Suicide Overview Dataset Report

Mark Movh

ABSTRACT

This statistical analysis explores a Kaggle dataset, revolving around suicide numbers and rates, involving other socio-economic factors, ranging from 1985 to 2016. Questions were presented with the report, in hopes of gaining insight into important information such as recognition of patterns, and developments that may decrease the number of suicides. It was found that using the Spearman's correlation, relationships existed between HDI vs GDP/Capita (0.93) and Suicide Count vs GDP (0.864). Additionally, looking into sexes, various graphs demonstrated a visual difference and further performing a 2-sample t-test at significance level 0.05, mathematically showed a significant difference between all generations of males and females. Finally, Norway, Lithuania, South Africa, Thailand and Guatemala all showed an increasing HDI, but suicide rates varied across nations. Indicating how suicide rates may be impacted by socio-economic factors, and were generally decreasing throughout 2 decades.

1 INTRODUCTION

Suicide is one of the important factors when it comes to a society's issues. Several aspects can impact individuals in a negative way, with extreme cases ending in suicides. According to a WHO report from 2019 based on suicide, the number of global suicides is estimated to be 703,000 per year, with 1.3% of all deaths (one in every hundred) resulting from someone taking their life. Albeit each one of the numbers represents a tragedy, the report explains and demonstrates that in almost a 2-decade period, suicides have decreased by 36% globally across all ages [3] Looking into socio-economic data in regards to numbers of suicides and their rates are important because it is a measurement that can provide an insight into a society's treatment of individuals and how these numbers could potentially be decreased as a result. The premise of this report takes this into account, and the decreasing suicide rates, and investigates further into social and economic factors through a dataset and how they may impact suicide numbers.

2 DATASET DESCRIPTION

This report is based on a Kaggle dataset called "Suicide Rates Overview 1985 to 2016" [5] compiled in 2018. The creator was a Kaggle user, Rusty, who collected and organised information from 4 other sources. These included the UN Development Program for HDI values, World Bank for economic values such as GDP, another Kaggle dataset focused on suicides in the 20th century, and lastly the World Health organisation focused on suicide prevention. The dataset used in this statistical report is a combination of the other datasets mentioned. The dataset is filled with information about the socio-economic factors (HDI, GDP for year, GDP per capita, populations), countries, years and suicides (number of suicides, suicides per 100k pop). Due to a wide-ranging area of fields, several questions were composed to be able to investigate key topics and present meaningful thoughts. The report aims its focus towards prediction from some correlation, differences between groups and change over time. These aims are explored through the three questions presented below:

(1) To what extent do, if any, correlations exist between continuous variables of the dataset?
(2) How can a difference be seen between males and females of various generations?
(3) Can the statistics based on suicide rate be an accurate representation of a developing/developed society?

Question 1 was deemed appropriate to use in this statistical report because it looks at a variety of factors and variables provided by the dataset. Furthermore, by focusing on correlations specifically, it is possible to see every relationship between the variables. Noting this, it can be possible to extract useful information about how variables impact each other. How does one react as the other increases, or decreases? Such information is important to collect because it builds the foundation for looking into specific correlations, and then further exploring them to gain some form of insight. Suicides and their correlated variables can be determined, then in the future it could be possible to determine such values as predictions.

In regards to importance of question 2, differences among groups and sexes are important to identify because it can lead to better decision making regarding what groups need support, and how can they be helped, so that numbers may decrease. Males and females are subject to different lives, with different societal expectations. These expectations and cultural backgrounds create various environments that impact them on a group level. Furthermore, various generations are included in the question because it shows how certain groups (based on generations) experience higher number of suicides, implying that further studies must be gone into why this happens.

An interesting factor of the dataset was including the Human Development Index (HDI). The reason that this was deemed interesting was because, whereas GDP revolved around the economical factor of the country, the HDI includes other variables such as education and health, in addition to the income. Meaning that, the question arose because it could be important to extract useful information from how a society has developed (through changes in HDI) and how those measures against suicide rates. It can lead to then exploring how a country's stability and society can impact suicide rates. Does a less progressed society result in harder and tougher conditions on a human? These questions can then help explain some of the suicides numbers gathered from the dataset.

3 DATA ANALYSIS

3.1 Question 1

Hypothesis: There are several variables that range from social development measures to different ways of presenting suicides. Taking this into account, it is believed that there will be some correlation between some of these development measurements (HDI and GDP) as for example, HDI takes into account income which means that at least one measurements of economic health should impact HDI in some way. Furthermore, since the social development variables say something about a society, it is believed that the statistics on suicide might have some correlation to one of these social factors, as they take into account status of employment, income, healthcare and education. [1] As such linear relationships are expected.

Firstly, considering that the correlation wants to be explored, the actual correlation method needed to be chosen. In this report, to investigate the correlations, the Spearman's correlation is used because it is not known whether the data is normally distributed. Using the Spearman's correlation, a heatmap was created which calculated all Spearman's correlations between all continuous variables. The correlation ranges from -1 to 1, which -1 indicating a perfect negative relationship, and 1 representing a perfect positive relationship. Through this map, it is possible to explore what the relationship between these continuous variables are, in terms of mathematics. See below to find the formula for calculate Spearman's correlation coefficient, and the matrix heatmap.

$$r = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

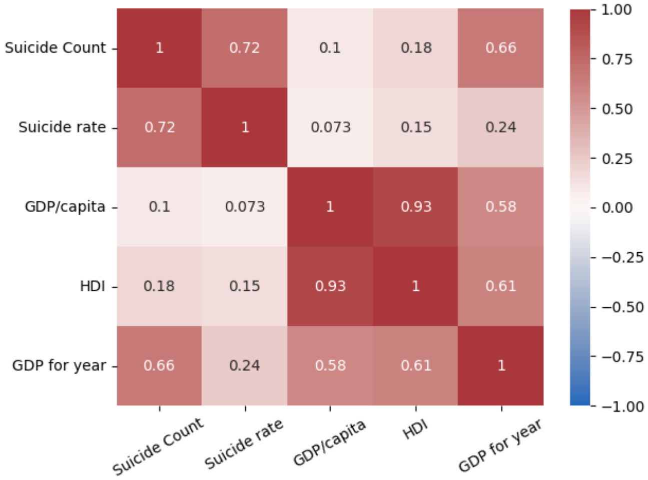

Figure 1: Matrix Heatmap of all continuous variables in the dataset

Before the interpretation of the matrix can begin, there are some important factors that need to be highlighted. Firstly, correlations between similar variables should be excluded. For example, suicide rate is a result of suicide count, therefore it should be excluded from the analysis because they are already extremely similar. Furthermore, the correlation between GDP for year and GDP per capita is excluded because GDP/capita comes as a result from GDP for year. Lastly, the correlations between these continuous variables are unorganised and have not been gone through to gain effective results. This will be explained further in a section below.

There are two points then which become of interest once these factors have been eliminated. To begin, the correlation between GDP for year and Suicide Count have a value of 0.66, which isn't necessarily the strongest correlation but does warrant further investigation. Perhaps there were some other factors which impacted this correlation; hence this was one of the correlations chosen. Additionally, the coefficient between GDP per capita and HDI had a value of 0.93, indicating a very high correlation. This was a very strong correlation, and was most likely because HDI does take into account wealth of a country among it's other factors, therefore it makes sense that such a correlation is strong. However, income is only one of the many variables involved in calculating a nation's HDI. Therefore, it was deemed interesting to explore. While these correlations were shown to exists mathematically, it would still be important to look at each correlation individually in graphical notation. To do this, two scatter plots were created to show a relationship between the variables.

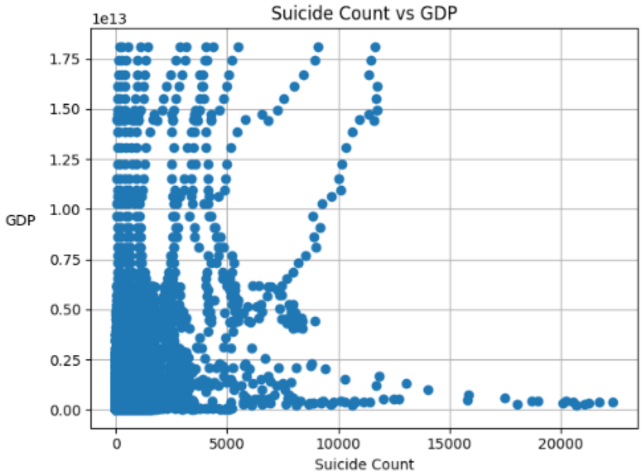

Figure 2: Unorganised data put into scatterplot, showing no sign of correlation between GDP and suicide count

As mentioned above, the data was unorganised and was taken directly from the dataset, to find correlation coefficients. This created a problem because although there was a correlation mathematically, it was not the same for it visually. In figure 2, the data is all over the place. There is no actual correlation to be found graphically, which meant that although there was a correlation of 0.66 found on the matrix heatmap, it was not confirmed by the scatterplot. To investigate further, the data for the scatterplot was computed in a different manner. Instead of taking in raw data, the GDP of every nation would be stored for every year, and then the average suicide count with be taken for that year, and then plotted on the scatterplot. By collection and combining the two variables this way, the correlation coefficient rose to 0.804, and the graph presented showed a more visible correlation than before. The two correct scatter plots are visible in Figures 3 and 4.

Mainly, what can be found from each scatter plot is that generally, as variable x increases, so does y. However, this relationship is more prominent in the scatterplot between HDI and GDP per capita. The relationships found are mostly non-linear; this is especially visible in figure 4. This indicates that the change between x and y are either slow or more rapid-like. In figure 4 the change in x causes a slow

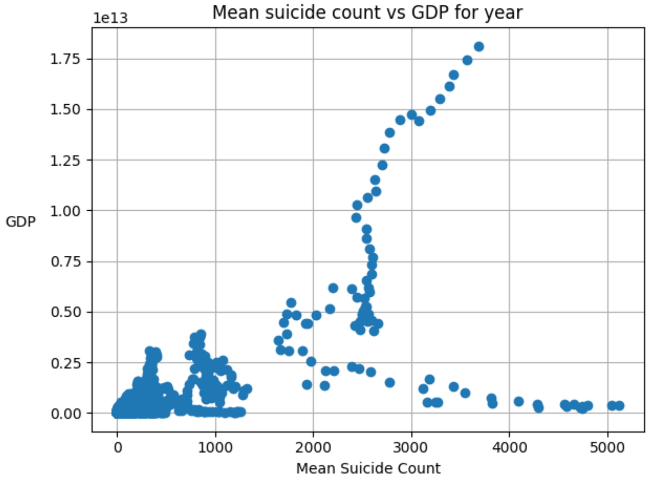

Figure 3: Scatterplot showing Mean GDP vs Mean Suicide Rate of a nation (Each year having one point)

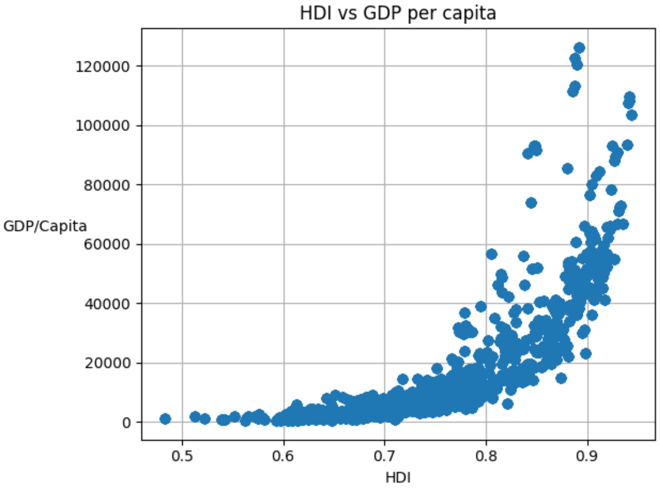

Figure 4: Scatterplot between HDI values and GDP per capita, demonstrating a slow increase as HDI increases

increase in y. This means that the relationship could be determined an exponential growth pattern. There is a slow start, then a rapid change. What this implies in terms of data is that while the HDI is lower, the GDP/Capita will also stay lower, until around 0.8 which is usually at the value when nations are seen as developed. Figure 3 however, takes a more unique shape. In figure 3, the graph splits up into two directions.

All the data is clustered in the beginning, but then the GDP increases, yet only with slight increases in suicide counts, whereas in the other direction GDP remains low while average suicides remain higher. It cannot be exactly confirmed that a correlation exists between these two variables, because there could be other factors impacting this data. One country has several points, which means that these patterns could be a result from one nation only over the years. Furthermore, GDP for year does not necessarily take into account individuals and instead groups up income of an entire nation, not per capita. This would mean that a more detailed analysis would need to be performed on this area alone to get more clear results, even after the data was more organised. However, if only one of these splits is taken, for example the increase, then it can be stated that some correlation exists between GDP for year and suicide count. As such, the correlation of determination may be applied to measure what percentage of variance in one variable is explainable in the other.

The two Spearman correlation coefficient that were collected were 0.93 for HDI against GDP per capita, and for Mean Suicide Count vs GDP for year it was 0.804. The correlation of determination is the correlation squared. For HDI against GDP/Capita, the determination is 0.864 which means that there is a strong correlation as 86.4% of all values in one variable are explainable in the other. This can be confirmed by the scatterplot, as generally the values increased in a similar pattern, until the very end. Values outside of the pattern may be those 13.6% that are unexplainable. Meanwhile for Suicide Count vs GDP for year, the correlation of determination is 64.6%, which meant that slightly over half the values are explained in variable one from variable two. Again, this number corresponds to the graph, because the split means that only roughly a half of the values are correlated in some way. However, the 64.6% does imply that the correlation might not be as strong as it was originally thought. As said, more research would need to be put into the correlation between Suicide Count and GDP for year.

To conclude on the findings in this section, the question investigated correlation between the continuous variables found in the dataset. Certain correlations between related variables (suicide count vs suicide rate and gdp for year and gdp per capita) were removed as they were already related. Predictions involving the question stated that some correlation would be present between developmental measurements such as HDI and GDP (either), while also stating that such measurements would also have some correlation on the suicide statistics. The investigation into this question deemed that there was sufficient evidence to support the premise that a relationship does exist between HDI and GDP/Capita, as was initially seen from heatmap, and then confirmed via scatterplot. However, what was not expected was the positive correlation between Suicide Count and GDP for year. While some evidence was collected, that showed such a relationship; through the coefficient of determination, it was concluded that there was some evidence to support the correlation, but further investigations would be required to come to more accurate conclusions. Overall, the statistical evidence gathered in this section shows that various correlations can be found between the continuous variables, with some being more prominent and stronger, than others.

3.2 Question 2

Hypothesis: It is predicted that males will have a higher number of suicides, compared to females. While males and females are each subject to upholding certain standards in society, causing stress and unnecessary expectations; females are offered larger support by society. Males are thought to emphasize independence, strength and individuality. This may be especially prominent throughout different generations. Due to amount of data, it is predicted that the younger generations will have less numbers of suicides recorded, while older generations will have more. However, the middle generations are predicted to have larger numbers because not only is there more data available about them, but also because the time periods these generations lived through cause difficult situations and hard times on individuals. Regardless, it is overall predicted that there will be a significant difference between males and females, regardless of generation.

To begin, a multibar chart was created out of these results, to view an overall difference of the total numbers which was done through all generations. Not only can this demonstrate the distribution of total number of suicides, but also allows for a comparison to be made. Furthermore, by having a bar chart for each gender, it is possible to see the actual numbers of total suicides, meaning total population suicide count against generation. To be noted, the y-axis "Suicide Count", is reduced by 100,000 for readability purpose Each y value should be multiplied by 100,000 for an accurate valu

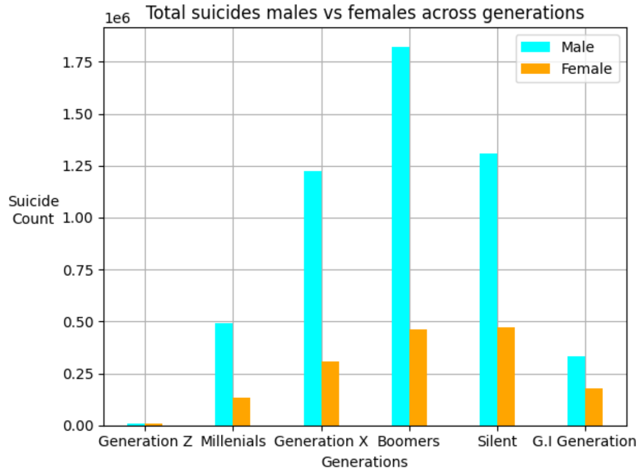

Figure 5: Total males vs females suicide count per generation

It is clearly visible that the figure shows two distributions of suicide counts in terms of male and female. The Male category had a symmetric distribution, while the Female category witnessed a more negative skew, skewing to the left. The figure immediately demonstrates that the total number of male suicides are vastly greater in all generations, except in Generation Z (Males – 9765 vs Females – 6141). The lack of difference, or number of suicides within this area could be a result from insufficient data due to being the youngest generation. For males, the "Boomers" generation had the highest number of suicides, standing at 1,823,530, whereas for females the "Silent" generation was the highest (with Boomers close behind) at 472,289. The graph in the figure above shows that a difference can clearly be visible between not only the male and female genre, but also specifically through generations. Males are predominantly the ones who have a higher tendency to commit suicide than that of females. Interestingly, the later generations are those with the highest values; such data might imply that this generation of people might have suffered through the most difficult times. However, it could be a result of these generations having most data points. Ultimately, with this graph there is definite reason to believe a difference exists between males and females regardless of generation. However, distributions of these can be further analysed to reach a more concrete conclusion with more data supported evidence.

Box and whisker plots were determined to be used which would best complement the bar chart in the figure presented above. Each generation contains a boxplot, for each sex. However, the boxplot in this case would be comparing all data points of the variable "suicide rate per 100,000 people". Not only does this provide an overview on distributions within individual pieces of data, instead of the total number that figure 5 presented, but also potential outliers and other information which may seem interesting. The two box plots can be seen in figures 6 and 7.

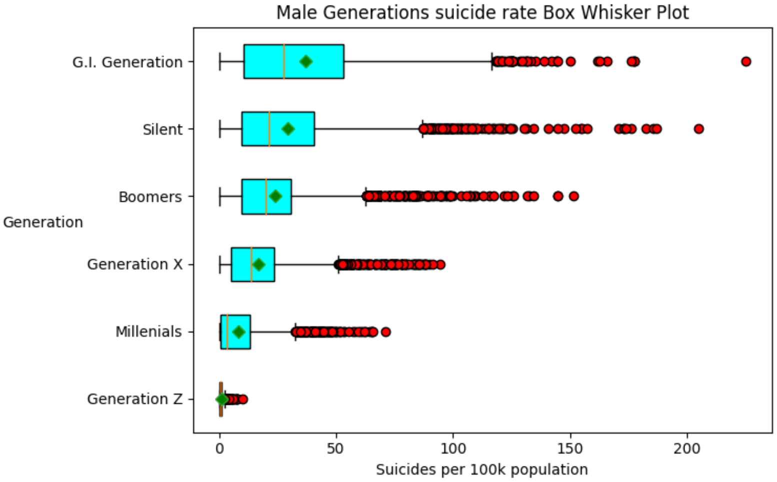

Figure 6: Boxplot containing distributions of male generations

As can be seen from the boxplots, there are several pieces of data (red circles) found outside the upper fence. While these points would by regular definition be considered outliers, in this analysis these points may be considered as abnormal observations. They are still part of the data, and represent important information. Perhaps, what could potentially be considered outliers are the points above 200 suicides per 100k population in the males box plot, and those over 100 suicides per 100k population in the female boxplot. Virtually all means across all plots are larger than the median, indicating that each distribution is skewed positively to the right. Each point beyond the fence is increasing the mean, and causing such skewed distributions. However, due to the size of the total records collected (over 24000 rows from the dataset) removing such a small number of data points would be trivial. Additionally, it would be redundant as the point of this section is to look at differences between the female and male plots.

From comparing the two plots above, it is possible to witness several crucial observations. Firstly, in this box plot the G.I. Generations had the highest numbers. The contrast between the two

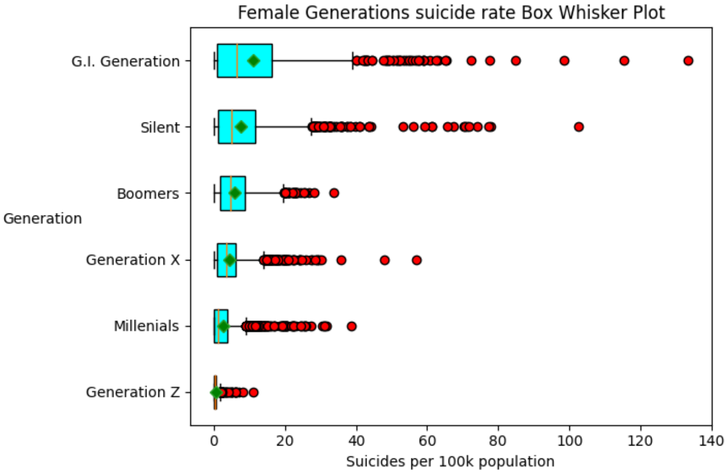

Figure 7: Boxplot containing distributions of female generations

generations is visible as the upper fence of the female G.I. generation is less than half that, almost one third of that of the male G.I. generation. Secondly, the overall interquartile ranges are larger with males, meaning that the data points are more spread out than the females. Furthermore, these ranges are always larger in male generations than that of the female generations, indicating that the 50% of suicide values are generally higher in males. Finally, when it comes to the abnormal observations mentioned, males can be seen having a higher number of these points beyond the upper fence, which leads to more extremes. For comparison, the largest value for the female G.I generation was close to 140, whereas for the male G.I generation it was beyond 220. The mentioned observations are clear demonstrations that visually, the data indicates a difference between males and females in terms of suicide. What can be further confirmed, is whether this difference can be deemed significant.

To test whether a difference exists and to determine it being significant, a null hypothesis is presented between each generation. H_0 is the null hypothesis, which states that there is no difference between means. H_1 being the alternate hypothesis, states that there will be a mean difference between suicide numbers. This way, a significant difference can be determined mathematically. Table 1 holds the values for the means and standard deviations of all generations, which will be the input values for the formula:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

2-Sample t-tests were carried out, with a t-test being done between every generation between the sexes (e.g. Female Generation Z vs Male Generation Z). To be noted, this was done at a significance level of 0.05/2 = 0.025, which would mean a z value of ±1.96 from the z-table. Any value that surpassed or was under this value means that it does not fall under the null hypothesis, and would need to be rejected due to insufficient evidence. Another table filled with results can be found below.

Table 1: Male and Female generations containing the mean (x), Standard Deviation – SD(s) and Total sample number (n)

Sex	Generation	Mean	SD	Sample
Male	Generation Z	13.29	31.77	735
Male	Millenials	167.64	578.44	2922
Male	Generation X	381.7	1084.97	3204
Male	Boomers	730.87	2047.74	2495
Male	Silent	411.52	1160.87	3182
Male	G.I. Generation	242.97	637.06	1372
Female	Generation Z	8.36	18.27	735
Female	Millenials	45.73	131.26	2922
Female	Generation X	96.7	271.91	3204
Female	Boomers	184.76	457.38	2495
Female	Silent	148.43	405.62	3182
Female	G.I. Generation	128.76	357.42	1372

Table 2: Table containing Z-scores and conclusions between generations

Generation	Z-Score	Critical Range	Conclusion
Generation Z	3.65	±1.96	H_0 Rejected
Millenials	11.11	±1.96	H_0 Rejected
Generation X	14.42	±1.96	H_0 Rejected
Boomers	13	±1.96	H_0 Rejected
Silent	12.07	±1.96	H_0 Rejected
G.I. Generation	5.79	±1.96	H_0 Rejected

As can be seen from the data, every value of Z-score lays in the extreme z-score of the right tail, with some values having extremely high z-scores. Although the numbers here are high, it makes sense because in the boxplot it was possible to see how heavy tailed the data was distributed, which impacted the standard deviation value. Furthermore, the large amount of data had further affected the results as the denominator became more minuscule. The abnormal observations additionally add onto this. All null hypotheses are rejected, due to a lack of sufficient evidence. What can be concluded, is that there is a definite significant difference between males and females, across generations. As has now been confirmed mathematically, through the 2-sample t test.

To summarise, this section explored and investigated how a difference may be found between the suicide statistics on males and the suicide statistics on females. The prediction mainly stated that males would be seen having larger numbers, while furthermore emphasising that the generations will see differences not only because of social factors, but because there is more available data presented. As table 1 saw, Generation Z only had 725 entries for each sex, whereas Generation X had 3204. Regardless of this difference, it was possible to see that in terms of total numbers the later generations, including Boomers and Silent, witnessed the highest number of suicides. In every generation in both number of suicides and suicides per 100,000 people, the male categories always ranked higher, which indicated visually from the data that there is quite a difference. The data from the boxplot showed distributions which were heavily affected by abnormal observations. As such 2-sample t-tests were used to mathematically confirm a significant difference. All null hypotheses were rejected, which meant there was sufficient evidence to conclude that a significant difference between males and females does exist.

3.3 Question 3

Hypothesis: Generally speaking, what is to assume to come up from the results of the analysis of question 3 is that suicide rates will be seen decreasing. Due to the large range of years chosen to investigate questions, there will be some form of decrease because the societies will see improvements in various areas in terms of medical care, life expectancy, quality of life, and other factors such as knowledge/education among the people. The higher ranking such factors are, the better a society is predicted to represent its people, hence leading to less struggle, and more opportunities for a person to be helped. Additionally, perhaps there could be some major events through the years that cause some rises such as the finance crisis in late 2000s.

5 nations were selected that would vary across HDI index. In 1996, the number of nations that were ranked on the Human Development Index was 148. These 5 countries would vary and at least have a difference of 20 rankings between each other. The countries selected were as follows in order of ranking (in 1996): Norway, Lithuania, South Africa, Thailand and Guatemala. Due to the lack of available information on HDI from the available dataset, HDI rankings for the listed nations were collected online from the United Nations Development Programme, Human Development Reports [4]. To note, every country had different ranges for their available data. For example, Lithuania only had information available from 1996, meaning that there was no point in taking data before this era for other countries (if a fair analysis was to be done). As a result of the range was set to 20 years, ranging from 1996 to 2015. Using the information from the reports, 4 pieces of information were collected. These pieces of information can be visible in the table below.

Table 3: Year 1996 vs 2015: HDI values and their corresponding HDI ranking

Country	1996 HDI	1996 Ranking	2015 HDI	2015 Ranking
Norway	0.887	2	0.947	1
Lithuania	0.719	46	0.862	34
South Africa	0.65	73	0.701	108
Thailand	0.622	84	0.749	86
Guatemala	0.518	124	0.652	124

The most important factor when selecting the countries was the HDI ranking. Nations were attempted to be picked which were available in the given dataset, while also being somewhat distributed across the HDI rankings. Spread out choices included looking towards higher, middle and lower ends of the spectrum. For this analysis, the contents in the table above were crucial as they would act as a reference point for how a nation has developed across two decades, without direct implication to number of suicides or suicide rates. What can be seen is that, although some nations fall down the actual ranking, it's critical to distinguish that the HDI increases everywhere. This implies that the standard of living had improved over the course of the years. The next step would then be to demonstrate that the suicide rates have also gone down as a result of a developing society.

Since more than one group is being used, a line chart was chosen to visualise the data tracked over two decades. To expand on the time period, having more time allows for the production of results which can lead to more evident and supported conclusions. 20 years was deemed a good amount to demonstrate some form of change over time. The data was organised and collected into individual years of every country, combining the suicide rates and getting their mean average. By plotting every group onto one chart, it is possible to see how countries of lower HDI differ from those with a higher HDI. This can help determine whether there has been any improvement over the last two decades, or if suicide rates had remained constant or even increasing.

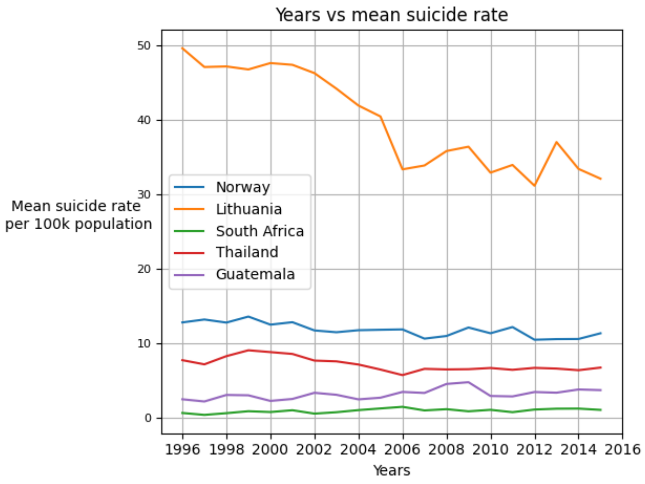

Figure 8: Line chart plotting mean suicide rates over the course of two decades

The prominent factor of the line chart above shows that Lithuania had always remained above other countries by quite a fair amount. This point is brought up because if referred back to the table, Lithuania was on the 34th place in terms of HDI ranking. Furthermore, Norway which was ranked 2nd in the HDI ranking is seen above all other nations. These two lines indicates that suicide rates might not necessarily fully represent a nation's development accurately, and it can be stated that other societal factor's may be involved that drive these numbers up; some that aren't considered such as the psychological mentality of populations (Is the person happy?). Although the main premise that the question does focus on, was improvement overtime. How have suicide rates changed throughout the course of two decades. Although it may not be entirely clear, most countries do witness numbers going down, even if gradual.

From the data collected, the results are present in the tables 4 and 5. First table looks at the suicide rates roughly every 10 years. From this, it is possible to see how data has changed over time. Taking into account HDI, it seems that more developed countries, such as Norway and Lithuania, saw decreases in suicide rates at every interval. Thailand however over also lowered this number over the course of 20 years; however, it is still possible to see that it had lower rates in other years. Table 5's values show that Thailand had its lowest suicide rate during 2006. Such values indicate that although HDI may not always largely change the values drastically (as was the case with Lithuania), it could mean that the nation is less stable. Hence, the suicide rate is also not completely stable and tends to have such uncertain fluctuations. For Thailand's case, other factors such as the 2014 coup d'état may have had some impact on values in 2014 and 2015. Ultimately, what the numbers indicate is that more developed, and higher-ranking nations on HDI, show more stability and in turn see decreases in their suicide rates.

Table 4: Countries and mean suicide rates between 10-year intervals

Country	1996 Rate	2005 Rate	2015 Rate
Norway	12.875	11.781	11.369
Lithuania	49.34	40.483	32.123
South Africa	0.669	1.274	1.083
Thailand	7.754	6.499	6.763
Guatemala	2.504	2.718	3.738

Table 5: Maximum and Minimum mean suicide rates for every country, accompanied by year of occurrence

Country	Max	Min
Norway	13.6 (1999)	10.5 (2012)
Lithuania	49.634 (1996)	31.177 (2012)
South Africa	1.274 (2005)	0.413 (1997)
Thailand	9.086 (1999)	5.753 (2006)
Guatemala	4.798 (2009)	2.209 (1999)

This can additionally be confirmed by values in table 5. The values for Norway and Lithuania experience their maximums during the late 90s, whereas other countries experience their maximums later into the years, showing that even though the HDI generally increased, the suicide rate remained growing overall. Thailand's minimum was found at 2006, which shows that there was stability during the mid-2000s, but some other factors caused the numbers of suicides to increase. Overall, the decreases were found to be slight in the case of Norway (11.7% decrease), and Lithuania seeing the most prominent decrease at 34.9%. Thailand, remained at almost the same HDI ranking, yet became the middle ground in terms of HDI by 86, which can be demonstrated from its general decrease (12.8%) in suicide rates. On the other hand, less developed nations such as South Africa and Guatemala saw increases in their suicide rates. Each of these two countries dropped in their ranking, despite the HDI increase. South Africa witnessed a 38.2% increase, and Guatemala witnessed a 33% rise in suicide rates. These two values and changes over time show that HDI rankings and values contain some evidence for how suicide rates may be interpreted from them.

In essence, the data visualisations presented allow for the interpretation of suicide rate change over time. The values and numbers extracted supported the thought that more stable, and developed countries saw decreases over the 20-year course, despite them having the highest rates. Whereas countries with a lower ranking and HDI saw increases in suicide rates by a third or more, showing that a nation's growth in terms of HDI value may also be reflected by the numbers in suicides to some degree. All countries selected for investigation saw increases, however even with these increases it did not show the decrease the hypothesis was hoping for. Most countries' suicide rates remained stable, with some slightly decreasing. Lithuania was the only data point which witnessed major decreases, while also having one of the greatest HDI value increase. Contrary, Guatemala also had a large HDI increase, but still had an overall increase in suicide rates in 2015. Indicating that HDI values themselves are not completely accurate, and their change should also be monitored. In terms of external events, there wasn't sufficient evidence which supported the idea that other factors or major occurrences caused suicide rate fluctuations. Further research would need to be conducted to reach any concrete conclusions on external events. Nations standards of living, health and education systems have increased showing social progress; the data extracted somewhat supports the idea that the more a country improves and develops, there will be at least some slight decrease in suicide rates.

3.4 Limitations

Each section, while providing results, can be subject to speculation. Questions are explored and investigated in what is the best thought way, however there are certain aspects which may go unnoticed or unaccounted for in the report. Hence, these limitations must be looked into and brought up to discuss the validation of results.

For question 1, the correlations were gathered and then two were selected. The GDP for year and Suicide Count data needed to be reorganised in a way that would better show some visual relationship between variables. Although the data used was confirmed to give the correct number of actual results, there could have been potential mistakes in the code that was used to calculate the actual data. It could be inaccurate, despite checking over it. Furthermore, the HDI vs GDP per capita correlation was found to be strong, but it should be considered that many HDI values were missing, which meant that a lot of the rows that did not include the HDI values needed to be omitted to calculate the actual correlation and create the scatterplot. This exclusion of data points may have led to an incomplete analysis, as including the entire dataset values could have brought about different results. Therefore, taking all this into account, there may still lay certain features of this statistical analysis report which may be subject to reviews and readjustments.

There are certain aspects which should be considered for question 2, which include amount of data available to each group, and the existent outliers that were deemed abnormal in the box and whiskers plot. Firstly, the groups had various amounts of data, which meant that despite some mention of this, it would ultimately mean that the comparisons between generations are less informative because such factors imply an unfair comparison. Some bias would be included as one generation is given more information than the other. However, it does not impact the overall difference between males and females, since it is still visible despite generation. Furthermore, the abnormal observations were chosen to be included as part of the analysis, however the report does not go into a solid discussion of these point. A further discussion would be necessary to determine how the data is impacted at a greater level, and what this means.

Finally, in regards to question 3, there are some things to highlight which may be interpreted as limitations. Firstly, the choice of a nation's selection was not completely at random. Originally, it was decided that nations would be based on HDI placement, however due to a lack of available data or none-existent entirely data, several nations had to be cut from this choice, leaving a less spaced-out end choice. This means limits are placed in regards to cultural background, and other environmental factors which better play into how society works across different regions. Furthermore, as mentioned although some had lack of data, Thailand was missing the data from the year 2001. Several sources were analysed, with various numbers appearing. The number chosen was based on an academic article, "Suicide in Thailand during the period 1998-2003" as it was deemed most credible. [2] Nevertheless, this piece of information may not be an accurate number compared to the dataset used, thus impacting the conclusions on Thailand's development.

4 CONCLUSION

To conclude on the findings of this report; suicide number and rates are important to look at because they can demonstrate possible issues and unfairness in a society. These numbers represent a problem that needs investigating, so specific measures may be installed as a result to attempt to decrease them and improve the life of all people. The dataset that was explored in this statistical report was posted on Kaggle in 2018, and spanned across the years from 1985 upward of 2016, giving large amounts of information on every nation for many years. To extract information from this data, three questions were generated, aiming at analysing correlations, differences among sexes and generations, and how statistics can help visualise a society's development. It was found that a strong correlation and a moderately strong correlation existed between HDI - GDP/capita and Suicide Cunt – GDP for year respectively, implying that various socio-economic factors impact each other and actual suicide numbers. Moreover, there was significant evidence presented mathematically, through rejection of all null hypotheses, that a difference existed between males and females spanning across all generations. Lastly, data collected across 5 picked countries demonstrated how changes in HDI and improvement of a nation's development has some affect on suicide rates. The statistical analysis is however subject to some speculation on potential limitations and hinderences. Keeping this in mind is crucial as conclusions may have overlooked potential mistakes. Irrespectively of such, data analysed exhibits correlations, differences and socio-economic developments. Fortunately, over the course of the years the general number of suicides has decreased overall.

REFERENCES

[1] Tony Blakely, S Collings, and J Atkinson. 2003. Unemployment and suicide. Evidence for a causal association? *Journal of epidemiology and community health* 57 (09 2003), 594–600. https://doi.org/10.1136/jech.57.8.594

[2] Manote Lotrakul. 2006. Suicide in Thailand during the period 1998–2003. *Journal of Neuropsychiatry and Clinical Neurosciences* 60 (2 2006), 90–95. https://doi.org/10.1111/j.1440-1819.2006.01465.x

[3] World Health Organisation. 2019. Suicide worldwide in 2019: global health estimates.

[4] United Nations Development Programme Human Development Reports. 2020. Human Development Index (HDI). https://hdr.undp.org/en/indicators/137506

[5] Rusty. 2018. Suicide Rates Overview 1985 to 2016. https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016/metadata