CIS 671 - Semester Project

James Stewart

`stewajam@mail.gvsu.edu`

Abstract

This paper will go over the creation of two visualizations that explore the relationship between international suicide rates, different demographic groups, and human freedom index score.

1. Proposal

According to the world health organization, over 700,000 people lose their lives to suicide each year [4]. However, the rate of suicide in the population is vastly different from country to country. Suicide rates more than double when looking specifically at European white men. Statistics like these and their variability among different societal groups are the motivation behind the visualizations in this task. This paper hopes to explore possible correlations between suicide rates and demographic information including age, gender, country, and the countries human freedom index score.

1.1. Tasks

This section will list off the tasks that the following report will hope to accomplish with it's visualizations. A more in-depth description of each task will follow.

1. Have the suicide rates (suicide deaths / populations) for different demographic groups changed over time and has this change been universal?

2. Was their a correlation between a countries suicide rate in 2010 and their human freedom index score from that same year?

Both tasks will be using suicide rate which will be a created column based on the (suicide deaths / population) x 10,000 (see data overview for more information). The first task will use demographic information including country lived in, age group, and gender. It will also look at suicide rates from 1979 - 2017. The goal of this task will be to allow for exploration of different combinations of demographic groups to help look for any possible correlations/interactions. The second task will look more closely at how national suicide rates correlated with the human freedom index score (see literature for more information).

1.2. Literature Review

As stated by Ajit Shah in their paper, "Suicide rates: age-associated trends and their correlates", it was normally thought that suicide rates increased with age [7]. However, Shah noticed that this wasn't always the case and they looked to explore on what caused these discrepancies. in the same paper, they found that the change in suicide rates and age differed greatly depending on their country and showed gender as an additional interacting variable [7]. Another study done by Rehkopf and Buka found that there were socio-economic factors involved with this relationship [5]. Studies like these point towards a relationship between demographic groups and suicide rates, which warrants looking further into correlations between such factors and suicide.

1.3. Data Overview

Two separate data sets were used. One with data from the world health organization on suicide statistics around the world from years 1990 - 2017 [8]. The second data set holds information on countries human freedom index from the years 2008 - 2021 [6]. All columns from the original suicide statistics data set will be used, however only a few columns will be used from the human freedom index data set. A breakdown of the data used is provided in the following tables.

World Health Organization - suicide Statistics			
Variable	Data Type	Range	Description
Country	Character	165 different countries	The country that the data is taken from
Year	Numeric	1979-2017	The year that the data is from
Gender	Factorial	Female, Male	The gender of the population in question
Age group	Factorial	4-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, 75+ years	The age group of the population in question
Suicide Num	Numeric	0-22.3k	The number of suicides in the population
Population	Numeric	259-43.8k	The number of people living in that country

Human Freedom Index			
Variable	Data Type	Range	Description
Year	Numeric	2008-2021	The year that the scoring and ranking took place
Countries	Character	165 different countries	The country that is being scored
hf score	Numeric	3.49-9.15 (0-10)	The human freedom index score given to them by the Cat Institute, Fraser Institute, and the Liberales Institute. Measures how free the people of the country are (higher = more freedom).
hf rank	Numeric	1-165	The ranking of how free the countries people are (1 = the most free)

Both data sets were obtained from Kaggle.com and were uploaded by their corresponding authors. Links to the specific web pages can be viewed in the reference section.

2. Design

This section will go over the design of each visualization, the reasoning behind the decisions made for this graph, and a simplified example.

2.1. visualization 1

For our first visualization, we hope to explore whether the national suicide rates for different demographic groups changes over time and whether these trends are universal or country specific. The demographics that we will be looking into are age group and gender. Suicide rates will be calculated by the number of suicides divided by the population multiplied by 10,000 to get the number of suicides per 10,000 people.

The visualization that we chose to use is an interactive, multiple line time series graph. When looking at changes over time, a time series graph is the most obvious choice.

To make it interactive, the user will be able to choose a grouping variable and what levels of this variable to look at. Each individual group will be a single line on the chart. This will be similar to the wide-form data shown by Mahbubul Alam in their article [1]. There were other time series options shown by Alam to choose from including bar plots and box-plot graphs. However, the need to be able to group the data into multiple variables will be clearest to the user in line form. Not only does this allow for a clear progression over time, but with multiple groups, the principal of continuation will allow the user to see which lines are which when they cross over. Having large box plots for each year would overlap and could be confusing, especially with multiple attribute levels.

The reason we chose to do an interactive visualization is to allow the user to explore the data more in-depth. A similar visualization experiment done by Kajla and Yeh [2] looked at suicide rates and happiness scores in which they included grouping variables. Our visualization will also implement slicing so the user can look more closely at specific information. This method was also shown by Alam in their article [1].

The main part of the graph is placed in the center-right and encompasses a much larger area than the rest of the visualization. This will help draw the user's attention to the basic information. The buttons/drop down menus are placed on the left hand of the screen. The user can use these to choose different grouping variables and levels to see the effects they have on the suicide rate. Finally, the legend for the different groupings will be placed on the top right. The color palette chosen from this graph was categorical so that it was easier for the different levels to stand out from one another. Figures 1 - 4 show visualization 1 at different points of its use. Figure 1 is the start screen that is shown when the user first starts the program. In the starting position, there is no grouping variable. Figure 2 and 3 then go on to show an example of choosing a grouping variable, in this case age, and then selecting multiple different levels of that variable. Finally, Figure 4 shows an example of a user reducing the year range that is being looked at.

The two main gestalt principles utilized by this visualization are proximity(similarity) and continuation. When looking at differences between trends, those lines that are closer together are more similar than other while lines that are further apart have a reason to be different. This is the basis of proximity. For continuation, the trend lines for each group will cross over each other so the continuation principle and the color schemes chosen will help organize the different lines.

2.2. visualization 2

The second visualization looks at whether there is a correlation between suicide rates and the human freedom in-

dex score of the country. Suicide rate will be calculated the same way as in visualization 1. As described in the data overview, the human freedom index score is a number between one and ten (ten being the highest) given to each country by an independent organization. HFI measures how free the people in the country are based on different principles.

As stated by Mindrila and Balentyne, the best way to show a relationship between two quantitative variables is by using a scatter plot [3]. This is because it allows for both variables to be plotted on an axis and the data points to be grouped together. Another visualization type that was considered was parallel lines graph. However, due to the number of countries included in this data, we believed the scalability issues of this graph could pose usability issues for our users. Because we are looking to view correlation between these two factors (or the lack thereof), a linear line of best fit will be included to help show the correlation.

The visualization for this task is relatively simple in comparison to our first graph. The placement of each point, along with the grouping of points in certain areas of the graph, will help relay not only the countries HFI and suicide rate, but will also help show the grouping and trends that we predict to see. The area between the points in the visualization will help show how closely related the different countries are.

The Cleveland's rule most utilized in this graph is position (common scale). The position of each point will help display (or not display) the correlation between the two variables. For gestalt principles, this graph will utilize good figure(connection) and similarity. Even though the data will be shown by individual points, correlation will be determined by how "close" the points are (along with the included line). The closer that the points are towards each other/the more in line they are, the higher the correlation that we hope to find will show. Also, how similarly or not similarly placed the points are on both the x and y scale will help show relationships between different countries and the two quantitative factors. Interesting outliers might then intrigue the user to explore our other visualization more to find out what the issues are.

3. Analysis

The task for the first visualization aimed at showing if there was a difference between suicide rate over time for different demographic groups. Figure 3 and 4 both give a good clear answer. Figure 3 displays multiple different age groups which all have vastly different suicide rates.However,it looks like the differences between these age groups stay relatively constant over the years as they all follow a similar pattern. This differs however when you look at Figure 4 which displays male vs female suicide rates. Males have a drastically higher suicide rate than girls.

They also differ in the change from year to year. Where females stay relatively stable over the years. The visualization makes really good use of proximity to help show this off.

The second visualization was hoping to find a correlation between human freedom index score (HFI) and suicide rate. Using the gestalt principles proximity and good figure, you can see an upward swing that the data makes. This, along with the line of best fit, shows a very positive correlation (linearly or exponentially) between the two variables. This is an extremely interesting finding because it means that, as the people of a country become more free, their likelihood of killing themselves goes up.

References

[1] M. Alam. Formating and visualizing time series data, 2020. https://towardsdatascience.com/formating-and-visualizing-time-series-data-ba0b2548f27b.

[2] A. Kajla and N. Yeh. Visualizing suicide rates and world happiness, 2019. https://towardsdatascience.com/visualizing-suicide-rates-and-world-happiness-bc4e6411732f.

[3] D. Mindrila and P. Balentyne. *The basic Practice of Statistics*. New York, NY: W.H. Freeman and Company, 6 edition, 2013.

[4] W. H. Organization. World health statistics data visualization dashboard, 2016. Suicide.

[5] D. H. Rehpkof and S. L. Buka. The associatin between suicide and the socio-economic characteristics of geographical areas: a systematic review. *Psychological Medicine*, 36(1):147–157, 2006.

[6] G. S. Schneider. The human freedom index, 2021. https://www.kaggle.com/gsutters/the-human-freedom-index.

[7] A. Shah. Suicide rates: age-associated trends and their correlates. *Injury Violence*, 4(2):79–86, 2012.

[8] Szamil. Who suicide statistics, 2019. https://www.kaggle.com/szamil/who-suicide-statistics.
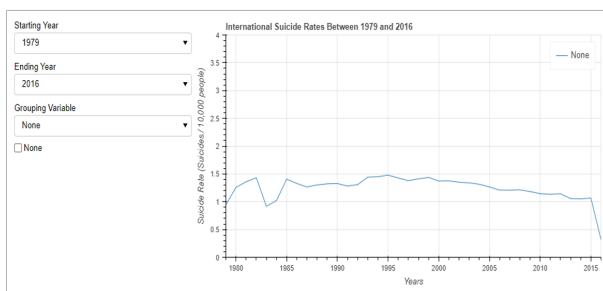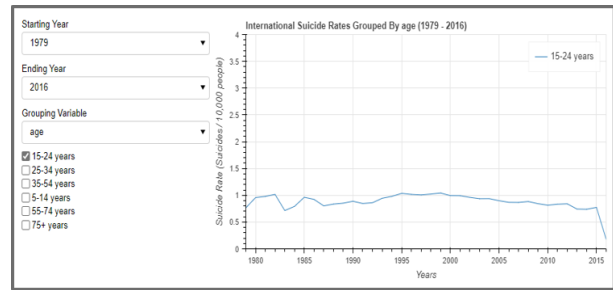
Figure 1. Start Screen

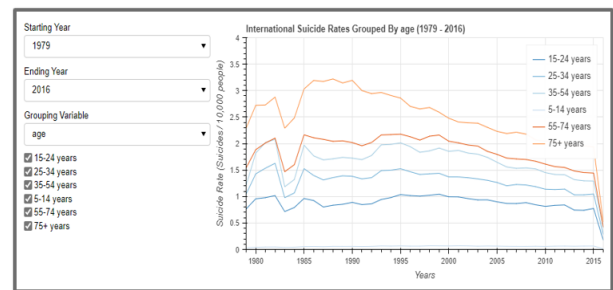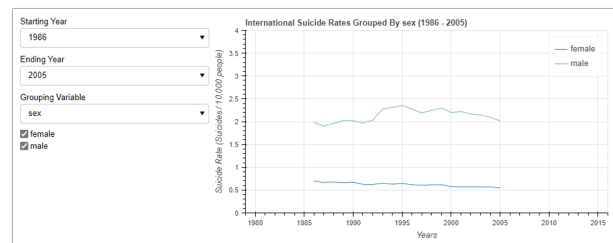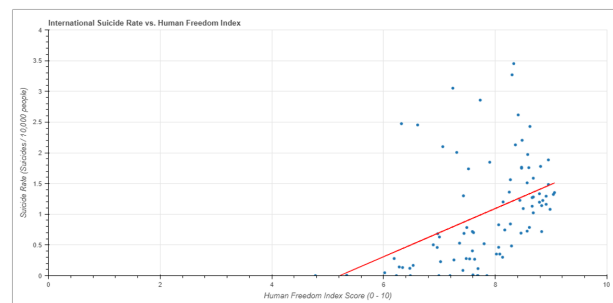Figure 2. Chosen Grouping Variable

Figure 3. Chosen Attribute Levels

Figure 4. Adjust Year Range

Figure 5. Visualization 2