A Study of Suicide Rates in Relation with Economic Indicators

Omer Cagri Aydin

02 February 2021

Introduction

A sensitive topic such as suicide of course has many levels and factors to it that must be individually studied for each different case. However, the purpose of this research is to find out if there exists any general correlation between economic indicators and suicide rates. I believe that there might be a correlation between these variables and determining a trend could help suicide prevention groups to target specific groups of people that are in increased risks of self-harm due to the surrounding economic factors.

The dependent variable in this data set will be suicides.100k.pop (suicide rates per 100.000 people) while my independent variables will be HDI, GDP per capita, and generation. I will study the correlations between my independent variables and suicides.100k.pop. Moreover, I will construct graphs in order to

visualize the realtionship between the variables. Finally, I will conduct a regression analysis of relevant variables, conduct a T test, and a F test in order to study the significance. I will also figure out a way to use generation variable in my regression. To do so, I must manipulate this data, which is categorical, into being numerical.

I believe that there might be a relationship between these valuables due to the way they effect the lifestyle. However, I would not be surprised to find a positive relationship (for example as GDP per capita increases so does suicides per 100k) because of the so called American Paradox.

Null Hypothesis: H0: u=0 Alternative Hypothesis: H1: u≠0

I will use the following packages for the given reasons

dplyr: rename variables

stargazer:create tables

corrplot: create muti-correlations

ggplot2: plot graphs

The data set

I obtained the dataset that I will be using from kaggle.com and the exact source is stated at the bottom of the page. I selected this data set because I am interested in finding out if there exists a relationship between suicide rates and HDI, GDP per capita, and generations. This data is adequate to answer the reseach question because it has a large number of observations and incorporates many countries throughout a large number of years. However, the data first needs some manipulation before it can be used.

```
#setting working directory
```

```
setwd("~/Desktop/OCA61/University/Junior/Spring 2019/
S19_Eco320L/Final Project")
#importing the data
master <- read.csv("~/Desktop/OCA61/University/Junior/
Spring 2019/S19_Eco320L/Final Project/master.csv")
#data manipulation
#selecting only the columns that will be used from the
data and creating a new dataset named master1
master1 <- master %>%  select(suicides.100k.pop,
HDI.for.year, gdp_per_capita...., generation)

#I realized that there were a lot of "0" values in the
data which corrupted it. This was something I have not
realized in my preliminary submission. So I set all of
these "0" values to NA
master1[master1==0] <- NA

#Creating a new dataset named master2 in which the NA
variables are omitted
master2 <- na.omit(master1)

#This code converts categorical value of generation
into numerical. After this I am creating the dataset
that I will be using throughout,
master2p1 <- master2 %>% mutate_if(is.factor,
as.numeric)

#Renaming the variables for convinence using dplyr
name<-c(suicides100k
="suicides.100k.pop",hdi="HDI.for.year",
gdp.per.capita= "gdp_per_capita....",
generation="generation")
master2p1<-rename(master2p1,!!!name)

#summary stats
```

```
summary(master2p1)
##   suicides100k          hdi        gdp.per.capita
generation
##  Min.   :  0.020   Min.   :0.483   Min.   :   313
Min.   :1.000
##  1st Qu.:  2.615   1st Qu.:0.711   1st Qu.:  4547
1st Qu.:3.000
##  Median :  7.640   Median :0.782   Median : 12584
Median :4.000
##  Mean   : 13.917   Mean   :0.779   Mean   : 21332
Mean   :3.765
##  3rd Qu.: 17.995   3rd Qu.:0.864   3rd Qu.: 31518
3rd Qu.:5.000
##  Max.   :187.060   Max.   :0.944   Max.   :126352
Max.   :6.000
str(master2p1)
## 'data.frame':    7207 obs. of  4 variables:
##  $ suicides100k  : num  5.58 5.06 4.9 4.59 4.56 3.98
3.72 2.65 2.24 1.59 ...
##  $ hdi           : num  0.619 0.619 0.619 0.619
0.619 0.619 0.619 0.619 0.619 0.619 ...
##  $ gdp.per.capita: int  835 835 835 835 835 835 835
835 835 835 ...
##  $ generation    : num  3 6 2 3 3 2 1 3 1 5 ...
# using stargazer to do a table of the summary
statistics of the dataset
stargazer(master2p1,type="text",title="Descriptive
Statistics",out="table.txt")
##
## Descriptive Statistics
##
==========================================================
==================
## Statistic        N      Mean     St. Dev.   Min
Pctl(25) Pctl(75)   Max
```

```
##
-------------------------------------------------------
------------------
## suicides100k   7,207   13.917     17.973   0.020
2.615    17.995  187.060
## hdi            7,207   0.779      0.096    0.483
0.711    0.864    0.944
## gdp.per.capita 7,207 21,331.820 22,776.030  313
4,547    31,518  126,352
## generation     7,207   3.765      1.720      1
3        5        6
##
-------------------------------------------------------
------------------
```

The generations that correspond with the numeric values are:

1=Boomers

2=G.I Generation

3=Generation X

4=Generation Z

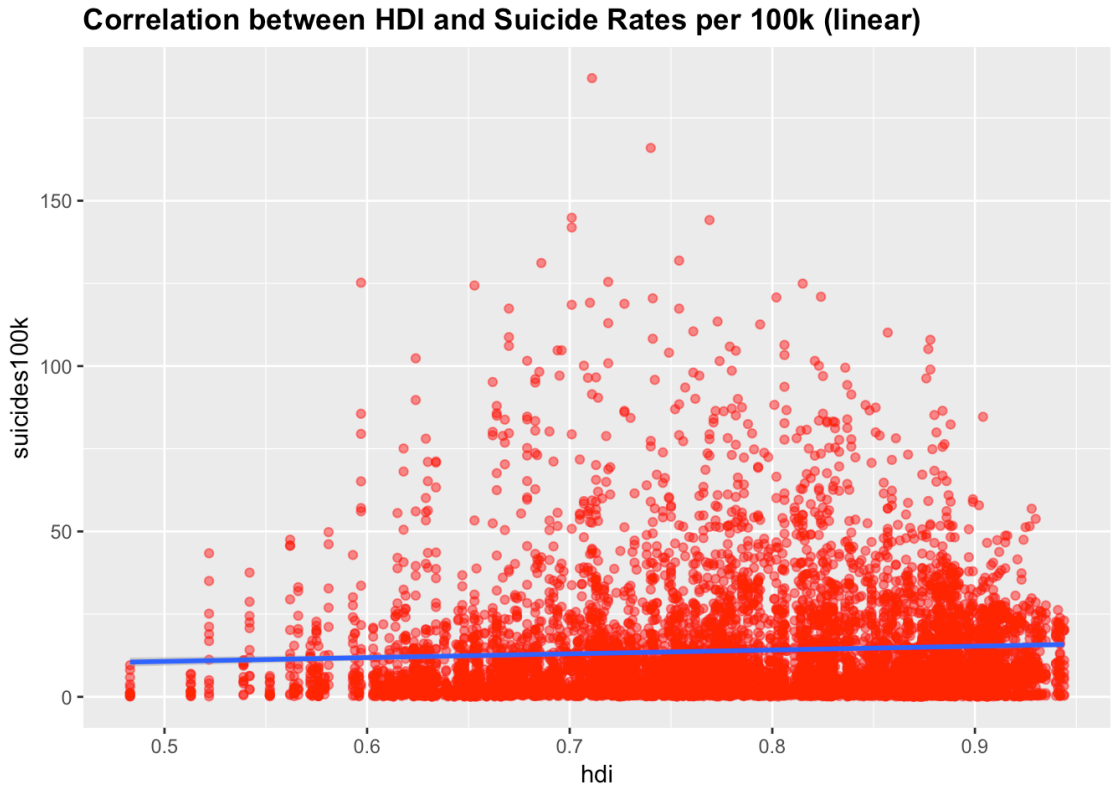
5=Millenials

6=Silent

Graphs and Descriptive Statistics

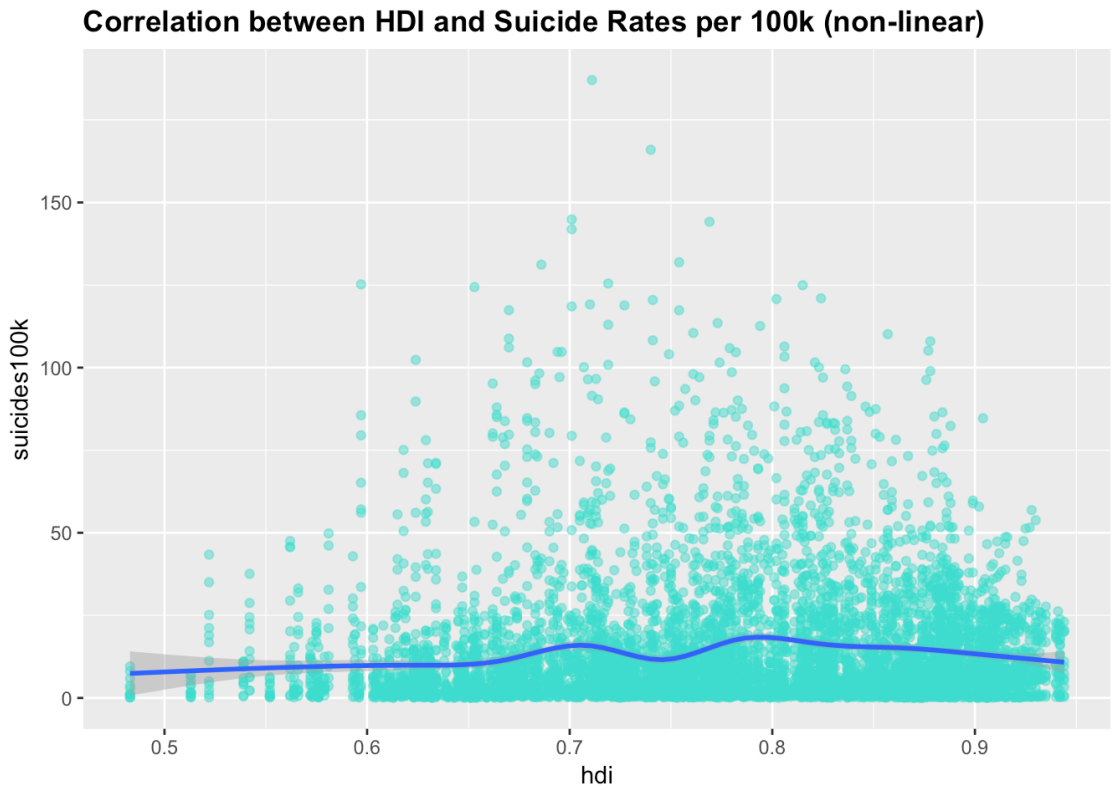
```
#Creating a Scatterplot for suicides100k and hdi
ggplot(data=master2p1, aes(x=hdi, y=suicides100k)) +
geom_point(alpha=0.5, color="red")+
  geom_smooth(method='lm') + ggtitle("Correlation
between HDI and Suicide Rates per 100k (linear)") +
```

```
    theme(plot.title = element_text(face="bold"))
```



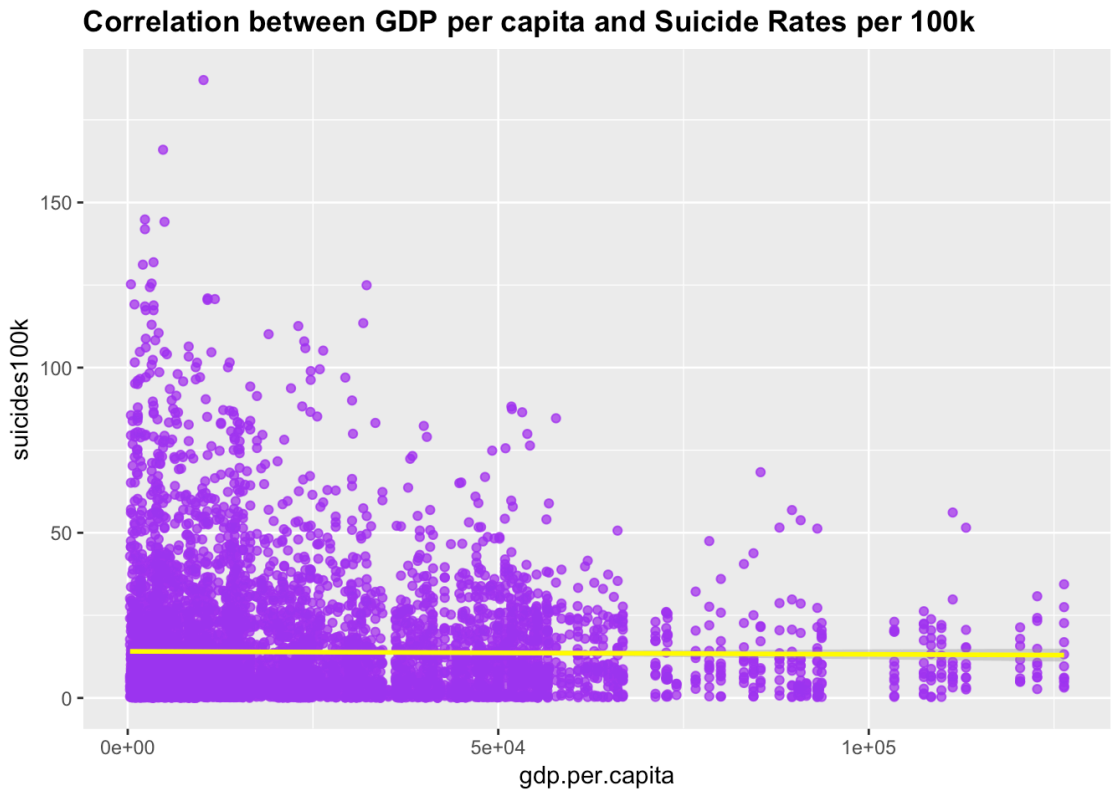
The scatter plot above shows the correaltion between HDI and the suicide rates per 100k. There seems to be a weak upward trending relation between Suicide Rates and HDI. The data may have no significance. A second scatterplot must be constructed with a non-linear trend line to see in more detail.

```
#Recreating the scatterplot above, however, this time
the trend line is nonlinear
ggplot(data=master2p1, aes(x=hdi, y=suicides100k)) +
geom_point(alpha=0.5, color="turquoise")+
  geom_smooth() + ggtitle("Correlation between HDI and
Suicide Rates per 100k (non-linear)") +
    theme(plot.title = element_text(face="bold"))
## `geom_smooth()` using method = 'gam' and formula 'y
~ s(x, bs = "cs")'
```



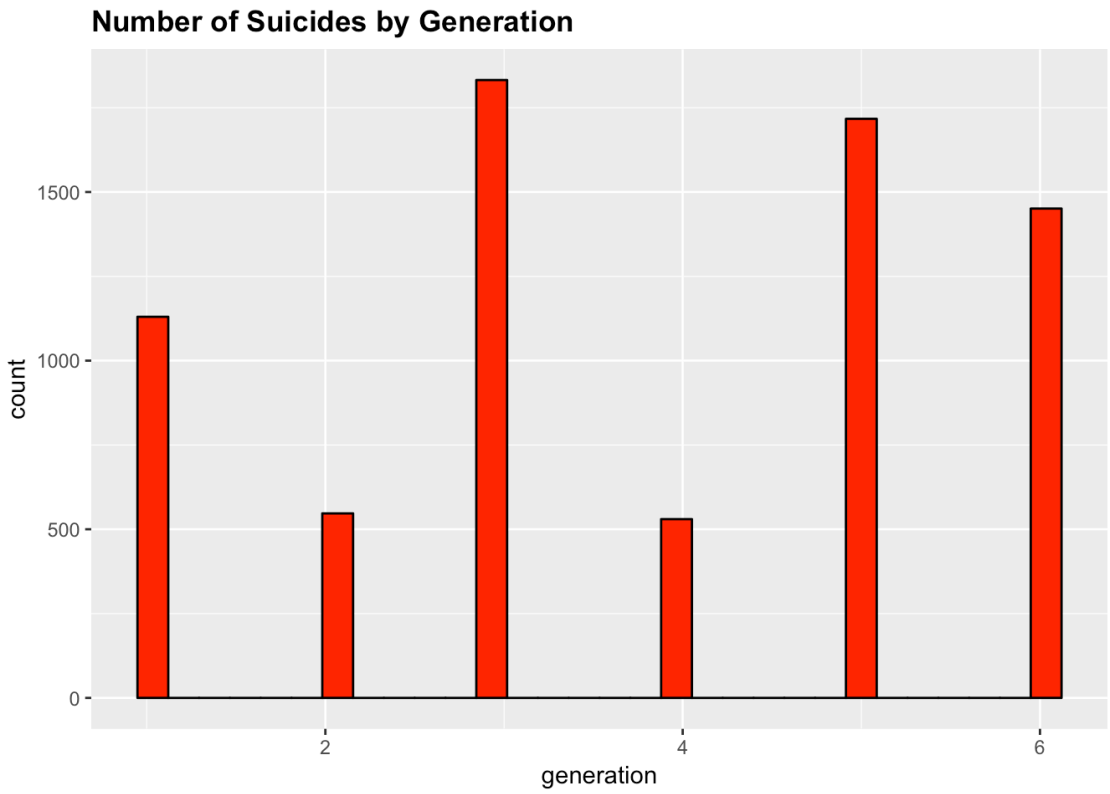
There seems to be two spikes on this trend with spikes on 0.7 and 0.8 HDI. However, just by looking at the plot, there still seems to be no significant relation between HDI and suicide rates.

```
#Creating a scatterplot with GDP per capita and sucides
per 100k
ggplot(data=master2p1, aes(x=gdp.per.capita,
y=suicides100k)) + geom_point(alpha=0.7,
color="purple")+
  geom_smooth(method='lm', color="yellow") +
ggtitle("Correlation between GDP per capita and Suicide
Rates per 100k") +
     theme(plot.title = element_text(face="bold"))
```



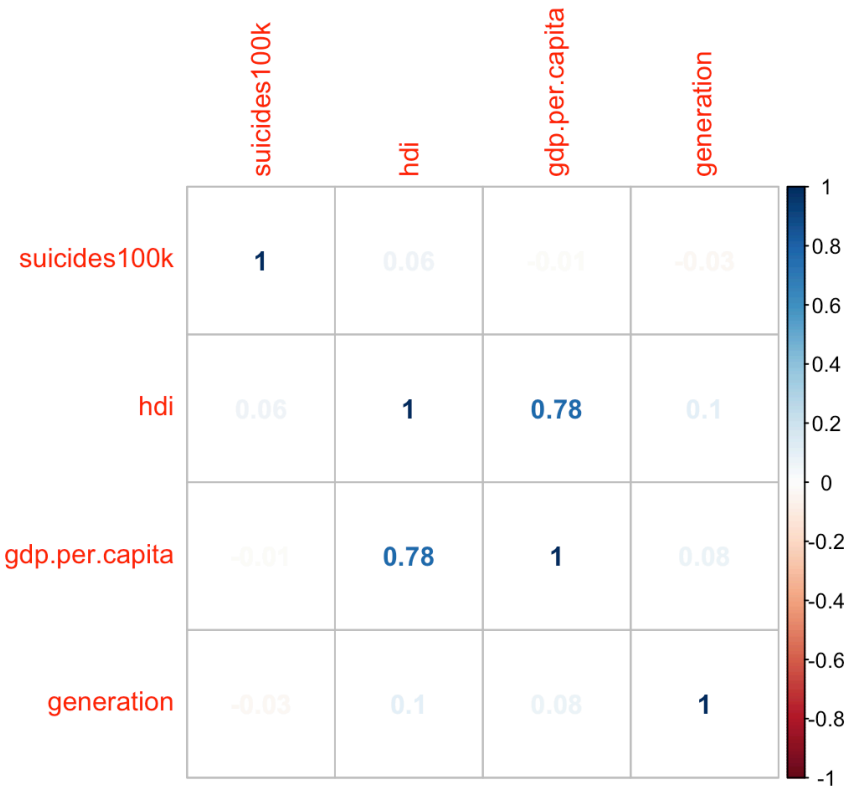
There seens to exist a weak downward trend on the line that probably will not be significant. This plot may be tricky. There seems to be a lot more suicide values on the lower side of the GDP per capita. In reality, although there is more data on the left side of the graph, a lot of them come from low suices per 100k values. Because of this, the trend line does not increase a significant decrease in suicide rates.

```
#Creating a histogram for showing the counts of suicide
cases for each generation
ggplot(data=master2p1) +
  geom_histogram(mapping=aes(generation), bins = 30,
color="black", fill = "red") + ggtitle("Number of
Suicides by Generation") +
     theme(plot.title = element_text(face="bold"))
```



This histogram shows the number of suicides by generation. Looking at the data, generations 2 and 4, G.I. Generation and Generation Z respectfully, seem to have the least count of recorded suicides.

```
#Constructing a multi-correlation
cor1<-cor(master2p1)
corrplot(cor1,method="number")
```



The multiple correlations above shows that suicides per 100k has no correlation with any of the independent variables. The only correlationthere exits is between hdi and gdp per capita, which is not significant in this study.

Regression Analysis

The following models will be created and tested: \[y = \beta_0+ \beta_1*hdi+u\] \[y = \beta_0+\beta_1*hdi + \beta_2*gdp.per.capita+u\] \[y = \beta_0+\beta_1*hdi + \beta_2*gdp.per.capita+\beta_3*generation+u\]

```
#Creating linear models
model1<-lm(suicides100k~hdi,data=master2p1)
model2<-
lm(suicides100k~hdi+gdp.per.capita,data=master2p1)
model3<-
```

```
lm(suicides100k~hdi+gdp.per.capita+generation,data=master2p1)

#Estimating the models
stargazer(model1,model2,model3,type="text",title="Models for estimation of Suicide
Rates",out="model.txt",dep.var.labels ="Suicide Rates per 100k (suicides100k)",align=TRUE)
##
## Models for estimation of Suicide Rates
##
=============================================================================
##
Dependent variable:
##
-----------------------------------------------------------------------
##                                          Suicide Rates per 100k (suicides100k)
##                                (1)                      (2)                      (3)
##
-----------------------------------------------------------------------
## hdi                          11.508***                33.312***                34.125***
##                               (2.193)                  (3.475)                  (3.480)
##
## gdp.per.capita
-0.0001***                -0.0001***
##
(0.00001)                (0.00001)
##
```

```
## generation
-0.431***
##
(0.123)
##
## Constant                    4.952***
-9.502***                 -8.509***
##                              (1.722)
(2.480)                   (2.494)
##
##
------------------------------------------------------
--------------------------------------
## Observations                 7,207
7,207                     7,207
## R2                           0.004
0.013                     0.014
## Adjusted R2                  0.004
0.012                     0.014
## Residual Std. Error    17.940 (df = 7205)
17.861 (df = 7204)        17.847 (df = 7203)
## F Statistic         27.530*** (df = 1; 7205)
46.422*** (df = 2; 7204) 35.100*** (df = 3; 7203)
##
======================================================
======================================
## Note:
*p<0.1; **p<0.05; ***p<0.01
```

This estimation generates the following models:

\[y = 4.952+ 11.508*hdi + u\] \[y = -9.502+33.312*hdi + -0.0001*gdp.per.capita+u\] \[y = -8.509+34.125*hdi + -0.0001*gdp.per.capita+-0.431*generation+u\]

Looking at these models it could be said that the most accurate model is model2 because model1 has too few variables and the

authencity of generation variable in model3 could be disputed. Looking at these modesl, we see that a small incease in hdi can cause to a positive effect in suicide rates, that is as hdi increases so does suicde rates per 100k. Moreover, we see a negative relation with gdp.per.capita. That means that the more the gdp per capita the less the suicide rates. The effect of this could seem small due to the small integer in the front. However, gdp per capita in this data measures as high as 126,352. This means that for certain values above a threshold, the gdp per capita definetely effects the outcome. model3 is no good because I was not able to reflect the generation variable effectively into the equation. The values of this variable hold no significance.

Inference

```
#running the F-test on all the models
anova(model1,model2,model3)
## Analysis of Variance Table
##
## Model 1: suicides100k ~ hdi
## Model 2: suicides100k ~ hdi + gdp.per.capita
## Model 3: suicides100k ~ hdi + gdp.per.capita +
generation
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   7205 2318913
## 2   7204 2298155  1   20757.9 65.172 7.983e-16 ***
## 3   7203 2294234  1    3921.3 12.311  0.000453 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

Looking at the results of the F statistic, it could be said that none of the models are able to reject the null hypothesis. model2's F statistic is 65.172 which is significantly larger than 0.05. We fail to reject the

null hypothesis. Therefore, it could be said that there is no clear trend between suicide rates per 100k and hdi, gdp per capita, and or generation that the individual belongs to.

Sources

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