Jason Ficorilli

Prof. Pelechrinis

INFSCI 1040

5 December 2019

Final Project Report: Suicide Rates

For my final data analytics project, I chose to find out whether the conclusions drawn from Emile Durkheim’s foundational 1897 book *Le Suicide* still hold up today. I found this question interesting because Durkheim is considered one of the founding fathers of sociology, which I am studying as a minor. *Suicide* is a significant text because it was one of the first works to use data to explain a social phenomenon. Since suicide data is much more easily obtainable today than it was in the late 1800s, Durkheim’s study is an ideal candidate to be explored from a modern perspective.

My first step for the project was finding a dataset that I could draw conclusions from. I found an adequate one on Kaggle, a data science website. The dataset was called “Suicide Rates Overview 1985 to 2016,” and it contained 12 columns: country, year, sex, age group, count of suicides, population, suicide rate, country-year composite key, Human Development Index (HDI) for year, GDP for year, GDP per capita, and generation. The set contained data for North America, Europe, most of South America, the Caribbean, and Oceania. Most of Asia and Africa were unavailable.

Before beginning my analysis, I had to determine which of Durkheim’s findings I would be following up on given the data I had on hand. The most easily answerable one was his conclusion that men commit suicide at a higher rate than women, since the data was already divided into categories of men vs. women. The next conclusion I chose to investigate was Durkheim’s finding that Protestants commit suicide at higher rates than Catholics. Since my dataset didn’t contain any information about religion, I had to do some research of my own. I used information from the Wikipedia pages “Protestantism by Country” and “Catholic Church by Country” to find which countries were majority-Protestant and which were majority-Catholic. My criteria for this was if over 50% of a country followed either religion.

With my research questions defined, I imported the dataset into Jupyter Notebooks in order to perform my analysis using Python. After reading the CSV file into a dataframe, I determined the number of null cells and plotted a histogram to find the overall shape of the data. I found that the only null cells came from the HDI column, which wasn’t too much of a hindrance since it wasn’t essential for the purposes of my analysis. The histogram revealed some limitations of the data, such as the fact not all countries had data available for the oldest and most recent segments of time (i.e., around 1985 and around 2016). I then created basic scatterplots to see trends in the data. HDI and suicide rate didn’t seem to have much of a correlation, but suicide rates seemed to be higher in countries with low GDP per capita (though there were much more data points for low-GDP countries). With an idea of the general shape of the data in mind, I then began to investigate whether Durkheim’s findings still hold up.

I first compared suicide rates between men and women since it was the easiest analysis to perform. I grouped the data according to the “sex” column and used Python’s sum function to add up the total number of suicides and the total population of both men and women. I then used these values to calculate the number of suicides per 100,000 people for both men and women. The result I got was 20.7 suicides per 100,000 men and 5.9 suicides per 100,000 women. I then created a simple bar chart to visualize the difference in these values. My calculation more than reaffirmed Durkheim’s conclusion that men commit suicide at a higher rate than women, as the rate I arrived at for men was nearly four times higher than that of women.

To compare suicide rates between Protestants and Catholics, I made an array listing the names of majority-Protestant and majority-Catholic countries, then created a new column in the dataframe and assigned the countries in both arrays as either “Catholic” or “Protestant.” The Catholic countries were mainly located in South America and Western and Central Europe. There were much fewer Protestant countries, with all of them coming from either Scandinavia or the Caribbean. With the two religious categories created, I then made a scatterplot that showed the distribution of suicide rates in Protestant and Catholic countries. I made the scatterplot a function of suicide rate per 100,000 people vs. GDP per capita. The data showed most of the data for Catholic countries bunched on the low end of GDP per capita, while Protestant countries were more equally distributed across the spectrum of GDP, with the largest chunk of them being in the middle. Suicide rates seemed to be higher for Catholic countries, particularly those with lower GDP.

With a broad idea of the distribution of suicide rates for Catholic and Protestant countries, I began to calculate an overall aggregated suicide rate for both categories of country. Due to the large amount of data on hand, I thought it would be more descriptive if I limited the data I’d be analyzing to the 21st century. With all data prior to 2000 excluded, I then created separate dataframes for Catholic and Protestant countries. I filled in the null HDI cells with the average HDI value for each given country. I then found the sum of the populations and suicides committed in each Catholic and Protestant country from 2000-2015 to find the overall suicide rates. In the end, I came up with 12.4 suicides per 100,000 people in Catholic countries compared to 6.56 suicides per 100,000 people in Protestant countries. I then created a bar chart displaying these two values side by side, and then made a line graph that showed the suicide rates in Catholic and Protestant countries over the years from 2000 to 2015. My conclusions directly contradict Durkheim’s findings that Protestants commit suicide at a higher rate than Catholics. This shows that the social context must’ve changed in both categories of countries in the hundred years since Durkheim wrote *Le Suicide*.