Analyzing Trends in Suicide Based Off Demographics

Lia Ran, Mitchell Rapaport

Abstract

This paper addresses how suicide trends are affected by the socio-economic demographics of a country, including: GDP, sex, age group, generation, population, and year. Our questions of interest involve observing the correlation of these features with suicide rates, and with each other. Our chosen data set has over 25,000 observations, spanning 100 different countries. We mostly focused on using visualizations and linear regression to help analyze our data and answer our questions of interest. At the end of our analysis we found that age, population and sex are the best predictors for suicide rate in this data set. However given ethical considerations how this affects bias, there is possible error in the predictions using our model. There are more predictors not included in the given data set that could be considered to increase our accuracy.

1 Introduction

Our primary goal of this project is to analyze socio-economic trends in relation to suicide rates to gain insight into possible correlations that may contribute to suicide rates. Our primary motivation for looking into this dataset is to gain insight on this worldwide phenomena and tragedy, and attempt to deduce a general trend. This helps shed understanding on approaching suicide prevention. Are there any upcoming trends we can predict and thus take preventative action? We want to identify patterns in suicide with socio-economic circumstances. Understanding these trends is a key factor in addressing and handling suicide prevention.

There is an overwhelming global stigma around mental health. Everyone understands the trauma of suicide, either directly or indirectly. According to the World Health Organization (WHO), one person commits suicide every 40 seconds leading to an annual total of around 800,000 lives lost. And this doesn't even account for the 20 attempted suicides for every successful one. In fact, in 2016, WHO reported that suicide was the 18th leading cause of all deaths globally. According to *suicide.org*, the rate of suicide has increased globally by a whopping 60% in the past 45 years, highlighting the ever-pressing and increasing importance of understanding key factors in its causation.

We want to use data exploration to provide insight into the demographics and correlation between external factors and suicide. In particular, we will be examining the Kaggle data set- *Suicide Rates Overview from 1985 to 2016*. This data set compares socio-economic information with suicide rates by year and country. This dataset is licensed by the World Bank Dataset Terms of Use, a reliable source of authority. The inspiration behind the creation of this data set was suicide prevention, and it contains multiple authentic sources as references: United Nations Development Program (2018) for human development index, the World Bank (2018) for GDP by country, Szamil for a previously sourced data set on suicide in the twenty-first century, and the World Health Organization for suicide prevention. Given these reliable and trusted sources, this dataset establishes itself as a dependable source to analyze suicide trends.

2 Questions of Interest

1. How have suicide levels changed throughout the years and how can we predict future suicide rates based off of demographics?
2. How is worldwide suicide affected by gender?
3. What influence does a country's economic level(GDP per capita) have on overall suicide trends?
4. Is suicide correlated by generation?
5. How are suicide trends affected by age group?

3 Data and Methods

3.1 Data

Our dataset was sourced from Kaggle:

https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
License: World Bank Dataset Terms of Use

The variables we will be investigating in order to analyze correlation with suicide rates are: GDP per capita, sex, age, generation, and average suicide rates per 100k pop. When the initial

pre-processing of the data, outliers are removed that may have an impact, considering potential ethical issues. We transform sex into a binary variable in order to better analyze trends. Additionally, we will use one-hot encoding to convert categorical variables like generation and age group to better train our linear model. Missing values are contained within single variable: Human Development Index (HDI).

Some relevant principles of measurement to consider are GDP, and Human Development Index. Since GDP per year for each country is converted into dollars, inflation of the US dollar can affect GDP through distortion and lack of precision. Human Development Index does not seem to be a relevant variable, as around 70% of the observations of this particular variable are missing. This also has to do with cost, as it may not be feasible for small or poor countries to invest to get their HDI report.

Some ethical biases to take into consideration when analyzing this data are inaccuracies in reporting suicide rates. Certain environments can cause suicides to get falsely reported, or on the flip side not reported at all . Another anomaly in the data could be physicians assisted suicide (PAS), which could explain the tendency that older individuals have towards high suicide rates. Can we really categorize this with other suicides? Additionally, one must consider a country's social desirability bias, which is a bias caused by desire to present information in a desirable light. Also, non-binary sex is not taken into account in this data set and could possibly affect the overall findings regarding sex. While this data set has authentic sources, we have no way of knowing the accuracy of the numbers reported.

3.2 Methods

1. To better determine suicide trends throughout the years, we can use visualizations to determine which demographics are most affected. Most of this is done through exploratory data analysis. We can grasp a pretty solid understanding of the demographics through multiple visualizations, and we gain insight on how to further develop questions of interest after experimenting with the data. Also, creating a linear model creates a way to see how accurately we can predict suicide rates based off specific demographics.
2. To discover the role sex plays in suicidal tendencies, we will create a visualization using our given data set to see the correlation between sex and overall suicide rates. Something to consider, however, is the existence of non-binary sexes and how this data is not included in the data set. We can also use a linear model as a means of identifying the use of sex in predicting suicide rate.
3. Similarly to sex, we will create a visualization that compares how average suicide rates differ among countries with different average GDP's. We can also use a linear model similar to the way we do for sex.
4. To examine the relationship between suicide rates through generations, we will use one hot encoding to create a new feature matrix. One hot encoding is used to turn categorical variables and 'binarize' them in order to use certain features algorithmically. This way, we can train our linear model using categorical variables. This will help us run multiple linear regression through training our model on this new covariate matrix.We will also use visualizations.
5. To analyze the relationship between suicide rates and age range, we will add new columns into our previously used covariate matrix, again using one-hot encoding. This should minimize our loss function and train our linear model even better. Using this linear model now, we will be able to more accurately predict certain tendencies towards suicide, further contributing to our original and first question. This will also tell us if analyzing by age range provides more insight towards suicide trends than just using generational data to train our linear model. We will also use visualizations.

3.3 Exploratory Analysis

Exploratory plots of data to gather a sense of relationships between relevant variables. Questions to ask:

1. Any transformations of any variables (ex. subbing in numerical values)
2. Dimension reduction (PCA) to identify interesting patterns
3. Outliers / measurement concern
4. Interesting or counterintuitive observations

Upon inspection, we noticed some outliers in the data through its extremeties. Additionally, when taking ethical considerations into issue, we removed all countries with a suicide rate of 0. The thought process behind this was the intuition that less developed countries have less access and reporting capacity to large institutions like WHO, as well as less reliable information when it comes to cause of death. Upon cleaning the data of these outliers, we notice the data became further standardized. We will be using the clean version of the dataframe to run the rest of our analyses. As we can see in Figure 1 and Figure 2, after we clean our data frame and remove outliers, the data appears more centered and normalized.

	year	suicides_no	population	suicides/100k pop	gdp_per_capita ($)	gdp_for_year ($)
count	27820.000000	27820.000000	2.782000e+04	27820.000000	27820.000000	2.782000e+04
mean	2001.258375	242.574407	1.844794e+06	12.816097	16866.464414	4.455810e+11
std	8.469055	902.047917	3.911779e+06	18.961511	18887.576472	1.453610e+12
min	1985.000000	0.000000	2.780000e+02	0.000000	251.000000	4.691962e+07
25%	1995.000000	3.000000	9.749850e+04	0.920000	3447.000000	8.985353e+09
50%	2002.000000	25.000000	4.301500e+05	5.990000	9372.000000	4.811469e+10
75%	2008.000000	131.000000	1.486143e+06	16.620000	24874.000000	2.602024e+11
max	2016.000000	22338.000000	4.380521e+07	224.970000	126352.000000	1.812071e+13

Figure 1: original data

	year	suicides_no	population	suicides/100k pop	gdp_per_capita ($)	gdp_for_year ($)
count	23353.000000	23353.000000	2.335300e+04	23353.000000	23353.000000	2.335300e+04
mean	2001.317390	282.258211	2.172742e+06	14.293268	17288.815570	5.248001e+11
std	8.463061	935.354871	4.178816e+06	17.215690	19152.428349	1.573480e+12
min	1985.000000	1.000000	1.003000e+03	0.020000	251.000000	4.691962e+07
25%	1995.000000	8.000000	2.423000e+05	2.740000	3299.000000	1.599989e+10
50%	2002.000000	42.000000	5.934090e+05	8.140000	9773.000000	7.714800e+10
75%	2009.000000	172.000000	2.142391e+06	19.190000	25848.000000	3.253583e+11
max	2016.000000	21262.000000	4.380521e+07	99.990000	126352.000000	1.812071e+13

Figure 2: cleaned data

To better analyze the dataset, we transformed some variables into numerical values. One transformation of a variable that we applied in order to better numerically grasp and manipulate the data was converting GDP per capita into a numerical value, versus a string. Additionally, to better analyze trends between male and female suicide rates, we used a binary conversion by setting {Male: 0, Female: 1} to better understand and numerically analyze the difference. Another method we used to further numerically categorize the data is through one-hot encoding. This allowed us to develop more complex linear models to answer our questions of interest, namely when consideration the correlation between suicide rates and generation, as well as categorization by age.

When exploring this data, we wanted to get a feel for the average values of each column, based on age and year. Thus, we group our data frame by the columns 'age' and 'year', take

the column-wise mean, and reset our index. This helps condense our data frame vastly and gives us a general feel for the values of each country, to better provide intuition into our guiding questions.

To continue our data exploration, we then created a box-plot to further explore if there was a relation between age and suicide rates. As illustrated in the box plot below, we found that the age range of 5-14 years had abnormally low values, probably due to the fact that it is extremely uncommon for people that young to have the means or even the emotional comprehension to commit suicide. Something interesting to note was that there seemed to be an increase in average suicide rates the older the age group got. Topping off the suicide per 100k population were adults 75+. This may have to do with a lot of factors, including the fact that Physician's Assisted Suicide may be legal in a lot of countries, whereas in the Unites States it is only legal in a couple of states. Another interesting visualization this box plot highlighted was that it seems that average suicide rate is positively correlated with age, minus the 5-14 year old age group of course.

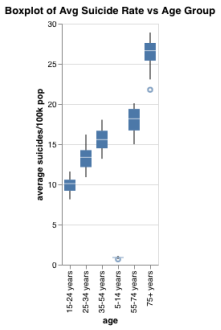

Figure 3: Boxplot of Average Suicide Rate Per Age Group

To explore the relationship between gender and tendencies toward suicide, we took our data frame and grouped by year and sex. Then, we summed up total suicides and manually calculated the suicide rate per 100k population, to standardize our values. Interestingly enough, the rate of women that committed suicide was consistently lower than men, and remained relatively constant throughout the 32 year span hovering at about 6 suicides/100k, with tendency of lowering in more current years. However, the average suicide rate of men fluctuated quite often, with a steady decrease in more recent years.

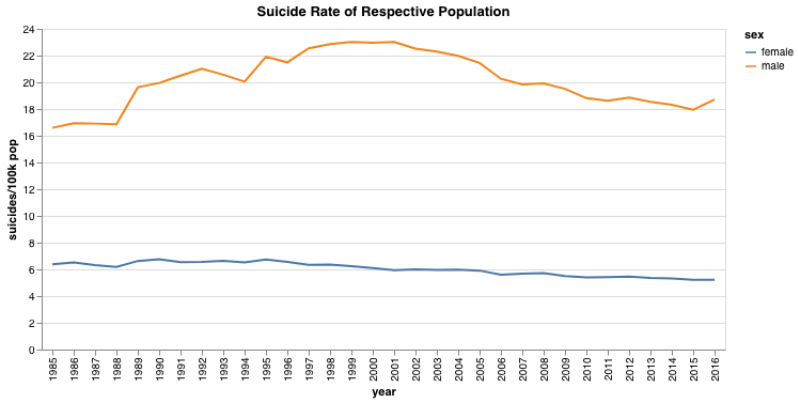

Figure 4: Progression of Suicide Rates Between Different Sexes

3.4 Inferential or Predictive Methods/Models

We used a Linear Regression model to predict suicide rates using the variables age, sex, and population as predictors. We used exploratory data analysis to create various scatter plots and boxplots to help identify trends in our data set. These visualizations help guide our decisions in what variables correlated or not. We used the scikit-learn package on Jupyter to our full advantage to more easily create linear models, and fit them with the corresponding data.

4 Analysis, Results, Interpretation

4.1 Details

Our assumptions were somewhat plausible. We assumed that suicide rates would be higher for men because men generally seem to be more reckless. We figured that age would have some relationship with suicide rate but didn't think they would completely positively correlate the way they did. We assumed that GDP would negatively correlate with suicide rate such that as one increased the other would decrease. This seems plausible because money may not be able to buy happiness but it does create more opportunity. However, the results proved otherwise. We thought that Gen Z would have the highest suicide rate because it is arguably the most depressed generation given the rise of social media and FOMO(Fear of Missing Out). However, suicide rate does not necessarily correlate with depression and this is shown in the results of our analysis.

There were many outliers in the inital imported dataset. Upon cleaning our data, we dropped about 4,000 observations. The intuition behind dropping these certain observations followed the following constraints: first, we wanted to drop outliers in the feature 'suicide/100k pop'. The outliers in this scenario were caused by countries with very small populations and a small amount of suicides, but an abnormally large suicide rate. If we blow up a ratio of a country with, say nine-hundred people, to one-hundred thousand people, the number of suicides gets inflated and thus causes an outlier in the dataset for the suicide rate. Upon removing this, we wanted to take into consider certain ethical issues. After contemplation, we decided to drop all observations of countries which had a suicide rate of 0. A recurring trend in these countries was that they tended to not only have smaller populations, but a lower overall GDP. This intuition leads to the belief that many smaller and/or poorer countries do not have the same

access to report cause of death with great accuracy and especially to such a large organization such as WHO, from which the majority of the data was pulled from. The probability of any given country having no suicides at all is very unlikely.

The only missing values in this data-set was the variable Human Development Index (HDI), and there were quite a lot (about 70% of our observations were missing a value in this column). One of the reasons behind this is because HDI was first introduced in 1990, while our dataset spans back to 1985. Additionally, HDI is notoriously hard to measure, and when we consider many impoverished and small countries, this kind of large scale reporting might be hard to achieve with the readily available data. To combat this, we removed the column in its entirety and instead focused on GDP to measure a countries economic well-being.

In terms of our questions of interest, we determined that suicide levels have been most affected by age, population, and gender. We can use these variables to attempt to predict suicide rates using a linear model with these predictors as our input. We came to the conclusion that males have a greater tendency to commit suicide over females by plotting suicide rates grouped by gender with respect to time. We deduced that GDP cannot be used to accurately predict the suicide rate of a certain demographic. Suicide largely seemed correlated by generation, but this could be biased by the fact that certain generations have a larger time-span of prevalence. We found this trend also in age group. Older people had more of an inclination to have higher suicide rates, but we acknowledge the bias based off of a larger time-span. Also, The R-Squared value associated with the model is not very close to 1, so this model is not great at predicting suicide rate based on the demographics in this data set. The Average Squared Loss of the predicted suicide rate(shown in Figure 8) compared to the true suicide rate and the nonlinear plot of true vs predicted suicide rate also show that our model is not accurate enough.

```
from sklearn.metrics import r2_score
r2_score(Y,Y_hat2)
```

```
0.36139441833557717
```

Figure 5: R-Squared Value of Linear Regression Model

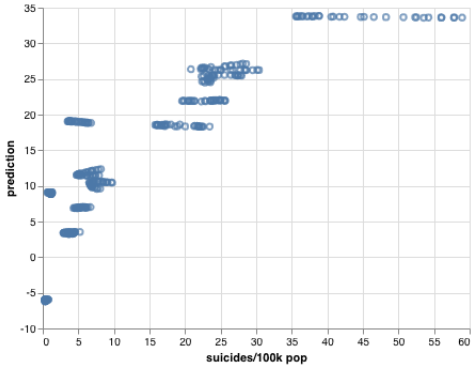

Figure 6: Predicted vs Observed Suicide Rates

An interesting feature that we deduced through running a linear model on the data is the similarity that generation and age group had on the variance of the data set. This implies that there is not a need to include both, and below you can see how the loss function for the two

covariate matrices is very similar. The left image shows the average squared loss of the model with both generation and age included. The right shows the model with generation excluded.

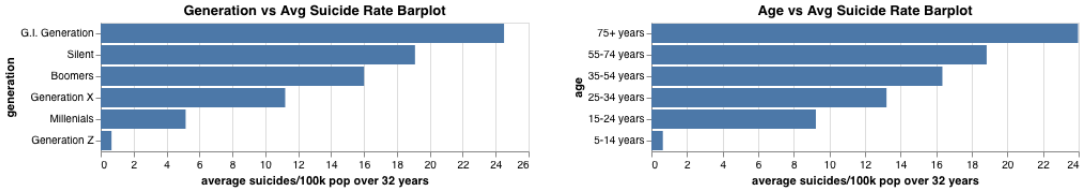

Figure 7: Similarities Between Age and Generation versus Suicide Rate

Sex is also an interesting feature as males consistently had higher rates of suicide(as shown in Figure 4), so this must attribute to some biological or social aspect that differs between men and women.

Another surprising conclusion we found was that GDP per capita had almost no effect on predicting our outcomes,so it was ineffective. The regression coefficient for GDP was small enough to the point where it did not have a large effect on our predictions. However, this proves somewhat surprising as there is contrary analysis online that suggests more suicides happen in low and middle income countries. Something to note, however, is the bias that could happen when adapting currency in our GDP variable to dollars.

```
Y_hat=model.predict(X)
loss=avg_squared_loss(Y,Y_hat)
loss
```

```
188.88685427874285
```

```
model2 = LinearRegression()
model2.fit(X2,Y)
Y_hat2=model2.predict(X2)
loss2=avg_squared_loss(Y,Y_hat2)
loss2
```

```
189.26179545435585
```

Figure 8: Comparing Models With/Without Generation Predictor

Ethical considerations took a large part in grappling with this data and was a big factor deciding how to manipulate it. Upon first inspecting the data, we removed outliers in order to more efficiently represent the data. Some main things we took into consideration were overall access to resources, or economic access. For example, lower income or smaller countries likely have less resources to properly account for total death numbers, as well as causation. They would also have less opportunity to communicate with large-scale organizations, such as the World Health organization. Additionally, under the impression of social desirability bias, a country might be inclined to adjust their numbers in order to appear more favorable.

Something that also must be taken into consideration is inaccuracies in the reporting of suicide due to social norms. For example, in cases of police brutality, often a murder can go reported as a suicide in order to been inconspicuous. We can expect this in countries with a higher police presence. This event of a fake suicide could also be common when countries are experiencing turmoil such as wars. War crimes could often be covered up and passed off with suicide. On the other hand, we need to consider countries where suicide is considered extremely taboo. This would most likely occur in extremely religious countries or regions. The disgrace of having a family member commit suicide might be so great that the death is passed off as a natural cause. Thus, we must be extremely sensitive to these inaccuracies when drawing

conclusions and performing analysis on our data. On the flipside, we must consider economic risk that a country might be considering. This could lead to inaccuracies in reporting data in order to embellish a certain image.

Non-binary sexes are also something to take into consideration. In today's world, non-binary sexes are becoming more recognized and suicides in this community would give more accurate results when analyzing the relationship between sex and suicide rate. Many online sources would say that this community actually has the highest suicide rate, so it would be very important to include in the data.

Finally, something to take into account in our observations is a possible explanation between the extremely high rates of suicide in older individuals and generations. This could largely have to do with physicians assisted suicide (PAS), which is when an older individual chooses to pass peacefully and deliberately. Can we consider it morally correct to group this with all other instances of suicide? Of course, PAS differs between country and regions, but it is hardly an obscure concept. This could greatly affect our reports and bias in our data analyses.

5 Conclusion

Our model has failed to reach our main goal of being able to predict suicide rate given the features in this data set. We found the best possible model given our predictors, but the model is not accurate enough for the reasons given in the analysis. In order to better predict suicide rate, more features have to be considered. It is also possible that determining suicide rate has to do with the specific culture of a person in a specific community, which is almost impossible to quantify as a predictor.

The model may have been a bit more accurate if we had considered removing more outliers, but removing anymore 'outliers' could possibly distort our data even moreso than keeping all of them. We would be taking out too big of a chunk of our data. Either way, features/demographics not included in this data set would have to be added to better our model. However, just because our model should not be used to predict suicide rates, does not mean that the predictors are not correlated to suicide. We found that older men are the highest risk group for committing suicide. Suicide is one of the leading causes of death in the world, and it is important to identify the causes of it and the groups most at risk, as well as to find the best ways to prevent it.

6 Bibliography

1. https://www.oreilly.com/library/view/statistics-in-a/9781449361129/ch01.html
2. https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f
3. https://www.who.int/mental_health/prevention/suicide/suicideprevent/en/
4. https://qz.com/1456012/the-3-key-problems-with-the-uns-human-development-index/
5. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html
6. https://ds100.lsit.ucsb.edu/user/mrapaport/notebooks/ds100-s20-content/labs/lab6/lab06.ipynb
7. https://ds100.lsit.ucsb.edu/user/mrapaport/notebooks/ds100-s20-content/labs/lab8/lab8.ipynb
8. https://ds100.lsit.ucsb.edu/user/mrapaport/notebooks/ds100-s20-content/hw/hw3/hw3(1).ipynb
9. https://williamsinstitute.law.ucla.edu/publications/suicidality-transgender-adults/