# Final Project

## Suicide Rates Overview 1985 to 2016 Dataset

The dataset was imported from a popular online dataset sharing community named Kaggle. Kaggle allows users to find and publish data sets, explore and construct data science models in a web-based data science environment, work with other data scientists and machine learning engineers, and compete to solve data science challenges.

Suicide Rates Overview 1985 to 2018 dataset provides us information about country, age range, income for the respective years to understand the trend in suicide rate and hidden factors.

***Deliverable 1***

**Briefly describe the dataset: Size (Required Storage), Metadata (Data items, meanings and types), Structure.**

Size: The dataset has 12 columns and 27821 rows (2.58 MB file).

Metadata: The dataset has country (Nominal), year (Nominal), sex (Nominal), age (Ordinal), suicides_no (Interval), population (Ordinal), suicides/100k pop (Ratio), HDI for year (Ratio), gdp_for_year ($) (Ratio), gdp_per_capita ($) (Ratio), generation(Nominal).

**Who collected the data? Who they are, what they do?**

The dataset is collected by a Kaggle user from various sources naming as the following:

- United Nations Development Program. (2018). Human development index (HDI)
- World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016.
- (2017). Suicide in the Twenty-First Century [dataset].
- World Health Organization. (2018). Suicide prevention.

Kaggle is an online community owned by Google that offers sharing of datasets and build models in different environments.

**What is their role or purpose?**

Main purpose in blending all the datasets from disparate sources is to understand the trend in suicide rate for each country from the previous years and conclusions to prevent the same.

**Why did they collect the data?**

To find the signs that correlate with higher suicide rates across the global socio-economic spectrum among different cohorts.

**Describe any privacy, quality, ethical or other issues with this dataset?**

The data is directly collected from the Data bank, Official WHO, UNDP and Kaggle websites where most of the data is published publicly. There are no privacy issues with this dataset as it is open source.

**What potential value can be obtained by studying this data? List some specific questions and plan to answer them in your analysis?**

We can analyze many things from this dataset. Few among that are:

- What are the various aspects involved in suicide rate?
- How socio-economic conditions leads to suicide?
- Who commit suicides most? Is it Men or Women?
- Are the teenagers who commit the most suicides?
- How to prevent suicides?

**What software and hardware resources will you need to study this data?**

Hardware resources:
All the analysis for this project was performed in a laptop with the following configurations:

- System Type: 64-bit operating system, x64-based processor
- Installed RAM: 4 GB
- Processor Name: AMD A6-5200 APU

Software requirements are R, Python, Tableau, SQL to study this data.

**Identify and briefly discuss one or more other similar studies that were done in the domain of your project.**

There was a study done similar to this in that the people analyzed the data of the WHO Suicide Analysis.

***Deliverable 2***

## Data Exploration

**The following analysis are performed using "Python"**

A brief summary of different attributes in the dataset and summary of the dataset.

<u>country</u>

```
count          27820
unique           101
top        Mauritius
freq             382
Name: country, dtype: object
```

year

```
count    27820.000000
mean      2001.258375
std          8.469055
min       1985.000000
25%       1995.000000
50%       2002.000000
75%       2008.000000
max       2016.000000
Name: year, dtype: float64
```

sex

```
count      27820
unique         2
top       female
freq       13910
Name: sex, dtype: object
```

age

```
count           27820
unique              6
top       15-24 years
freq             4642
Name: age, dtype: object
```

suicide_no

```
count    27820.000000
mean       242.574407
std        902.047917
min          0.000000
25%          3.000000
50%         25.000000
75%        131.000000
max      22338.000000
Name: suicides_no, dtype: float64
```

population

```
count    2.782000e+04
mean     1.844794e+06
std      3.911779e+06
min      2.780000e+02
25%      9.749850e+04
50%      4.301500e+05
75%      1.486143e+06
max      4.380521e+07
Name: population, dtype: float64
```

suicides/100k pop

```
count    27820.000000
mean        12.816097
std         18.961511
min          0.000000
25%          0.920000
50%          5.990000
75%         16.620000
max        224.970000
Name: suicides/100k pop, dtype: float64
```

country-year

```
count           27820
unique           2321
top        Turkey2012
freq               12
Name: country-year, dtype: object
```

HDI for year

```
count    8364.000000
mean        0.776601
std         0.093367
min         0.483000
25%         0.713000
50%         0.779000
75%         0.855000
max         0.944000
Name: HDI for year, dtype: float64
```

gdp_for_year ($)

```
count                27820
unique                2321
top        397,558,094,270
freq                    12
Name:  gdp_for_year ($) , dtype: object
```

gdp_per_capita ($)

```
count     27820.000000
mean      16866.464414
std       18887.576472
min         251.000000
25%        3447.000000
50%        9372.000000
75%       24874.000000
max      126352.000000
Name: gdp_per_capita ($), dtype: float64
```

generation

```
count               27820
unique                  6
top          Generation X
freq                 6408
Name: generation, dtype: object
```

Descriptive statistics

| | year | suicides_no | population | suicides/100k pop | HDI for year | gdp_per_capita ($) |
|---|---|---|---|---|---|---|
| **count** | 27820.000000 | 27820.000000 | 2.782000e+04 | 27820.000000 | 8364.000000 | 27820.000000 |
| **mean** | 2001.258375 | 242.574407 | 1.844794e+06 | 12.816097 | 0.776601 | 16866.464414 |
| **std** | 8.469055 | 902.047917 | 3.911779e+06 | 18.961511 | 0.093367 | 18887.576472 |
| **min** | 1985.000000 | 0.000000 | 2.780000e+02 | 0.000000 | 0.483000 | 251.000000 |
| **25%** | 1995.000000 | 3.000000 | 9.749850e+04 | 0.920000 | 0.713000 | 3447.000000 |
| **50%** | 2002.000000 | 25.000000 | 4.301500e+05 | 5.990000 | 0.779000 | 9372.000000 |
| **75%** | 2008.000000 | 131.000000 | 1.486143e+06 | 16.620000 | 0.855000 | 24874.000000 |
| **max** | 2016.000000 | 22338.000000 | 4.380521e+07 | 224.970000 | 0.944000 | 126352.000000 |

## Data Visualization

**The following visualizations are performed using "Tableau"**

**Graph set 1:**

Used **Tableau** to create these **Maps** which represent the Population density distributed across different Countries for the years 1985 and 2016.

Population size chart for each Country in the year 1985


Map based on Longitude (generated) and Latitude (generated). Size shows sum of Population. The marks are labeled by Country. Details are shown for Country. The data is filtered on Year, which ranges from 0 to 1985.

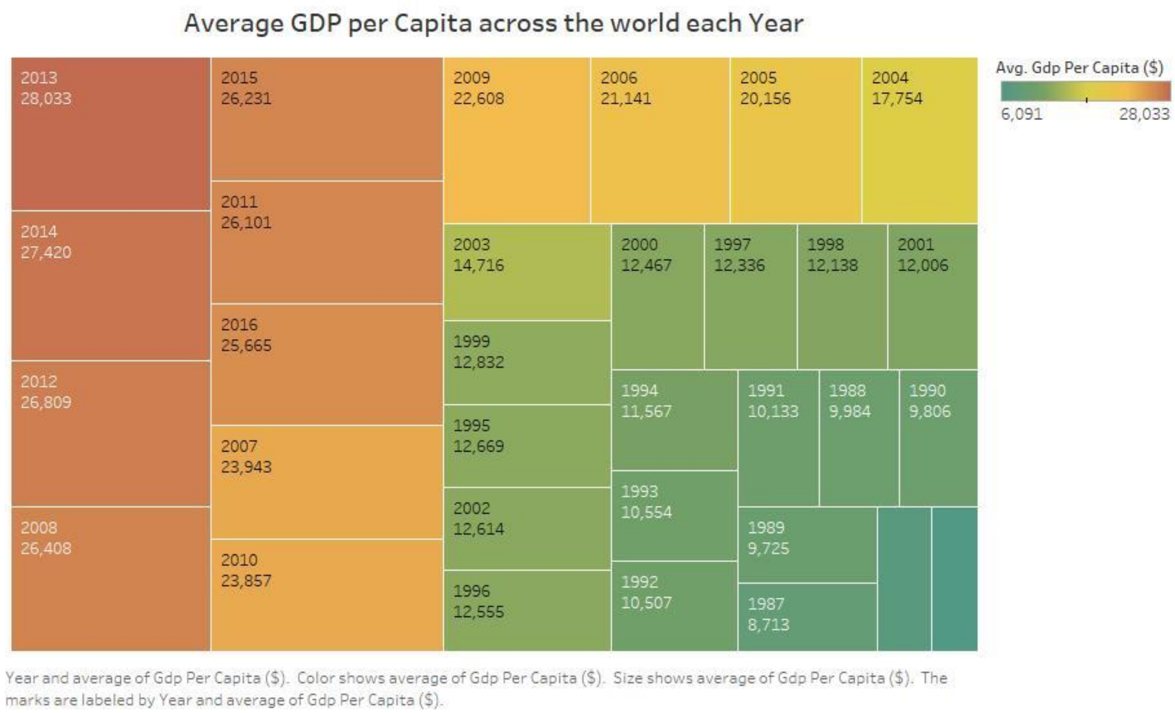
Population size chart for each Country in the year 2016


Map based on Longitude (generated) and Latitude (generated). Size shows sum of Population. The marks are labeled by Country. Details are shown for Country. The data is filtered on Year, which ranges from 0 to 2016.

Above graphs show us that Population increase in the years 1985 and 2016. More analysis had been done by including additional Countries in the year 2016 than that of 1985.

**Graph 2:**

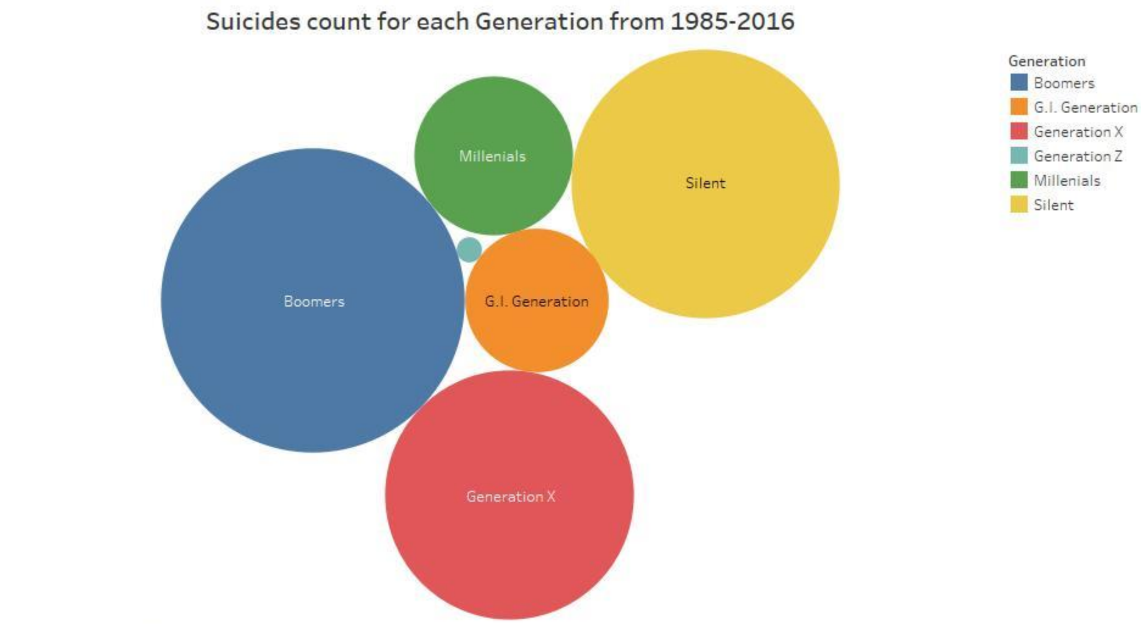
Used **Tableau** to create this **Tree Map** that shows Average GDP per Capita across the world for each Year in the dataset.




The tree map indicates that the Year 2013 has the maximum average GDP per Capita compared to the other years in the dataset.

**Graph 3:**

Used **Tableau** to create this **Bubble chart** that depicts the Suicide rate among various generations from the years 1985 - 2016.

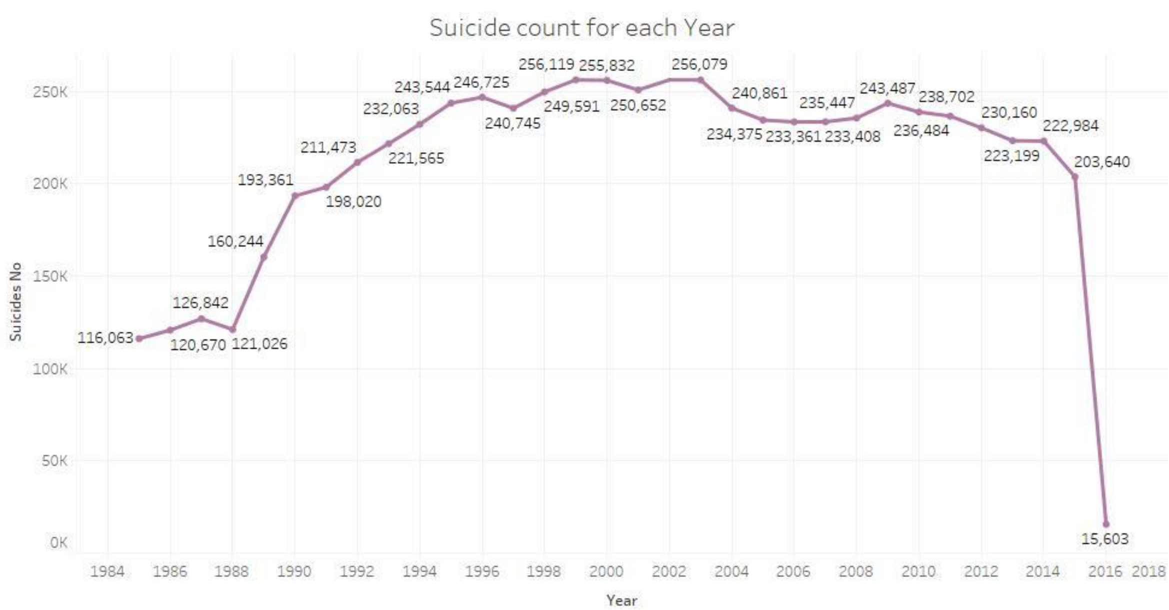



Generation. Color shows details about Generation. Size shows sum of Suicides No. The marks are labeled by Generation.

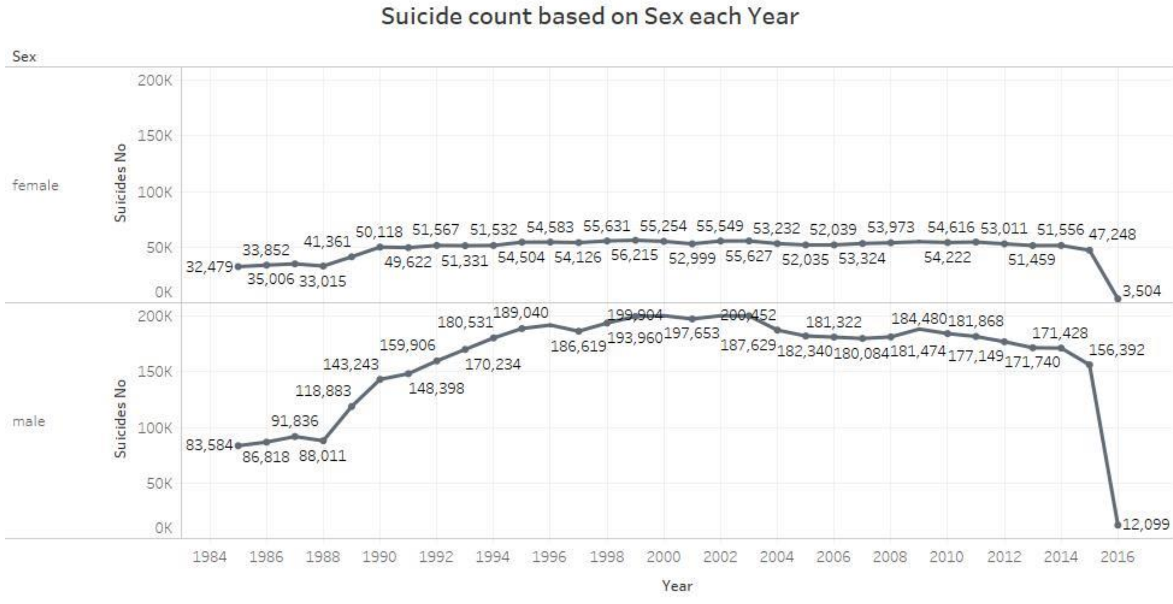
The above bubble chart reveals that Boomers tend to commit suicide more often than other generations. Whereas, Generation Z less likely committed suicides.

**Graph set 4:**

Used **Tableau** to create these **Line chart** that depicts the Suicide rate among various generations from the years 1985 - 2016.




The trend of sum of Suicides No for Year. The marks are labeled by sum of Suicides No.



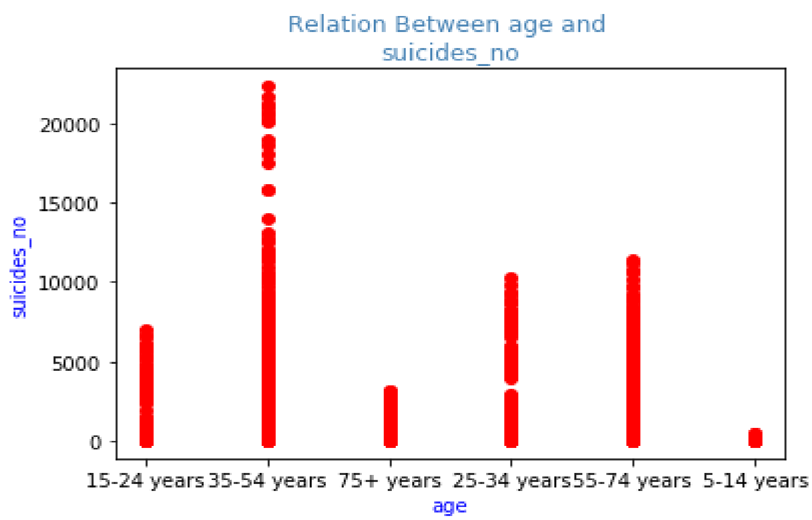

The trend of sum of Suicides No for Year broken down by Sex. The marks are labeled by sum of Suicides No.

The above graph set shows that the highest number of suicides happened in the year 1999 with a count of 256,119. The female suicide count ratio was constant for most of the time but, the male suicide count has a lot of fluctuations in it.

**The following visualizations are performed using "Python"**

**Scatter plot:**

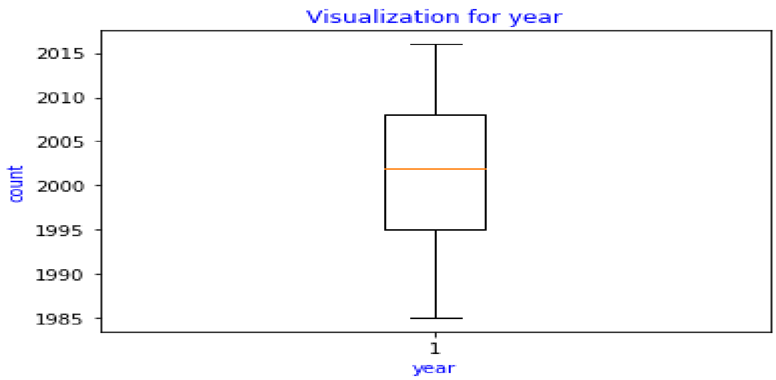
Used **Python** to create this **Scatter plot** which shows relation between Age and Suicide_no.



From the above graph, we can observe that people aged between 35 -54 years committed suicides more frequently than any other age group.

**Box plot:**

Used **Python** to create this **Box plot** which represents visualization for the column Year.




The above box plot conveys us that Suicide analysis has been conducted for the years ranged from 1985 to 2016 having them represented as minimum and maximum in the box plot respectively. And the year 2002 is just situated in the center being the median in the plot.

**The following analysis are performed using "R"**

**Correlation Analysis:**

Correlation coefficient "r" signifies the strength of a linear relationship. If "r" value is positive, it shows the attributes have strong linear relationship, if in case it is negative, that shows weak linear relationship. In addition to that p-value should always be less than 0.05, that implies the relation is linear and significant.

Strong linear relationship:

```
> cor.test(mydata$population,mydata$suicides_no)

        Pearson's product-moment correlation

data:  mydata$population and mydata$suicides_no
t = 130.48, df = 27818, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6088195 0.6233995
sample estimates:
      cor
0.6161623
```

Here, the correlation coefficient “r” value is 0.6161623 which is positive and close to 1 that clears out that it shows a strong linear relationship between Population and Suicide_no. The p value is extremely small which represents highly significant result.

Weak linear relationship:

```
> cor.test(mydata$suicides.100k.pop,mydata$year)

        Pearson's product-moment correlation

data:  mydata$suicides.100k.pop and mydata$year
t = -6.5158, df = 27818, p-value = 7.353e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.05076446 -0.02729837
sample estimates:
       cor
-0.0390368
```

Here, the correlation coefficient “r” value is negative, that shows a weak linear relationship between Suicides per 100k population and Year. Having an extremely small p value shows it’s a significant result.

**Regression Analysis:**

Regression analysis is usually run to understand how independent variables help to predict a dependent variable.

The adjusted R-squared is for the multi- variable model which represents how variables are in place and not randomly distributed. Even the p value conveys the same. More specifically it shows the non-randomness (if it’s value is less than ) and significance of the variable. In addition to the above constraints(p value and the adjusted R-squared), there is one more important feature known as Significant F that conveys significance of the overall model (including dependent and independent variables).

```
> #Regression analysis
> linearModel <- lm(formula = mydata$suicides_no~ mydata$age+mydata$year+
+                     mydata$population,data = mydata)
> print(linearModel)

Call:
lm(formula = mydata$suicides_no ~ mydata$age + mydata$year + 
    mydata$population, data = mydata)

Coefficients:
          (Intercept)  mydata$age25-34 years  mydata$age35-54 years  
            2.146e+03              7.419e+01              1.792e+02  
 mydata$age5-14 years  mydata$age55-74 years    mydata$age75+ years  
           -1.573e+02              1.782e+02              1.489e+02  
          mydata$year      mydata$population  
           -1.117e+00              1.416e-04  
```

```
> summary(linearModel)

Call:
lm(formula = mydata$suicides_no ~ mydata$age + mydata$year + 
    mydata$population, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-3285.8  -104.4   -35.0    84.8 19543.4 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            2.146e+03  9.923e+02   2.162   0.0306 *  
mydata$age25-34 years  7.419e+01  1.454e+01   5.103 3.36e-07 ***
mydata$age35-54 years  1.792e+02  1.460e+01  12.274  < 2e-16 ***
mydata$age5-14 years  -1.573e+02  1.456e+01 -10.801  < 2e-16 ***
mydata$age55-74 years  1.782e+02  1.454e+01  12.259  < 2e-16 ***
mydata$age75+ years    1.489e+02  1.460e+01  10.198  < 2e-16 ***
mydata$year           -1.117e+00  4.958e-01  -2.253   0.0243 *  
mydata$population      1.416e-04  1.093e-06 129.608  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 700.3 on 27812 degrees of freedom
Multiple R-squared:  0.3974,	Adjusted R-squared:  0.3973 
F-statistic:  2620 on 7 and 27812 DF,  p-value: < 2.2e-16
```

The above model has Age, Year, Population (independent variables) columns to predict the Suicide_no (Suicide number) (dependent variable). The p value for all the variables is less than 0.05, that means all variables are fit to predict the Suicide number. But the Adjusted R-squared values is not encouraging.

**Hypothesis test**

Chi-square test

```
> chisq.test(mydata$age,mydata$suicides_no)

	Pearson's Chi-squared test

data:  mydata$age and mydata$suicides_no
X-squared = 15584, df = 10415, p-value < 2.2e-16
```

We have a chi-squared value of 15584. Since we get a p-value less than the significant level of 0.05, we then reject the null hypothesis and conclude that the two variables, low confidence and high confidence, are dependent.

Wilcox test

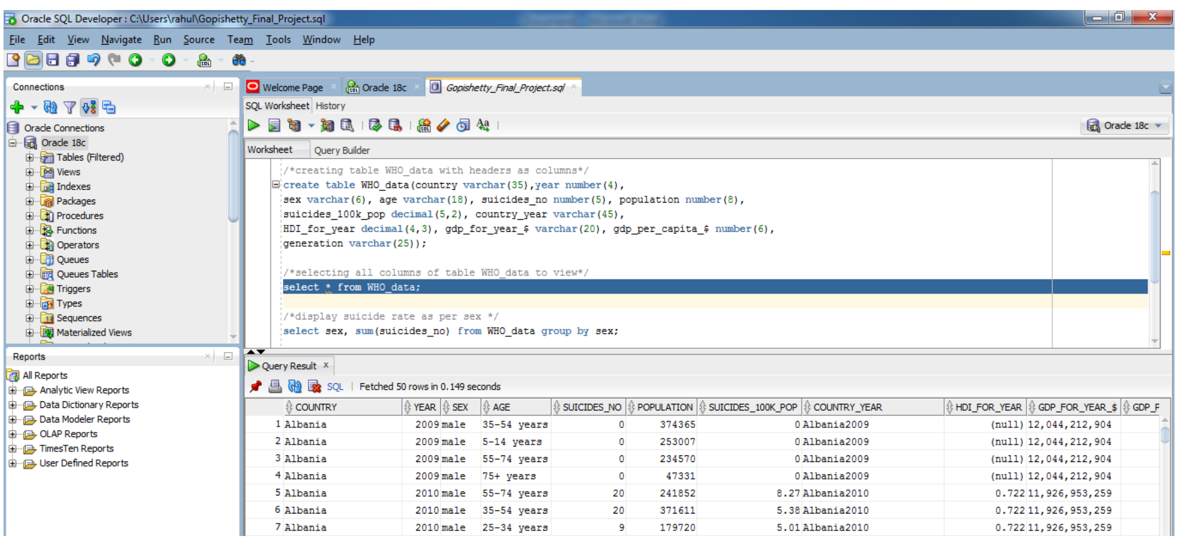
```
> wilcox.test(mydata$population,mydata$gdp_per_capita...., paired = TRUE)

        Wilcoxon signed rank test with continuity correction

data:  mydata$population and mydata$gdp_per_capita....
V = 381080000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

The null hypothesis is that the gdp_per_capita ($) and population are identical populations. At 0.05 significance level, we can conclude that gdp_per_capita ($) and population from the dataset data are non-identical populations.

**The following analysis are performed using "SQL"**

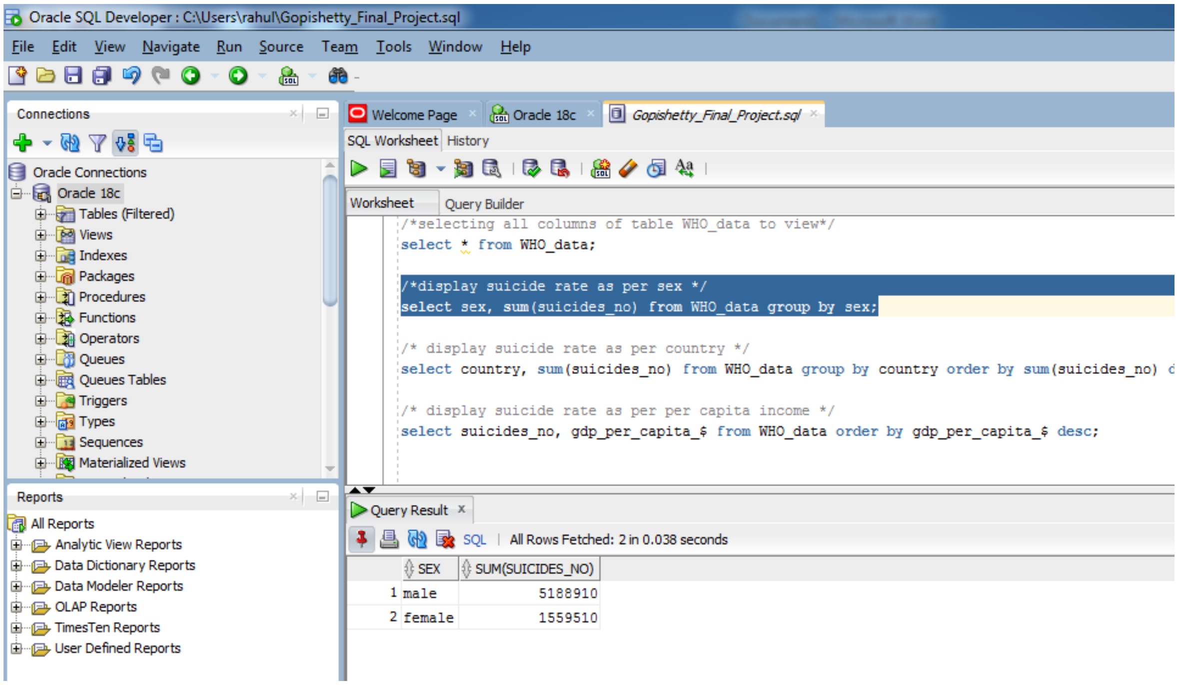
The first step is to create a table in order to execute the SQL queries. Here, the table is created with the name 'WHO_data' having all the columns as headers. Thereafter CSV file is imported into the table.

```
/*creating table WHO_data with headers as columns*/
create table WHO_data(country varchar(35),year number(4),
sex varchar(6), age varchar(18), suicides_no number(5), population number(8),
suicides_100k_pop decimal(5,2), country_year varchar(45),
HDI_for_year decimal(4,3), gdp_for_year_$ varchar(20), gdp_per_capita_$ number(6),
generation varchar(25));
```



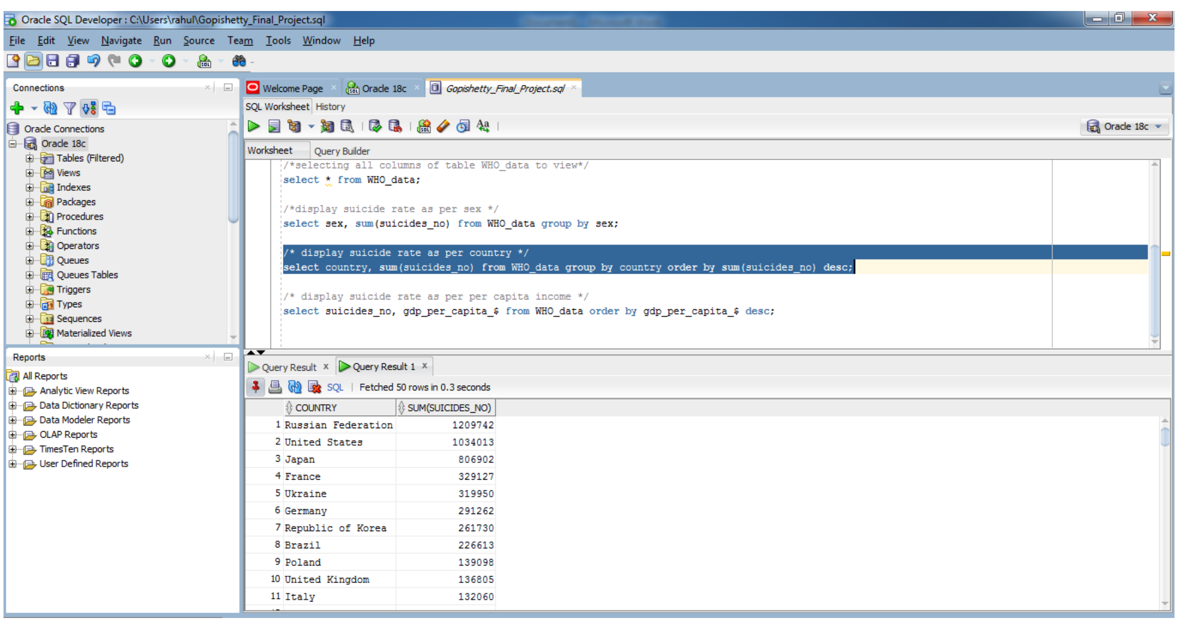

Now, the SQL queries are written to retrieve data from the table.

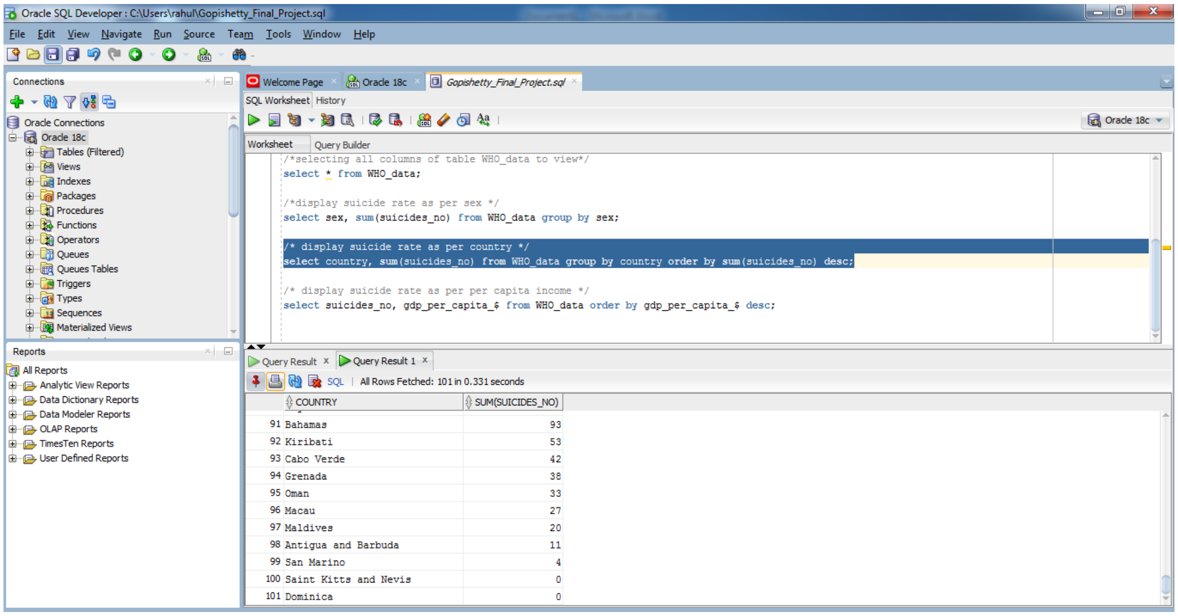
**Query1 :** It is based on suicide rate as per sex.




From the above result we can understand that males committed suicides in more number than females in the years 1985-2016.

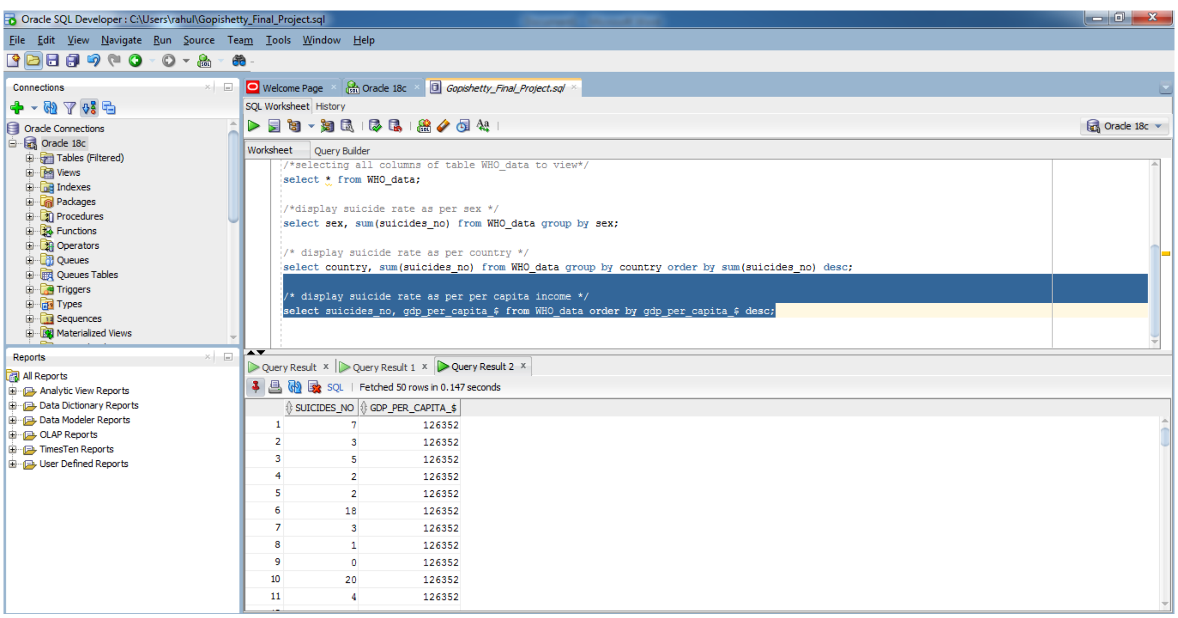
**Query2:** It is based on suicide rate as per country.

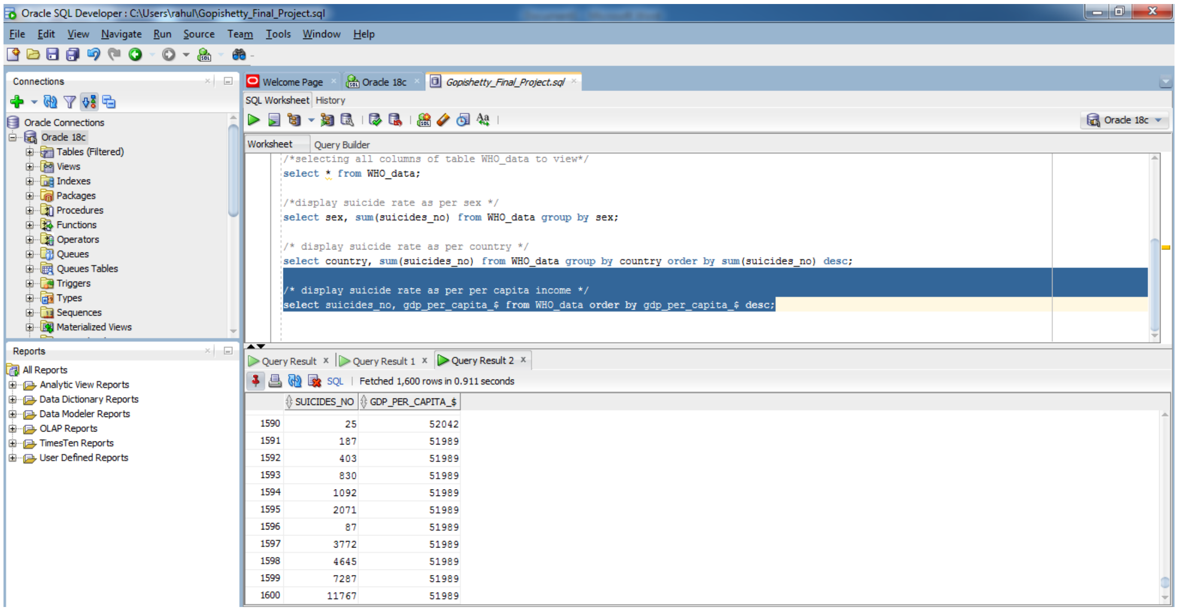





The above query result shows that the most number of suicides happened in 'Russian Federation' with  a count of 1209742, while the least in 'Dominca' and 'Saint Kitts and Nevis'.

**Query3:** It is based on suicide rate as per GDP per Capita income.





The above results show that people with high GDP rate committed less no. of suicides than people with low GDP rate. That means people from poor countries have committed suicides the most.

**Results Interpretation**

*This should reflect answers to the specific questions specified above*

Python was used to calculate the descriptive statistics and visualizations such as scatterplots and boxplots.

Tableau was used to create visualizations such as line graph, bubble graph, tree map, maps for the study.

R was used to calculate tests like Hypothesis test, correlation analysis and regression analysis are performed.

From the study we can conclude that from 1985 to 2016, there are correlations among various socio-economic aspects on suicide rates.

By studying this data, some potential questions can be answered such as:

- How socio-economic scenario of a country effects suicide rate?
- What is the trend in suicide rates across the world from past 3 decades?
- Which countries need more help and support to raise HDI?

While this data provides a quick access to few causes that effect the suicide rate. It throws light on areas where suicide rate can be prevented like income, gender, age range. Less GDP capita rate implies poor country people have higher chances of committing suicides than rich country people, Males definitely constitute more in suicide count than females. The misconception that teenagers commit more suicides has been disrupted by this study.

*Describe the value obtained from the study*

Insights obtained from this dataset analysis helps to work on the areas where suicides are to be prevented. By raising awareness on depression and various kinds of other causes for suicides; by uplifting poorer nations with financial and educational support ; by supporting and taking care most effected generation (Boomers) individuals and especially males.

The visualizations and tests are done in order to understand the data in the study.

**Explain/define terms:**

GDP: Gross Domestic Product.

GDP per Capita: per capita shows a country's GDP divided by its total population

WHO: World Health Organization.

HDI: Human Development Index.

Null Hypothesis: The null hypothesis is the initial statistical claim that the population mean is equivalent to the claimed.

Boxplot: It is a method for graphically depicting groups of numerical data through their quartiles displaying the five-number summary of a set of data.

Scatterplot: It is a two-dimensional data visualization and uses dots to represent the values obtained for two different variables along the x-axis and y-axis.

**References:**

Suicide Rates Overview 1985 to 2016, Retrieved from kaggle: [Online]. Available: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016 [Accessed: 20-Nov-2019].

Python. (n.d.). Retrieved from Python: https://www.python.org/ [Accessed: 5-Nov-2019].

RStudio. (n.d.). Retrieved from RStudio: https://www.rstudio.com/ [Accessed: 5-Nov-2019].

tableau. (n.d.). Retrieved from tableau: https://www.tableau.com/ [Accessed: 2-Nov-2019].