An Analysis on Suicidal Tendencies Based on Socio-economic Information of Different Cohorts Gathered from 1985 to 2016

M.L.M Fernando
Department of Computer Science
University of Moratuwa
Sri Lanka
fernandomlm.20@uom.lk

T.M.T.A Gunarathna
Department of Computer Science
University of Moratuwa
Sri Lanka
gunarathnatmta.20@uom.lk

R.R.N.P.A.B.W.M.P.A Galagoda
Department of Computer Science
University of Moratuwa
Sri Lanka
galagodarrnpabwmpa.20@uom.lk

***Abstract*—Suicides are a major issue observed across the world. We conducted this analysis with the intention of finding any dependency of the social and economic status of a country with the suicidal tendencies. As a country the rate of suicides have a direct impact on it's economy and growth. We have extended our analysis to find patterns between the Gross Domestic Product (GDP) of the country and the suicidal rate, the age groups of the victims and the generation they belong to. Also at the later part of the report we have predicted the safe zone and danger zone countries to live based on their past records of suicides. We also predicted the suicidal trends for the upcoming years using time series forecasting.**

***Index Terms*—suicides, GDP, generation, population, country**

I. Introduction

Committing suicide is a fatal self-injurious act with some evidence of intent to die. It is considered as one of the major currently faced issue in all the countries across the world as it directly impacts the economy of a country and it's productivity. Suicides have become one of the leading causes of death recently. WHO[1] like organizations has identified the seriousness of this problem, and has called for an expansion of data collection on suicides to aid the planning of health strategies and health-care policies for the public.

In light of all these calls, in this paper we provide a analysis of the suicidal behavior with different social and economic factors. First, we provide an overview of the used dataset on the prevalence of suicidal behavior over the past years for a period of 32 years from 1985. In this section we observe the collected data and tune the data to be prepared for the analysis. Before the actual analysis begin the data was pre-processed.

Second, most part of the analysis is focused on the identifying trends between suicidal rates with country (e.g., the United States), subgroup (e.g., adolescents), generation (e.g., generation X), or gender of the victim. We review data from multiple countries, on all age groups, on both the genders and with the GDP of the particular country in the same year.

Third, we go over some statistical information of the data like mean, median and correlation of numeric fields in the collected dataset. Finally, we predict the safe zone countries and danger zone countries to live in based on their past records for three decades on the suicidal rates. We predicted the suicidal trends for the upcoming years using time series forecasting.

This compiled dataset is pulled from Kaggle[2].As per the description of the dataset, this has been generated by using four other datasets linked by time and place, and was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum[3].The dataset consists of 12 columns namely; country, year, sex, age group, count of suicides, population, suicide rate, country-year (composite key), HDI for year, GDP for year, GDP per capita and generation(based on age grouping average).

II. Pre-Processing

Dataset is pre-processed before it is used for analysis to identify anomalies and tune the dataset. This is done to improve the data quality. When we went through the dataset carefully, we observed some missing data fields for some entries, incomplete and missing records for some countries and years. The challenges we identified the regarding the dataset can be listed as follows,

- Identify inconsistencies in the dataset. Missing data in certain countries for a certain period of time.
- Remove duplicate or redundant data in the dataset.
- Manage resources for processing and manipulating the dataset.
- Fill the gaps in missing features of the selected dataset.

Following measures were taken to overcome the above listed challenges.

A. Data Validity

The data validity is the degree to which the data conform to defined business rules or constraints. Data type constraints are values in a particular column must be of a particular data type, e.g., boolean, numeric, date, etc. Before analysing the data the data-type constraints must be checked.

In our dataset as the Fig.1 shows the data type of the column "gdp_per_year ($)" is of type "object". But it is be a numeric

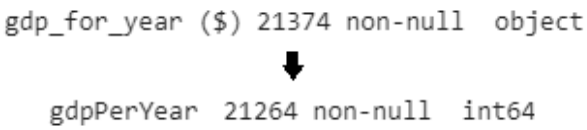

Fig. 1. Before and After data type conversion

value and is used in calculations in analysis tasks. This column was converted in to the type "int" before processing.

B. Data Completeness

Data completeness denotes the degree to which all required data are available in the dataset. First we checked for null values in the selected dataset. As in Fig.2, only the column 'HDI_for_year' has null values and the null values count is very high. For higher accuracy we did not consider 'HDI_for_year' column while performing analytics.

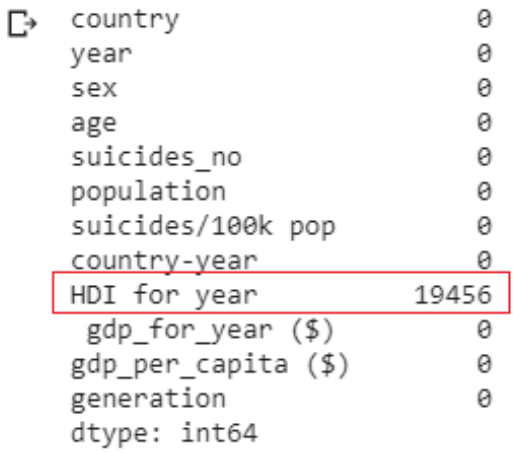

Fig. 2. Columns with null records

Next we calculated the percentage of missing data entries, year wise , country wise and gender wise. Incomplete data was found in, with respect to country wise and years wise only.

Second, we plotted a graph to check the incompleteness of the data country wise. The expected number of records per country is,

$$\text{expected\_record\_count\_per\_country} = \text{years} * \text{sex\_count} * \text{age\_count}$$

Where 'sex_count' is the number of gender types and 'age_count' is the number age groups which the data was collected. So the total number of expected records per country with respect to our dataset is 384. As depicted in Fig.3, the records count of certain countries are very low. This is due to missing records in the dataset.

We calculated the third quartile of the expected number of records per country, which is 288 and eliminated all the countries having a records count lesser than that. We achieved 75% completeness of the data(country wise) through this approach. After this step the number of countries remaining for analysis purposes is 62.

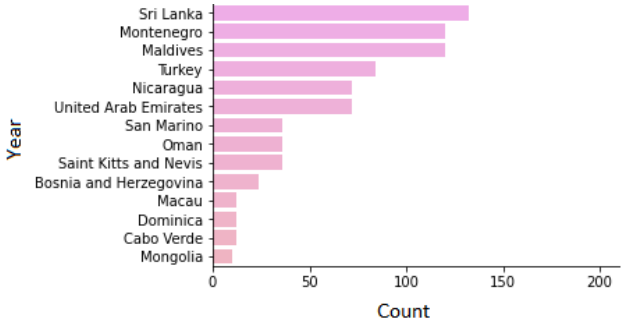

Fig. 3. Records count per country

Third, we plotted a graph to check the incompleteness of the data year wise. The expected number of records per year is,

$$\text{expected\_record\_count\_per\_year} = \text{countries} * \text{sex\_count} * \text{age\_count}$$

So the total number of expected records per country with respect to our dataset is 1212. As depicted in Fig.4, the records count of 2016 is significantly low. This is again due to missing records in the dataset.

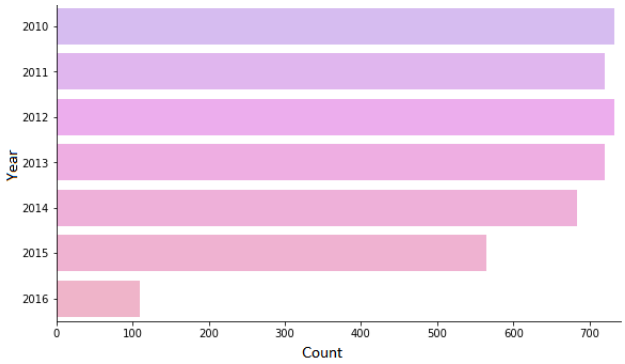

Fig. 4. Records count per year

So we dropped records belonging to 2016 from the dataset to be used for analysis.After this step the number of years remaining for analysis purposes is 31.

After pre-processing the data set, then we carried out analytics on the modified dataset.

III. Analysis

Data analysis is a process where you need to extract useful information by inspecting, transforming data for supporting decision making. In this data set, we were able to create graphs/visualizations, we were able to get more statistical figures on the data. by analyzing through descriptive, diagnostic and predictive.

A. Descriptive Analysis

By the end of the data manipulation phase, the data distribution of suicides has been spread over 1985 to 2015. In the Fig.5, the bar chart shows the total number of suicidal cases reported over the years in the selected countries. According

to the figure, the suicidal tendency has been grown to its maximum in the year 2012. Even though the suicidal tendency is fluctuating around 1985 to 1990 in a lower amount, it has been grown considerably from 1990 to 2015. The absence of reported suicidal cases in some countries could be coursed the distribution of the number of suicides from 1985 to 1990 to be lower to some extent.

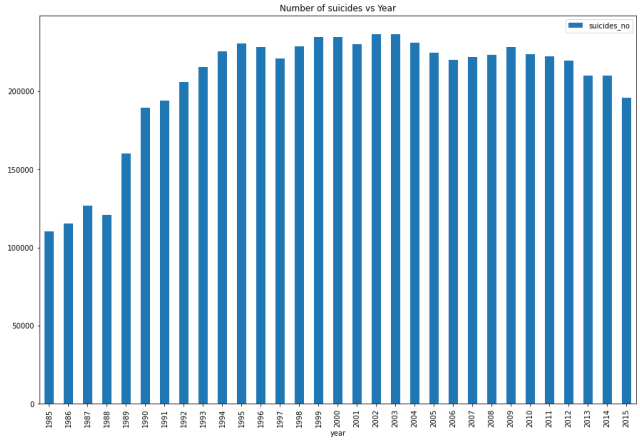

Fig. 5. Number of suicides vs year

The Fig.6, shows that the suicidal tendency according to the gender of the victims. It is critically biased to the male gender group when compared with the female gender group. The total number of suicidal cases of males reported is more than doubled by the total number of suicidal cases of females.

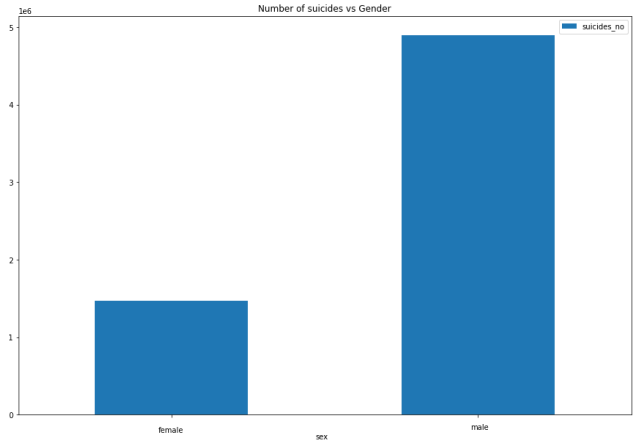

Fig. 6. Number of suicides vs gender

The Fig.7 and Fig.8, shows how the victims are spread over the age range and the number of suicidal cases reported. According to the horizontal bar chart, the age group, 35 - 54 years has affected majorly. The age group, 35 - 54 years can be mainly taken as the working force of a country with huge responsibilities to take care of their families. According to the pie chart, 36.4 percent of the total suicidal cases are affecting the countries' workforce.

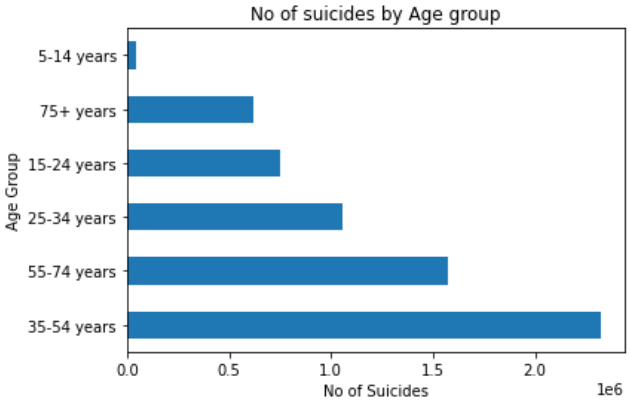

Fig. 7. Number of Suicides by Age Group Horizontal Bar

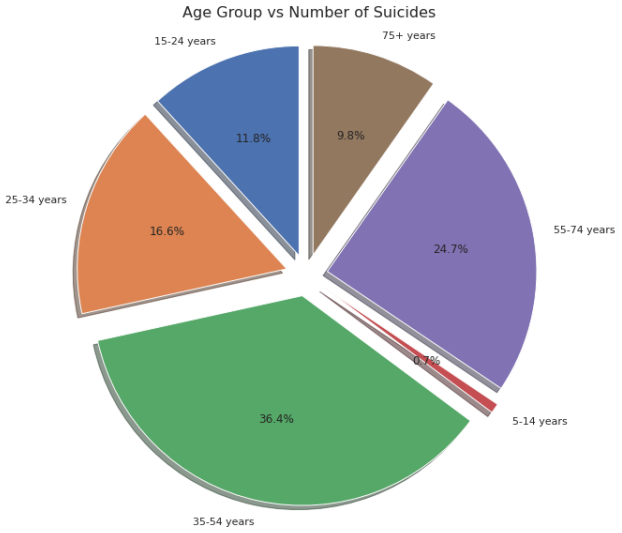

Fig. 8. Number of Suicides by Age Group Pie Chart

When taking the suicidal cases according to the gender, age group and the number of suicides, the male age group is more prominent over every age group that has been considered to be prone to be victims. The number of suicides of males is more than double when compared with the number of suicides of females in the same age group.

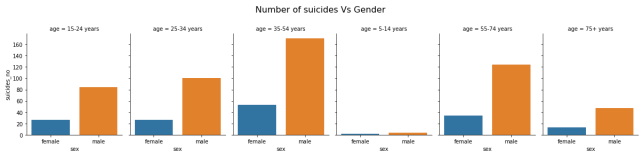

Fig. 9. Number of Suicides vs Gender

When we are considering the number of suicides over the years and the age groups, that shows the age group 35 - 54 years are keeping higher peeks comparing with other age groups. On the other hand, suicidal cases are highest in the middle of 2000 to 2005 according to the Fig.10 in the age

group 35 - 54 years. But after 2005, the suicidal cases are fluctuating slowly over about 10 years of time.

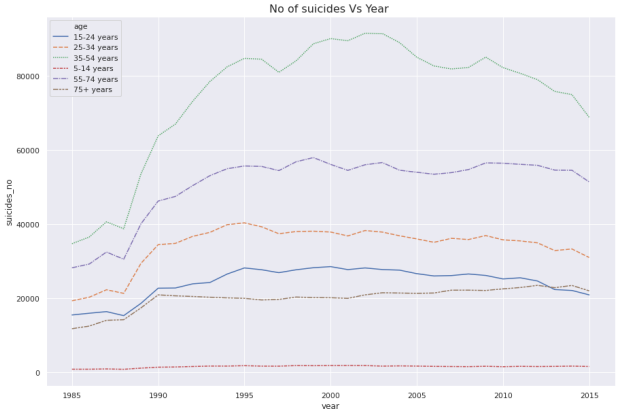

Fig. 10. Number of Suicides vs Year - Line Chart

The Pie chart Fig.11 shows the relationship of the results between the generation and the number of suicides. Mainly the generation is divided into 6 main parts;

- G.I. Generation - born between 1901 and 1927
- Silent - roughly from the mid-1920s to the mid-1940s
- Boomers - born between 1946 and 1964
- Generation X - early 1960s to late 1970s
- Millennials - reaching adulthood in early 21st century
- Generation Z – born 1995 to 2010

From the above-mentioned graph it is clear that the majority of the suicides belonged to the boomers. A nearly equal amount of generation X and silent generation tends to suicides. Only a small minority of Generation Z affected on suicide. According to this data it seems that more aged people tend to suicide.

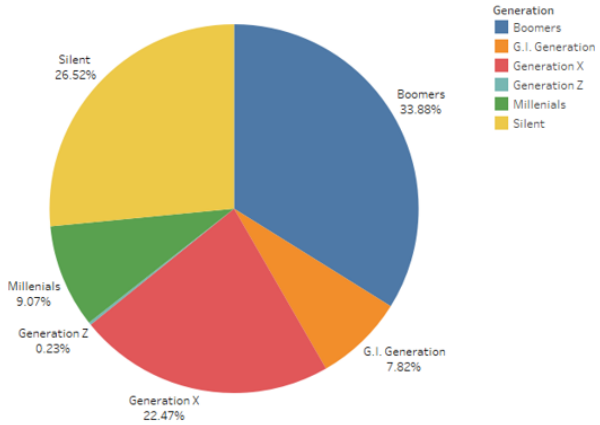

Fig. 11. Generations vs Number of Suicides

The Fig.12 shows the relationship between a numerical(suicidal Number) and categorical variable(Gender) according to the generation. As previously mentioned there is a higher tendency of generation boomers effect suiciding. As the graph depicts there is a significantly higher amount of males are suiciding than the females for each and every generation.

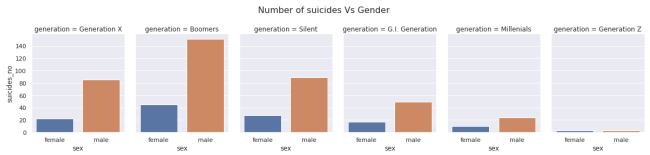

Fig. 12. Suicidal Rates vs Gender

Below Fig.13 shows how each generation's suicide tendency for each year. There is a sudden increment in Boomers suicidal rate during the 1990 to 1995 time period and in 2010 there is an increment in generation x as well. For the silent generation, there is a decreasing trend after 2010. While G.I generation was over in 2000, during this time Millennials affected in suiciding starting from the early 90's and Generation z in mid 20th century.

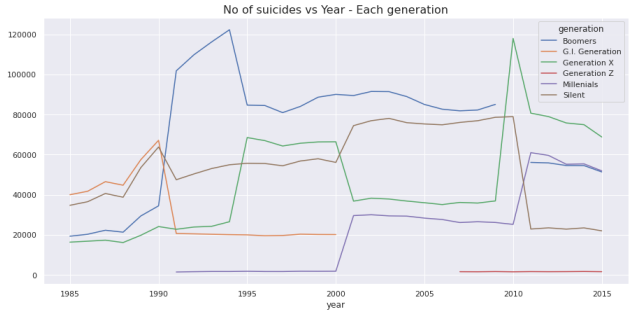

Fig. 13. Number of Suicides vs Each Generation

Heat map Fig.14, Fig.15, Fig.16 shows the distribution of suicidal rates across the world. Most suicidal rate in Russia Federation(11,305 suicidal/100k population).where united states and japan followed as 2nd and 3rd position. the rest of the countries shows comparatively less amount of suicidal rates compared to the top ones.Trinidad and Tobago shows least number of suicides.

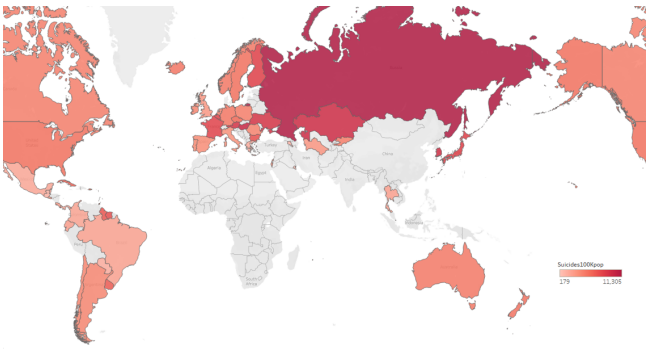

Fig. 14. Heat Map For Country vs Suicides/100k Population

B. Diagnostic Analysis

In this data analysis, we are focusing on observing the variables that could affect to increase the suicidal tendency

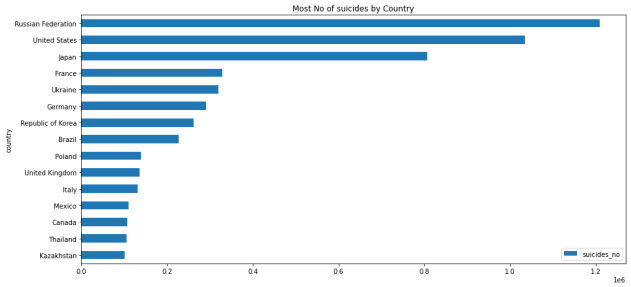

Fig. 15. Most Number of Suicides by Country

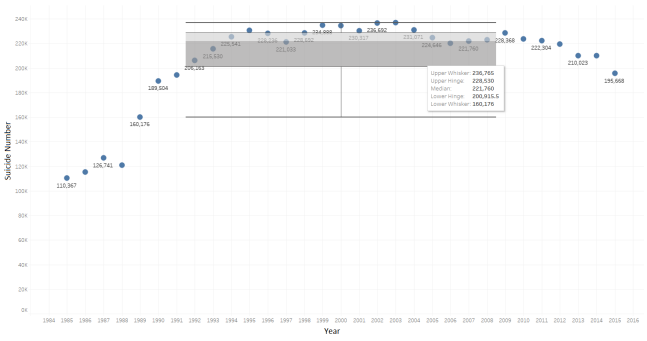

Fig. 16. Box Plot Year vs Number Suicides

by finding out the correlation in-between them. For that, we are observing the correlations of gdp_per_capita vs suicides_no, gdp_per_capita vs suicides100k and population vs suicides_no.

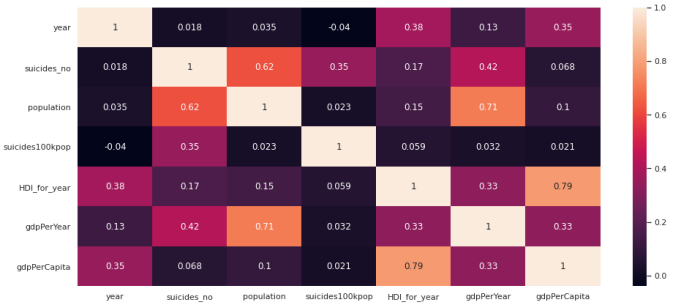

Fig. 17. Heat Map

In the Fig.17, the heat map shows the Pearsons correlation between the metric features in the dataset. According to that gdp_per_capita and suicides_no is having the Pearson correlation of 0.068 and that shows that according to the Pearsons correlation, those two features are more toward zero. So that gives us gdp_per_capita and suicides_no have no correlation. To elaborate on it further, the scatter plot in the Fig.18 has plotted with gdp_per_capita and suicides_no and that shows the data point scattering is clustered by the X-axis and the data points distribution is not linear with each other. So that we used Spearmans correlation and that gives the correlation of 0.149. When compared with the two methods that we have followed, Spearmans correlation gives a higher value for the correlation but that also suggests that gdp_per_capita has no relation upon suicides_no occurred.

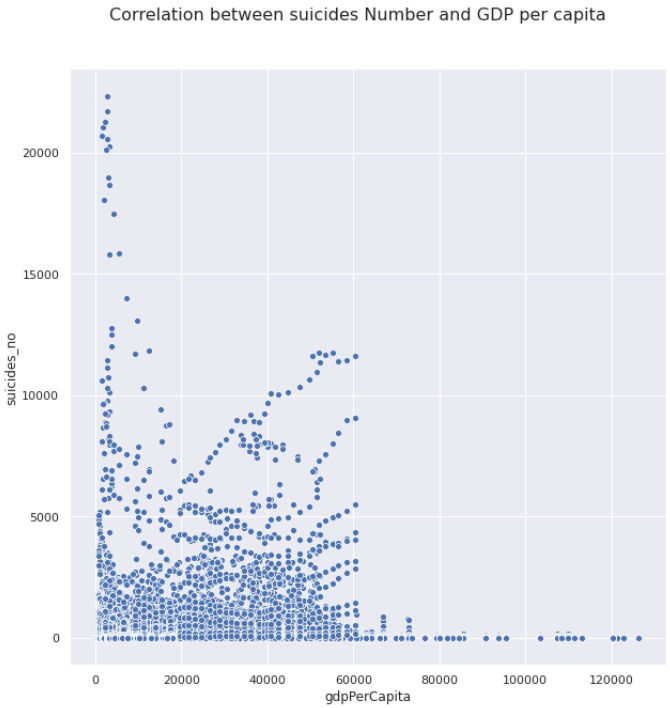

Fig. 18. Correlation between suicides Number and GDP per Capita

The next correlation analysis has been done gdp_per_capita and suicides100k. According to the figure(heat map figure number), the heat map shows the Pearsons correlation between those features as 0.021 and that more towards the zero. So that according to the correlation suggested by Pearsons method, suggest that those two features also have no correlation in between. To elaborate this further, the scatter plot in the Fig.19 has plotted with gdp_per_capita and suicides/100k and it shows that the data points are clustered by the X-axis and Y-axis. So then we used Spearman correlation and that gives the correlation of 0.088. When compared with the two methods that we have followed, Spearmans correlation gives a higher value for the correlation but that also suggests that gdp_per_capita has no relation upon suicides100k.

As the next correlation analysis has been done population and suicides_no. According to the figure(heat map figure number), the heat map shows the Pearsons correlation between those features as 0.62 and that more towards the one and that suggests it does have a positive correlation. To elaborate this further, the scatter plot in the Fig.20 has plotted with population and suicides_no and it shows that the data points are scattered in a linear distribution. For further, we used Spearmans correlation and that gives the correlation of 0.789. When compared with the two methods that we have followed, Spearmans correlation gives a higher value for the correlation and which also suggests that the population has a strong correlation upon suicides_no occurred.

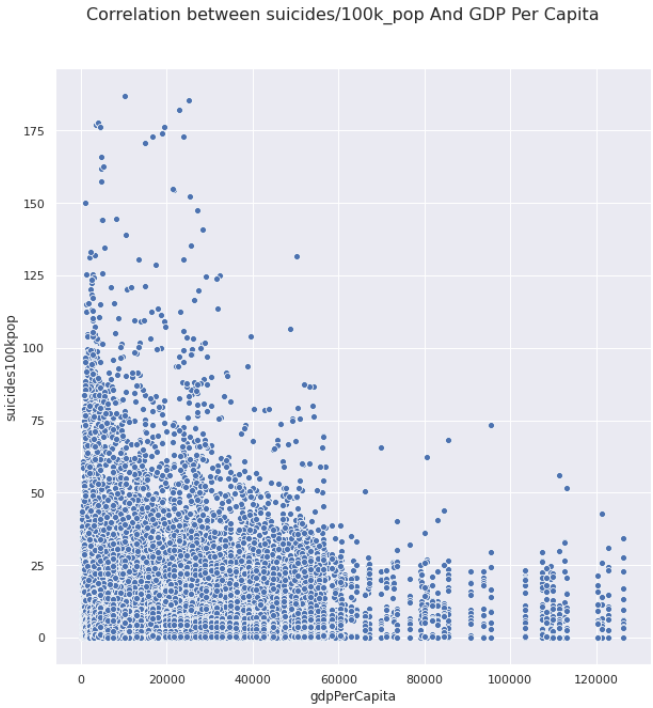

Fig. 19. Correlation between suicides/100k_pop And GDP Per Capita

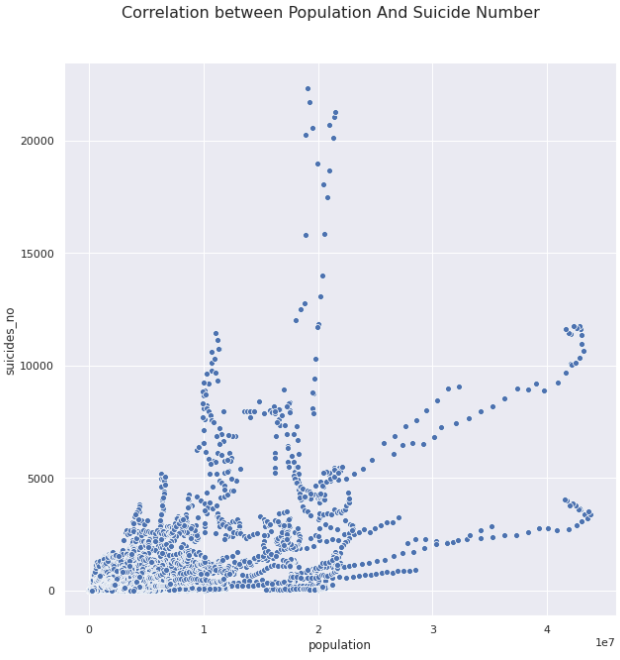

Fig. 20. Correlation between Population And Suicide Number

C. Predictive Analysis

Predictive analysis is the type of data analysis which makes predictions about the future based on historical data and predictive models. As the last part of our analysis we predict two sets of countries as Safe Zone countries and Danger Zone countries. Safe zone and danger zone countries are determined as follows. The countries with an increasing suicidal rate for three successive decades are grouped as danger zone countries. The safe zone countries are the countries having a decreasing rate of suicides over the last three decades. In this section we have considered three decades, 1987-1996, 1997-2006 and 2007-2015. The data was first modified to contain the respective decade for each entry in the dataset, and later grouped by the country and decade to take the summation of the suicide counts. With this information countries with increasing and decreasing suicidal rates were plotted as in the Fig.21 and Fig 22.

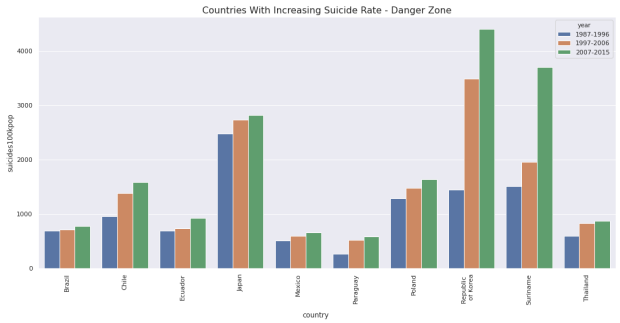

Fig. 21. Countries With Increasing Suicide Rates - Danger Zone

With this analysis we can predict which countries are susceptible to have higher rates of suicides and the respective governments/states can take steps to try to minimize it by root causing the problems. This information can be used while framing public health policies and health-care principles.

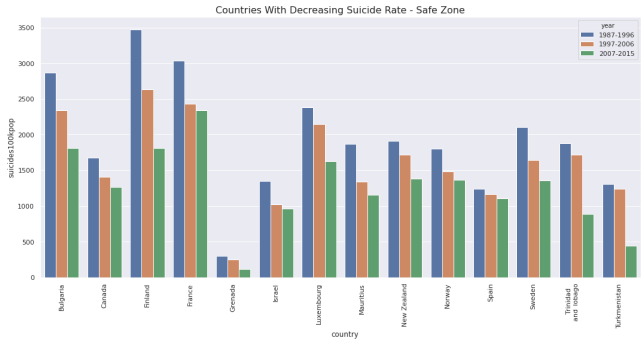

Fig. 22. Countries With Decreasing Suicide Rates - Safe Zone

For countries with a decreasing suicidal rate, we can look into what actions that have been taken by the governments/states in order to decrease the suicidal tendencies in those specific countries. This is an area where this analysis can be further expanded.

To predict the number of suicides in the coming years, we have used a Auto Regressive Integrated Moving Average (ARIMA). Here we assumed that the time series in stationary and the suicidal rates in the past years can alone be used to predict the future suicidal rates because the correlation we expected to prove between socio-economic status and the suicidal rate was not evident from the analysis.The mean squared error of the predictions was 388684.863. The predicted curve and the expected curve for the suicidal trend with time is shown in the Fig.23.

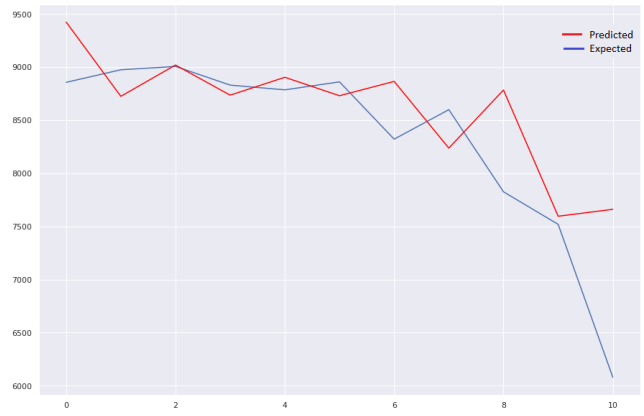

Fig. 23. The Predicted and the Expected Suicidal Rate

Conclusion

In our analysis, the main objective was to observe a trend between socio-economic status and the suicidal tendencies of different cohorts. For this purpose first, we followed preprocessing of the data to improve the data quality. Then we performed analysis under three main fields namely, descriptive, diagnostic and predictive analysis according to Gartner's Maturity model.

In the descriptive analysis, we used visualizations to understand the dataset and observe any visually clear trends between fields such as age group of victims, generation of victims and the GDP of respective countries and years. Under this, we calculated median, mean and quartiles of numeric data fields to find the distributions of data.

We observed most of the suicides were prevalent among males compared to females. Also, most of the victims were from the age group 35-54 years. The Boomers seemed to have the highest suicidal rate among generations.

Secondly in diagnostic analysis, we used Pearson correlation[4] and Spearman correlation[5] to find out the relationship between selected features like number of suicides vs GDP per capita, number of suicides/100K population vs GDP per capita and number of suicides vs population.

In this analysis, we observed there is no clear relationship between the suicidal rate and the GDP per capita. The calculated correlation between the two features was of type neutral. The number of Suicides vs population proved to have a positive correlation.

In the last part of our analysis, we carried out a predictive analysis to predict safe zone countries/danger zone countries and built a model to forecast the suicidal rates using ARIMA.

These predictions can be used when framing public health policies and health-care principles.

References

[1]"Home", Who.int, 2020.[Online].Available: https://www.who.int/.[Accessed: 05-Apr-2020].

[2]"Kaggle: Your Machine Learning and Data Science Community", Kaggle.com, 2020.[Online].Available: https://www.kaggle.com/.[Accessed: 10-Apr-2020].

[3]"Suicide Rates Overview 1985 to 2016", Kaggle.com, 2020.[Online].Available: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016.[Accessed: 15-Apr-2020].

[4]"Data Analysis-Pearson's Correlation Coefficient",Learntech.uwe.ac.uk,2020.[Online].Available: http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442. [Accessed:05-Apr-2020].

[5]"Spearman's Rank-Order Correlation - A guide to when to use it, what it does and what the assumptions are.", Statistics.laerd.com, 2020.[Online].Available: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php. [Accessed:05-Apr-2020].