A visual analysis of suicide rates across the World

Siddharth Madan*
Department of Computer Science, School of Science and Technology and
City, University of London

(Dated: August 11, 2023)

This is a new report submitted to fulfill the coursework requirements of INM433 Visual Analytics (PRD1 A 2022/23).

Keywords: Fuzzy C-Means Clustering, Machine Learning, Unsupervised Learning.

I. PROBLEM STATEMENT

In this work, we attempt to explore the possible connections between suicide rates and other common parameters for various countries worldwide. Changes in the human lifestyle have been unprecedented during the past two centuries. From the electrification of households to global connectivity via the internet, there are certain common threads that we all share across the globe. It would be safe to say that from the point of view of our ancestors, the economic and technological growth in the past few decades have made human survival fairly easier and guaranteed. Despite all these pros we still have to deal with a unique set of challenges such as suicide rates amongst people of all age groups, different nationalities, financial backgrounds, etc. This research paper attempts to find the correlation between suicide rates for a hundred different countries from 1985 to 2016 and various other parameters such as age groups, gender, nationality, human development index (HDI), and GDP per capita during those years. Here, we also attempt to predict commonalities between countries based on numerous factors from the considered dataset with the aid of an unsupervised learning technique called Fuzzy C - Means (FCM) clustering. We have chosen FCM over other clustering algorithms for a couple of key reasons, first, due to its ability to be able to handle both numerical and categorical data, and more importantly, during the clustering phase of FCM, inclusion degrees are continuously updated, aiding the algorithm's convergence to optimised solutions.

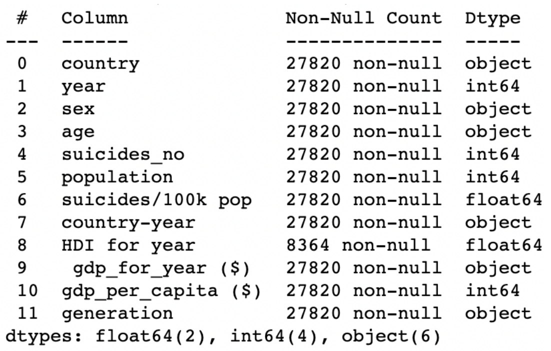
II. STATE OF THE ART

In order to identify the patterns of suicide rates amongst various countries there have been a lot of similar research works conducted. In [1], the author explores the possible relationships between suicide rates and age for a dataset of 62 countries. In this work, the authors found a significant increase in suicide rates with age for almost half of the countries and no significant increase in suicide rates for the rest of the half. Moreover, the author highlighted the fact that in some countries the suicide rates declined with increasing age, and also in some countries suicide rates were higher amongst youngsters. The author also highlighted some of the common traits among old people who had a propensity towards committing suicide such as the inability to form close relationships, tolerate change, loss of control, feeling of loneliness, despair and dependency on others, loss of income, retirement, social isolation or any other bereavement. The author shared another intriguing insight that suicide rates were found to be declining with age for immigrants living in countries like US, UK, and Australia.

In [2], researchers study the suicide rates in relation to Human Development Index (HDI) and other health-related aspects for 91 countries. This paper mentions the average suicide rate globally to be 10.5 for every hundred thousand people in each country. Also, they found that the average suicide rate amongst men is 4 times larger than the same in women. This work majorly emphasises identifying the impact of HDI at a macro or national level i.e. by generalising a country and all its population as equal and suggests that this work could be carried further by considering different strata of society with different socio-economic backgrounds within a country. Some of the tenets the authors have considered to define HDI are life expectancy at birth, mean years of schooling, expected years of schooling, and GDP per capita. The authors have claimed to have found suicide rates to have positive correlations with urbanisation, and life expectancy at birth and especially amongst females. Also, they have found suicide rates to have negative correlations with obesity, unemployment rate, and total fertility rate. In conclusion, they highlighted that suicide rates increased with the increasing levels of HDI, and the rates were higher amongst men as compared to women.

In this research work, we aim to not only conduct visual exploratory data analysis but also attempt to predict the commonalities between various countries by using an unsupervised learning clustering algorithm called Fuzzy C - Means.

* siddharth.madan@city.ac.uk



```
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   country              27820 non-null  object 
 1   year                 27820 non-null  int64  
 2   sex                  27820 non-null  object 
 3   age                  27820 non-null  object 
 4   suicides_no          27820 non-null  int64  
 5   population           27820 non-null  int64  
 6   suicides/100k pop    27820 non-null  float64
 7   country-year         27820 non-null  object 
 8   HDI for year         8364 non-null   float64
 9    gdp_for_year ($)    27820 non-null  object 
 10  gdp_per_capita ($)   27820 non-null  int64  
 11  generation           27820 non-null  object 
dtypes: float64(2), int64(4), object(6)
```

III. PROPERTIES OF DATA

The dataset used for our analysis has been taken from Kaggle [3]. The publishers of this Kaggle page have assimilated this dataset from various resources such as HDI data from the United nation development program (2018) [4], GDP-related data from World Bank [5], and country-wise suicide-related data from another Kaggle page [6]. The table on the top shows the basic statistics for each 12 different columns in our CSV file. The dataset consists of data from 100 different countries, and the data for most of the countries in Africa and Asia is missing from the list of all countries. The span of the dataset is from 1985 to 2016 in time and for each year we have the data for both genders and for each gender, we have six different age groups i.e. 5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, and above 75 years. Corresponding to each age group in the last column is the generation of that group indicating the generation that group of individuals come under, following are the six generations mentioned in our dataset: G.I. Generation, Silent, Gen. X (Boomers), Gen. Y (Millenials), and Gen. Z. The column for human development index (labeled HDI) has certain missing values for certain years for some countries and we have used interpolation to impute those values. We also have columns for GDP, GDP per capita, population, suicide numbers, and suicide per 100 thousand of the population of that country corresponding to each year, age group, and gender. This makes the total entries to be 27820 in total. In order to study the basic correlation amongst all the columns we plotted the correlation matrix and found only two significant but obvious correlations i.e. population with a suicide number of about 0.62 which indicates the higher the population the higher the number of suicides and another correlation of HDI with GDP per capita of about 0.74 which is also understandable since the GDP per capita itself is a considerable factor in the calculation of HDI and this also indicates the consistency of data from both World Bank (GDP related) and UN development program.

In order to apply the FCM clustering algorithm we have used label encoding for all the categorical columns in our datasets and more about the same has been elaborated in the next section. Also, in order to understand the basic overview of suicide rates over the years we plotted this Fig. 1. In this plot, we can see the overall trend of suicide rates which rises upwards from 1988 to 2009 and gradually falls back to the level of 1985 by 2015. A sudden drop in suicide rates indicates that the data for the whole of 2016 might not be considered during cataloging and we haven't clipped 2016 from analysis for consistency and to avoid complication.

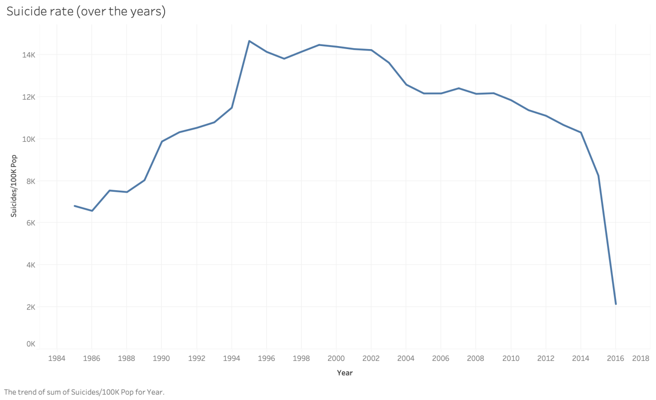


FIG. 1. Suicide rates over the years.

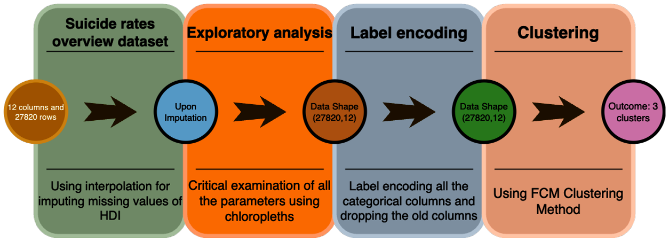
IV. ANALYSIS

This section, which has been further divided into three subsections, primarily focuses on the analytical components of the methodology.

A. Analysis Approach

On the following page, you can see the flow diagram we created to further explain our methodology. It lists each step we took to accomplish this analysis in order.

As seen from the diagram our first step was to impute data in the missing column of HDI which was the only column with missing values for certain years for a few countries. In order to do that we used interpolation and considered HDI values for those same countries but during some other years where the HDI values were calculated and transcribed. In the next step, we began with exploratory data analysis and we plotted numerous parameters corresponding to their counterparts in order to gain an overview of the patterns lurking in the dataset. Suicides occurring across the globe at any given time might have a myriad of reasons as their cause or triggers. It could be due to social or political unrest in a region or it could be an economic slowdown like the one we faced in 2008 and we tried to visualise each important parameter in order to reveal or recognise any such pattern. Upon doing so we had a decent idea about what common traits could be affecting the suicide rates across these countries. Then in order to apply Fuzzy C - Means


Suicide rates overview dataset
Exploratory analysis
Label encoding
Clustering
12 columns and 27820 rows
Upon Imputation
Data Shape (27820,12)
Data Shape (27820,12)
Outcome: 3 clusters
Using interpolation for imputing missing values of HDI
Critical examination of all the parameters using chloropleths
Label encoding all the categorical columns and dropping the old columns
Using FCM Clustering Method

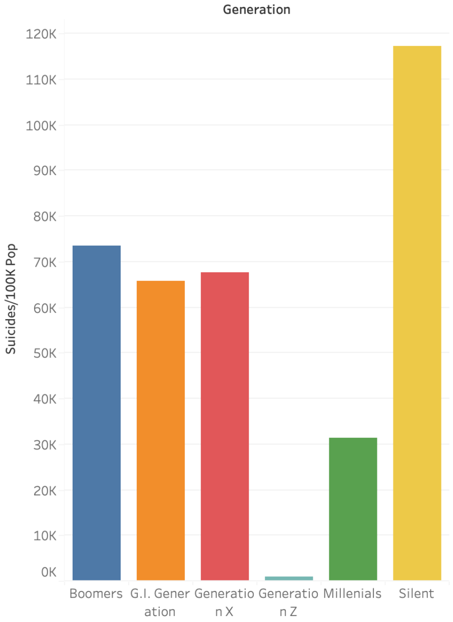


FIG. 2. Bar chart showcasing suicide rates for different generations.

clustering algorithm we had to use label encoding on all the categorical variable columns and eventually drop the old columns. Finally, as the last step of this endeavour, we applied the FCM clustering algorithm and obtained the clusters of countries. We then honed our algorithm with the aid of comparing silhouette scores for different numbers of clusters using the FCM clustering algorithm.

B. Analysis Process

Now we begin with the exploratory data analysis, first, we plot the bar chart showing the suicide rates for all the generations i.e. Fig 2. In this bar chart, we can see that the suicide rates amongst the silent generation i.e. the generation born between 1927-45 is the highest for all the generations. It is trumping the suicide rates mark to approximately 120 thousand deaths, nearly twice as much as the second highest number of deaths, the Boomers. This surge of suicide rates amongst the silent generation can be attributed to the times of World Wars i.e. the aftermath of the first world war and the grim conditions it brought right before their birth and the ongoing second world war during their formative years. The suicide rates for Boomers born between 1946-65, the G.I. generation (touted as the greatest generation) born between 1901-26, and Generation X born between 1966-80 all lie roughly between 65 thousand deaths and 75 thousand deaths. From this bar chart, it seems that this should be the norm for all the generations in the 20th century. For the case of Millennials born between 1981-96, and Generation Z born between 1997- 2010, it would be appropriate to consider that not enough time has elapsed for them to consider such drastic steps with their lives. This also indicates that the probability of committing suicide grows as one age.

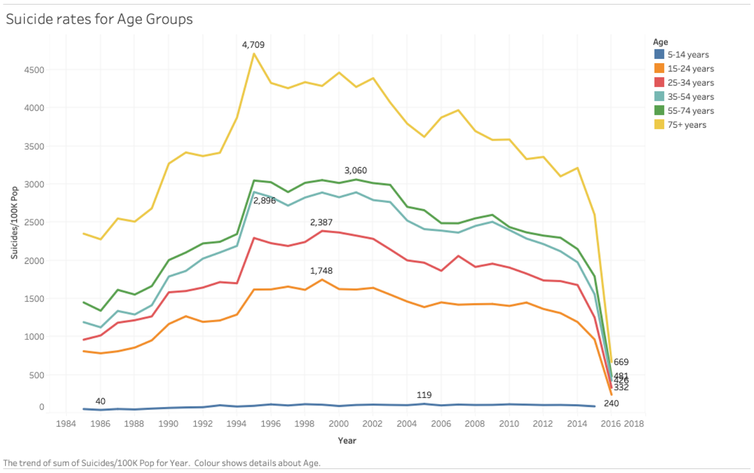
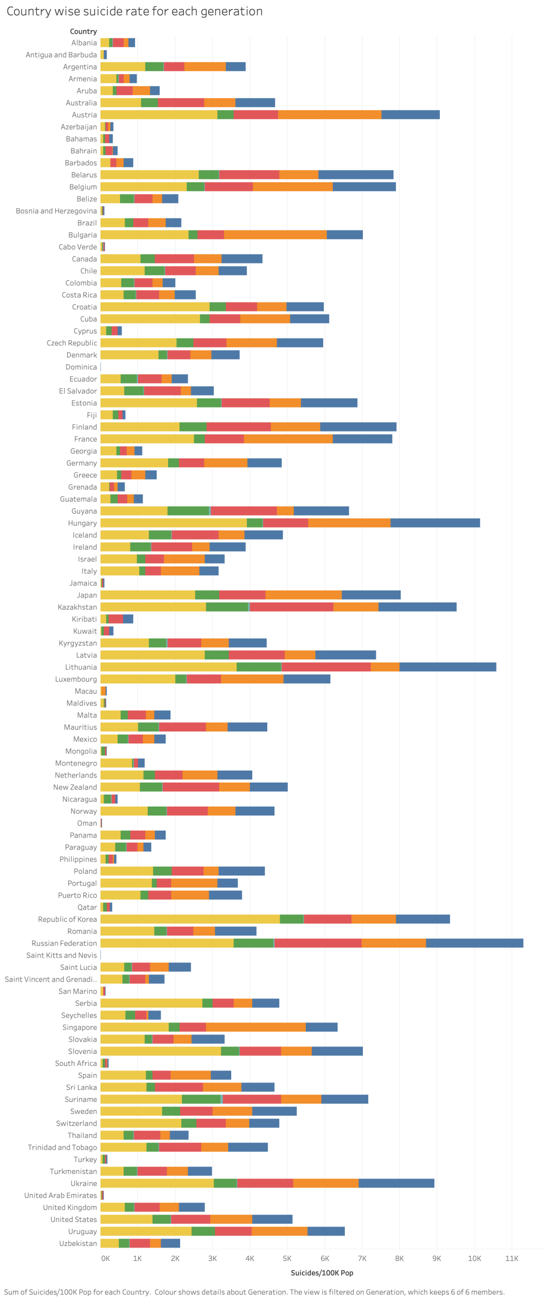


FIG. 3. Plot showcasing fluctuating suicide rates for different age groups over the years.

In order to investigate this assertion we plotted the suicide rates for different age groups over the years of our dataset as shown in Fig. 3. This plot reinforces our assertion that the probability of committing suicide increases with age as each older age group has a higher suicide rate throughout than its adjacent younger age group. This plot also replicates the curve trend as seen in Fig. 1., i.e. there is an evident rise in suicide rates from 1988 onwards which peaks at 1994 for the age group of 75 and above in age which makes them from the G.I. generation, the peak for 55-74 years old ones is around 2000 and so on for the various age groups. This finding also correlates with the one mentioned by authors in [1], in which the authors highlighted the common traits for the elderly to have a propensity towards committing suicide. This revelation is also contrary to a common belief that suicide rates are higher among youngsters as they are immature in handling stress and can be whimsical in taking such drastic steps.

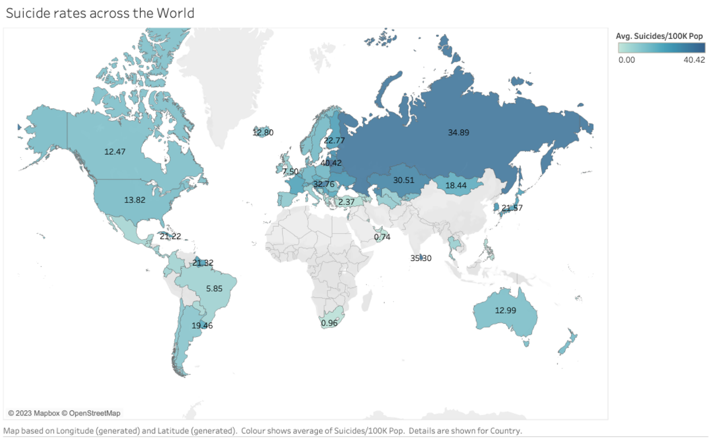
Keeping these considerations in mind we then check go on to examine the country-wise suicide rates for each generation, as shown in Fig. 4. Here, we can see in this bar representation that the silent generation (in yellow) has the highest share of suicides per 100 thousand indi-



Sum of Suicides/100K Pop for each Country. Colour shows details about Generation. The view is filtered on Generation, which keeps 6 of 6 members.

FIG. 4. Bar chart representing country-wise suicide rate for each generation.

viduals of each country's respective population. Countries like Lithuania, Hungary, Austria, Belgium, Russia, Kazakhstan, Ukraine, Latvia, etc., in the European region, have had fairly high suicide rates amongst the G.I. generation which also correlates to the repercussions of the World Wars on those countries. Countries like Singapore, Bulgaria, Japan, France, and Austria have had a significant amount of suicides among their G.I. generation. Also, Lithuania, Kazakhstan, Russia, Belarus, Luxembourg, Guyana, Suriname, and Ukraine are one with equally high suicide rates amongst their Generation X which is not the norm for the rest of the world, this can attributed to be coinciding with the fall of the communist party in USSR during the end of the cold war in the late 80s and early 90s. On the other hand countries like France, Germany, the UK, Canada, Australia, the USA, etc., have quite evenly spaced suicides across all the generations.

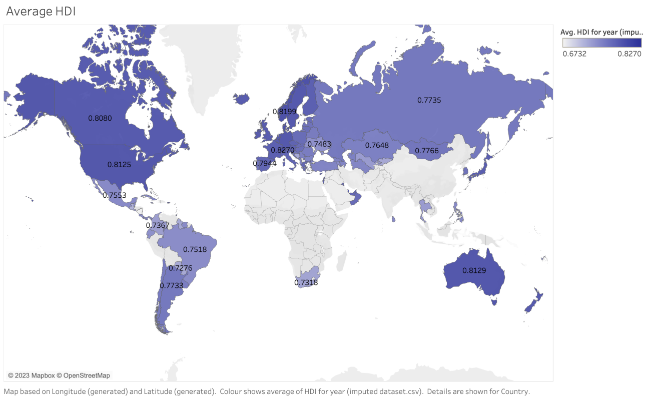


Map based on Longitude (generated) and Latitude (generated). Colour shows average of Suicides/100K Pop. Details are shown for Country.

FIG. 5. Chloropleth map depicting average suicide rates worldwide.

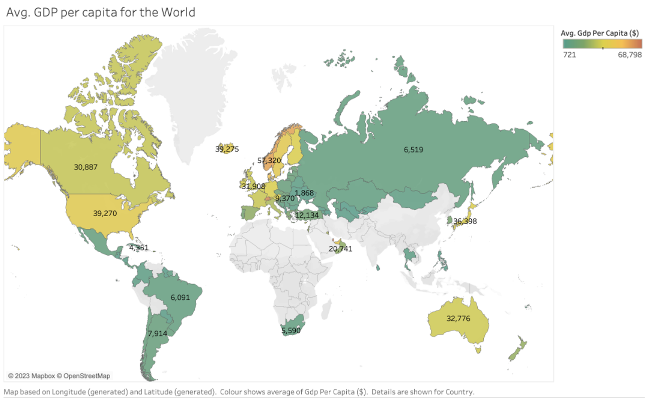
In Fig. 5., we then plotted a chloropleth map depicting the average suicide rates over the years for each country considered in this study. We have also highlighted some of the key countries with their respective suicide rate numeric values on this map to create some context. The highest suicide rate among all the countries is in Lithuania which is 40.42, followed by countries like Sri Lanka (35.30), Russia (34.89), and Hungary (32.76). And the lowest suicide rates are for the countries like Oman, South Africa, Turkey, Greece, and Mexico, etc., the only possible correlation amongst these countries could be their weather being temperate and reception of bright sunny days throughout the year instead of being gloomy like some of the further northern countries leading to individuals being less depressed, but obviously we do not have any such metric in our dataset to explore that possibility. Countries like the US, Canada, UK, France, Australia, Iceland, Argentina, Spain, Portugal etc., have suicide rates closer to the world average which is 10.5 per 100 thousand individuals.

In Fig. 6., we then plotted the chloropleth depicting the human development index (HDI) for all the countries considered in our study. This also highlights a possible loophole in our analysis that we have considered the suicide rates for countries having fairly above average HDI and is excluding most of the low HDI countries, had this not been the case and the dataset for all countries across the HDI spectrum would have been accessible then it might result in some concrete and conclusive findings. Nevertheless, from this plot, we can glean the fact that there is a very low correlation between suicide



Map based on Longitude (generated) and Latitude (generated). Colour shows average of HDI for year (imputed dataset.csv). Details are shown for Country.

FIG. 6. Chloropleth map highlighting the HDI for various countries.

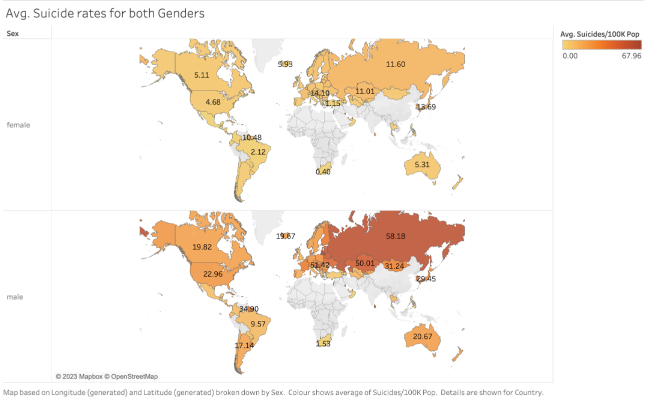


Map based on Longitude (generated) and Latitude (generated). Colour shows average of Gdp Per Capita ($). Details are shown for Country.

FIG. 7. Visual depiction of GDP per capita for all considered countries.

rates and the HDI of a country. Switzerland is the highest HDI country, with Australia, Scandinavian countries, the USA, Canada, the UK, etc., not being too far from it in terms of HDI. Almost all former USSR countries like Russia, Kazakhstan, Ukraine, etc., have fairly high HDI which might indicate the fact mentioned in [2], that HDIs are calculated at a macro level and the disparity of socioeconomic factors aren't considered accordingly. Eventually, this might lead to the wrong interpretation of the well-being of a common individual within a country. This same argument can be carried further for the interpretation of the GDP per capita of a nation and in order to explore that possibility we plotted another chloropleth depicting GDP per capita for all these countries as shown in Fig. 7.

GDP per capita tells a completely different story in the context of suicide rates as we can see that countries like South Africa, Argentina, Brazil, Russia, and Mexico have their GDP per capita between 5000 - 8000 (USD) but their suicide rates have no correlation whatsoever. On the contrary, one can argue that countries like Western European countries, North America, and Australia despite having GDP per capita in the highest bracket, have suicide rates which are significantly higher than the global average. This indicates ambiguity in claiming any direct correlation between GDP and HDI with suicide rates. In the final step of us playing around with the dataset, we plotted a chloropleth again to visualise country-wise suicide rates for men and women. This plot endorses the claim mentioned by authors in [2] about the suicide rates being four times higher in males than that for females. In fact, suicide rates for males in countries like Russia, Kazakhstan, Hungary, and Lithuania are almost five to six times higher than their female counterparts, indicating higher levels of stress, anxiety, and emotional turmoil among men. Upon this, we moved ahead towards the implementation of the FCM clustering algorithm and then eventually compared the outcome from the knowledge gleaned from exploring the intricacies of the dataset.



Map based on Longitude (generated) and Latitude (generated) broken down by Sex. Colour shows average of Suicides/100K Pop. Details are shown for Country.

FIG. 8. Depiction of average suicide rates for males and females.

C. Analysis Results

As mentioned in the previous subsection we will now ponder on the results obtained via application of the FCM clustering algorithm. In order to determine the optimum number of clusters we have taken the aid of comparing the silhouette scores of various clusters as shown in Fig. 9. It is evident that the silhouette scores are dropping with the increase in the number of clusters and even though we get the highest silhouette scores for two clusters we have chosen three as our choice for the number of clusters in order to classify countries with a diverse group range. The silhouette score for three clusters was 0.88 which can be considered to be good enough. In Fig. 10, which is the chloropleth map depicting all three clusters, we can see that countries like Lithuania, Hungary, Latvia, Austria, Sri Lanka, Kazakhstan, Ukraine, Mongolia, etc., all have been grouped in cluster 1. And countries like the USA, UK, Canada, Brazil, Japan, etc.,

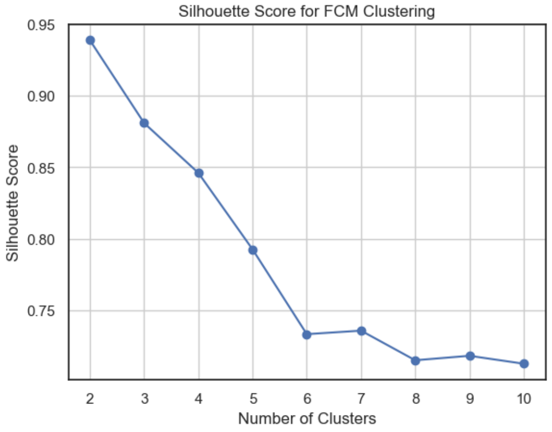
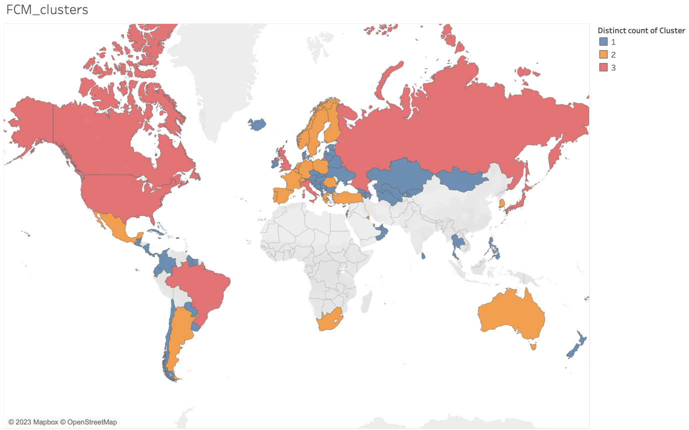


FIG. 9. Silhouette score comparison for different numbers of clusters.



Map based on Longitude (generated) and Latitude (generated). Colour shows details about distinct count of Cluster. Details are shown for Country.

FIG. 10. Depiction of clusters from FCM algorithm.

have grouped in cluster 3. And finally, countries like Switzerland, France, Germany, Greece, Mexico, Scandinavian countries, South Africa, Australia, etc., have grouped together as cluster 2. Commonalities within all these clustered together countries can be discussed in the next section.

V. CRITICAL REFLECTION

In this research work, we attempted to study the commonalities between countries based on their HDI, GDP, suicide rates for various ages and generations, and genders. With information gleaned from comparing all these parameters, we can group certain countries together in various groups to identify and understand certain attributes which affect the suicide rates within a country. In order to critique our analysis we may discuss the outcome of the FCM clustering algorithm. Cluster 1 (in blue) is a group of countries like Lithuania, Hungary, Latvia, Austria, Sri Lanka, Kazakhstan, Ukraine, Mongolia, Guyana, and Columbia, etc., all these countries have relatively high suicide rates, and many of them are former USSR countries and countries with high suicide rates for both silent and G.I. generations. The next cluster is Cluster 2, which mainly consists of countries like Switzerland, France, Germany, Greece, Mexico, Scandinavian countries, South Africa, and Australia, etc., all these countries have higher GDP per capita and HDI than Cluster 1 countries and have reasonably lower suicide rates than Cluster 1 countries.

Finally, the third cluster which consists of countries like Canada, the UK, the USA, Japan, Brazil, Russia, and Italy. All these countries have a lot in common such as Canada, the US, the UK, and Japan have quite a similar GDP per capita and HDI. But Brazil, Russia, and Italy are outliers in that regard. Brazil and Italy could have been included in another cluster of their own with other countries like New Zealand, Turkey, etc., as they all have similar suicide rates, GDP per capita, HDI, and lesser influence of the cold war politics. Also, Russia could have been included in Cluster 1 as all its counterpart countries with similar attributes have been included in that cluster.

Even though all three clusters group most similar nations into their respective groups, it is imperative to address the fact that not all nations in those groups belong to those groups. This problem can be tackled by adding data from more middle to low HDI and GDP per capita countries, in order to not only give us a better context but also to provide our clustering algorithm with a wholesome picture of the problem. These measures will yet be dealing at the macro level as mentioned in [2], and only if one could gather or obtain data for different socioeconomic strata even for a few countries then any such clustering algorithm might surely perform better than this algorithm. We can use transfer learning (reinforcement learning technique) in order to estimate the missing socioeconomic strata data for the rest of the countries and then possibly apply any suitable clustering algorithm.

As a form of future work, we can consider the application of hierarchical clustering, DBSCAN, or Gaussian Mixture Models in order to cluster these countries and possibly compare the results of all these clustering algorithms.

VI. REFERENCES

[1] Ajit Shah, 2007 The relationship between suicide rates and age: an analysis of multinational data from the World Health Organization, International

Psychogeriatrics, 19(6), 1141-1152 pages=1141–1152 https://doi.org/10.1017/S1041610207005285

[2] Khazaei S, Armanmehr V, Nematollahi S, Rezaeian S, Khazaei S. Suicide rate in relation to the Human Development Index and other health related factors: A global ecological study from 91 countries. J Epidemiol Glob Health. 2017 Jun;7(2):131-134. doi: 10.1016/j.jegh.2016.12.002 Epub 2017 Feb 7. PMID: 28188120; PMCID: PMC7320427.

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

[4] United Nations Development Program. (2018). Human development index (HDI). http://hdr.undp.org/en/indicators/137506

[5] World Bank. (2018). World development indicators: GDP (current US Dollars) by country:1985 to 2016. http://databank.worldbank.org/data/source/world-development-indicators

[6] [Szamil]. (2017). Suicide in the Twenty-First Century [dataset]. https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook

VII. WORD COUNT

Section	Expected Words	Actual Words
Problem Statement	250	242
State of the Art	500	416
Properties of Data	500	464
Analysis Approach	500	280
Analysis Process	1500	1220
Analysis Results	200	200
Critical Reflection	500	474