**Data Preparation & Information Modeling Report**

**Human Freedom vs Suicide Rates**

This project will use two datasets obtained from kaggle.com to explore relationships between a country’s human freedom – measured by their Human Freedom Index (HFI) score – and their suicide rates, population, and gross domestic product (gdp) per capita (Sutter, 2020; Yates, 2018). I joined the two datasets by country and year to determine if there is any link between freedom, number of suicides scaled by population, a country’s gdp per capita, and its total population. More specifically my analysis examined the links between population size and freedom, female freedom scores and female suicide rates, and important factors that contribute to human freedom. I posited that high female freedom scores will have less than average female suicide rates and that small population countries will have more freedom than large population countries.

**Data Preparation**

Before beginning exploratory data analysis (EDA), it’s important to assess the initial data format and clean up variable types and names. By default when R reads in a csv file, it assigns categorical variables a character format when often a factor format is more appropriate (R Core Team, 2021). Applying this approach, the categorical variables *country, region, sex,* and *age* were reassigned to be factors. Each *age* value initially had the suffix ‘years’ - like ‘5-14 years’ for the first age bucket - but it was chopped off as ‘years’ doesn’t add any value to the analysis. In comparing the HFI and suicides datasets, I noticed that HFI records start back in 2008, while data on suicides began in 1987. However, my project is primarily focused on the HFI freedom dataset, so the suicides data was filtered to remove records before 2008 prior to exploration and joining. Suicide counts initially were broken down by age and sex categories (in addition to year and country). Some analyses compare suicide rates to freedom scores and do not involve sex or age, and this required aggregating suicides and population into overall counts and dividing them to get a scaled rate of suicides per one hundred thousand persons. This scaling and formatting is common in the medical and sociological fields to give a rate proportional to a country’s population in a usable format - simpler values like 5.6 rather than long values like 0.0000056. For a later analysis on female freedom, a few variables that measure female freedom – inheritance rights, movement, female to female marriages, and divorce – are averaged together to resolve some missing values in each individual variable and provide a better correlation plot (Vásquez & McMahon, 2020).

*Joining Datasets.* While joining two datasets on the same two variables – country and year – may seem simple, it did involve a fair amount of work to identify and fix country name mismatches, like Cabo Verde vs Cape Verde. I checked country names with the Duckduckgo search engine and Wikipedia entries ("Wikipedia", 2021). I was able to fix six country names so they could join correctly, but over fifty more were present in only one dataset or were present in both but the years were not aligned. In total, eighty-four countries were joined together to create a clean dataset consisting of seventeen variables across almost seven thousand rows.

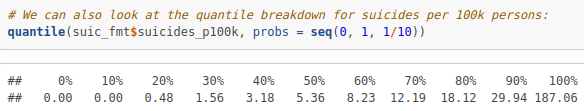
**Suicides Data Exploration**

I enjoy using R for data analysis projects as I’ve done for past courses in this masters program, but this time I wanted to learn a new way to do EDA with R Markdown (Xie et al., 2018). Doing EDA this way took longer, but the upside is the exploration process is more detailed and accessible for other data analysts. R Markdown combines code with markup

comments and code results all in one output so that analysts can show their thought process and results of each step.

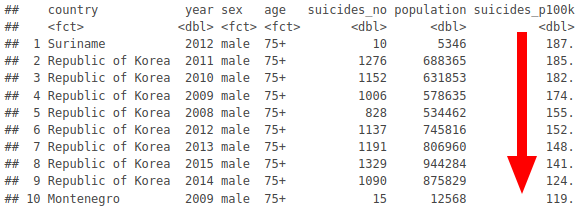
*Suicides data.* I started with the typical *summary, head,* and *tail* functions and noticed that there were no missing values in the data and nothing out of the ordinary looking at the top or bottom of the dataset. The value in checking the top and bottom is verifying that all records were loaded and no additional metadata – column names, number of rows, date createed, etc – snuck in from the top or bottom of the file. I first sorted the data by suicide counts descending to see what years and countries had the highest suicides. It turns out that the US and Russian men in the age bracket 35-54 were at the top, but after moving to the scaled variable *suicides\_p100k* they were no longer present. This is due to the large populations that US and Russia have. Without scaling, their populations would incorrectly weight their suicide counts.

*Scaled suicide rates.* R’s *summary* gives some detail on the distribution of suicides per one hundred thousand persons, but it’s *quantile* and *boxplot* functions provide more resolution on the shape of suicide rates in the data. From Figure 1 below I can see that a significant portion of the records have a count of zero and those that don’t remain small until the last quantile between ninety-one and one hundred percent.

  
Figure 1: Quantile breakdown for Scaled Suicide Rates

This indicates a right skewed distribution with the majority of values between zero and twenty with a few large positive values – coming mainly from elderly males from the Republic of Korea (Table 1).

*Zero suicides?* There are quite a few records – greater than ten percent – that have recorded zero suicides for a given year and country. R counted the number of records each country has with zero and non-zero suicide counts and those with the highest percentage of zero suicide counts – Antigua and Barbuda, Grenada, Barbados, Maldives – were not included in the analysis due to their not being present in the HFI dataset.

  
Table 1: Descending Scaled Suicide Rates, Korea at top

*Age and Sex.* Though age and sex are not featured in this analysis, besides female suicide rates and freedom correlation, I decided to explore their contributions to the suicides data as part of the EDA process. Suicides seem to be more common in men than women and in elderly vs young folk, consistent with what Table 1 above is showing. For a more appealing visual of this trend and all other project resources, please see my [github](https://github.com/doug-cady/gmu_daen/tree/master/AIT664_InfoReprProcVisuals/project/bin) page.

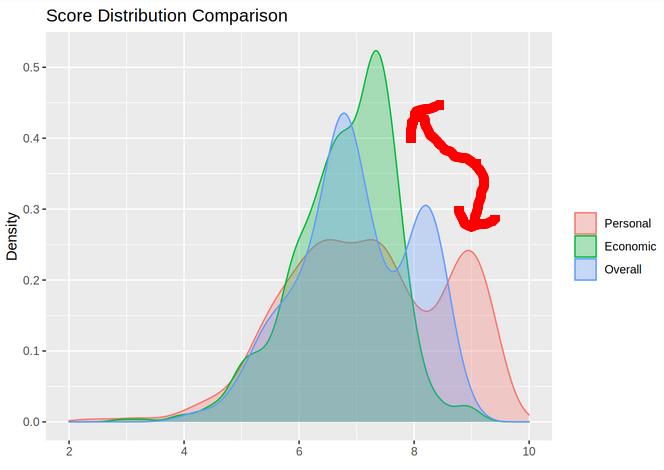
*Distributions.* Each of the numeric variables – *suicides\_no*, *population, suicides\_p100k,* and *gdp\_per\_capita* – had a heavy right skew enough to where each histogram looked a lot like a letter L. While log transformations are a bit harder to interpret, I felt it was necessary to use them here to get a better view the variable distributions.

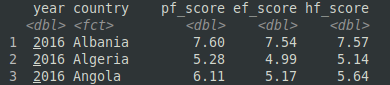
*Time-series.* Lastly, I wanted to get some feel for suicide rate trends over time and created an interactive plotly graphic. It is quite busy since there are ninety-four countries being displayed at one time. Fortunately, the interactive nature allowed me to selectively show only a few countries at a time with a few clicks of the mouse. While this dataset didn’t have any missing data in the form of NAs, I could see from this plot that several countries suicide records were not included in the dataset. These tended to be in the smaller countries that were ultimately removed after joining to the HFI dataset.

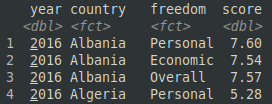
**HFI Data Exploration**

The HFI dataset is much larger and contains over one hundred variables, so it was important to me to limit the scope and focus in on the ones that would help answer the questions I laid out. I narrowed down the variables to year, country, region,5 female freedom variables, personal freedom score, economic freedom score, and overall human freedom score. Like before with the suicides dataset, I started with *summary, head,* and *tail* functions to check if all data was loaded in and was in a good format.

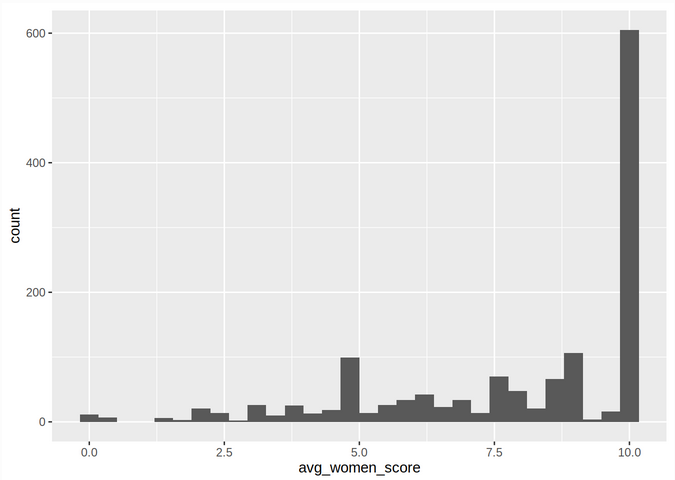
*Freedom scores.* Since the overall freedom score, will be the outcome variable in our modeling step, I looked at its distributions first compared to its personal and economic subscores with a density plot using R’s ggplot2 package (Wickham, 2016). To construct a superposed plot like this, R needs data to be in a tidy long format. The *pivot\_longer* function can pivot from wide to long formats as shown in Figures 2 and 3 above. Figure 4 below is a plot showing the distribution of scores for each of the three types of scores and highlights that economic scores tend to be in a lower tighter range than either personal or economic freedom scores. As the arrow annotations indicate, personal freedom scores have a greater percentage of high scores than economic scores.

  
Figure 4: Score Distribution by Type

  
Figure 2: Before: Wide with 3 Score Variables

  
Figure 3: After: Long format

*Female freedom.* Turning our attention to the five female freedom variables, there are quite a few missing values ranging from around five to sixty percent. This led me to average each row’s five scores together to partially resolve the missing values problem and lead to a simpler plot representation. Surprisingly, the most common average female freedom score was ten or very near ten as shown in Figure 5 below.

  
Figure 5: Average Women Score Distribution

**Information Modeling**

The earlier data preparation section details joining the datasets together, so now that the two datasets have fused into one I can start trying correlations and other related techniques to answer questions. <corrs freedom suicide gdp>

<pop and freedom>

<female freedom, suicide rates>

**References**

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/.

Sutter, G. (2020). The Human Freedom Index. Retrieved 28 October 2021, from https://www.kaggle.com/gsutters/the-human-freedom-index

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>

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

Wikipedia. (2021). Retrieved 13 November 2021, from https://www.wikipedia.org/

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown*. CRC Press.

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