Suicide, Happiness, and Non-Socioeconomic Variables

Team Mountain Tiger: Katey Forsyth, Chris O’Neil, Tucker Paron, Nolan Jimmo

12/15/2021

Abstract

Suicide has always been a tragedy across the globe that has touched everyone in some way. Currently, amidst a global pandemic, while all members of our society may be struggling mentally in their own way, mental health issues are at the forefront of our minds. Often, suicide is the culmination over a long period of time of someone lacking the support or professional help that they need in order to combat those struggles. This research endeavor aims to show evidence of factors that have not historically been present in the process of identifying and helping those in need of clinical mental support. One goal of this project is to potentially identify human factors beyond the current, often strictly socio-economic, criteria that can indicate a person being at higher risk for mental health struggles and ultimately suicidal thoughts/actions.

Introduction

In a time where many people have been forced to face mental health struggles because of a global pandemic, it has become abundantly clear that while suicide is something that has continued to improve in treatment and prevention, there are still massive institutional and cultural gaps that need to be filled in order to better assess and identify those at risk of, or who are currently experiencing, suicidal thoughts/actions. In many cases, even if a person actively visits an emergency department or recovery institution, the current standards for assessing a person are strongly influenced by the fact that a person does or does not have a suicide note, or an actionable plan (Hall et al, 2011). While it may be the current standard of the field, the idea of "we can't help you because a bad thing hasn't happened yet" is an extremely flawed mentality that requires research and proof of invalidity to begin to change. The following data analysis of non-socio-economic factors and how they relate to suicide counts across countries over the 10-year period from 2005 to 2015 hopes to shed light on new, innovative ways to better identify people who may be at risk of, or experiencing, suicidal thoughts/actions. By potentially curating a list of more inclusive factors, this research could potentially help current mental health institutions better identify people who need help.

Data

The dataset that was used for the base suicide data is a dataset from Kaggle called "Suicide Rates Overview: 1985 to 2016" that was contained in the csv called master.csv. The second data set, called "Life Expectancy (WHO)", was contained in the life_expectancy.csv. The third data set was a World Happiness report from the Sustainable Development Solutions Network that was contained in the happiness_report.csv. The data that was extracted from master.csv was the data on suicide counts and population. These counts were broken up by country, year, sex, age, and by generation. The data in life_expectancy.csv and happiness_report.csv was only characterized by country and by year; the sex and generation characteristics had to be collapsed into just country and year for the population and suicide counts to match corretly (`data_wrangling.py`, lines 50-56). Also, the range of data in life_expectancy.csv only had complete data from 2005 to 2016 so the master.csv was subsetted to those years (`data_wrangling.py`, lines 47-48) The happiness_report.csv was subsetted to just look at the year 2015 because the analysis will be comparing variables in the happiness_report with statistics from the other data sets for the year 2015 (`data_wrangling.py`, line 60). Once these changers were complete, it was possible to merge master.csv with the life_expectancy.csv. There was originally a population column in life_expectancy.csv but large discrepancies were found in this column. Some years countries had hundreds of thousands for their population, then the next year had millions, and then went back down to thousands. With this knowledge the population column was dropped and replaced with the newly generated population column from master.csv (`data_wrangling.py`, lines 80-98). With this newly merged data frame, a new data frame was created by subsetting the newly merged frame by the year 2015 (`data_wrangling.py`, line 101). The new data frame was then merged with the happiness_report dataset (data_wrangling.py, line 101). During the merging of the master.csv, the life_expectancy.csv, and the 2015 frame with happiness_report.csv, countries differed between the data sets, the countries between them were compared with python set data-types (example of this in data_wrangling.py, lines 183-189). Once the countries that didn't match were found they were dropped from the data frames. Once all of this was completed there were two data frames; one that contained all of the suicide counts along with the variables from life_expectancy and another data frame which only had data from 2015 and was merged with the happiness report data from 2015. The variables in both data frames were then explored with histograms to see if any variables required cleaning or if there were any large discrepancies in each (data_wrangling.py, lines 131-157). It was found in the dataframe containing data about only 2015 that the columns alcohol, total expenditure, and percentage expenditure were either missing large portions of data or were completely empty, so they were dropped from the frame (data_wrangling.py, lines 160-165). For the data frame containing data from 2005 to 2016 and no data from happiness_report.csv, it was decided that countries and columns would only be removed if necessary because there was a large loss of data points when merging because of the first step in the data extraction in master.csv. With this knowledge the countries with massive missing values in nearly all columns were dropped from the frame (data_wrangling.py, lines 175-187). After the exploration and cleaning processes were complete, the two frames were then saved to csv files for analysis (data_wrangling.py, lines 205-206).

Results

NON-SOCIO-ECONOMIC VARIABLES

The three main variables that were focused on in the early stage of the data analysis were alcohol consumption, average BMI across a country, and the total expenditure on health care as a percentage of a country's total government spending.The goal of the project was to use a number of analytical tools to show evidence that one or more of these non-traditional variables was related to the change in suicide counts per country over time. The over-time factor is extremely important to the analysis of data. All of the variable data is comparable to suicide counts because of this timeseries variable that links everything together and allows for relationships to exist over time.

Additionally, five countries and their corresponding variables were used as the sample for this part of the analysis. The plotting output that would have been produced from using every available country would have been disproportionately large and difficult to sift through. To combat this issue while still providing an accurate representation of the whole dataset, the sample of five countries is made up of two countries that had the highest suicide counts from 2005 to 2015, the country with the median suicide count, and the two countries with the lowest suicide counts over the given time period. Those countries are Japan, the Russian Federation, Finland, Grenada, and Seychelles (from highest suicide counts to lowest).

Figure 1 (below) shows the plots of each variables' count per year, plotted over the time period of interest, as well as a plot demonstrating the suicide counts per country over the time period of interest (data_analysis.py, lines 226-253). These plots are an extremely important first step because they show that the data clearly changed over time in some predictable way. Having a clear model of how the data is changing in each variable makes it easier to compare that model to the suicide data per country. For example, if both alcohol consumption and suicide counts increase in a linear fashion over 10 years in a country, then there is strong supporting evidence that alcohol consumption may be a factor that can help professionals get a better idea of who is at a higher risk for suicide and then accurately assess who needs help and in what ways.

As seen in the plots, not every country and variable combination has a strong model that represents the variables change over time. However, Japan, the Russian Federation, and Finland (the three countries with the highest suicide counts) show the most consistent changes over time across all variables. This fact suggests that in at least three of the five sample countries there is a distinct possibility of finding relationships between suicide counts and the variables of interest.

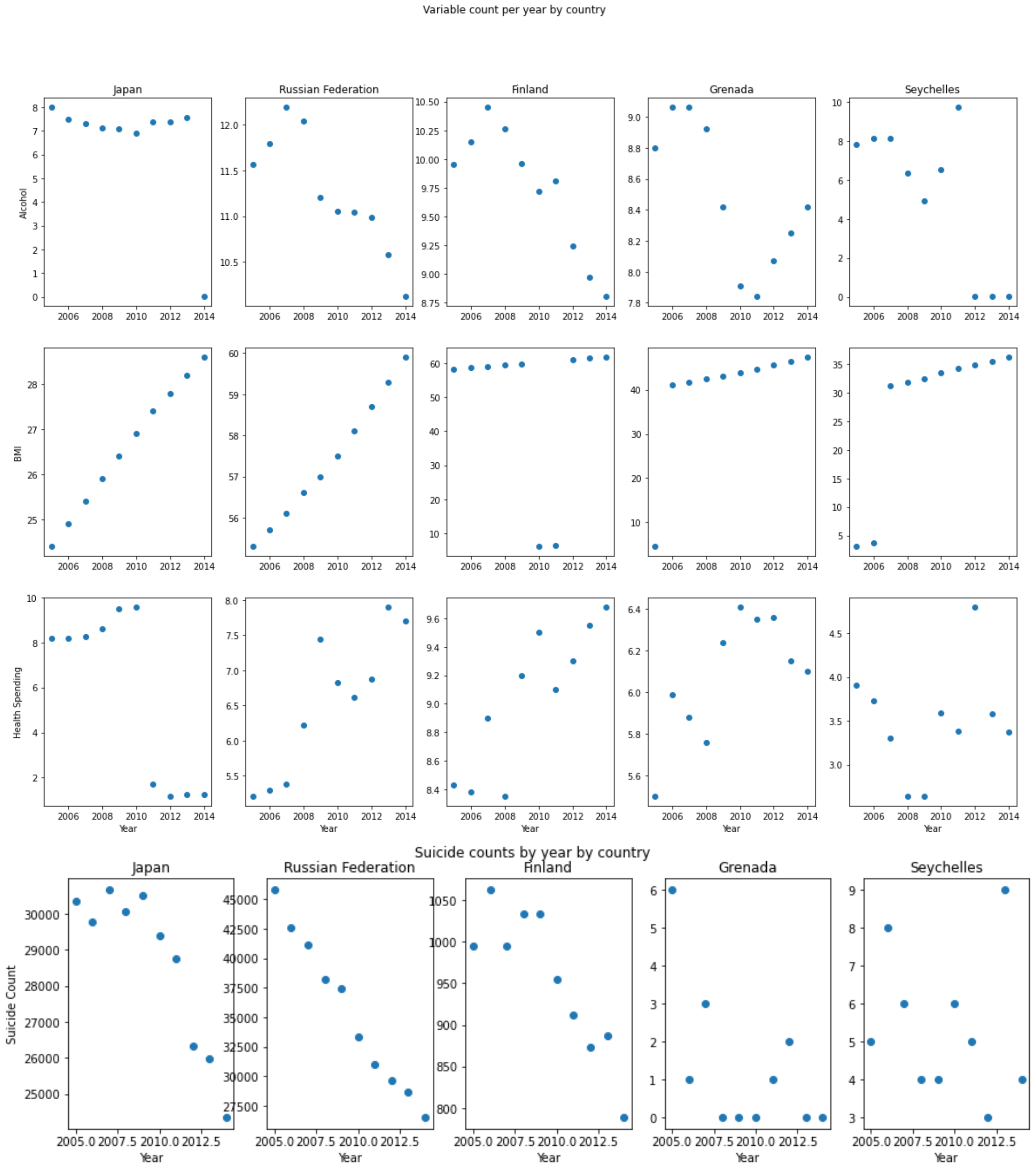


Figure 1 Raw plots of all variables over time. Alcohol is measured in liters, BMI in BMI index and health spending is measured in millions of dollars, USD.

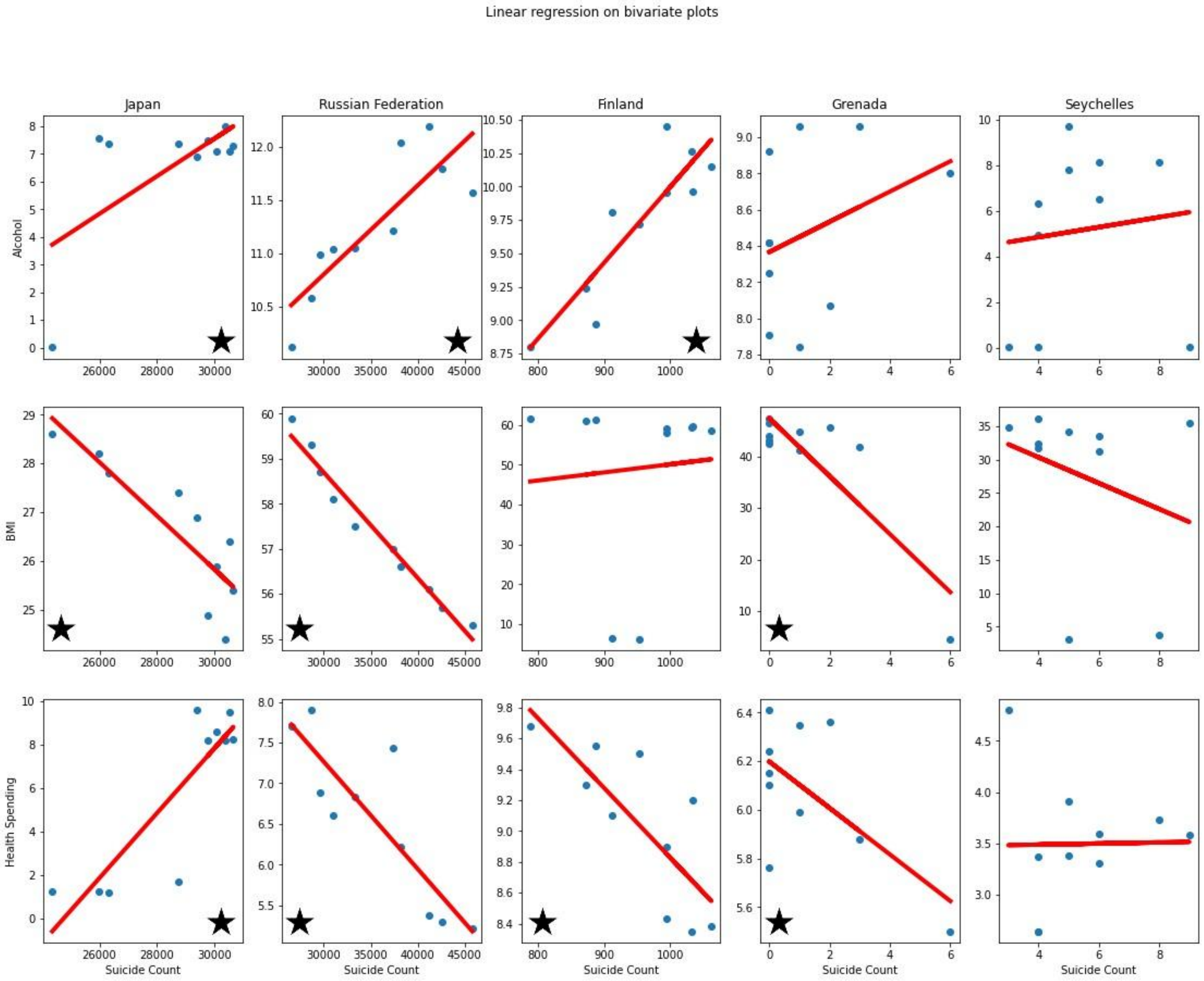


Figure 2 Linear regressions of bivariate plots with all of the country and variable combinations. Again, alcohol is measured in liters, BMI in BMI index and health spending is measured in millions of dollars, USD.

R values (Pearson Correlation Coefficients)

	Japan	Russian Federation	Finland	Grenada	Seychelles
Alcohol	0.64931	0.84118	0.88941	0.35058	0.11007
BMI	-0.85788	-0.98457	0.07836	-0.86431	-0.28548
Health Spending	0.86932	-0.86471	-0.78031	-0.63346	0.01640

Table 1 Table depicting r-values (a.k.a Pearson Correlation Coefficients) for the linear regression models in Figure 2. This is provided to quantify the significance of each regression.

Based on the confidence gained from the initial look at the five sample countries and their variable data, the decision was made to explore linear regression models. Linear regression was the correct method for this experiment because when observing Fig. 1, most variables that do have some resemblance of following an equation of some kind, follow a linear one (not all, but most). On top of that, because the data is based in the timeseries realm which changes linearly, it makes logical sense that linear regression would be a valuable tool in this circumstance.

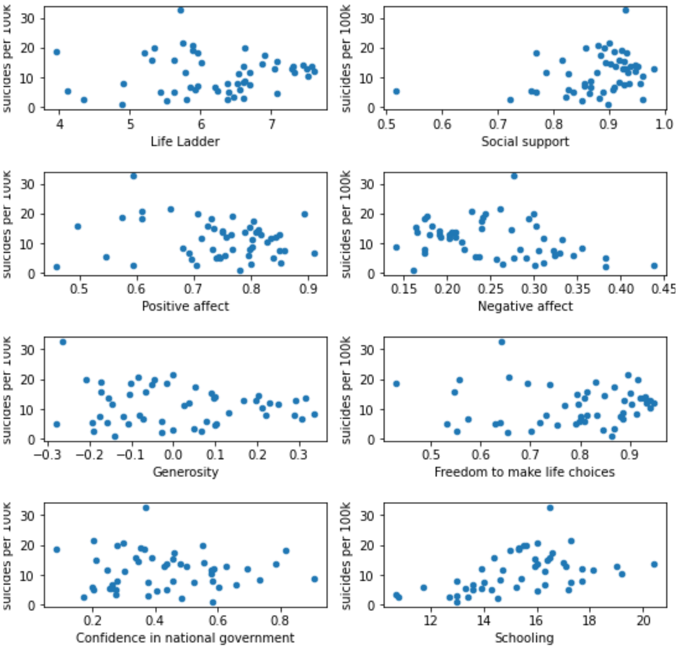
Looking at Figure 2, the choice to use linear regressions has paid off. Out of the 15 country/variable combination plots, 10 of them have absolute value r-values above .60 (denoted with stars), and eight of those have r-values at or above 0.80. Furthermore, Table 1 shows the specific r-values for each regression (`data_analysis.py`, lines 305-318).

Overall, the results are strong indicators that in at least some of the countries, the possibility of alternative risk indicators or behaviors are well-supported in the data and certainly deserve to be looked at further.

HAPPINESS VARIABLES

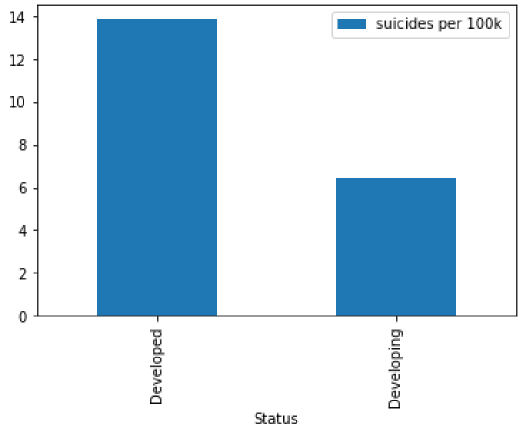
As previously mentioned, happiness is an alternative variable that was investigated for possible impact on suicide rates. Data from 2015 across fifty countries was observed. Given the subjective nature of happiness, it was recorded through several different metrics, mostly based on survey responses. Figure 4 depicts the correlation of these metrics with suicides per one hundred thousand people (within each country's population) (`happiness_analysis.py`, lines 21-43). The results were rather surprising with most happiness metrics yielding very weak correlation or none at all. Life ladder, for instance, a numerical measure that asks individuals to rate their percieved standing in life on a scale of 1 to 10, seemed to have no relationship with suicide rates (r ≈ .054; `happiness_analysis.py`, lines 29-30). Positive effect, an average of two yes or no questions about if the respondent 'laughed' or 'enjoyed a lot' in the day prior to the question, also had little to no correlation. Negative affect (similar to positive but asking about 'worry', 'sadness', and 'anger'), seemed to have a mildly negative relationship, yielding a correlation coefficient of -0.3077 (`happiness_analysis.py,` lines 35-36). While the relationship is far from substanstial, this suprisingly points to countries with more 'Negative' moods having lower rates of suicide. Similarly, social support, a yes or no question answering if the respondent has someone to 'count on' in times of need, had a mildly positive relationship with a coefficient of .2822 (`happiness_analysis.py`, lines 31-32). The rest of the variables had essentially zero relationship, except for Schooling which measures

Figure 4 Scatterplots of happiness metrics against suicide rates



the number of years an individual is in school (averaged for the country). This variable produced a correlation coefficient of .4714, indicating a weak positive relationship; however, while minimal, of all variables tested, this one had the strongest correlation to suicides per one hundred thousand people (`happiness_analysis.py`, lines 42-43). From this data alone, it is unclear if any of the suspected variables actually have any sort of relationship with suicide rates. To test further, the average number of suicides per one hundred thousand people was tested between developed nations and developing nations. This factor was calculated by summing the total number of suicides for each country in either group (developed and developing), then dividing by the sum of the total populations, and finally multiplying by one hundred thousand (`happiness_analysis.py`, lines 46-47). Figure 5 depicts how these calculations showed developed countries have nearly double the average number of suicides per one hundred thousand people (~14) than that of developing nations (~7) (`happiness_analysis.py`, line 37). Based on these analyses, it appears that the tradionally intuitive ‘causes’ of suicide have little to no effect on suicide rates and that developed countries, with higher incomes and more industrialization, actually tend to have higher suicide rates. Most importantly, the aforementioned happiness measures such as Positive affect, Social support, and Life ladder, seemed to have very weak or no correlation at all.

Figure 5 Histogram of suicide rates per 100,000 by development



OTHER VARIABLES

Potential relationships between sucide rates and other factors in the data were also explored. For single variable linear regression models, a correlation with suicide counts only existed with a few of these alternative factors. One such factor is the Gini of household income in the GWP (Gallup World Poll), which is essentially the average household income in a given country in international dollars. The coefficient in the linear model was 2,3870 ($p < 0.05$; `other_vars_analysis.py`, lines 126-144) with a Pearon’s r value of 0.30 ($p < 0.05$; `other_vars_analysis.py`, lines 126-144).

Another variable that correlated with suicide counts is the infant death rate. In the linear model, this factor had a coefficient of 261.49 ($p < 0.05$; `other_vars_analysis.py`, lines 126-144) and a Pearson’s r value of 0.36 ($p < 0.05$; `other_vars_analysis.py`, lines 126-144). A somewhat similar factor, the mortality rate for children under the age of five, also showed significant results. The linear model using this variable had a coefficient of 216.17 ($p < 0.05$; `other_vars_analysis.py`, lines 126-144) and a Pearson’s r coefficient of 0.35 ($p < 0.05$; `other_vars_analysis.py`, lines 126-144).

Multi-variable linear regression models were also tested but none showed statistically significant results.

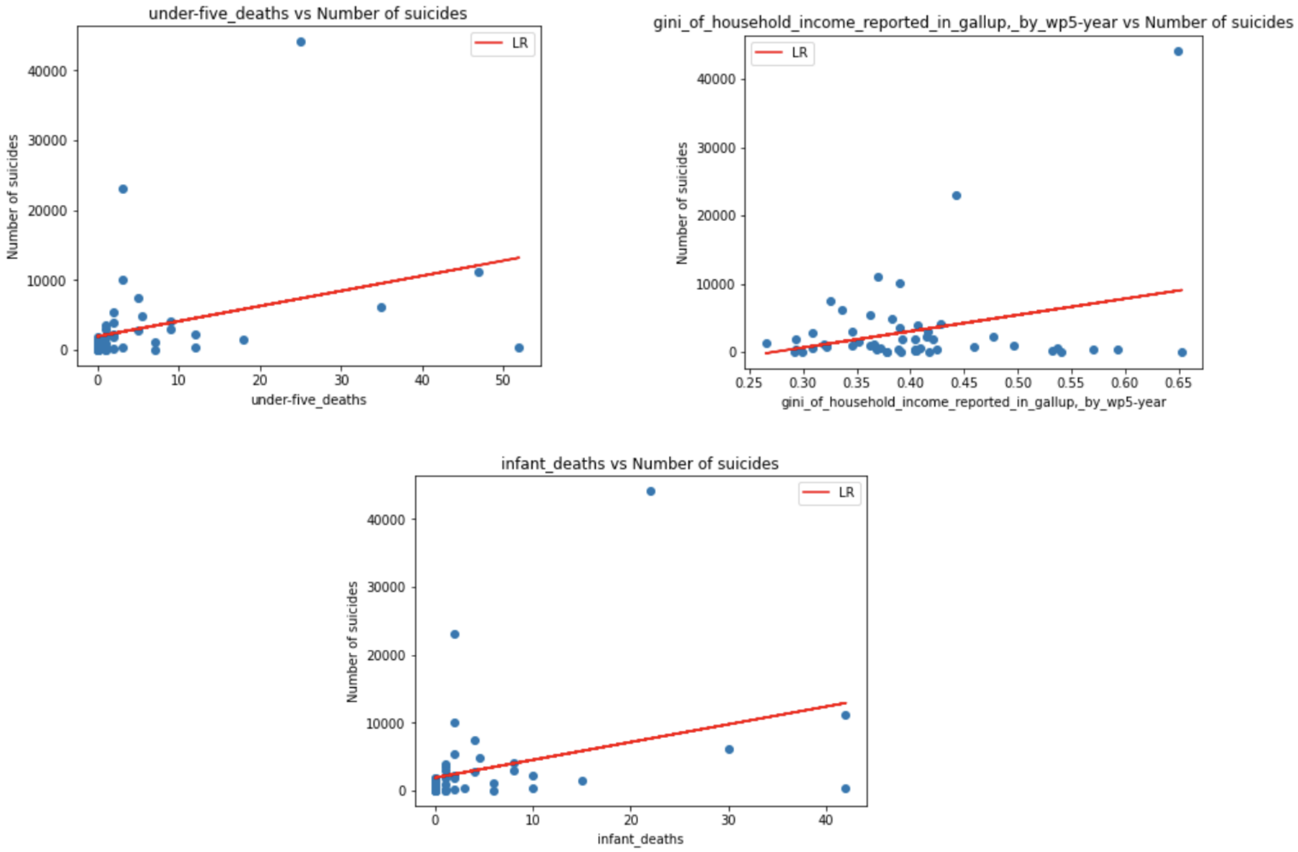


Figure 6 Graphs depicting linear model between number of suicides and average household income, infant mortality rate, and mortality rate for children under five

Discussion

NON-SOCIO-ECONOMIC VARIABLES

In general, with regards to the non-socio-economic factors, the analysis results seem to be quite strong. The main goal of the entire project really revolves around trying to find strongly correlated variables to suicide rates and, as shown in Figure 1 and Figure 2 above, there are a number of variables that have statistically significant correlations to suicide rates in a number of different countries. On top of that, it is worth mentioning that in the cases where little to no correlation was found between the chosen variables and suicide rates, there were very few suicide observations for the country in question.

Most important to talk about is the correlations between the highest r-value performers and suicide rates. In almost all three variables (alcohol, BMI and health spending), three of the countries have a strong correlation absolute value of .7 or greater. Among industry standards, a value of .7 indicates very strong correlation. Furthermore, in countries like the Russian Federation, every variable shows a correlation coefficient of .8 or above (data_analysis.py 305-318). These strong correlations are a sign that this research may have uncovered strong evidence for new, more accurate factors to consider when evaluating people for suicide risk.

The other significant factor to understand is that the sample consists of five vastly different countries. In the sample countries, there is an Eastern Asian country (Japan), a Central Asian country (Russian Federation), a Scandinavian Country (Finland), a Caribbean Country (Grenada) and an Eastern African island country (Seychelles) (`data_analysis.py` 112-115). This *gigantic* spread of countries and their cultures, population, government type, etc. makes it easy to assume that there are no similar factors across each of them that would indicate suicide risk, especially when it comes to non-socio-economic factors. However, as shown, in four out of five of those countries (all but Seychelles), there is at least one variable, highly correlated to suicide count, that is also present in the other countries (and in almost all four, there are at atleast two variables) (`data_analysis.py,` lines 273-318). That is to say that in many cases, regardless of culture, government type, population, economy, etc. the indicators of alcohol consumption, BMI and the health spending of a country are all highly correlated to its suicide count. This information can now be extrapolated and researched on smaller level such as between social classes, demographics, locations, etc. and tuned even further to the point where eventually, ideally, a physician or health profession can use a simple patient survey incorporating these factors (and others) to get an even better understanding of a person's needs, *especially* in sucide recovery or emergency room situations where a patient is actively *seeking* help and may historically have been turned away due to outdated criterion for suicidal thoughts/actions.

The last aspect of the non-socio-economic data is the fact that about a third of the variable/suicide count combinations seemed to not be significantly correlated. It is equally as important to try and understand why that may be the case, and consider if those influencing factors are more significant factors than those influencing the strongly correlated variables. In the case of this data, it seems to very much be a case of few observations influencing the correlations of variables. This behavior is also a product of an analysis design choice that was made to use two of the highest suicide count countries, a middle country, and then the two lowest suicide count countries in the dataset. In doing this, the idea is to use the three larger count countries as the main point of analysis because they have vastly more data points and much like flipping a coin, give better indications of global trends because the number of observations is so much greater. The two smallest suicide count countries are then used to put a 'spin' on those initial results. It is common for these small observation variables to not support the larger observation variables, but when they do, it gives even more validity to the larger observation results. Again, following the coin-flip analogy, if a person flips a coin 10 times, there might be seven heads, but the coin can't be labeled "biased" yet. It just has the possibility to act in a biased way at low observation values. If the coin continues to stay at a .7 proportion when observations are increased to 1000, then there would be far more justification in saying the coin is legitimately biased because higher observation values provide more stability in results.

HAPPINESS VARIABLES

With regards to the 'happiness' variables, it was evident from the plots that the metrics used to measure happiness were not good predictors of suicide counts by country. There is some possibility that the questions were interpreted differently through different translations for each country, therefore, somehow impacting the scores. Regardless, the absence of correlation is a valuable conclusion as it contradicts the intuitive hypothesis that 'happier' populations of people (by these metrics) are less likely to have high suicide rates. Rather, this data proposes that these measures of happiness have no impact on suicide rates. Additionally, countries with higher incomes and higher rates of industrialization (developed nations) actually have significantly higher rates of suicide than developing countries do. This result in particular would be a great point of future research as the specific variables that decide what constitutes a developed country versus a developing country could be tested.

OTHER VARIABLES

The results from the single variable linear regression model suggests that the average household income, the infant death rate, and the mortality rate of children under five have positive correlations with the number of suicides in a country. The coefficient from the model looking at average household income in a country indicates for each additional dollar in a country's average household income, there are an additional 2,3870 suicides per year (`other_vars_analysis.py`, lines 126-144). The fact this relationship is positive seems odd, considering one would expect financial struggles and all the stress that accampies them to increase the number of suicides. Therefore, an increase in financial resources would lead to a decrease in suicide rates. However, the model has a Pearson's r value of 0.30 (other_vars_analysis.py, lines 126-144), so this positive correlation is in fact very small and may just be a fluke in the data, despite being statistically significant.

The other two findings, regarding the rate of infant deaths and the mortality rate of children under five, look very similar to each other. The infant death rate has a slope of 261.49 (`other_vars_analysis.py`, lines 126-144) and the mortality rate of children under five has a slope of 216.17 (`other_vars_analysis.py`, lines 126-144). The implication that babies and/or child deaths has a positive correlation with suicide rates makes sense logically. However, it is worth noting that the Pearson's r coefficient for both these relationships is small (less than 0.40 for both variables), suggesting any linear relationship between these factors and suicide counts in a country is also very small (`other_vars_analysis.py`, lines 126-144).

Conclusion

This project investigated various factors that might impact suicide rates within a country and in doing so, revealed potential avenues for future research. The correlation between suicide rates and BMI, alcohol use, and health spending suggest these factors may be useful in identifying populations at higher risk for suicide and therefore, provide guidance on how to disperse mental health resources. The lack of correlation between happiness and suicide counts also provides a focus for more research. For example, do all metrics of happiness lack correlation with suicide counts or is this unique to the happiness metrics used in this data set? Also, the higher suicide rates within developed nations than developing nations implies some aspect of the difference between the definitions of "developed" and "developing" may be related to suicide and therefore, deserves further exploration. Overall, the project's emphasis on non-socioeconomic factors and their correlation with suicide revealed relationships and potential areas of future research that may help the healthcare system better serve and treat people who are suicidal.

References

Borges, G., Benjet, C., Orozco, R., Medina-Mora, M.-E., & Menendez, D. (2017, March 1). Alcohol, cannabis and other drugs and subsequent suicide ideation and attempt among young Mexicans. Retrieved November 26, 2021, from https://www.sciencedirect.com/science/article/abs/pii/S0022395616304253.

Gusmão, R., Ramalheira, C., Conceição, V., Severo, M., Mesquita, E., Xavier, M., & Barros, H. (2021). Suicide time-series structural change analysis in portugal (1913-2018): Impact of register bias on suicide trends. *Journal of Affective Disorders*, 291, 65-75. doi:http://dx.doi.org/10.1016/j.jad.2021.04.048

Hall, R. C. W., Platt, D. E., & Hall, R. C. W. (2011, April 28). *Suicide risk assessment: A review of risk factors for suicide in 100 patients who made severe suicide attempts: Evaluation of suicide risk in a time of managed care*. Psychosomatics. Retrieved November 26, 2021, from https://www.sciencedirect.com/science/article/pii/S0033318299712673.

Helliwell, J., Layard, R., & Sachs, J. (2018). World Happiness Report 2018, New York: Sustainable Development Solutions Network. Retrieved from https://worldhappiness.report/ed/2018/

Jayasinghe, N. R. M., & Foster, J. H. (2011). Deliberate self-harm/poisoning, suicide trends. the link to increased alcohol consumption in sri lanka. *Archives of Suicide Research, 15*(3), 223-237. doi:http://dx.doi.org/10.1080/13811118.2011.589705

Koivumaa-Honkanen, H., Honkanen, R., Koskenvuo, M., & Kaprio, J. (2003). Self-reported happiness in life and suicide in ensuing 20 years. *Social Psychiatry and Psychiatric Epidemiology: The International Journal for Research in Social and Genetic Epidemiology and Mental Health Services, 38*(5), 244-248. doi:http://dx.doi.org/10.1007/s00127-003-0625-4

Lester, D. (2002). National ratings of happiness, suicide, and homicide. *Psychological Reports, 91*(3), 758. doi:http://dx.doi.org/10.2466/PR0.91.7.758-758

Norris, D. R., & Clark, M. S. (2012, March 15). *Evaluation and treatment of the suicidal patient*. Evaluation and Treatment of the Suicidal Patient. Retrieved November 26, 2021, from https://www.aafp.org/afp/2012/0315/p602.html.

Pendergast, P. M., Wadsworth, T., & Kubrin, C. E. (2019). Suicide in happy places: Is there really a paradox? *Journal of Happiness Studies: An Interdisciplinary Forum on Subjective Well-being, 20*(1), 81-99. doi:http://dx.doi.org/10.1007/s10902-017-9938-y

Pompili, M., Innamorati, M., Lamis, D. A., Lester, D., Di Fiore, E., Giordano, G., . . . Girardi, P. (2016). The interplay between suicide risk, cognitive vulnerability, subjective happiness and depression among students. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues, 35*(3), 450-458. doi:http://dx.doi.org/10.1007/s12144-015-9313-2

Suicide Prevention: Risk and Protective Factors. (2021, May 13). Center for Disease Control and Prevention. https://www.cdc.gov/suicide/factors/index.html

van Vuuren, C. L., van der Wal, Marcel Franciscus, Cuijpers, P., & Chinapaw, M. J. M. (2021). Sociodemographic differences in time trends of suicidal thoughts and suicide attempts among adolescents living in amsterdam, the netherlands: Time trends of suicidal behaviors among adolescents. *Crisis: The Journal of Crisis Intervention and Suicide Prevention, 42*(5), 369-377. doi:http://dx.doi.org/10.1027/0227-5910/a000735