Suicide Rates Database Project

Gregory Nau and Jay Sherman

DS 4100- Spring 2019

April 17, 2019

The main motivation for this project is to investigate correlations between a countries' culture, economic standing, values, and population to determine potential risk factors associated with higher suicide rates between 1985 and 2016. Through the analysis of a large number of statistics for over one hundred countries for the past few decades, our goal was to create a database that would allow for exploration into what attributes are the most strongly associated with suicide rates in countries. It is important to emphasize that our goal is not to predict, label certain factors as definite causers of suicide, or claim that this data is the perfect insight into suicide rates, as this is inappropriate. The associations discovered do not necessarily denote causation, as the decision to end one's life is made on the individual level, which means the many forces affecting an individual's probability of dying by suicide may not be visible in the data collected at the country level. The purpose of this project is to collect as many sets of data as possible to compare by country by year and to store them for easy retrieval, such that associations between these attributes and suicide rates can be broadly studied in a very quick manner at any time. The information revealed by this research will hopefully inspire others to do further research, looking into data on the level of individual people, in a causal manner if possible. Eventually, this further research may be published and used to develop coping strategies for suicidal mentalities. For instance, if a negative relationship between percent of GDP spent on education and suicide rate is discovered in the database, it might indicate that spending more of the national budget on education is one way to increase the quality of life of the population of a country, and reduce the suicide rate. This could motivate studies between individual's level of education or value of their academic competency and their mental state, or an experiment might observe the effect of temporarily removing educational resources from someone's life on their mood levels.

Some of the data was collected based on intuition that mental illness might be a significant factor in people's desire to live. Data relating the percent of people with depression, anxiety disorders, bipolar disorder, schizophrenia, eating disorders, drug addictions, and alcoholism was therefore collected.

We also wanted to assess how safety and legal security concerns influence suicide rates. For instance, people might be more likely to consider taking their own lives if their life expectancy is lower, or if they are already at relatively high risk of dying another way. Knowing that one's legal rights (if limited) can indicate institutional bias against them and might be indicative of a larger social issue, we also evaluated the general strength of individuals' legal rights, which was stored in the database as a measure from 0 to 1, with 1 representing more legal rights. This section also included mortality rate caused by traffic incidents, as the amount of deaths per 100,000 lives.

Another possible factor in whether or not someone dies by suicide is the degree to which their country is developed. It seems reasonable to assume that one's life is affected dramatically by the technological and economic standing of their nation, and that such aspects may influence their contentment with their life. Statistics that were used to evaluate the developmental level of the country include the HDI itself, the amount of governmental spending spent on education (by percent of the nation’s GDP), the percent of people with access to electricity, the fertility rate, and infant mortality rate (represented by amount of deaths by age 5 per 1000 births). Relatedly, economic factors were evaluated, including the average amount of hours that people worked during the year, the GDP, GDP per capita, the degree of institutional gender bias in wages (on a scale from 0 to 1, with 0 representing complete equality), and the percent of the total labor force that is unemployed.

Statistics related to family dynamics were also collected, with the intuition that a person's upbringing and home life can affect their mental state, and that relatives can contribute to their support group. To this end, we collected data relating to the average house size of the countries for each year, how many children were born out of wedlock, the marriage and divorce rates, how equal the rights are between men and women (on a scale from 0 to 1, with 0 representing complete equality), the mean age where mothers have their first child, the age of heads of households, and how many children have one parent, or two parents, or another guardianship arrangement.

Lastly, general population dynamics were observed, as an attempt to see if they had a strong correlation with suicide rates. The values collected include the number of displaced individuals in the country, the percent of people living in particular age ranges, the percent of the population living in rural areas, and the percent of people living in urban areas. It is worth noting that although we collected data related to age ranges and stored it in the database, and the number of people living within certain age ranges for each country by year is obtainable from the raw data we collected for suicide rates, we decided that an in-depth analysis of the way that age is associated with suicide rates is beyond the scope of this project. Thus, information related to age ranges of the countries' populations is present in the suicide_rates.csv file provided with this report, as well as age data stored in the avgNumByAge, perAgeRange, perYoungHead, and perOldHead tables in the database can be used for future analysis by people who wish to explore the database for trends, but we will not be investigating these attributes.

We chose to store the data in a relational database, as almost all of the data collected represents a specific country for a specific year, so the relationship between different attributes (via foreign key in the database) is clear and definite. Specifically, SQLite was used, because the

RSQLite package allows for easy creation of a database within the R code. Although SQLite is often criticized for not allowing the user to create a server for the database that would allow multiple users to develop accounts for and gain access to the data storage, this is not a deterrent for our project. Our goal is to investigate and research relationships and trends involving suicide rates by country-year, and to provide a tool for other researchers to do the same. We do not expect this database to be used by a large number of people, and there is no reason many people would need to use it concurrently and see the queries that others make on the database (because they should all be reads, and no writes). Thus, hosting it on a server would not provide much functionality that would change what is done with our project, so a SQLite database is an acceptable means of data storage.

```
create_table <- function(is_suicide = FALSE, df, column_names, data_types = c("VARCHAR(64)", "NUMERIC", "NUMERIC"), pk = c("Country",
"Year")) {
  print(column_names)
  print(is_suicide)
  fields <- "";
  for (i in seq(1:length(column_names))) {
    fields <- paste0(fields, column_names[i], " ", data_types[i], " NOT NULL, ")
  }
  header <- paste0("CREATE TABLE ", deparse(substitute(df)), "( ")
  pk <- paste0("PRIMARY KEY(", pk[1], ", ", pk[2], ")")
  query <- paste0(header, fields, pk)
  if (is_suicide) {
    query <- paste0(query, ", ")
    for (str in c("education_funding", "electricity", "female_life_expectancy", "fertility_rate", "infant_mortality_rate",
"male_life_expectancy", "traffic_mortality", "pop_density", "pop_growth", "rural_pop", "legal_rights", "unemployment", "urban_pop",
"income_inequality", "marriage_inequality", "hours_worked", "gdp", "mental", "hdi", "housedemo", "displacedpersons", "averageHouseSize",
"avgNumByAge", "bastardChild", "marriage", "meanAgeAtBirth", "parents", "perFHead", "perMember", "perOldHead", "perYoungHead",
"perAgeRange")) {
      query <- paste0(query, "FOREIGN KEY (Country, Year) REFERENCES ", str, " (Country, Year)")
      if (str != "perAgeRange") {
        query <- paste0(query, ",")
      }
    }
  }
  query <- paste0(query, ");")
  dbSendQuery(db, query)
}
```

the code to abstract the creation of tables based on their column names and data types


```
create_table <- function(is_suicide = FALSE, df, column_names, data_types = c("VARCHAR(64)", "NUMERIC", "NUMERIC"), pk = c("Country",
"Year")) {
  print(column_names)
  print(is_suicide)
  fields <- "";
  for (i in seq(1:length(column_names))) {
    fields <- paste0(fields, column_names[i], " ", data_types[i], " NOT NULL, ")
  }
  header <- paste0("CREATE TABLE ", deparse(substitute(df)), "( ")
  pk <- paste0("PRIMARY KEY(", pk[1], ", ", pk[2], ")")
  query <- paste0(header, fields, pk)
  if (is_suicide) {
    query <- paste0(query, ", ")
    for (str in c("education_funding", "electricity", "female_life_expectancy", "fertility_rate", "infant_mortality_rate",
"male_life_expectancy", "traffic_mortality", "pop_density", "pop_growth", "rural_pop", "legal_rights", "unemployment", "urban_pop",
"income_inequality", "marriage_inequality", "hours_worked", "gdp", "mental", "hdi", "housedemo", "displacedpersons", "averageHouseSize",
"avgNumByAge", "bastardChild", "marriage", "meanAgeAtBirth", "parents", "perFHead", "perMember", "perOldHead", "perYoungHead",
"perAgeRange")) {
      query <- paste0(query, "FOREIGN KEY (Country, Year) REFERENCES ", str, " (Country, Year)")
      if (str != "perAgeRange") {
        query <- paste0(query, ",")
      }
    }
  }
  query <- paste0(query, ");")
  dbSendQuery(db, query)
}
```

the code to abstract the creation of tables based on their column names and data types

```
insert_data <- function(df, is_hdi = FALSE) {
    for (i in 1:nrow(df)) {

    query <- paste0("INSERT INTO ", deparse(substitute(df)), " VALUES (");
    for (j in seq(1:ncol(df))) {
      if (j == 1 || (is_hdi && j == ncol(df))) {
        query <- paste0(query, "'");
      }
      query <- paste0(query, df[i, j])
      if (j == 1|| (is_hdi && j == ncol(df))) {
        query <- paste0(query, "'");
      }
      if (j == ncol(df)) {
        query <- paste0(query, ");")
      } else {
        query <- paste0(query, ", ")
      }
    }
    print(query)
    dbSendQuery(db, query)

  }
```

The code to abstract insertions into the database

Another important decision we needed to make was how to deal with NA values. Because the primary key for our tables was the composition of the country and the year, we decided that it was not important to portray every country for the same range of years, depending on the data present for them. This is to say that if one country is represented for a shorter time range than a second country, it is acceptable to portray both countries for the range that is appropriate for each individually, and it is preferable to do that than to extrapolate in order to estimate more data for the first country or to limit the representation of the second country unnecessarily. When there is data missing for a country for a specific year, but there is data present for that country in at least one later year and one earlier year, then a linear model based on the other data present for that country will be created and used to impute the missing values. In other words, for each country, “trailing” NA values were removed, and other NA values were imputed with a linear model. (Our first attempt to impute data involved the use of the mice package, but due to difficulties discussed later in this report, we decided to use the lm function from the stats package to create a linear model instead.)

One of the difficulties of this project came from how the suicide data was stored in the csv file that we downloaded. Our desire was to store the suicide rate by country and by year, and

to have a representation of the total suicide rate, male suicide rate, and female suicide rate. The data that we used for suicide rates did not report a suicide range, instead giving an amount of suicides and a population for the appropriate group, and stratified the data by country, year, sex, and age range. We decided that we were not interested in exploring the age distribution of populations when evaluating the suicide rates, so getting the data table into an easily usable format involved accumulating rows for the same combination of country, year, and sex, spreading the rows for different sex into their own columns, and dividing the number of suicides for a group by the population of that group to convert the measures from counts to rates.

```
for(index_country_year in seq(1:nrow(suicide_rates_inter))) {
  only_one_country <- suicide_rates %>% filter(suicide_rates$country.year ==
suicide_rates_inter$country.year[index_country_year])
  only_one_country_m <- only_one_country %>% filter(sex == "male")
  only_one_country_f <- only_one_country %>% filter(sex == "female")

  sum_suicides <- sum(rm.na = TRUE, x = only_one_country$suicides_no)
  sum_suicides_m <- sum(rm.na = TRUE, x = only_one_country_m$suicides_no)
  sum_suicides_f <- sum(rm.na = TRUE, x = only_one_country_f$suicides_no)

  population_total <- sum(rm.na = TRUE, x = only_one_country$population)
  population_m <- sum(rm.na = TRUE, x = only_one_country_m$population)
  population_f <- sum(rm.na = TRUE, x = only_one_country_f$population)

  suicide_rate <- sum_suicides / population_total
  suicide_rate_m <- sum_suicides_m / population_m
  suicide_rate_f <- sum_suicides_f / population_f

  suicide_rates_inter$male_suicide_rate[index_country_year] = suicide_rate_m
  suicide_rates_inter$female_suicide_rate[index_country_year] = suicide_rate_f
  suicide_rates_inter$total_suicide_rate[index_country_year] = suicide_rate
  suicide_rates_inter$population[index_country_year] = population_total
  suicide_rates_inter$male_population[index_country_year] = population_m
  suicide_rates_inter$female_population[index_country_year] = population_f

}
```

the code to aggregate suicide rates by male, female, and total across age ranges

ï..country	year	sex	age	suicides_no	population	suicides.100k.pop	country.year	HDI.for.year	gdp_for_year....	gdp_per_capita....	generation
Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NA	2,156,624,900	796	Generation X
Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NA	2,156,624,900	796	Silent
Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NA	2,156,624,900	796	Generation X
Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NA	2,156,624,900	796	G.I. Generation
Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NA	2,156,624,900	796	Boomers
Albania	1987	female	75+ years	1	35600	2.81	Albania1987	NA	2,156,624,900	796	G.I. Generation
Albania	1987	female	35-54 years	6	278800	2.15	Albania1987	NA	2,156,624,900	796	Silent
Albania	1987	female	25-34 years	4	257200	1.56	Albania1987	NA	2,156,624,900	796	Boomers
Albania	1987	male	55-74 years	1	137500	0.73	Albania1987	NA	2,156,624,900	796	G.I. Generation
Albania	1987	female	5-14 years	0	311000	0.00	Albania1987	NA	2,156,624,900	796	Generation X

the first 10 rows of the suicide rates table before the cleaning code is run

Country	Year	male_population	male_suicide_rate	female_population	female_suicide_rate	population	total_suicide_rate
Albania	1987	1392701	3.518343e-05	1316901	1.974332e-05	2709601	2.731029e-05
Albania	1988	1420701	2.956287e-05	1343601	1.711818e-05	2764301	2.315233e-05
Albania	1989	1439801	3.750518e-05	1363301	1.173622e-05	2803101	2.461560e-05
Albania	1992	1399301	2.429785e-05	1423201	1.053962e-05	2822501	1.700619e-05
Albania	1993	1379901	3.406041e-05	1427401	1.961607e-05	2807301	2.635984e-05
Albania	1994	1404201	2.563736e-05	1445101	1.107189e-05	2849301	1.789913e-05
Albania	1995	1429601	3.847227e-05	1473801	2.374812e-05	2903401	3.065371e-05
Albania	1996	1444201	3.531364e-05	1496001	2.673795e-05	2940201	3.061015e-05
Albania	1997	1448401	8.215957e-05	1528901	3.466542e-05	2977301	5.743457e-05
Albania	1998	1475401	6.574484e-05	1537301	3.837895e-05	3012701	5.144885e-05

the first 10 rows of the suicide rates table after the cleaning code is run

When ascertaining which attributes we should accumulate into our linear model, we first had to determine the attributes that would be worthwhile to include. To do this, we built a series of linear models for each characteristic that we collected data for, each predicting the total suicide rate for a country for a specific year based only on that one characteristic. (It is worth noting that we had to do an inner join between the suicide rates table and the table that carried that characteristic, so the set of country years represented in the data is different in each case- sometimes this difference is slight, but in some cases the year range is severely limited, or there are a notable number of countries missing.) To determine that the characteristic was significant enough to include in our model, we abided by the standard that the p-value had to be less than 0.05, as this means that there is a less than 5 percent chance that there is no relationship between the suicide rate for the country and the value of the studied characteristic. We noted after creating these models that almost all of our attributes satisfied this condition, but the r-squared value of the linear models were extremely low. The r-squared value of a linear model measures the degree to which the variation in the predicted value can be associated with variation in the predictor value, meaning that even if there is a relationship between the attribute and suicide

rates, these values are not going to be useful in a predictive model. We thus also determined that we would only include attributes for which the r-squared value was larger than 0.1.

```
electricity_join <- dbGetQuery(db, "SELECT * FROM electricity JOIN suicide_rates ON electricity.country = suicide_rates.Country AND
electricity.year = suicide_rates.Year")
pred_electricity <- lm(data = electricity_join, formula = total_suicide_rate ~ percent_of_people_with_access_to_electricity)
summary(pred_electricity)
```

the code for assessing electricity access for placement in the linear model

```
Call:
lm(formula = total_suicide_rate ~ percent_of_people_with_access_to_electricity,
    data = electricity_join)

Residuals:
       Min         1Q     Median         3Q        Max
-1.300e-04 -6.408e-05 -1.641e-05  4.341e-05  3.801e-04

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -1.037e-04  2.518e-05  -4.118 3.98e-05 ***
percent_of_people_with_access_to_electricity  2.341e-06  2.602e-07   8.996  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.919e-05 on 1881 degrees of freedom
Multiple R-squared:  0.04125,   Adjusted R-squared:  0.04074
F-statistic: 80.93 on 1 and 1881 DF,  p-value: < 2.2e-16
```

The output of the linear model created by the code above- note that the p-value is less than 0.05, so it is significant, but the r-squared value is less than 0.1, so we will not include it in the linear model.

name_of_predictor	p_value	r_squared_value
pred_AgeRange_7	8.009698e-44	0.2914131
pred_AgeRange_8	4.415479e-46	0.3043993
pred_AgeRange_9	3.782871e-44	0.2933009
pred_alcohol	1.780848e-101	0.2181590
pred_averageHouseSize	4.571545e-61	0.2749214
pred_divorce	1.277693e-33	0.1518748
pred_fertility_rate	2.073877e-95	0.1823476
pred_FifteenMinus	3.003515e-47	0.3110188
pred_perFHead	3.189342e-34	0.1766435
pred_perMemberOne	9.059674e-66	0.2861161
pred_perMemberTwoOrThree	6.633818e-52	0.2320019
pred_perMemberFourOrFive	2.013930e-58	0.2579274
pred_perMemberSixPlus	4.903922e-57	0.2524856

Some of the attributes with a usable p-value and r-squared value, such that they can be used in the linear model.

Our process of transforming the data and removing variables from the linear model based on their distributions is explained further in the Rmd near the code used to develop the normal model. We also removed all of the data that was related to the age range of people in the country, consistent with our decision not to analyze age distributions. After training data and testing data were separated, we went through the process of improving these linear models as described in class, removing the attribute that had the highest p-value until all of the attributes in the linear model had a p-value of less than 0.05.

The equation for the linear model was $Suicide\ Rate\ Estimate = -2.969 * 10^{-3} + 1.933 * 10^{-4}\ (Average\ House\ Size) + 4.051 * 10^{-5}\ (Divorce) + 6.827 * 10^{-6}\ (Percent\ of\ People\ Living\ Alone) + -3.478 * 10^{-5}\ (Rate\ of\ Population\ Growth) + 2.329 * 10^{-5}\ (Percent\ of\ Children\ Living\ With\ Two\ Parents) + 2.723 * 10^{-5}\ (Percent\ of\ Children\ Living\ With\ A\ Single\ Parent)$. The model estimated a suicide rate within 5% of the actual suicide rate in approximately 10% of cases, within 10% of the actual suicide rate in approximately 26% of cases, and within 20% of the actual suicide rate in approximately 53% of cases. Although we would hope that the model would predict more accurately than this, this is effective enough for understanding the trend of the relationship these attributes have with the suicide rate, and thus serves to identify which attributes are worthy of future research. (Indeed, even the process of iterating through linear models as we attempt to make it more accurate is, in a sense, ranking the attributes based on level of indication that there is a relationship worth studying, based on the p-value. These other variables are clearly shown within the code when developing the linear model.) We also calculated the root mean squared error, which is the standard deviation of the residuals that the model created. Our value for the

root mean squared error was $4.658 * 10^{-5}$, which is fairly large for the data we collected, indicating that our linear model was not incredibly accurate.

```
Call:
lm(formula = total_suicide_rate ~ AVGHouseSize + Divorce + PercentOne +
    population_growth + TwoParents + SingleParent, data = mass_join_without_ages_training)

Residuals:
       Min         1Q     Median         3Q        Max
-1.011e-04 -3.079e-05 -1.451e-06  2.016e-05  1.360e-04

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.969e-03  6.497e-04  -4.569 1.55e-05 ***
AVGHouseSize      1.933e-04  4.658e-05   4.150 7.53e-05 ***
Divorce           4.051e-05  8.804e-06   4.602 1.37e-05 ***
PercentOne        6.827e-06  1.928e-06   3.542 0.000633 ***
population_growth -3.478e-05  5.721e-06  -6.078 2.88e-08 ***
TwoParents        2.329e-05  6.019e-06   3.869 0.000207 ***
SingleParent      2.723e-05  6.520e-06   4.177 6.83e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.836e-05 on 90 degrees of freedom
Multiple R-squared:  0.5699,    Adjusted R-squared:  0.5412
F-statistic: 19.87 on 6 and 90 DF,  p-value: 1.182e-14
```

The summary of the linear model

For the sake of ease of presentation, the section of this report that visualizes the attributes that we discovered were relevant for suicide rates (as well as presenting a few other visualizations) is included in VisualPlayground.Rmd, or the knitted PDF, VisualPlayground.pdf. Please look to this file to see the visualizations.

Another difficulty we encountered was using the mice package to impute data. Although this is a commonly-used package with a lot of customizability (if not user-friendliness) for how one can replace NA values in their data, there were significant barriers to using the functions provided by this library. Notably, if the function is called when there is no data to impute, instead of doing nothing, the function throws an error, stopping the entire data imputation process. Furthermore, although we did not want the missing values of the attributes to be affected by the year or the country name for that observation, the mice function will not run on only a single column. We did not know how the year or country name would affect the value of the attribute,

which made us weary about how mice would replace the NA values in the data. For these reasons, we thought it would be more appropriate to use a linear model to replace the missing data values.

Because of this project's exploratory nature, we needed to gather data on country years over many different variables. Unfortunately, not all of this data was easy to obtain. While most of our data allowed for downloads of the dataset, and then just had to be cleaned, the OECD database for family structure did not follow this ruleset. While we could download its dataset, it would only have a single country over all the years recorded, and not all the countries in the database. Because of the number of countries this database held, individually downloading them was just not feasible. While web scraping with rvest was possible, the data for any given country was not accessible from unique urls, nor all in one url. So we needed a way to automatically web scrape while retaining the ability to activate javascript commands on the webpage, these commands modify the html document and are thus required to extract all countries' data. R-Selenium allowed us to do this quite easily. Once set up, it allows very easy commands such as 'getElementByID' that allows us to identify the element we need to activate to run the javascript. In addition, the 'getPageSource' command allowed us to pull all the html data to then use in RVest. This was necessary because R-Selenium is relatively slow compared to RVest because all commands to identify data used in R-Selenium are http requests, and require communication between the client and the server. This allowed us to only have to communicate with the server when needing a new html document, and never communicate when just pulling data from that document. All this in mind, R-Selenium was quite difficult to set up. We had to install an older version of Chromedriver, set up that older version in my system directory, and specify that version explicitly for R-Selenium in order to run. While this fix was simple, figuring out it was

necessary took a lot of time. A more fatal problem with R-Selenium, however, is that it does not have a concept of 'ReadyState', only a 'implicit wait timer'. The difference is that a readystate allows us to read the current status of the webpage (whether it be loaded, or javascript still running, etc) while an implicit wait timer only lets me us set a maximum amount of time to wait before it times out. In addition, R-Selenium only has a notion of 'script implicit wait time' and 'load implicit wait time'. Load refers to the webpage loading, and script is your r-code's script run time. Neither covers javascript loading, which is the time it takes for the javascript command to query the server and then update the html document. Without this ability, there was no automated way (outside of some type of thread.sleep command) to prevent the immediate running of the next r-script command, which will inevitably through an error as the html document has not yet been updated. To get around this problem, we set up a user input where the r-code stops running until the user presses enter. This allows us to wait for the javascript to finish running and the http request to be fulfilled without first predicting how long it will take.

Some of the insights that we gained from this project came from learning to work with the technology. We are more familiar with how to use R-Selenium now, which will be useful in future projects scraping data from websites with confusing layouts. We also gained knowledge on how to perform queries on a SQLite database using the RSQLite package, experience with doing data manipulation in R to tidy and clean ill-formed tables, and an appreciation for the vast functionality offered by ggplot for producing professional and suggestive visualizations. We also gained a lot of insight into the relationships in the data. Although we cannot name every notable relationship that we discovered due to the amount of observations and attributes we collected data for, one interesting finding was that suicide rates for males was consistently around 4 times the size of suicide rates for females across all countries for all years. The attribute for which we

saw the starkest difference between male suicide rate and female suicide rate was divorce. We also noted that depression was not as closely related to suicide rate as we originally believed, which is surprising, given how associated people tend to think they are. We noted that five of the six attributes present in the final iteration of the linear model have to do with family dynamics. These five traits are the percent of people living alone, the average house size, the divorce rate, the percent of children living one parent, and the percent of children with living two parents. (Note that the last two columns are not inverses of one another, as children with non-traditional families, homeless children, children living with non-parent family members, and children that are unaccounted for do not fit into either of these categories.) From this, we believe that family dynamics and individual's perspective on their families would be a good area of future research.

There are many other directions for further exploration and research related to this project as well. One potential source of trends that we did not uncover might be in the data related to age, such as the percent of households with young heads, the percent of households with old heads, and the general age distribution within the population. This data is already collected in the database, so performing these analyses would be simple. Also, as has been referenced earlier in this report, the trends that we discovered in this project should be investigated and research on the individual level to indicate more definitively an association for a person's choice to end their lives. Although we judged these to be beyond the scope of our analysis for this project, further trends might be noticeable when observing just the relationship between the attributes and the male suicide range or the female suicide range. Because the suicide rates table already has male suicide rates and female suicide rates separated for the countries by year, users could easily create linear models to determine which attributes have a p-value of less than 0.05 and an r-squared value of greater than 0.1, then iterate through the process of determining which variables

are appropriate for keeping in the linear model (as in, they are normally modeled well enough and have a p-value of less than 0.05 when in the linear model with the other attributes) for male suicide rate and female suicide rate. Furthermore, because the hdi table categorizes countries by their HDI, users can find data only related to countries within a specific HDI category (being Very Low, Low, Medium, High, and Very High). Linear models can then be created from these separations, and it might be discovered that developing countries are more influenced by different variables than countries that are already developed. For a different goal using this data set, a knn analysis can be performed to determine the HDI category of a country based on other attributes of the country. Lastly, if an attribute that we did not collect data is desired for inclusion in the exploration (such as any of the Hofstede insights values of hours of sunlight during the year), it can be added to the database fairly easy by reading the file in and tidying it by using tidy_data, removeTrailingNA, and cleancountry functions, and then inserted into the database using the create_table and insert_table functions.

Works Cited

"Access to Electricity (% of Population)." *The World Bank*, The World Bank, data.worldbank.org/indicator/EG.ELC.ACCS.ZS?view=chart. ***

"Government Expenditure on Education, Total (% of GDP)." *The World Bank*, The World Bank, data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS?end=2018&start=1998&view=chart. ***

"Household." *United Nations*, United Nations, population.un.org/Household/index.html#/countries/840.

"Human Development Reports." *Human Development Index (HDI)*, United Nations Development Programme, hdr.undp.org/en/content/human-development-index-hdi.

"Life Expectancy at Birth, Female (Years)." *The World Bank*, The World Bank, data.worldbank.org/indicator/SP.DYN.LE00.FE.IN?view=chart. ***

"Life Expectancy at Birth, Male (Years)." *The World Bank*, The World Bank, data.worldbank.org/indicator/SP.DYN.LE00.MA.IN?view=chart. ***

"Mortality Caused by Road Traffic Injury (per 100,000 People)." *The World Bank*, The World Bank, data.worldbank.org/indicator/SH.STA.TRAF.P5?view=chart. ***

"Mortality Rate, under-5 (per 1,000 Live Births)." *The World Bank*, The World Bank, data.worldbank.org/indicator/SH.DYN.MORT?end=2017&start=1985&view=chart. ***

OECD (2019), Discriminatory family code (indicator). doi: 10.1787/7ce07d5f-en (Accessed on 17 April 2019) (https://data.oecd.org/inequality/discriminatory-family-code.htm)

OECD (2019), Hours worked (indicator). doi: 10.1787/47be1c78-en (Accessed on 17 April 2019) (https://data.oecd.org/emp/hours-worked.htm)

OECD (2019), Income inequality (indicator). doi: 10.1787/459aa7f1-en (Accessed on 17 April 2019) (https://data.oecd.org/inequality/income-inequality.htm)

Oecd. "Family Database: By Country - The Structure of Families." *Family Database : By Country - The Structure of Families*, Organization for Economic Co-Operation and Development, stats.oecd.org/index.aspx?queryid=68249.

"Population Growth (Annual %)." *The World Bank*, The World Bank, data.worldbank.org/indicator/SP.POP.GROW?view=chart. ***

Ritchie, Hannah, and Max Roser. "Mental Health." *Our World in Data*, University of Oxford, Apr. 2018, ourworldindata.org/mental-health.

Ritchie, Hannah, and Max Roser. "Natural Disasters." *Our World in Data*, University of Oxford, 3 June 2014, ourworldindata.org/natural-disasters.

Rusty. "Suicide Rates Overview 1985 to 2016." *Kaggle*, Kaggle, 1 Dec. 2018, www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016. ***

"Rural Population." *The World Bank*, The World Bank, 2018, data.worldbank.org/indicator/SP.RUR.TOTL?end=2017&name_desc=false&page=1&start=1972. ***

"Strength of Legal Rights Index (0=Weak to 12=Strong)." *The World Bank*, The World Bank, data.worldbank.org/indicator/IC.LGL.CRED.XQ?name_desc=false&page=1. ***

"Unemployment, Total (% of Total Labor Force) (Modeled ILO Estimate)." *The World Bank*, The World Bank, data.worldbank.org/indicator/SL.UEM.TOTL.ZS?view=chart. ***

"Urban Population (% of Total)." *The World Bank*, The World Bank, data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?end=1973&name_desc=false&page=1&start=1972. ***

******* Note: These datasets were originally collected by the World Health Organization.