R bootcamp - suicide

Eva-Maria Bonin

Last edited Sunday 12/03/2023

Contents

Introduction

Suicide is a significant public health issue globally, with an estimated 800,000 deaths annually.[1] Suicide is the second leading cause of death among individuals aged 15 to 29 years old. Mental health issues, including depression and substance abuse, are significant risk factors for suicide, and early intervention and treatment for these conditions can help prevent suicidal behavior.

Certain groups are more vulnerable to suicide, including individuals who have previously attempted suicide, individuals with a family history of suicide, and those who have experienced trauma or abuse. Men are also more likely to die by suicide than women, with suicide rates for men consistently higher in most countries.

There are also significant economic factors that contribute to suicide rates, such as poverty, unemployment, and income inequality. Studies have shown that economic recession and job loss can lead to an increase in suicide rates. Social isolation and lack of access to healthcare resources can also contribute to the risk of suicide.

Switzerland has a relatively high suicide rate, with approximately 1,200 suicides per year. Suicide rates in Switzerland are highest among men, individuals aged 45 to 54 years old, and individuals living in rural areas.

In this report, I explore data from the World Bank and the Global Burden of Disease Study to look at global trends in suicide rates, and compare the case of Switzerland to two countries with a higher and a lower suicide rate for males.

Data sources

I used data on suicide rates, population and economic indicators from the World Bank (WB), and data on mental health problems from the Global Burden of Disease study (GBD).

Pre-processing

Obtaining and merging data files

I downloaded selected indicators (suicide rate, economic indicators, population) from the WB data using the *wb_data* function, and obtained the GBD data by using the web tool to select and download the data.

Next, I created a column to capture the gender associated with an indicator value and removed the string indicating gender from the indicator descriptions. This allowed me to structure the data in a convenient way for grouping by gender later on.

```r
# Creating gender column in WB data
tmp_wb_long <- tmp_wb_long %>%
  mutate(sex = case_when(
    str_detect(indicator, "female") ~ "Female",
    str_detect(indicator, "male") ~ "Male"),
    sex = ifelse(is.na(sex), "Both", sex))

# Remove "male" and "female" from indicator descriptions
tmp_wb_long$indicator <- gsub("_male","",as.character(tmp_wb_long$indicator))
tmp_wb_long$indicator <- gsub("_female","",as.character(tmp_wb_long$indicator))
```

Similarly, the GBD data was processed and structured to match the WB data, ready for merging.

[1] World Health Organization. (2021). Suicide. https://www.who.int/news-room/fact-sheets/detail/suicide

The datasets were merged and a variable for the continent included.

During the initial cleaning, I generated a list of all unique country names in the combined dataset to check for duplicates caused by variance in spelling. These issues were resolved by manually modifying country names in the GBD datafile, to reflect the names used in the WB data.

I renamed the variables in the wide dataset to ensure consistency between data coming from the WB and GBD datasets.

Addressing missing data

Now the data were ready to address missing data, either by removing unusable rows or by performing imputation. This consisted of three different approaches:

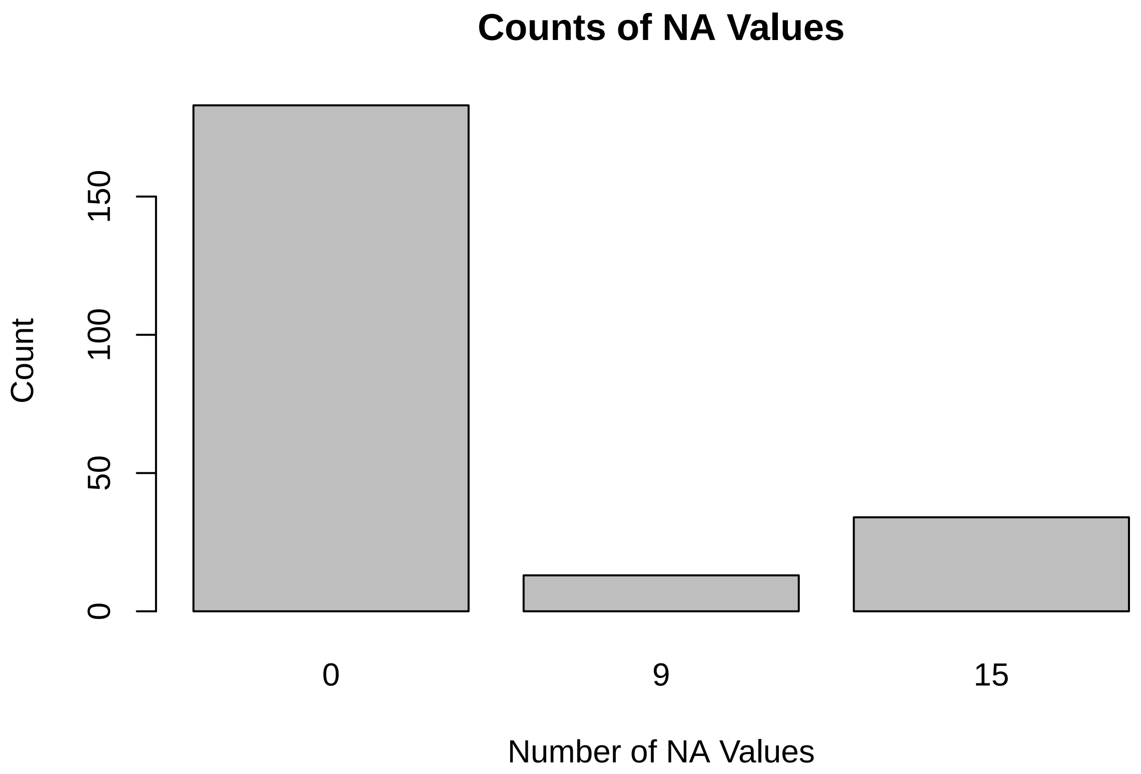
1. GDP per capita, years of compulsory education and total population were available overall, but not by sex. This overall value was used to fill missing values in "male" and "female".
2. Countries with no data at all on suicide rates were removed from the dataset.
3. For all over variables, missing data were replaced by manifestation of the *sex* variable. In a first step, the mean for the country and year was used to replace missing values. Where missing values remained, these were replaced by the mean for the country overall, then for the region, and where missing values still remained, they were replaced by the mean for the continent.

```r
# First step: Replace missing values by male and female with overall value

## Filter out values for sex == "Both"
tmp_wb_gbd_wide.both <- tmp_wb_gbd_wide %>%
  filter(sex == "Both") %>%
  select(!sex)

## Join the data, making sure that country and year are matched correctly.
tmp_wb_gbd_wide.tmp2 <-
  left_join(x = tmp_wb_gbd_wide, y = tmp_wb_gbd_wide.both,
            by = c("country", "year"),
            suffix = c("", ".y"))


# Replacing missing data in original variable with overall value.
tmp_wb_gbd_wide <- tmp_wb_gbd_wide.tmp2 %>%
  plyr::mutate(sui = coalesce(sui, sui.y),
               gdp_pc = coalesce(gdp_pc, gdp_pc.y),
               edu = coalesce(edu, edu.y),
               pop_t = coalesce(pop_t, pop_t.y))


# Remove columns with .y suffix
columns_to_remove <- grep("\\.y", names(tmp_wb_gbd_wide))
tmp_wb_gbd_wide <- tmp_wb_gbd_wide[,-columns_to_remove]
```

After this operation, we have data for 230 countries.

```r
# Second step: Drop countries with no data on suicide rates
# Counting missing values on suicide rates by country
```

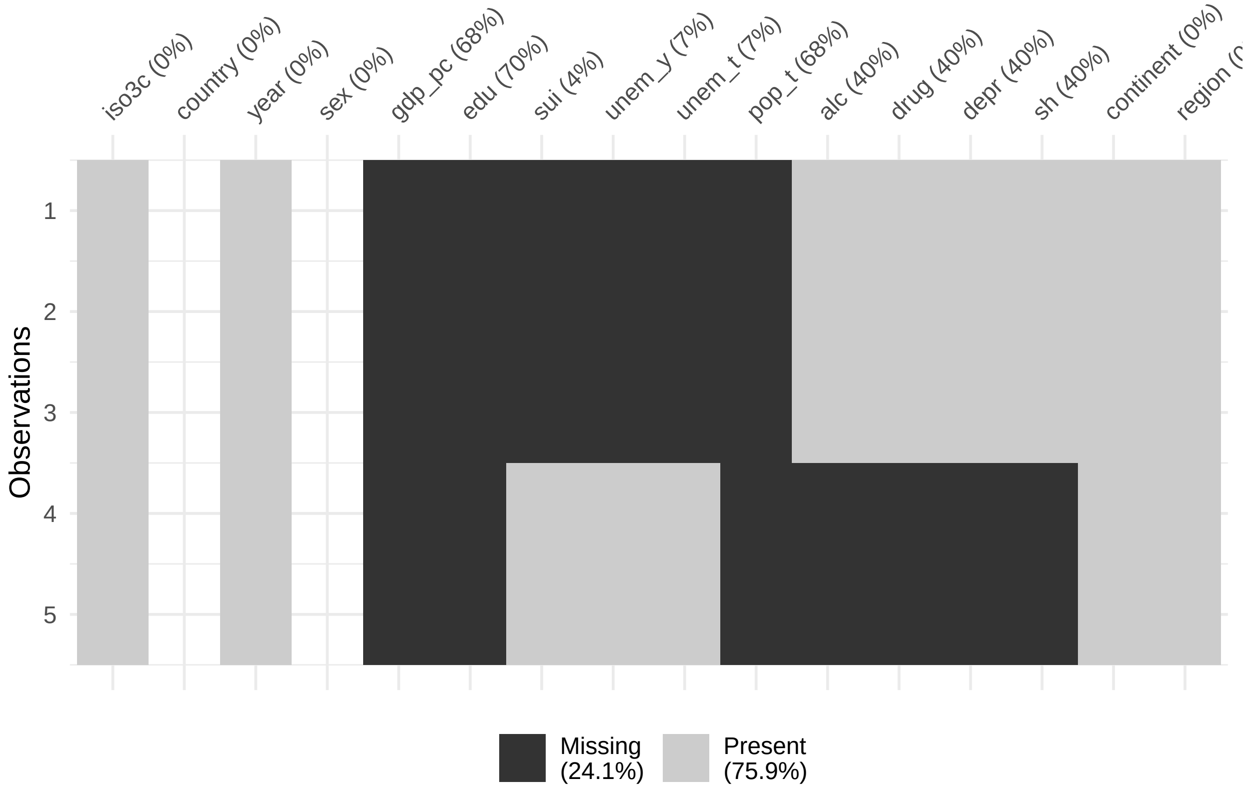
```
tmp_na_count <- tmp_wb_gbd_wide %>%
  group_by(country) %>%
  dplyr::summarize(count_na = sum(is.na(sui)))

# Count the occurrences of each value in count_na
tmp_counts <- table(tmp_na_count$count_na)

# create a bar plot of the counts
barplot(tmp_counts, xlab = "Number of NA Values", ylab = "Count", main = "Counts of NA Values")
```



After some investigation to confirm that these countries really had no data on the suicide rate, I generated a list of countries that were missing 9 or more instances of the suicide rate and dropped them from the dataset.

We now have 196 countries in the dataset.

Prior to performing the imputation in the third and final step, I visualised the distribution of missing values using the *vis_miss* library.



Most data are missing in the economic indicators and mental health data obtained from the GBD dataset.

Data were then imputed to address this.

Before the first imputation, there were 1.1015×10^4 missing values in the dataset. After the imputation, there were 5655 missing values in the dataset.

After the imputation, there were 1407 missing values in the dataset.

After the imputation, there were 120 missing values in the dataset.

After the final imputation, there were 0 missing values in the dataset.

Final datasets

The final dataset was saved in wide format (*suicide_final*) and in long format (*suicide_final_long*). Final preparations included

- Adding a variable of *date* type, based on the variable *year*, to facilitate the use of time series functions;
- Adding a calculated column showing total estimated deaths from suicide (*deaths*);
- Labelling variables;

Finally, I wrote the data to disk to preserve it.

Exploratory data analysis

Summarising the dataset

An overview of the variables in the dataset is provided using the *table1* function. Note that this is filtered for values that apply to both genders (sex == "Both"), but shows the average of data for all countries and years, by continent.

	Africa	Americas	Asia	Europe	Oceania
	(N=273)	(N=177)	(N=247)	(N=198)	(N=59)
Suicide mortality rate (per 100,000 population)					
Mean (SD)	8.93 (10.2)	7.61 (7.79)	7.01 (5.08)	13.6 (5.38)	13.0 (8.25)
Median [Min, Max]	6.60 [1.50, 78.3]	6.00 [0, 40.3]	5.60 [1.50, 28.6]	12.2 [0.600, 29.7]	12.5 [2.80, 29
Deaths by suicide					
Mean (SD)	1590 (2290)	2740 (8930)	8490 (28800)	2800 (6200)	371 (877)
Median [Min, Max]	943 [3.12, 14300]	385 [0, 53400]	1460 [11.8, 179000]	816 [25.7, 40700]	52.5 [3.99, 32
Total population					
Mean (SD)	24500000 (35100000)	28500000 (65100000)	94600000 (273000000)	18900000 (29200000)	3600000 (720
Median [Min, Max]	13100000 [95800, 213000000]	6440000 [91100, 332000000]	19000000 [430000, 1410000000]	7050000 [343000, 144000000]	290000 [10500
Compulsory education, duration (years)					
Mean (SD)	8.35 (1.74)	11.5 (3.00)	9.47 (1.95)	10.6 (1.39)	10.2 (1.80)
Median [Min, Max]	8.00 [5.00, 12.0]	12.0 [6.00, 17.0]	9.00 [5.00, 15.0]	10.0 [8.00, 13.0]	10.0 [8.00, 15
GDP per capita (constant 2015 US$)					
Mean (SD)	2430 (2940)	11400 (11400)	12600 (14900)	29700 (25100)	11200 (17800
Median [Min, Max]	1340 [261, 17000]	8410 [1280, 61900]	5210 [426, 66200]	19900 [2250, 108000]	4420 [1390, 5
Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate)					
Mean (SD)	17.9 (15.7)	19.6 (9.85)	14.7 (8.94)	17.8 (9.50)	10.5 (4.98)
Median [Min, Max]	11.0 [0.735, 78.8]	20.1 [3.25, 44.7]	13.1 [0.380, 42.3]	14.9 [5.62, 47.3]	11.8 [1.36, 20
Unemployment, total (% of total labor force) (modeled ILO estimate)					
Mean (SD)	9.11 (7.14)	8.32 (4.28)	5.98 (4.16)	7.38 (4.39)	4.16 (2.18)
Median [Min, Max]	5.81 [0.542, 28.8]	7.95 [1.20, 20.6]	4.67 [0.100, 19.2]	5.75 [2.01, 22.8]	4.29 [0.698, 9
Alcohol abuse					
Mean (SD)	450 (266)	959 (296)	626 (469)	1200 (253)	489 (346)
Median [Min, Max]	421 [159, 1190]	871 [431, 1720]	439 [180, 2270]	1190 [710, 1860]	362 [291, 147
Drug abuse					
Mean (SD)	139 (42.9)	230 (78.2)	197 (41.2)	268 (66.0)	249 (98.9)
Median [Min, Max]	122 [86.1, 271]	218 [150, 583]	189 [126, 324]	264 [162, 420]	206 [177, 523
Depression					
Mean (SD)	4080 (1010)	4080 (902)	3690 (1330)	4400 (1130)	2820 (980)
Median [Min, Max]	3870 [2240, 6620]	3960 [2100, 6390]	3500 [1360, 6920]	4290 [2070, 7470]	2440 [1950, 5
Self harm					
Mean (SD)	33.0 (17.2)	37.0 (38.5)	53.3 (33.1)	87.6 (33.8)	77.6 (26.6)
Median [Min, Max]	26.2 [15.5, 99.7]	20.0 [8.87, 168]	40.6 [18.3, 177]	85.4 [33.1, 198]	84.6 [21.9, 11

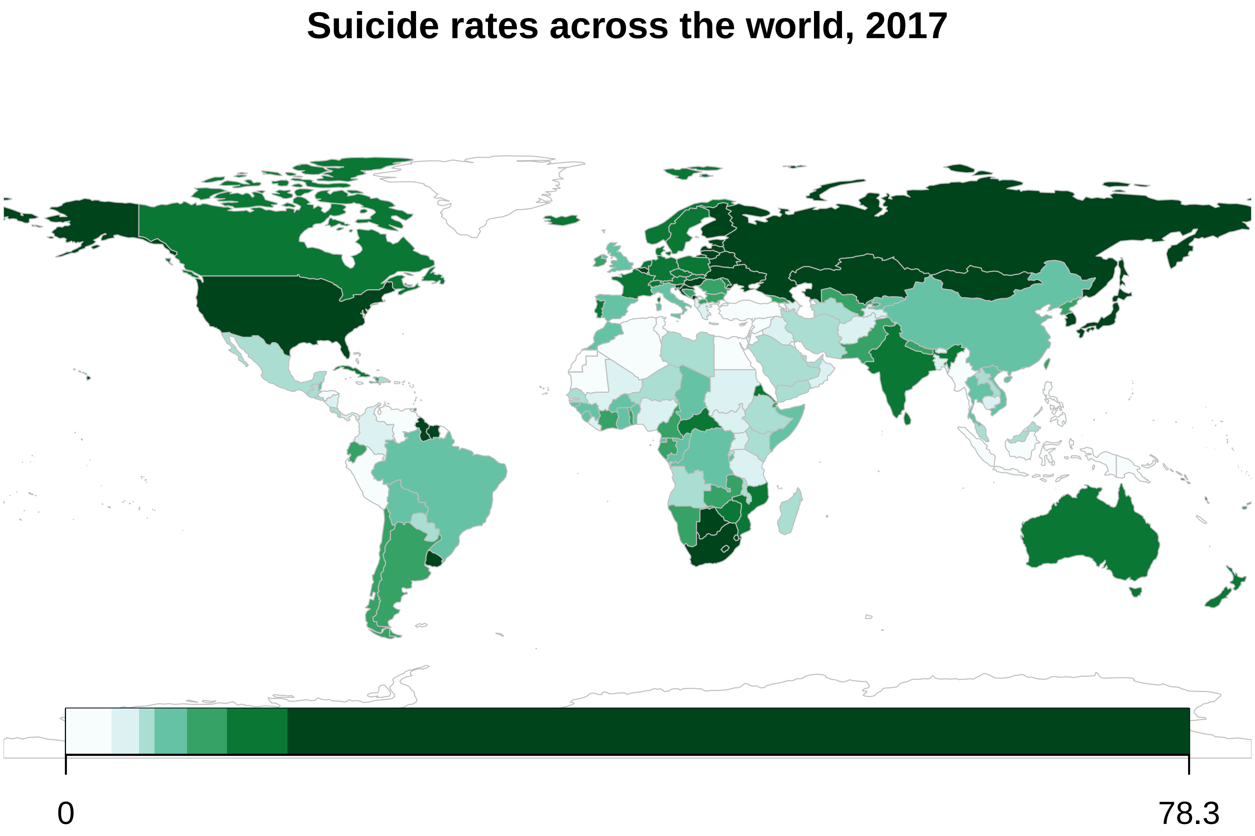
Mapping suicide rates

World maps showing suicide rates by country for 2017 and 2021 were created. A darker colour signifies a higher suicide rate.

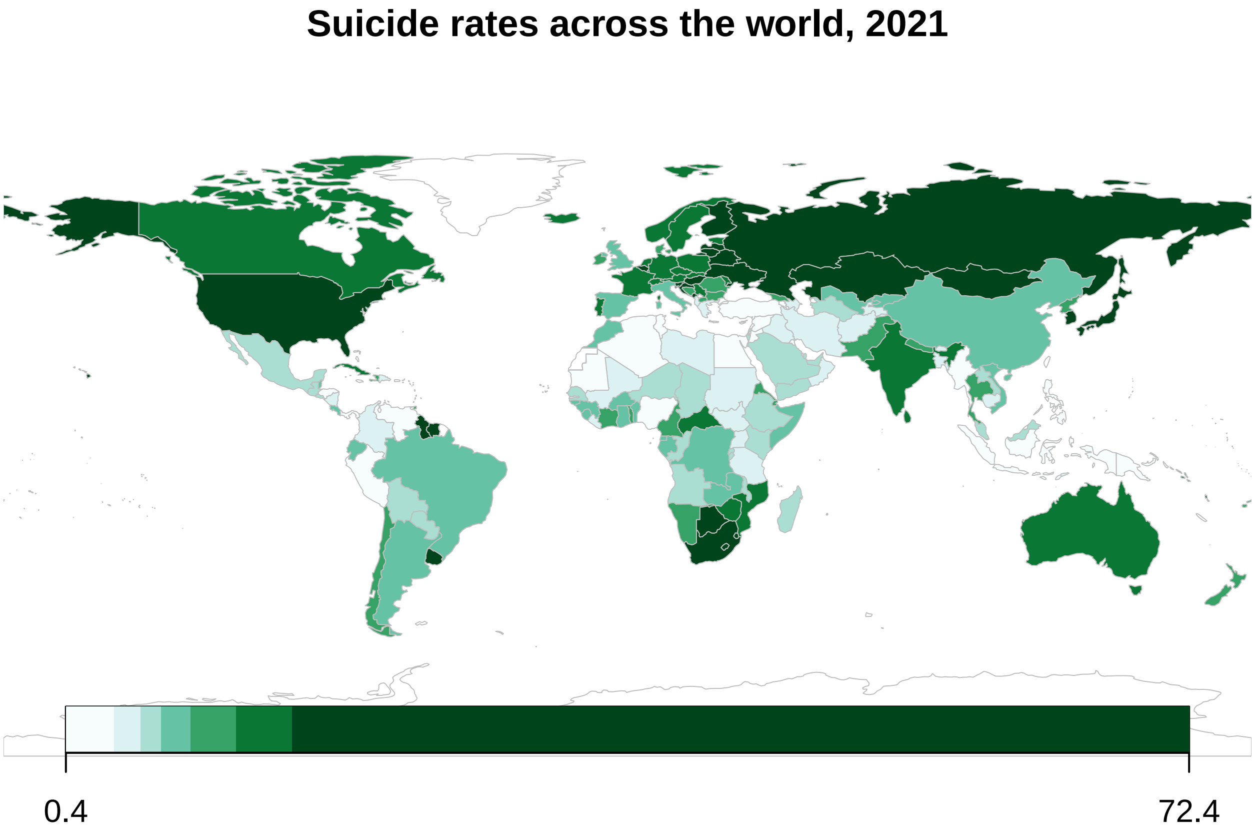
```
df1 <- suicide_final %>%
 filter(year == 2017) %>%
  drop_na()
data(df1)
colpal <-  c('#f7fcfd','#e5f5f9','#ccece6','#99d8c9','#66c2a4','#41ae76','#238b45','#006d2c','#00441b')
sPDF <- joinCountryData2Map(df1, joinCode = "ISO3", nameJoinColumn = "iso3c")
```

```
## 585 codes from your data successfully matched countries in the map
## 3 codes from your data failed to match with a country code in the map
## 54 codes from the map weren't represented in your data
```

```
par(mai=c(0,0,0.2,0),xaxs="i",yaxs="i")
mapCountryData(sPDF, nameColumnToPlot="sui", mapTitle = "Suicide rates across the world, 2017", colourP
```



```
## 549 codes from your data successfully matched countries in the map
## 0 codes from your data failed to match with a country code in the map
## 60 codes from the map weren't represented in your data
```



To show the top 20 countries with the highest average suicide rates at any point, I calculated the average by country and sex and sort by the mean:

```
## # A tibble: 588 x 3
##    country               sex        mean_sui
##    <labelled>            <labelled>    <dbl>
##  1 Lesotho               Male          120.
##  2 Lesotho               Both           74.4
##  3 Guyana                Male           62.1
##  4 Eswatini              Male           55.3
##  5 Kiribati              Male           49.0
##  6 Lithuania             Male           46.9
##  7 Russian Federation    Male           45.5
##  8 Micronesia, Fed. Sts. Male           43.1
##  9 Guyana                Both           39.8
## 10 Korea, Rep.           Male           39.7
## 11 Ukraine               Male           38.9
## 12 South Africa          Male           38.5
## 13 Suriname              Male           38.5
## 14 Belarus               Male           38.1
## 15 Latvia                Male           35.5
## 16 Uruguay               Male           34.4
## 17 Montenegro            Male           31.8
## 18 Mongolia              Male           31.4
## 19 Slovenia              Male           31.1
## 20 Kazakhstan            Male           30.7
## # ... with 568 more rows
```

We can see that most of the 20 highest values are for males, or are overall rates driven by very high rates for males.

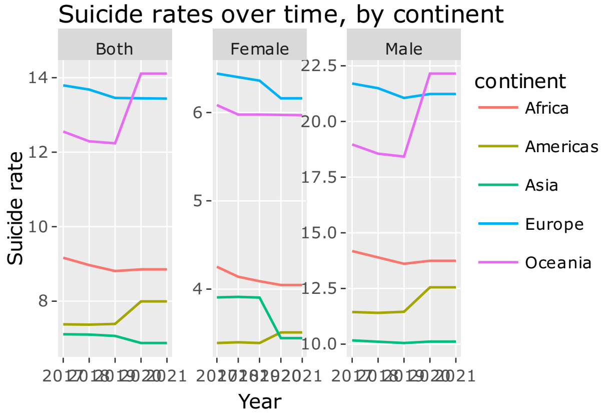
The highest value is for the country of Lesotho, the lowest for the country of Barbados. For comparison, the value for males in Switzerland is . These three countries are chosen for the case study comparisons.

Exlploring data by continent

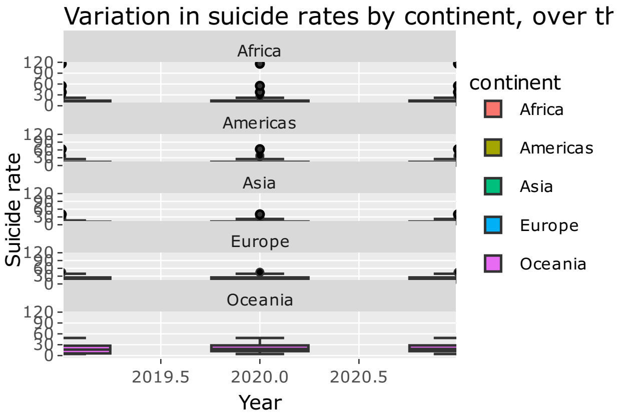
Due to the many countries in the data set, visual analysis is challenging. An overview is provided by showing the average data by continent.

- A line graph shows trends in suicide rates over time.
- Box plots for the last 3 years show the distribution of the variable, to give an idea of the variability within continents.
- A cumulative graph shows total deaths by suicide.

I used the *plotly* package with *ggplot2* to make the graphs more interactive.



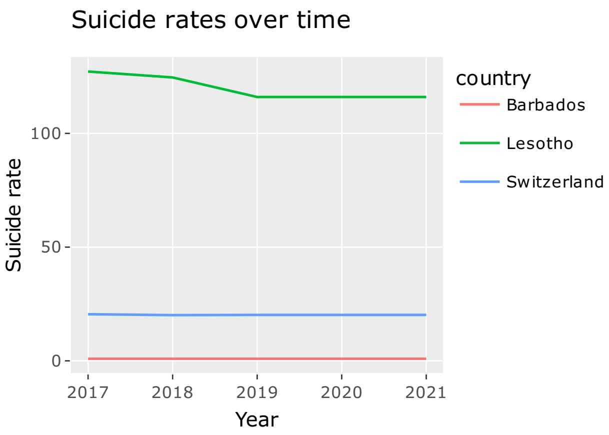
Suicide rates between men and women diverge for every continent. Rates are highest in Europe and Oceania, and lowest in Asia.



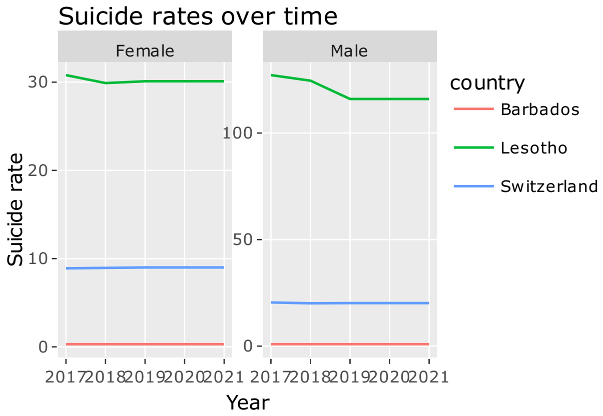
There is a lot of variation in suicide rates within continents. Visual inspection of the plots suggests that there are more outliers on those continents that contain more countries. Most distributions do not indicate normality.

Case study Switzerland

In a case study, I compare Switzerland to the countries with the highest and lowest male suicide rates.



As expected, rates in Lesotho are much higher than in the other countries, and Barbados has rates close to zero.



While trends for females are steady in all three countries (although at different levels), trends for males appear to be downward. However, it is unclear without further investigation if this could be an artefact of the imputation procedure.

Regression models

Given the findings above, does the suicide rate differ over time for males or females? I ran and compared 6 regression models to find out.

1. Regression of suicide rade on *sex*, *year*.
2. Adding fixed effect variable *country* to see if effects differ by country.
3. Including economic data such as population, education, GDP and unemployment.
4. Including mental health data.
5. Adding a fixed effect for the year to account for time trends in the data.
6. Removing variables with coefficients that are zero or very close.

I used the *feols* function for all models because it has easy options for fixed effects and clustering, and is compatible with the *modelsummary* and *modelplot* functions I used to visualise the results.

```
# Simplest model
reg1 <- suicide_final %>% filter(sex != "Both") %>% feols(sui ~ sex + year)

# Fixed effect by country
reg2 <- suicide_final %>% filter(sex != "Both") %>% feols(sui ~ sex + year | country)
```

```
# Including economic data
reg3 <- suicide_final %>% filter(sex != "Both") %>% feols(sui ~ sex + year + pop_t +
                                                            edu + gdp_pc + unem_y + unem_t | country)
# Including mental health data
reg4 <- suicide_final %>% filter(sex != "Both") %>% feols(sui ~ sex + year + pop_t
                                                          + edu + gdp_pc + unem_y + unem_t+
                                                            alc + drug + depr + sh | country)
# 4 plus fixed effect for year
reg5 <- suicide_final %>% filter(sex != "Both") %>% feols(sui ~ sex + pop_t
                                                          + edu + gdp_pc + unem_y + unem_t+
                                                            alc + drug + depr + sh | country + year,
                                                          cluster = ~ country)
# 5 minus vars with zero coefficient
reg6 <- suicide_final %>% filter(sex != "Both") %>% feols(sui ~ sex +
                                                          edu + unem_y + unem_t+
                                                            alc + drug + depr + sh | country + year,
                                                          cluster = ~ country)
```

Special chapter: modelsummary

To provide an overview of the models and display the most salient results in a way that is both attractive and easy to read, I used the library *modelsummary* to combine results from the six models. I then used *modelplot*

Overview of models

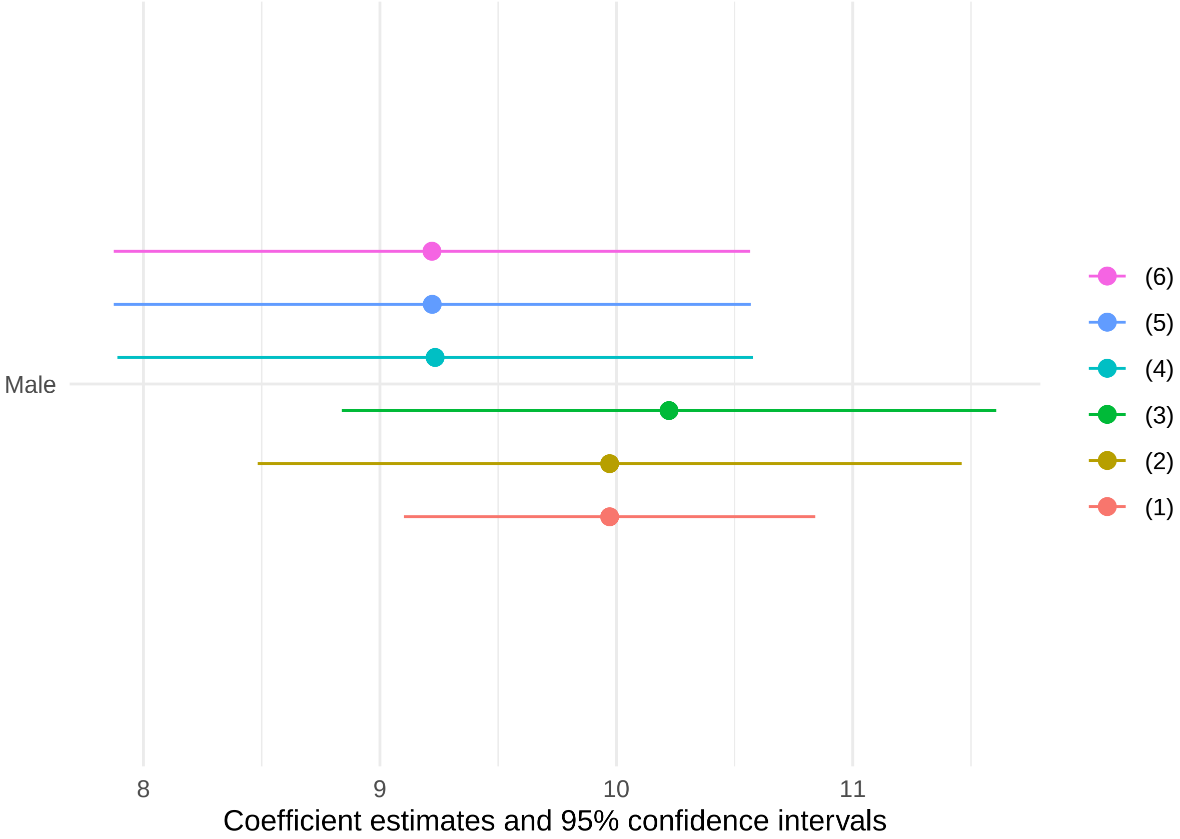
An overview of the coefficients

There appears to be a negative (but not statistically significant) coefficient on *year*, pointing towards a downward trend in suicide rates. Unemployment, alcohol misuse and self harm are significantly related to suicide rates. Rates for males are higher than for females in all models.

	(1)	(2)	(3)	(4)	(5)	(6)
Male	9.972***	9.972***	10.223***	9.233***	9.221***	9.220***
	(0.444)	(0.755)	(0.702)	(0.681)	(0.683)	(0.682)
Year	0.004	−0.056**	−0.066	−0.045		
	(0.157)	(0.021)	(0.042)	(0.039)		
Total population			0.000	0.000	0.000	
			(0.000)	(0.000)	(0.000)	
Years compulsory education			0.115	0.044	0.038	0.031
			(0.100)	(0.103)	(0.109)	(0.113)
GDP per capita			0.000	0.000	0.000	
			(0.000)	(0.000)	(0.000)	
Youth unemployment			−0.306	−0.255	−0.258	−0.258
			(0.212)	(0.192)	(0.195)	(0.195)
Total unemployment			0.731*	0.500*	0.499*	0.498*
			(0.293)	(0.248)	(0.248)	(0.248)
Alcohol misuse				0.002+	0.002+	0.002+
				(0.001)	(0.001)	(0.001)
Drug misuse				0.007	0.007	0.007
				(0.013)	(0.013)	(0.013)
Self harm				0.201***	0.201***	0.201***
				(0.034)	(0.034)	(0.034)
Num.Obs.	1908	1908	1908	1908	1908	1908
R2	0.210	0.765	0.772	0.815	0.815	0.815
Std.Errors	IID	by: country	by: country	by: country	by: country	by: country
FE: country		X	X	X	X	X
FE: year					X	X

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Model plots



The plot shows the coefficient on *Male* and the standard errors. Model 1, surprisingly, shows the smallest standard errors.

Conclusions

Suicide continues to be a public health concern across the world. There is a lot of variation in rates between countries, and between and within continents. Men continue to be at higher risk of suicide than women. There may be a downward trend overall, but the fact that much of these data were imputed means that interpretation has to be very cautious.