Data Visualization Project - Data Acquisition Report

Focus Area

With limited time on this earth, I am motivated by a philosophy that attempts to minimize human suffering and maximize human freedom, happiness, and health. In past data analysis projects for this masters program, I've analyzed income inequality, police stops, and Dota 2 video game data. Specialization and subject matter expertise is definitely valuable, but I like the idea of exploring analyses on varying topics. A project using stock market data will require different techniques than one using genome sequencing data. With that in mind, I browsed available datasets on kaggle.com and found two that were interesting - the Human Freedom Index (HFI) and a suicide rates overview. This data will help me investigate factors that contribute to a human's freedom and its relationship to suicide, which could be a proxy for happiness and health as well.

HFI Context. The HFI "presents a broad measure of human freedom" for many countries around the globe, representing 94% of the world's population (Vásquez & McMahon, 2020). The authors of the HFI define freedom as the absence of coercive constraint, commonly known as negative liberty. This is in contrast to positive liberty – "the removal of constraints that impede one's personal improvement or fulfillment of his or her potential as the individual understands it" – as it can be difficult to measure and its meaning varies from person to person (Berlin, 1969). In my ideal index both negative and positive freedom would be accounted for, but I can understand their reasoning for not including it at this time. An example of negative versus positive liberty is when a poor person is not by law prohibited from attending college or starting a business, but their lack of financial resources keep them from fulfilling their dreams. I am planning to join the HFI data to the suicides dataset to determine if there is any link between freedom, number of suicides, and a country's gdp per capita and total population.

Initial Requirements

To evaluate a country's freedom, I will look at input variables like HFI score, economic freedom subscore, and the personal freedom subscore by year, region, and country. In the suicide dataset, I plan to use the variables country, year, sex, suicides per one hundred thousand persons, and gdp per capita. With data that has broken down freedom scores and suicides, population, and gdp per capita, many questions can be explored. For this analysis, I'd like to answer if a country's freedom, number of suicides, and gdp per capita are related, if a country's population impacts its freedom, if a country's freedom for women is related to its female suicide rates, and what are important factors in determining a high freedom country. Correlations between variables will be calculated to identify existing relationships and regression models will be employed to aid in finding key factors for countries with high freedom.

Hypothesis

This is a topic where I am not very knowledgeable, but some educated guesses can be made to opine about the direction of the above relationships. Specifically, I believe that female suicide rates is inversely correlated to women's freedom, that is female suicide rates should decline with increasing women's

freedom in a country. With lower confidence, I also posit that countries with smaller populations tend to have more freedom. Both of these hypotheses can be tested with simple linear regression models since the input and output variables are numerical, but more complex models may be employed if too many assumptions are violated. As my student license for Tableau has expired, I plan to use R (R Core Team, 2021) to conduct my analyses as it is a favorite of mine for analysis and visualization with its tidyverse suite of packages including ggplot2 (Wickham, 2016).

Data Sources

Both the HFI and suicide rates datasets have indexes of year and country tied to numerical measures like freedom indicators across many categories and population and suicide counts, respectively. HFI also groups the countries by global region, while suicides are further broken down by sex and age. Both were downloaded from the kaggle.com datasets page in a comma delimited file format (csv).

HFI data. The HFI dataset has a total of 123 columns and 1,458 rows and covers the time period 2008 to 2018 for 162 unique countries. The majority of columns are from the 76 numerical indicators and their category subtotals that average together to create each country's freedom score. Each of the 76 indicators are rated on a scale of 0 - 10 with 10 representing the most freedom, along with subtotals for the individual categories[1]. The numerical columns have varying degrees of missing values, but are especially pronounced in the categories rule of law, inheritance rights, freedom to establish/operate religious organizations, and freedom to establish/operate political parties. HFI contains 4 categorical variables – year, ISO_code (country abbreviation), country name, and region. Since the HFI report was compiled by two large North American think tanks, the data seems to be high quality, on consistent scales, but with some missing data. The referenced report from Vásquez and McMahon is quite long and contains the index context, detailed column descriptions, and many tables and graphics for regions across the globe. The HFI dataset can be joined with the second suicide dataset on the country and year variables so that records can be matched up appropriately, though issues in joining may occur as the suicide dataset has collected data from fewer countries.

Suicide data. This dataset has 12 columns and 27,820 rows and covers the years 1985 to 2016 for 101 countries. From kaggle's convenient map of the country column, much of the suicide data from African, Asian, and Middle Eastern countries is not included in this dataset. While this will hamper comparison between freedom and suicide rates for some countries in those regions, I think there is sufficient overlap in the remaining countries to facilitate a satisfactory analysis. Among the columns, 6 are categorical – year, country, sex, age category, country_year, and generation. The remaining 6 columns are numerical – suicide counts, population, suicides per one hundred thousand persons, human development index (HDI), gdp for year, and gdp per capita. All columns are complete – no missing data – except for the HDI column where 70% of the values are missing. The metadata states that it is a compiled dataset combining four other datasets from the likes of the United Nations Development Program, World Bank, kaggle.com, and World Health Organization (Yates, 2018). I don't have a good reason not to trust the data published by these organizations, so I believe this suicide dataset to be of good quality and validity. The kaggle source page is lacking column descriptions, but all

[1]See page 15 in the HFI 2020 report for the structure and categories of the HFI data

except HDI were fairly self-explanatory. The human development index is discussed further in Suicide Rate, Depression and the Human Development Index: An Ecological Study from Mexico article in the Front Public Health journal (Cabello-Rangel, Márquez-Caraveo & Díaz-Castro, 2020). The suicide rates and global domestic product (gdp) have already been scaled by each country's population, making for easy comparison among countries.

[1]See page 15 in the HFI 2020 report for the structure and categories of the HFI data

References

Berlin, Isaiah, “Two Concepts of Liberty,” in Isaiah Berlin, *Four Essays on Liberty* (Oxford: Oxford University Press, 1969).

Cabello-Rangel, H., Márquez-Caraveo, M., & Díaz-Castro, L. (2020). Suicide Rate, Depression and the Human Development Index: An Ecological Study From Mexico. *Frontiers In Public Health*, *8*. doi: 10.3389/fpubh.2020.561966

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Vásquez, I., & McMahon, F. (2020). *the Human Freedom Index 2020*. CATO and Fraser Institutes. Retrieved from https://www.cato.org/sites/cato.org/files/2021-03/human-freedom-index-2020.pdf

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Yates, R. (2018). Suicide Rates Overview 1985 to 2016. Retrieved 28 October 2021, from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

[1]See page 15 in the HFI 2020 report for the structure and categories of the HFI data