Study of Global Suicide Rates (2005-2015)

Akshara Kannan (2045986)
University of Warwick
MSc Computer Science, 2020.
January 5, 2021

Abstract— **Suicide is one of the leading causes of premature death, that we hear about in the news everyday- from farmers in developing countries, to celebrities in developed countries. While loss of life of any age-group has a gigantic impact on the economic development of a country in terms of loss of human resources, studies show that the low-income countries possess a risk of higher suicide rates than their high-income counterparts. Moreover, religiously and politically, suicide is still a hugely stigmatised and worse, criminalised act. In order to prevent such a catastrophic event in the lives of large number of people, organisations and governments are urgently required to collaborate and work towards suicide prevention. This can be achieved by raising awareness, and also scientifically studying the factors that cause suicidal events and the far reaching effects of nationality, race, culture and economic and social welfare on it. The goal of this report is to study the causal relationship between the suicide rates of a country with respect to its economic standpoint. Machine learning and visualisation techniques are employed to study this relationship on the Weka Tool, considering the effect of age and sex of individual as well. The report also supplements these findings with a brief discussion on the challenges posed to data analysis of suicide-related data, and the different analysis paths that can be taken within the purview of suicide-related data and suicide prevention.**

Index Terms—**Suicide, Economy, Python, Weka, Classification, Regression, Clustering, Data Visualisation**

I. INTRODUCTION

According to the World Health Organisation(WHO), approximately 800,000 people die by suicide, globally every year. It has been established that suicide, if not a mental disorder in itself, is heavily interlinked with extreme severity of mental stress among people. [1]

Suicide Prevention has always been one of the forerunners of the fight against mental health. Mental Health activists are moving away from strategies to create awareness only within the community, in an attempt to eradicate the stigma surrounding mental health and/or suicide; and are now working in tandem with governments and policy makers to bring in effective policy changes to save people's lives. This is due to the rise in statistics showing an increased correlation of **suicide rates to the social and economic disposition of countries**. With each day, more organisations are researching to understand the various kinds of factors that exacerbate the suicide rates, to recommend to local governments and international health bodies such as WHO- a framework to set up a multi-thronged suicide prevention support system suitable for the behaviour of each country in this aspect. This involves attenuating the hardships in lives of people because of social issues such as lack of freedom and hate crimes, or economic issues such as drought-wrought or worst-hit pandemic countries.

To make matters worse, according to researchers from International Association for Suicide Prevention(IASP), [2] suicide is considered as an illegal act in 22 countries, with more countries banning suicidal behaviour quoting the Sharia Law. Such criminalisation of suicide directly sabotages and threatens to compromise the hard work performed by grassroot level workers to arrest the growth of suicide rates, while also impairing the vulnerable individuals from accessing the suicide helplines or other mental health services.

Owing to these challenges, it is imperative and utterly urgent to contribute to making suicidal behaviour as a public health issue. With the impending sense of doom during the COVID-19 pandemic, there has been a dip in economic wellfare of people, causing a boom in mental health issues as well. Now more than ever, there is a need to make suicidal behaviour an urgent public health issue, and work towards decriminalisation of suicide.

Looking at this in an academic point of view, it is not surprising that research in this area necessitates the availability of large data of good quality, which considering the illegality and stigma shrouding suicidal behaviour, is a monumental task at hand.

This paper seeks to demonstrate the relationships between suicide rates and age, sex and the country they belong to, and hence attempt to debunk myths and previously held beliefs on suicides.

II. BACKGROUND

Suicide refers to the act of ending one's own life, and suicidal tendencies refer to talking about or taking actions related to taking their own life. People who suffer from severe mental health illnesses sometimes believe that taking away one's life serves as the only escape from reality. [3] Suicide inflicts immense misery on families, communities and entire countries as well. Throughout the years, the deaths of various celebrities and brilliant academicians was what it took for toxic and overworked environments in every industry to be brought to light and was the much needed wake-up call for the need for mental health services to be provided at every institution/organisation. [4]

A. *Causes of Suicides*

Originally, the cause of suicides was attributed to severe mental disorders, especially depression, alcohol and drug abuse among people in high-income countries who have access to higher education, high-end jobs etc. [5] But, in reality, a great deal of suicide deaths are observed in low-income countries, as an act of spontaneity, when people (with no history of mental illness) face an extreme breakdown due to financial problems, societal humiliation or family issues. More importantly, the rates are higher among vulnerable groups such as refugees and migrants; indigenous peoples; lesbian, gay, bisexual, transgender, intersex (LGBTI) persons; and prisoners who meet with discriminatory hate crimes in their day-to-day lives.

Such findings led to the shift in focus of suicide prevention organisations to inculcate the nature of countries and their economic and social circumstances into study to provide better care.

B. Misconceptions about Suicide and its Criminalisation

One of the biggest misconceptions about suicide is that, "talking about suicide to create awareness may encourage impressionable people to think about it and act towards it." This school of thinking has led to many countries criminalising suicides, based on the strong belief that if suicide was a crime, it would deter people from actually committing it. Inadvertently, this hampers those in need from asking for help, while punishing and penalising suicide-attempt victims, and families who have lost a loved one to suicide. In contrast, there is no proof that suicide rates are any lesser in countries which have criminalised suicide. [2] [6]

C. Towards Suicide Prevention

Globally, on the 10th of September,[7] people celebrate the World Suicide Prevention Day (WSPD) to highlight that that suicide is a vastly preventable cause of premature death. During the Suicide prevention Week, organisations from different parts of the world collaborate to analyse and study the various factors causing suicides, and ways to mitigate them and address victims and survivors.

Some examples are :-

- International Association of Suicide Prevention
- Samaritans [8]
- Mental Health Foundation [9]
- Suicide Prevention Lifeline [10]
- BC Partners for Mental Health and Substance Use [11]
- ... and many more.

III. DATA PRE-PROCESSING

A. Data Source

The dataset used for this project was chosen from Kaggle Datasets. **Suicide Rates Overview 1985 to 2016** [12] provides the rate of suicides by country, sex and age-group while also specifying the socio-economic state of that country through attributes such as GDP and HDI, as elaborated below.

The original dataset consists of 12 attributes:

- **Country** is a list of countries in alphabetical order, and **Population** shows the total population of the country in any given year.
- textbfYear depicts the year in $YYYY$ format, and contains values from 1985 to 2016.
- **Sex** and **Age Group** specify sex and age-group of suicide victims. Another attribute called **Generation** is also provided.
- While **Number of suicides** specifies the total count of suicide victims in a particular country, belonging to a sex and falling under a particular age group, the normalised value of **Suicide per 100k population** is also provided as an attribute.
- **Country-year** depicts a grouping of country and the year.
- Finally, the last three attributes include **HDI**, **GDP** and **GDP per capita**. HDI and GDP stand for Human Development Index and Gross Domestic Product respectively. They reflect the social and economic state of the country in that year.

Attributes and their type

Name of Attribute	Type of Attribute
Country	Nominal
Year	Nominal
Sex	Nominal
Age Group	Nominal
Number of Suicides	Numeric
Population	Numeric
Suicides per 100k	Numeric
Country-year	Nominal
HDI	Numeric
GDP	Numeric
GDP per capita	Numeric
Generation	Nominal

B. Need for Pre-Processing

Data Pre-processing as the name suggests, refers to a Data Analysis Technique that is performed prior to the said analysis to ensure that the data extracted is of the right format and value, suitable for obtaining accurate results.

The chosen dataset as specified in [12] was compiled from different datasets and was found to contain some inconsistencies and missing data. Two data pre-processing techniques were performed- **Data Cleaning** and **Dimensional Reduction**, which are elaborated in the sections III-D and III-E.

C. Platforms used Data Preprocessing

The following different data exploratory and programming platforms were utilised in this Project.

1) Microsoft Excel: Kaggle hosts most datasets as CSV Files- Comma Separated Values files. Since our dataset was downloaded as a *.csv* file, there was an increased tendency to use Microsoft Excel as the primary data viewing software.

2) Jupyter Notebook: Python provides various functionalities through its famed Numpy and Pandas frameworks [13] to perform a myriad techniques of data analytics. Specifically, Pandas is a framework that is much used by students and researchers alike, to perform data preprocessing tasks. Dataframe of Pandas, is a tabular structure which is flexible in its size and data-typing capability, that stores a relational dataset such as those hosted on $.csv$ files with rows and columns. In-built functions enable us to read data stored on $.csv$ files into dataframes, manipulate the data read into dataframes, and then finally writing the cleaned data back to a $.csv$ file.

3) Weka Tool: For exploratory purposes, the PreProccess Tab provided by Weka Tool[14] was also used. This tab is the opening page of the Weka Explorer Tool, into which the dataset file is uploaded. Once uploaded the user obtains an overview of the dataset with details such as number of instances, number of attributes, mean and standard deviation in numeric attributes, and the labels and their count in each categoric attribute.

D. Dimensionality Reduction

Dimensionality Reduction refers to the transformation performed on data such that the transformed data still exhibits the behaviour of original data, but with reduced number of columns, thus eliminating redundancy. Attributes such as GDP, and Country-year were dropped from the dataset because of the clear redundancy in their values (country and year are individual attributes in the dataset, and GDP per capita provides a better intuitive understanding in our context as it is defined as country's economic output per person).

E. Data Cleaning

Completeness is an important criterion that has to be satisfied for data to be of good quality. It refers to the extent to which the dataset contains all data to be known to perform analysis. In effect, to ensure completeness, the dataset has to be checked for missing data. Once such instances are found, they have to either be removed from the dataset, or be filled with a value appropriate to that attribute.

In our dataset, am attribute that contained a lot of missing values was **HDI**. A majority of the instances did not have a value of HDI for the years specified in the dataset. The HDI value is a characteristic measure of every country, and hence the missing values cannot be filled with an aggregate measure such as mean or median of other non-null value. Therefore, the HDI attribute was removed from the dataset. In Python, the instances with missing values are found by using is $isna()$ function, and attributes are removed using the $drop()$ function.

Another reason to perform data cleaning is to maintain the balance and uniformity of the data. The **Age-group**, which was nominal in nature, contained the word *years* in every value, which was trivial to its purpose. This word was removed from the values of that attribute.

Further, consider the **Year** attribute, ranging from 1985 to 2016, thus owing to the sheer magnitude of the dataset. Therefore, for the sake of the this project, the instances with year less than 2005 was excluded from the dataset. On visualising this attribute on Weka's Preprocess tab, it was observed that the number of suicides recorded for 2016 were very less. To ensure that a balance of distribution of values in each label, the instances of 2016 were also removed from the dataset.

Lastly, the values of **Generation** attribute were modified slightly to remove white spaces.

F. Converting to Weka Data file

Weka accepts data files saved using the *.arff* file extension. The format of the file consists of three sections- $@relation$, $@attribute$ and $@data$. The file starts with a one line definition of the name of the relation using the $@relation$ command. The attributes of the dataset are defined with the type and the labels suitably in the $@attribute$ section, and the data is pasted below that in the $@data$ section.

Now, the data is ready to be studied and analysed.

IV. EXPLORATORY DATA ANALYSIS

A. Exploratory Data analysis

Our cleaned dataset now consists of 9 attributes, **5 nominal** attributes **Country, Year, Sex, Age and Generation** and **4 numeric** attributes **NoOfSuicides, Population, Suicides/100k, and PerCapitaGDP**. At the end of the pre-processing of data, we have 10848 instances of the suicide data of 96 countries corresponding to 11 years- 2005 to 2015 with respect to all age groups and both sexes. On an average, globally we have 233 suicides per country, in a particular age group, belonging to one sex, in a year.

First of all, when we study the population and number of suicides visualisations, we can see clearly that majority of the instances belong to the lower most spectrum of the these attributes. In other words, majority of countries for most years have recorded a zero-suicidal death rate. While this indeed seems like something to be rejoiced about, we need to also study the state of these countries. Most countries which are in deep economic stress and is undergoing political turmoil, cannot and does not release trustworthy data regarding its citizens. [2] Thus, those instances which show a very less suicide rate, are of great importance in this context, and have thus not been removed from the dataset.

B. Sex of Victims

It is to be noted that the dataset contains equal number of instances for females and males respectively. From the legend of Sex attribute in the Preprocess Tab of Weka, blue corresponds to Male and Red corresponds to Female(Fig. 1), In each year and in each age-group, approximately equal number of males and females are recorded in our dataset. (Refer to Fig. 3 and 2)

The dataset is structured in a way that both male and female counts of suicides are recorded for all countries (for most years). it is essential to study the relationship between Suicides/100k and Sex due to understand the gender disparity better. As in seen in the visualisation in Figure (4), where X axis is the Sex attribute and Y axis is the SuicidesPer100k, **we can see the more blue points(male) in the higher reaches of the Y axis, than the red points(female)**. In other words, suicide rates are seen to be higher for males than females.

Fig. 1. Legend of Male and Female

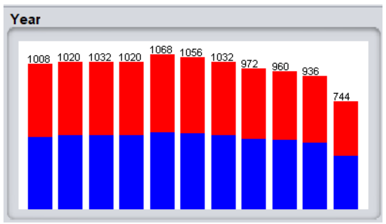

Fig. 2. Year vs Sex

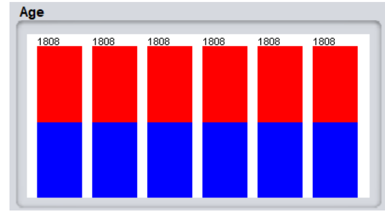

Fig. 3. Age vs Sex

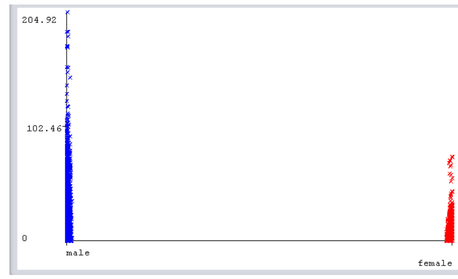

Fig. 4. Relationship between Sex and SuicidesPer100k

It was predominantly established that males are more vulnerable to premature death by committing suicide, than women.[17]

- One reason for this contrast could be under-reporting of female suicides due to domestic abuse by their family members, and other social issues.
- Transpeople are one of the most vulnerable groups to commit suicide(due to the mental stress caused by their gender dysphoria, hormone replacement therapy, hate crimes, unemployment, non-existent health care etc). The victims could be mis-gendered [15] when their suicides are registered, causing disparities in the counts of male and females. This is clearly reflected in the way datasets released by organisations do not inherently contain any information regarding this aspect.

C. *Generational trends*

In the research community for studying suicide rates, one of the most happening conversations revolve around the apparent link between the generations of people who are showing higher signs of suicides. For a layman, it may seem like that there has been a rise in suicides among the GenZ generation- the younger people; the section of the society that has been experiencing toxic social networking environments and entertainment media. Nonetheless, research has shown that the Millennial generation has grabbed the top spot in terms of suicide rates, when compared to previous generations. Experts bring to the foreground the difference between the way GenZ and Millennials interact with social networking sites such as Facebook, highlighting that since the latter were the first generation to be using such platforms, a greater majority of them have developed a unhealthier relationship with comparing themselves to their peers on the Internet. In addition to this, they have faced devastating events such as the 9/11 terrorist attack in the Us, a loss of employment as well as fighting wars. [16]

In accordance to this trend, our dataset contains the largest count for zero suicides in GenZ, while the graph shows higher suicide rate for the Millenials and the Silent generation, as seen in the image below.

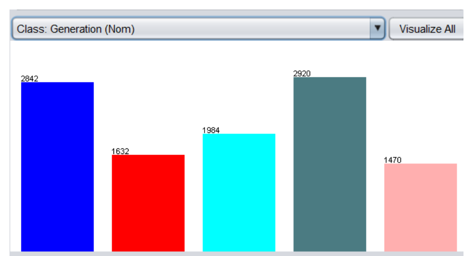

Fig. 5. Legend of Generations

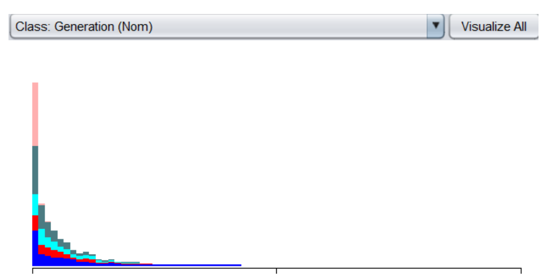

Fig. 6. Generations vs SuicidePer100k

V. RELATIONSHIP OF SUICIDE RATE WITH OTHER PARAMETERS

A. *Predicting Number of Suicides*

Predicting the number of suicides that could occur in a year or country is one of the leading tasks that we can perform with this dataset. Such a prediction would help the suicide prevention organisations to help countries with higher predicted rates better. A 10 fold cross validation evaluation mode was adopted for Regression and Classification of this dataset.

1) Regression: To regress the **NoOfSuicides** attribute, linear and polynomial techniques were considered. But, suitable results were obtained only for Linear Regression, as elaborated below.

- Firstly, **NoOfSuicides** was individually predicted as a function of Population, and Generation. But, the results obtained were unimpressive with low correlation coefficients.
- Then, attributes such as **Age** and **Sex** were included into the attribute list. In this case, a mediocre value of 0.5454 was obtained as the correlation coefficient.
- In an attempt to improve the efficiency of the model, factors about the country were also added such as PerCapitaGDP. In this scenario, a correlation coefficient of 0.7293 was obtained, highest value yet.
- Finally the attributes describing the country name and year of suicide were incorporated into the model. This modification did not improve the model by even a small margin, rather produced a greater relative absolute error by 0.1%

Observations :

1) Surprisingly, for Regressing NoOfSuicide, attributes **Population** and **PerCapitaGDP** were not useful for predicting the value of suicide count, with a value of 0.6741 as Correlation Coefficient. In other words, the linear regression equation as a function of these attributes gave trivial values for the slope. Therefore, it is necessary to cluster the suicides with respect to the these factors to identify a relationship.
$NoOfSuicides = 0.0001 * Population +$
$0.001 * PerCapitaGDP - 50.8623$
2) Sex=male fetched a high value of slope in the linear regression equation equalling 13.7023. This can be attributed to the high amount of male suicide victims in the dataset, as evidenced in Section III.

B. Studying relationship between Generation and Age

Included in this dataset are attributes that describe the general mindset of the victims at the time of their suicide- the generation to which they belong. Given that suicides occur due to paramount emotional turmoil in people, from the life experiences building up to that point, the differences in the way present and past generations perceive difficulties in life, contribute to similar trends in suicide rates with respect to this attribute.

On that premise, we perform classification of the Generation attribute with respect attributes using the K-Nearest Neighbour Classification algorithm.

Let us go through the steps used to attain the maximum possible accuracy for classifying Generation attribute. Classifying **Generation** only based on the **SuicidesPer100k** attribute gave very poor results. When **Year** was used, the model was found to be overfitting due to the direct correlation. Therefore, Year was not considered.

1) Since Age is an important characteristic that determines the generation a person falls in, in any year, this attribute was considered next. The best correctly classified instances was 74.3086 for a k value of 8. On studying the confusion matrix, it was understood that the classification into the *Boomer* label is poor, as its true positive rate was only 0.20.
2) When **Sex** was included, no increase in percentage of correctly classified accuracies was recorded.
3) Consequently, **Country** attribute was take into account for classification, because the distribution of generations differs among countries. For a k value of 1, the maximum accuracy was 74.7603 was obtained. (Table A)
4) On the other hand, when **PerCapitaGDP** attribute was added, the maximum accuracy obtained was 76.4288%. (Table B)
5) With a difference in accuracy of 1.6%, there is no remarkable difference in the values of true and false positives in the confusion matrix.
6) Finally, a model was built bringing together all these above attributes- **PerCapitaGDP, Country, Age and SuicidesPer100k**. The accuracy for this model was 86.3293, the highest obtained till now. (Table C)
7) The distance measure considered until now is Euclidean Distance Measure (Refer Eq. 1). The model can be built using the Manhattan Distance Measure (Refer Eq. 2). The accuracy for same k value, and the same set of independent variables is 86.8455%, an increase from its Euclidean counterpart by 0.5%.

For a 2-dimensional Euclidean space, and for two points (x_1, y_1) and (x_2, y_2),

$$EuclideanDistance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad (1)$$

$$ManhattanDistance = |(x_1 - x_2)| + |(y_1 - y_2)| \qquad (2)$$

Table A - Confusion Matrix :-

a	b	c	d	e	<– Classified as
2469	373	0	0	0	a = Silent
423	839	370	0	0	b = Boomers
0	369	1256	359	0	c = GenX
0	0	399	2247	274	d = Millenials
0	0	0	171	1299	e = GenZ

Table B - Confusion Matrix :-

a	b	c	d	e	<– Classified as
2477	365	0	0	0	a = Silent
362	894	376	0	0	b = Boomers
0	369	1224	391	0	c = GenX
0	0	375	2388	157	d = Millenials
0	0	0	162	1308	e = GenZ

Table C - Confusion Matrix :-

a	b	c	d	e	<– Classified as
2620	222	0	0	0	a = Silent
222	1145	265	0	0	b = Boomers
2	249	1500	233	0	c = GenX
0	0	233	2667	30	d = Millenials
0	0	0	37	1433	e = GenZ

Thus, the K Nearest Neighbour Classification algorithm was executed on this set of attributes [PerCapitaGDP, Country, Age,SuicidesPer100k] for classification of **Generation** attribute, resulting in an optimal accuracy of 86.3293%.

C. Use of Clustering Techniques

Clustering is an unsupervised learning technique, that can be used to find clusters of similar points in a dataset, and analysing the inter- and intra- cluster patterns.

In this section we perform clustering on the dataset, by utilising the **Generation Attribute**, with "classes to clusters" as evaluation mode.

1) When Simple K-Means algorithm was executed with $k = 5$, the highest accuracy for clustering was obtained when the attributes used were **Year** and **Age** apart from **Suicideper100k**. The incorrectly classified instances were 37% of the dataset, thus correctly classified instances were 63%.
2) When the Farthest First clustering algorithm was performed, with $k = 5$, the incorrectly classified accuracy obtained for the above attributes was 47.7231%

We can see that Classification performs with better accuracy than Clustering for Generation attribute.

Finally, in this section we bring the spotlight back to the study of suicide rates with respect to the country's economical state of affairs. While we were not able to predict the suicide rate exclusively in terms of population and percapitaGDP, we perform clustering on the data against these attributes, and then examine the cluster assignments for trends.

- **Attributes Considered**: SuicidePer100k, Population, PerCapitaGDP
- **Clustering Mechanism**: Simple K Means
- **Number of clusters Considered**: 3

According to the **Department of Economic and Social Affairs of the United Nations Secretariat (UN/DESA)**[17], countries of the world are classified into three main categories, and hence we choose the number of clusters to perform clustering to be 3.

Categories of countries according to their economic situation:-

- Developed Economies
- Economies in Transition
- Developing Economies

.

*1) **Population vs Per Capita GDP**:* To visualise the cluster assignments, let's first choose the X-Axis as Population attribute, and Y-Axis as the PerCapitaGDP attribute. Consider the graph given Figure (7), the following important observations were drawn.

- The three clusters denoted by Blue, Red and Green correspond to the clusters localised to the lower, middle and the higher portions of the Y Axis- which is Per capita GDP value. These three clusters are analogous to the three categories of economies that countries are grouped into.
- The X- Axis (Population) ranges from 908 to 43,805,214- sparsely to densely populated. There are more points in the lower-right quadrant of the graph which corresponds to high Population and low per capita GDP.
- Similarly, the left top quadrant of the graph depicts those points belonging to countries that have a low population and high per capita GDP.

This shows that countries grouped into the three categories stated above have a negative correlation with Population, high Population corresponds to low PerCapitaGDP and low population corresponds to high PerCapitaGDP.

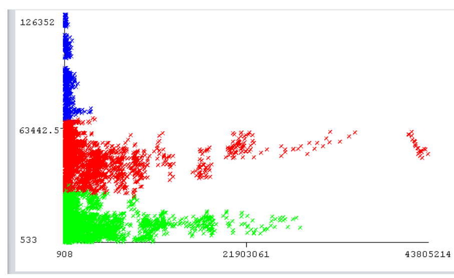

Fig. 7. Population VS Per Capita GDP

*2) **SuicidePer100k Vs PerCapitaGDP**:* Moving on, let us study the relationship between Suicide rate per 100k population and the Per Capita GDP of the country. X axis is chosen to be SuicidePer100k and Y axis as PerCapitaGDP. The visualisation of the clusters are displayed in Figure (8). From the figure, we can see that:

- The data is divided into three clusters, depicted by Green, Red and Blue.
- The green cluster represents those instances which have a low per capita GDP; this is the cluster that has values of high suicideper100k. Points belonging to this cluster have high rates of suicide.
- The other two clusters have progressively lesser suicide rates when compared to the green cluster.

In other words, countries with low PerCapitaGDP exhibit higher rates of suicide (Green cluster) than countries that boast a high PerCapitaGDP. The results from executing the K-Means clustering algorithm on our dataset proves our hypothesis that

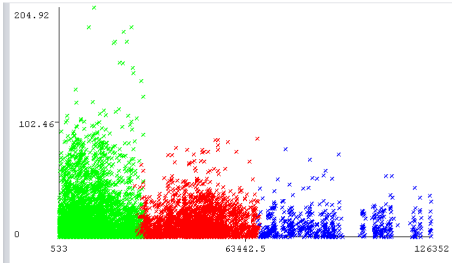

Fig. 8. Per Capita GDP VS SuicidePer100k

a negative relationship exists between the economic state of a country and the rate of suicides that it exhibits.

No doubt, the mental condition of any suicidal victim is unimaginable and does not serve to be probed in the context of the country's economic state, but while aiming to cater to a larger section of the global community, in this case as big as a country, organisations such as WHO need to acknowledge the role played by the monetary state of affairs prevailing in the country, affecting its people on a generation-to-generation basis.

VI. CHALLENGES IN ANALYSING SUICIDE RATES

Prior to identifying the areas that future work could focus on, it is vital to understand the challenges that data analysts eager to study this issue face.

The challenges posed here are of two classes- one relating to the quality of data, and the other relating to the attributes of the day. The first set of challenges are taboo and societal stigma. Not only does the stigma surrounding the word, hinders those in need to ask and access help, but it also majorly affects the way the academia functions in its attempt to acquire good quality data.

[2]As such, only 80 member states of WHO have frameworks put in place for collection of data directly useful to estimate suicide rates. In most other countries, especially developing countries that are at increased risk, have few to no systems in hospitals, psychiatric clinics, mortuaries etc., to register suicidal attempts or mortalities. As discussed in Section II, people with suicidal tendencies show clear indicators and the data regarding their mental state can be recorded under monitored and safe environments only if they are duly taken note of in the first place.

Secondly, and most importantly, the process of data collection needs to incorporate different institutions that are involved with every person's life. To be able to predict whether a particular person would take the fateful decision, we need to build datasets that have attributes such as *presence of intrusive thoughts, feelings of depression, reasons of emotional breakdowns, history of suicides/suicide attempts in the family, history of suicide attempts for the individual, etc.* This calls for enhanced vigil and supervision of suicide attempts, at home, school and old age homes too, apart from hospitals alone.

Moreover, as has been the theme of this report, attributes that quantitatively factor in the cross-national differences that exist between countries and cultures, differences in races and religions in the understanding of suicides have to drafted and brought into practice for an exhaustive study of suicide-related data.

VII. SCOPE FOR FUTURE WORK

This section introduces some directions that the study of suicide-related data can take.

- Due to non-availability of data for matching years, this paper has not explored the links between the happiness scores for each country and the extent of suicide rates recorded in that country for that year. Released by the UN Sustainable Development Solutions Network, the Happiness report consists of attributes such as *GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust Government Corruption* that provide one score for the happiness level of that country. Correlating these two datasets, in addition to complementing each other in their analysis, would also bring to light inadequacies and flaws in the way the individual datasets are recorded and analysed.
- With the rise of media representation and detailed coverage of the aftermath of a headlining case, a new class of copycat suicides are being recorded in the young people of the country. The risk to suicide rates among younger generation from the entertainment industry- suicide rates of each year can be correlated against viewership and popularity of movies and shows that depict suicidal/mental health issues in each country.
- Moreover, due to the escalating popularity of social networking and live-video-sharing sites, glorification of suicidal events and its aftermath has caused many attention seeking people to resort to a public display of their intent to engage in suicide. Video analytic software that is built to identify such events can lead to timely support for the individual in question.
- Study of prevalence of suicide-related terms on the internet- Research argues that the immense information available through a simple google search with suicide-related terms, has engendered pro-suicidal behaviour.
- Correlation of suicide rates among vulnerable groups such as suicides among transpeople, racially discriminated people, physically disabled people etc.

VIII. CONCLUSION

The aim of this report has been to study the relationship between rate of suicides and various other factors such as age and sex, with a special focus on economic condition of the country in which the event was recorded. Data pertaining to 96 countries in the years 2005 to 2015 was used, for each age group. Data Pre- processing was performed to ensure the data was of good quality. In its attempt to identify trends linking suicide rates to different factors, the report both verified and debunked previously established facts from official mental health and suicide prevention organisations from the outcomes of the classification, regression and clustering techniques demonstrated on Weka Tool.

The report catalogues immediate courses of action to be undertaken in terms of data features and data quality, in order to refine the quality of analysis and results of suicide-related data. Finally, an the report concluded by proposing a non-exhaustive list of future analytic work that can be undertaken using suicide-related data.

IX. REFERENCES

[1]https://www.who.int/health-topics/suicide#tab=tab_1

[2]https://www.iasp.info/

[3]https://www.heretohelp.bc.ca/infosheet/what-is-suicide

[4]https://www.timeshighereducation.com/news/lecturers-suicide-wake-call-overworking-academia.

[5]https://www.who.int/news-room/fact-sheets/detail/suicide.

[6]https://www.nami.org/Blogs/NAMI-Blog/September-2020/5-Common-Myths-About-Suicide-Debunked.

[7]https://www.iasp.info/wspd2019/

[8]https://www.samaritans.org/about-samaritans/research-policy/

[9]https://www.mentalhealth.org.uk/statistics/mental-health-statistics-suicide

[10]https://suicidepreventionlifeline.org/

[11]http://www.bcmhsus.ca/our-research

[12]https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016.

[13]https://pandas.pydata.org/pandas-docs/stable/index.html

[14]https://www.cs.waikato.ac.nz/ml/weka/

[15]https://deconforming.com/misgendering/

[16][https://www.theatlantic.com/health/archive/2020/06/why-suicide-rates-among-millennials-are-rising/612943/]

[17]https://www.verywellmind.com/gender-differences-in-suicide-methods-1067508

[18] Secretariat, U.N., Statistical Annex-Country classification Data sources, country classifications and aggregation methodology [Internet]. Statistical Annex. 2012 [cited 2019 Dec 3].