Suicides Around the World: A Visual Analysis

Vighnesh Mane

Abstract—Suicide rates appear to be rising, indicating a pressing need for more effective preventive efforts. Concern for suicide analysis should be taken very seriously, as it will give appropriate preventive measures depending on the circumstances. Analysis of the Kaggle's 31-year-long data set of suicide cases is used in order to reduce the number of suicides. In addition to showing the temporal pattern of suicides, the data also shows the geographical pattern throughout the globe. Temporal patterns are represented using the python packages seaborn and matplotlib, which are both written in the programming language Python. Line charts and bar charts are used to illustrate the results of the temporal analysis. The geographical analysis is shown visually by a density map. Spatial analysis is carried by using the Plotly library. For the purpose of locating and grouping comparable data points, K-means and hierarchical clustering are used. Suicide rates can be reduced, in my opinion, through the use of predictive analytics. Therefore, linear regression is used to predict the occurrence of suicides.

1 Problem Statement

As per WHO (World Health Organization's), Suicide claims the lives of more than 700,000 individuals each year, or one person every 40 seconds. Self-harm is a worldwide epidemic that affects people of all ages and backgrounds. When it comes to preventing suicide and suicide attempts, evidence-based interventions can be conducted at the population, sub-population, group, and individual levels. There is evidence to suggest that for every adult who died by suicide, there were more than 20 others who attempted to take their own lives at some point [8].

I believe that the use of temporal and spatial analysis can assist authorities in reducing the suicide rate around the world. Additionally, the clustering will group the relevant data points that will form a certain pattern, and the linear regression will tell us what the future scenario of suicides will be. The following are the general trends, temporal trends, and spatial trends that will be identified:

1. What is the suicide tendency among people of different ages groups?
2. Which generation is responsible for the highest number of suicides?
3. Has the trend accelerated or slowed throughout the period 1985-2016?
4. In terms of the number of male candidates, what is the overall trend?
5. What is the pattern of female candidates?
6. Is there any correlation between the number of suicides and the suicide rate across generations?
7. In the future, what can be predicted about the number of suicides between generations?
8. Which country has the highest suicide rate?

2 State of the Art

Mrs. B. Ida Seraphim, Subroto Das, and Apoorv Ranjan [2] began by creating a bar graph depicting the suicide rate per 100,000 persons. Additionally, it indicates the distribution of the various age groups and the gender of the individual. One striking tendency is that the male suicide rate is significantly greater than the female suicide rate. The ratio is greatest between the ages of 35 and 54 years. Suicide is most likely to occur between the ages of 35 and 54. Another surprising result is that women over the age of 75 are more likely to commit suicide, reversing a previous tendency. Next, they take a look at the 10 countries with the most suicides. Russia, the United States, and Japan are the three countries with the most suicides, while the United Kingdom has the fewest. After that, they looked at the trends in suicides from year to year. 1997 features several peaks and valleys followed by consistent declines from 2002 to 2008, which is followed by a rise and fall from 2009 to 2015. After absorbing some of the information in this paper, I plan to use a line plot to see whether there is a pattern in the number of suicides among the various age groups and compare that to other factors. This is followed by an analysis of suicide rates by generation using a line chart. For the sake of historical perspective, I'll examine the trend of suicides over the previous 31 years. Suicide statistics for male and female candidates will then be compiled for comparison purposes. In addition, for spatial analysis, I will be plotting on the suicides around the world and suicides per hundred thousand population.

Sujatha, S. Sree, E. P. Ephzibah, and R. Kiruba [3] have used K-means and hierarchical clustering in their work. One of the most effective clusters was created using Silhouette scores for age and gender scatterplots and K-means boxes, which may be applied to any group. A scatter plot shows the results of hierarchical clustering as well. As a result of learning from this paper, I plan to use K-means clustering and plot it on a pair plot, as well as hierarchical clustering.

Wang N., Luo F., Shivtare Y., Badal V. D., Subbalakshmi K. P., Chandramouli R., & Lee E [1] have performed various techniques such as KNN, SVM, random forest, and linear regression for the prediction of suicides has been investigated extensively. After reading this work, I discovered that they attained f1 and f2 scores of 0.563 and 0.584, respectively, for linear regression in their experiments. I've opted to use linear regression to predict suicides among different generations, but I'll calculate the accuracy of the model and implement it.

3 Properties of the Data

The data has been obtained from Kaggle. In Kaggle this dataset was made from four other datasets connected by time and location was used in the Kaggle competition, which was designed to uncover signals that were associated with rising suicide rates among different cohorts around the planet and across the socio-economic spectrum. The data set contains

27821 occurrences that occurred in various locations around the world between 1985 and 2016. Such occurrences are appropriate for the purposes of deriving the conclusions of this paper. A variety of variables are included in this data, including the following: country, year, sex, age, suicides_no, population, suicides/100k, country-year, HDI for year, gdp_for_year ($), gdp_per_capita ($), generation (based on age grouping average).

The variables generation and count of suicides are employed in my analysis in order to establish the overall trend of the various age groups. I will use the same variables, which were generation and the suicides_no committed by each generation, for the goal of studying the trend in suicides among different generations. For temporal analysis, the year variable is included in the data set to aid in the analysis of trends in suicides between 1985 and 2016. This will allow for a more accurate comparison of patterns between 1985 and 2016. Also taken into consideration is the year variable, which will be used to examine the trends in suicides among the male and female candidates. I also decided to use the generation variable to perform K-means clustering, and the count of suicides and the population variables to perform hierarchical clustering, in addition to the generation variable. The country variable is appropriate for use in geographic analysis. Because of this, I will utilise the country variable to display suicide rates around the world as well as suicide rates per 100 thousand people in different countries. Finally, I will use the generation variable to make predictions about suicides.

3.1 Missing Values

It was discovered when I looked at the dataset that the HDI variable, which stands for Human Development Index, had 19456 missing values when I checked the missing data. The HDI variable, on the other hand, is not significant in any of the analyses presented in this research. Consequently, I chose to leave out the variable from any form of analysis rather than filling in the missing values by computing the mean of a variable that I had included.

3.2 Outliers

In a spontaneous small group of participants, outliers are data points that are statistically significantly different from the rest of the data points in the group. As a result, it is required to exclude the outliers from the population. I used the Z-score technique to look for outliers in this data set, and I was able to eliminate few of them.

4 Analysis

4.1 Approach

Fig. 1. depicts the schematic workflow of the analysis. Firstly, I collected the data set from the Kaggle. This data set is having the information which is required for the analysis. This data set consists of 12 columns and 27821 rows. The country variable contains 130 unique countries. The year variable consists of years from 1985 to 2016. The sex variable consists of two unique values that is male and female. The age column

Fig. 1. Flowchart of analysis

contains various age groups out of which the age group between 15 to 24 and 35 to 54 hold 17% each respectively and the rest age groups hold 67%. The suicide_no represents the suicide counts and the population variable represent the population of 130 countries. The suicides/100k pop represents the suicide rate per 100k population. The country-year show the combined countries and year. HDI for year represents the human development index rate and the gdp_for_year ($) represents the GDP for all the years and the gdp_per_capita ($) shows the GDP divided by the population and the generation represents the various generations.

Second, I will use Python programming to perform data pre-processing on the information. Due to the large number of missing values in the HDI variable during data pre-processing, I will opt to ignore that column and instead detect and remove outliers rather than fill it with the mean of the column. Further, I will conduct a temporal analysis, which can reveal a general trend in suicide rates between 1985 and 2016. In addition, I will check a trend in both males and girls committing suicide. In addition, I will utilise plotly to plot the locations of suicide cases all around the world when conducting spatial analysis. In addition, I will plot the number of suicides per 100,000 people using the plotly programme.

In addition, I will adjust the dataset since I wanted to concentrate on the data generated throughout the generating process. As a result, I will separate the generation column

from the original data set and then consider each generation separately. Following that, I will perform feature engineering on the improved data set. On this data set, I will run the correlation analysis. After getting results of the correlation were such that I will discover that all of the generations were substantially associated. After that, I will choose the highly correlated variables and run the hierarchical clustering and K-means analyses. As it is necessary to pick all of the highly correlated variables from the changed data set. Predictions concerning suicide are made with the help of linear regressions. So, I will choose the Independent and dependent variables which will be two elements. Then I will Split the dataset in 80 percent and the 20 percent which will be training set and the testing set respectively. Further I will import important libraries in order to perform the linear regression so that I can predict the suicide cases.

4.2 Process

4.2.1 Temporal analysis

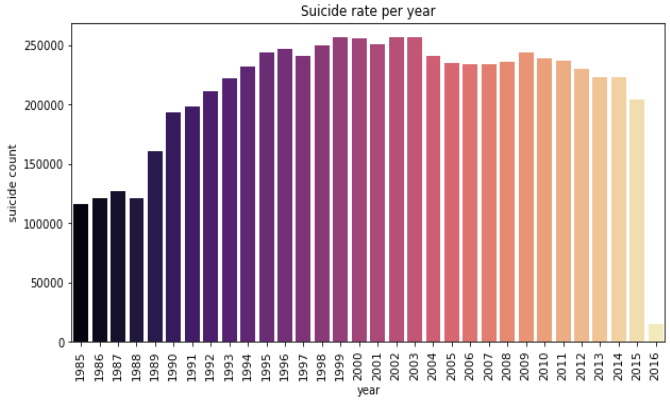

Fig. 2. Trend of suicides from 1985-2016

In accordance with the illustration in Fig. 2. Between 1985 and 1988, the annual suicide rate was somewhat higher than one lakh people. The number of suicides reached one and a half lakhs in 1989, and it began to rise steadily from that point on, eventually reaching two and a half lakhs in 2000. As a result, from 1985 to 2000, there was an upward trend in the index. After falling below two and a half lakhs in 2001, it rose to two and a half lakhs in 2002 and 2003, before sinking again in 2004. It was lower in 2004 than it had been in 2003. In 2005, the number of suicides fell slightly below the number of suicides that occurred in 2004, and from 2005 to 2007, the number of suicides stayed constant. With the exception of a minor increase in 2008 and 2009, the suicide rate has been declining since 2010. As a result of the downward trend that began in 2010, the number of suicides fell below two lakhs in 2015. Due to the fact that the data for 2016 was not totally available, I only used the data that was available for 2016.

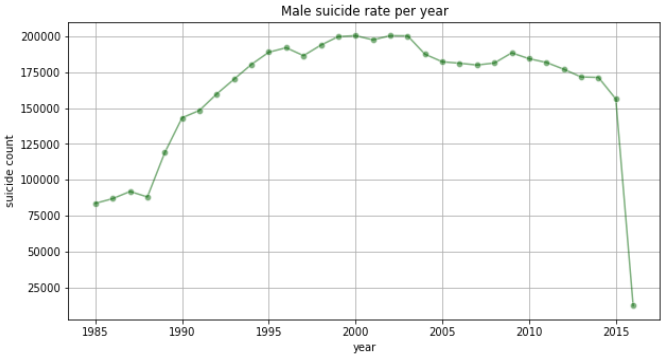

Fig. 3. Trend of male suicides from 1985-2016

As seen in the depiction in Fig. 3, Annual suicide rates for male candidates were somewhat greater than seventy-five thousand between 1985 and 1988. After reaching around one lakh and twenty-five thousand suicides in 1989, the number of suicides proceeded to grow consistently from that point on, eventually reaching two lakhs in 2000. As a result, the index showed an upward trend from 1985 to 2000. The number of suicides fell below two lakhs in 2001, but increased to two lakhs in 2002 and 2003, before declining again in 2004. In 2004, it was lower than it had been in 2003. In 2005, the number of suicides decreased slightly below the number of suicides that occurred in 2004, and the number of suicides remained stable from 2005 to 2007. After a brief spike in 2008 and 2009, the suicide rate has been steadily dropping since 2010. Suicides fell below one lakh and seventy-five thousand in 2015 as a result of a declining trend that began in 2010.

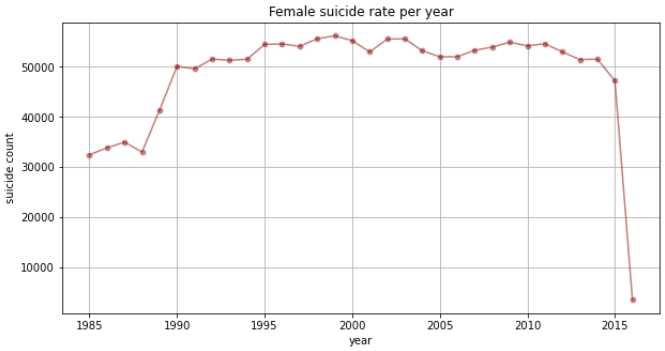

Fig. 4. Trend of female suicides from 1985-2016

As illustrated in Fig. 4, the annual suicide rates for female candidates were somewhat more than thirty thousand between 1985 and 1988, as shown in the graph. After surpassing 40000 suicides in 1989, the number of suicides continued to rise steadily from that point on, eventually reaching roughly 55000. Because of this, the index experienced an upward trend from 1985 to 2000. Despite a decrease in the number of suicides in 2001, the number of suicides climbed to over 55000 in 2002 and 2003, before decreasing again in 2004. It was lower in 2004 than it had been in 2003. When compared to 2004, the number of suicides reduced by a small margin in 2005, while the number of suicides stayed steady from 2005 to 2007. Suicide rates have progressively declined since 2010, following a brief rise in 2008 and 2009. Suicides fell below 50,000 in 2015, continuing a downward trend that began in 2010.

4.2.2 Spatial analysis

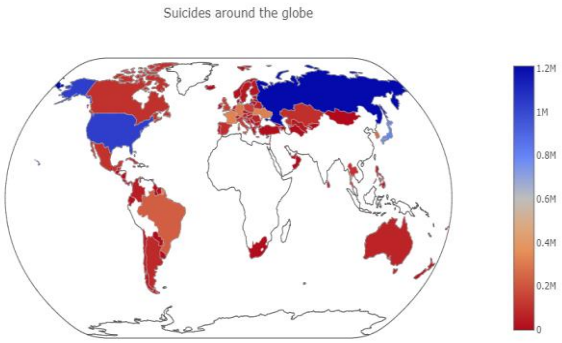

Fig. 5. Suicides around the world

Fig. 5. depicts the number of suicides that have occurred over the world. It is not possible to include suicide cases in the blank nations because the data for those countries was not accessible. In accordance with Fig. 5., Russia has the highest number of suicide cases, with more than 1.2 million cases reported in the country. The number of suicides in the United States was somewhat higher than one million. Japan was the third most suicide-prone country, with an estimated 800,000 people taking their own lives. Suicide rates in European countries such as France, Germany, and Ukraine are approximately 900,000. Other European countries, such as Norway, Sweden, the United Kingdom, Portugal, Spain, and Finland, have a suicide rate of approximately 346 thousand. Brazil, a South American country, has a suicide rate that is slightly higher than 225 thousand. The number of suicides in other South American nations including Argentina, Chile, and Paraguay is approximately 125 thousand. Australia alone has around 70 thousand suicides, while the Sultanate of Oman has only 33 such deaths.

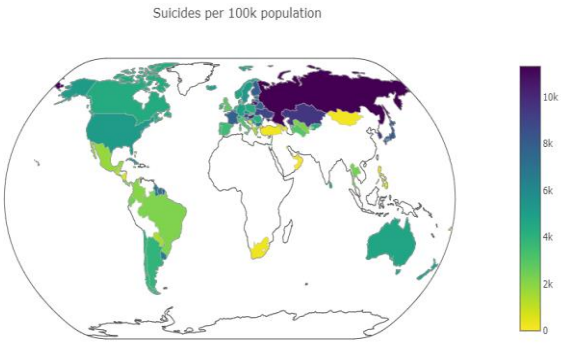

Fig. 6. Suicides around the world per 100 thousand people

Fig. 6. shows the number of suicides that have happened per 100,000 persons around the world since 1985. According to Fig. 6, Russia has the largest number of reported suicides, with more than 11 thousand cases documented in the country. There have been more than 6 thousand to 10 thousand suicides in countries such as Ukraine, Kazakhstan, Japan, the Republic of Korea, Belarus, Latvia, France, Finland, Suriname, Austria, Bulgaria, Guyana, and Uruguay. The number of suicides in countries such as the United States, Canada, Sri Lanka, Australia, Romania, New Zealand, Poland, Germany, Sweden, Norway, and Iceland is between 4 thousand and 6 thousand per 100,000 population. Suicide rates in countries such as Argentina, Chile, Brazil, and Columbia range between 2 thousand and 4 thousand per year. The remaining nations, which include Paraguay, Nicaragua, South Africa, Bosnia and Herzegovina, Turkey, Oman, Mongolia, the Philippines, Greece, and Montenegro, have suicide rates ranging from zero to two thousand per 100,000 population.

4.2.3 Clustering

Unsupervised learning techniques such as clustering are among the most widely used.

4.2.3.1 Hierarchical Clustering

I started by performing feature selection in order to perform hierarchical clustering. After performing the correlation, I discovered that the suicides_no and the population variables are highly correlated in feature selection. The hierarchical clustering is depicted in Figure 7. It demonstrates whether or not the population and suicide rates are comparable. Figure 7 clearly illustrates four distinct clusters. The first cluster, which is crimson in color, indicates that the suicide rate for a population of 0 to 1 million people is between 0 and 20000. The second cluster, which is violet in hue, indicates that the suicide rate is between 4000 and more than 60000 in a population of 0.9 to 2 million people. The dark green color of cluster 3 indicates that the suicide rate for the population of 0 to 0.3 million people is between 0 and 10000. The final cluster, which is cyan in hue, depicts a suicide rate ranging from 30000 to 50000 in a population of 2.5 to 3 million.

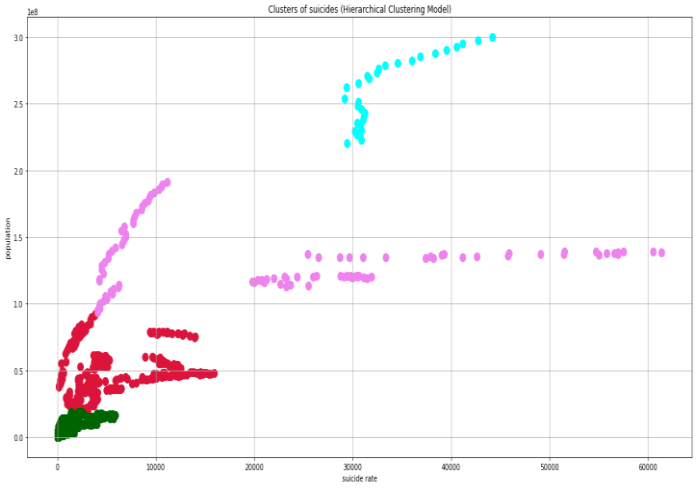

Fig. 7. Hierarchical clustering of suicide counts and population

4.2.3.2 K-Means Clustering

Fig. 8. K- Means clustering of generations

To perform K-Means clustering on the generation variable I first made a new data set and the separated each generation and then performed correlation on the data set in which I found that the all the generations was highly correlating with each other. Fig. 8. depicts the K-Means clustering. For boomers and silent generation, the cluster 0 is forming between 0 to 0.5 for boomers and 1 to 4 for silent, the cluster 1 is forming between 0.5 and 1 for boomers and between 1 to 8 for the silent. The cluster 2 is forming at 0 for boomers and between 0 to 1 for silent. For the boomer's generation the cluster 2 is above 1, while the cluster 0 was above 0.1 and the cluster 1 is at zero. For G.I. Generation the cluster 2 is above 5 and the cluster 0 is at zero and cluster 1 is also at zero. For Generation X the cluster 2 touched the 1.2 mark and the cluster 0 is at 0.1 and the cluster 1 is at zero. For Generation Z the cluster 2 is at 4 and the rest two clusters are at zero. For millennials the cluster 2 is above 8 and the cluster zero is at 0.1 and the cluster 1 is at zero. For the silent generation the cluster 2 is touching 8 and cluster 1 is at 0.1 and the cluster zero is at zero.

4.2.4 Prediction of suicides

The linear regression method is used to make predictions about suicides. To apply linear regression, I chose the year variable as the dependent variable and the remaining factors as the independent variables. After that, I divided the data set into two parts: the training set and the testing set. I preserved 80 percent of the data for training purposes and 20 percent for testing purposes. The model is then deployed after I loaded the library and finished with the model. Following the deployment of the model, I do an accuracy check on the model. The model's accuracy is 57.5 percent on the average. Furthermore, I calculated the root mean square error (RMSE) for the training set, which was 7.04, and the root mean square error (RMSE) for the testing set, which was 7.39.

4.3 Results

For the purposes of temporal analysis, the trend in suicides was relatively flat for the majority of the years after initially climbing during the early stage. Similarly, the pattern for male suicides followed a similar pattern to the general trend, and the pattern for female suicides followed a similar pattern to the overall trend as well. According to spatial analysis, the number of suicides is exceptionally high, having surpassed one million. Furthermore, the number of suicides in Russia exceeds 11 thousand per 100 thousand population, which is a cause for concern. In addition, I discovered the age-based suicides and the generation-based suicides, which can be seen in Figs. 9 and 10, respectively, of the report. In accordance with Fig. 8., the highest number of suicides occurring in the age group of 35-54 years being the most common. Additionally, as illustrated in Fig. 10, the highest number of suicides occurring among the Boomers generation. Also, I predicted the suicides using linear regression and according to it there might be chance of 57.5 percent that the history might repeat.

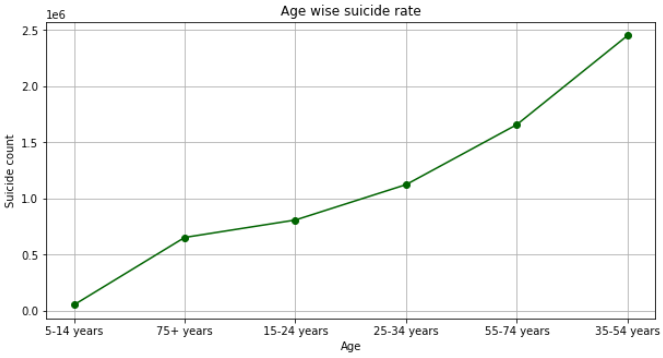

Fig. 9. Age wise suicide rate

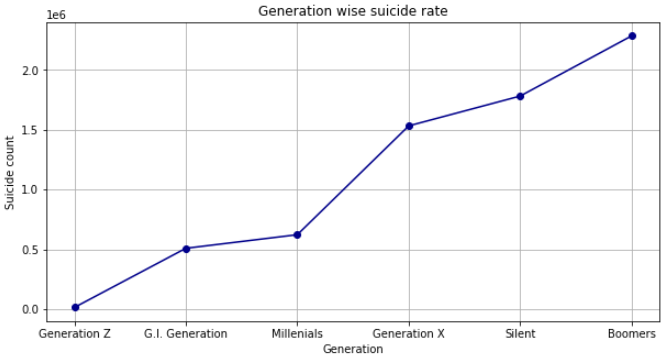

Fig. 10. Generation wise suicide rates

5 Critical reflection

After pre-processing, I moved on to the analysis portion of the project. Following a thorough investigation, I discovered a trend in the number of suicides among people of various ages groups. I discovered that the tendency of suicide among the for the age group between 35 and 54 years was particularly high, owing to the high level of job pressure experienced by people in this age group. In addition, the suicide rate for children between the ages of 5 and 15 was the lowest. After that, I conducted an analysis of suicides by generation, and discovered that the Boomers generation, which ranged in age from 58 to 76 [9], had the highest number of suicides, while the Generation Z, which ranged in age from 10 to 25, had the lowest number of suicides when compared to other generations.

In order to do a temporal analysis, the I discovered the pattern in suicides from 1985 to 2016, which spanned three decades. Following the visualisation of the trend. I discovered that the pattern was rising at the beginning stage, but that the trend was relatively flat during the middle years, and that the pattern ended up being flat at the conclusion. In addition, I look at the patterns of suicides for both men and women. In my research, I discovered that the tendency for males was upward throughout the early stage and then acted flat after that. The pattern for girls was very similar to the pattern for males in that it rose during the initial stage and then acted flat. Based on spatial research, I discovered that Russia had the highest number of suicides, with more than 1.2 million cases, which might be attributed to the difficult socioeconomic situations that the Russian people were experiencing at the time. As well as this, I looked at the number of suicides per 100 thousand people in the population, and discovered that Russia was once again at the top of the list with more than 11 thousand suicides per 100 thousand people.

In order to accomplish the clustering, I first updated the original data set and then did feature selection on the modified data set, which revealed that the population, suicides_no, and all of the generations were all highly correlated with one another. I used Hierarchical clustering to divide the data into four groups, and the variables population and suicide no were used in the analysis. When I used hierarchical clustering to analyse the data, I observed that cluster 2 had the largest number of suicides when compared to the other groupings of individuals. I used all of the generations to perform the K-Means analysis, and I examined three different clusters for it. I discovered parallels between the millennial and baby boomer generations using K-Means clustering. Finally, in order to forecast suicides, I ran a linear regression analysis, which revealed that the model worked accurately in 57.5 percent of the instances.

Table of word counts

Problem statement	249
State of the art	450
Properties of the data	458
Analysis: Approach	487
Analysis: Process	1474
Analysis: Results	189
Critical reflection	487
Total	3794

References

[1] Wang, N., Luo, F., Shivtare, Y., Badal, V. D., Subbalakshmi, K. P., Chandramouli, R., & Lee, E. (2021). Learning Models for Suicide Prediction from Social Media Posts. arXiv preprint arXiv:2105.03315.

[2] Mrs. B. Ida Seraphim , Subroto Das , Apoorv Ranjan, 2021, A Machine Learning Approach to Analyze and Predict Suicide Attempts, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 04 (April 2021).

[3] Sujatha, R., S. Sree Dharinya, E. P. Ephzibah, and R. Kiruba Thangam. K-Means and Hierarchical based Clustering in Suicide Analysis.

[4] Boudreaux, E. D., Rundensteiner, E., Liu, F., Wang, B., Larkin, C., Agu, E., Ghosh, S., Semeter, J., Simon, G., & Davis-Martin, R. E. (2001, January 1). Applying Machine Learning Approaches To Suicide Prediction Using Healthcare Data: Overview And Future Directions. Frontiers. https://www.frontiersin.org/articles/10.3389/fpsyt.2021.707916/full.

[5] Torok, M., Konings, P., Batterham, P.J. et al. Spatial clustering of fatal, and non-fatal, suicide in new South Wales, Australia: implications for evidence-based prevention. BMC Psychiatry 17, 339 (2017). https://doi.org/10.1186/s12888-017-1504-y.

[6] Pérez-Costillas L, Blasco-Fontecilla H, Benítez N, Comino R, Antón JM, Ramos-Medina V, Lopez A, Palomo JL, Madrigal L, Alcalde J, Perea-Millá E, Artieda-Urrutia P, de León-Martínez V, de Diego Otero Y. Clusters de casos de suicidio espacio-temporal en la comunidad de Antequera (España) [Space-time suicide clustering in the community of Antequera (Spain)]. Rev Psiquiatr Salud Ment. 2015 Jan-Mar;8(1):26-34. Spanish. doi: 10.1016/j.rpsm.2014.01.007. Epub 2014 Jun 27. PMID: 24986472.

[7] M. E. Larsen, M. Torok, K. Huckvale, B. Reda, S. Berrouiguet and H. Christensen, "Geospatial suicide clusters and emergency responses: An analysis of text messages to a crisis service," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 6109-6112, doi: 10.1109/EMBC.2019.8856909.

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

[9] https://www.beresfordresearch.com/age-range-by-generation/