**Text

Description automatically generated**

**GLOBAL SUICIDE RATE ANALYSIS**

**FINAL PROJECT**

**Submitted By:**

Durga Bhanu Nayak

Class: ALY6010-70516

College of Professional Studies, Northeastern University

Professor: Amin Karimpour

Date: 12/15/2021

# **Introduction**

According to World Health Organization 2021 statistics, there are more than 700,000 people suicide each year and 77% of suicides happen in the countries with low or middle income ("Suicide", 2021). Especially among young people whose age are between 15 and 29, suicide is the second main factor to (World Health Organization, 2014). This is a public health problem which negatively affect countries and the whole world. However, suicide is a complex phenomenon. It could be caused by different reasons, e.g., mental health and personal economic status (Bhatia et al., 2006), macro environment, society issues etc. To prevent suicide, one of the most significant steps is to find the reasons behind different groups of people. This report wants to analyze a data on global suicide rate and to find different suicide rate of different group of people, also try to figure out the causes.

Many countries have realized the problem of suicide and take several actions. From a global survey conducted by IASP and WHO in 2013, 31% of the countries’ government have adopted strategies and plans to prevent suicide (Arensman, 2017). In this report, we also want to compare the suicide result in different countries and their trends.

**Purpose of this dataset:**

* Explore the suicide trends in different countries.
* Find out the factors which correlate with suicide.
* Compare the suicide rate of different groups of people and find the riskiest kind of person.

**Data Source:**

this data set is downloaded from Kaggle which combines the sources from United Nations Development Program (2018), World Bank (2018), World Health Organization and a data set named “Suicide in the 21 Century”.

**Data Description**

* **Data title**: The Suicide Rate in Different Countries from 1987 to 2014

This data set is about the suicide rate in different countries from 1987 to 2014, with detailed information on suicide people persona and macro environment indicators.

* **Data type**: By using glimpse(suicide\_rate) and summary(suicide\_rate), we could find,
* There are 27,820 rows and 12 columns (variables) in this data set.
* 6 Categorical variables: Country, Year, Sex, Age (range), Country-Year, Generation
* 2 Discrete variables: Suicide Number, Population
* 4 Continuous variables: Suicides per 100k Population, HDI for Year, GDP for Year, GDP per Capital

# **Data Cleaning Process**

Following are the steps involved in cleaning the data set:

* **Step 1 Summary of the raw dataset**

A screenshot of a computer

Description automatically generated with medium confidence

This helps in getting the summary of the whole dataset including all the variables involved. From this we can see the variable classification easily.

* **Step 2 Glimpse of the raw dataset**

A close-up of a document

Description automatically generated with medium confidence

This gives us the total rows and columns present in the current dataset. Also, it does enables us to know about the data type of each variable and their sample data in a clean manner.

* **Step 3 Dropping null rows where HDI.for.year field is having NA**

Now that we are aware of of the null values in the dataset for column HDI.for.year we are going to remove these values by dropping the records with the null values using the below command.



* **Step 4 Renaming required columns**

This step is required to rename the exisiting columns with logical names.



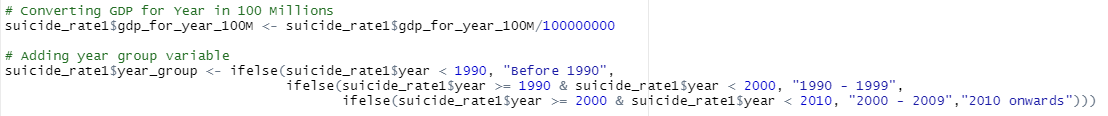
* **Step 5 Fixing variables**

Text

Description automatically generated with low confidence

Using above steps, we have fixed some of the common formatting issues to make the values more readable.

* **Step 6: Transforming GDP for year field and adding Year\_Group field**



Here, we have converted the GDP for year variable values in 100 million so that the values does not goes out of scope. Also, we have added a new year group and categorized the Year values based on that.

* **Step 7: Making variable factors (nominal and ordinal)**

Text

Description automatically generated

We have converted te categorical variables into factors which is further segregated into nominal and ordinal categories.

* **Step 8: Dropped redundant column**

Text

Description automatically generated with medium confidence

In this step, we have removed the redundant column which is ‘Country\_year’ from the dataset and saved the values in a new dataframe named suicide\_rate2 which can be coonsidered as a clean dataset to work with.

# **Exploratory Data Analysis**

1. **Descriptive Statistics of the Dataset (Numerical Variables)**

Graphical user interface

Description automatically generated

**Observations:**

1. Average suicide number throughout the dataset can be seen as 206.12.
2. Also, the range of the suicide number can be seen as 11767.
3. As the skew value is coming around 8.48, this shows the dispersion of the suicide rate data is positively skewed which also mean that most of the suicide data points are lesser than the mean value.
4. The overall dispersion of the data is widely spread due to the high value of the Standard deviations.
5. Out of the average population of 1852K people, 206 people on an average are committing suicide.
6. Average Human Development Index is coming around 0.78 and the standard deviation is 0.09 which shows that the dispersion is very compact.
7. Average GDP per capita is found to be 21074.37 and looking by its skewness which is 1.76, the distribution of data points is close to the normal distribution.
8. **The trend of suicide rate of different genders from 1985 to 2015**

Chart, line chart

Description automatically generated

**Observations:**

1. From 1985 to 2015, the suicide rate of male is higher than female in each of the years.
2. The suicide rate of female is below 10 persons out of 100K population from 1985 to 2015, while male has the suicide rate above 15.
3. There is a peak suicide rate for both female and male in 1995. The male suicide rate in that year reached nearly 25 persons out of 100K population
4. From 1985 and 1995, we could see an increase of the suicide rate for both male and female, while from 1995 onwards, the suicide rate started to fall.
5. **Box plot of GDP per year per 100 million Population**

**Chart

Description automatically generated**

**Observations**

1. The GDP per year per 100 million population is concentrated mostly between 0-3500.
2. Both the above plots show a positive skew in the distribution i.e., most of the data points are situated on left sides.
3. Although most of the data shows positive skew, the outliers value goes up to 200000.
4. 50% of all the GDP distribution lies between 143.0751 - 3115.3950.
5. The data remains mostly in the range of 0 - 7422.9345.
6. **Box plot and histogram of GDP per Capita**

Timeline

Description automatically generated with low confidence

**Observations**

1. The distribution of the GDP per capital from 1985 to 2015 is wide. 50% of the GDP per capital is above 12,584.
2. The average GPD per capital from 1985 to 2015 is 210,744 which is higher than the mean value. The mean value has been pulled over by the outliers.
3. The GPD per capital value below 12,584 is more compact than those higher than 12,584.
4. The values between 0 and 6000 have the highest value of GDP per capital.
5. **Box plot and histogram of Human Development Index for the Year**

Chart, box and whisker chart

Description automatically generated

**Observations**

1. The HDI values in the period from 1985 to 2015 have a wide distribution.
2. The average HDI is close to median, while the distribution is still slightly negatively skewed, which means most of the HDI values are higher than 0.8.
3. There are several outliers in the boxplot, but the outlier value is 0.483. Here, we would need to detect which year and country these indicates and how the suicide rates are.
4. The HDI values are compact in the range between 0.7 and 0.9.
5. **Clustered bar plot of Suicide per Generation**

Chart, bar chart

Description automatically generated

**Observations:**

1. From the above graph we can see that, people from Generation X have committed the highest suicides than other generations with global suicides more than 500K.
2. Generation Z people seems to be the happy people who has committed the least suicides among all.
3. Of all the people, male suicides are visibly much higher than female ones across all the generations.
4. Female suicides across ‘Silent’, ‘Boomers’ and ‘Generation X’ generation type can be seen close to each other.
5. **Bar plot of suicide rate in each year group**

Chart, bar chart

Description automatically generated

**Observations:**

* People have committed more suicides after the year of 2010, which can be considered as the global digital era.
* It can also be observed that suicides in the decade 2000 – 2009 has slightly increased from its past decade.
* There has been an increasing trend in global suicide rate with the time.

1. **Bar plot of suicide rate in each age group**

Chart, funnel chart

Description automatically generated

**Observations**

* The highest number of suicides recorded were from the age group of 35-54 years, which is the age group for common workforce.
* The population belonging from the age group of 5-14 years are recorded with least number of suicides.
* We see the decreasing in the age group of 55-74 years where the older generation is also doing suicides.
* There is a decreasing trend as the population age increases from 55-74 years and 75+ years
* There’s an increasing trend of suicide rates from the age of 5-54 years after 54 years when the population grows older there’s a decreasing trend.

# **Hypothesis Test**

**Question1:**

**Is the average suicide rate in the year group 2010 onwards higher than the population mean?**

Chart, bar chart

Description automatically generated

**Reason:** from the above bar chart, we see that the 2010 onwards has the highest suicide rate, therefore, we want to see whether the suicide rate in this group year is higher than the population mean.

**One-sample t-test**

**Step 1: State the hypothesis and Claim**

Null Hypothesis: The average suicide rate in the year group 2010 onwards has the same average suicide rate of the whole population.

Alternative Hypothesis: The average suicide rate in the year group 2010 onwards is higher than the average suicide rate of the whole population. (Claim)

* H0: μ1 = μ0
* H1: μ1 > μ0 (Claim)

**Step 2: Check t-test assumptions**

1. ***To check the normality, I used QQ Plot and histogram with density curve.***

**Text

Description automatically generated**

Chart, line chart

Description automatically generated

From the above graph, we could see that the sample data is skewed.

However, since the sample size is 4740 which is much greater than 30, then it is justified to use t-test.

1. ***The data are from the random sample***

Therefore it is justified to use one-sample t-test to compare the year group average suicide rate with the population mean.

**Step 3: T-Test Result**

Graphical user interface, text, application, email

Description automatically generated

* **Analysis**

With the confidence level of 0.95 (**α = 0.05)**

* The test value is -5.723;
* The p-value is greater than the significance level of 0.05;
* The confidence interval is 95% IC[10.36, ∞];

Since p-value is greater than the significance level, the null hypothesis should not be rejected.

* **Result**

There is not enough evidence to support the claim that the average suicide rate in the year group 2010 onwards is higher than the average suicide rate of the whole population 11.99.

**Question2:**

**Is the average suicide rate of male GenX higher than the average suicide rate of female GenX?**

Chart, line chart

Description automatically generated

**Reason:** from the above line graph, we see that male has higher suicide rate than female over the 30 years. Also from the previous report, it is shown that GenX has the highest suicide rate. Therefore, this report want to test whether male GenX has higher suicide rate than female GenX.

**Independent Two-sample t-test**

**Step 1: State the hypothesis and Claim**

Null Hypothesis : The average male GenX suicide rate has no difference with the average female GenX suicide rate.

Alternative Hypothesis: The average male GenX suicide rate higher than the average female GenX suicide rate. (Claim)

* H0: μ1 = μ2
* H1: μ1 > μ2 (Claim)

**Step 2: Check t-test assumptions**

1. ***To check the normality, I used QQ Plot and histogram with density curve.***

Text

Description automatically generated

Chart, line chart, histogram

Description automatically generated

From the above graph, we could see that the sample data is skewed.

However, since the sample sizes are much greater than 30, then it is justified to use t-test.

1. ***Variance check***

Text

Description automatically generated

From the variance test, it is shown that the p-value is smaller than the significance level of 0.05, therefore, the variances of males suicides rates and female suicide rates are different.

1. ***The data are from the random sample***
2. ***The 2 sample are independent with different subjects.***

Therefore it is justified to use independent two-sample t-test to compare male GenX and female GenX suicide rates. Since the variances are not equal, this report will use Welch two-sample t-test.

**Step 3: T-test result**

Graphical user interface, text, application, email

Description automatically generated

* **Analysis:**

**With the confidence level of 0.95 (α = 0.05)**

* The test value t is 25.721 which is in the critical region.
* The p-value is small than the significance level of 0.05.
* 0 is not contained in the confidence interval 95%CI [12.423, ∞].

**Therefore, the null hypothesis should be rejected.**

* **Results**

This report decides to reject the null hypothesis. There is enough evidence to support the claim that the average male GenX suicide rate higher than the average female GenX suicide rate.

**Question 3:**

**Is the HDI of GenX lower than the HDI of GenZ and Boomers?**

Chart, bar chart

Description automatically generated

**Reason:** from the above bar chart, it is shown that GenX has the highest suicide rate, followed by Boomers. GenZ has the lowest suicide rate. This report want to check whether HDI has some influence on the suicide rate. Therefore, we will compare the HDI of GenX and Boomers, also the HDI of GenX and GenZ.

**Step 1: State the hypothesis and Claim**

1. ***GenX vs GenZ***

Null Hypothesis : The average HDI of GenX suicide rate has no difference with the HDI of GenZ

Alternative Hypothesis: : The average HDI of GenX suicide rate is lower than the HDI of GenZ (Claim)

* H0: μ1 = μ2
* H1: μ1 > μ2 (Claim)

1. ***GenX vs Boomers***

Null Hypothesis : The average HDI of GenX suicide rate has no difference with the HDI of Boomers

Alternative Hypothesis: : The average HDI of GenX suicide rate is lower than the HDI of Boomers (Claim)

* H0: μ1 = μ2
* H1: μ1 > μ2 (Claim)

**Step 2: Check t-test assumptions**

1. ***To check the normality, I used QQ Plot and histogram with density curve.***

Chart, line chart

Description automatically generated

From the above graph, we could see that all the 3 samples are approximately normal.

1. ***Variance Check***

Text

Description automatically generated

The p-value is smaller than the significant level of 0.05, therefore the variances of GenX HDI and GenZ HDI are different.

Text

Description automatically generated

The p-value is greater than the significant level of 0.05, therefore the variances of GenX HDI and Boomers HDI are equal.

1. ***The data are from the random samples;***
2. ***The 3 samples are independent with different subjects.***

Therefore, it is justified to use independent two-sample t-test to compare GenX HDI and GenZ HDI and also to compare GenX HDI and Boomers HDI.

**Step 3: T-test results**

Graphical user interface, text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

* **Analysis:**

**With the confidence level of 0.95 (α = 0.05)**

* The 2 test values are in their critical regions accordingly.
* The p-values are smaller than the significance level of 0.05.
* 0 is not contained in the confidence interval 95%CI [12.423, ∞].

**Therefore, both the null hypothesis in the above 2 tests should be rejected.**

* **Results**

This report decides to reject the 2 null hypothesizes . There is enough evidence to support the claim that the average GenZ HDI is lower than GenX HDI, and GenZ HDI is lower than Boomers HDI.

**Question 4:**

**For the year 2014, does the average suicide rate of Asia continent equals to the global average suicide rate?**

**One-sample t-test**

**Step 1: State the hypothesis and Claim**

***Null Hypothesis*** **H0**: For year 2014, average suicide rate of Asia per 100K population is **equal** to the global average suicide rate.

***Alternative Hypothesis*** **H1**: The average suicide rate of Asia is **not equal** to the global average suicide rate

* H0: μ1 = μ0 (Claim)
* H1: μ1 ≠ μ0

**Step 2: Check t-test assumptions**

1. ***To check the distribution, we have plotted graph using histogram with density curve.***

Text

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated

From the above graph, we could see that the sample data is positively skewed.

However, since the sample size is 192 which is greater than 30, then it is justified to use t-test.

1. ***Testing strategy applied***

We will be using one-sample t-test to compare the average suicide rate of Asia continent with the global population mean.

**Step 3: T-Test Result**

Text

Description automatically generated

**Analysis**

With the confidence level of 0.95 (**α = 0.05)**

* The t value is -6.3092
* The p-value is coming as 0.000000001909 which is much smaller than the significance level of 0.05
* The confidence interval is 95% IC [5.029, 7.706]

Since p-value is smaller than the significance level, we can reject the Null Hypothesis.

**Result**

As we have got smaller p-value, we will be rejecting the H0 and state that, average suicide rate for Asia continent is not equal to the global average suicide rate. Here, from our data, average suicide rate for Asia is coming as 6.36 per 100K population for year 2014.

**Question 5:**

**For the year 2010, does countries with higher GDP per capita has lower suicide rate than the countries with lower GDP per capita?**

**Reasoning:**

To test this hypothesis, we have taken a sample country (‘United States’) which is having high GDP and a country (‘Ukraine’) which is having the lower GDP per capita income. Also, we have taken the data from year 2010, where global suicide rate went high compared to previous years among all the generations.

**Two-sample t-test**

**Step 1: State the hypothesis and Claim**

***Null Hypothesis*** **H0**: Average suicide rate United States is **less than** the average suicide rate of Ukraine.

***Alternative Hypothesis*** **H1**: Average suicide rate United States is **greater than** the average suicide rate of Ukraine

* H0: μ1 < μ0 (Claim)
* H1: μ1 > μ0

**Step 2: Check t-test assumptions**

1. ***To check the distribution, we have plotted graph using histogram with density curve and checked the normality using Shapiro Wilk test.***

Chart, histogram

Description automatically generated

Text

Description automatically generated with medium confidence Text

Description automatically generated with medium confidence

From the above data distribution and normality test, we can check that data distribution of United States suicide rate is close to normal distribution as the p-value is higher than 0.05. Whereas, the data distribution of Ukraine is skewed distribution as its p-value is slightly lower than the significance level.

1. ***Testing strategy applied***

We will be using two-sample t-test to compare the average suicide rate of United States with the average suicide rate of Ukraine as these represents two different groups.

**Step 3: T-Test Result**

Text

Description automatically generated

**Analysis**

With the confidence level of 0.95 (**α = 0.05)**

* The t value is -1.2201
* The p-value is coming as 0.8806 which is much greater than the significance level of 0.05
* The confidence interval is 95% IC [-21.04417, ∞]
  + Since p-value is higher than the significance level, it suggests that we do not have enough evidence to reject the Null Hypothesis.

**Result**

As we have got higher p-value, we will not be rejecting the H0 which states, average suicide rate for United States was lesser than the average suicide rate of Ukraine. Here, from our data, average suicide rate for United States is coming as 13.24 per 100K population and average suicide rate for Ukraine is 21.92 per 100K population for year 2010.

**Question 6**

**Is the mean suicide no of Generation X and Silent similar?**

Chart, box and whisker chart

Description automatically generated

**Reason:** The graph shows that both for Generation X and Silent generation the data lies mostly in the bottom and is positively skewed. This test is to check if there is any difference between both the generation’s mean.

**Step 1: State the hypothesis and Claim**

***GenX vs Silent***

Null Hypothesis : The average suicide\_no of GenX has no difference with the suicide\_no of GenZ

Alternative Hypothesis: : The average suicide\_no of GenX isn’t equal with the suicide\_no of GenZ.

* H0: μ1 = μ2
* H1: μ1 ≠ μ2 (Claim)

**Step 2: Check t-test assumptions**

**Graphical user interface, text, application

Description automatically generated**

The t-test is done taken mean of Generation X suicide\_no and Silent generation suicide\_no into consideration.

1. The data from the dataset is stored in data frame and then filtered to two different objects
2. The suicide number of Generation X is saved in fil\_genxno and suicide number of Silent is stored in fil\_gensno
3. T-test is carried out considering CI of 0.95.

Graphical user interface, text, application, email

Description automatically generated

**Analysis**

With the confidence level of 0.95 (**α = 0.05)**

* The t value is 1.6822
* The p-value is coming as 0.09262 which is much greater than the significance level of 0.05
* The degree of freedom is 3716.9
  + Since p-value is higher than the significance level, it suggests that we do not have enough evidence to reject the Null Hypothesis.

**Result**

As the p-value is higher, we will not be rejecting the H0 which states, average suicide rate for Generation X is similar with Silent generation. Here, from our data, average suicide numbers for Generation X is coming as 262.8456 and average suicide number of Silent is 221.5607.

# **Correlation and Regression Analysis**

**Question 1:**

**Will high GDP per Capita contribute to better Human Development?**

**Reason:**

Human development index (HDI) is a reflection of 3 aspects of human: life expectancy, education and the living standard ("Human Development Index (HDI) | Human Development Reports", n.d.). This report wants to check the correlation between GDP per Capita and Human Development to see whether these 2 factors have any linear relationship.

1. **Correlation Coefficient: GDP per Capita & HDI**



The correlation coefficient of the GDP per Capita and HDI is 0.831 which very close to 1. This means GDP per Capita and HDI has very strong positive linear correlation. When GDP per Capita increases, HDI increases.

1. **Hypothesis test to check the correlation in population**

Text

Description automatically generated

**Analysis:**

The Pearson’s product-moment correlation test show the p-value is close to zero, which means the correlation between GDP per Capita and HDI in the population also significant.

1. **Regression Analysis**
2. ***Normality Check***

Chart, line chart

Description automatically generated

**Analysis:**

From the above QQ plots and histogram, it is shown that:

* GDP per Capital values are slightly positively skewed.
* HDI values are approximately from a normal distribution.

1. ***Regression Analysis***

Text

Description automatically generated

**Analysis:**

* The slope of the formula is 0.000004, and intercept is 0.692. Therefore the formula should be Y = 0.000004X + 0.692, where X is GDP per Capita and Y is HDI.
* The F-statistics p-value is close to 0, which means the correlation between the 2 variables is significant.
* R-squared = 0.6914 which means 69.14% of the variability of HDI could be explained by the changes of GDP per Capita.

1. ***Homoscedasticity Check***

Chart, scatter chart

Description automatically generated

**Analysis:**

The relationship between the x and y in the above plot is nearly linear, so the regression line in the regression analysis could be used to make predictions.

1. **Visualization: scatter plot of GDP per Capita and HDI**

Chart, scatter chart

Description automatically generated

69.14% of the HDI value could be predicted by GDP per Capita by using the formula **Y = 0.000004X + 0.692**, where X is GDP per Capita and Y is HDI.

**Question 2:**

**How the GDP per Capita affects the suicide rate for Generation X**

**Reason:** As the seen in graphs of Question 1 and Question 3, that there is increase in suicide rates after the year 2010 and over the years the highest number of suicides were by Generation X followed by Millennials. This test checks if GDP plays a major factor in suicide rates.

**Step 1: Removing the outliers**

Graphical user interface, text, application, email

Description automatically generated

* Above codes represented are used in removal of outliers for representing the regression
* Filtered the information about Generation X and found the minimum outlier value using min() function and stored the information with no outliers in an object.
* Removed the outlier using the filter technique, and checked the outliers using boxplot.

**Step 2: Generating the regression plot and the correlation data**

Graphical user interface, text, application

Description automatically generated Text

Description automatically generated

**Step 3: Graph representation of GDP per Capita’s effect on the average suicide rate for Generation X**

Chart, scatter chart

Description automatically generated

* Using scatterplot() to create the regression graph between GDP per capita and average suicide rates of Generation X.
* regLine() is used to add visualization in the regression line, “smooth” is used to add a smooth graph providing the concentration of the datapoints.
* The regression analysis among the variable is obtained using lm() and is stored in an object, when summarized, the p value is less than 0.05 which shows there’s a correlation between GDP per capita and average suicide rates.
* The value is 0.07097 which is 7.097% which shows degree of variation which can be observed from this model for average suicide rate corresponding to GDP per capita for Generation X.

**Question 3:**

**How the GDP per Capita affects the suicide for Millennials (generation)**

**Step 1: Removing the outliers**

Graphical user interface, text, application

Description automatically generated

* Above codes represented are used in removal of outliers for representing the regression
* Filtered the information about Millennial and found the minimum outlier value using min() function and stored the information with no outliers in an object.
* Removed the outlier using the filter technique, and checked the outliers using boxplot().

**Step 2: Generating the regression plot and the correlation data**

Graphical user interface, text, application

Description automatically generated Text, letter

Description automatically generated

**Step 3: Graph representation of GDP per Capita’s effects on the average suicide rate for Millennials**

Chart, scatter chart

Description automatically generated

* Using scatterplot() to create the regression graph between GDP per capita and average suicide rates of Millennials.
* regLine() is used to add visualization in the regression line.
* The regression analysis among the variable is obtained using lm() and is stored in an object, when summarized, the p value is less than 0.05 which shows there’s a correlation between GDP per capita and average suicide rates.
* The value is 0.03531 which is 3.531% which shows degree of variation which can be observed from this model for average suicide rate corresponding to GDP per capita for Millenials.

**Question 4:**

**Will GDP per Capita affect the suicide rate in Generation X in different Gender Group?**

**Reason:**

From the above analysis, the correlation between the GDP per Capita and Suicide rate among Generation X is weak. To further explore, this report also check the correlation in 2 sub gender group, male and female to see if the correlation will have any changes and whether linear regression model could be used to predict the suicide rate based on the GDP per Capita.

1. **Correlation coefficient: GDP per Capita & Suicide Rate**
2. ***Generation X Female***



The correlation coefficient of **GDP per Capita and GenX female Suicide Rate** is 0.321 which is not very high. This means the correlation between this 2 variables are not very strong, but they have a positive correlation. Therefore when GDP per Capita increases, the suicide rate of GenX female will increase.

1. ***Generation X Male***



The correlation coefficient of **GDP per Capita and GenX male Suicide Rate** is 0.167 which is also small. This means the correlation between this 2 variables are not very strong, but they have a positive correlation. Therefore when GDP per Capita increases, the suicide rate of GenX male will increase.

1. **Hypothesis test to check the correlation in population**
2. ***Generation X Female***

Graphical user interface, text, application

Description automatically generated

**Analysis:**

The Pearson’s product-moment correlation test show the p-value is close to zero, which means the correlation between GDP per Capita and suicide rate of Generation X female in the population is also significant.

1. ***Generation X Male***

Graphical user interface, text, application

Description automatically generated

**Analysis:**

The Pearson’s product-moment correlation test show the p-value is close to zero, which means the correlation between GDP per Capita and suicide rate of Generation X female in the population is also significant.

1. **Regression Analysis**
2. **Normality Check**
3. Generation X Female GDP per Capita & Suicide Rate

Chart, line chart

Description automatically generated

**Analysis:**

From the above QQ plots and histogram, it is shown that:

* GDP per Capital values are slightly positively skewed.
* Generation X Female suicide rate values are approximately from a normal distribution.

1. Generation X Female GDP per Capita & Suicide Rate

Chart, line chart

Description automatically generated

**Analysis:**

From the above QQ plots and histogram, it is shown that:

* GDP per Capital values are slightly positively skewed.
* Generation X Male suicide rate values are approximately from a normal distribution.

1. Regression Analysis
2. ***Generation X Female***

Text

Description automatically generated

**Analysis:**

* The slope of the formula is 0.00006, and intercept is 2.745. Therefore the formula should be Y = 0.00006X + 2.745, where X is GDP per Capita and Y is GenX female suicide rate
* The F-statistics p-value is close to 0, which means the correlation between the 2 variables is significant.
* R-squared = 0.1033 which means 10.33% of the variability of GenX female suicide rate could be explained by the changes of GDP per Capita.

Text

Description automatically generated

**Analysis:**

* The slope of the formula is 0.00006, and intercept is 5.030. Therefore the formula should be Y = 0.00006X + 5.030, where X is GDP per Capita and Y is GenX male suicide rate
* The F-statistics p-value is close to 0, which means the correlation between the 2 variables is significant.
* R-squared = 0.028 which means 2.8% of the variability of GenX male suicide rate could be explained by the changes of GDP per Capita.

1. **Visualization: scatter plot of Generation X suicide rate by Gender Group**

Chart, scatter chart

Description automatically generated

* There is significant correlation between GDP per capita and suicide rate in both gender among Generation X.
* However, the relationships are not linear: we could not use linear regression model to precisely predict the suicide rate based on the value of GDP per Capita.

**Question 5:**

**How HDI is affecting the average suicide rate in Europe?**

Chart, bar chart

Description automatically generated

**Observation:**

* Europe is having the highest suicide number recorded till now among all the continents
* Africa is having the least suicide number among all.
* Female from Asia have committed the highest suicide among females from other continents.
* Male suicide rate is higher than the female suicide rate across all the continents.

As we saw from the above graph that Europe is having the highest suicide rate among all the populations, we are curious to see how Europe is trying to handle this situation.

For this purpose, we are going to see the correlation between average suicide rate in Europe till now with its Human Development Index.

Text

Description automatically generated

***Normality check for average suicide rate and HDI in Europe***

After preparing the data for Europe continent, we will check the normality of the data distribution in Europe for variable ‘avg\_suicide\_rate’ and ‘hdi’.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

* Here the above average suicide rate and human development index data seems to be close to normally distributed.

***Checking correlation between the average suicide rate with the human development index:***



***Pearson’s Product-Moment Correlation test:***

Text

Description automatically generated

Plotting the correlation with a linear regression in a graph:

Chart, scatter chart

Description automatically generated

Here we can clearly see that there is a weak negative correlation present between the average suicide rate and the human development index in Europe.

This suggests that Europe is trying to deal with the suicide rate issue by working over its Human Development Index.

As the HDI value is increasing, the suicide rate is gradually decreasing.

**Regression Analysis:**

Here, we will be checking on the linear regression line that can be plotted between average suicide rate and the human development index.



A screenshot of a computer

Description automatically generated with medium confidence

**Observations:**

* Here, we can see that after using the regression formula, we are getting the intercept value of **46.76** and slope as **-36.918**.
* We can prepare an equation using the above values as:

***y’ = 46.76 + (-36.918) \**** *HDI*

*Here, y’ represents the estimated suicide rate based on the input HDI value to the equation*

* We are also getting the R2 value as 0.08391
* This suggests that our linear regression model is about 8% efficient in predicting the estimated value for average suicide rate based on the HDI.

# **Conclusion**

1. The average suicide rate in the year group of 2010 onwards is greater than the population mean value of 11.99. There might be some reasons behind it, like in some countries, the suicide rates increased due to the micro or macro reasons.
2. Male has higher suicide rate than female.
3. HDI seems have influence on the suicide rate, and from the 2 t-tests, we see that the higher suicide rate generation has lower HDI. Therefore, we may need to further check the correlation between the suicide rate and HDI.
4. Even though Asia covers the largest population of the whole world, its average suicide rate per 100K population comes to be lower than the global suicide rate recorded for the year 2014. This suggests that countries from different continents are also having suicide rates comparable to Asia.
5. As from the results obtained after applying the t-test, we can assume that countries with higher GDP per capita income are having lesser suicide rates than the countries with lower GDP per capita. These assumptions can also be influenced by various other factors like work stress, loans etc.
6. The generations are divided into 6 categories, where the highest number of suicides were made by Generation X, followed by Boomers, Silent, Millennials, G.I. Generation, Generation Z.
7. Although, there is a difference in the total suicide numbers, and the T-test shows the mean of generation X is 262.8456 and Silent is 221.5607.
8. The p value is 0.09262 which is above 0.05, so we will be accepting the null hypothesis, which is the mean suicide numbers of both the generations.
9. As GDP increases, Human Development Index(HDI) also increases.
10. The average suicide rate increases among the Generation X with increase in GDP.
11. GDP per Capita has strong correlation with Suicide rate in both Generation X Male and Female Group, while the linear regression could not explain the variability.
12. Suicide rate in Europe is also displaying a negative correlation with its HDI.

# **Reference**

* Arensman, E. (2017). Suicide Prevention in an International Context. *Crisis*, *38*(1), 1-6. https://doi.org/10.1027/0227-5910/a000461
* Bhatia, M., Verma, S., & Murty, O. (2006). Suicide Notes: Psychological and Clinical Profile. *The International Journal of Psychiatry In Medicine*, *36*(2), 163-170. https://doi.org/10.2190/5690-cmgx-6a1c-q28h
* Human Development Index (HDI) | Human Development Reports. Hdr.undp.org. Retrieved 14 December 2021, from http://hdr.undp.org/en/content/human-development-index-hdi?utm\_source=EN&utm\_medium=GSR&utm\_content=US\_UNDP\_PaidSearch\_Brand\_English&utm\_campaign=CENTRAL&c\_src=CENTRAL&c\_src2=GSR&gclid=Cj0KCQiAweaNBhDEARIsAJ5hwbcuztp3Q5WjvyH6LhjmOaTDQZO8etcPN6CbPwyOFebT3iotAqrs9nkaArSPEALw\_wcB.
* Suicide. World Health Organization. (2021). Retrieved 15 November 2021, from https://www.who.int/news-room/fact-sheets/detail/suicide.
* World Health Organization. (2014). *Preventing suicide: a global imperative.* Switzerland.
* Home | The World Happiness Report. (2021, March 31). WHR. https://worldhappiness.report/

# **Appendix**

* The corresponding R script to create the above report is attached seperately.



* The dataset used to analyze and populate the report is included in the report.

