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**

```
> summary(suicide_rate)
  i..country            year          sex                 age             suicides_no       population        suicides.100k.pop country.year
 Length:27820       Min.   :1985   Length:27820       Length:27820       Min.   :    0.0   Min.   :     278   Min.   :  0.00    Length:27820
 Class :character   1st Qu.:1995   Class :character   Class :character   1st Qu.:    3.0   1st Qu.:   97498   1st Qu.:  0.92    Class :character
 Mode  :character   Median :2002   Mode  :character   Mode  :character   Median :   25.0   Median :  430150   Median :  5.99    Mode  :character
                    Mean   :2001                                         Mean   :  242.6   Mean   : 1844794   Mean   : 12.82
                    3rd Qu.:2008                                         3rd Qu.:  131.0   3rd Qu.: 1486143   3rd Qu.: 16.62
                    Max.   :2016                                         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
```

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**

```
> glimpse(suicide_rate)
Rows: 27,820
Columns: 12
$ i..country          <chr> "Albania", "Albania", "Albania", "Albania", "Albania", "Albania", "Albania", "Albania", "Albania", "Albania", "Albania", "Alban~
$ year                <int> 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1988, 1988, 1988, 1988, 1988, 1988, 1988, 1988, 1988, 1~
$ sex                 <chr> "male", "male", "female", "male", "male", "female", "female", "female", "male", "female", "female", "male", "female", "male", "~
$ age                 <chr> "15-24 years", "35-54 years", "15-24 years", "75+ years", "25-34 years", "75+ years", "35-54 years", "25-34 years", "55-74 year~
$ suicides_no         <int> 21, 16, 14, 1, 9, 1, 6, 4, 1, 0, 0, 0, 2, 17, 1, 14, 4, 8, 3, 5, 5, 4, 0, 0, 2, 18, 15, 6, 12, 7, 5, 2, 1, 0, 0, 0, 12, 9, 5, 7~
$ population          <int> 312900, 308000, 289700, 21800, 274300, 35600, 278800, 257200, 137500, 311000, 144600, 338200, 36400, 319200, 22300, 314100, 140~
$ suicides.100k.pop   <dbl> 6.71, 5.19, 4.83, 4.59, 3.28, 2.81, 2.15, 1.56, 0.73, 0.00, 0.00, 0.00, 5.49, 5.33, 4.48, 4.46, 2.85, 2.71, 2.03, 1.91, 1.79, 1~
$ country.year        <chr> "Albania1987", "Albania1987", "Albania1987", "Albania1987", "Albania1987", "Albania1987", "Albania1987", "Albania1987", "Albani~
$ HDI.for.year        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ gdp_for_year....    <chr> "2,156,624,900", "2,156,624,900", "2,156,624,900", "2,156,624,900", "2,156,624,900", "2,156,624,900", "2,156,624,900", "2,156,6~
$ gdp_per_capita....  <int> 796, 796, 796, 796, 796, 796, 796, 796, 796, 796, 796, 796, 769, 769, 769, 769, 769, 769, 769, 769, 769, 769, 769, 769, 769, 833, 83~
$ generation          <chr> "Generation X", "Silent", "Generation X", "G.I. Generation", "Boomers", "G.I. Generation", "Silent", "Boomers", "G.I. Generatio~
```

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.

```
# Dropping null rows for HDI.for.year as it contains lots of null values
suicide_rate1 <- suicide_rate[is.na(suicide_rate$HDI.for.year) == FALSE,]
```

- **Step 4 Renaming required columns**

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

```
suicide_rate1 <- suicide_rate1 %>%
  rename(gdp_for_year_100M = gdp_for_year....,
         gdp_per_capita = gdp_per_capita....,
         country_year = country.year,
         country = names(suicide_rate[1]))
```

- **Step 5 Fixing variables**

```
suicide_rate1$age<- gsub(" years", "", suicide_rate1$age)
suicide_rate1$gdp_for_year_100M <- gsub(",", "", suicide_rate1$gdp_for_year_100M)
suicide_rate1$gdp_for_year_100M <- as.numeric(suicide_rate1$gdp_for_year_100M)
suicide_rate1$sex <- ifelse(suicide_rate1$sex == "male", "Male", "Female")
```

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**

```
# Converting GDP for Year in 100 Millions
suicide_rate1$gdp_for_year_100M <- suicide_rate1$gdp_for_year_100M/100000000

# Adding year group variable
suicide_rate1$year_group <- ifelse(suicide_rate1$year < 1990, "Before 1990",
                                   ifelse(suicide_rate1$year >= 1990 & suicide_rate1$year < 2000, "1990 - 1999",
                                          ifelse(suicide_rate1$year >= 2000 & suicide_rate1$year < 2010, "2000 - 2009","2010 onwards")))
```

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)**

```
# Factors:

# Nominal
lst <- c("country","sex")
suicide_rate1[lst] <- lapply(suicide_rate1[lst], as.factor)

# Ordinal
#unique(suicide_rate1$age)
suicide_rate1$age <- factor(suicide_rate1$age, ordered = T,
                            levels = c("5-14","15-24","25-34","35-54","55-74","75+"))

#unique(suicide_rate1$generation)
suicide_rate1$generation <- factor(suicide_rate1$generation, ordered = T,
                                   levels = c("G.I. Generation","Silent","Boomers","Generation X","Millenials","Generation Z"))

suicide_rate1$year_group <- factor(suicide_rate1$year_group, ordered = T,
                                   levels = c("Before 1990","1990 - 1999","2000 - 2009", "2010 onwards"))
```

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

- **Step 8: Dropped redundant column**

```
# dropping the column country_year as it is redundant
suicide_rate2 <- subset(suicide_rate1, select = -c(country_year))
```

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)

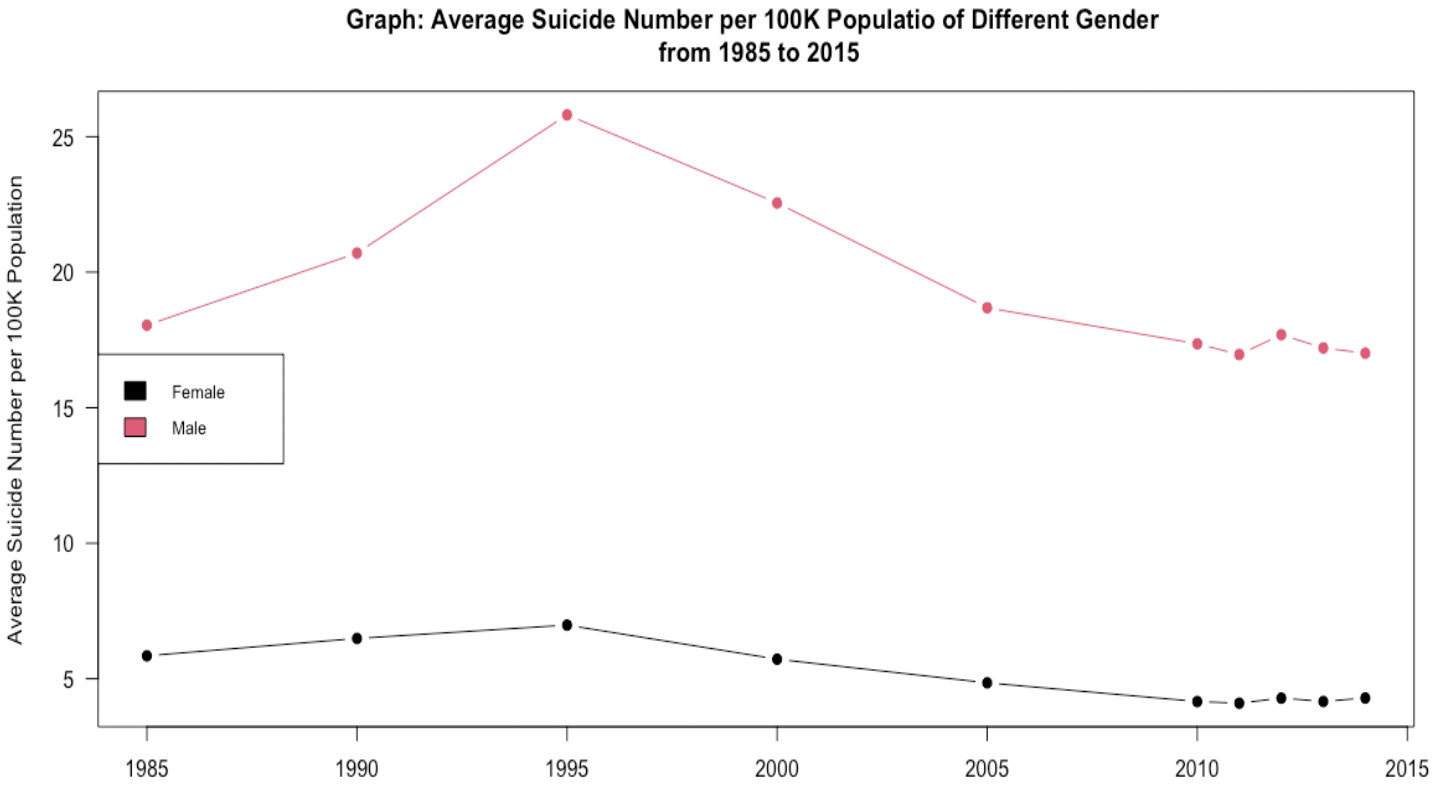
Descriptive Statistics of Numerical variables

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
suicides_no	1	8364	206.12	681.00	27.00	69.36	40.03	0.00	11767.00	11767.00	8.48	97.64	7.45
population	2	8364	1852173.38	3969753.63	472250.50	926648.22	626207.99	875.00	43509335.00	43508460.00	4.73	30.48	43406.68
suicides.100k.pop	3	8364	11.99	17.36	5.72	8.27	8.04	0.00	187.06	187.06	2.90	11.56	0.19
HDI.for.year	4	8364	0.78	0.09	0.78	0.78	0.11	0.48	0.94	0.46	-0.30	-0.65	0.00
gdp_for_year_100M	5	8364	5476.64	17201.06	617.58	1802.63	887.28	3.96	174276.09	174272.13	6.74	53.55	188.08
gdp_per_capita	6	8364	21074.37	22579.19	12584.00	17137.23	13639.92	313.00	126352.00	126039.00	1.76	3.40	246.89

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.

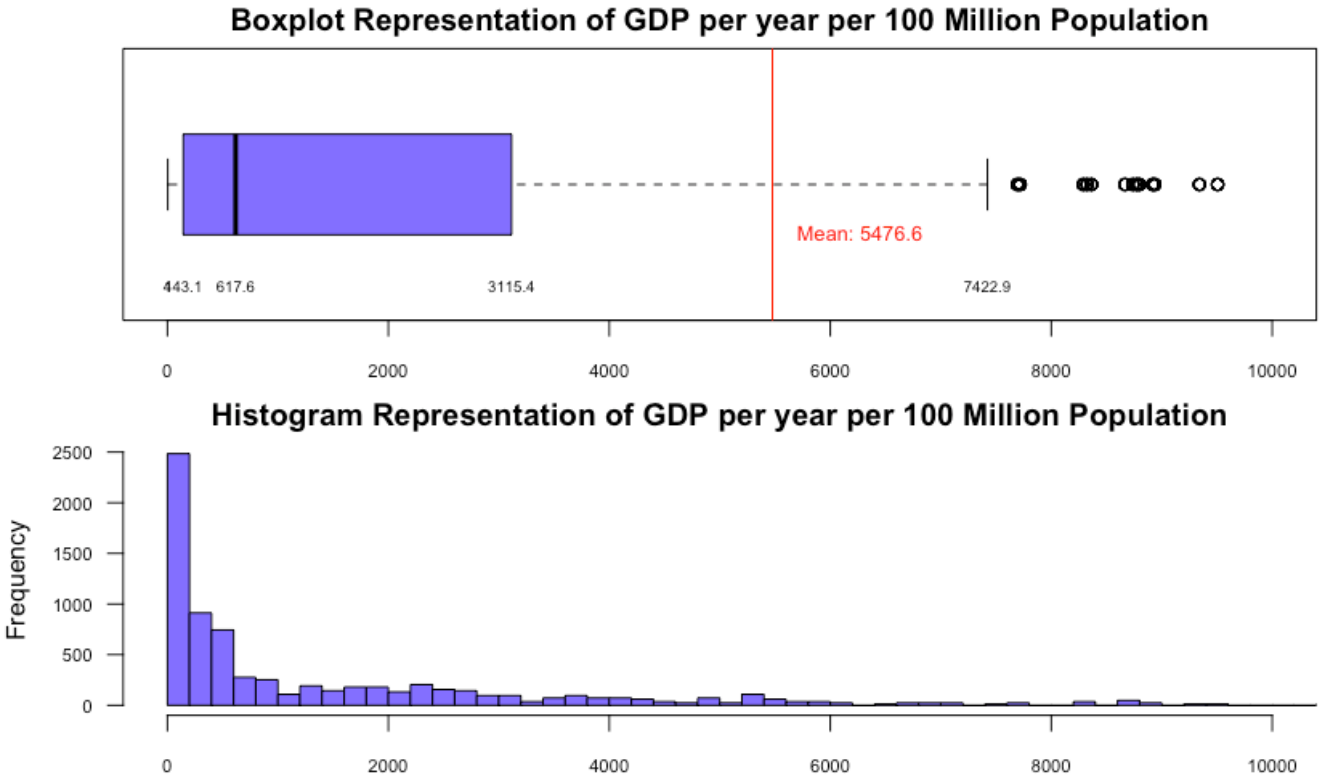
2. The trend of suicide rate of different genders from 1985 to 2015



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.

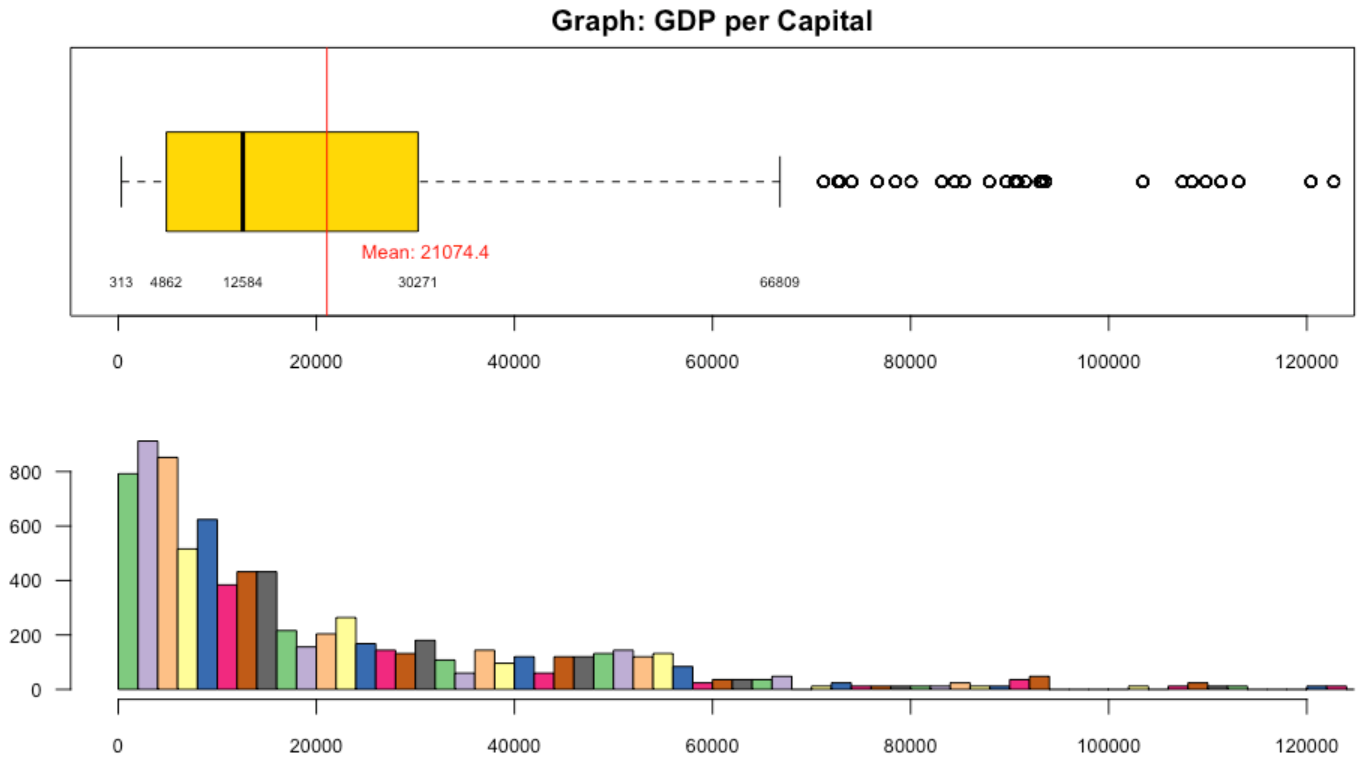
3. Box plot of GDP per year per 100 million Population



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.

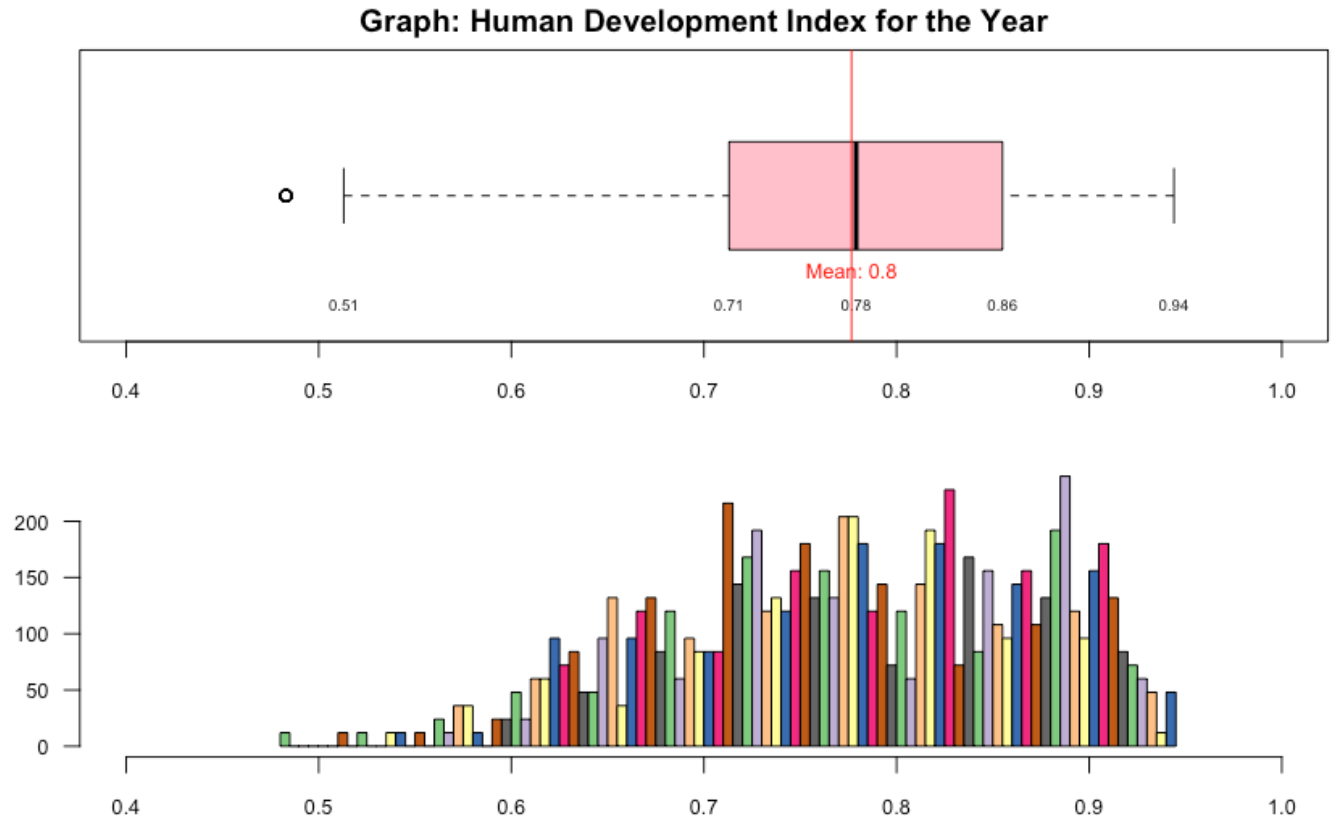
4. Box plot and histogram of GDP per Capita



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.

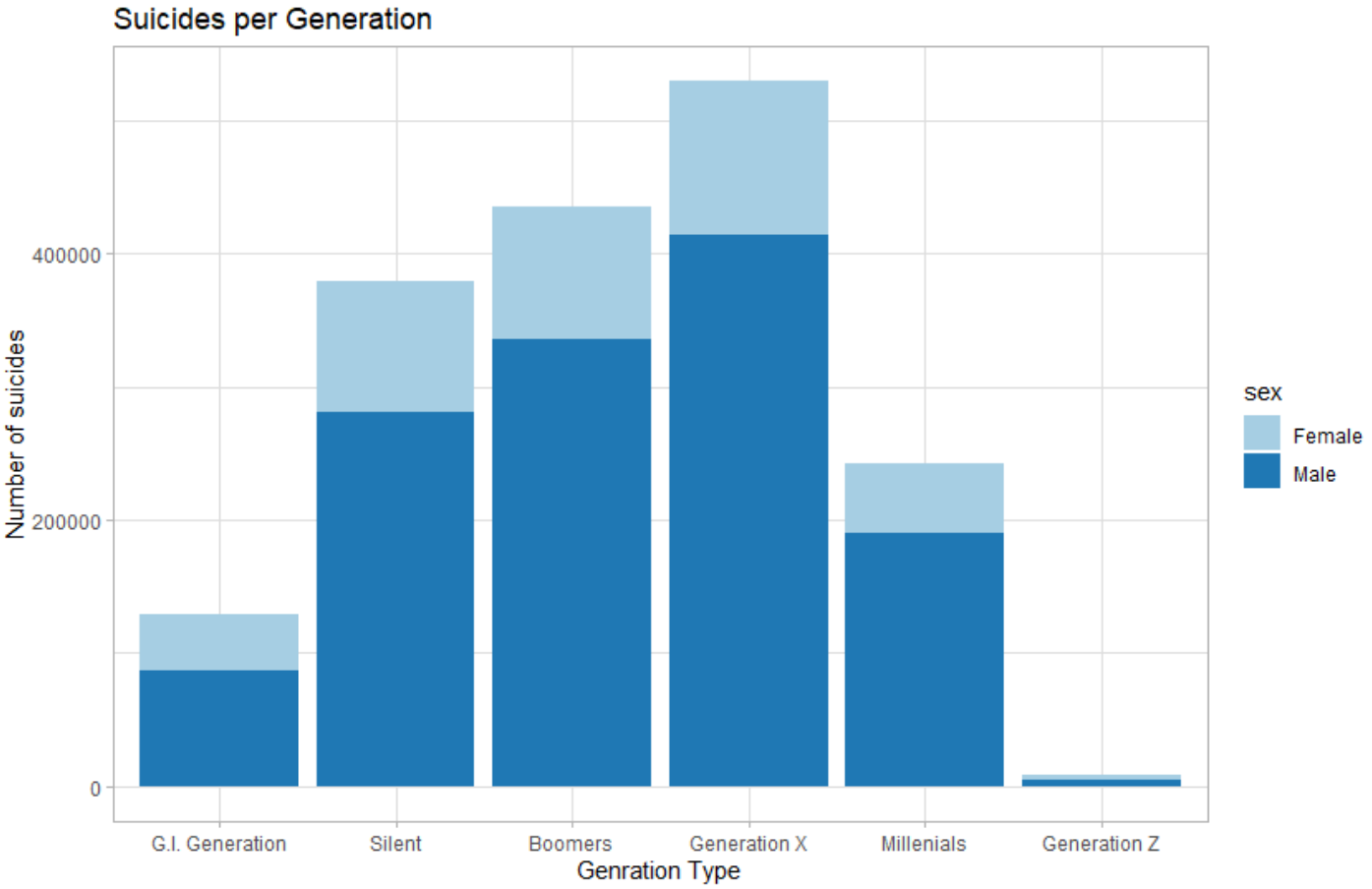
5. Box plot and histogram of Human Development Index for the Year



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.

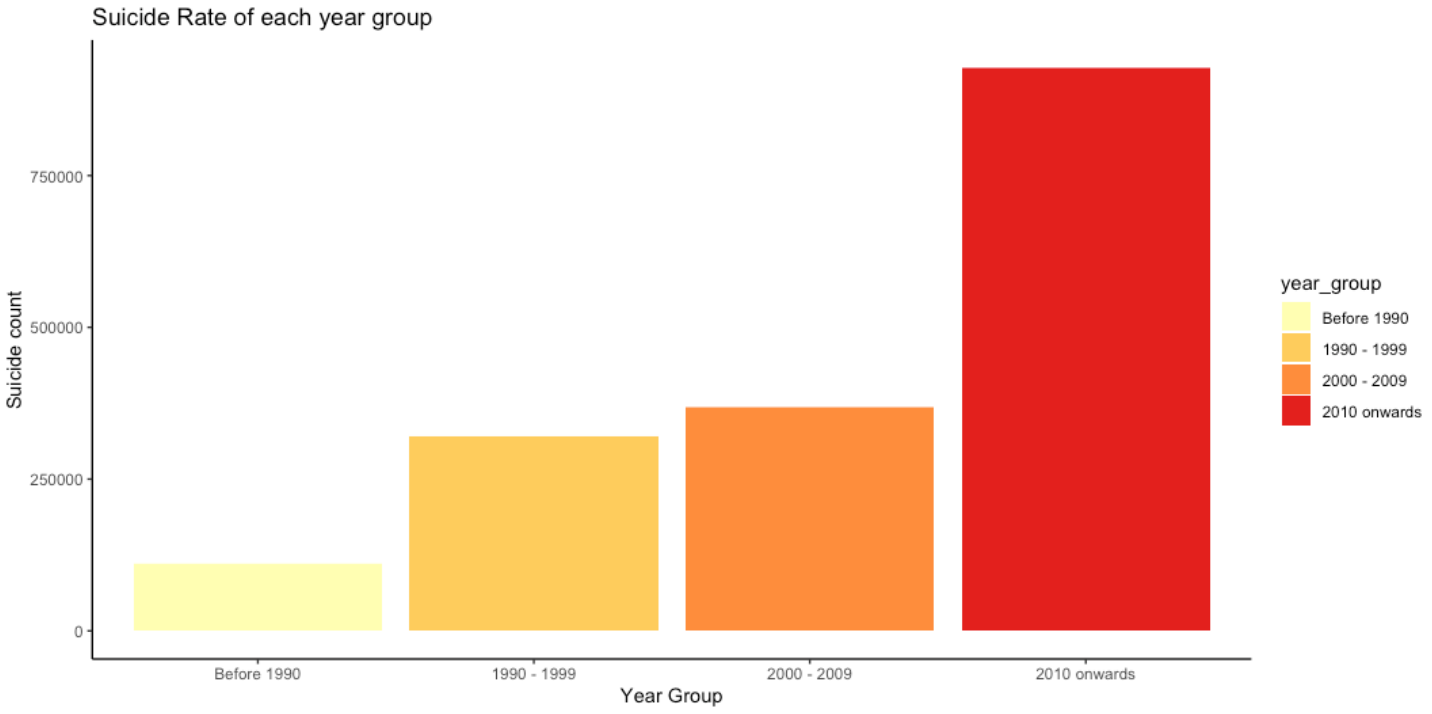
6. Clustered bar plot of Suicide per Generation



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.

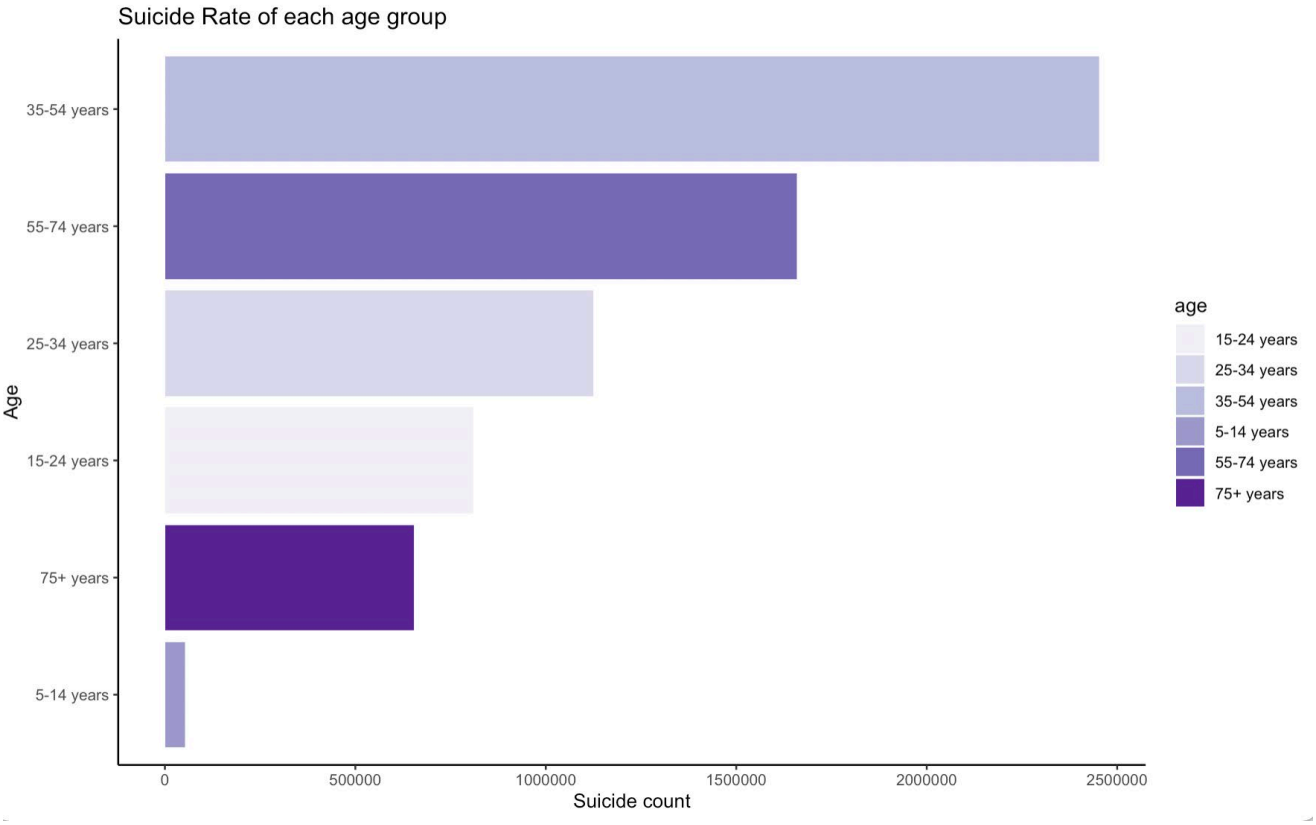
7. Bar plot of suicide rate in each year group



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.

8. Bar plot of suicide rate in each age group

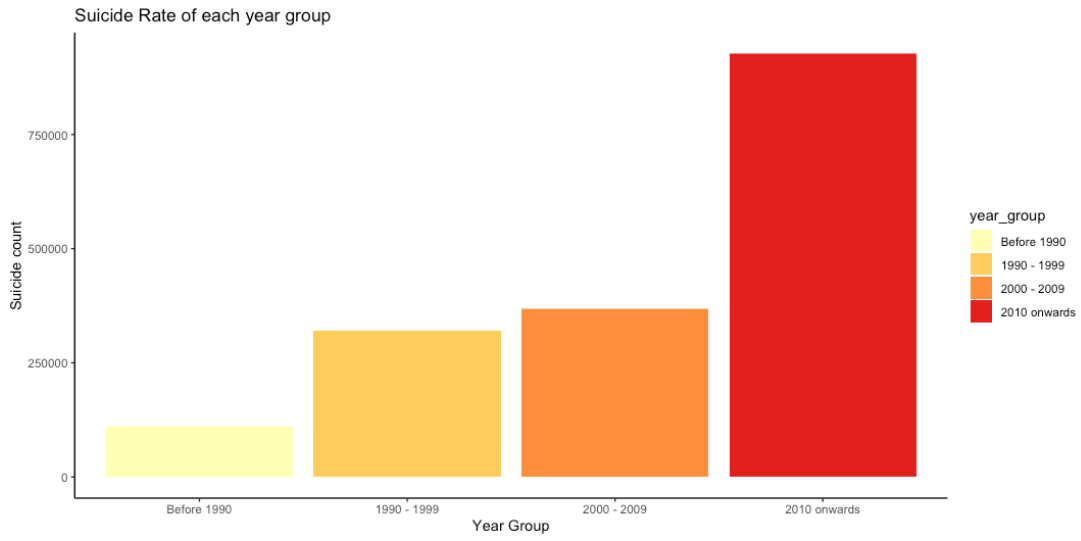


Observations

- 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?



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.

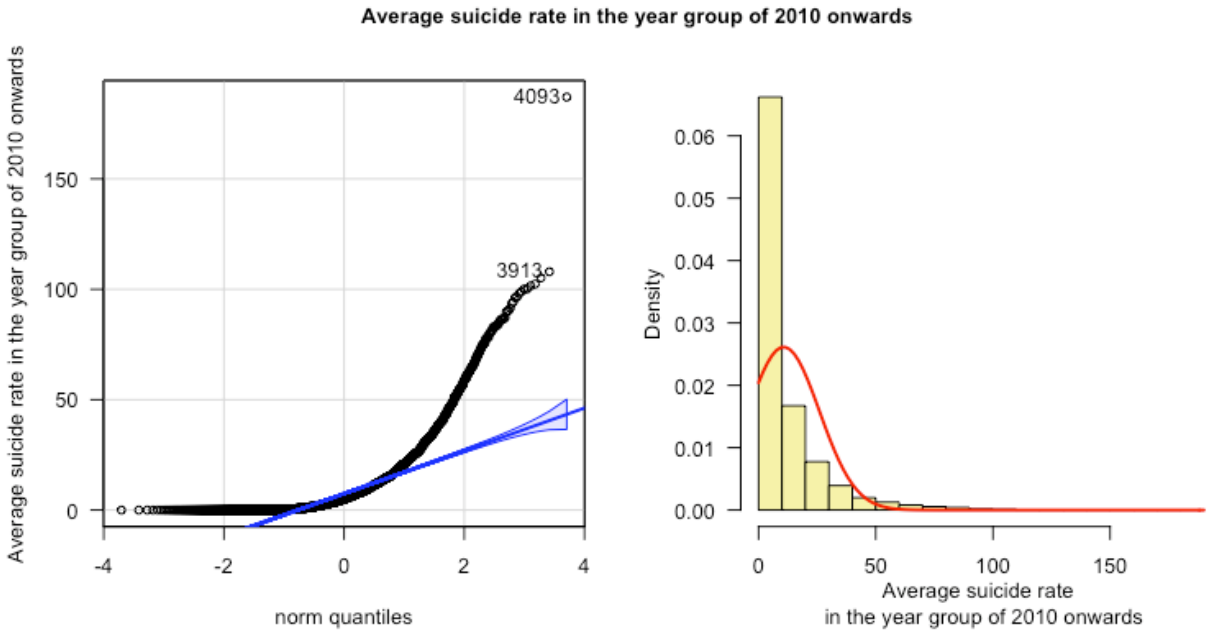
Alternative Hypothesis: The average suicide rate in the year group 2010 onwards is higher than the average suicide rate of the whole population. (Claim)

- H_0: $\mu_1 = \mu_0$
- H_1: $\mu_1 > \mu_0$ (Claim)

Step 2: Check t-test assumptions

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

```
year<-subset(suicide_rate2, subset = (suicide_rate2$year_group == "2010 onwards"))

par(mfrow= c(1,2), cex= 0.8, mgp = c(3,1,0))
qqPlot(year$suicides.100k.pop, ylab = "Average suicide rate in the year group of 2010 onwards", las = 1)
hist(year$suicides.100k.pop, freq = FALSE, las = 1, col = "#F5EFA4", xlab = "Average suicide rate \n in the year group of
2010 onwards", main = "")
curve(dnorm(x, mean= mean(year$suicides.100k.pop), sd = sd(year$suicides.100k.pop)), col = "red", lwd =2, add = TRUE)
mtext("Average suicide rate in the year group of 2010 onwards", side = 3, line = -2, outer = TRUE, cex = 0.8, font = 2)
```



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.

2) ***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

```
> t.test(year$suicides.100k.pop, mu = 11.99, alternative = "greater", conf.level = 0.95)

        One Sample t-test

data:  year$suicides.100k.pop
t = -5.723, df = 4739, p-value = 1
alternative hypothesis: true mean is greater than 11.99
95 percent confidence interval:
 10.35715      Inf
sample estimates:
mean of x
 10.72174
```

- **Analysis**

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

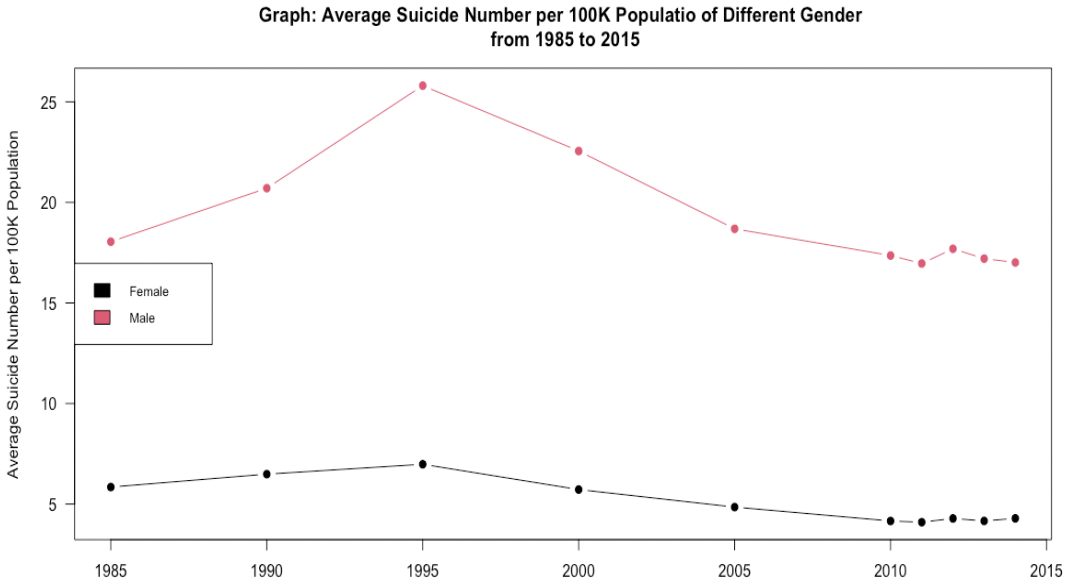
- The test value is -5.723;
- The p-value is greater than the significance level of 0.05;
- The confidence interval is 95% IC[10.36, ∞];

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?



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.

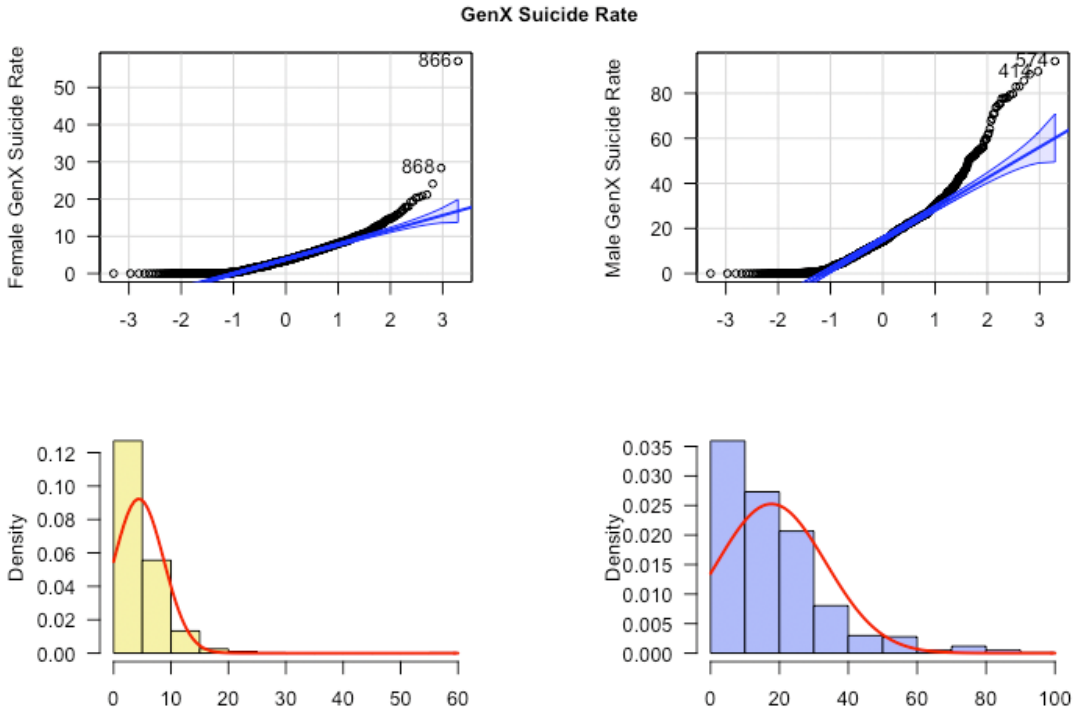
Alternative Hypothesis: The average male GenX suicide rate higher than the average female GenX suicide rate. (Claim)

- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 > \mu_2$ (Claim)

Step 2: Check t-test assumptions

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

```
par(mfrow= c(2,2), cex= 0.8, mgp = c(3,1,0), mai = c(0.5,1,0.5,0.5))
qqPlot(fgenx$suicides.100k.pop, ylab = "Female GenX Suicide Rate", las = 1)
qqPlot(mgenx$suicides.100k.pop, ylab = "Male GenX Suicide Rate", las = 1)
hist(fgenx$suicides.100k.pop, freq = FALSE, las = 1, col = "#F5EFA4", xlab = "Female GenX Suicide
Rate", main = "")
curve(dnorm(x, mean= mean(fgenx$suicides.100k.pop), sd = sd(fgenx$suicides.100k.pop)), col = "red",
lwd =2, add = TRUE)
hist(mgenx$suicides.100k.pop, freq = FALSE, las = 1, col = "#A4B3F5", xlab = "Male GenX Suicide
Rate", main = "")
curve(dnorm(x, mean= mean(mgenx$suicides.100k.pop), sd = sd(mgenx$suicides.100k.pop)), col = "red",
lwd =2, add = TRUE)
mtext(" GenX Suicide Rate", side = 3, line = -2, outer = TRUE, cex = 0.8, font = 2)
```



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.

2) Variance check

```
> var.test(mgenx$suicides.100k.pop, fgenx$suicides.100k.pop)

        F test to compare two variances

data:  mgenx$suicides.100k.pop and fgenx$suicides.100k.pop
F = 13.38, num df = 1006, denom df = 1006, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 11.82367 15.14123
sample estimates:
ratio of variances
          13.38002
```

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.

3) The data are from the random sample
4) 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

```
> t.test(mgenx$suicides.100k.pop, fgenx$suicides.100k.pop, alternative = "greater",
 conf.level = 0.95, var.equal = F)

        Welch Two Sample t-test

data:  mgenx$suicides.100k.pop and fgenx$suicides.100k.pop
t = 25.721, df = 1155.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 12.42263      Inf
sample estimates:
mean of x mean of y
17.680735  4.408679
```

- **Analysis:**

With the confidence level of 0.95 (α = 0.05)

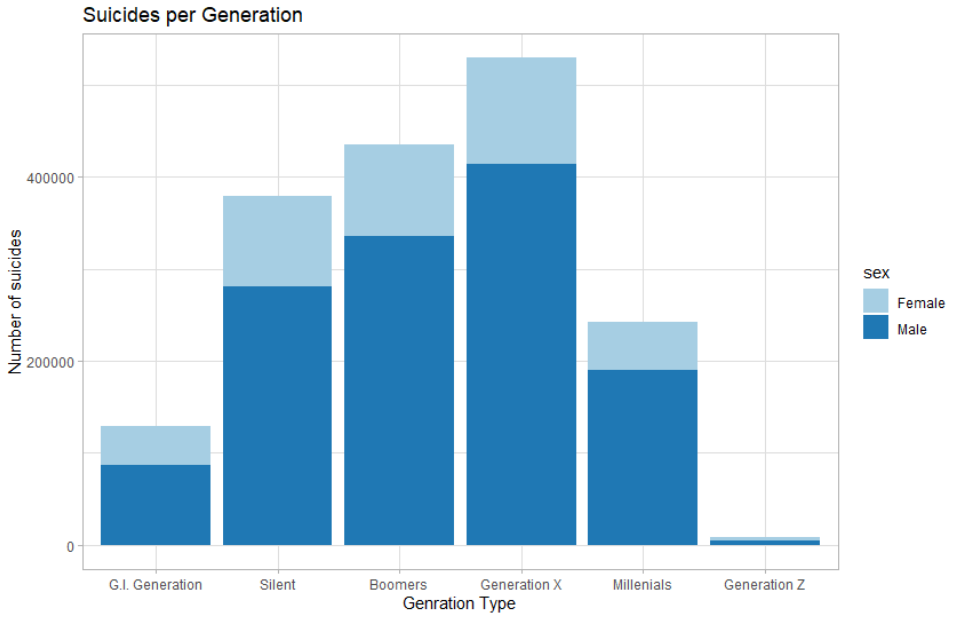
- The test value t is 25.721 which is in the critical region.
- The p-value is small than the significance level of 0.05.
- 0 is not contained in the confidence interval 95%CI [12.423, ∞].

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?



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)

- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 > \mu_2$ (Claim)

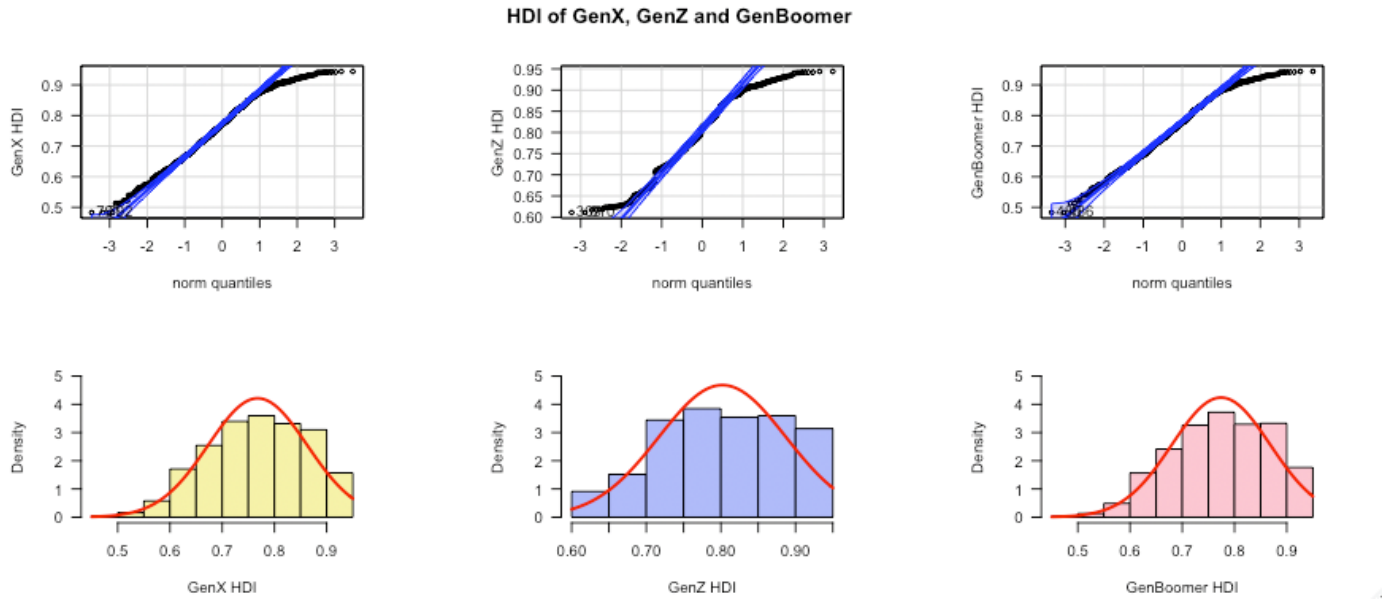
2) GenX vs Boomers

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)

- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 > \mu_2$ (Claim)

Step 2: Check t-test assumptions

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



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

2) Variance Check

```
> var.test(GenX$HDI.for.year, GenZ$HDI.for.year)

        F test to compare two variances

data:  GenX$HDI.for.year and GenZ$HDI.for.year
F = 1.2348, num df = 2013, denom df = 789, p-value = 0.000493
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.097333 1.385223
sample estimates:
ratio of variances
          1.234757
```

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

```
> var.test(GenX$HDI.for.year, GenBoom$HDI.for.year)

        F test to compare two variances

data:  GenX$HDI.for.year and GenBoom$HDI.for.year
F = 1.0123, num df = 2013, denom df = 1225, p-value = 0.8143
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9149845 1.1186387
sample estimates:
ratio of variances
           1.01233
```

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

3) The data are from the random samples;
4) 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

```
> t.test(GenX$HDI.for.year, GenZ$HDI.for.year, alternative = "less", conf.level = 0.95, var.equal = F)

        Welch Two Sample t-test

data:  GenX$HDI.for.year and GenZ$HDI.for.year
t = -9.2282, df = 1592, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        -Inf -0.02800376
sample estimates:
mean of x mean of y
0.7678798 0.8019620
```

```
> t.test(GenX$HDI.for.year, GenBoom$HDI.for.year, alternative = "less", conf.level = 0.95, var.equal = T)

        Two Sample t-test

data:  GenX$HDI.for.year and GenBoom$HDI.for.year
t = -1.8573, df = 3238, p-value = 0.03168
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
         -Inf -0.000725242
sample estimates:
mean of x mean of y
0.7678798 0.7742349
```

- **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 $\mathbf{H_0}$: For year 2014, average suicide rate of Asia per 100K population is **equal** to the global average suicide rate.

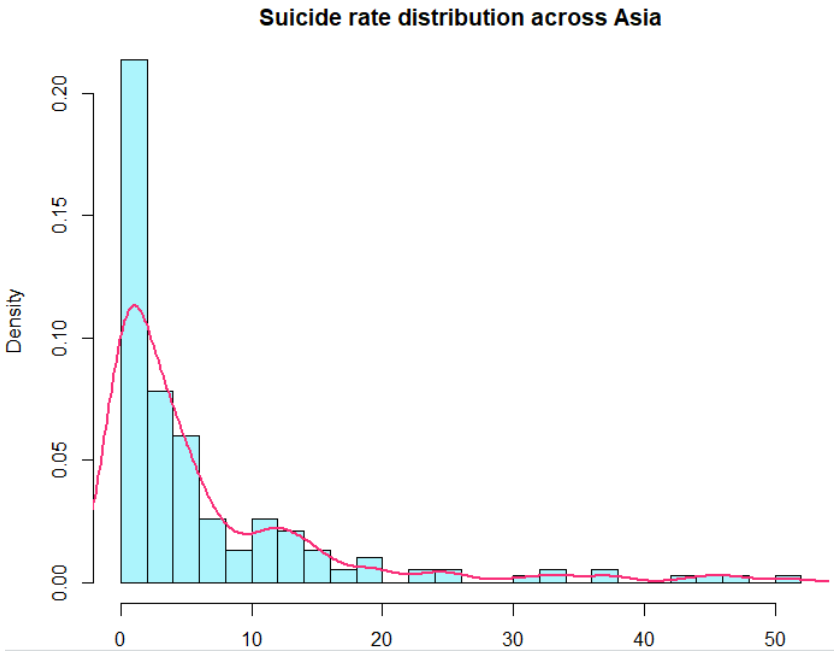
Alternative Hypothesis $\mathbf{H_1}$: The average suicide rate of Asia is **not equal** to the global average suicide rate

- H_0: $\mu_1 = \mu_0$ (Claim)
- H_1: $\mu_1 \neq \mu_0$

Step 2: Check t-test assumptions

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

```
df_year_2014 <- suicide_rate2[which(suicide_rate2$year=="2014"),]
mean_global_srate <- mean(df_year_2014$suicides.100k.pop)
srate_aisa_2014 <- df_year_2014[which(df_year_2014$continent == "Asia"),"suicides.100k.pop"]

# Distribution of Suicide rate in Asia continent
par(mfcol = c(1,1))
hist(srate_aisa_2014, freq = F, breaks = 20, col = "#ACF4FC", main = "Suicide rate distribution across Asia",
     xlab = "Suicide rate per 100K population")
lines(density(srate_aisa_2014), col = "#F92F79", lwd = 2)
```



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.

2. 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

```
        One Sample t-test

data:  srate_aisa_2014
t = -6.3092, df = 191, p-value = 0.000000001909
alternative hypothesis: true mean is not equal to 10.6501
95 percent confidence interval:
 5.029370 7.706776
sample estimates:
mean of x
 6.368073
```

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 H_0 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

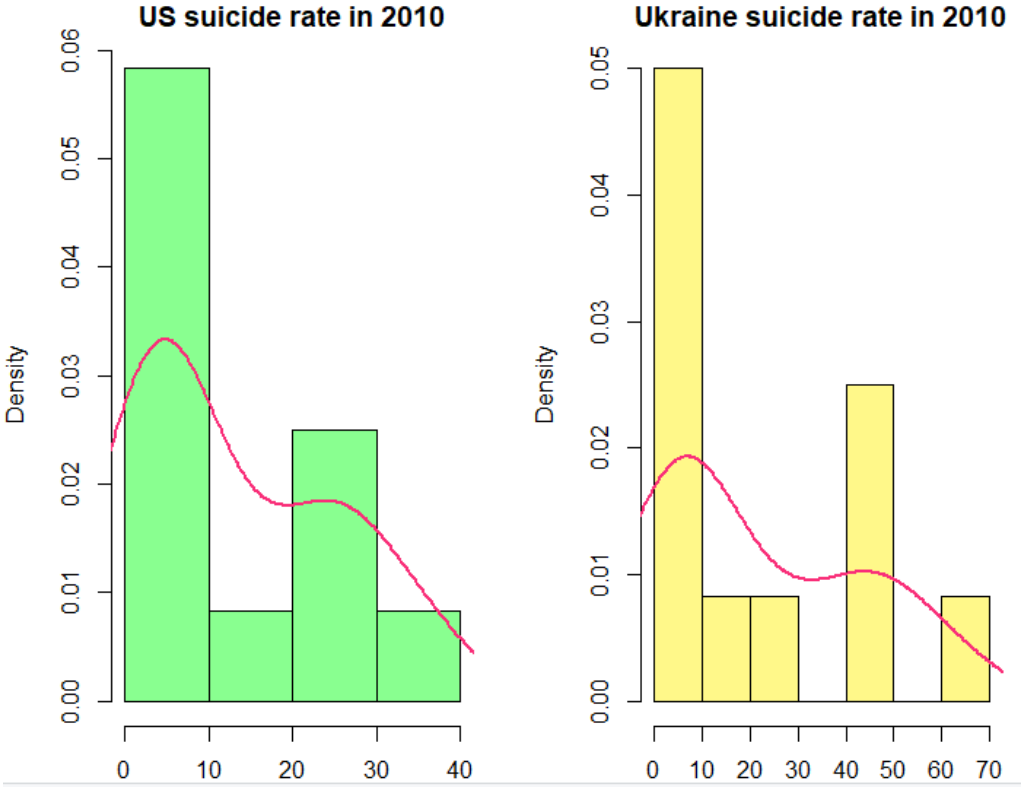
Step 1: State the hypothesis and Claim

Null Hypothesis $\mathbf{H_0}$: Average suicide rate United States is **less than** the average suicide rate of Ukraine.
Alternative Hypothesis $\mathbf{H_1}$: Average suicide rate United States is **greater than** the average suicide rate of Ukraine

- H_0: $\mu_1 < \mu_0$ (Claim)
- H_1: $\mu_1 > \mu_0$

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.***



```
> shapiro.test(us_srate)

        Shapiro-Wilk normality test

data:  us_srate
W = 0.88061, p-value = 0.08924
```

```
> shapiro.test(ukraine_srate)

        Shapiro-Wilk normality test

data:  ukraine_srate
W = 0.85326, p-value = 0.04029
```

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.

2. 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

```
> # Hypothess testing
> t.test(us_srate,ukraine_srate, alternative = "greater", conf.level = 0.95)

        Welch Two Sample t-test

data:  us_srate and ukraine_srate
t = -1.2201, df = 17.385, p-value = 0.8806
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -21.04417       Inf
sample estimates:
mean of x mean of y
 13.24083  21.92250
```

Analysis

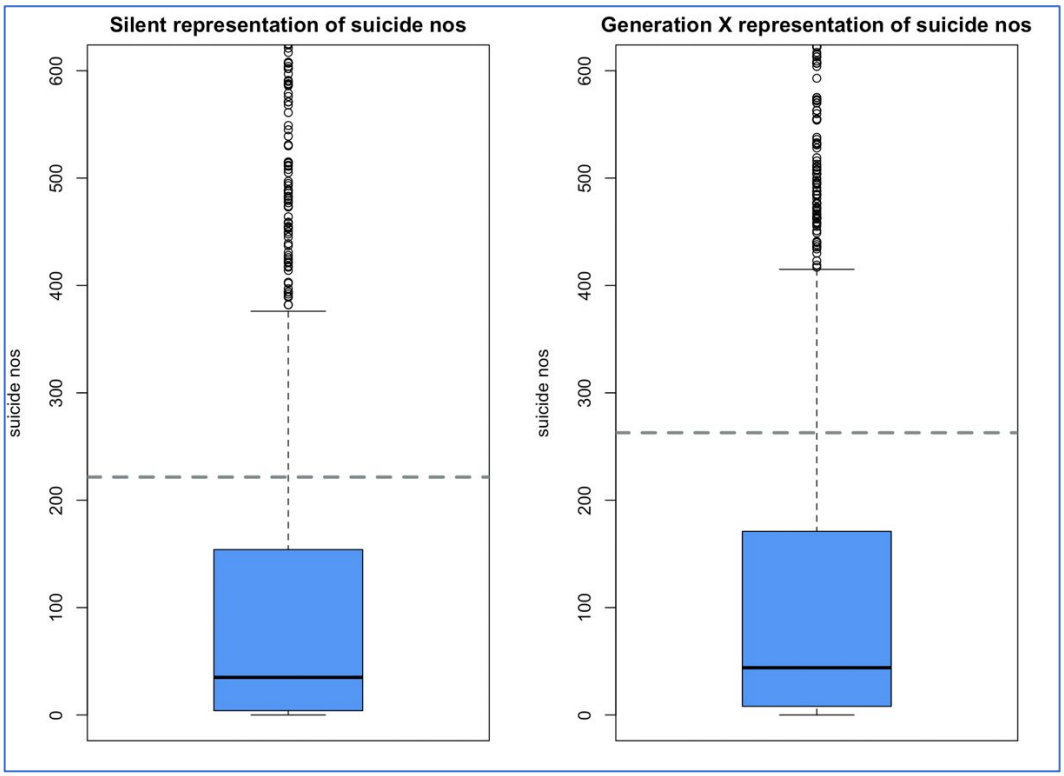
With the confidence level of 0.95 (**α = 0.05)**

- 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 H_0 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?



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.

- H_0: $\mu_1 = \mu_2$
- H_1: $\mu_1 \neq \mu_2$ (Claim)

Step 2: Check t-test assumptions

```
# MILESTONE 2

#H0 <- Generation X suicide no = Silent generation suicide no
#H1 <- Generation X suicide no ≠ Silent generation suicide no

# Assigning the variables in a data frame
gen_suicideno <- data.frame(suicide_rate2$generation, suicide_rate2$suicides_no)
gen_suicideno <- setNames(gen_suicideno, c('generations', 'suicide_nos'))

# Filtering different generations and assigning to different objects
fil_genx <- filter(gen_suicideno, generations == c('Generation X'))
fil_gens <- filter(gen_suicideno, generations == c('Silent'))

fil_genxno <- fil_genx$suicide_nos
fil_gensno <- fil_gens$suicide_nos

# T-test for suicide_no of Generation X and Silent
t.test(fil_genxno, fil_gensno)

meanx <- mean(fil_genxno)
means <- mean(fil_gensno)
par(mfcol = c(1:2))
boxplot(fil_gensno,ylim = c(0, 600), col = 'cornflowerblue')
abline(h = means, col=c("azure4"), lwd=3, lty=2)
title(main = "Silent representation of suicide nos", ylab = "suicide nos")
boxplot(fil_genxno,ylim = c(0, 600), col = 'cornflowerblue')
abline(h = meanx,col=c("azure4"),lwd=3, lty=2)
title(main = "Generation X representation of suicide nos", ylab = "suicide nos")
```

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.

```
> # T-test for suicide_no of Generation X and Silent
> t.test(fil_genxno, fil_gensno)

        Welch Two Sample t-test

data:  fil_genxno and fil_gensno
t = 1.6822, df = 3716.9, p-value = 0.09262
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.833577 89.403385
sample estimates:
mean of x mean of y
 262.8456  221.5607
```

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 H_0 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**

```
> cor(gdphdi$gdp_per_capita , gdphdi$HDI.for.year)
[1] 0.8315324
```

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.

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

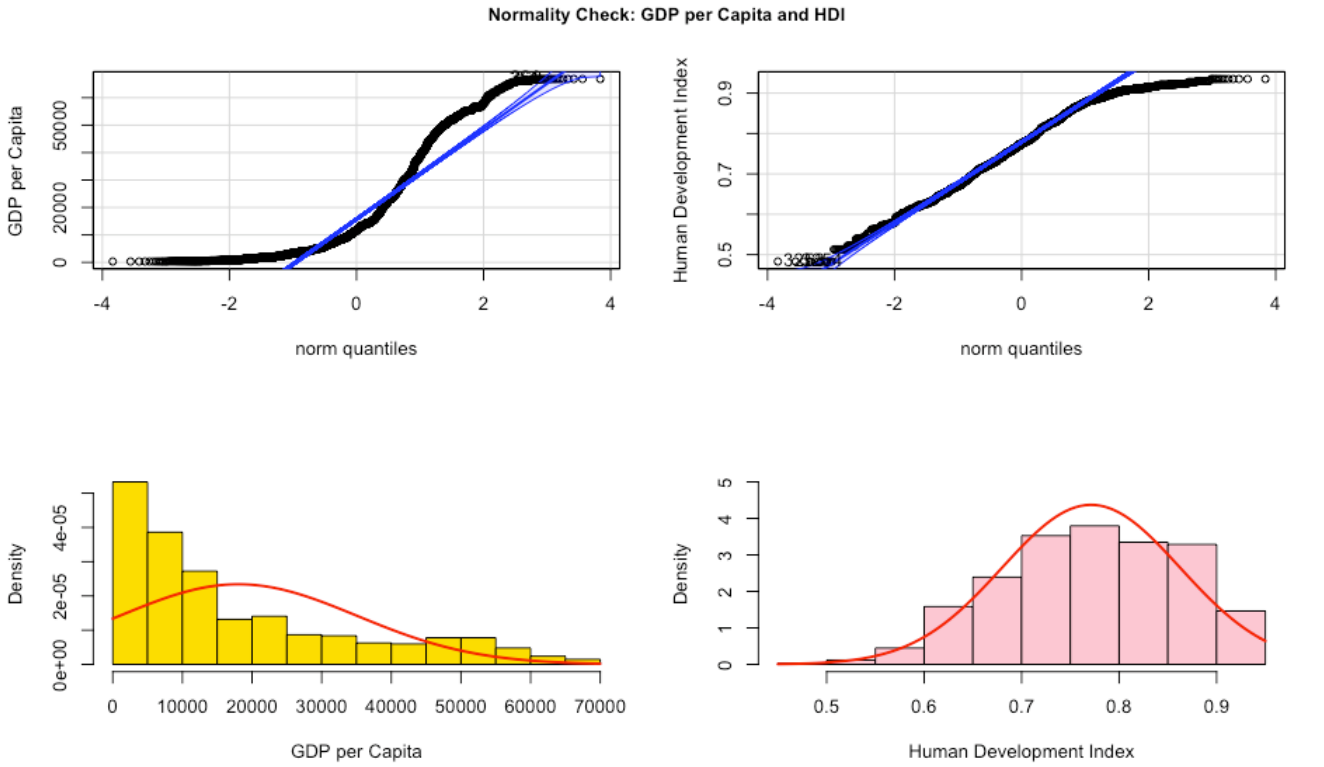
```
> cor.test(gdphdi$HDI.for.year , gdphdi$gdp_per_capita)

        Pearson's product-moment correlation

data:  gdphdi$HDI.for.year and gdphdi$gdp_per_capita
t = 134.01, df = 8014, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8246523 0.8381664
sample estimates:
      cor
0.8315324
```

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.

3. **Regression Analysis**

   1) Normality Check



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.

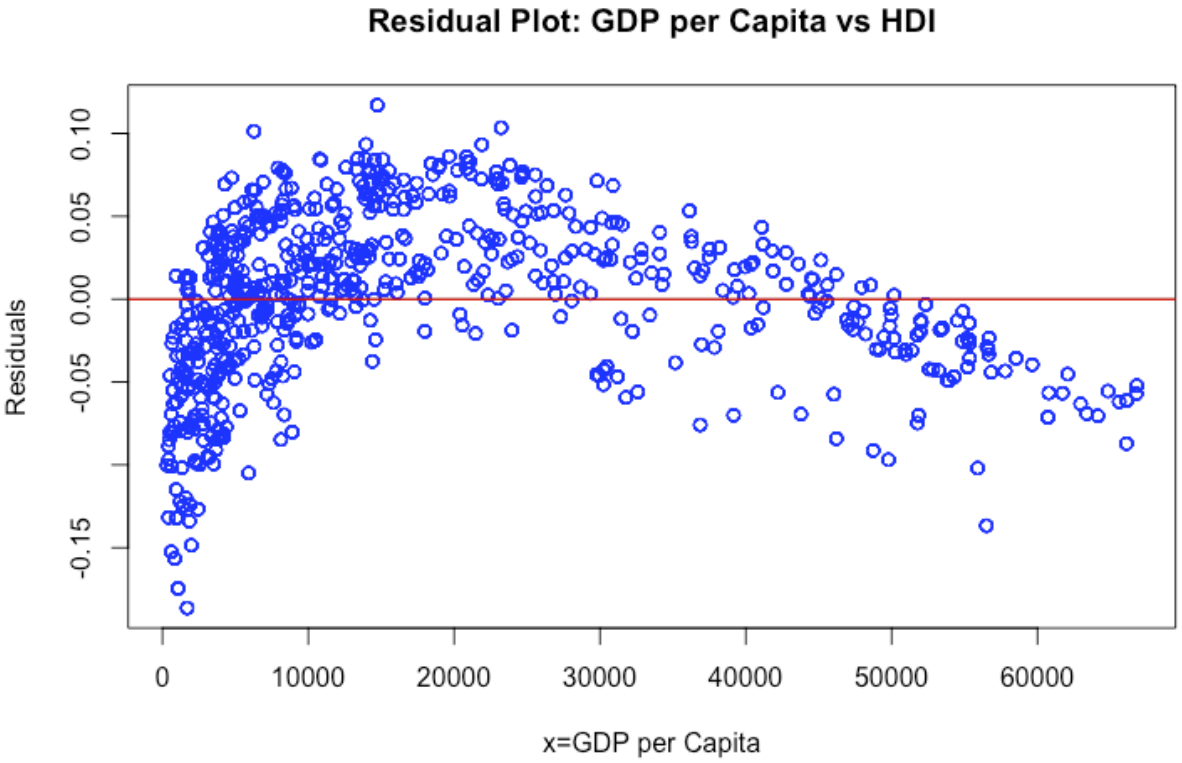
2) Regression Analysis

```
> GHlm <- lm(gdphdi$HDI.for.year ~ gdphdi$gdp_per_capita)
> summary(GHlm)

Call:
lm(formula = gdphdi$HDI.for.year ~ gdphdi$gdp_per_capita)

Residuals:
      Min        1Q    Median        3Q       Max
-0.186338 -0.031252  0.005145  0.036001  0.117052

Coefficients:
                          Estimate  Std. Error t value            Pr(>|t|)
(Intercept)           0.691838506 0.000819215   844.5 <0.0000000000000002 ***
gdphdi$gdp_per_capita 0.000004422 0.000000033   134.0 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05034 on 8014 degrees of freedom
Multiple R-squared:  0.6914,    Adjusted R-squared:  0.6914
F-statistic: 1.796e+04 on 1 and 8014 DF,  p-value: < 0.00000000000000022
```

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.

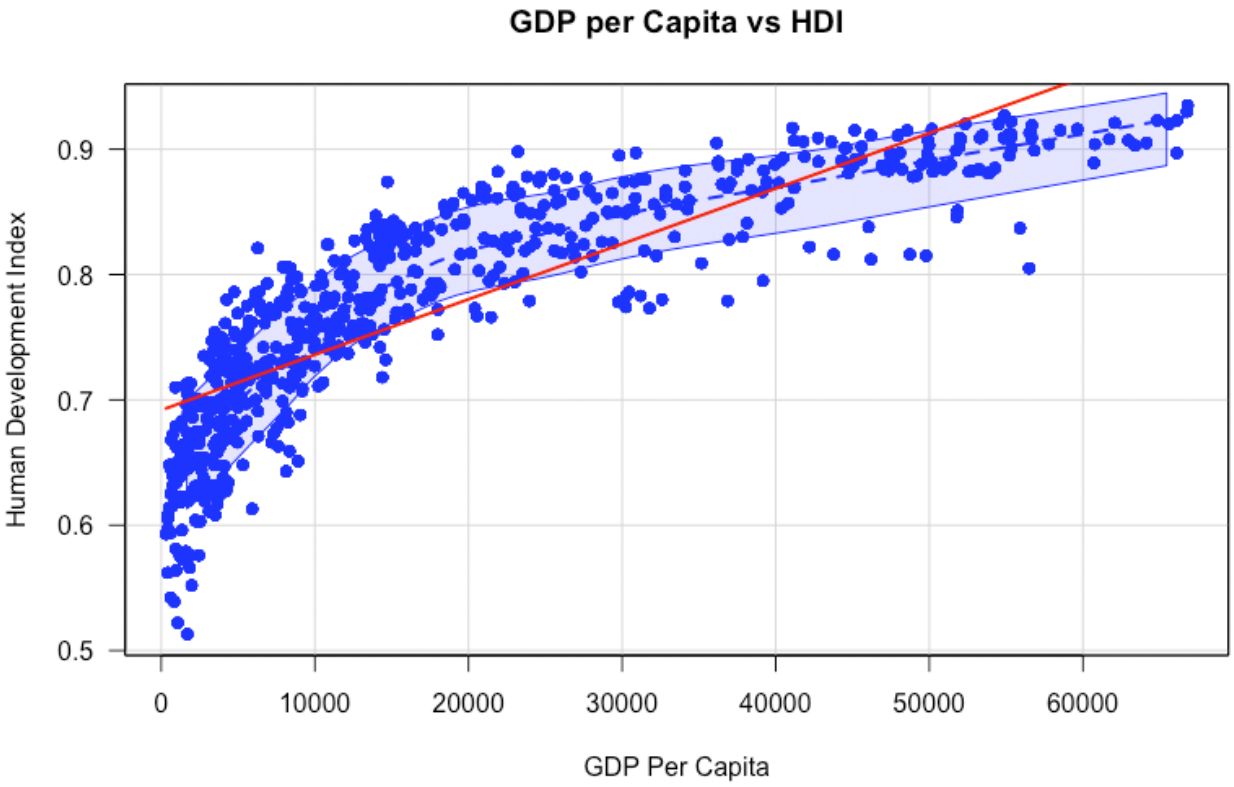
3) Homoscedasticity Check



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.

4. Visualization: scatter plot of GDP per Capita and HDI



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

```
#------------------------------------------------

# Removing outliers

df_gen <- suicide_rate2 %>% group_by(country, year) %>% filter(generation == "Generation X") %>%
  summarise(
    avg_suicide_no = mean(suicides_no),
    avg_suicide_rate = mean(suicides.100k.pop),
    hdi = HDI.for.year,
    gdp_per_capita
  )

min_out <- min(boxplot.stats(df_gen$avg_suicide_rate)$out)

df_gen <- df_gen[df_gen$avg_suicide_rate<min_out, ]

boxplot( df_gen$avg_suicide_rate)

cor(df_gen$gdp_per_capita,df_gen$avg_suicide_rate)
```

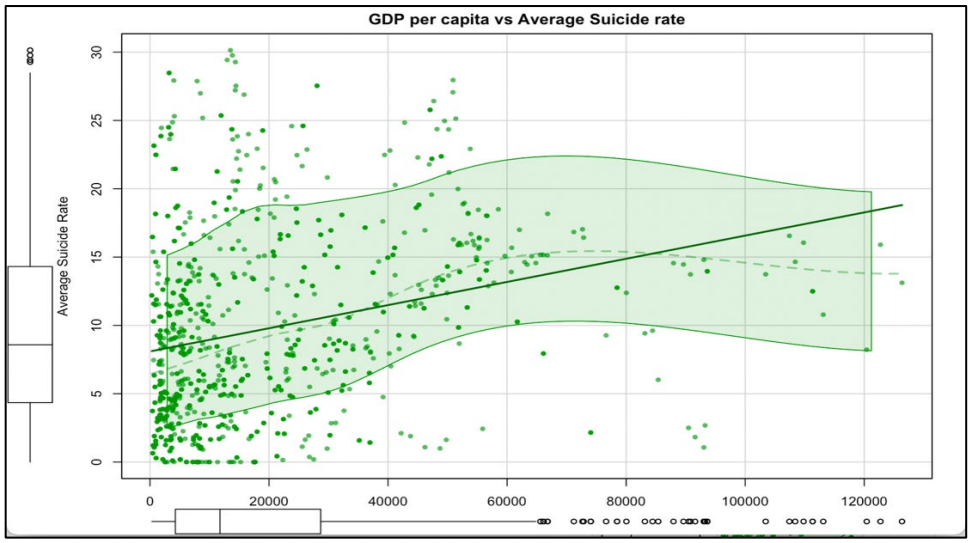
- Above codes represented are used in removal of outliers for representing the regression
- Filtered the information about Generation X and found the minimum outlier value using min() function and stored the information with no outliers in an object.
- Removed the outlier using the filter technique,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

Step 2: Generating the regression plot and the correlation data

```
# Plotting GDP per Capita vs Average Suicide rate for Generation X

scatterplot(df_gen$gdp_per_capita,df_gen$avg_suicide_rate,
            regLine=list(method=lm, lty=1, lwd=2, col="darkgreen"),
            legend = c(title="cyl", coords="topright"),
            main = "GDP per capita vs Average Suicide rate",
            ylab= "Average Suicide Rate",
            xlab = "GDP Per Capita",
            col=rgb(0,70,0,50,maxColorValue=120), pch=16, frame = TRUE, smooth = TRUE)

lmgen <- lm(df_gen$avg_suicide_rate~df_gen$gdp_per_capita)

summary(lmgen)
```

```
> summary(lmgen)

Call:
lm(formula = df_gen$avg_suicide_rate ~ df_gen$gdp_per_capita)

Residuals:
    Min      1Q  Median      3Q     Max
-14.917  -4.790  -1.300   3.818  20.898

Coefficients:
                          Estimate  Std. Error t value            Pr(>|t|)
(Intercept)           8.095320538 0.204122623   39.66 <0.0000000000000002 ***
df_gen$gdp_per_capita 0.000084797 0.000007002   12.11 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.574 on 1920 degrees of freedom
Multiple R-squared:  0.07097,   Adjusted R-squared:  0.07049
F-statistic: 146.7 on 1 and 1920 DF,  p-value: < 0.0000000000000022
```

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



- 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 R^2 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

```
# Removing outliers

df_mil <- suicide_rate2 %>% group_by(country, year) %>% filter(generation == "Millenials") %>%
  summarise(
    avg_suicide_no = mean(suicides_no),
    avg_suicide_rate = mean(suicides.100k.pop),
    hdi = HDI.for.year,
    gdp_per_capita
  )

min_out_mil <- min(boxplot.stats(df_mil$avg_suicide_rate)$out)

df_mil <- df_mil[df_mil$avg_suicide_rate<min_out_mil, ]

boxplot( df_mil$avg_suicide_rate)

cor(df_mil$gdp_per_capita,df_mil$avg_suicide_rate)
```

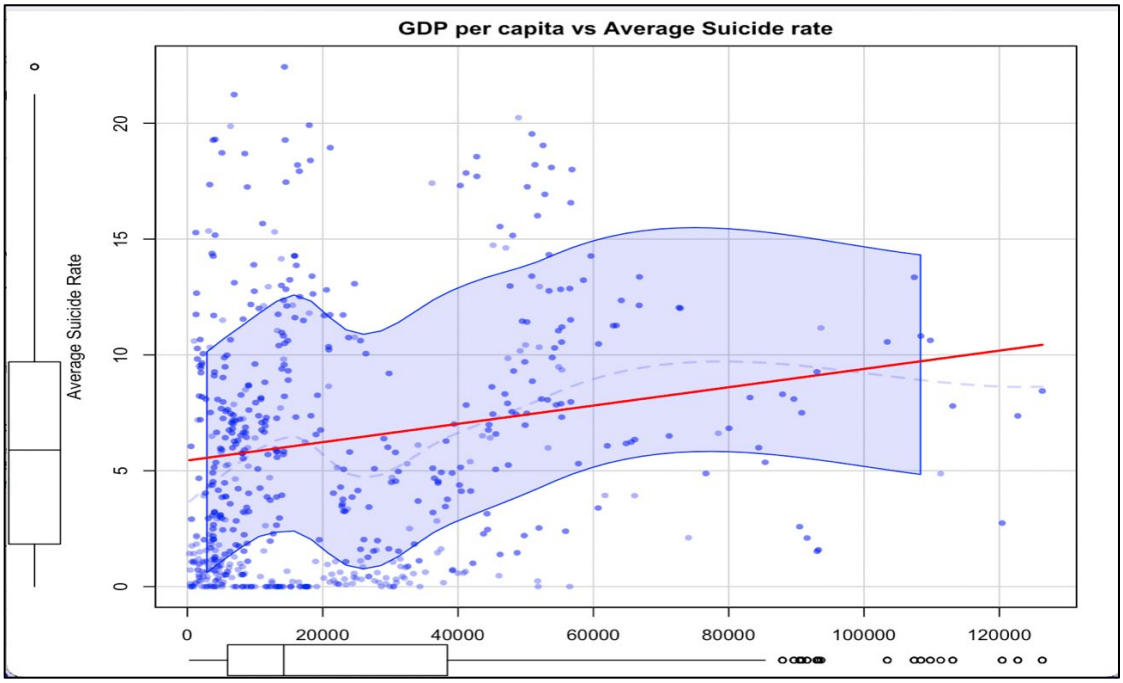
- Above codes represented are used in removal of outliers for representing the regression
- Filtered the information about Millennial and found the minimum outlier value using min() function and stored the information with no outliers in an object.
- Removed the outlier using the filter technique, and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

Step 2: Generating the regression plot and the correlation data

```
#Plotting GDP per Capita vs Average Suicide rate for Millenials

scatterplot(df_mil$gdp_per_capita,df_mil$avg_suicide_rate,
            regLine=list(method=lm, lty=1, lwd=2, col="red"),
            legend = c(title="cyl", coords="topright"),
            main = "GDP per capita vs Average Suicide rate",
            ylab= "Average Suicide Rate",
            xlab = "GDP Per Capita",
            col=rgb(0,0,90,20,maxColorValue=100), pch=16, frame = TRUE)

lmmil <- lm(df_mil$avg_suicide_rate~df_mil$gdp_per_capita)


summary(lmmil)
```

```
> summary(lmmil)

Call:
lm(formula = df_mil$avg_suicide_rate ~ df_mil$gdp_per_capita)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6772 -4.3940 -0.6746  3.3301 16.4238

Coefficients:
                          Estimate  Std. Error t value           Pr(>|t|)
(Intercept)          5.444009200 0.162621039  33.477 <0.0000000000000002 ***
df_mil$gdp_per_capita 0.000039526 0.000004705   8.401 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.064 on 1928 degrees of freedom
Multiple R-squared:  0.03531,   Adjusted R-squared:  0.03481
F-statistic: 70.58 on 1 and 1928 DF,  p-value: < 0.00000000000000022
```

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



- 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 R^2 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**

   1) Generation X Female

```
> cor(fgenx_suigdp$suicides.100k.pop , fgenx_suigdp$gdp_per_capita)
[1] 0.3213692
```

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.

2) Generation X Male

```
> cor(mgenx_suigdp$suicides.100k.pop, mgenx_suigdp$gdp_per_capita)
[1] 0.166685
```

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.

2. Hypothesis test to check the correlation in population

1) Generation X Female

```
> cor.test(mgenx_suigdp$suicides.100k.pop, mgenx_suigdp$gdp_per_capita)

        Pearson's product-moment correlation

data:  mgenx_suigdp$suicides.100k.pop and mgenx_suigdp$gdp_per_capita
t = 3.6571, df = 468, p-value = 0.000284
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.07740348 0.25331434
sample estimates:
     cor
0.166685
```

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.

2) Generation X Male

```
> cor.test(mgenx_suigdp$suicides.100k.pop, mgenx_suigdp$gdp_per_capita)

        Pearson's product-moment correlation

data:  mgenx_suigdp$suicides.100k.pop and mgenx_suigdp$gdp_per_capita
t = 3.6571, df = 468, p-value = 0.000284
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.07740348 0.25331434
sample estimates:
     cor
0.166685
```

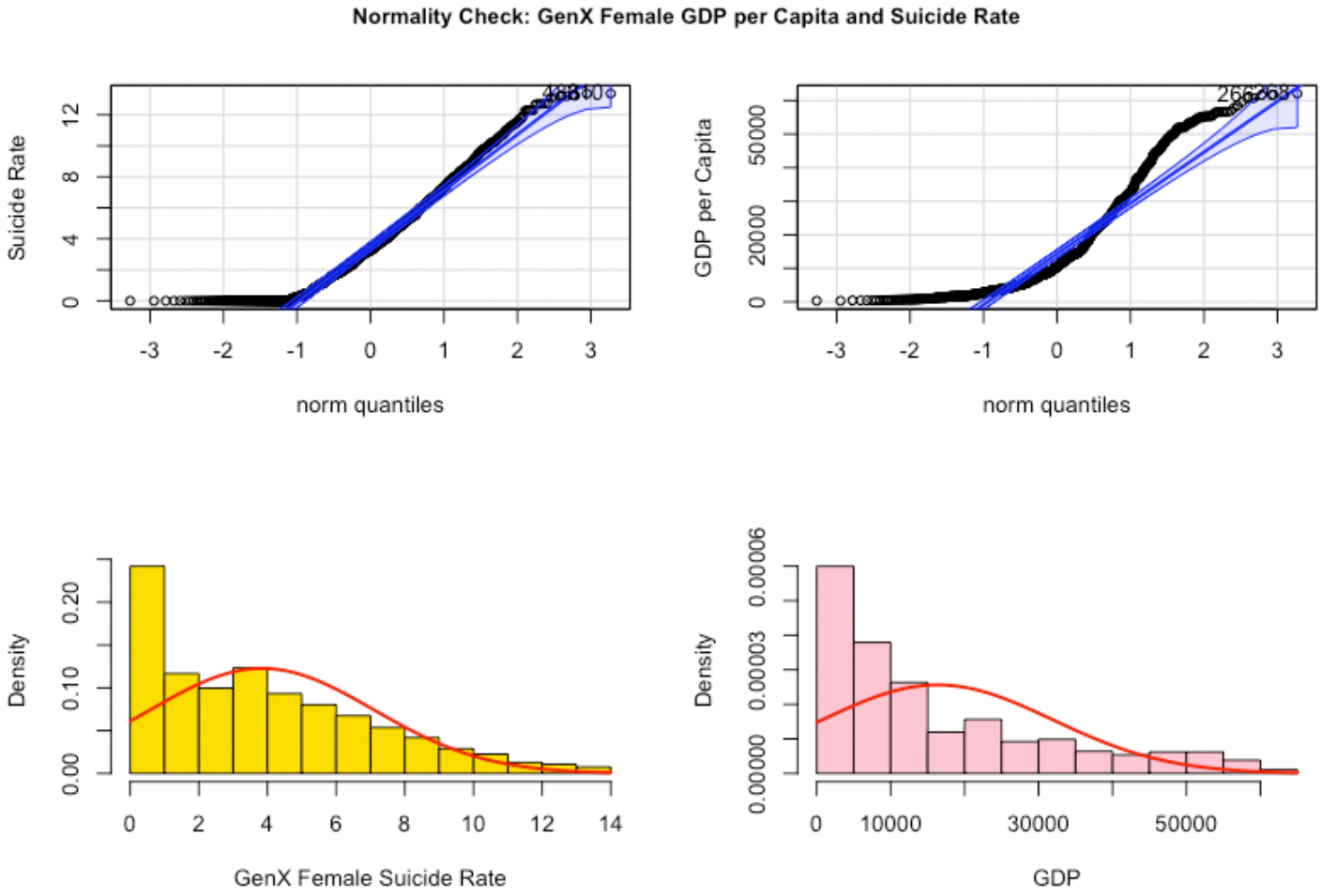
Analysis:

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.

3. Regression Analysis

1) Normality Check

a. Generation X Female GDP per Capita & Suicide Rate

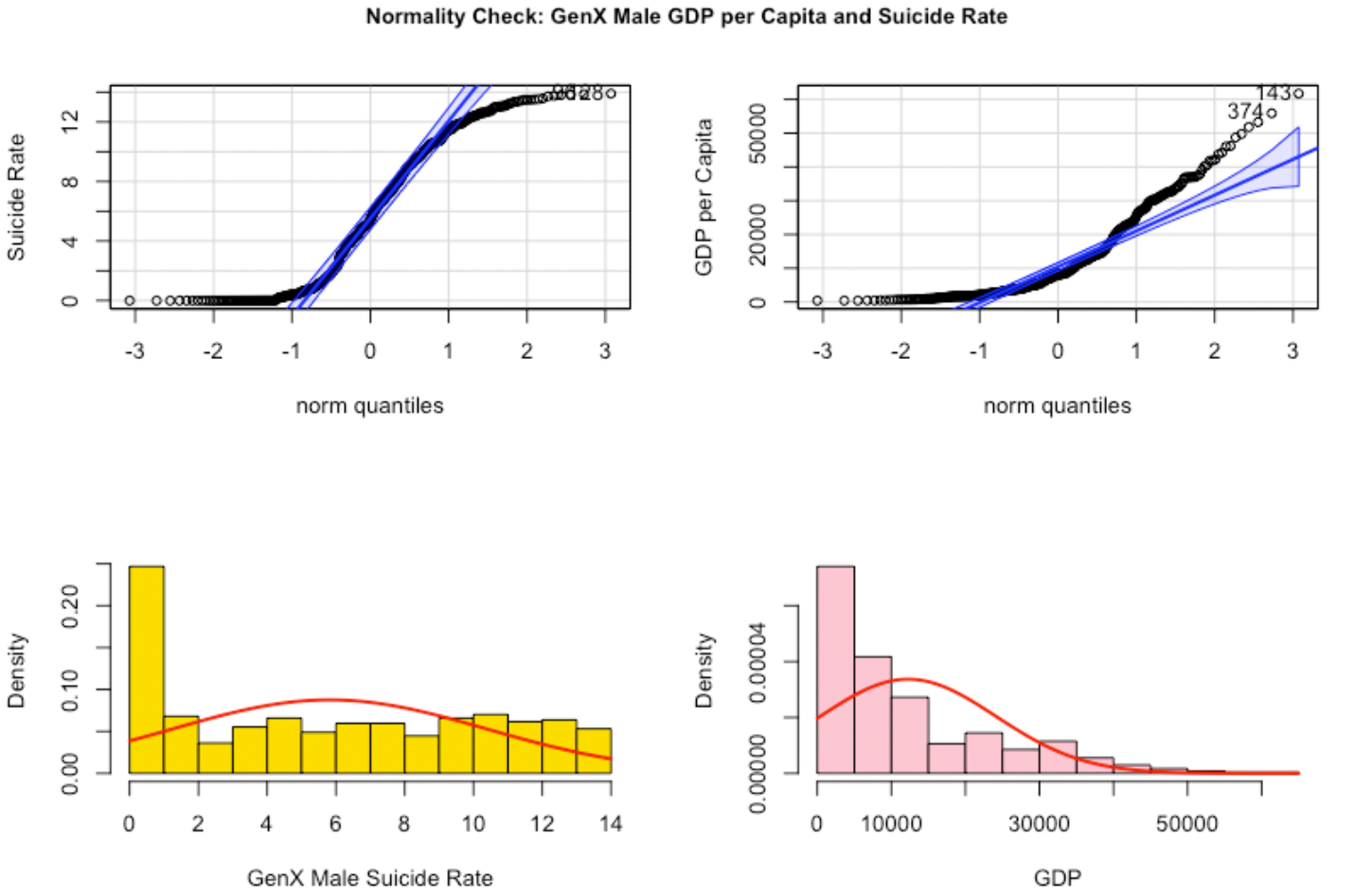


Analysis:

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.

b. Generation X Female GDP per Capita & Suicide Rate



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.

2) Regression Analysis

a. Generation X Female

```
> fgenx.lm <- lm(fgenx_suigdp$suicides.100k.pop ~ fgenx_suigdp$gdp_per_capita)
> summary(fgenx.lm)

Call:
lm(formula = fgenx_suigdp$suicides.100k.pop ~ fgenx_suigdp$gdp_per_capita)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8210 -2.5358 -0.5066  1.7819  9.8620

Coefficients:
                             Estimate Std. Error t value            Pr(>|t|)
(Intercept)                2.74490927 0.14624016   18.77 <0.0000000000000002 ***
fgenx_suigdp$gdp_per_capita 0.00006714 0.00000648   10.36 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.081 on 932 degrees of freedom
Multiple R-squared:  0.1033,    Adjusted R-squared:  0.1023
F-statistic: 107.3 on 1 and 932 DF,  p-value: < 0.00000000000000022
```

Analysis:

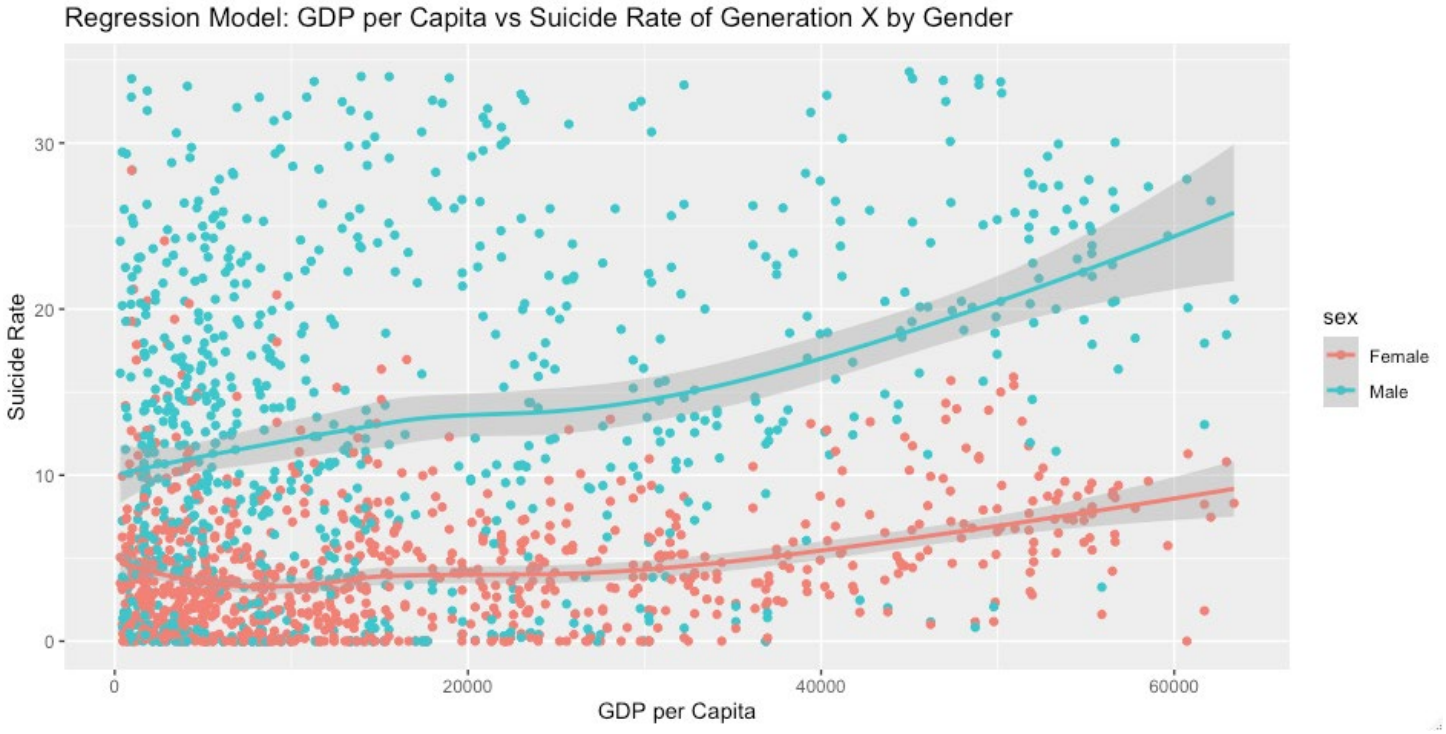
- The slope of the formula is 0.00006, and intercept is 2.745. Therefore the formula should be Y = 0.00006X + 2.745, where X is GDP per Capita and Y is GenX female suicide rate
- The F-statistics p-value is close to 0, which means the correlation between the 2 variables is significant.
- R-squared =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

```
> mgenx.lm <- lm(mgenx_suigdp$suicides.100k.pop ~ mgenx_suigdp$gdp_per_capita)
> summary(mgenx.lm)

Call:
lm(formula = mgenx_suigdp$suicides.100k.pop ~ mgenx_suigdp$gdp_per_capita)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3929 -4.6757 -0.0938  3.9966  8.7252

Coefficients:
                             Estimate Std. Error t value            Pr(>|t|)
(Intercept)                5.02998115 0.29859669  16.845 < 0.0000000000000002 ***
mgenx_suigdp$gdp_per_capita 0.00006409 0.00001752   3.657             0.000284 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.494 on 468 degrees of freedom
Multiple R-squared:  0.02778,   Adjusted R-squared:  0.02571
F-statistic: 13.37 on 1 and 468 DF,  p-value: 0.000284
```

Analysis:

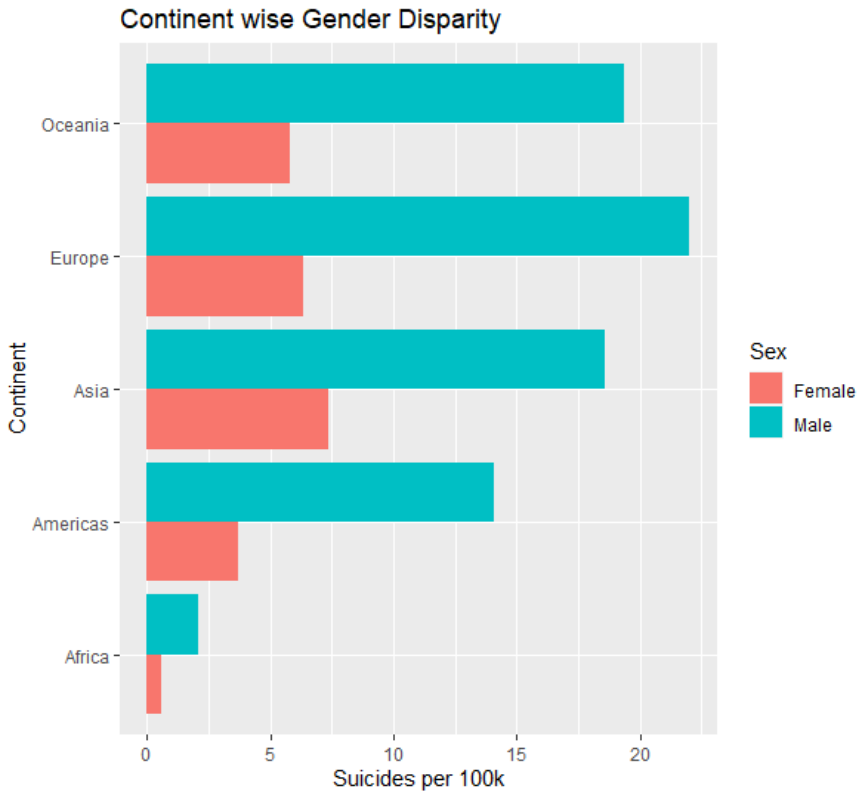
- The slope of the formula is 0.00006, and intercept is 5.030. Therefore the formula should be Y = 0.00006X + 5.030, where X is GDP per Capita and Y is GenX male suicide rate
- The F-statistics p-value is close to 0, which means the correlation between the 2 variables is significant.
- R-squared = 0.028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

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



- 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?



Observation:

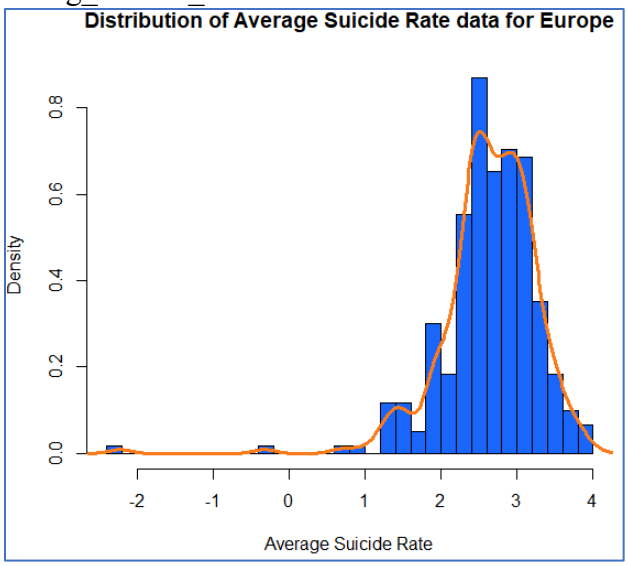
- Europe is having the highest suicide number recorded till now among all the continents
- Africa is having the least suicide number among all.
- Female from Asia have committed the highest suicide among females from other continents.
- Male suicide rate is higher than the female suicide rate across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

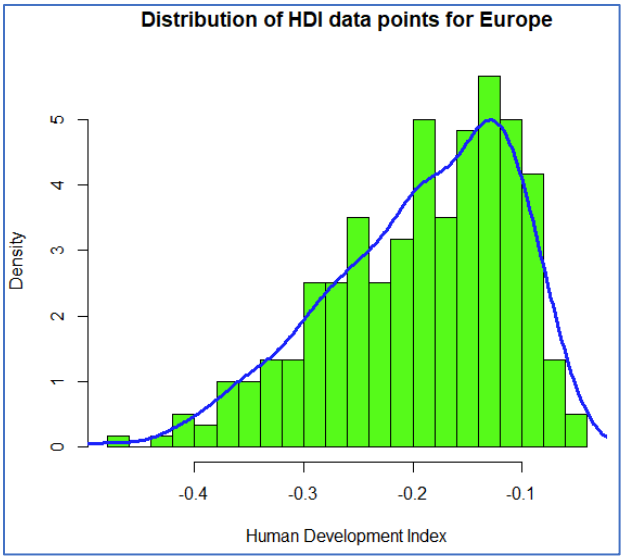
As we saw from the above graph that Europe is having the highest suicide rate among all the populations, we are curious to see how Europe is trying to handle this situation.
For this purpose, we are going to see the correlation between average suicide rate in Europe till now with its Human Development Index.

```
# Filtering and preparing data for Europe continent
df <- suicide_rate2 %>% group_by(country, year) %>% filter(continent == "Europe") %>%
  summarise(
    avg_suicide_rate = mean(suicides.100k.pop),
    hdi = HDI.for.year,
    gdp_per_capita
  )
df_unq <- unique(df)
```

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'.





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

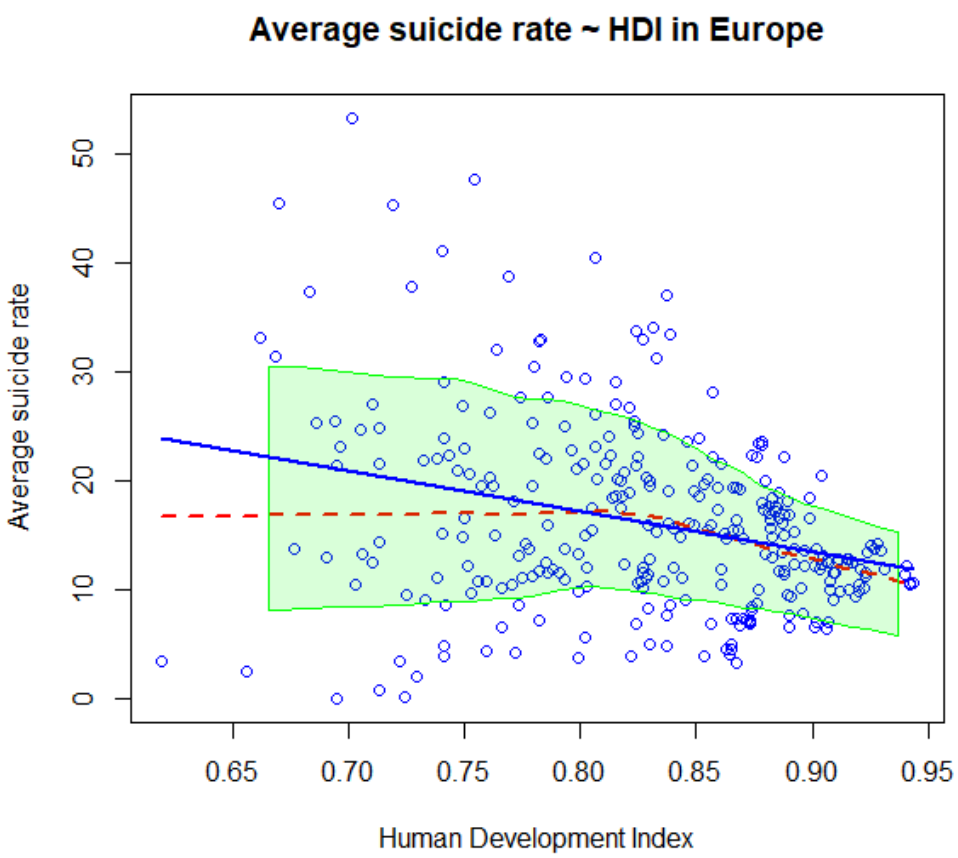
Checking correlation between the average suicide rate with the human development index:

```
> cor(df_unq$hdi, df_unq$avg_suicide_rate)
[1] -0.289673
```

Pearson's Product-Moment Correlation test:

```
> cor.test(df_unq$hdi, df_unq$avg_suicide_rate)

        Pearson's product-moment correlation

data:  df_unq$hdi and df_unq$avg_suicide_rate
t = -5.2245, df = 298, p-value = 0.0000003286
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3901169 -0.1824160
sample estimates:
      cor
-0.289673
```

Plotting the correlation with a linear regression in a graph:



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.

```
# Checking for regression
lin_reg <- lm(df_unq$avg_suicide_rate ~ df_unq$hdi, data = df_unq)
summary(lin_reg)
```

```
Call:
lm(formula = df_unq$avg_suicide_rate ~ df_unq$hdi, data = df_unq)

Residuals:
    Min      1Q  Median      3Q     Max
-21.110  -5.759  -0.186   4.185  32.387

Coefficients:
            Estimate Std. Error t value           Pr(>|t|)
(Intercept)   46.768      5.869   7.969 0.0000000000000343 ***
df_unq$hdi   -36.918      7.066  -5.225 0.0000003286456131 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.255 on 298 degrees of freedom
Multiple R-squared:  0.08391,   Adjusted R-squared:  0.08084
F-statistic:  27.3 on 1 and 298 DF,  p-value: 0.0000003286
```

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 R^2 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.

Group_Project_Final_Script.R

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

suicide_rate.csv