Emily Cao

April 14, 2022

Chris Teplovs

SI 330: Data Manipulation

How Happy are the Countries?

Motivation

The project I am presenting is comparing each country's suicide rate with the happiness score. While there is a lot of information on the country's relation to the happiness score, I want to also determine what regions of the countries that the world has identified to their cumulative happiness score. I decided to do this project because there has been an increase in the suicide rate in all countries, and I wonder how happy the country is if the people feel the desire to suicide. One goal I want to achieve is finding the top ten happiest countries based on their suicide rate and happiness score to determine how content each country is. The main goal of the project that I wanted to address was determining the happiness of each region and country with their suicide rate and happiness score.

Data Sources

One of the data sources that I plan to access, manipulate and bring together is "List of countries by suicide rate" from Wikipedia. I would be grabbing all 183 countries' total, female, and male suicide rates from the year 2015 to 2019, which is 2,745 records in total. The data source is located at https://en.wikipedia.org/wiki/List_of_countries_by_suicide_rate and can be accessed by going on the link through a device with internet access. The format of this data source is fetching and parsing the tables on the HTML page and making the data into a data frame with pandas' function read_csv to be used in the project. This data has consisted of various tables on the suicide rate that is categorized under all suicide rates from the most recent year that the page has been updated (2019), the male suicide rate from the year 2000 to 2019, and the female suicide rate from the year 2000 to 2019. The most important variables contained in this dataset are the countries and the suicide rate based on the year. I only retrieved each sexes' suicide rate from the year 2015 to 2019 because the second dataset I will be using only has information on the country's happiness from the year 2015 to 2019. There are 183 countries for each table that will be used.

The other data source is the "World Happiness Report" by Sustainable Development Solutions Network. I will be using all the CSV files, and the size is 80.86kB. There are five CSV files of 158 countries with 10 specific data for each country, resulting in 7,900 records, in total. The data is located at https://www.kaggle.com/unsdsn/world-happiness and can be accessed by going on the link through a device with an internet connection. First, an account must be made by registering with an email address to download the CSV files. On the page, the files can be downloaded and accessed by clicking on the button with the word "Download". The format of this data source is to read all of the CSV files and format the data into a data frame with the pandas' function read_csv to be used in the project. The most important variables contained in this dataset that I will be using are the Country, Region, Happiness Score, and Happiness Rank. However, other variables are significant to mention, which are economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute. The datasets from the year 2017 to 2019 don't have the region in their data, but I was able to manipulate with, which will be discussed late, to include the regions for these countries during these years.

Due to other data that needed to be collected to create more accurate calculations, I had to use other dataset resources. I grabbed 2015 and 2019's population data from their specific Wikipedia resource. For 2015, I used the website, https://en.wikipedia.org/wiki/List_of_countries_by_population_in_2015, to grab the population for each country. There are 198 countries and its population listed. There were other variables in the table, such as Change from 201, Area (km^2), and Population Density, but I only needed the population and the country. For 2019, I used the website, https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations), to grab the population for each country. There were other variables in the table, such as UN continental Region, UN statistical subregion, Population (1 July 2018), and Change. I could've used July 2018 population on the Kaggle dataset for the year 2018, but I wanted to only have the first and last files given by Kaggle to make it not seem random picking of the years. I couldn't find any data with lots of countries' populations for the years 2016 and 2017, so I only used the ones I could find. To clean up some data on the population column in the years 2015 and 2019, I had to grab population data for specific countries from https://countryeconomy.com.

Data Manipulation

For the Suicide Rate resource, I had to manipulate the data to grab only the second row and onwards because the first row was only null values. I noticed that some countries had an asterisk after it because, in Wikipedia, it means that the country's information was cited. To get rid of the asterisk, I used replaced the asterisk with a value of nothing. I noticed that some countries are aligned with one another, but because of the name that the resource set it as, the merge on the country didn't detect it. To make an accurate analysis and calculation, I had to change a lot of the countries' names to the correct title, the current title of the country. I had to Google search a lot of the countries to figure out what their current title is and some examples of countries I had to change are Swaziland to Eswatini and Timor-Leste to 'East Timor. I did this by converting the Countries into strings and replacing the old title with the new title. Since the main data is in the World Happiness Report Kaggle resource, a lot of the data manipulation was done on those data.

For the World Happiness Report Kaggle resource, a lot of manipulation had to be done. Most of the data manipulation consists of me having to change the country's name and for some of the data that doesn't have the region, I have to add the regions that are missing. I searched for the country with null values after merging the specific year data with the Wikipedia data to see what countries don't have data. The countries' name in the data is well-updated, so there were no changes to the correct title for the countries. However, there was some countries' name that needed to be adjusted to align with the Wikipedia page. Some examples are Congo (Kinshasa) to DR Congo and Congo (Brazzaville) to Congo, and I had to search up the country from the Wikipedia resource to see if there are countries with the title. I also noticed some regions were missing when I used .isnull() on the data frame, so I had to Google search each null value in the region to replace it with the accurate region, so future calculations were done accurately. I added the population data that I grabbed from the other Wikipedia resource to do accurate calculations. However, there was some population that wasn't shown, so by using .isnull() on the specific column on the data frame, I determined which populations I had to retrieve from outside resources, which is what I did with the help of https://countryeconomy.com.

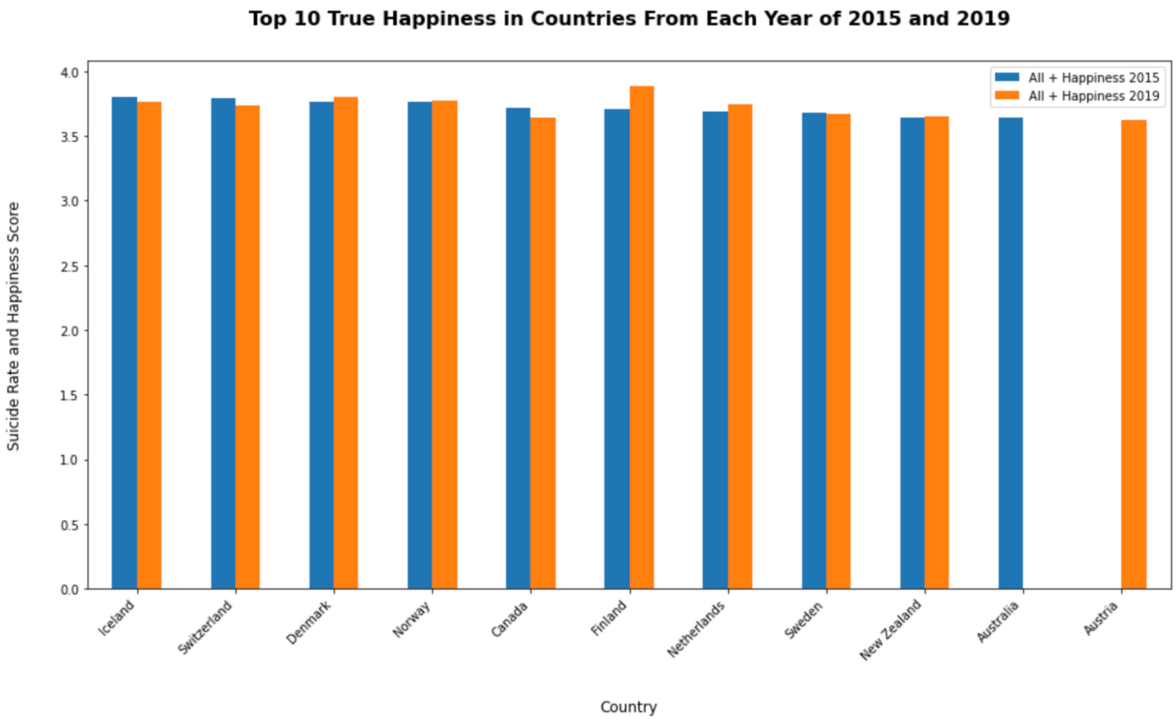
I joined the two data sets mainly the country because that was the only thing that both datas have in common. For retrieving data from the resources is to use .isnull() to retrieve any null value and determine if there is any way of grabbing data from a third or fourth resource to make accurate calculations and visualizations. The main challenge I encountered was to determine whether the country is related to each other because some of the countries such as Macedonia renamed themselves North Macedonia in 2019 and I have to change all the data frames to be named North Macedonia with .isin() by searching for Macedonia in the Country column and making the country equal to the value I want to change, "North Macedonia". These are minor challenges, but they are tedious because it results in me having to Google search up whether or not I should make the changes. One of the countries that I had to difficult with determining whether to change it and decided not to was "Somaliland region" in the Kaggle resource. I wasn't sure whether to change it to Somalia, which is what is listed in the Wikipedia resource, but I decided not to after a careful search of the Country's history.

The main workflow is to start downloading ipympl for the interactive scatterplot graph that will be used in the visualization. The start of the code is to grab all the data from their respective resource. From this, data manipulation of changing the country's name happens to make sure it is formatted to its current title. The population data is grabbed and there are populations for specific countries following the retrieval of the Wikipedia population data, so it can be used later on for making sure the calculations are done correctly. Afterward, the first data manipulation is made, in which the top 10 true happiest countries are found, in which a new data frame is created to be used for visualization by adding the population to the years 2015 and 2019 and calculating the average happiness score and true suicide rate. Following this, the second data manipulation for the ratio of the combination of the suicide rate for each region to the number of countries from the region for each year is created, in which I find the sum of the suicide rate for each region for each year. A new data frame is developed. Soon, the third data manipulation of finding each region's average happiness rank for each year is developed, in which I find the sum of the rank for each region and divide it by the number of countries within that respective region. Later on, the fourth data manipulation of finding the lowest-ranked and highest-ranked countries in their respective region for each year it used to compare their happiness rank. Additionally, each year from 2015 to 2019, the male suicide rate, female suicide rate, and happiness score for each country are created into a data frame.

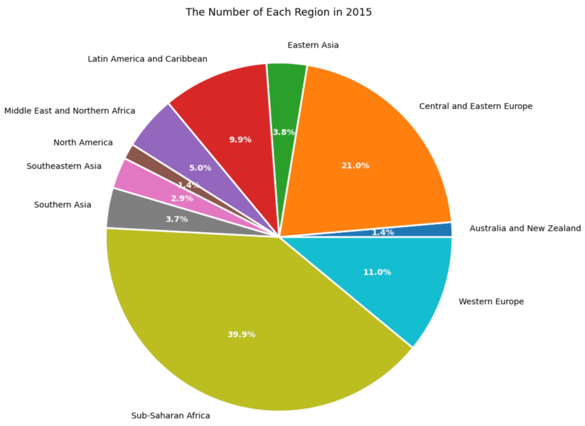
Finally, visualizations are made for each form of data manipulation, according to what the visualizations ask for. For example, data visualization #2 requires me to graph all the years, whereas #1 only needs me to graph one figure of the bar graphs for data manipulation #1. For visualization #3, I have to grab the USA happiness rank and I had to adjust the format of the data frame for the average happiness rank for each region. For visualization #4, I graphed the line graph as it is according to the data frame of lowest-ranked, highest-ranked, or the combination. For visualization #5, an interactive scatterplot graph is created to visualize the cluster of suicide rates and find the outliers based on the male, suicide rate, female suicide rate, and happiness score.

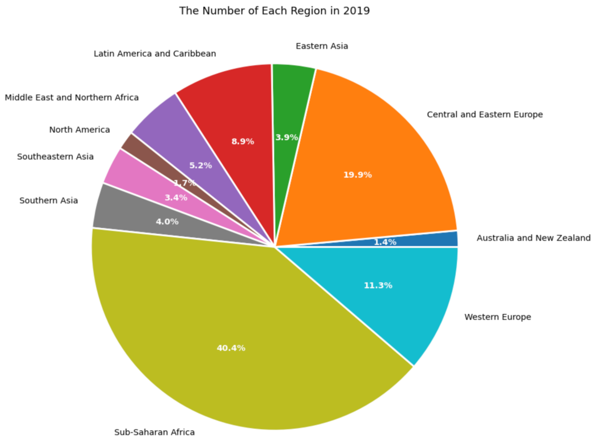
Analysis and Visualization

One interesting analysis and visualization I wanted to present is whether a country is truly as happy as Kaggle said it is with the combination of the suicide rate for all sexes and the happiness score. As seen in key 36 and key 50, I multiplied the suicide rate by .01 for a whole number rather than the rate. Then, I multiplied it by 100,000 because as indicated on the Wikipedia page, the rate is by every 100,000 people. To determine the true happiness, I added the number from the converted suicide number of people for 2015 and 2019 with the happiness score for their respective countries and year. I didn't have to divide the number by 2, but from my logic that there are two resources, I decided to do so. One interesting thing I noticed from this visualization is that most of the countries listed in 2015 are in 2019's list. As seen from the visualization, Australia in 2015 and Austria in 2019 are the ones listed in the top 10 that are not in both data's top 10. Another interesting thing I noticed is that the rank for each country listed are only a few margins off and are ranked high. For example, in 2015, Iceland is the first ranked in the suicide rate and happiness score, whereas in Kaggle, Iceland is ranked second. There are not a lot of drastic differences as seen in the highest-ranked countries with suicide rates, but it is interesting to see in 2019 that Canada ranked 9 in Kaggle is ranked #5 and Finland ranked #1 in Kaggle is ranked #6. This shows that the suicide rate does factor into the true happiness of each country.

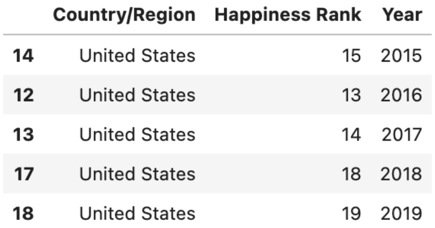


Another interesting analysis and visualization for the comparison of all the suicide rates of a region to the number of countries in the region each year is to see how each region compares to each other in terms of how many people suicide based on the cumulative suicide rate of the region. Since there are no specific data on the cumulative suicide rate for each region, I had to find it for each region and each year. To prevent having to do it repeatedly, a function was made to help get the sum for each region based on the year given (key 59). This was made into a dictionary and then was converted into a data frame, so it can be used for future mergers and calculations (key 61). After making a new data frame for each year with the region and cumulative suicide rate for each year, I merged the data frames to create a pie chart for each data frame on the suicide rate for each region based on its year. As seen in the pie charts, Sub-Saharan Africa has the most suicide rate each year. Based on Google Search, Africa is the most suicidal country, so it made sense for the pie chart to show this information. It is saddening to see this because the people's decision to suicide is due to the lack of resources available to them as they are one of the continents that is the most underdeveloped. As seen, North America seems the least suicidal rate. Also, it must be taken into account that North America has fewer countries than the other regions listed, so while this looks complete visually, it is not the most accurate.





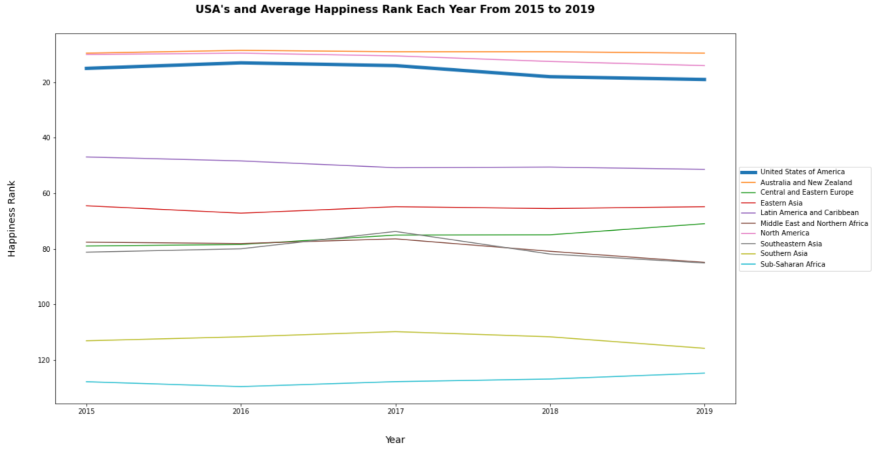
Additionally, I wanted to compare the USA with all the regions' happiness ranks. Each country has a happiness rank and I wanted to find the region's average happiness rank for each year to help visualize each region compare to each other in terms of its ranking. This will be later used in the future visualization to compare the USA's happiness rank to all the regions' average happiness rank for each year. To avoid having to repeatedly code on retrieving the average rank for each year and each region, a function was created to help seamlessly do that with only needing the data frame of Country, Happiness Rank, and Region for each year and the year for the particular data frame (key 69). I merged all the data frames and to calculate the average happiness rank to the number of countries and their happiness rank, a function was created to go through the merged data frame made before and create a new data frame to show the average rank. On keys 95-96, the USA data frame was created to grab the rank each year. Since the USA data frame has the country, happiness rank, and year indicated in the format shown



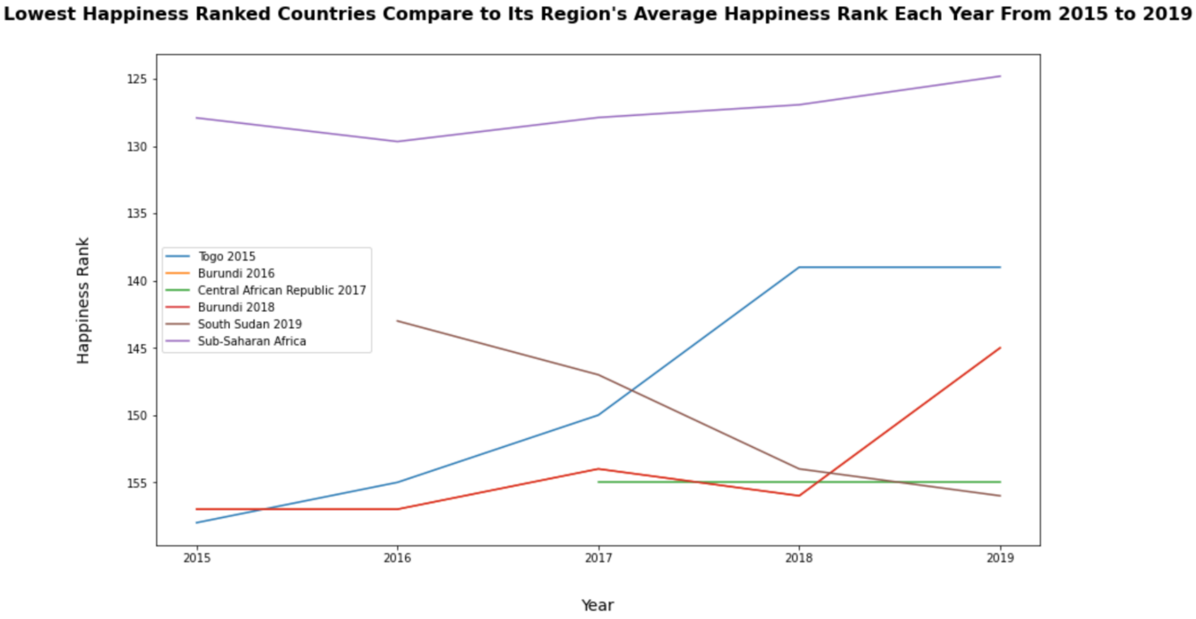
	Country/Region	Happiness Rank	Year
14	United States	15	2015
12	United States	13	2016
13	United States	14	2017
17	United States	18	2018
18	United States	19	2019

I had to convert the data frame for the average happiness rank for the regions to be the same, which is what I did in keys 98 - 100. It is interesting to see that the United States of America is right under North America's average happiness rank and higher than most other regions. The United States of America has a lot of resources from education to job opportunities, so it made sense for the US to be ranked high. However, it is surprising to see that they are ranked higher than Western Europe because most of the countries in Western Europe are highly ranked.

However, as discussed before, other regions have a lot more countries, so we have to take that into account.

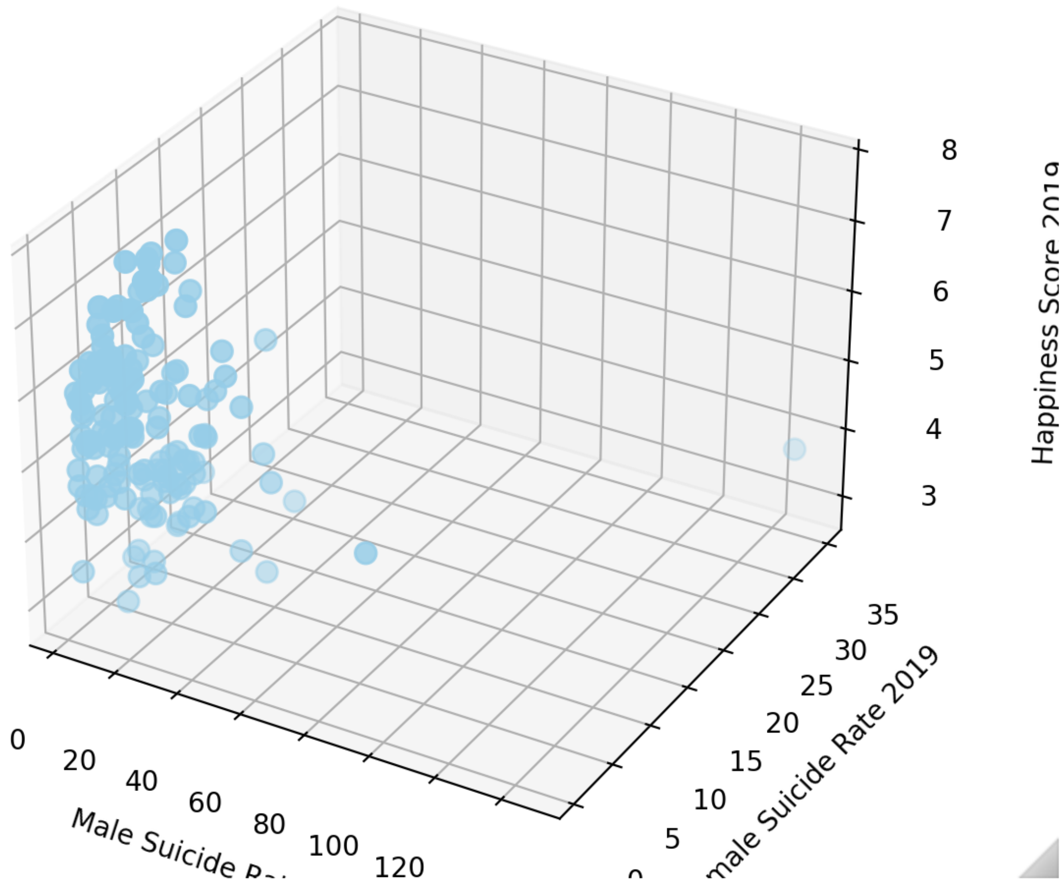


Furthermore, I wanted to compare the highest and lowest-ranked countries with their regions for each year. Each country has a happiness rank and I wanted to find the region's average happiness rank for each year to help visualize each region compare to each other in terms of its ranking. This will be later used in the future visualization to compare the USA's happiness rank to all the regions' average happiness rank for each year. To achieve this, a function was made to avoid having to duplicate the same codes. Since grabbing the specific country and region's rank is very specific, the first chunk before the for loop is used to grab only the happiness rank for each country and the region for each year (key 76). Surprisingly, the highest and lowest-ranked countries from each year were all from the same regions, so there was only one region when combing the data together for each the lowest and highest rank. After the for loop, the calculation of the average happiness rank of all the countries in each region was made, which became the last five columns. The new data frame became only the specific country of the year based on the "num", of which 1 is the highest-ranked and -1 is for the lowest-ranked country. After merging all the data into their respective data frame for low-ranked, highest-ranked, and low versus high, I plotted a line graph to show the change and differences. It is interesting to how different the lowest-ranked countries are plotted. The average rank for Sub-Saharan seems to be higher than all the lowest-ranked countries each year, which shows that there other countries within the region that are happier than their counterparts that are listed as the lowest ranked. It is also interesting to see that South Sudan started in 2016 and the Central African Republic started in 2017 in the plot, which shows that the countries weren't ranked in the top 183 the years before. This shows improvement in how happy the countries are. Other graphs were made, but I found the lowest-ranked to be the most interesting.


Lowest Happiness Ranked Countries Compare to Its Region's Average Happiness Rank Each Year From 2015 to 2019

I want to compare the male suicide rate and female suicide rate with the happiness score and males suicide rates with the happiness score for each country. This will help answer the question of how sexes are dealing with life in their country in comparison to the happiness score generated for a particular year. I merged the data on the country and grab all the suicide rates for each specific sex and its happiness score. Since this has three values to consider, happiness score, male suicide rate, and female suicide rate, I made a 3-D scatterplot with interactivity, so the viewer can rotate the graph to see all the data. It is interesting to see the two outliers, which are Eswatini and the most extreme outlier, Lesotho. Lesotho has a male suicide rate in 2019 of 146.9, a 34.6 female suicide rate in 2019, and a 3.802 happiness score in 2019. I would expect Lesotho to be the lowest-ranked country in 2019, but in 2019, the lowest-ranked country in South Sudan. It is interesting to find this out because Lesotho has such a high suicide rate, but possibly considering other factors that the World Happiness Report shows, such as Freedom, Health, etc., Lesotho is higher than South Sudan, giving it a better happiness score and rank.

Figure 1


Happiness Score 2019
male Suicide Rate 2019
Male Suicide Ra