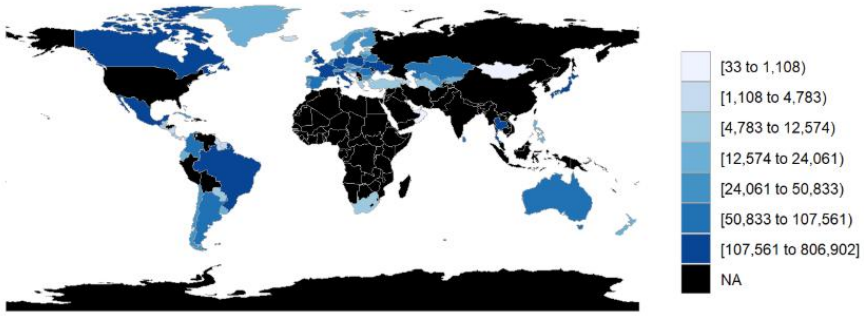
```
library(tidyverse)
library(ggplot2)
library(choroplethr)
data <-  read.csv('Master.csv')
```

Statistic of death per country

```
a <- as.data.frame( group_by(data , data$ï..country ) %>% summarise( sum (suicides_no)) )
x <-  a$`data$ï..country`
y <-  a$`sum(suicides_no)`
barplot((y),names.arg =  x,density=10)
```

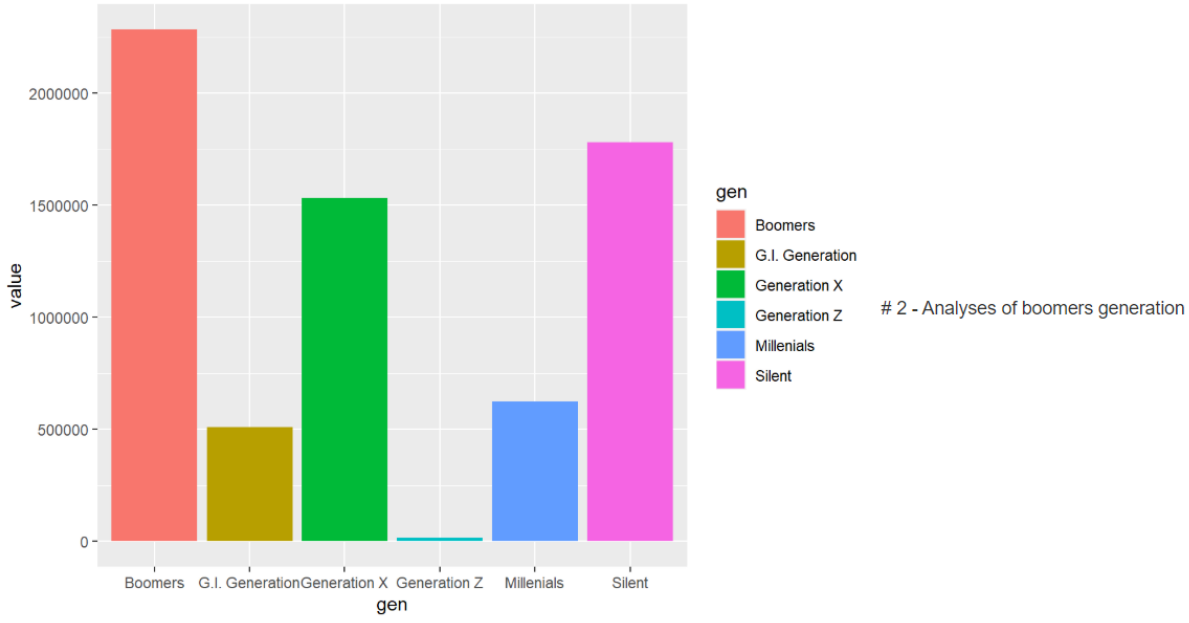


```
colnames(a) <- c("region", "value")
a$region <-  tolower(a$region)
country_choropleth(a , num_colors = 7)
```



Suicide base on generation

```
gen <- group_by(data,generation) %>% summarise(sum(suicides_no))
colnames(gen) <- c("gen", "value")
ggplot(gen , mapping = aes(x = gen,y = value,fill=gen)  ) + geom_bar(stat='identity')
```



2 - Analyses of boomers generation

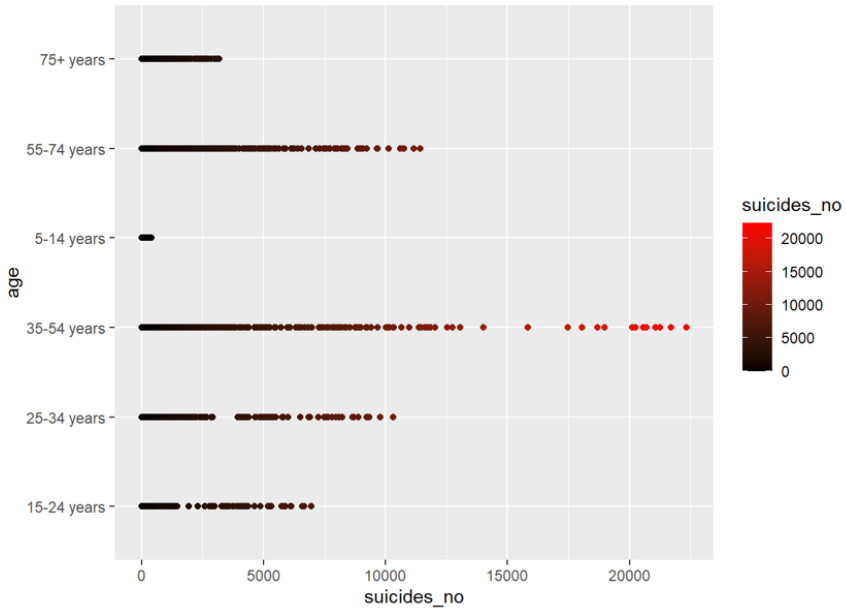
by country

Number of suicides by year

```
ggplot() + geom_point( data  = data , aes( suicides_no, year , color = suicides_no )) +
  scale_color_gradient(low="black", high="red")
```

Number of suicides by Age

```
ggplot() + geom_point( data  = data , aes( suicides_no, age , color = suicides_no )) +
  scale_color_gradient(low="black", high="red")
```

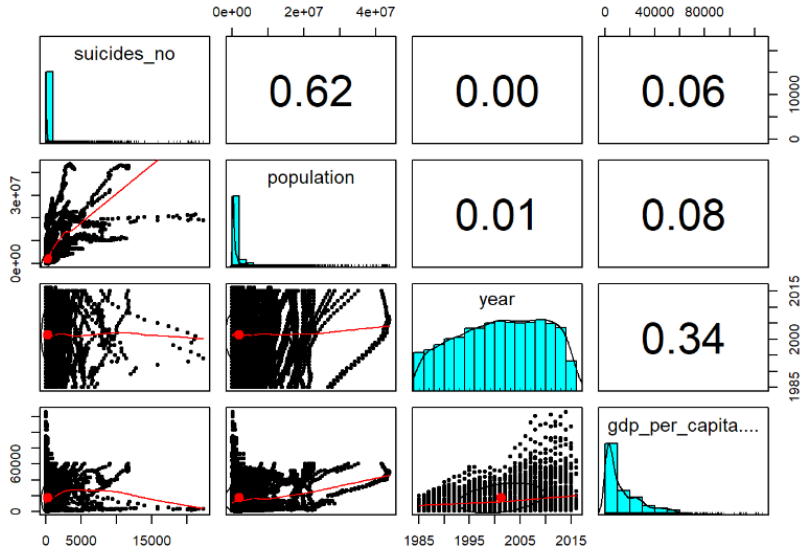


Correlation Checking

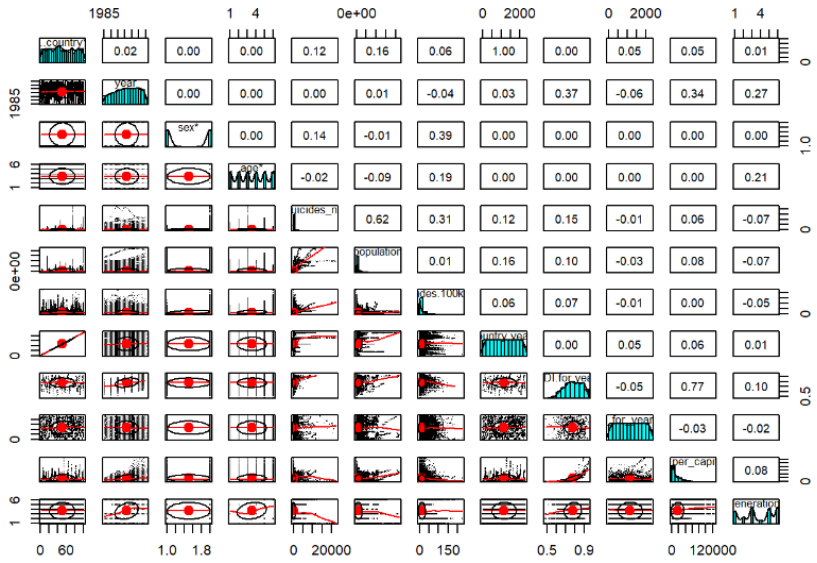
```
#check the correlation
cor(data[c("suicides_no","population","year","gdp_per_capita....")])
```

```
##                    suicides_no population         year gdp_per_capita....
## suicides_no         1.000000000 0.61616227 -0.004545958         0.06132975
## population          0.616162268 1.00000000  0.008850170         0.08150986
## year               -0.004545958 0.00885017  1.000000000         0.33913428
## gdp_per_capita....  0.061329749 0.08150986  0.339134280         1.00000000
```

```
pairs.panels(data[c("suicides_no","population","year","gdp_per_capita....")])
```



```
#Alternative Scatterplot Matrix Function
suppressMessages(library(psych))
pairs.panels(data, pch=".")
```



Exploratory Data Analysis in R - Dataset 2

From the Dataset 1, we have the overview of suicides problem around the world and we have seen the correlation between variables. In this next Dataset 2, we could have more insights about the reasons that lead people to commit suicides. This dataset used the data from India only.

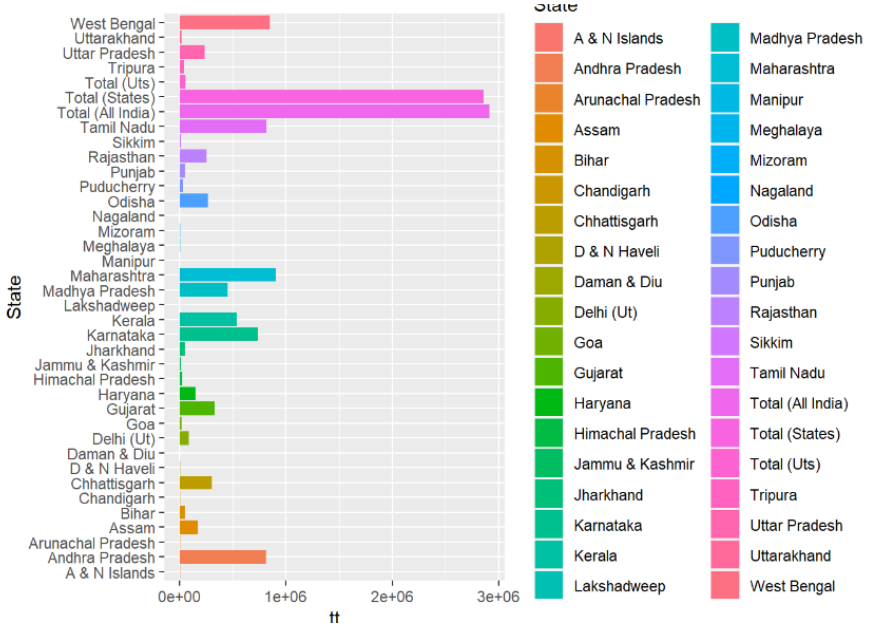
Importing the dataset & library

```
library(ggplot2)
library(psych)
```

```
library(class)
library(dplyr)

india = read.csv('Suicides in India 2001-2012.csv')
View(india)
summary(india)
```

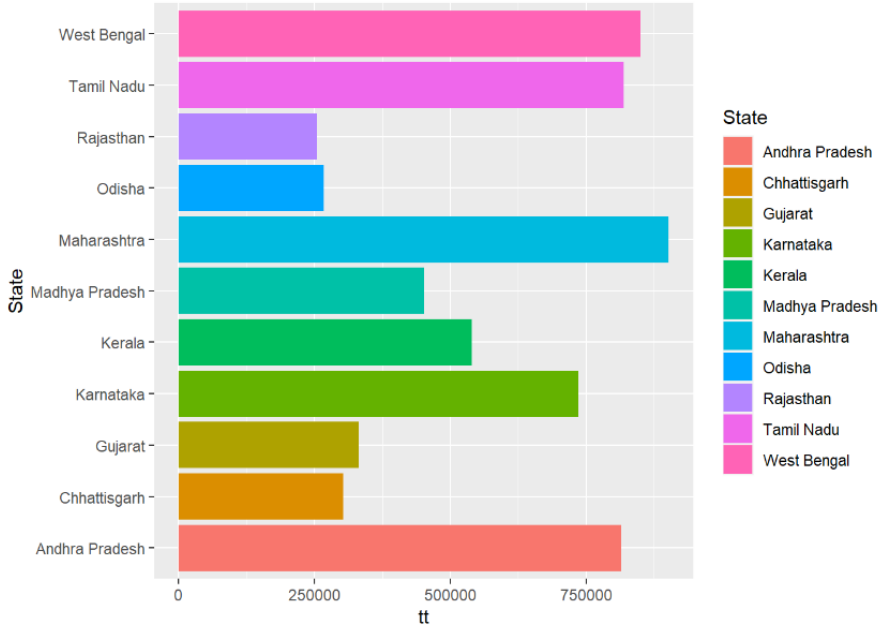
Suicide rate by State

```
# suicide rate by state

data1 <-  group_by(data,State)  %>%  summarise(tt =  sum(Total))

ggplot(data1 , mapping = aes ( State,tt ,fill = State ) ) + geom_bar(stat="identity" ) + coord_flip()
```



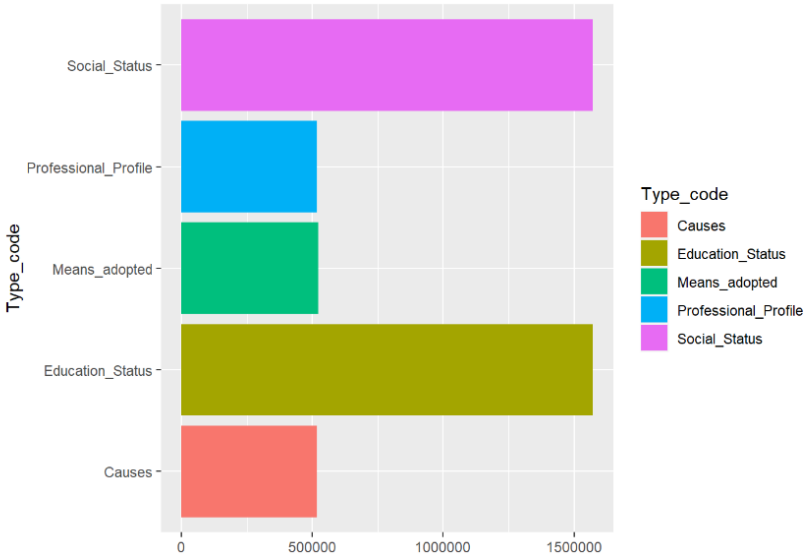
Top 10 suicide state

```
data1 <-  arrange(data1 , desc(tt))
#take the top 10
data1 <-  data1[3:13 ,]
ggplot(data1 , mapping = aes ( State,tt ,fill = State ) ) + geom_bar(stat="identity" ) + coord_flip()
```



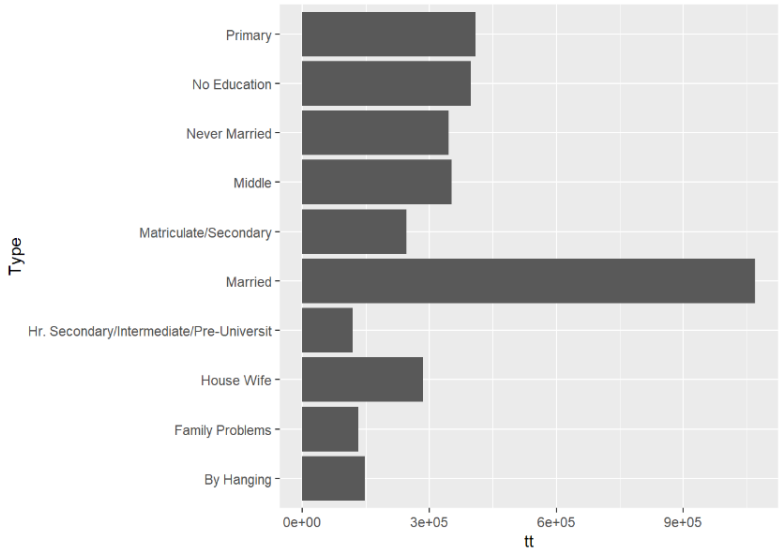
Reasons of a women committed suicide

```
# raison of a women committed suicide

data2  <-  filter(data , Gender == 'Female')
data2  <-  group_by(data2, Type_code) %>% summarise(tt = sum(Total))
ggplot(data2 , mapping = aes ( Type_code,tt ,fill = Type_code ) ) + geom_bar(stat="identity" ) + coord_flip()
```



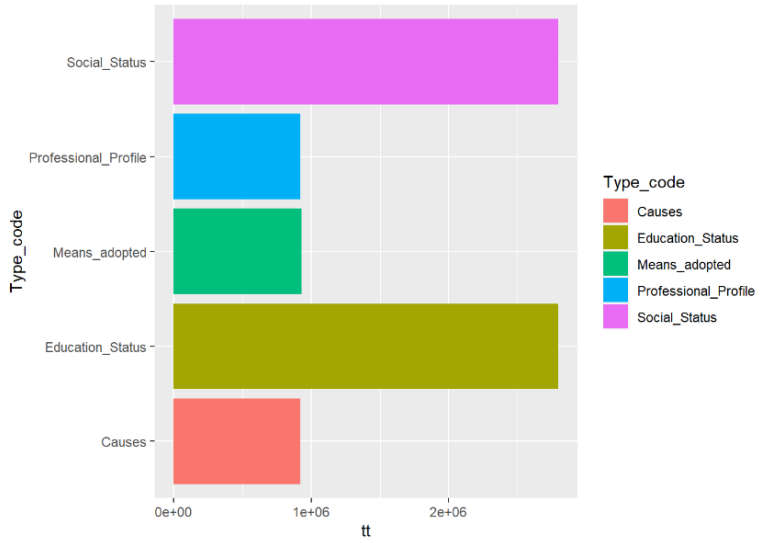
Specific reason of women committed suicide

```
#specific reason of man committed suicide
data5  <-  filter(data , Gender == 'Female')
data5  <-  group_by(data5, Type) %>% summarise(tt = sum(Total))

data5 <-  arrange(data5 , desc(tt))
#take the top 10
data5 <-  data5[1:10 ,]

ggplot(data5 , mapping = aes ( Type ,tt  ) ) + geom_bar(stat="identity" )+ coord_flip()
```



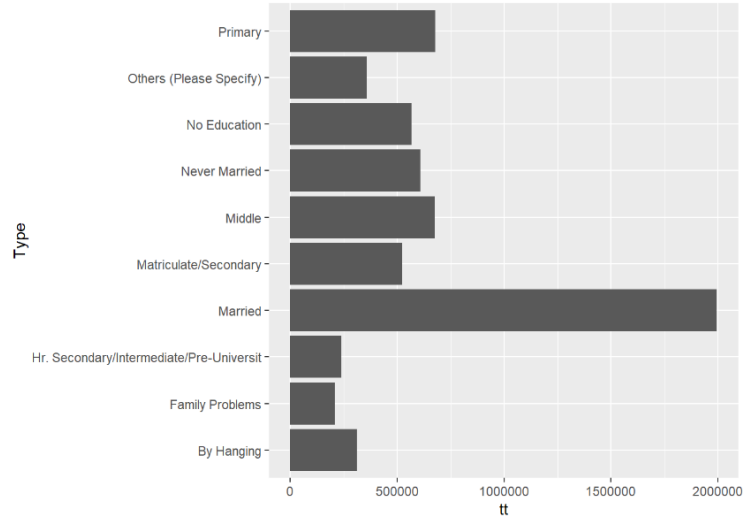
Raison of man committed suicide

```
data3  <-  filter(data , Gender == 'Male')
data3  <-  group_by(data3, Type_code) %>% summarise(tt = sum(Total))
ggplot(data3 , mapping = aes ( Type_code,tt ,fill = Type_code ) ) + geom_bar(stat="identity" ) + coord_flip()
```



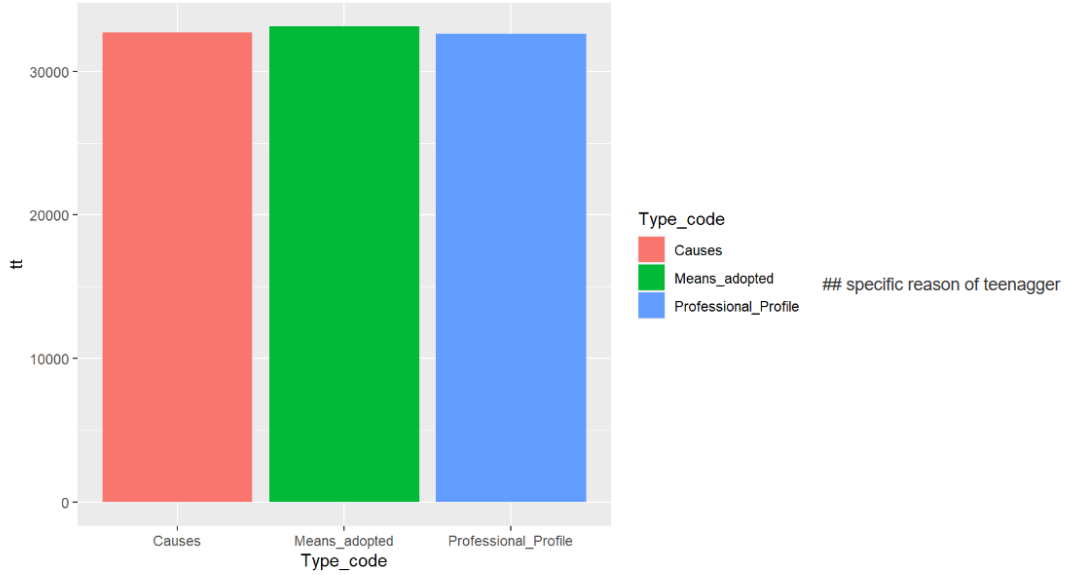
Specific reason of man committed suicide

```
#specific reason of man committed suicide
data4  <-  filter(data , Gender == 'Male')
data4  <-  group_by(data4, Type) %>% summarise(tt = sum(Total))

data4 <-  arrange(data4 , desc(tt))
#take the top 10
data4 <-  data4[1:10 ,]

ggplot(data4 , mapping = aes ( Type ,tt  ) ) + geom_bar(stat="identity" )+ coord_flip()
```

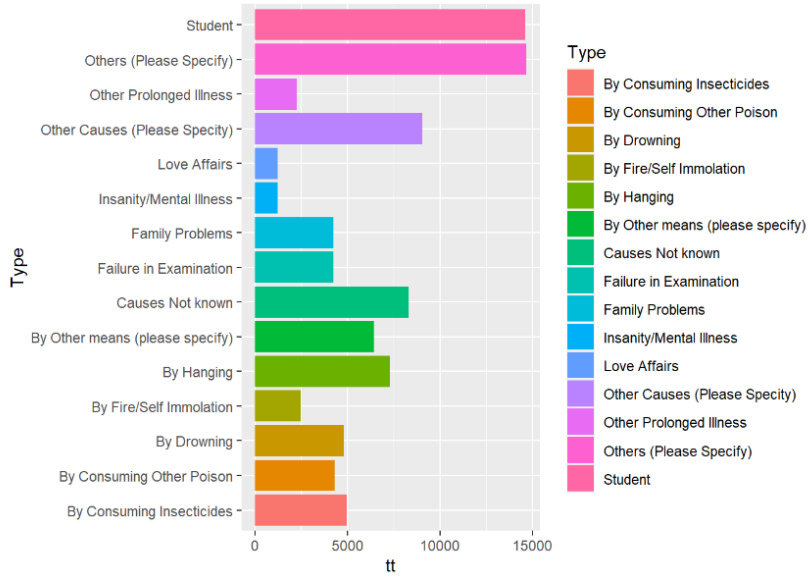


Why Children/Teenager commit suicide?

```
data6 <- group_by(data,Age_group,Type_code) %>% summarise(tt = sum(Total))
data6 <-  filter(data6, Age_group == '0-14')
ggplot(data6 , mapping = aes ( Type_code ,tt , fill = Type_code ) ) + geom_bar(stat="identity" )
```

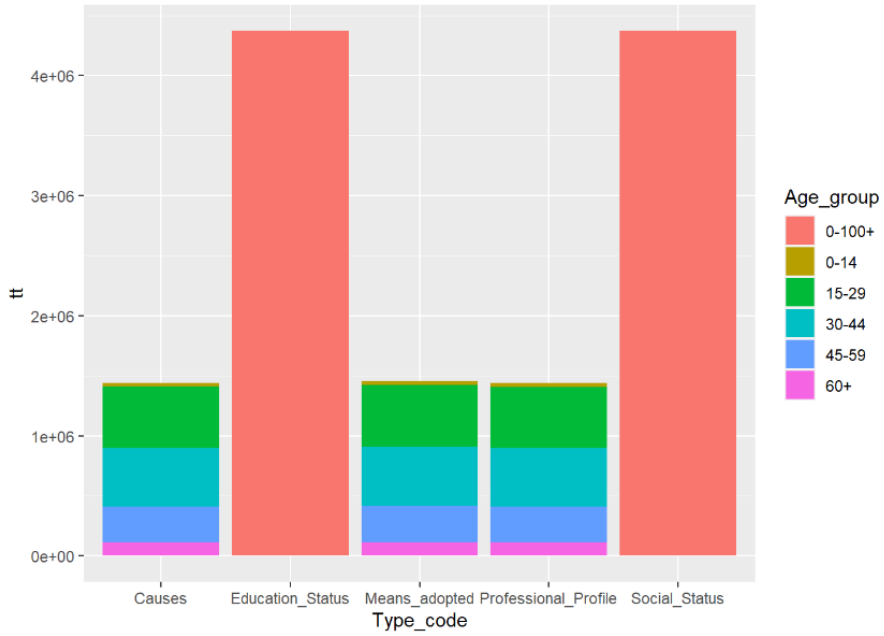



```
data8 <- group_by(data,Age_group)
data8 <-  filter (data8 ,Age_group == '0-14' )
data8 <- group_by(data8,Type) %>% summarise( tt = sum (Total))

data8 <-  arrange(data8 , desc(tt))
#take the top 10
data8 <-  data8[1:15 ,]

ggplot(data8 , mapping = aes ( Type ,tt , fill = Type ) ) + geom_bar(stat="identity" ) + coord_flip()
```



The cause of suicides by age

```
data7 <- group_by(data,Age_group,Type_code) %>% summarise(tt = sum(Total))
ggplot(data7 , mapping = aes ( Type_code ,tt , fill = Age_group ) ) + geom_bar(stat="identity" )
```



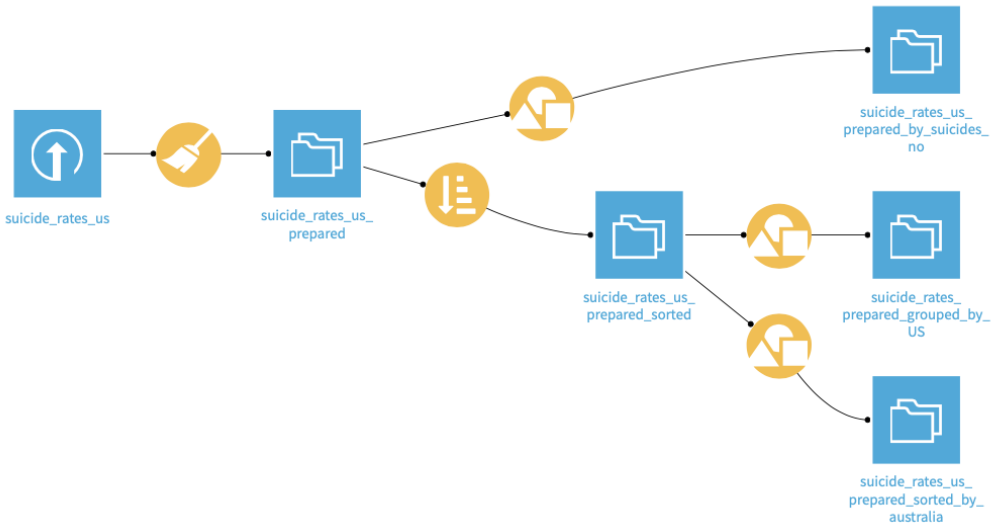
Exploratory Data Analysis in Dataiku – Dataset 1

Dashboard

Overall Dashboard which shows the Top 25 countries with highest number of suicide cases, Number of suicide cases by age & gender, Suicide cases by generation and Number of suicide cases by GDP per capital by country.

Details Analysis

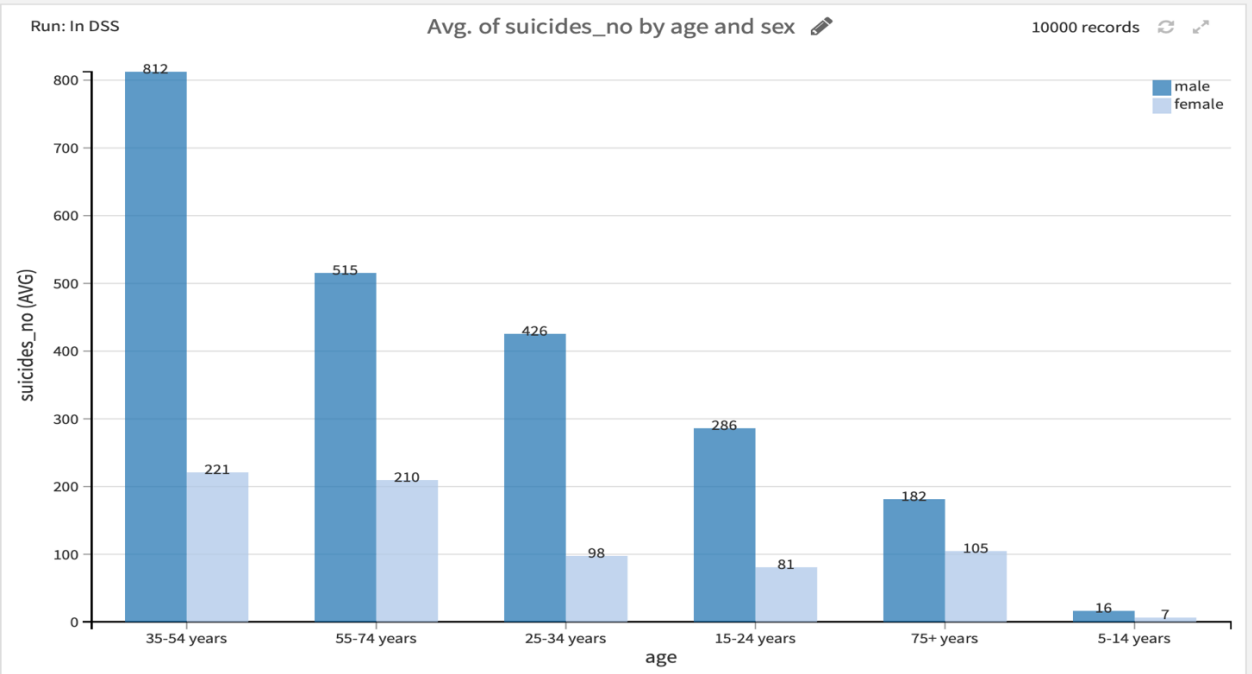
Flowchart of the analysis:



The dataset “suicide_rates_us” is sampled with a size of 10000 records and sorted by year. The analysis of this dataset and the datasets grouped by US and Australia are performed.

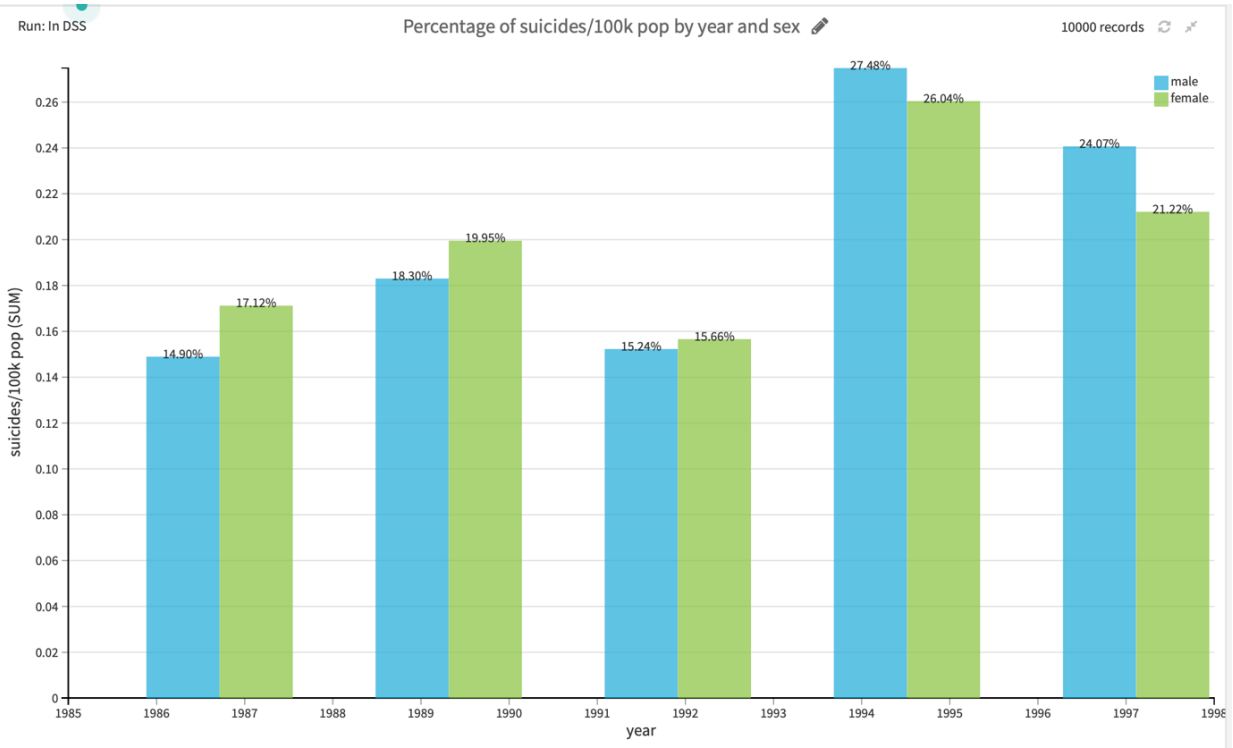
Analysis of the world: “Dataset: suicides_rates_us_prepared_sorted”

Parameters: Suicides_no vs age and sex



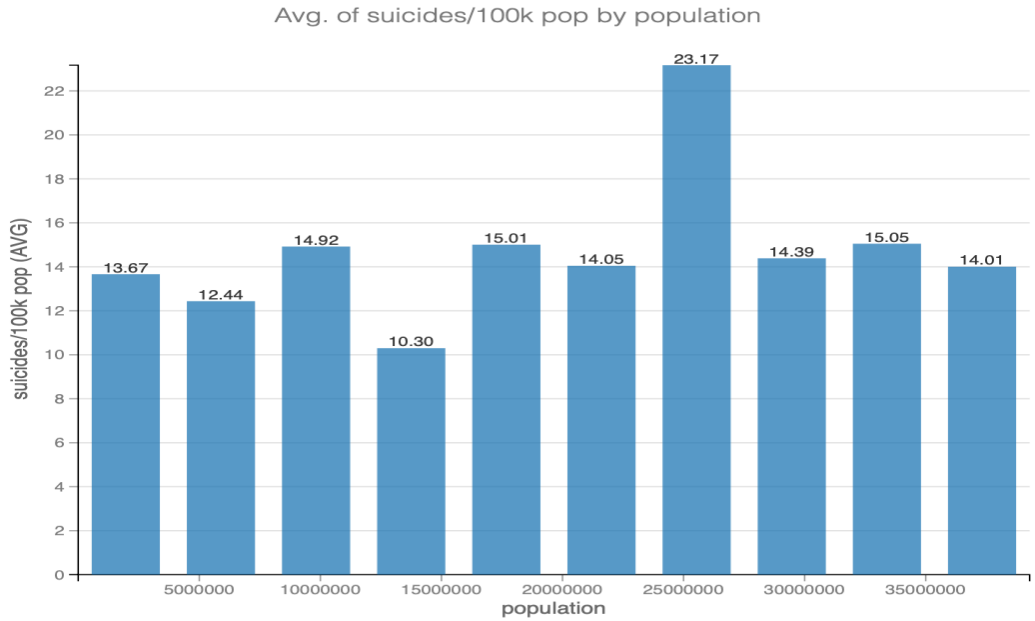
Observation: The age group of 35-64 years has the highest average number of suicides worldwide with the male average of 812 and female average of 221.

Parameters: Suicides/100k vs sex and year



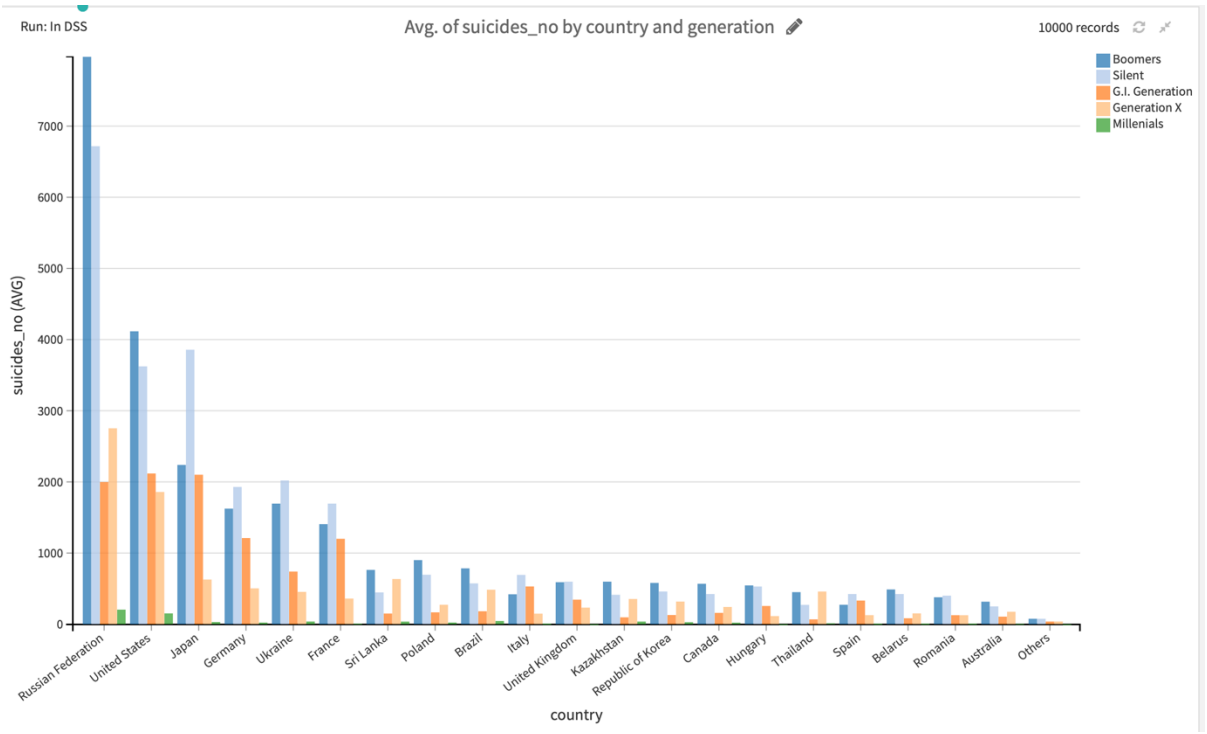
Observation: The number of suicides/100k is highest in the year 1994-1995. This year interval has 27.48% of all the male suicides and 26.04% of all the female suicides from 1985 to 1998.

Parameters : Suicides/100k vs population



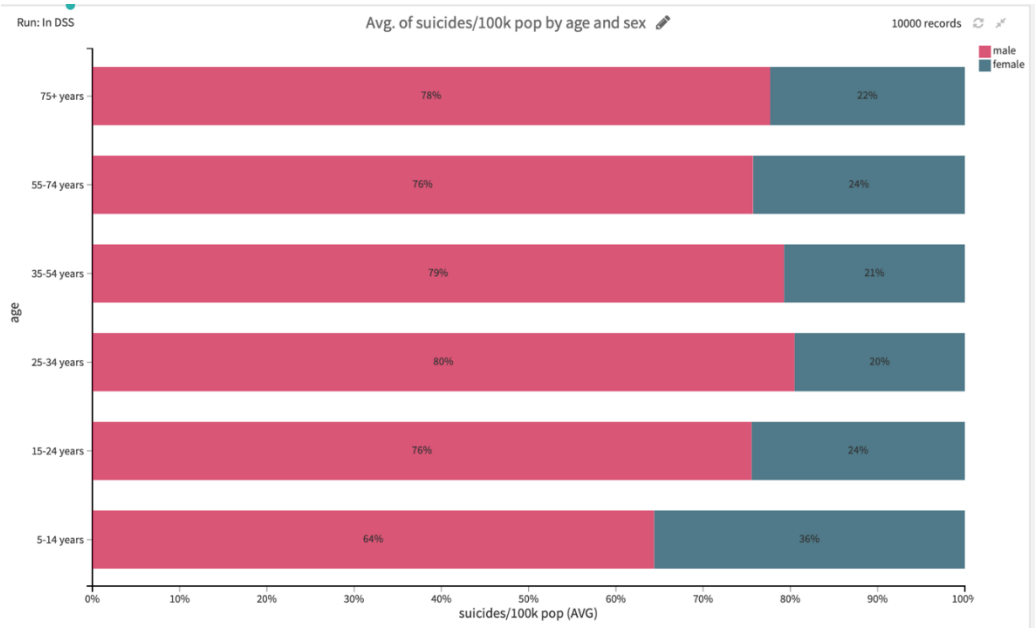
Observation: The average number suicides per 100k population is surprisingly highest at 250k population while the other population sizes tend to have similar number of suicides.

Parameters: Suicides/100k vs country and generation



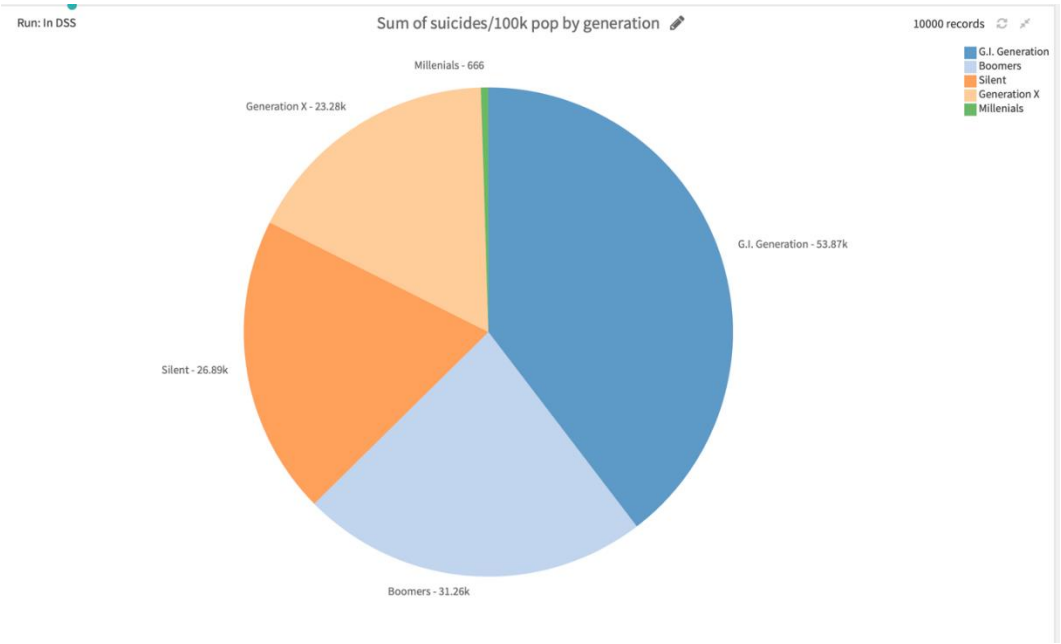
Observation: Russian Federation has the highest number of suicides/100k of all the countries. Among the age group, Boomers have the highest number of suicides for each of the countries except Japan, Germany, Ukraine, France and UK where Silent dominates Boomers.

Parameters: Suicides/100k vs age and sex



Observation: The ratio of male: female suicide ratio is highest in the 25-34 years group with the male suicides as high as 4 times the female suicides.

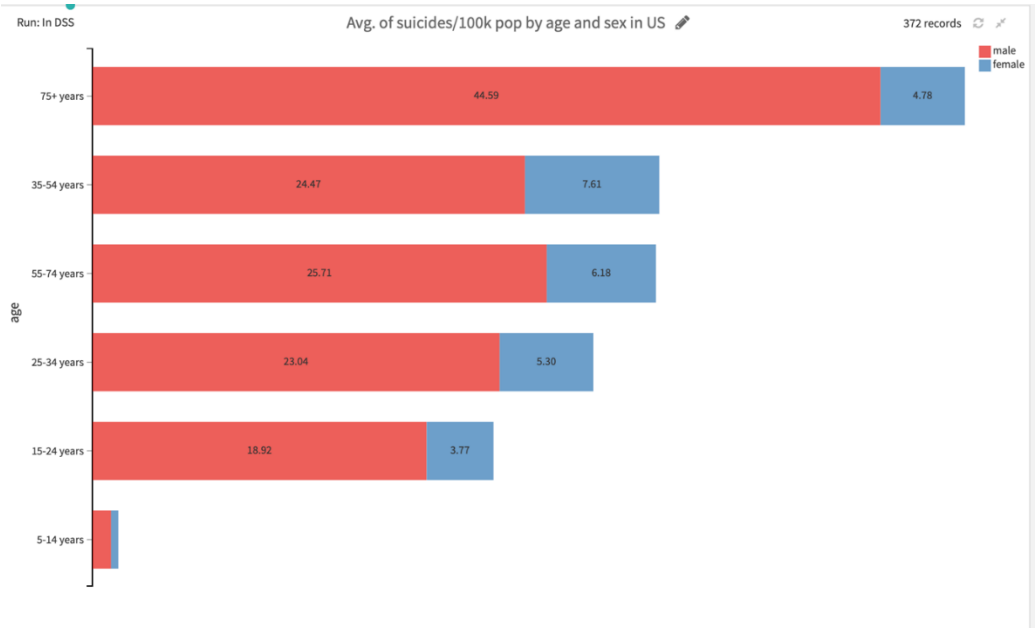
Parameters: Suicides/100k vs age-group



Observation: Out of the total number of suicides/100k by generation for all the countries combined, G.I. Generation is the highest with a total sum of 53.8k

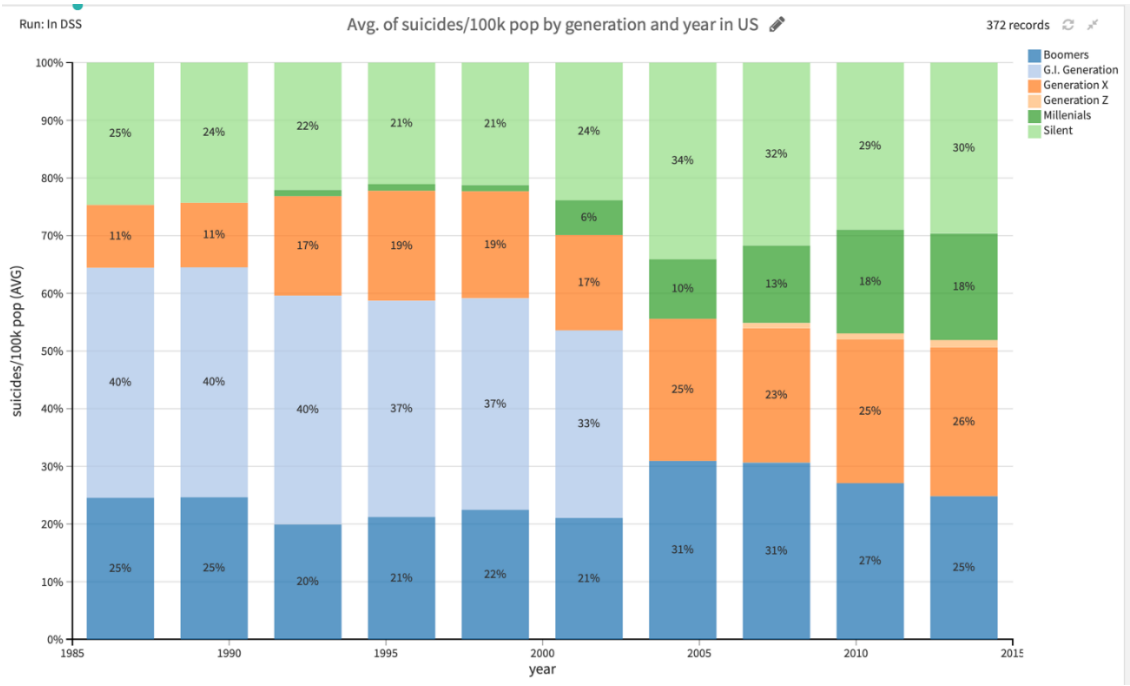
Analysis of US: “Dataset: suicides_rates_us_prepared_grouped_by_US”

Parameters: Suicides/100k vs age and sex



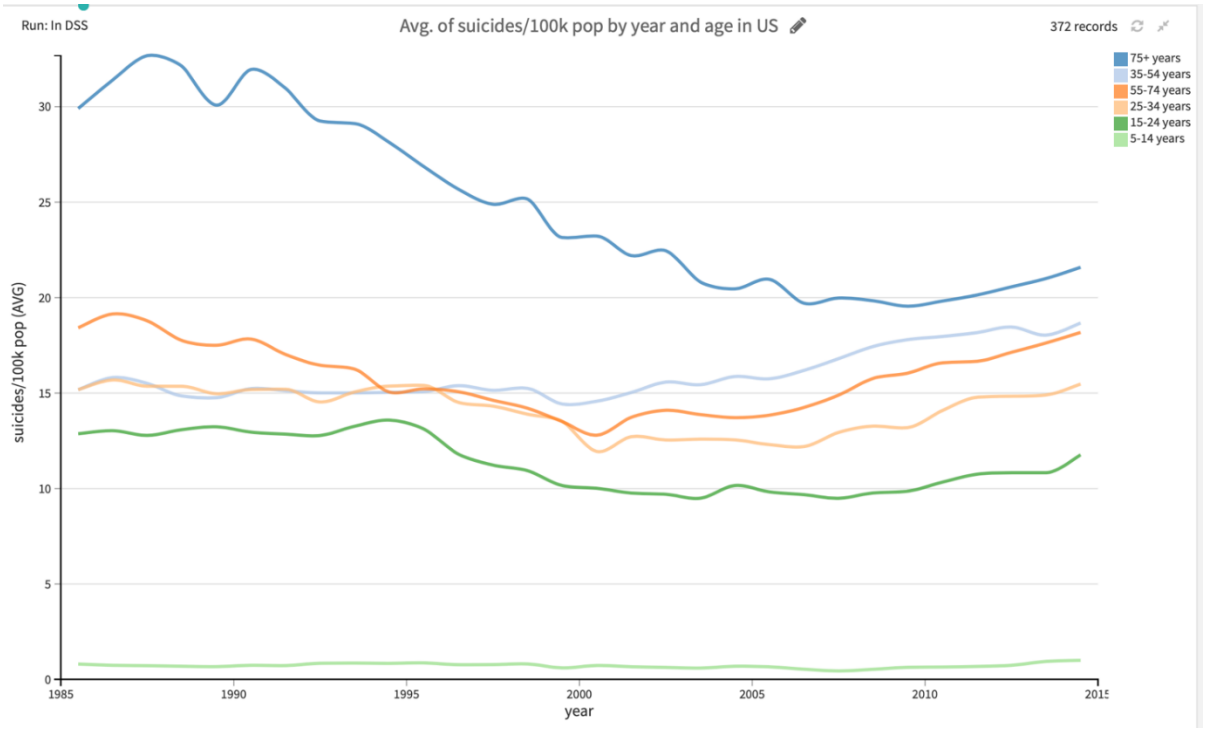
Observation: The age group 75+ years have the highest rate of suicides. The male suicides of this particular age group are phenomenally higher than the female suicides (approximately 10 times). This age group also has the highest male suicides (44.59%) of all group while the age group 35-54 years has the highest female suicides (7.61%) of all groups.

Parameters: Suicides/100k vs generation and year (year interval: 3 years)



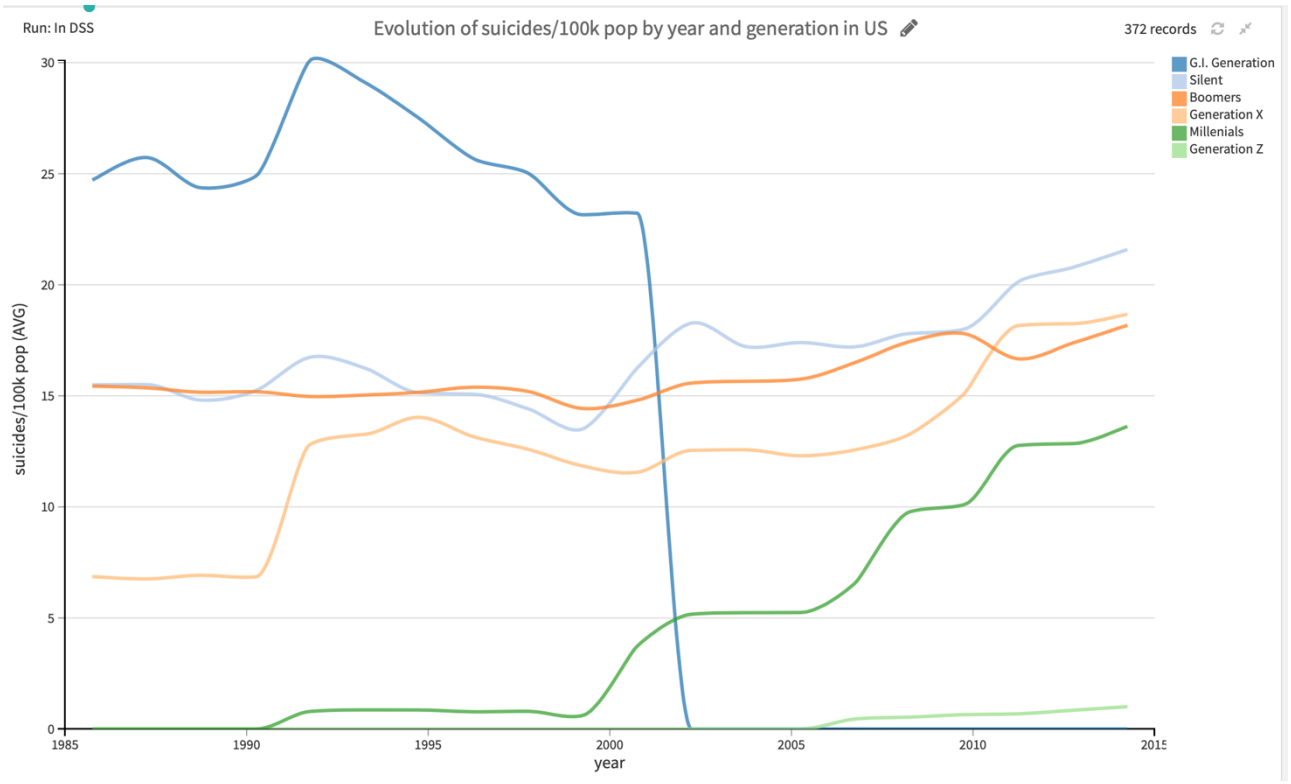
Observation: The amount of suicides/100k is significantly increasing over years for the generations: Generation X, Millennials and Silent. It is surprising to see that the most vulnerable generation G.I. has no suicides/100k after 2003.

Parameters: Suicides/100k vs year and age



Observation: It is evident from this diagram that the suicides/100k is constantly increasing for the age group 35-54 (Silent generation).

Parameters: Suicides/100k vs year and generation



Observation: This diagram re-justifies the sharp increase of suicides/100k for the generations: Millennials, Generation X and Silent. Also, the G.I. generation suddenly drops to the lowest at 2003 from being the highest suicide group according to the dataset.

These findings are coherent with the available information on the internet as available through the links below:

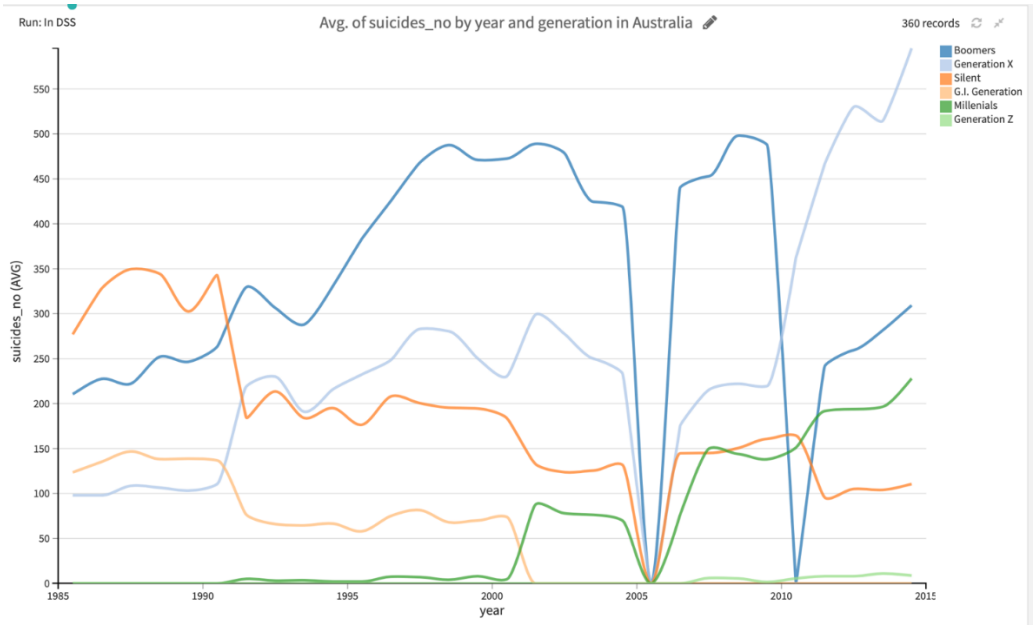
1. *"Baby Boomer Suicide Rate Rising, May Go Higher with Age"* available at: https://www.healthline.com/health-news/baby-boomer-suicide-rate-rising-031515#1

2. *"Study explores which generation of workers is most likely to consider suicide"* available at: https://www.safetyandhealthmagazine.com/articles/18536-study-explores-which-generation-of-workers-is-most-likely-to-consider-suicide

3. "*Generation Z Reported the Most Mental Health Problems, and Gun Violence Is the Biggest Stressor"* available at: https://www.sprc.org/news/generation-z-reported-most-mental-health-problems-gun-violence-biggest-stressor (In their survey, generation z is the same as generation x in our dataset)

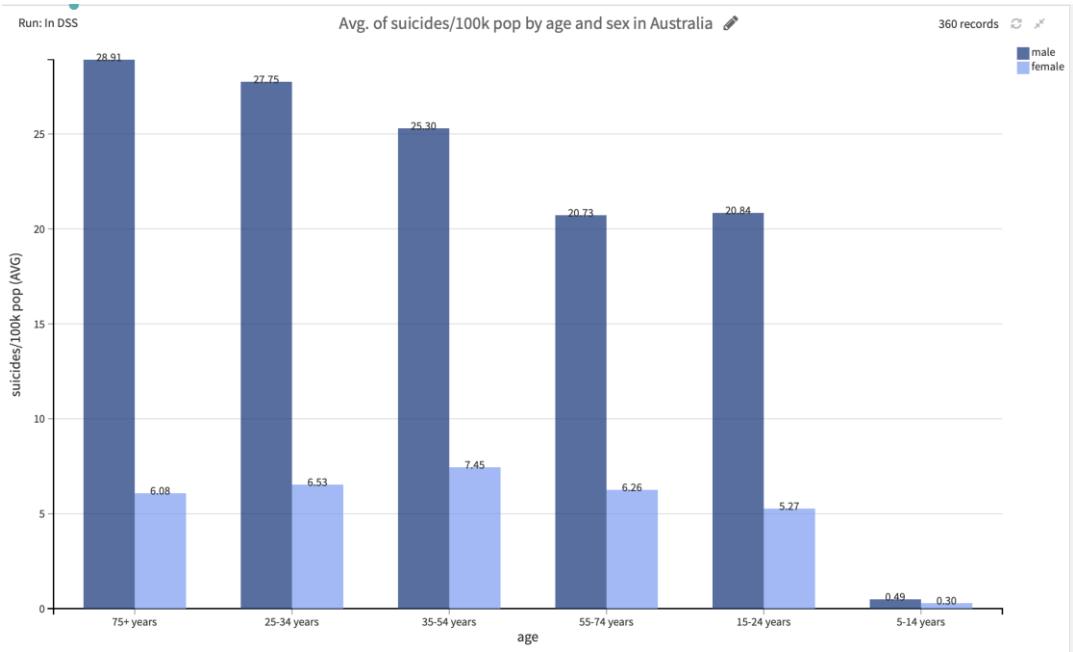
Analysis of US: “Dataset: suicides_rates_us_prepared_sorted_by_Australia”

Parameters: Suicides_no vs year and generation



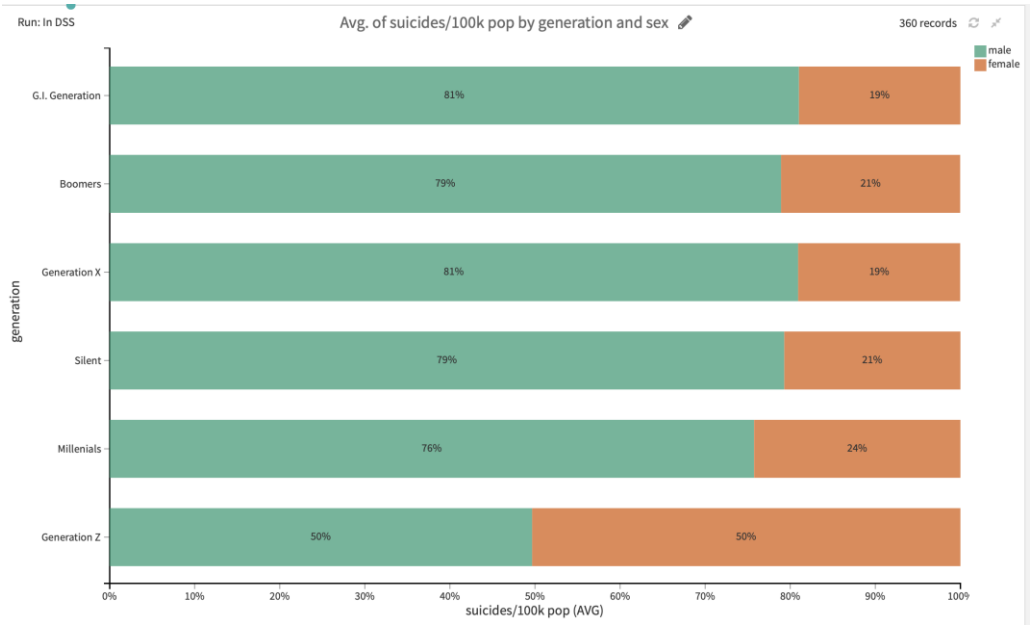
Observation: The suicides for all the age groups attains its lowest at 2006, at 10.2 per 100k according to the report of the Guardian [1].

Parameters: Suicides/100k vs age and sex



Observation: The age group: 75+ years has the highest suicides in Australia with a male average of 28.91 /100k and a female average of 6.08.

Parameters: Suicides/100k vs generation and sex



Observation: G.I. generation and Generation X have the same male: female suicides ratio while Generation Z has this ratio equals to 1.

These findings are also coherent with the available information on the internet as available through the links below:

1. *"Australia's suicide rate to rise 40% if emerging risks such as debt not tackled"* available at: https://www.theguardian.com/australia-news/2019/sep/10/australias-suicide-rate-to-rise-40-if-emerging-risks-such-as-debt-not-tackled

2. "*Australian men aged over 85 have the highest rate of suicide, ABS data shows*" available at: https://www.abc.net.au/news/2017-05-30/australian-men-aged-over-85-have-the-highest-rate-of-suicide/8569740

Prediction Model using R (Dataset 1)

Linear Regression

```
# building the model - linear regression
suicide_model<-lm(suicides_no~sex+year+population+gdp_per_capita....,data)
summary(suicide_model)
```

```
##
## Call:
## lm(formula = suicides_no ~ sex + year + population + gdp_per_capita....,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3082.7  -132.0   -27.6   134.4 19511.8
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.144e+03  1.049e+03   2.998 0.002723 **
## sexmale             2.734e+02  8.359e+00  32.712  < 2e-16 ***
## year               -1.656e+00  5.247e-01  -3.156 0.001604 **
## population          1.422e-04  1.072e-06 132.627  < 2e-16 ***
## gdp_per_capita....  7.802e-04  2.360e-04   3.306 0.000949 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 697 on 27815 degrees of freedom
## Multiple R-squared:  0.403,  Adjusted R-squared:  0.4029
## F-statistic:  4693 on 4 and 27815 DF,  p-value: < 2.2e-16
```

```
suicide_model2<-data[1:130,]
suicide_test2<-data[131:18000,]
suicide_train2<-suicide_model2
suicide_model2<-lm(suicides_no~sex+year+population+gdp_per_capita....,suicide_train2)
suicide_pred<-predict(suicide_model2,suicide_test2)

cor(suicide_test2$suicides_no,suicide_pred)
```

```
## [1] 0.4808525
```

```
head(suicide_pred)
```

```
##       131       132       133       134       135       136
## 12.866151 18.394220 16.181283 14.965523  6.910185 11.569854
```

```
head(suicide_test2$suicides_no)
```

```
## [1]  1  0 17 10  2  1
```

```
suicide_model2
```

```
##
## Call:
## lm(formula = suicides_no ~ sex + year + population + gdp_per_capita....,
##     data = suicide_train2)
##
## Coefficients:
##        (Intercept)             sexmale                year          population
##         -1.095e+03           4.947e+00           5.477e-01           2.227e-05
## gdp_per_capita....
##          4.572e-03
```

```
summary(suicide_model2)$coef
```

```
##                         Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)        -1.095286e+03 3.351013e+02 -3.268522 1.396421e-03
## sexmale             4.946909e+00 1.209158e+00  4.091200 7.636606e-05
## year                5.477075e-01 1.685113e-01  3.250272 1.481649e-03
## population          2.226668e-05 5.342767e-06  4.167631 5.700982e-05
## gdp_per_capita....  4.572314e-03 2.663776e-03  1.716479 8.855153e-02
```

```
#knn model

temp.data <- subset(data, select = -suicides_no )
summary(data)
```

```
##   ï..country             year          sex                age
##  Length:27820       Min.   :1985   Length:27820       Length:27820
##  Class :character   1st Qu.:1995   Class :character   Class :character
##  Mode  :character   Median :2002   Mode  :character   Mode  :character
##                     Mean   :2001
##                     3rd Qu.:2008
##                     Max.   :2016
##
##   suicides_no        population       suicides.100k.pop country.year
##  Min.   :    0.0   Min.   :     278   Min.   :  0.00    Length:27820
##  1st Qu.:    3.0   1st Qu.:   97498   1st Qu.:  0.92    Class :character
##  Median :   25.0   Median :  430150   Median :  5.99    Mode  :character
##  Mean   :  242.6   Mean   : 1844794   Mean   : 12.82
##  3rd Qu.:  131.0   3rd Qu.: 1486143   3rd Qu.: 16.62
##  Max.   :22338.0   Max.   :43805214   Max.   :224.97
##
##   HDI.for.year   gdp_for_year....   gdp_per_capita....  generation
##  Min.   :0.483   Length:27820       Min.   :   251     Length:27820
##  1st Qu.:0.713   Class :character   1st Qu.:  3447     Class :character
##  Median :0.779   Mode  :character   Median :  9372     Mode  :character
##  Mean   :0.777                      Mean   : 16866
##  3rd Qu.:0.855                      3rd Qu.: 24874
##  Max.   :0.944                      Max.   :126352
##  NA's   :19456
```

Conclusion

From WHO, suicide is one of the priority conditions in the WHO Mental Health Gap Action Programme (mhGAP) launched in 2008, which provides evidence-based technical guidance to scale up service provision and care in countries for mental, neurological and substance use disorders.

Through this project, we have the overview of the current global suicides problem and particularly the reasons behind of committing suicides in the case of India. Since the limitation of Dataset, we can not know the picture of other countries, we hope in the future there would be more insightful data from other high-suicides-rate countries in order to have the prevention action accordingly.