

## ***Analysis of Suicide Statistics***

*I provide an analysis of suicide data available from Kaggle (the data originates with the WHO). I focus on cross-country differences in the ratio of suicides by men vs. women. It turns out that this statistic is strongly correlated with the gap in life expectancy, but only weakly related to factors such as the incidence of depression.*

### **Data**

I will work with a data set that originates with the WHO. It records information on suicides for many countries in the world over a time horizon of 30+ years. This dataset is available from Kaggle:

<https://www.kaggle.com/szamil/who-suicide-statistics><sup>i</sup>

Next to the suicide statistics, I will also use information about life expectancy, mental health, and Gross National Income per capita that are available from the OurWorldInData website at the following pages:

<https://ourworldindata.org/life-expectancy><sup>ii</sup>

<https://ourworldindata.org/mental-health><sup>iii</sup>

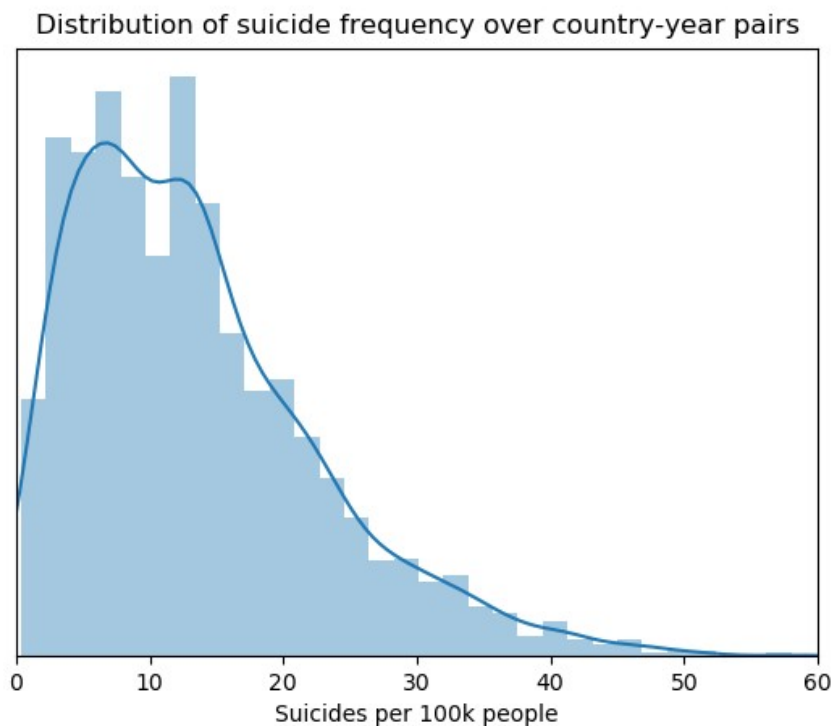
<https://ourworldindata.org/human-development-index><sup>iv</sup>

### **Pre-processing and first inspection of data**

After loading and preprocessing data, the first step is to create a dataframe that groups information about suicides by country and per year. In the resulting dataframe, each row corresponds to a country-year pair, and contains information about both men and women. I add a column that provides a measure for the male preponderance in suicides: It is the ratio of suicides per 100,000 men over suicides per 100,000 women. (This measure is so constructed so that a possible imbalance in the share of the population which is male / female does not distort the results on suicides.) Hence, if the male preponderance measure equals four, a man is four times as likely to commit suicide as a woman is.

## *Analysis of Suicide Statistics*

To gain a first feel for the data, one may wonder how many suicides per 100,000 people should be expected in general. Let us therefore visualize the distribution of suicides per 100,000 people over the country-year pairs in the newly created dataframe.



While the histogram ignores time trends, cross-country differences, or male preponderance in suicides, it can give us a good idea of the magnitude and distribution of the numbers we are dealing with. We see that the distribution skews right: We will mostly be dealing with about 5-20 suicides per 100,000 people, while in less frequent cases, much higher numbers also occur. To be more specific, the distribution of suicides across the country-year pairs can be described in the following summary statistics:

```
print(grouped_data["Suicides per 100k"].describe())
```

```
count    1983.000000
mean      13.528408
```

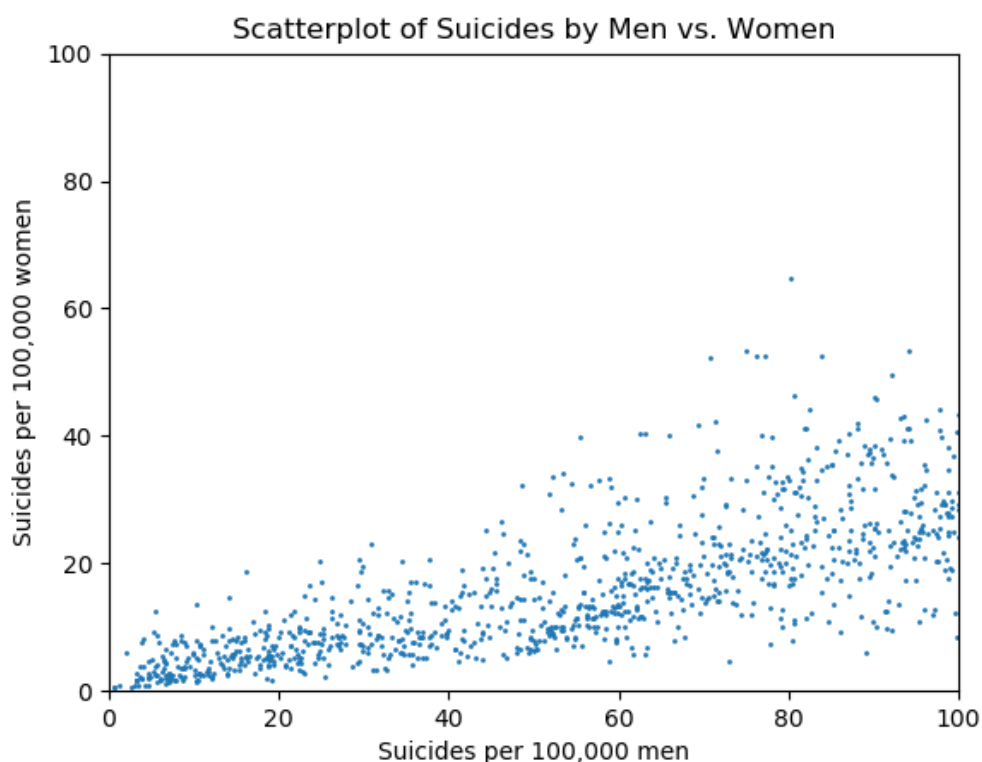
## *Analysis of Suicide Statistics*

```
std      9.244799
min       0.367630
25%       6.541924
50%      12.006743
75%      18.545301
max      58.013151
Name: Suicides per 100k, dtype: float64
```

The subject of interest that I have chosen for my analysis is the male preponderance in suicides:

*More men than women commit suicide.*

To take a first glance at what the data says about this issue, I use again the same set of country-year pairs and plot suicides per 100,000 men and suicides per 100,000 women against each other in a scatter plot. Each dot corresponds to a country-year pair from the dataset.

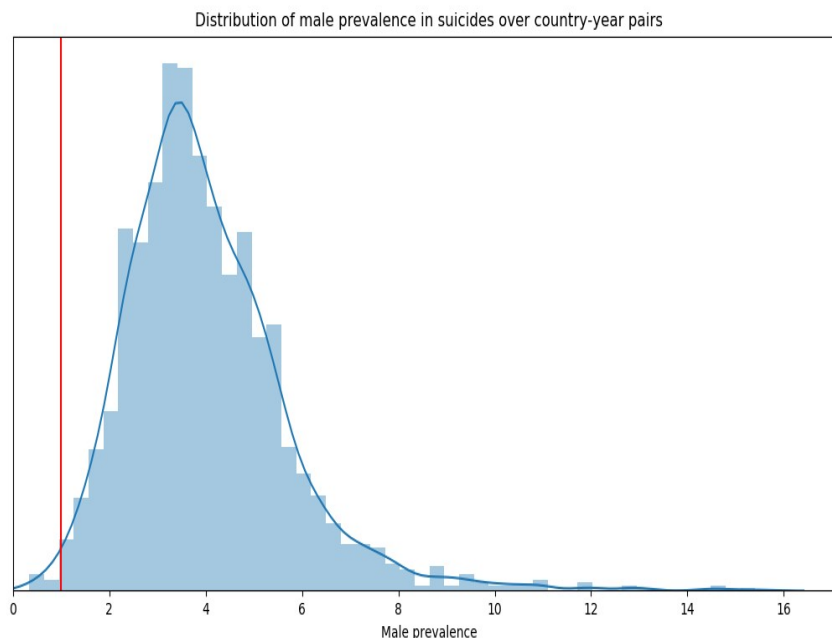


## ***Analysis of Suicide Statistics***

The scatterplot allows for two basic observations: First, the strong positive correlation indicates that deep underlying causes of suicides are not too specific to men or women. Second, since the scales of the two axes are equal, the concentration of dots in the lower parts of the figure indicates the male preponderance that we are interested in:

*Whatever the overall number of suicides is in a given country and year, men tend to be much more likely to commit suicide than women are.*

Let us now get a first impression of the numbers for male preponderance. First, consider a histogram of the distribution of male preponderance in suicides across all the country-year pairs that remain in the 2013 dataset. In the histogram, the red vertical line indicates a preponderance of one: That is, in the region to the right of the red line, men are more likely to commit suicide than women. (The histogram is normed so that the blue curve is essentially the distribution's density.)



## ***Analysis of Suicide Statistics***

Some basic statistics about the distribution are the following:

```
print(grouped_data["Male_Preponderance"].describe())
```

```
count    1983.000000
mean      4.012583
std       1.690178
min       0.338409
25%       2.909900
50%       3.745098
75%       4.854813
max       15.409283
```

```
Name: Male_Preponderance, dtype: float64
```

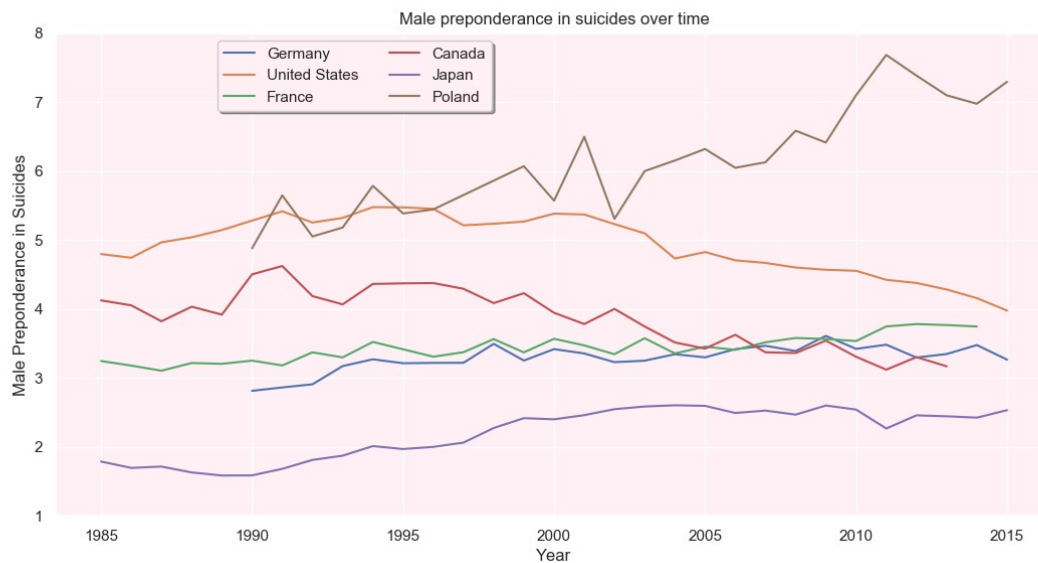
What it tells us is that male preponderance in suicides is a phenomenon that consistently occurs across time and across countries, although to varying degrees. On average, suicides by men outnumber those by women by a factor of about four.

As a point of caution, be aware that just looking at the country-year pairs is still a rather superficial approach: For instance, it treats populations of unequal size equally and thus any statement about “averages” is distorted by over-weighting smaller countries.

### **Comparisons across time and countries**

As a next step, I will dig deeper and try to understand what drives male preponderance in suicides. As a starting point, one may wonder whether there is a clear time trend in male preponderance. Since the data goes back to 1985, it captures the developments in more than a generation. First, create a plot to display time trends for various countries.

## *Analysis of Suicide Statistics*

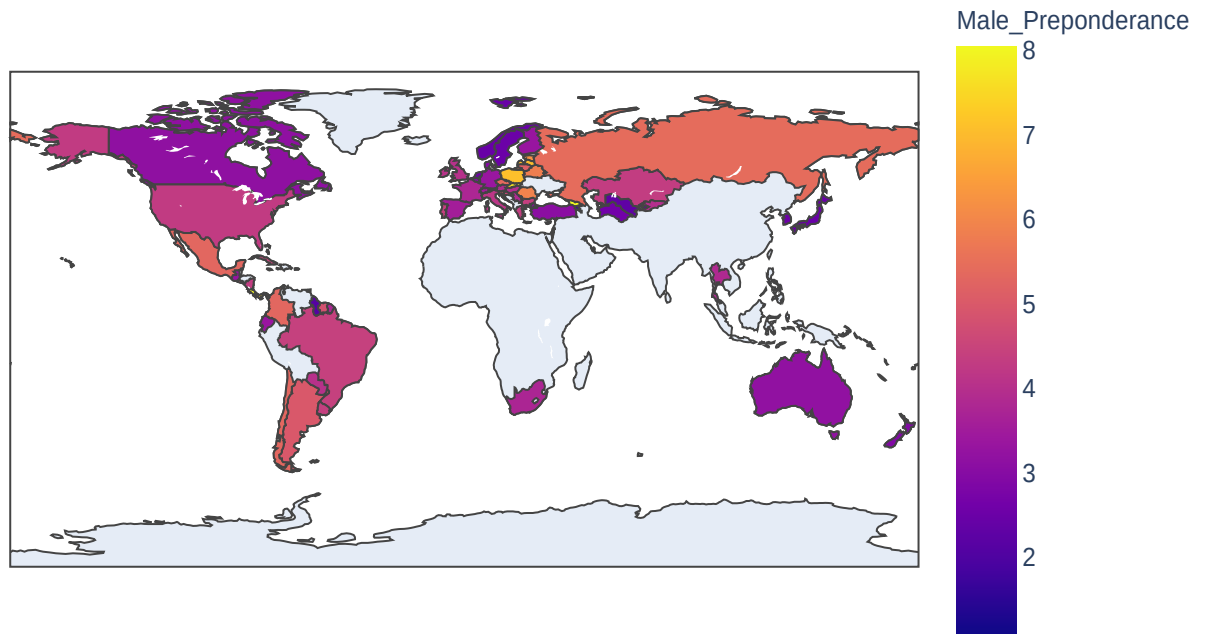


Not only the magnitude of male preponderance in suicides but also the changes over time seem to differ a lot across different countries, as exemplified by the six countries shown in the figure. Hence, it does not seem very promising to look for world-wide time trends in the variable of interest – instead, the cross-country differences could be more interesting. Therefore, I am now going to focus on a cross-country analysis for one particular year: I have chosen the year 2013 as a compromise between looking at recent data vs. complete data: For years following 2013, information for quite a few countries is missing from the dataset.

Creating a reduced dataset for the year 2013 only requires regrouping the earlier dataframes first by year and then by country. Then I use the plotly-library to create a map of the world that displays male preponderance in suicides across countries. This yields the following choropleth map:

## *Analysis of Suicide Statistics*

### Male preponderance in suicides throughout the world (2013)



While male preponderance in suicides occurs across all the countries considered (unfortunately, there is no information about most African and Asian countries in the dataset), its magnitude differs a lot across countries.

Studying the choropleth map gives some potential clues as to what might drive the cross-country differences:

1. One pattern that can be seen in the map is that male preponderance in suicides is **particularly pronounced in the Central and Eastern European countries**. Many media reports suggest that some of these countries have problems with a particularly low male life expectancy. It is known that men generally have lower life expectancy than women, but the size of this difference varies between countries. I will look at data on relative life expectancy

## ***Analysis of Suicide Statistics***

of men and women and see how it relates to the male preponderance in suicides.

2. Another well-known fact is that suicide is often the consequence of depression and that women are more often diagnosed with depression than men. I will also include in the analysis information about the *Female Preponderance in Depression* in various countries.
3. Another pattern that is clear from the world map as shown above is that male preponderance in suicide seems to be **less pronounced in some countries that are known to have a particularly high level of human development**, such as the Scandinavian countries or Japan. I include data about the Gross National Income (per capita) to see how it relates to male prevalence in suicides. (I do not take the Human Development Index because life expectancy, like Gross National Income, is one of the variables that are used to produce the HDI in the first place.)

### **Relating suicide data to data on life expectancy, mental health, and GNI**

First, turn to the data on life expectancy: For a first visual impression of the correlation between relative life experience of men vs. women and male preponderance in suicides, I create a scatter plot.



## *Analysis of Suicide Statistics*



The plot suggests a clear negative correlation:

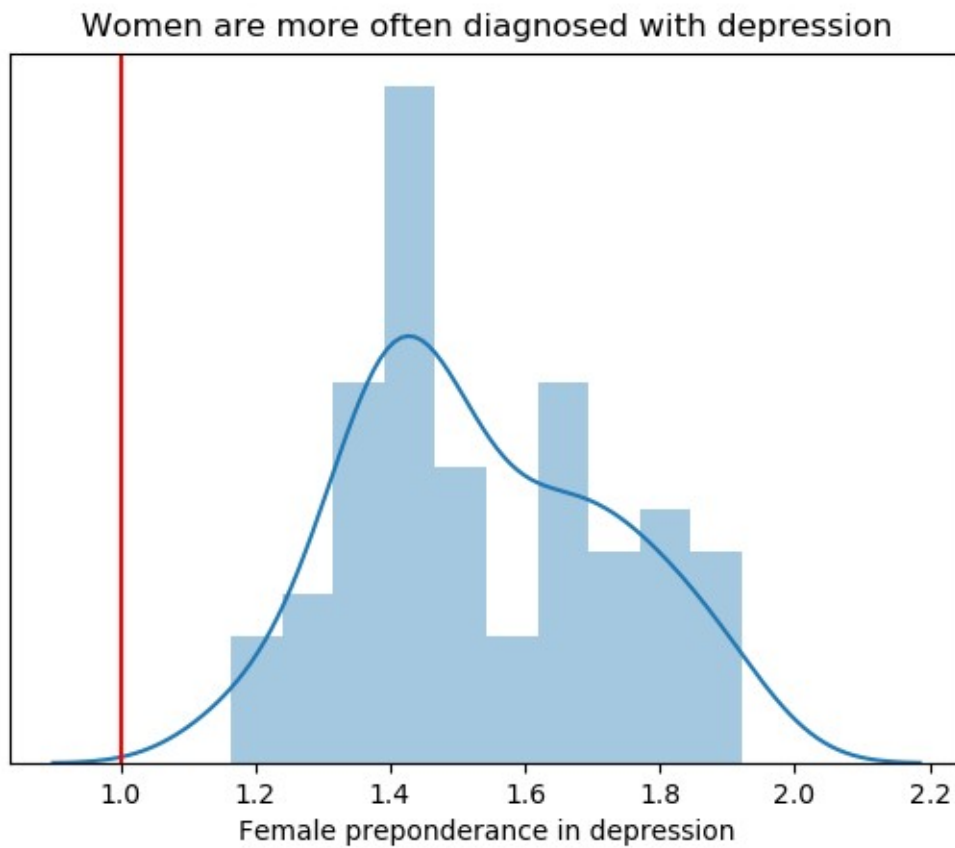
*Countries in which male life expectancy is low relative to female life expectancy, tend to have a high male preponderance in suicides.*

Since only a small share of the total population commits suicide, we can rule out the possibility that male preponderance in suicides causes lower relative life expectancy of men. Instead, it seems that there is a common driver behind the lower life expectancy of men and the male preponderance in suicides. One possible candidate would be a greater propensity towards all kind of self-destructive behaviors. That is, one interpretation consistent with the correlation could be this:

*There is an entire range of self-destructive behaviors, starting with things like smoking or drug use, and ending with an extreme action like suicide. Men are more likely than women to engage in these self-destructive behaviors. Therefore, their life expectancy is lower, and they are more prone to suicide. The degree to which men are more inclined to self-destructive behavior, however, differs greatly across countries.*

## ***Analysis of Suicide Statistics***

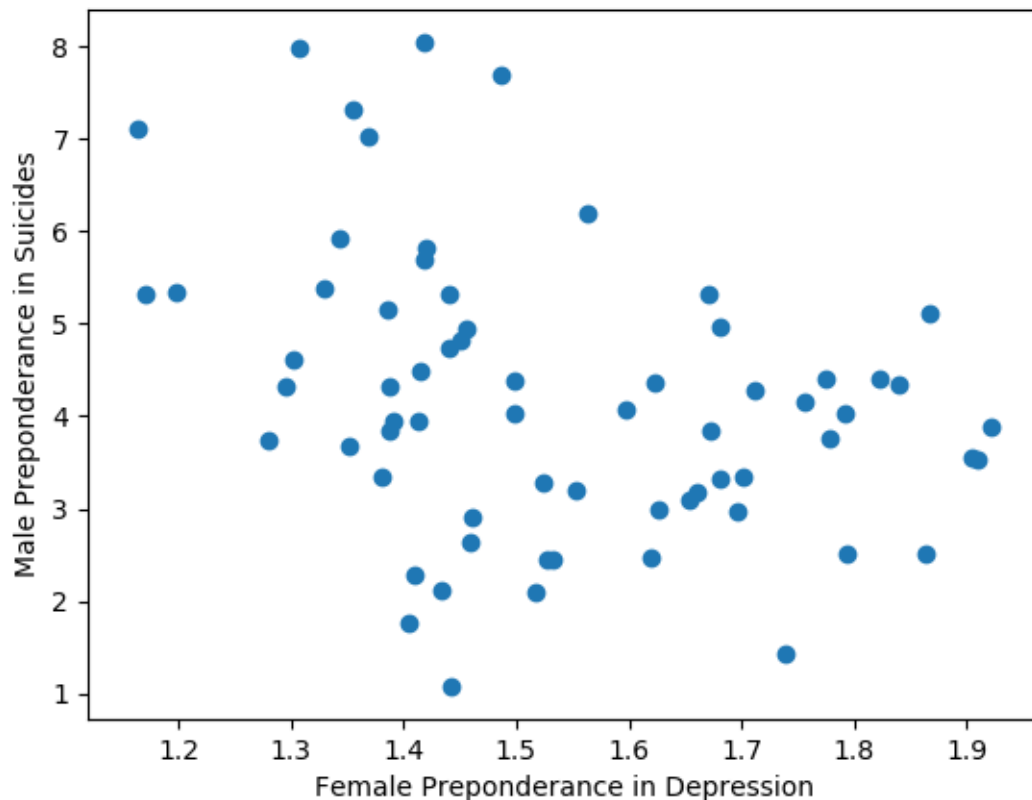
Other factors could be taken into account as well: In particular, one puzzle is that suicide is often said to be related to mental health problems such as depression, but it is also widely known that women are diagnosed with depression more often than men. This latter fact can be seen in the following chart, which depicts Female Preponderance in Depression across different countries. In the histogram, the vertical red line indicates equal incidence of depression among men and women.



## *Analysis of Suicide Statistics*

Next, I create a scatterplot that displays the relation between female preponderance in depression and male preponderance in suicides.

Female preponderance in depression vs. male preponderance in suicides

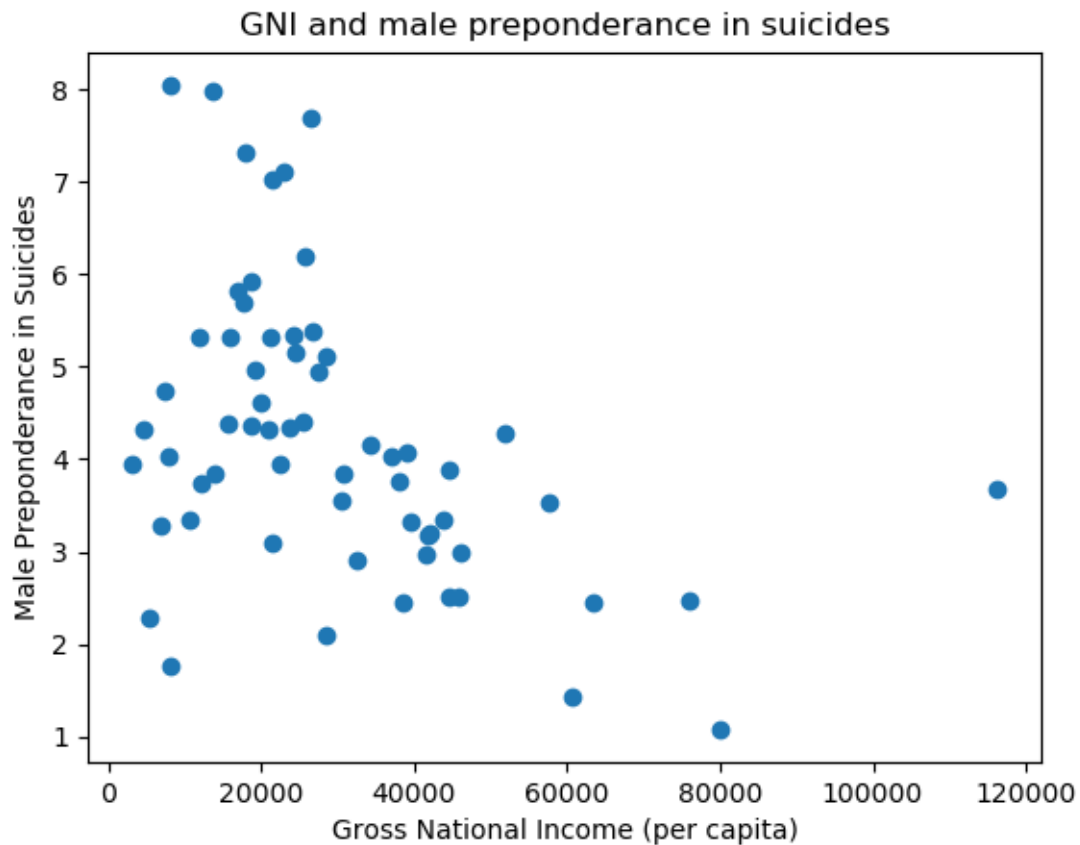


Inspecting the scatter plot suggests a negative correlation: Countries in which depression occurs more evenly in men and women tend to be countries in which the male prevalence in suicides is also lower. One possible interpretation could be the following:

*Across countries, there are differences in how reluctant men are to seek treatment for an existing depression. In those countries where they are more likely to seek treatment, they show up in greater numbers in the statistics, which brings down female preponderance in (diagnosed) depressions. Treatment is effective, which then also reduces male preponderance in suicides.*

## ***Analysis of Suicide Statistics***

Finally, consider the potential influence of gross national income (per capita) on cross-country differences.



It could be observed from the visualization on the world map that some highly developed countries (Scandinavia, Japan) have low male preponderance in suicides. The expectation was that there is generally a negative correlation between the GNI and the male prevalence in suicides. Indeed, the scatter plot displays such a negative correlation. One possible interpretation is along the following lines:

## Analysis of Suicide Statistics

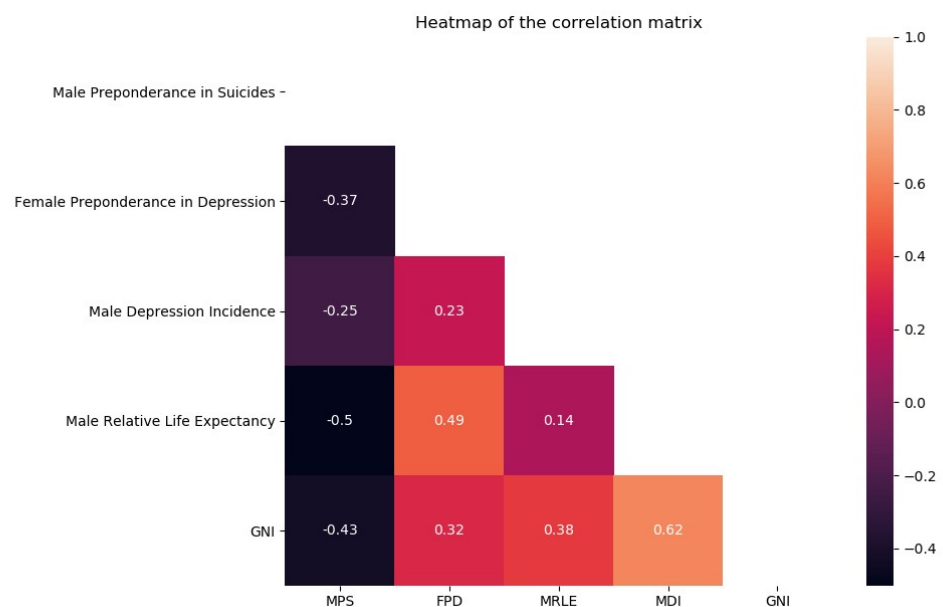
*In men, the propensity to commit suicide is more often driven by economic problems than it is in women.*

However, related to our previous point, it could also be true that countries with higher GNI are also the ones where it is easier (particularly for men) to seek treatment for depression. The factors that we have considered could also be correlated in various ways with each other: We will look at these relationships in the next section.

Next to the original data set on suicides, I have now used information on life expectancy, depression, and gross national income across countries. The data confirms some basic intuition gained from inspecting the map of the world created earlier.

### Correlations among all variables considered

So far, I have always focused on correlation between Male Preponderance in Suicides and one other variable. Let us now take a look at how all the different variables correlate with each other. To this end, I have created a correlation matrix and visualized it in a heatmap:



## ***Analysis of Suicide Statistics***

Male Preponderance in Suicides is correlated most strongly with Male Relative Life Expectancy. The latter also correlates strongly with Gross National Income. The correlation with the variables related to depression is weaker. It seems that the greater propensity of men to commit suicide and their shorter life expectancy are driven by common factors, and these do not seem to be primarily related to depression.

The first insight gained from the coefficients in the heatmap is also corroborated by looking at the coefficients and significance levels found in linear regression: Consider Female Prevalence in Depression and Male Relative Life Expectancy as explanatory variables (adding a constant), then we find a coefficient of -1.3 ( $p=0.19$ ) for Female Preponderance in Depression and -25 ( $p=0.002$ ) for Male Relative Life Expectancy.

### **Takeaway**

It is somewhat surprising that differences in the incidence of depression seems to play such a weak role since it is often cited as the cause or at least a contributing factor in most suicides. Of course, it could be that numbers are distorted by a lower male propensity to seek therapy for depression. However this might be, the main takeaway from the analysis is the following:

*In order to understand male preponderance in suicides, it may be necessary to look beyond the issue of suicide alone and ask instead under what conditions men have high propensity to commit an entire range of self-destructive behaviors, of which suicide is merely the most extreme case.*

## Original Data Sources:

- i The Kaggle dataset I refer to here was prepared by Kaggle user Szamil using data from the World Health Organization's Mortality Database, see <https://apps.who.int/healthinfo/statistics/mortality/whodpms/>.
- ii The website OurWorldInData.org lists as the original data source:  
James C. Riley (2005) – Estimates of Regional and Global Life Expectancy, 1800–2001. Issue Population and Development Review. Population and Development Review. Volume 31, Issue 3, pages 537–543, September 2005., Zijedeman, Richard; Ribeira da Silva, Filipa, 2015, "Life Expectancy at Birth (Total)", <http://hdl.handle.net/10622/LKYT53>, IISH Dataverse, V1, and UN Population Division (2019)
- iii The website OurWorldInData.org lists as the original data source:  
Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2017 (GBD 2017) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2018.
- iv The website OurWorldInData.org lists as the original data source:  
United Nations Development Programme (UNDP)