**A Final Report in Data Analysis of**

**Suicide trends in different country between 1985- 2015**

By Srijana Lawa

Contents

[**Introduction** 3](#_Toc7937444)

[**Programming Language:** 3](#_Toc7937445)

[**Data Set Information** 3](#_Toc7937446)

[**Methodology** 4](#_Toc7937447)

[**Result** 7](#_Toc7937448)

[**Conclusion** 38](#_Toc7937449)

# **Introduction**

According to WHO, close to 800,000 people die due to suicide every year, which is one person every 40 seconds. Effective and evidence- based interventions can be implemented at population, sub population and individual levels to prevent suicide and suicide attempts. There are indications that for each adult who died by suicide there may have been more than 20 others attempting suicide. Even the popular American Glamor model “**Stephanie Adams**”, popular singer of my country Nepal “**Anil Adhikar**i” famous as Yamma Buddha committed suicide in the peak of his career and many more other celebrities, businessman also have committed suicide. So, Reason of suicide may be different for each person. Hence, I am trying to find the trend of suicide for different country between 1985 to 2015 and find the reason behind these so that we can prevent suicide.

# **Programming Language: Python**

# **IDE: Atom**

# **Data Set Information**

I have used 5 data set and their information as below:

Master.csv

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

This data set has 12 columns as 27820 rows. Columns are country, year, sex, age, suicides\_no, population, suicides per 100k population, country-year, HDI for year, gdp\_for\_year ($), gdp\_per\_capital ($), generation.

Femalesuiciderate\_per\_100k.csv

Link: <http://hdr.undp.org/en/data>

This data set has information about the female suiciderate\_per\_100k for 4 year namely 2000, 2005,2010,2015

Malesuiciderate\_per\_100k.csv

Link: <http://hdr.undp.org/en/data>

This data set has information about the male suiciderate\_per\_100k for 4 year namely 2000, 2005,2010,2015

Gross\_national\_income\_per\_100k.csv

Link: <http://hdr.undp.org/en/data>

It is the data which is listed on the basis of the 2017 Human development index. It has information about the Gross National Income per Capita (2011 PPP $) Purchasing power parity per dollar for 4 year namely 2000, 2005,2010,2015.

Human\_Development\_Index.csv

Link: <http://hdr.undp.org/en/data>

This dataset consists of HDI value from 1990 to 2017 but I have just use data of 2000, 2005, 2010 and 2015.

# **Methodology**

Well I have two different sets of observation data so i will be explaining them in different category.

Methodology for master data set

Initially, I read master.csv file as original and masterdata and I have made changes in masterdata only. To get general idea about data I have printed the shape, top 5 data and datatype of dataset columns. Each row in this data set gives the information based on the particular country, year, agegroup and sex. And then i checked for the null value in data set using isnull method and found that HDI\_for\_year column has null value. In 27,820 rows, 19456 rows where empty for HDI\_for\_year column in master data set so, i dropped this column. I grouped the data on the basis of country\_year and used for loop to print each value with their count. Total there are 101 countries and more than 50 countries did not have data for 2016 year so i dropped value for 2016 year. I am considering data from 1985 to 2015 i.e. 31 years. I defined getcountryToDelete to get the array of country which has data for less than 10 years and dropped those countries. The dropped countries are Bosnia and Herzegovina, Cabo verde, Dominica, Macau, Nicaragua, Oman, Saint Kitts and Nevis, San Marino, Turkey, United Arab Emirates. Now, there are 11 columns and 27264 rows. I have defined getfilteredData method which requires four parameters namely: - dataset, country name, gender and agegroup. This method returns a row of data by filtering information through that requirement. I also have a getgender method which takes an array of sex of master data and returns an array of sex where male is replace by 0 and female is replaced by 1. To see the general statistics like mean, max, min and percentile value I have used describe method of Pandas library. The mean value of suicideper100k is 12.94 and the maximum value of suicide per 100k is 224.97(Aruba 1995) but the 75% of suicideper100k is 16.84 only so with this information I am assuming that there is only limited country with the suicide per 100k greater than 100.

I plotted count plot on the basis of gender, generation, country and year to see the distribution of data. I plotted bar plot to see the trend of suicide per 100k from year 2000 to 2015 on the basis of gender, age group and generation. I summed all the population and suicide number for each country and then plotted graph for top and least 15 countries. I also plotted plot of suicide per 100k of female, suicide per 100k of male, gdp\_per\_year, gdp\_per capita for USA, UK, Republic of Korea, Singapore and Norway from 1999 to 2015. USA has a powerful economy, UK is one of powerful country in Europe, Korea had highest female suicide for 2010 (which I found in Suicide\_Data\_for\_4 different\_year data set data set), Singapore has the third highest per-capita GDP in the world in terms of Purchasing Power Parity according to world bank (2017) and Norway has highest HDI index.

After these data visualization, I printed the correlation between the data column with numerical data. For regression analysis, I created dummy variable for gender using getgender method and for age group and generation I used get\_dummies inbuilt method and then dropped one column and concatenated with masterdata. I read continent file and created series for each continent and then created world dictionary with the value of the continent and then i created continent column in master data and assigned country value to them. I have created a series con which consist of all the country from that column and then removed the whitespace between the continent name.  I have used for loop to assign the continent value to con series and then i set the value on the continent column as the con. After this, i created dummy variable for continent and then concatenated with the masterdata as column. I created variable name for each dummy variable and interaction term of age and gender. I have created a variable array with all the dependent and independent array. I have transposed this array in order to interchange row and column and then created dataframe final data with this array. Basically, masterdata is an unbalanced panel dataset as this dataset consist of observations of multiple phenomena obtained over multiple time periods for same country, gender and agegroup and dataset does not have information about all the countries from 1985 to 2015. So, I have used PooledOLS method for the regression. In order to used PooledOLS I had to convert dataframe into multiindex dataframe. I called set\_index method using country (entity – outer index) and year (time- inner index) to create multiindex dataframe for my finaldata dataframe.

For all the model, Suicide per100k is dependent variable. I have defined exog\_vars dataframe with columns name which I will be using in model. For first model, the independent data are years, gend, age5, age15, age25, age35, age55, boomers, GIgen, genX, genZ, Millenials, population, gdpForYear, africa, europe, namerica, samerica, oceania. I called PooledOLS in built method of linearmodels. panel and then I used predict method for the prediction. I found residual by subtracting predicted suicideper100k from the actual data and then projected data to see the distribution. I also found Mean absolute error, mean squared error and root mean squared error. For all the models, I have repeated same process only the independent variable is different.

For second model, I have replaced gdpForyear and population with gdpPerCapita as GDP per capita is GDP per population. For third model, I have added interaction term of gender and agegroup in first model. Similarly, in fourth model in have added interaction term in second model. For fifth model, the independent data are years, gend, age5, age15, age25, age35, age55, boomers, GIgen, genX, genZ, Millenials, suicidenum, population, gdpForYear, africa, europe, namerica, samerica, oceania. At last I have used compare method to compare the value of all the models I have defined for this dataset.

Methodology for Suicide\_Data\_for\_4 different\_year data set

I had individual dataset for GNI, malesucide\_per100k, femalesucide\_per100k and HDI value. I merged these datasets on the basis of Country and HDI\_Rank (2017). I dropped GNI\_2016 and GNI\_2017 because other dataset did not have value for 2016 and 2017. After merging, I created a new csv file and named it as Suicide\_Data\_for\_4different\_year. I read this csv file as testData and it has 180 row and 18 columns.  I checked for the null value and some value of HDI and GNI for 2000 and 2005 where null so, i filled the empty data with the mean value of the same year. I assigned value of the HDI, GNI, female suicide per 100k and male suicide per100k of different year to an individual variable which data type was series. I have created three methods namely: getMaxValue, getMinValue and getfilteredData. getMaxValue method requires two parameters: - dataset and the parameters like HDI or GNI or Suicide etc. This method returns the row of the value with the maximum value for the filtered data. Similarly, getMinValue method also takes same parameter and return the row of value with the minimum value. getfilteredData method takes dataframe and country parameter and returns the row of data for the filtered country. I used a describe method to see the general statistics like mean, standard deviation, 25%,50%,75% for each column. I have called getMaxValue and getMinValue method for each year value of HDI, GNI, female suicide per 100k and male suicide per 100k.

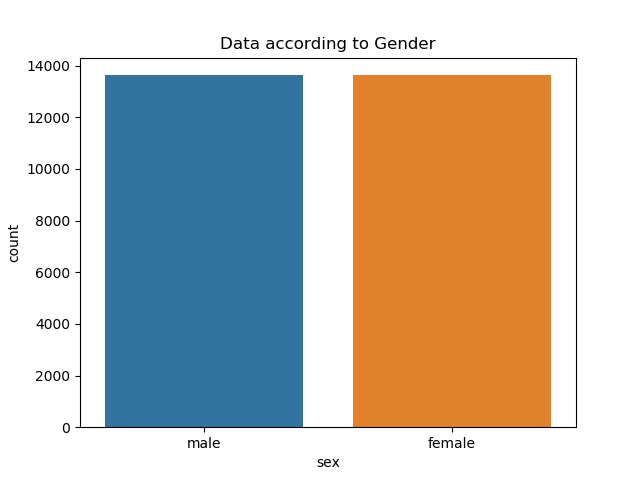
I have plotted the graph for Qatar as it has the highest GNI for 2000,2010 and 2015, Lithuania as it has maximum male suicide per 100k for 2000 and 2005 Then I appended each HDI, GNI, male suicide per 100k and female suicide per 100k of all the year and created one array for each parameter. I plotted the distribution plot for female and male suicide per 100k. Then I read the continent.csv file and created series for each continent. Furthermore, I created dictionary naming world with the value of the continent and continent column in test data. Then, assigned country value to them. I created a series con which consist of all the country from that column and then removed whitespace between the continent name.  I used for loop to assign the continent value to con series and then assigned value on continent column as the con series.

After this i created dummies for continent by calling get\_dummies default method in pandas. I deleted dummy for Asia as n categorical variable only needs n-1 dummy. Then, I concatenated these dummy variables as column in testData dataframe. Further, I created series of each continent and I had to append these values 3 more time as i had merged each value of 2000, 2005, 2010 and 2015 year into a single column. I created an array named ‘data’ which consist of GNI, HDI, africa, europe, namerica, samerica and oceania. I transposed this array so that the number of rows of these arrays matched with the rows of dependent variable female suicide per 100k and male suicide per 100k. I printed the correlation between items of data array. I used sklearn package linear\_model. Linear Regression method which is Ordinary Least Square Linear regression in sklearn. I created four model where independent variable is female or male. female is array of female suicide per 100k and male is array of male suicide per 100k. I used train\_test\_split method to split the data and dependent variable into X\_train and X\_test, y\_train and y\_test. Then, I defined a linear model and fitted the X\_train and Y\_train array. I printed the intercept and coefficient and in order to know standard error, t-value and probabilities I have written the formulas and rounded the values to 3 except for coefficient parameter. I found the y - prediction using predict method. I have printed R2, Mean Squared Error, Mean Absolute error and Root Mean Square Error for the evaluation of model. And finally, i plotted the distribution plot of residuals. For other three models, i used OLS method from statsmodels.api as it provided more detail information.

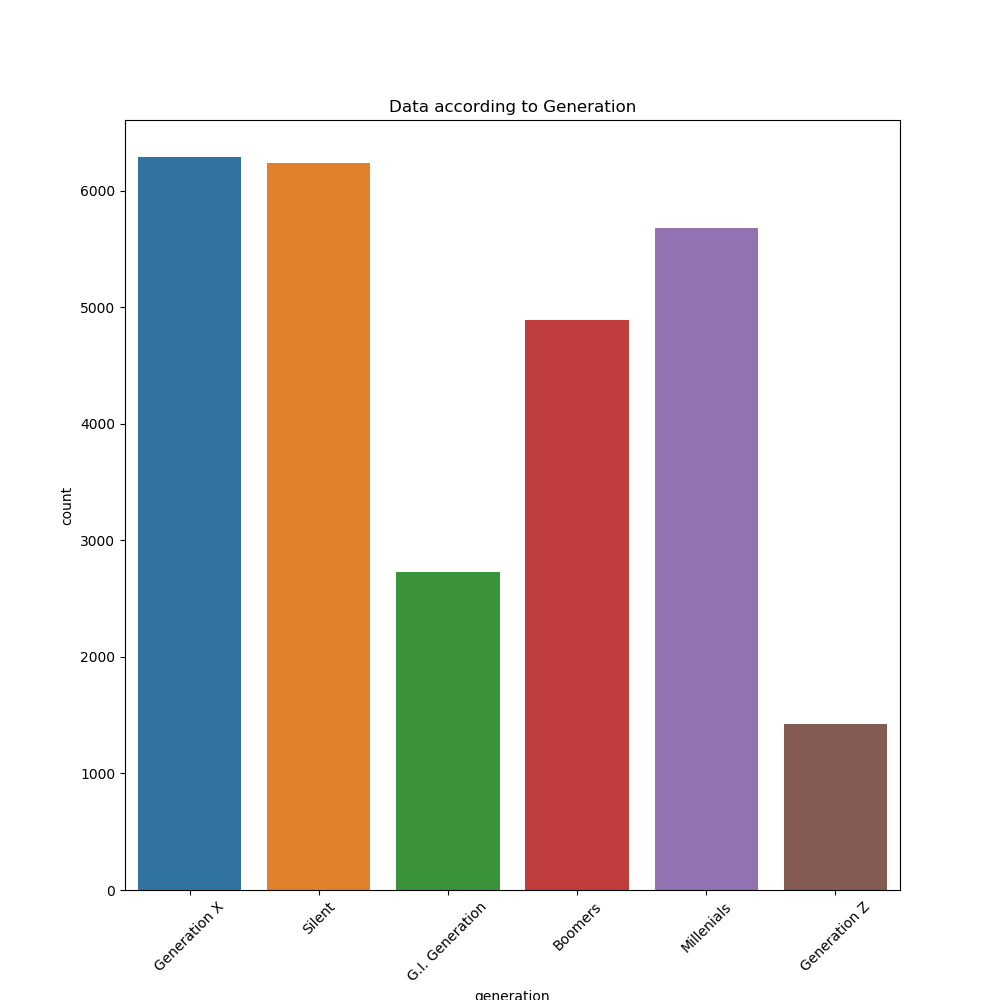
# **Result**

Master Data set

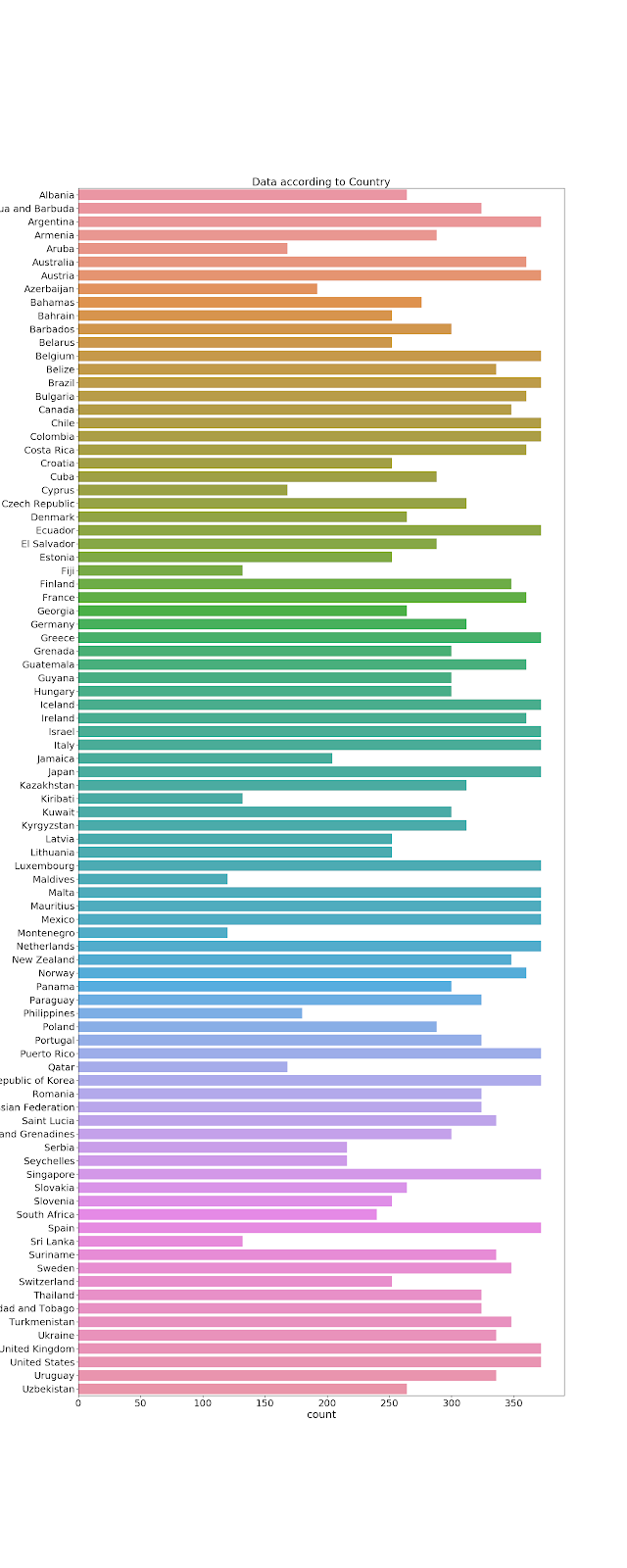
There was 27264 rows and 11 columns and 90 unique countries data from year 1985 to 2015 after cleaning data. The counter plot according gender is as follow:

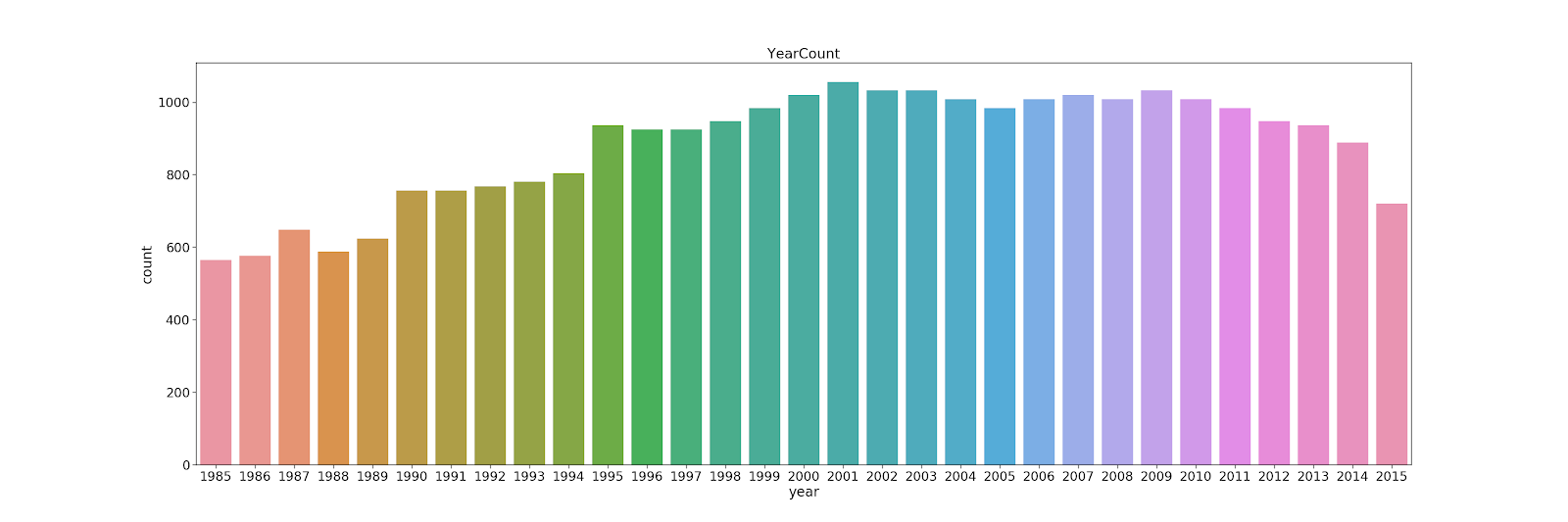


This count plot shows that the number of male and female data in the data set is same. Let’s see for the generation:

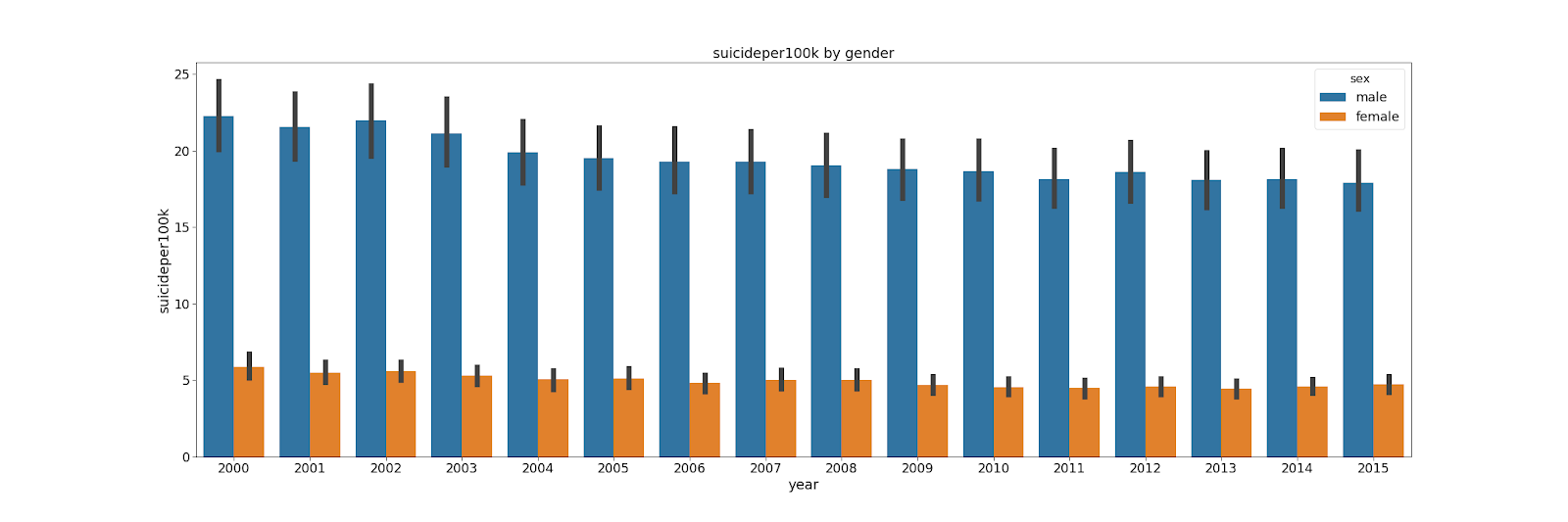


The dataset had large amount of data for generation X and silent and least data for generation Z. Data count on the basis of country:

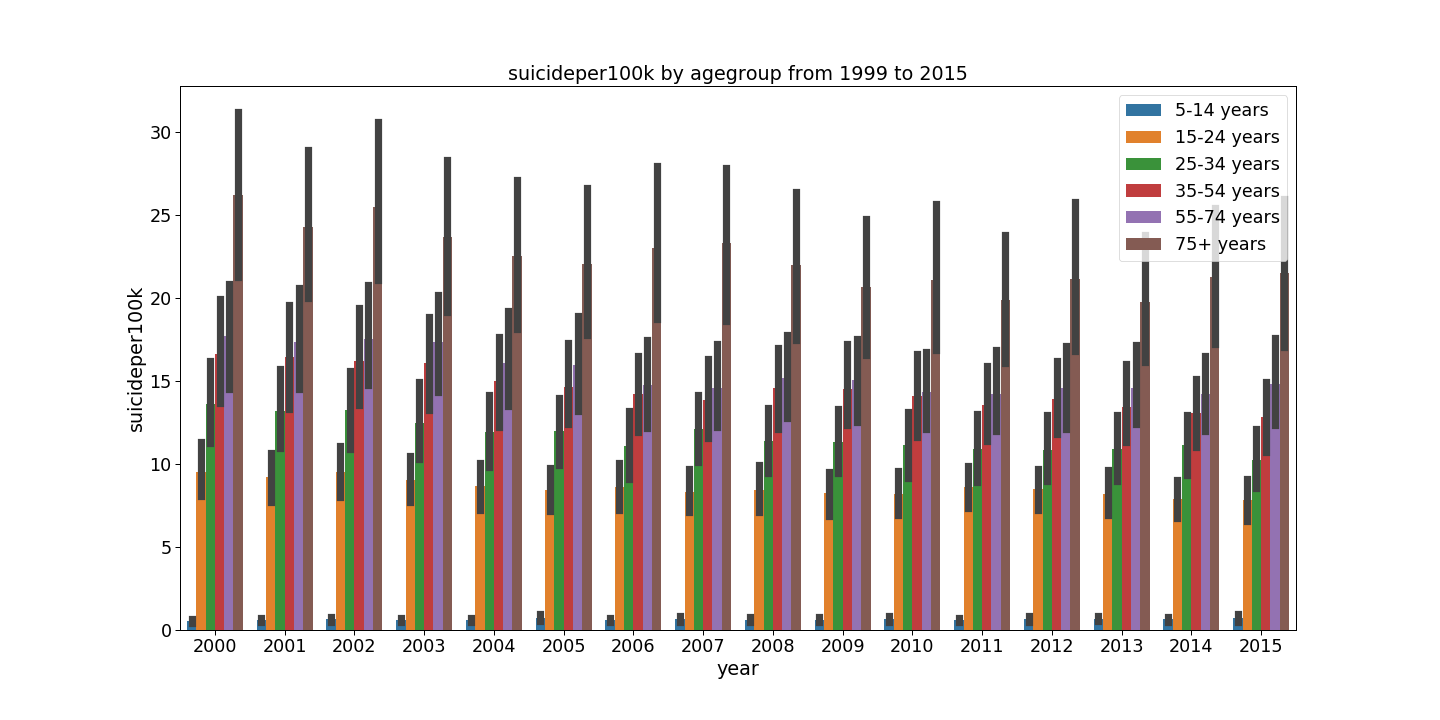


The number of data from each country is not equal. Now let’s see for year: -  


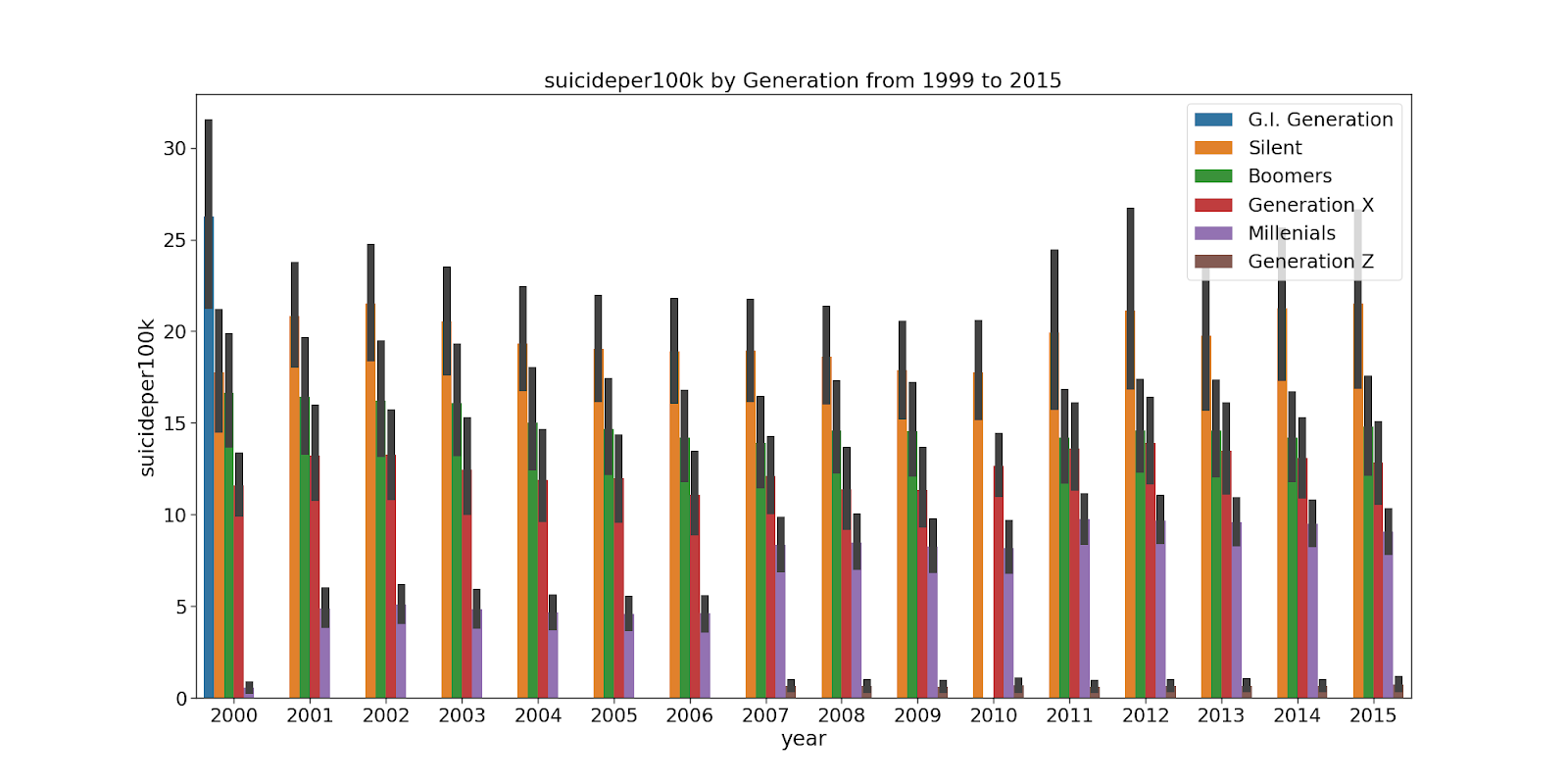
Each year does not have same amount of data and from all this counter plot, it can be said that every country did not have data for all year. The bar plot between suicideper100k and year from 2000 to 2015 differentiated by gender is as follow:



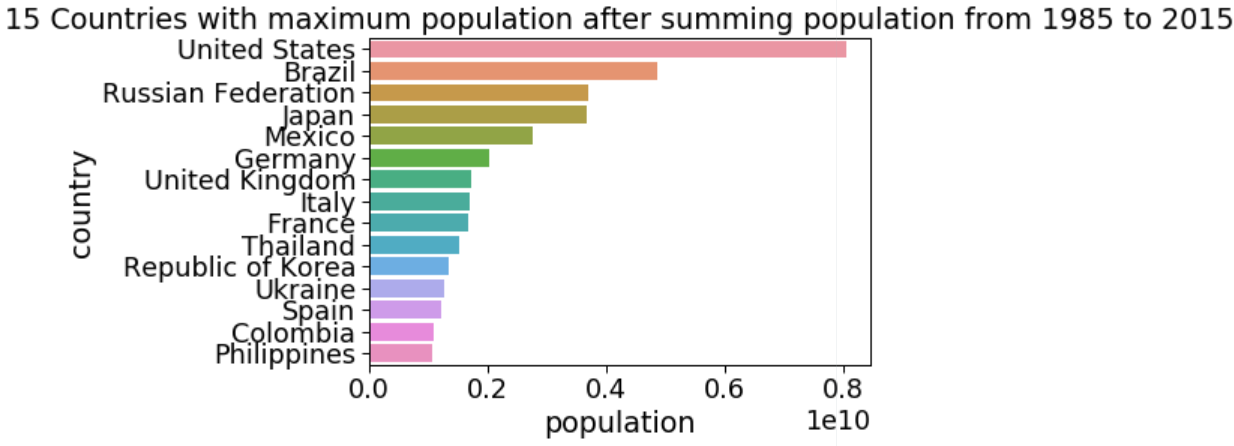
For all the years, the suicide per 100k of male is greater than that of female. Now, Let’s view the suicide trend by the age group



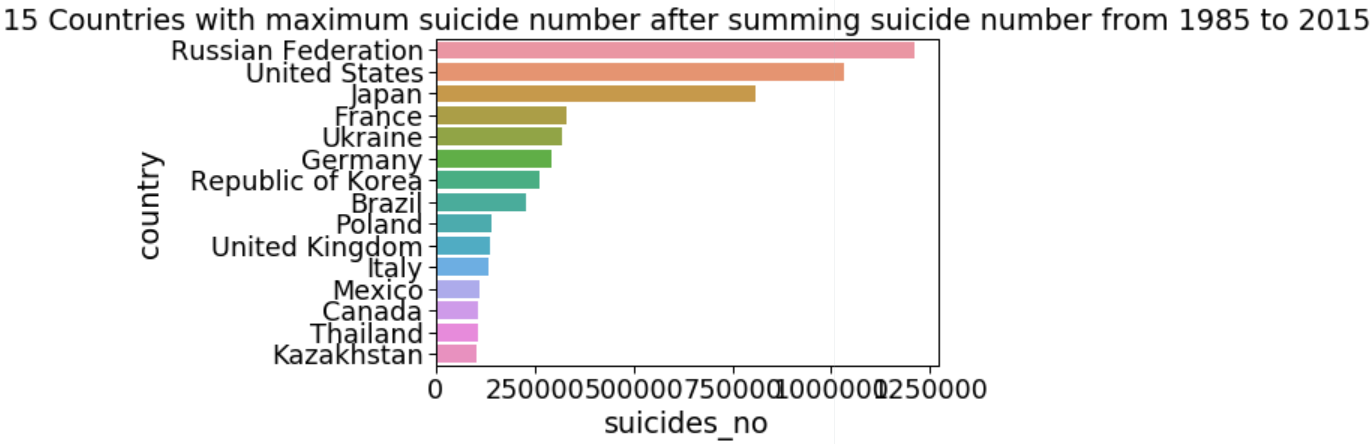
This 15-year information, shows that the age group 75+ did more suicide than other age group. From this graph it shows that with the increased in age group, the number of suicides is also increasing. Now, Let’s see the suicide per 100k bar plot according to generation.



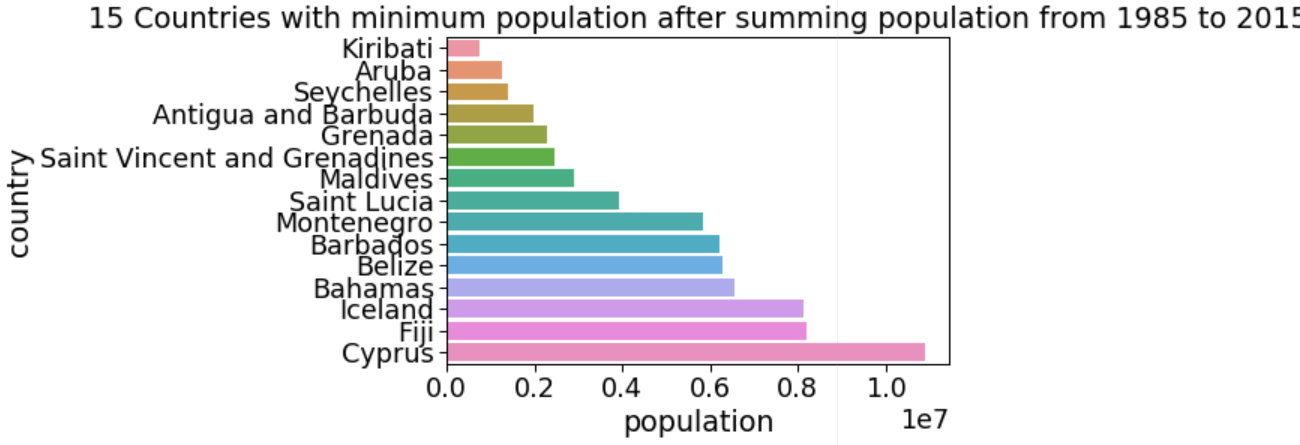
I have plotted from 2000 to 2015 and 2000 only has G.I. Generation. If i neglect that generation for this graph, i can say that generation silent has highest suicide and this is relatable to the age group information. Plot showing 15 countries with maximum population and maximum suicide number after summing the data for all year on the basis of country is as follow:

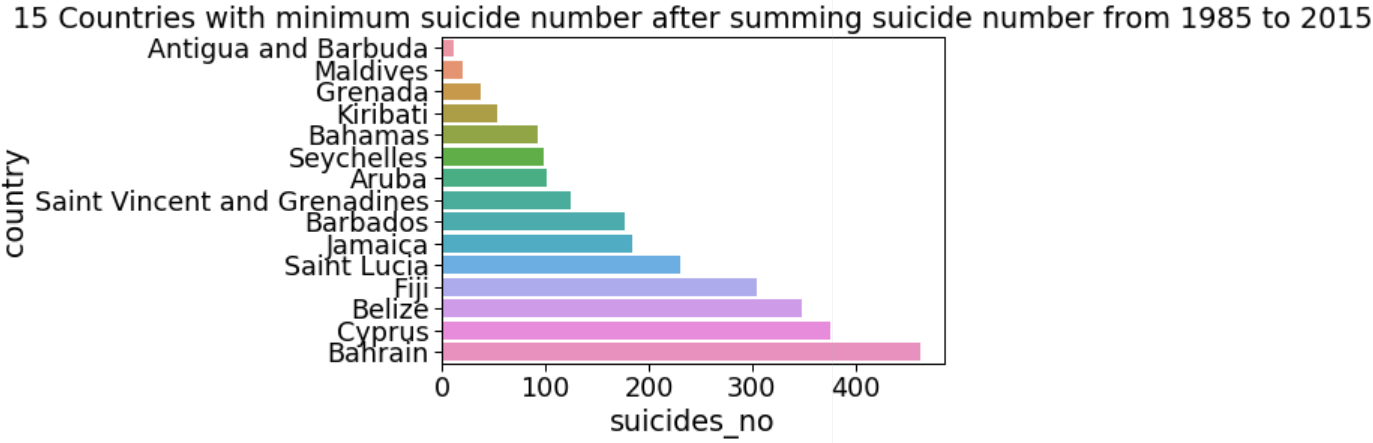


Russian Federation had third highest population sum but the suicide sum of Russian Federation is highest while the United states is in second position after having the highest sum of population.



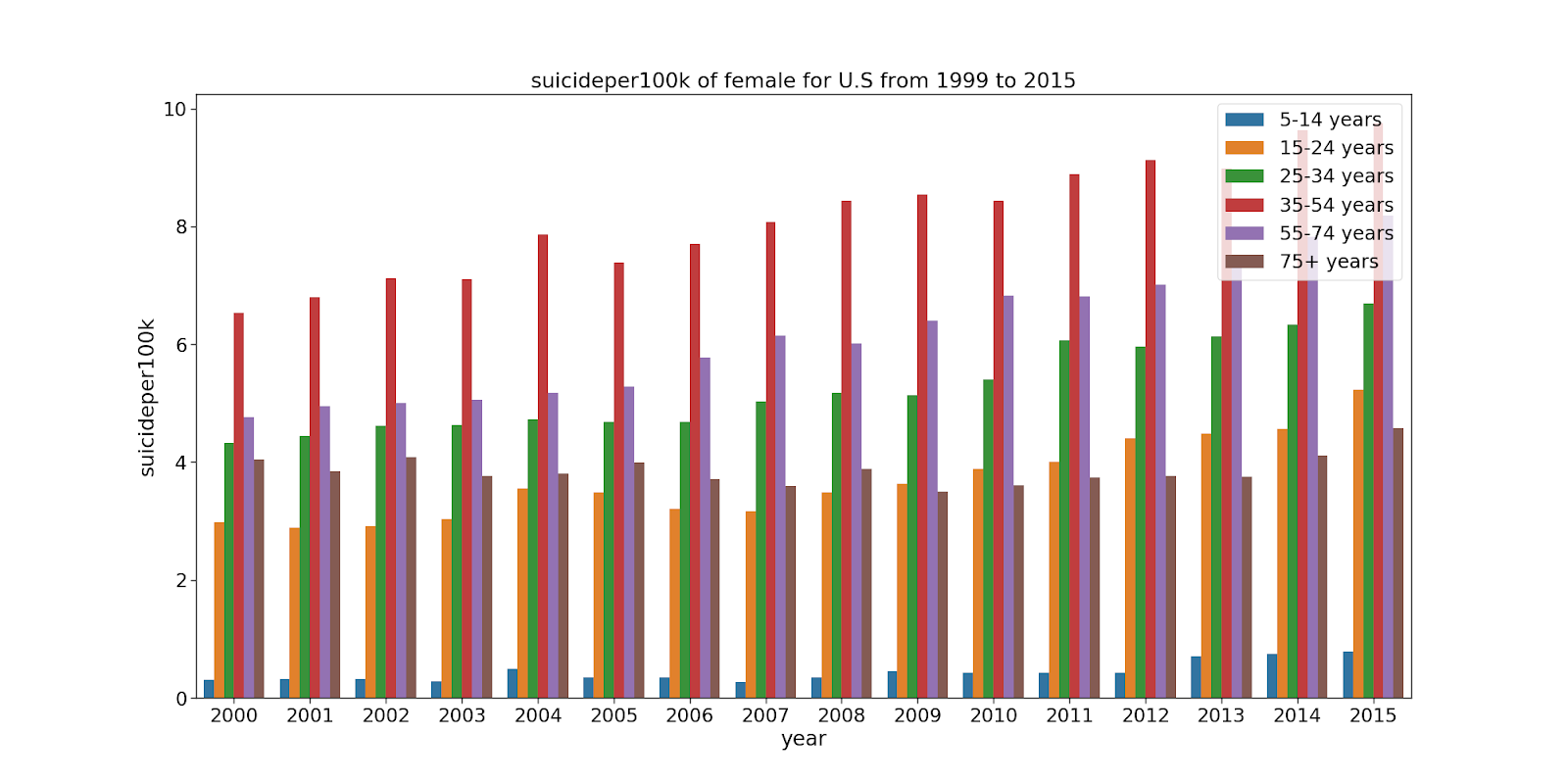
Now, Let’s see plot showing 15 countries with minimum population and minimum suicide number after summing the data for all year on the basis of country.



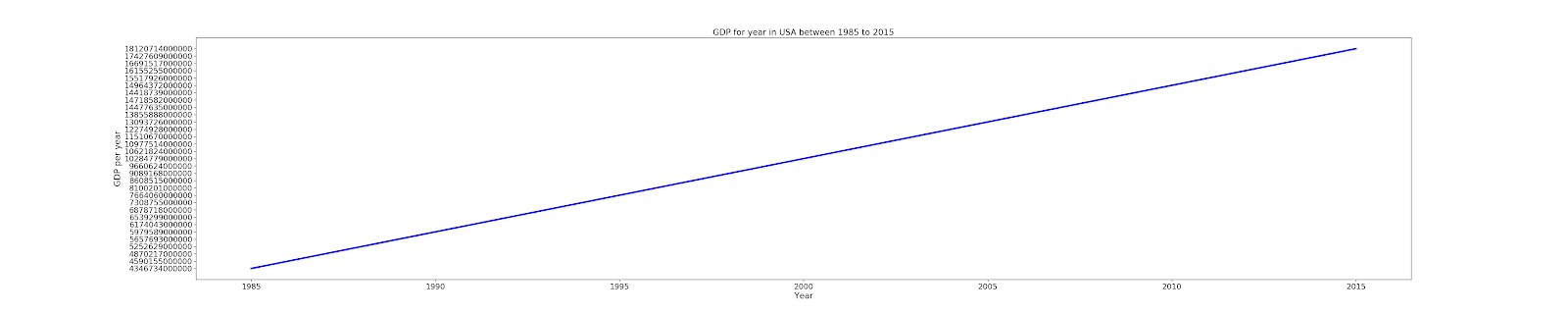


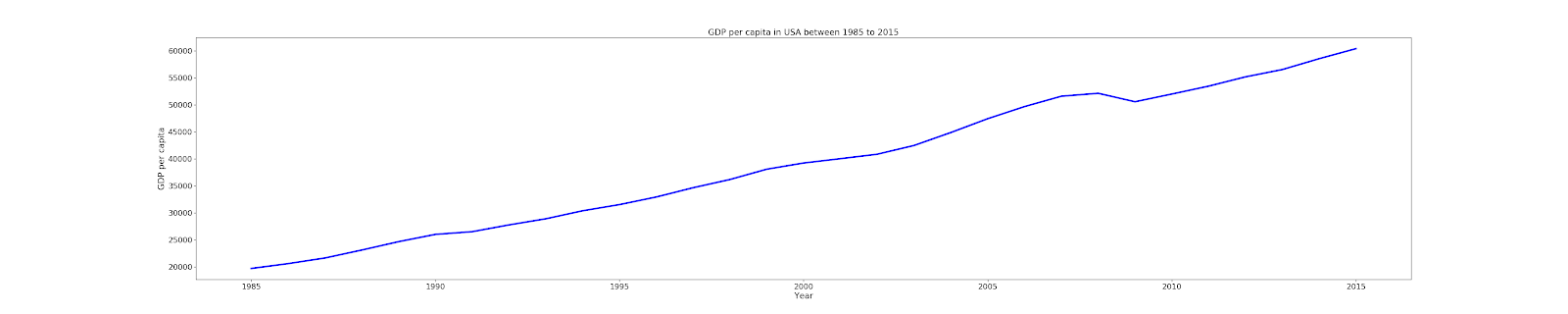
Antigua and Barbuda are the fourth least population country after summing and it is first country with least suicide after summing suicide number for all years.

Data visualization of United states is as follow: -



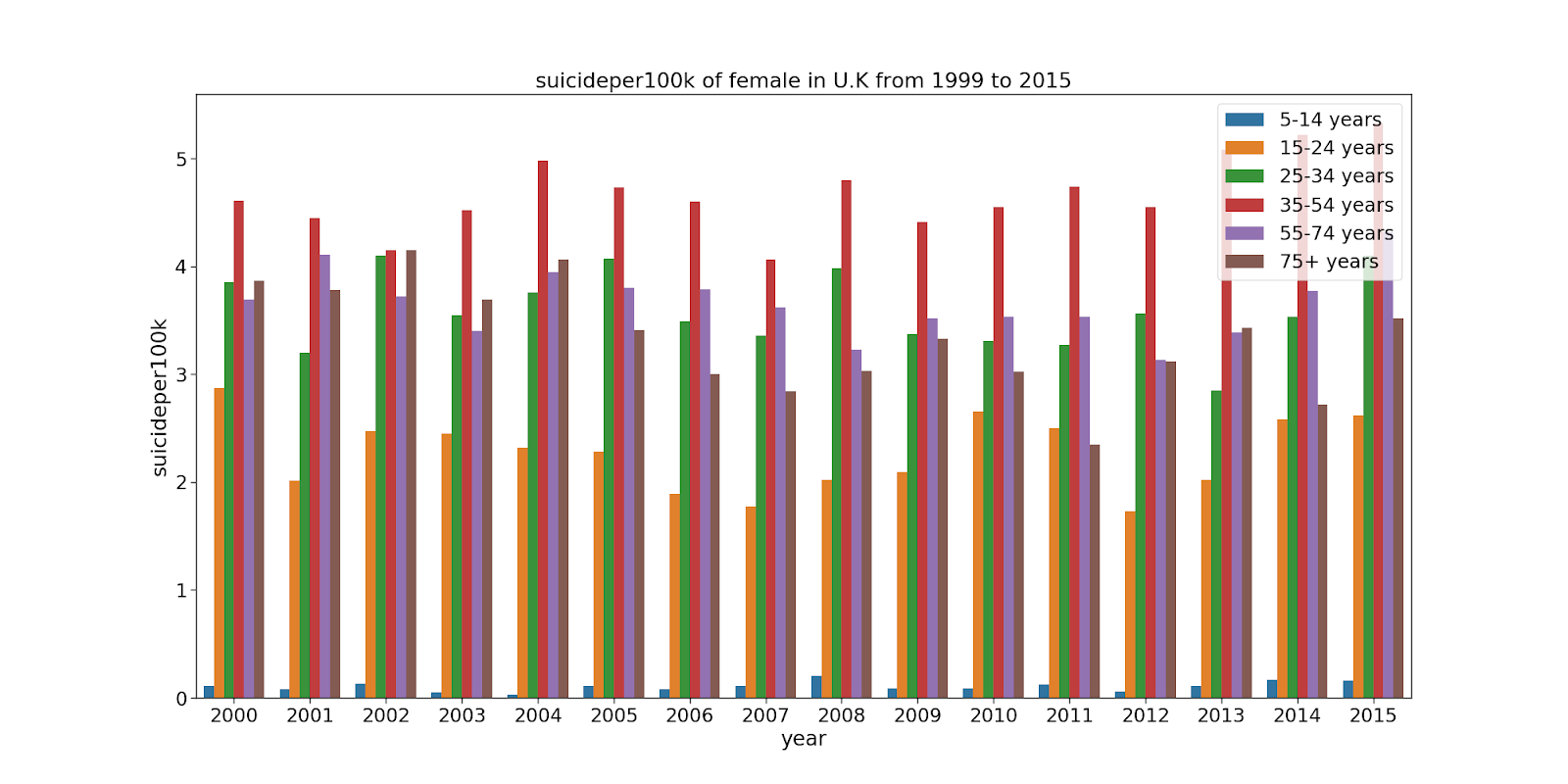


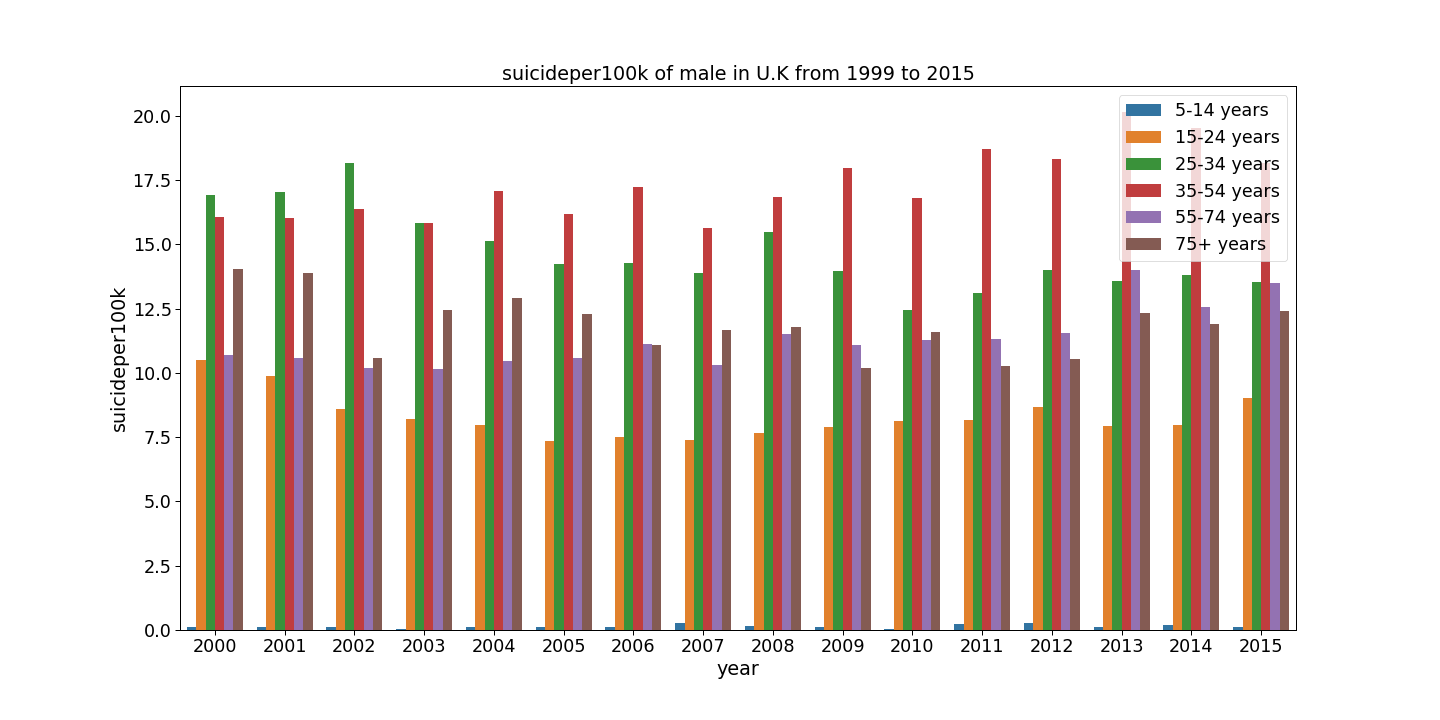


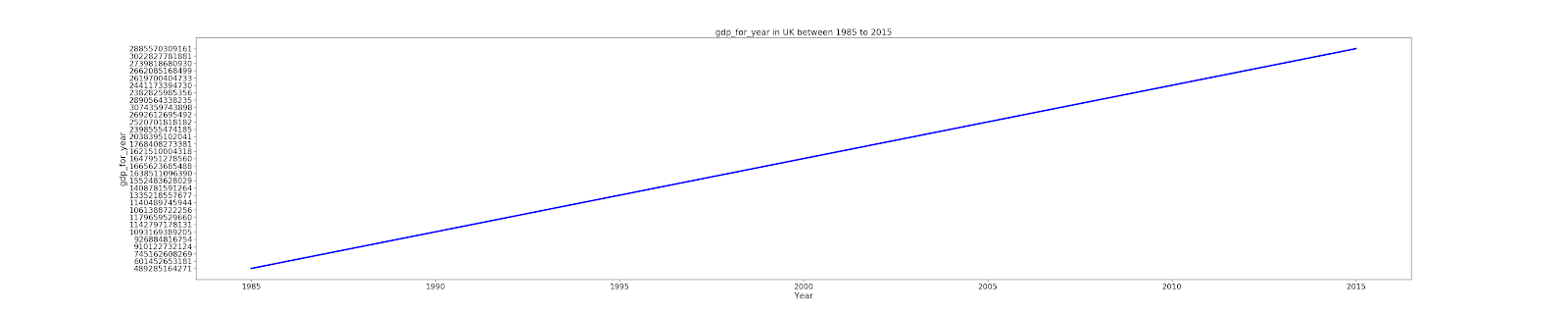


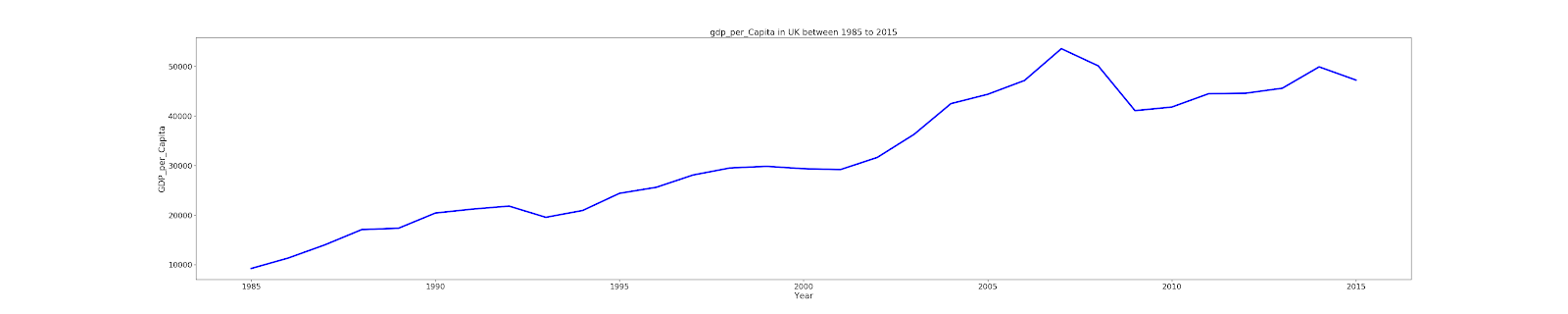
Suicide rate of male is higher than that of female for all the age group in USA. In female age group 35-54 has the highest suicide rate while in male 75+ age group has highest suicide rate. We can see small fluctuation in trend of suicide rate of both male and female. Difference between male suicide rate and female suicide rate is high. GDP for year is linear which means it is increasing with the year and GDP per capita is also increasing with year expect around 2010 where there was sudden decrease and after that it is increasing.

Data visualization of United Kingdom is as follow



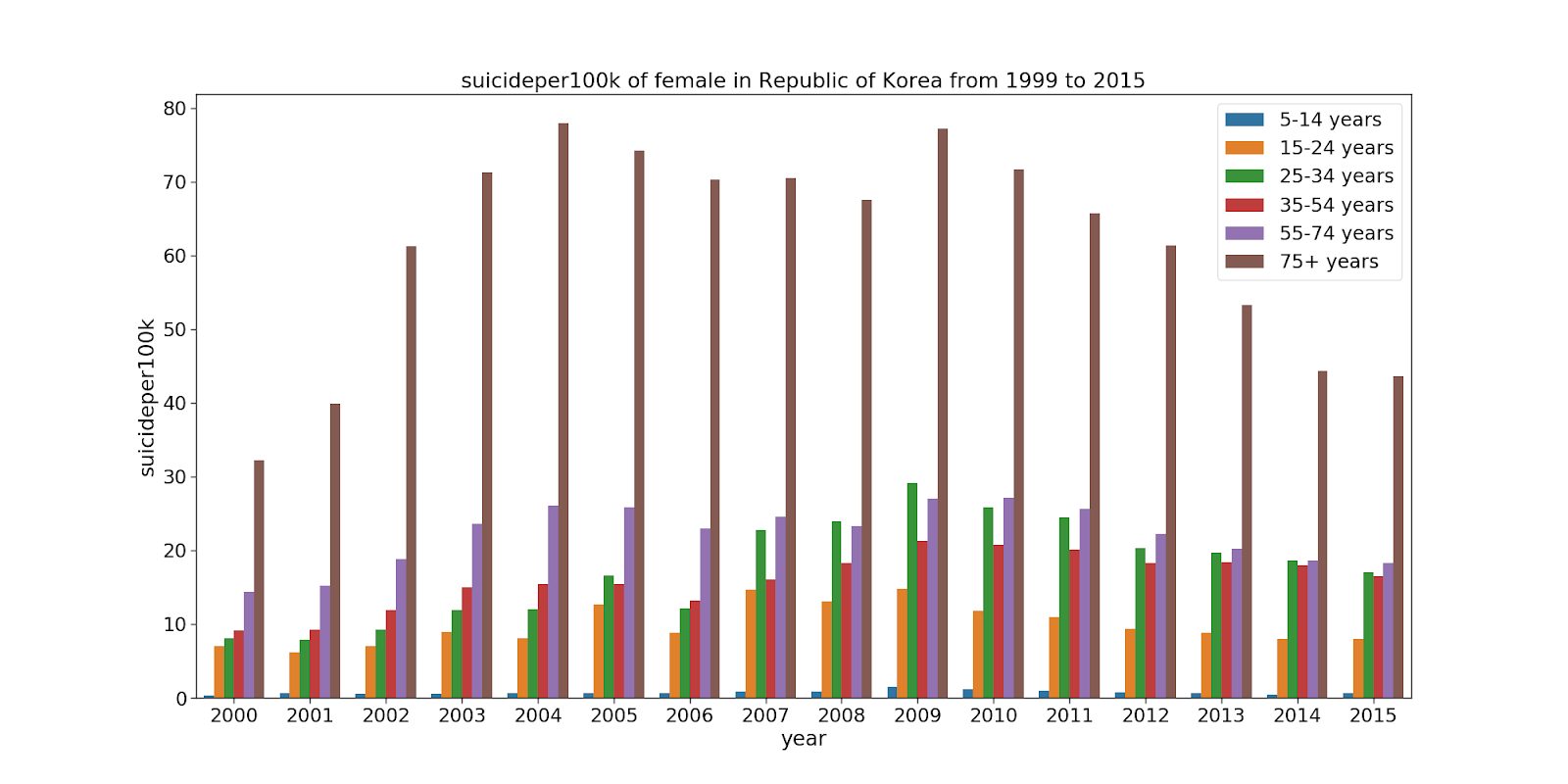




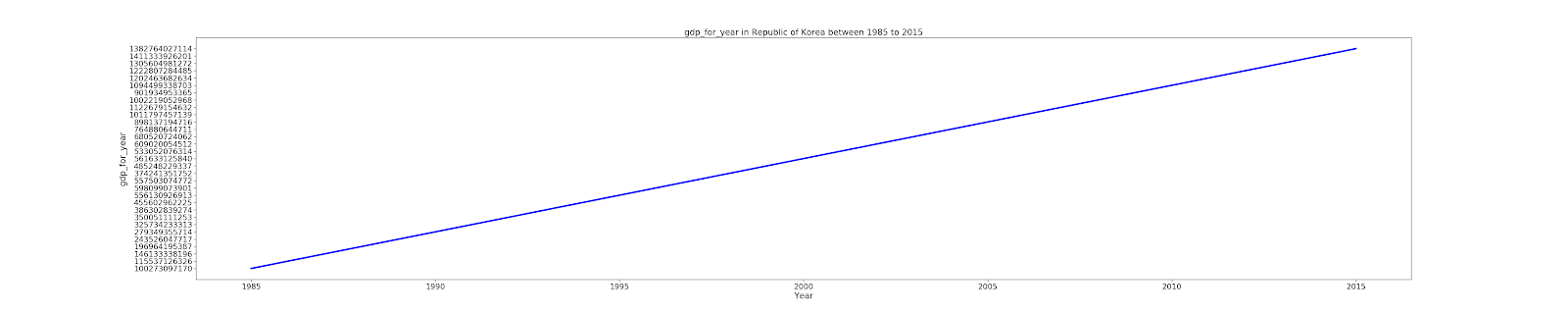


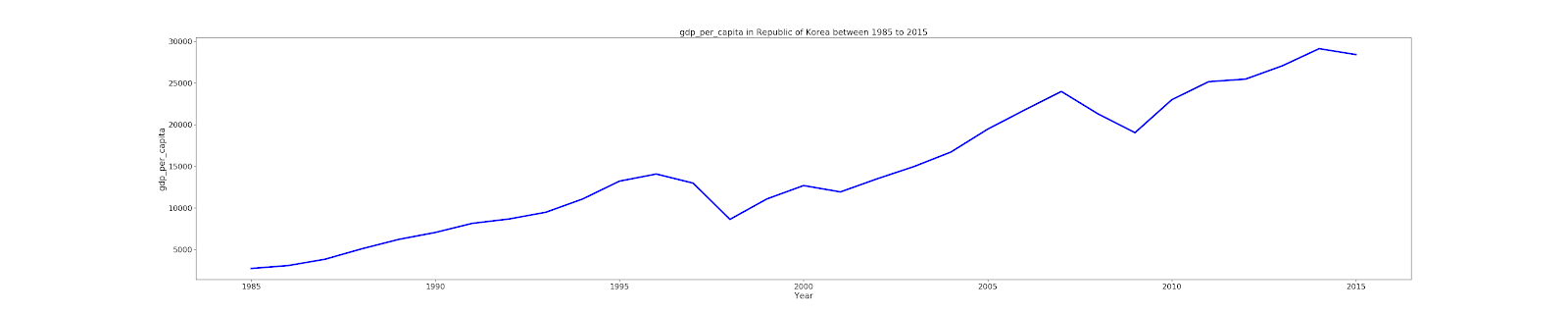
In UK, male suicide rate is higher than that of female for all the years and as in USA, the age group 35- 54 of female has the highest suicide rate in women age group. Except for 3 years, age group 35-54 of male have highest suicide rate. From this it can be said that people of age group 35-54 suicide more in U.K.GDP for year is increasing with year but GDP per capita is fluctuating a lot with the year.

Data visualization of Republic of Korea as follow



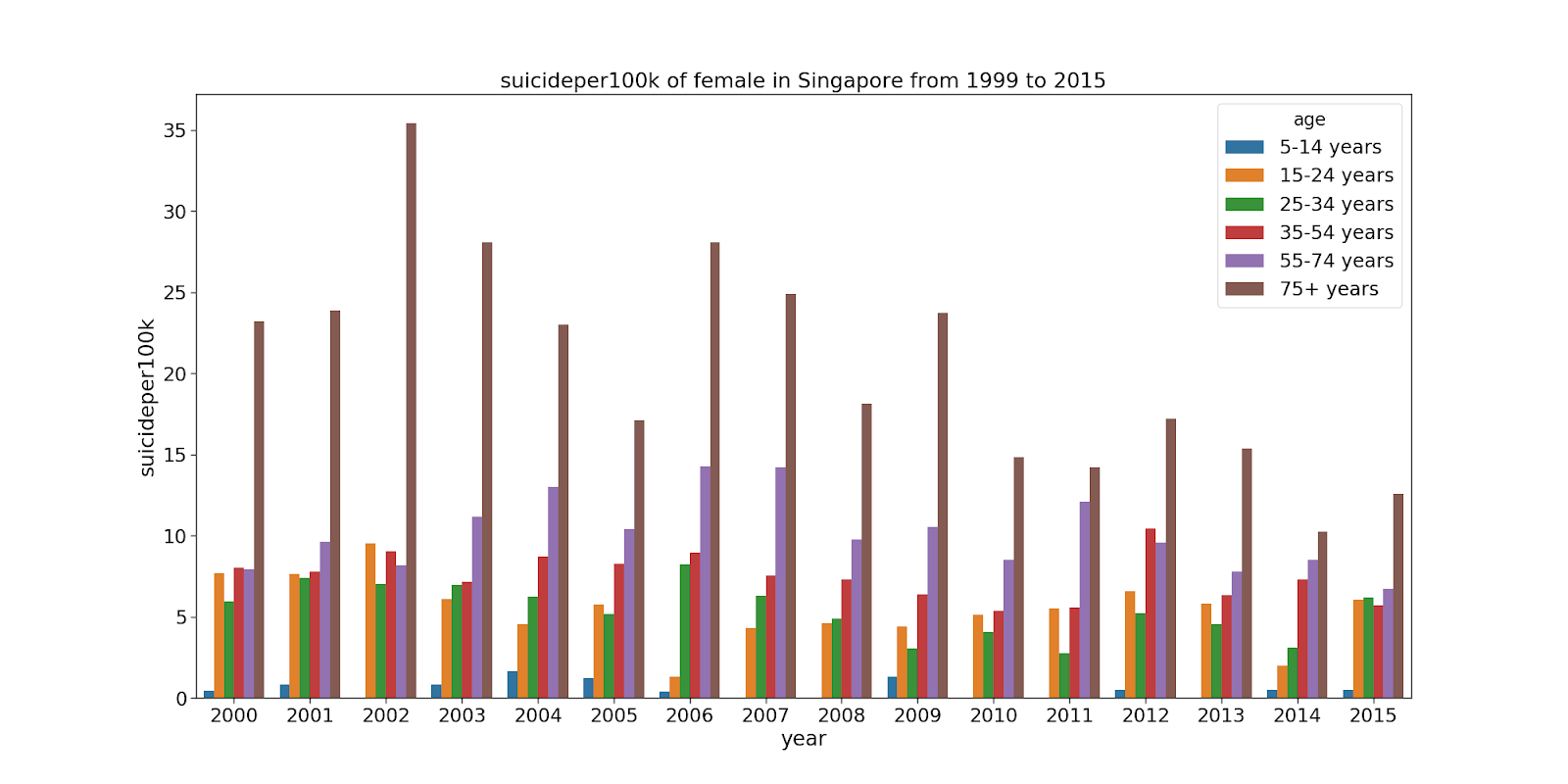


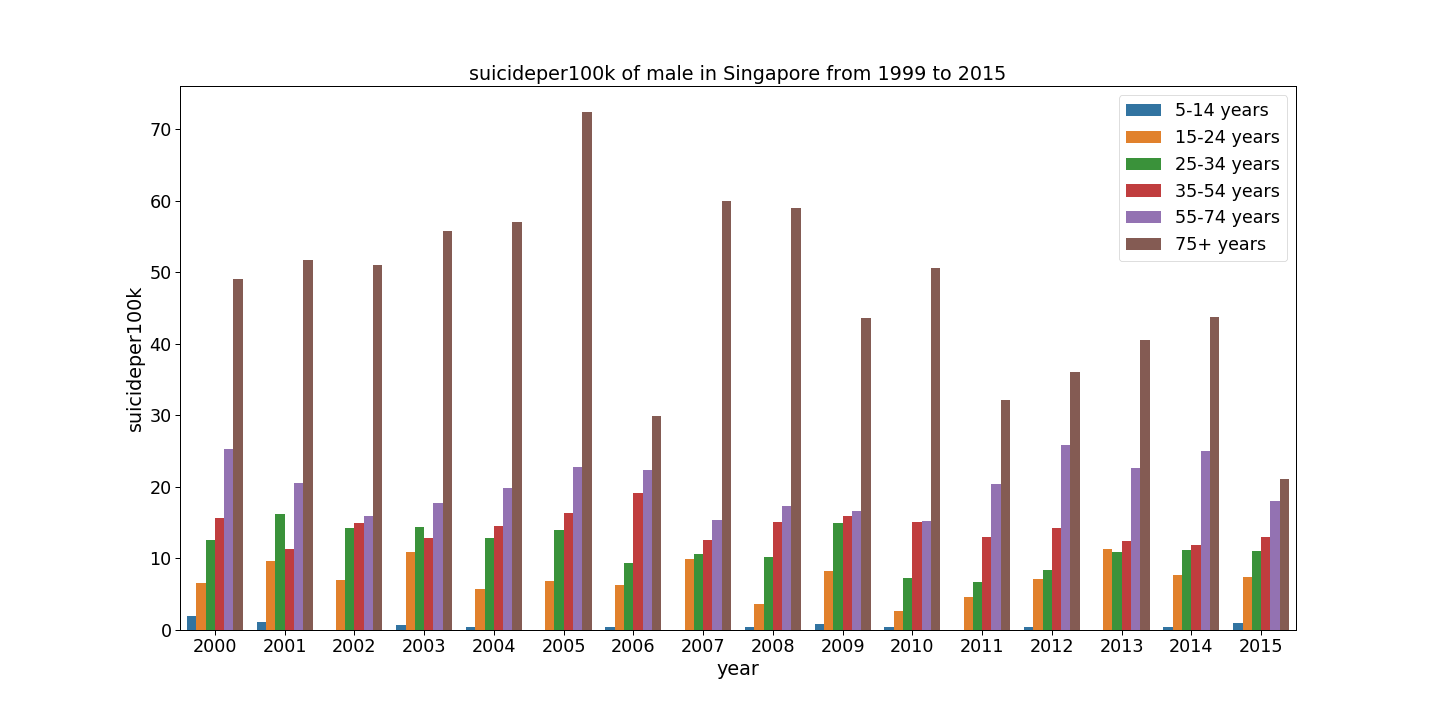


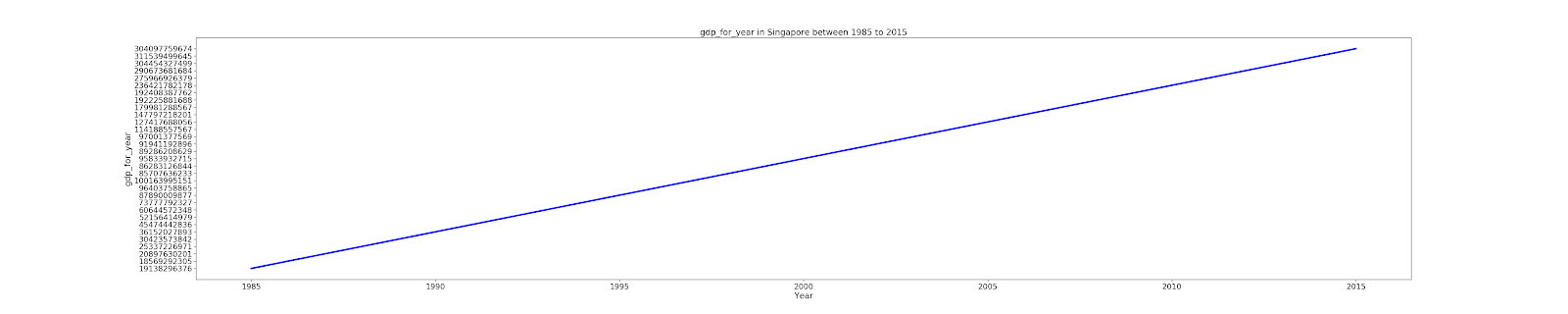


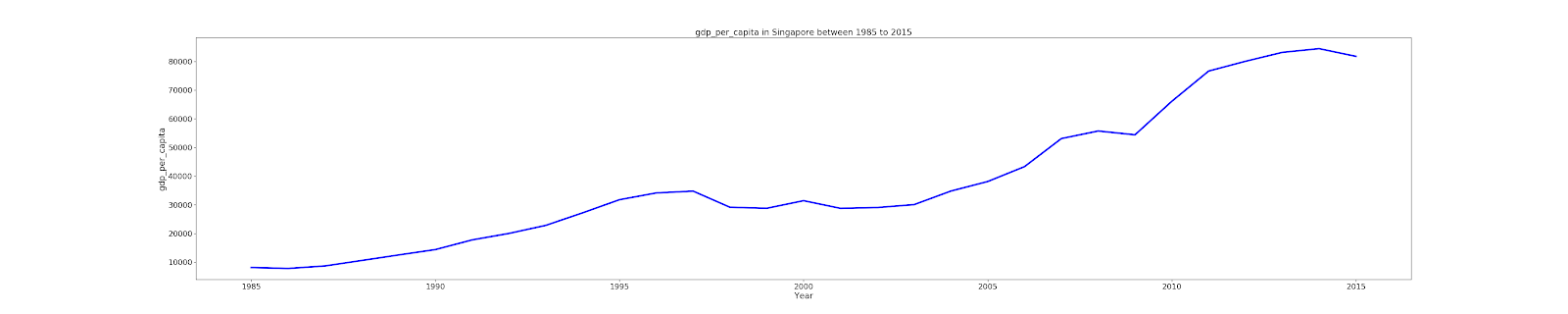
In Korea, age group 75+ had highest suicide rate in both male and female and as in above two country, this country also has high male suicide rate than that of female but the number of suicides is very high than that of above country. GDP for year has linear relation with the year while GDP per capita is fluctuating as there is increase for certain year and then decrease and again starts increasing.

Data visualization of Singapore as follow



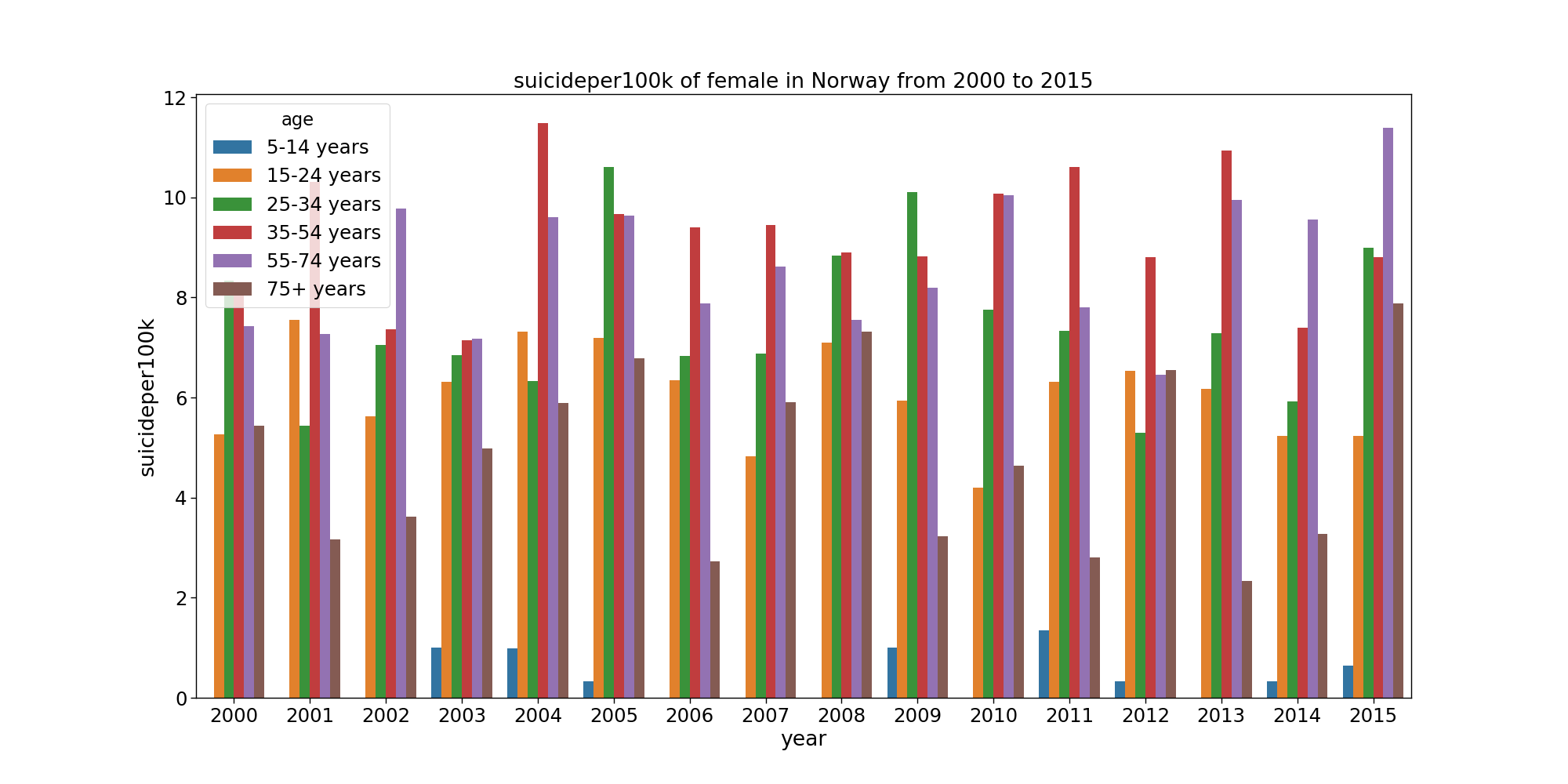


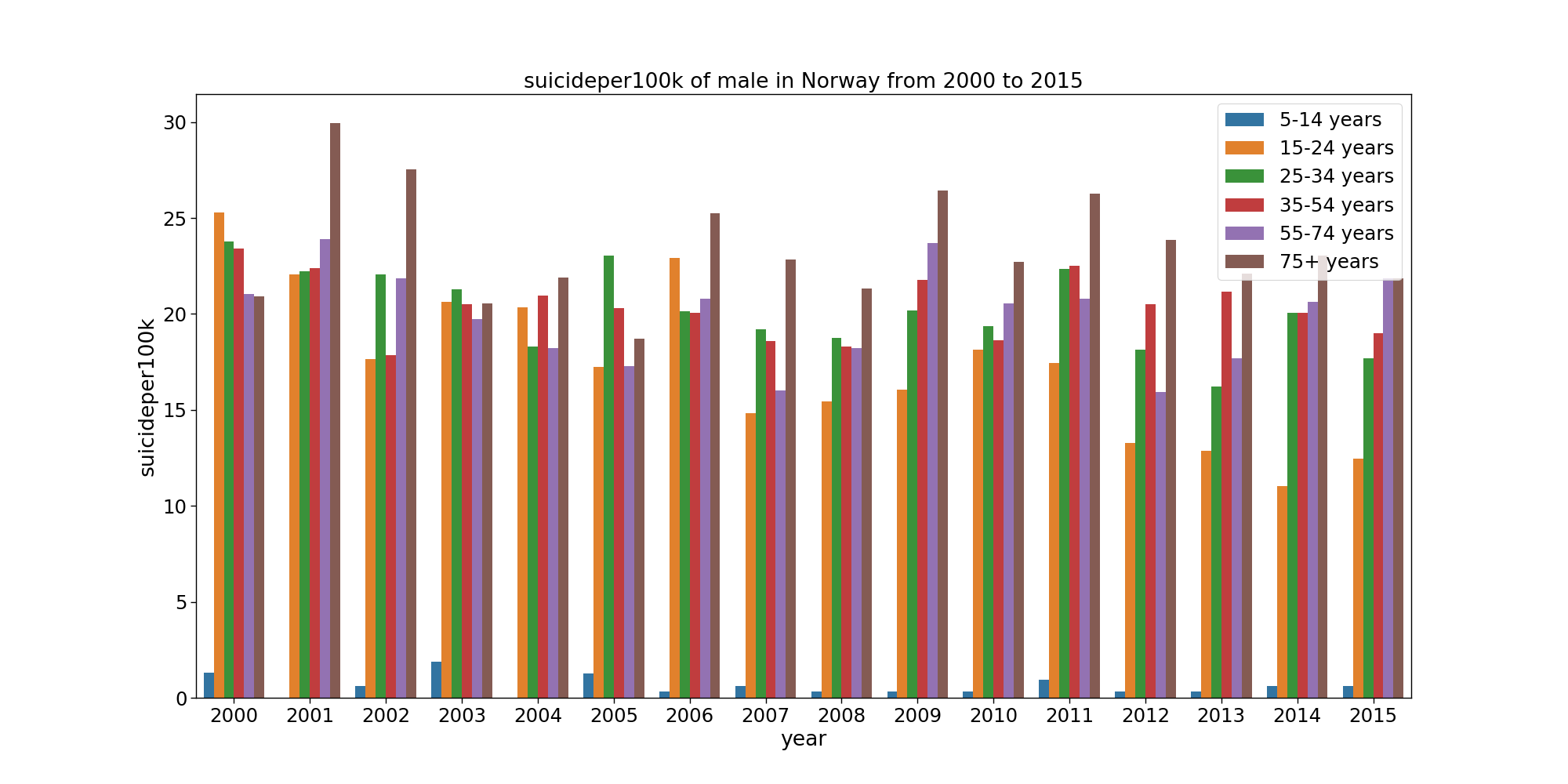


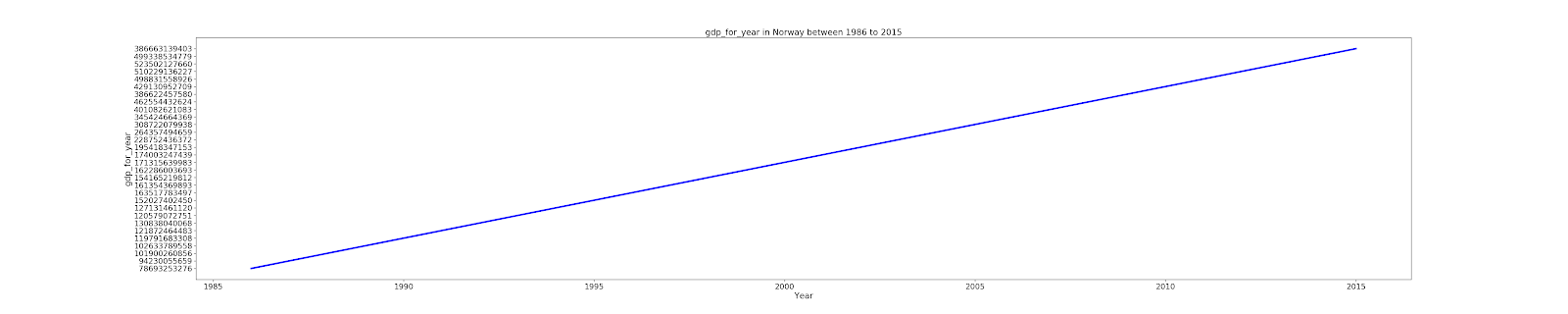


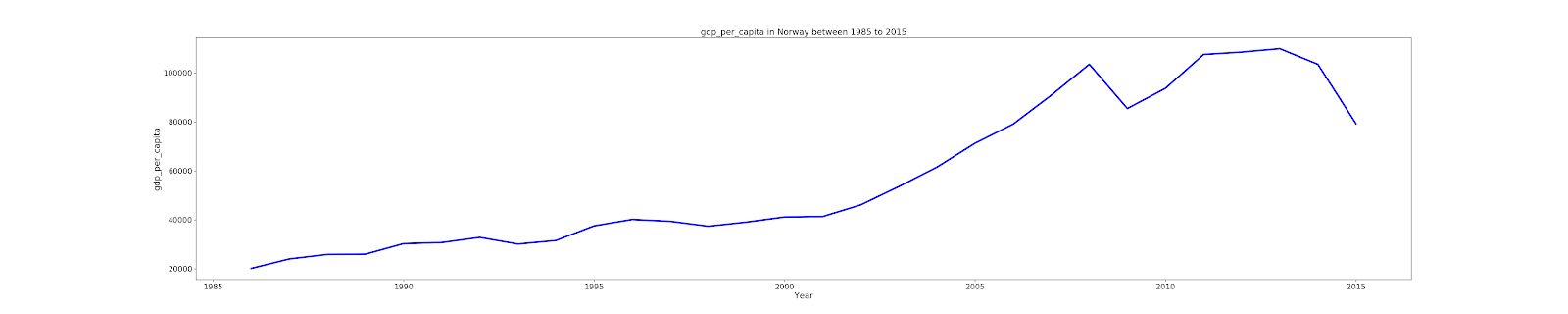
Singapore also has maximum male suicide compared to female. In both gender  age group 75 + had highest suicide and suicide trend is fluctuating . GDP for year has linear relation but the GDP per capita is fluctuating.

Data visualization of Norway as follow





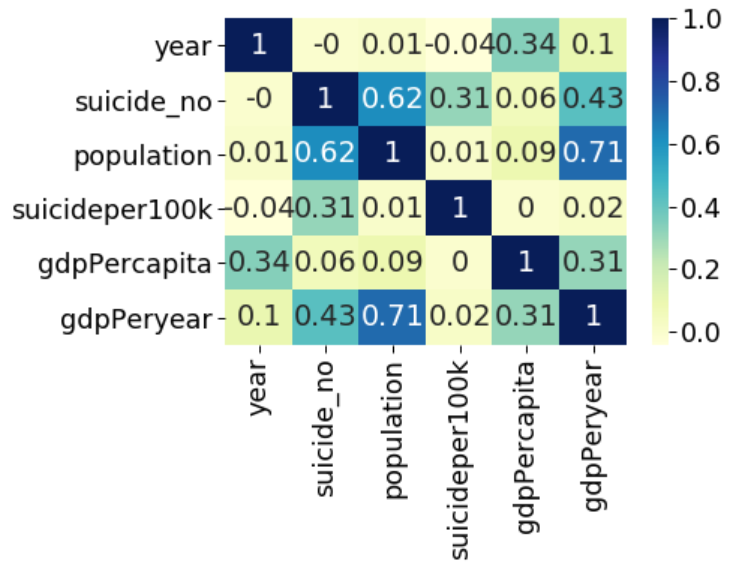




Norway is ranked in first position in HDI rank. This country also has highest male suicide compared to female. In male, 75+ age group have highest suicide but in case of female,

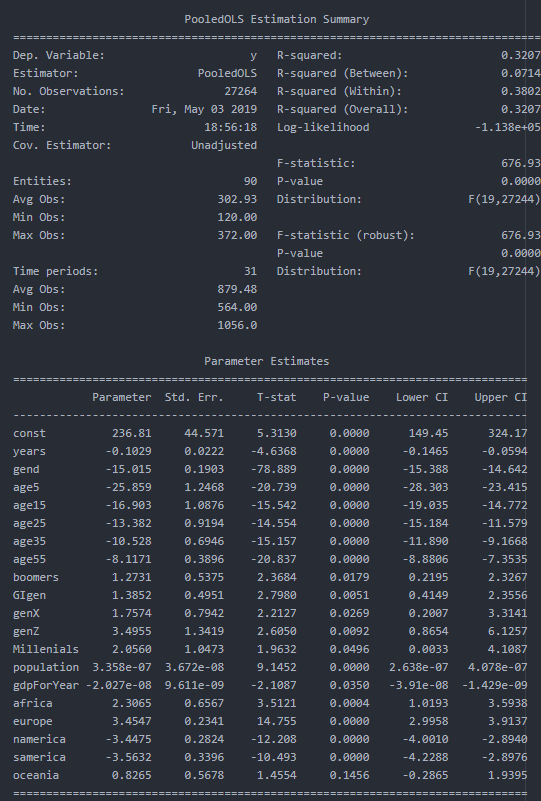
35-54 or 55-75 age group have high depending upon year. GDP for year has linear relation but GDP per capita has fluctuation.

The correlation between the numerical value of this data set is as follow:



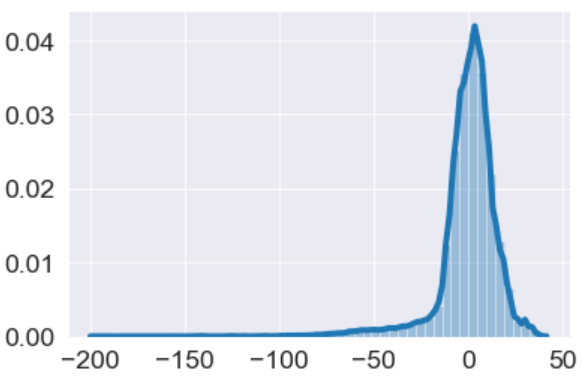
The first model is as follow:

Result of PooledOLS is as follow: -

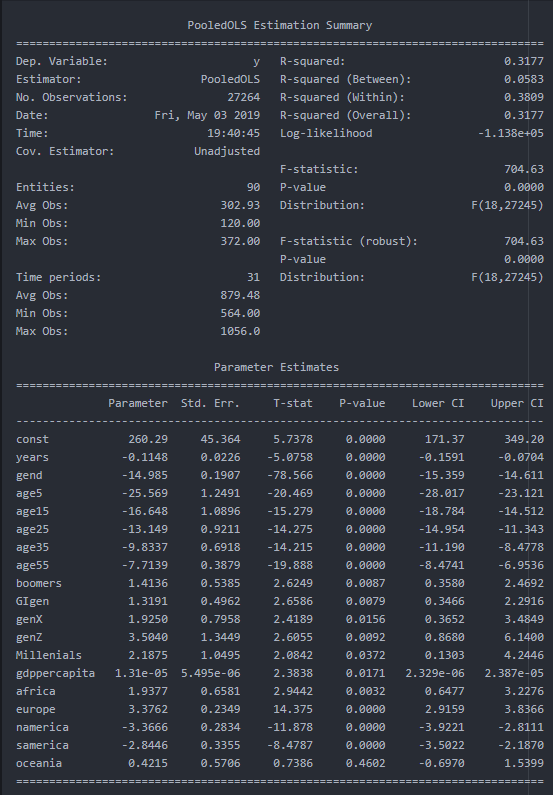


The first model can be written as: -

For master dataset, male is base gender, 75+ group is base age group, silent is base generation and Asia is a base continent. If everything is constant then 75+ age group females have 15.01 less suicide per 100k then that of male in Asia for silent generation. Similarly, if the age group is between 5-14 then suicide number per 100k will be 25.85 less than of 75+ age group in silent generation. If everything is same but the continent is Europe then the suicide per 100k will be 3.45 more than that of Asia. All the independent variables are statistically significant at 15% significance level.   
Here, the value of R-squared is 0.3207 which means that model explains 32.07% of variance in dependent variable by the explanatory variable. The Mean absolute error is 10.04, Mean squared error is 246.67 and Root mean squared error is 15.70. The distribution diagram of residual is as follow:

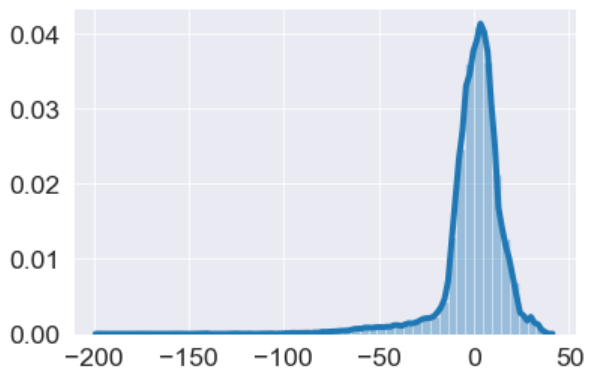


The Second model is as follow:

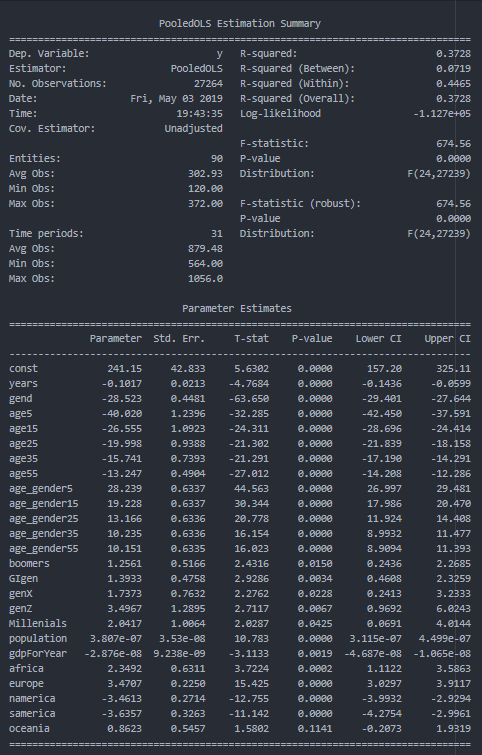


The model can be written as: -

If everything is constant then 75+ age group females have 14.98 less suicide per 100k then that of male in Asia for silent generation. Similarly, if the age group is between 5-14 then suicide number per 100k will be 25.56 less than of 75+ age group in silent generation. If everything is same but the continent is Europe then the suicide per 100k will be 3.78 more than that of Asia. If generation is genz (generation z) then female suicide per 100k will be 3.50 more than that of silent 75+ age group male. All the independent variables are statistically significant at 50% significance level.   
 Here, the value of R-squared is0.31 which means that model explains 0.31% of variance in dependent variable by the explanatory variable. The Mean absolute error is 10.03, Mean squared error is 247.67 and Root mean squared error is 15.74. The distribution diagram of residual is as follow:



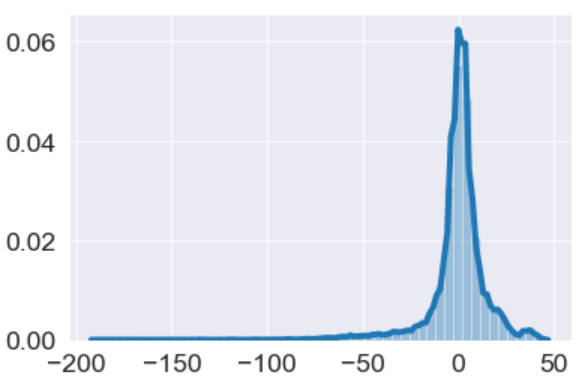
The third model is



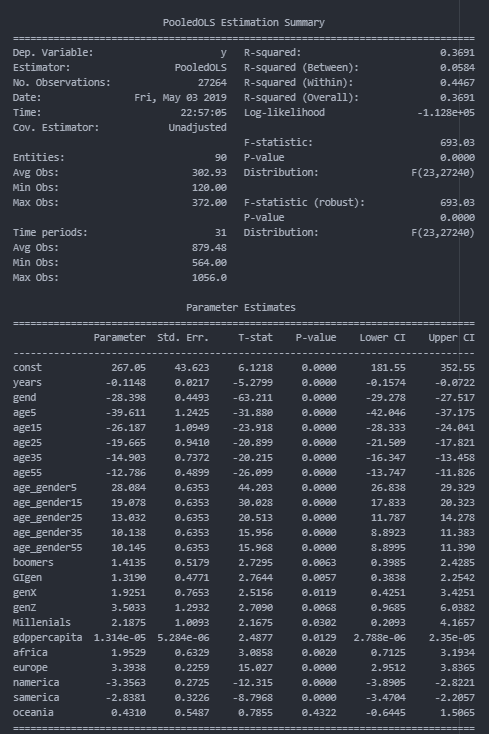
This model can be written as: -

The suicide per100k of female is 28.53 less than that of male of 75+ agegroup in silent generation in Asia. If the age range is 4-15 then, the suicide per 100k of female in silent will be (-28.523\*1-40.0201\*1+28.23\*1= -40.313) less by 40.313 then that of 75+ age group. European country will have 3.47 more suicide per 100k then that of Asian country but South America will have 3.63 less suicide per 100k then that of Asian country. The generation z will have 3.49 more suicide per 100k then that of silent generation. If population is increase by 10,000 then the suicide per 100k will be increased by 0.0038. In contrast, increase will GDP will decrease the suicide number. All the independent variables are statistically significant at 12 % significance level.

Here, the value of R-squared is 0.372which means that model explains 3.72% of variance in dependent variable by the explanatory variable. The Mean absolute error is 8.91, Mean squared error is 227.75 and Root mean squared error is 15.091. The distribution diagram of residual is as follow:

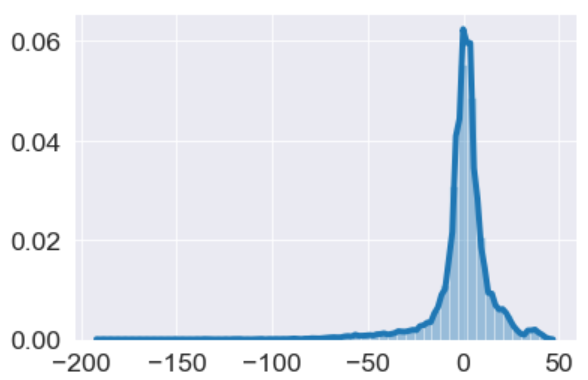


The fourth model is

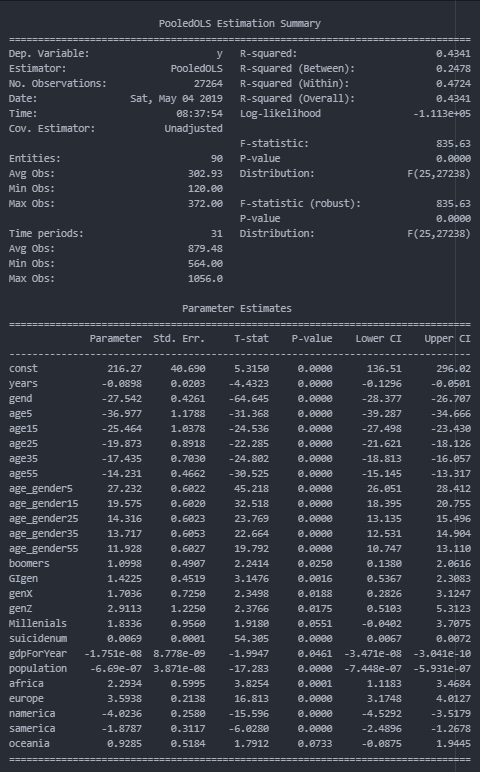


The model can be written as:

The suicide per100k of female is 28.39 less than that of male of 75+ agegroup in silent generation in Asia. If the age range is 4-15 then, the suicide per 100k of female in silent will be (-28.39\*1-39.61\*1+28.08\*1= -39.92) 39.92 less than that of 75+ age group. Increase in GDP per capita will increase the suicide per 100k. The European country will have 3.39 less suicide per 100k than that of Asia. The generation z has 3.50 more suicide per 100k than that of silent generation in Asia. All independent variables are statistically significant at 50% significant level.  
 Here, the value of R-squared 0.369 is which means that model explains 36.9 % of variance in dependent variable by the explanatory variable. The Mean absolute error is 8.91, Mean squared error is 227.755 and Root mean squared error is 15.091. The distribution diagram of residual is as follow:

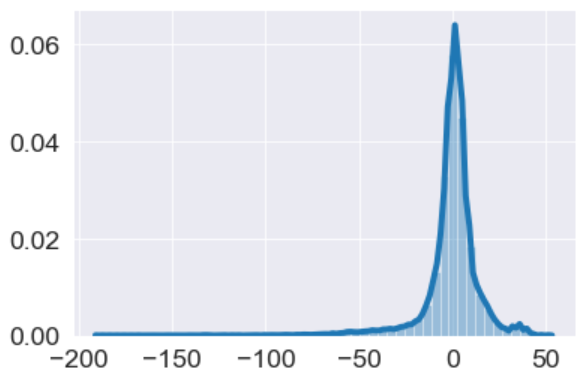


The fifth model is

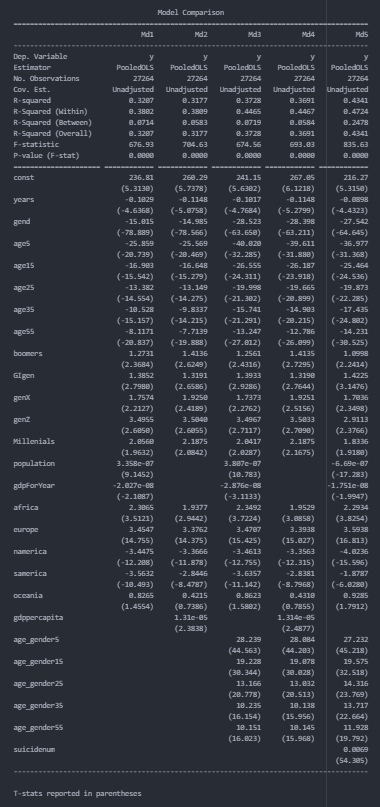


This model can be written as:

The suicide per100k of female is 27.54 less than that of male of 75+ agegroup in silent generation in Asia. If the age range is 4-15 then, the suicide per 100k of female in silent will be (-27.54\*1-39.97\*1+27.23\*1= -40.28) 40.28 less than that of 75+ age group. After adding the suicide number in the regression, the population have negative impact in the suicide per 100k. Increase in population and GDP will decrease suicide per 100k and increase in suicidenum will increase the suicide per 100k. With the increase in one year there will be decrease in suicide per 100k by 0.0898. All independent variables are statistically significant at 10 % significance level. Here, the value of R-squared is 0.434 which means that model explains 43.4% of variance in dependent variable by the explanatory variable. The Mean absolute error is 8.499, Mean squared error is 205.533 and Root mean squared error is 14.33. The distribution diagram of residual is as follow:



The result for comparison of model

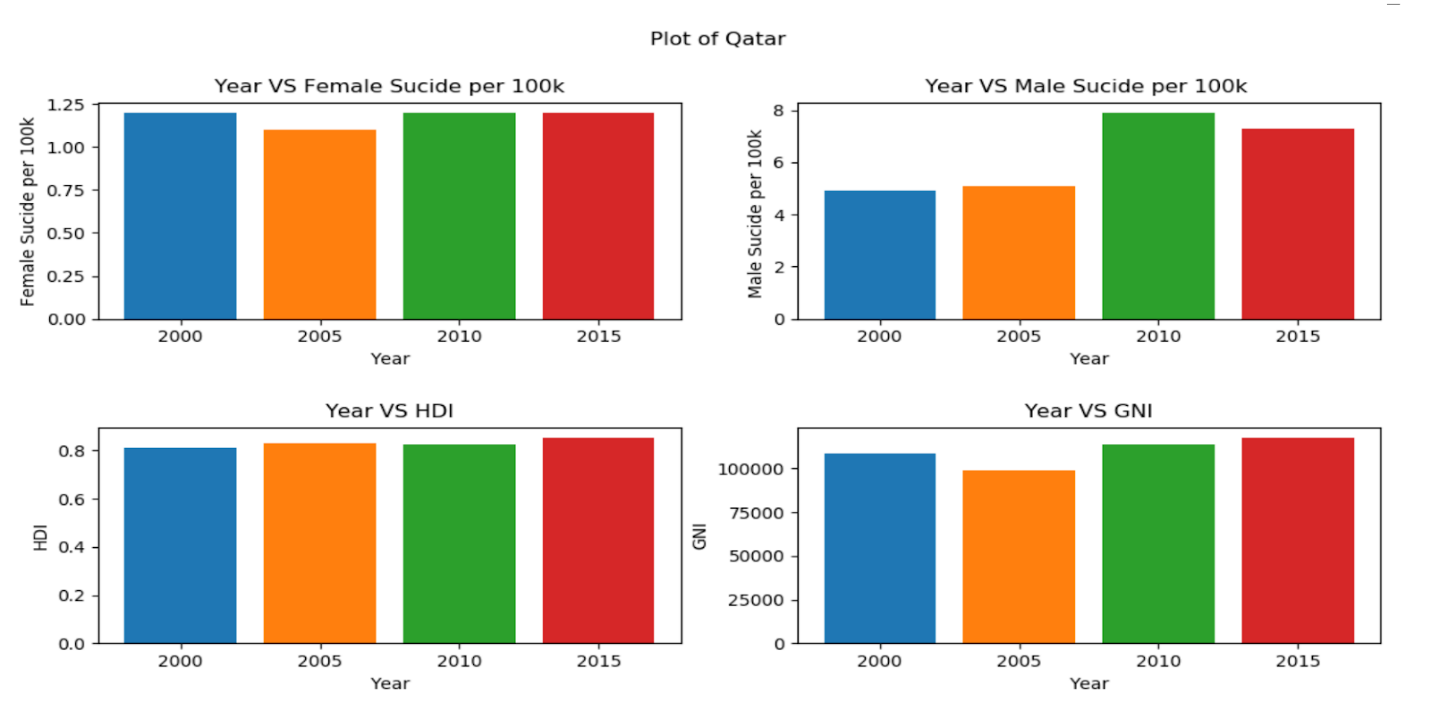


Here Md represents model and most of the t- statistic value of parameter is greater in model5. T- statistic is a measure of the likelihood that the actual value of the parameter is not zero so, larger t- value is somehow considered to be nice but other factors do matter. Similarly, R- squared value of model 5 is greater than that of other models which means this model explains higher percentage of variance in dependent variable by the explanatory variable than that of other models. Mean absolute error, Mean squared error and Root mean squared error of model 5 is less than that of another model. Considering all this, I think model 5 will make better prediction of suicide per 100k.

Suicide\_Data\_for\_4 different\_year data set

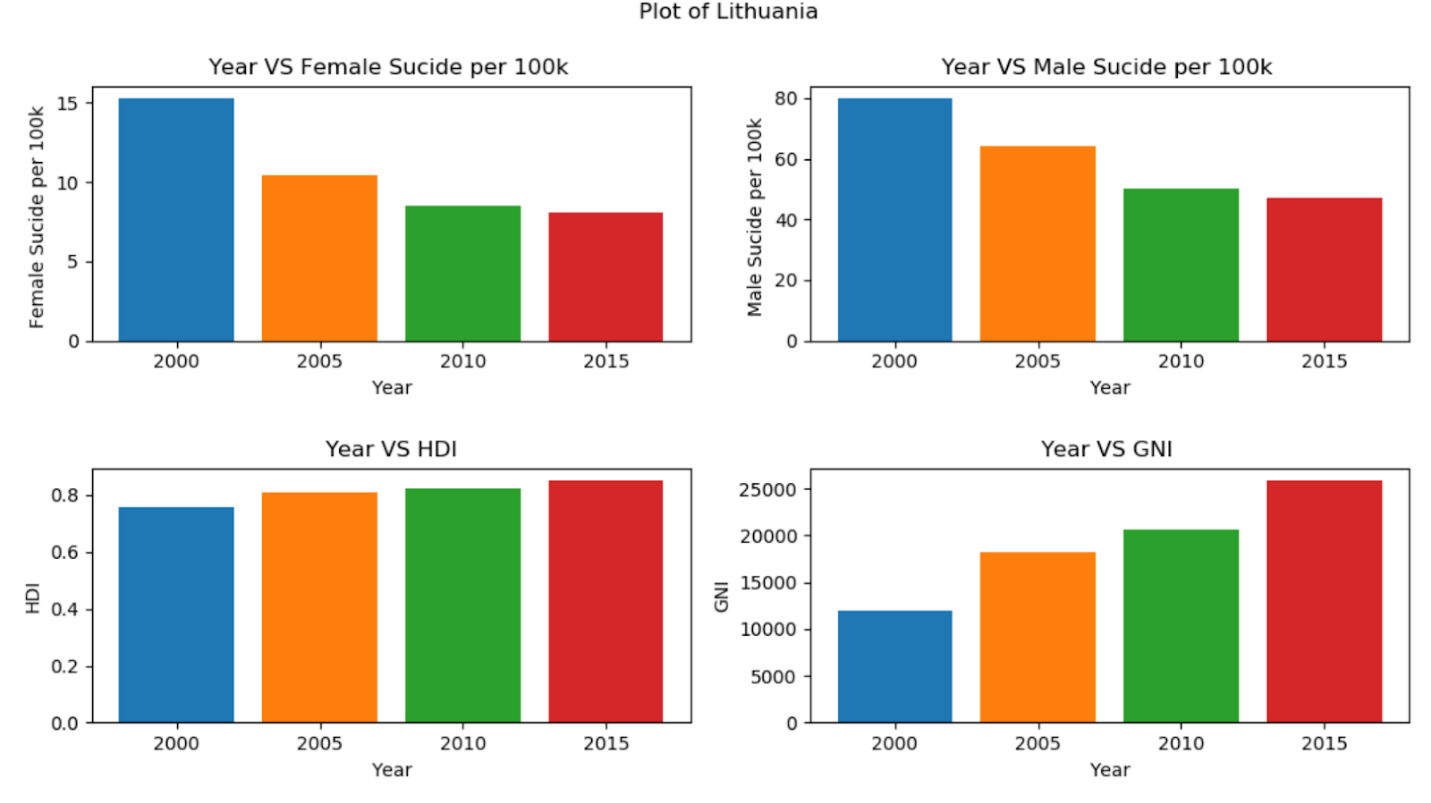
This data set has180 rows and 18 columns. From describe method output, I can say maximum value of suicide per 100k of female (23.4) and male (80) was high during year 2000 which eventually decreased in female with the succeeding years but in case of male there is fluctuation in the number but still it is less than that of year 2000. The mean value of female suicide per 100k for 2000,2005 ,2010 and 2015 is 6.146, 5.66, 5.35, 5.142 respectively and of male suicide per 100k for four year are 19.24, 17.74, 16.68, 15.93 respectively. Norway is ranked in position one in terms of Human development Index in all four years and there is fluctuation in suicide per 100k in both female and male. Highest female suicide per 100k was 7.1 in year 2005 and highest male suicide per 100k was 18.8 in year 2000. Nigeria is ranked at 189 position in terms of HDI. In order word it is the country listed with lowest HDI in this data set. Highest female suicide per 100k was 6.6 in year 2005 and highest male suicide per 100k was 12.2 during year 2005. Even the HDI ranked was higher for Norway the suicide number was more in Norway than in Nigeria. Gross National income of Qatar was highest for year 2000, 2010 and 2015. In 2005, female suicide per 100k is 1.1 which is less than that of other years while male had 5.1 which is greater than that of year 2000. For 2015, Africa has least GNI and female suicide per 100k is 9.3 which is less than year 2010(10.1) and male suicide per 100k is 30.3 which is less than year 20101(32.1). Female minimum suicide per 100k has zero value for some countries, I cannot certainly say that this may or may not be true because generally it impossible to have zero female suicide in a year in a country. They may have placed zero in the country for which they did not have suicide or it may not have been reported too. As i mentioned earlier, the maximum female suicide per 100k was in year 2000 with 23.4 and the country is Maldives but till 2015 this has been reduced to 8.9. It has been ranked as 101 position in HDI rank of 2017. The maximum male suicide per 100k was 80 in year 2000 in Lithuania. Lithuania had highest male suicide in year 2010 also but at 2015, Srilanka has the highest male suicide per 100k for year 2015 which is 58.8.

Data visualization of Qatar



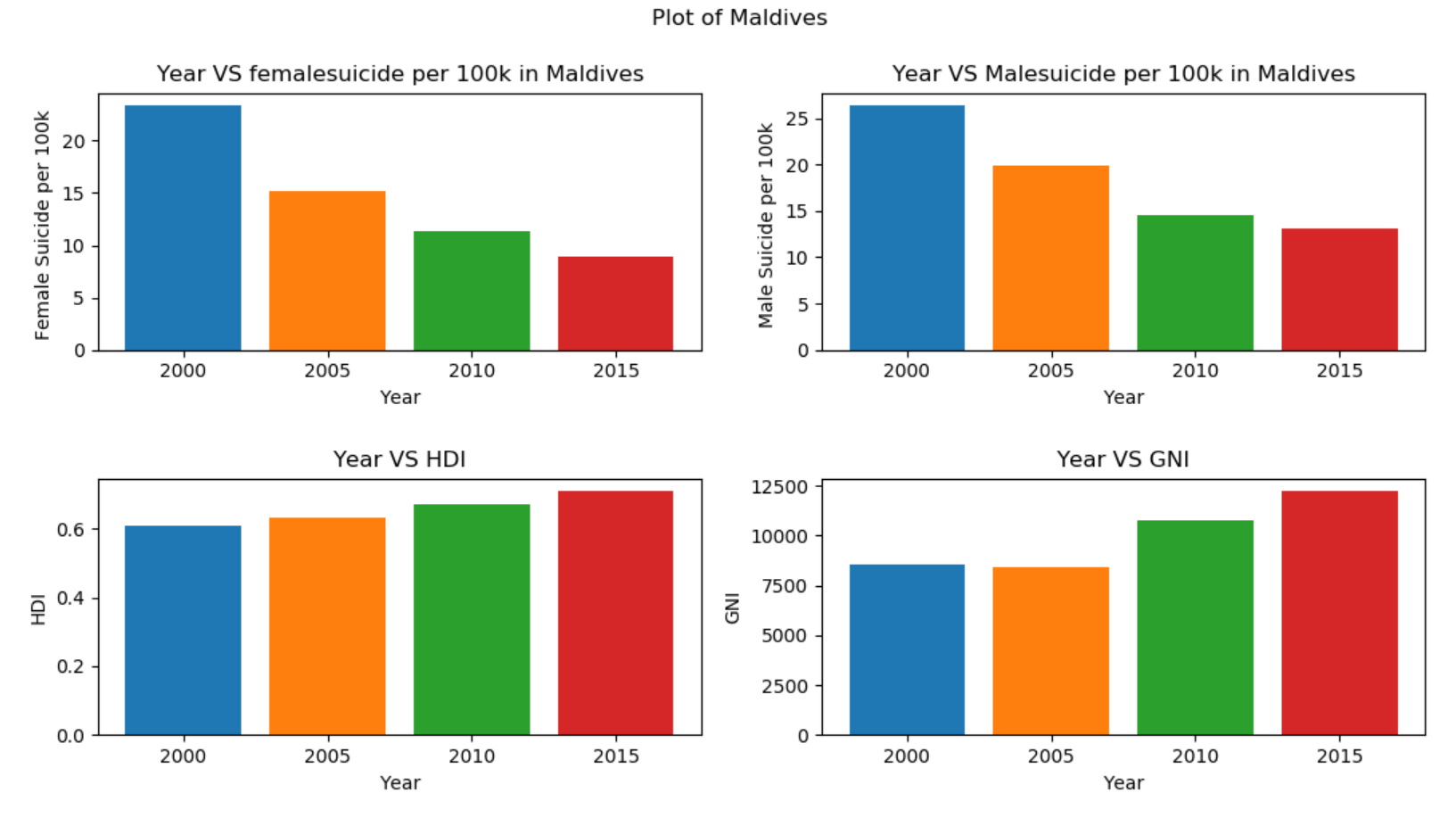
Qatar has highest GNI for three year and female suicide per 100k is below 1.25 and male suicide per 100k is below 8 for all four years. The mean value of suicide per 100k for four year for female is in between 5-7 and for male is in between 16-19. So, on basis of this it is possible to say that suicide in Qatar is less than average and there is huge gap between mean and its value. HDI value is increasing slightly and GNI value dropped in 2005 and after that it is increasing.

Data visualization of Lithuania



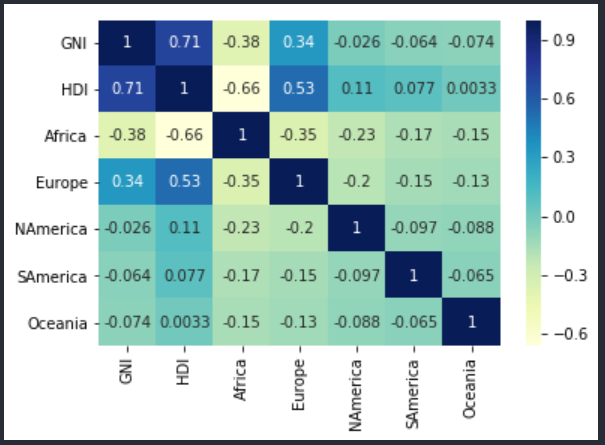
I have already mentioned that Lithuania had 80 for male suicide per 100k which is the greatest suicide value among the four year. The male suicide rate decreased by huge number from 2000 -2005 and 2005-2010. Similarly, female suicide per 100k is also about 15 which means it is thrice the mean value. The trend in female suicide value is also similar as that of male after 2000 it is decreasing. The HDI value increased from 0.756 to 0.809 and similarly GNI also increased from 11935 to 18273 in 2005. After this year, HDI and GNI both are increasing.

Data visualization of Maldives

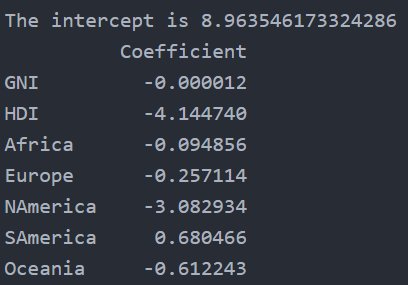
Female suicide per 100k is 23.4 and male suicide per 100k is 26.4 during 2000 for Maldives. In previous countries the difference between male and suicide per 100k was high. Male suicide per 100k was at least twice of female but in this country, it is just by 3 per 100k in 2000 and in other years also the gap is less. HDI value is increasing over the year while GNI decreased slightly from 8561 to 8388 in 2005 and after that it increased.

The first model is as below:

Here, female indicates female suicide per 100k. The correlation between independent variable is as follow:



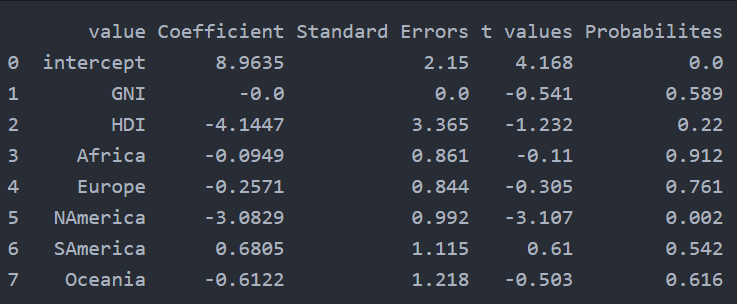
I used simple OLS method using sklearn package and the output for first model is as follow:



Now we can write model as:

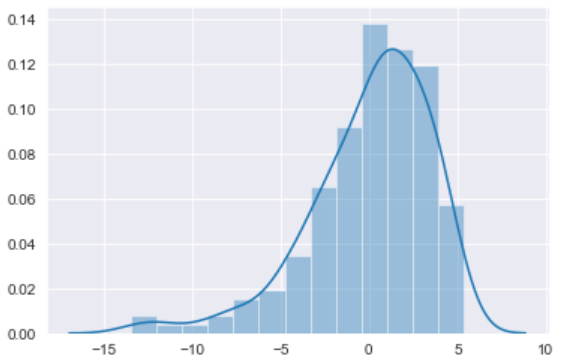
*female=8.96-0.000012GNI-4.144HDI-0.094Africa-0.2571Europe-3.082Namerica + 0.680Samerica - 0.612Oceania*

For all of the model Asia is a base continent. When every parameter of the model is constant and if HDI is increased by 0.1 in Asian continent, the suicide per 100k for female is decreased by 0.4144. Increase in both GNI and HDI will decrease the female suicide per 100k in Asia Continent. If GNI and HDI of two country is same but one is Asian and another is North American then North American country will have female suicide per 100k, 3.082 less than that of Asian country.

Coefficient with standards errors, t-values and probabilities values which has been rounded up are as follow: - 

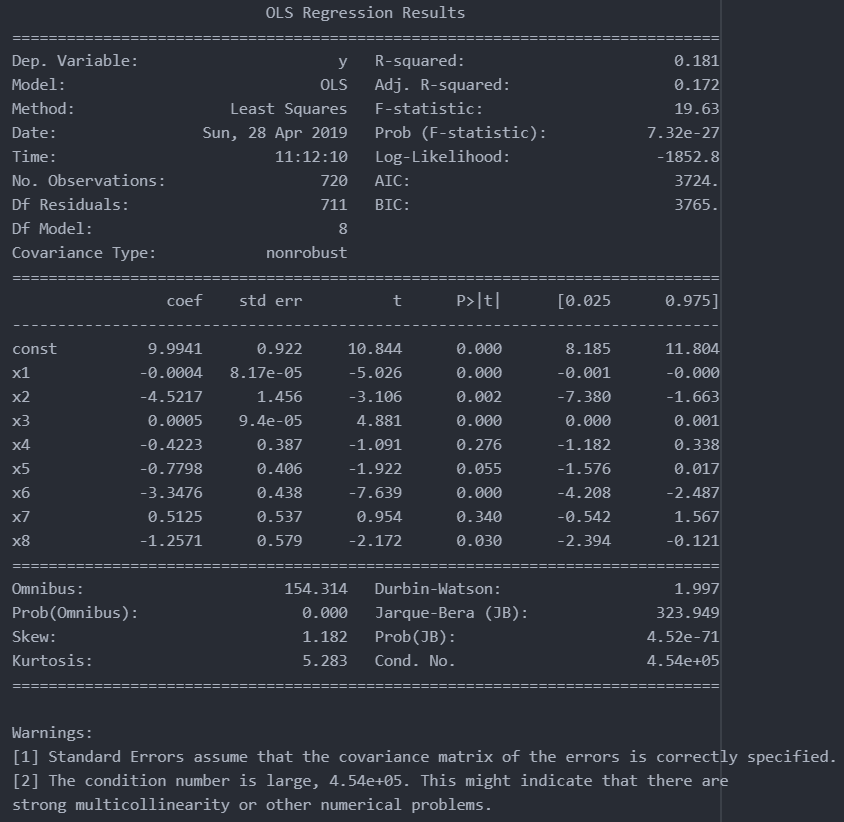
The coefficient of GNI in above image is showing zero because we have rounded up to four decimal places.

The R- squared value for the model is 0.1301 which means that model explains 13.01% of variance in dependent variable by the explanatory variable. Mean Absolute Error is 2.649, Mean Squared Error is 11.872 and Root mean Squared error is 3.446. The distribution plot of the residual is as follow:



For the next model, I have included interaction term of GNI and HDI to the first model. The third model in code is as follow:

The result from regression model is as follow:

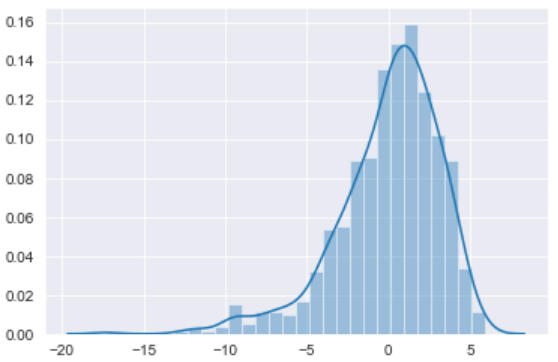


The model can be rewritten as:

*female=9.99-0.0004GNI-4.52HDI+0.0005HDI.GNI -0.422Africa-0.7778Europe-3.34NAmerica+0.51SAmerica-1.25Oceania*

The p-value for interaction term is 0.00, which means that interaction term of HDI and GNI are statistically significant in female suicide. So, it is better to include this term in model to make the study precise even the value of estimate is small. When every parameter of the model is constant and if HDI is increased by 0.1 in Asian continent, change in female suicide is (-)0.452 + 0.00005GNI. If we assume GNI of Asian country for year is zero then there will be decrease in female suicide by 0.452.

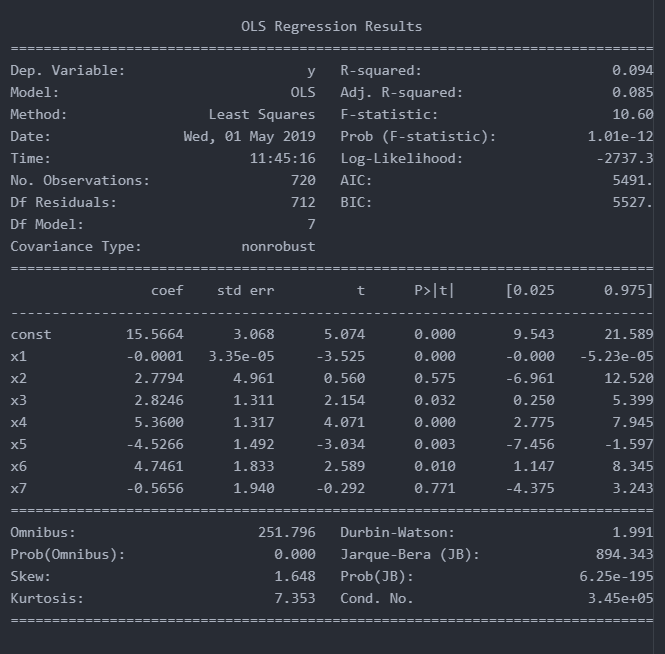
Here, the value of R-squared is 0.181which means that model explains 18.1 % of variance in dependent variable by the explanatory variable. The Mean absolute error is 2.39, Mean squared error is 10.06 and Root mean squared error is 3.17. The distribution diagram of residual is as follow:



After adding interaction term in female model, we can see that the R-squared value increased from 0.13 to 0.18 and the magnitude of error between observed and predicted value is also decreasing. So, this model is better for determining female suicide per 100k.

The second model is for male suicide per 100k

The result from regression model is as follow:

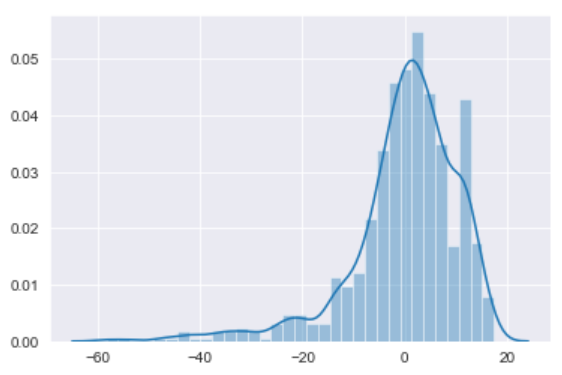


The model can be rewritten as:

*male=15.56-0.0001GNI+2.77HDI+2.82Africa+5.36Europe-4.52NAmerica + 4.74SAmerica - 0.56Oceania*

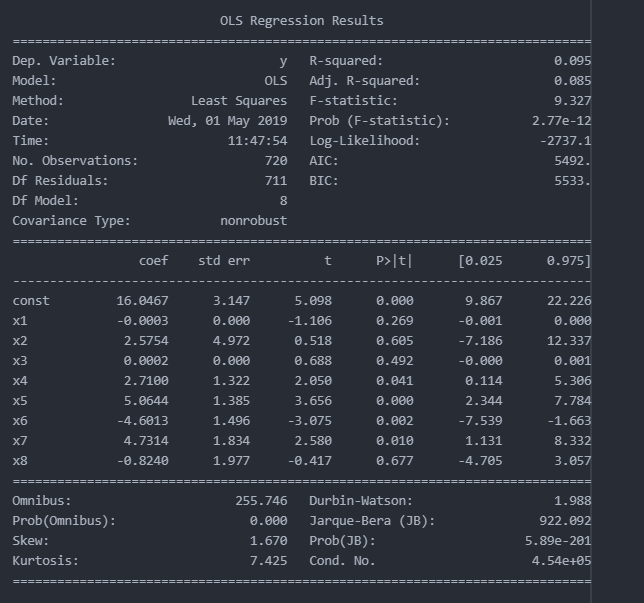
When every parameter of the model is constant and if HDI is increased by 0.1 in Asian continent, the suicide per 100k for male is increased by 0.277. Increase in will decrease the suicide of male but increase in HDI will increase the suicide number in male in Asian Country. If GNI and HDI of two country is same but one is Asian and another is Europe then Europe will have male suicide per 100k, 5.36 more than Asian country.

Here, the value of R-squared is 0.094 which means that model explains 9.4 % of variance in dependent variable by the explanatory variable. The Mean absolute error is 7.65, Mean squared error is 117.41 and Root mean squared error is 10.83. The distribution diagram of residual is as follow:



The fourth model I have included interaction term of GNI and HDI in third model and it is as follow

The result from regression model is as follow:

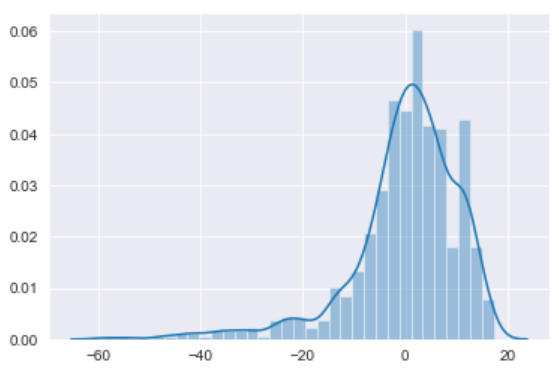


The model can be rewritten as:

*male=16.04-0.0003GNI+2.57HDI+0.0002HDI.GNI +2.71Africa+5.06Europe -4.60Namerica + 4.73Samerica - 0.82Oceania*

The p-value for interaction term is 0.492 and even at 10% significance level null hypothesis is rejected which means that interaction term of HDI and GNI are statistically insignificant in male suicide. When every parameter of the model is constant and if HDI is increased by 0.1 in Asian continent, change in male suicide is 0.257 + 0.00002GNI. If we assume GNI of Asian country for year is zero then there will be decrease in male suicide by 0.257.

Here, the value of R-squared is 0.095 which means that variable explains 9.5 % of variability in the model. The Mean absolute error is 7.63, Mean squared error is 117.34 and Root mean squared error is 10.83. The distribution diagram of residual is as follow:



After adding the interaction term in male model, R squared value increased by 0.001 and the interaction term is also statistically insignificant and the AIC value of model with interaction term is greater so, it is better not to include interaction term in male suicide model.

**Conclusion**

After studying both of the dataset, I can conclude that male suicide rate is higher than female suicide rate in all country. The difference between male and female suicide increased with the increasing age group. Surprisingly, suicide rate of age group above 75 was higher in most of all countries. Year individually has negative impact(decreasing) in suicide rate but when encounter with other factors it is difficult to say that suicide will be decreasing with the succeeding years. One of the surprising aspects in the result was the impact of GDP per capita. GDP per capita had positive impact (increasing)in suicide rate while GDP and Population of year had negative impact and GDP per capita is generally GDP divided by population. I think there are many other factors like education, individual income level, family environment, society condition and many more which has direct relation with suicide. GNI and HDI had negative impact in female suicide rate while their interaction term had positive impact but in male suicide HDI and interaction term had positive impact but GNI had negative impact. I cannot conclude with 100 % guarantee that higher HDI means lower suicide rate. Qatar HDI rank for 2017 is 37 while that of Norway is 1 but if we compare the suicide of male and female of these two countries, Norway has highest value for suicide than Qatar.

For linear regression we have the assumption of normality of residual and if the residual is normal then the assumptions are valid and model inference (confidence interval and model predictions) should be valid. The model which I have chosen best for these data set tend to have normal distribution but not exactly bell shaped because of some distortion. I believe that if I could include other factors which has impact in the suicide rate then I would have some better result then this but with this available data set only I cannot conclude that rich country and higher HDI rank country have less suicide and vice versa. I can also further used fixed effect model for the panel data and see whether any difference in result occurs. From all this study, there are only two things which can be concluded with accuracy i.e. male have higher suicide and with the increase in age group the suicide rate is increased. In other to conclude impact about HDI and GNI I think I should include other components in model.