Suicide Notebook

David Herel

14.12.2020

Introduction

In this notebook I want to make a closer look on Suicide dataset and analyze it.

Part I - Exploratory analysis of the suicide dataset

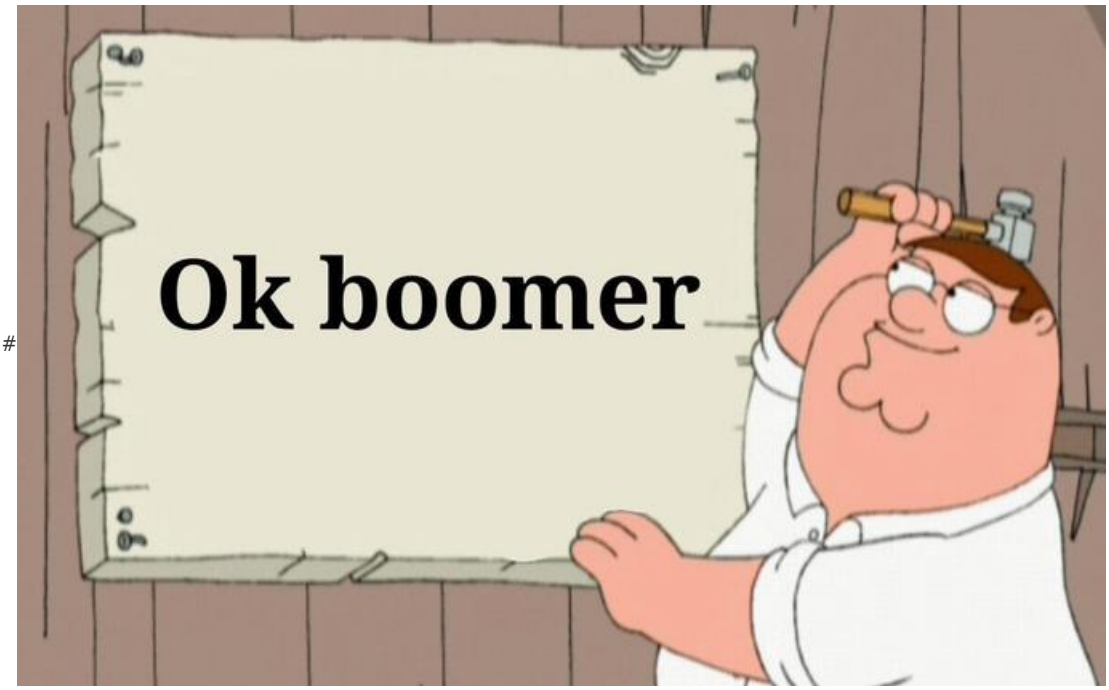
Load the dataset, visualize the main relationships and trends, preprocess the data, carry out dimensionality reduction and clustering. The main goal is to see whether there are regularities in suicide rates across countries, continents, periods of times, etc.

Analyze variables

So what about to start with analyzing dataset? At **first**, we look on interesting variables and their behavior.

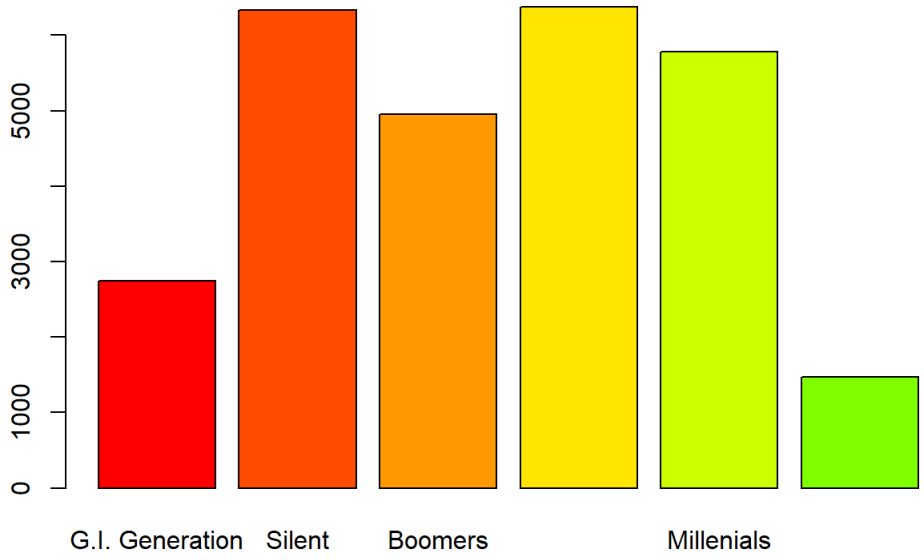
Let's start with most interesting one:

#



Generation

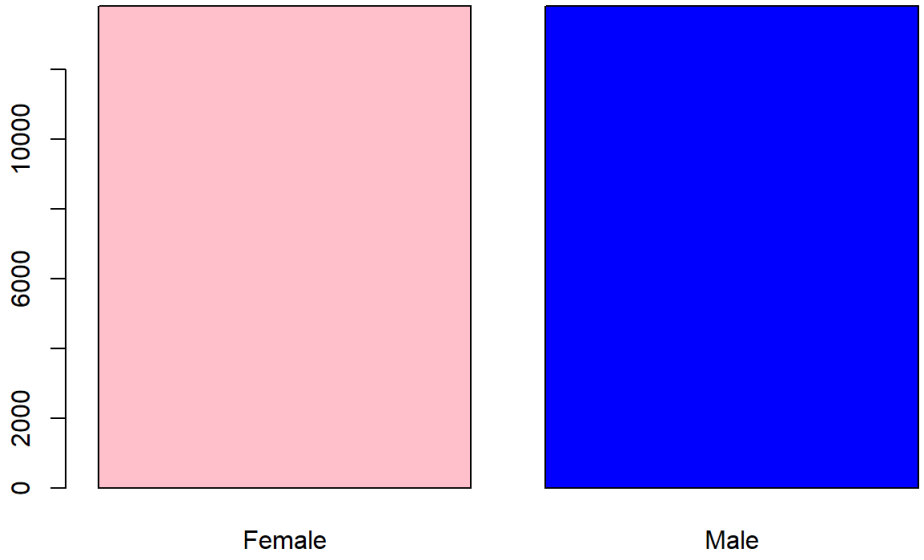
```
#load data in to "data"
load(file="suicide.RData")

barplot(table(data$generation),col = rainbow(20))
```



We can see that generation **G.I.** and **Z** are not so big as others. My explanation is that we can view all generations as **Gauss distribution**, where **G.I.** is at the start of distribution and **generation Z** is at the end of distribution. Both generation dont have so many people in their group, because either they are too old or too young

Sex

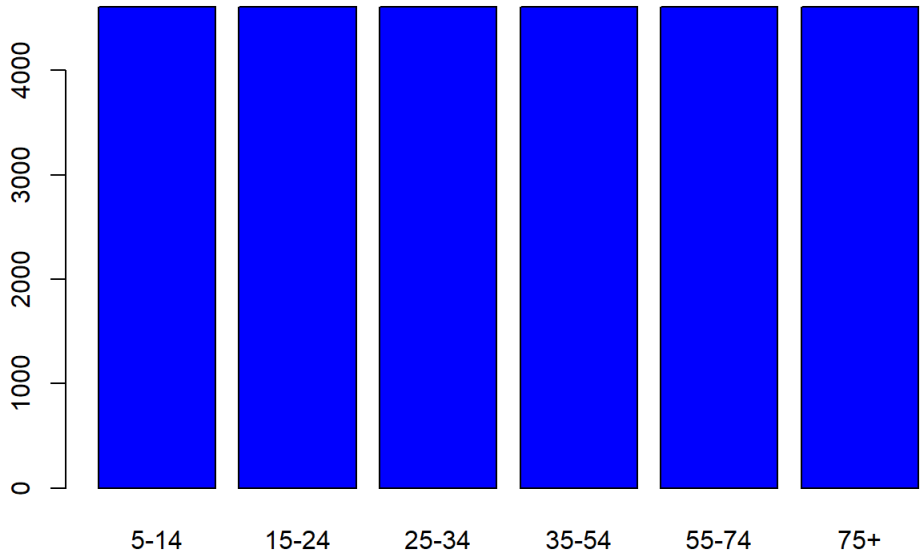
```
barplot(table(data$sex), col=c("pink", "blue"))
```



We can see we **same number** of males in females in dataset.

Age

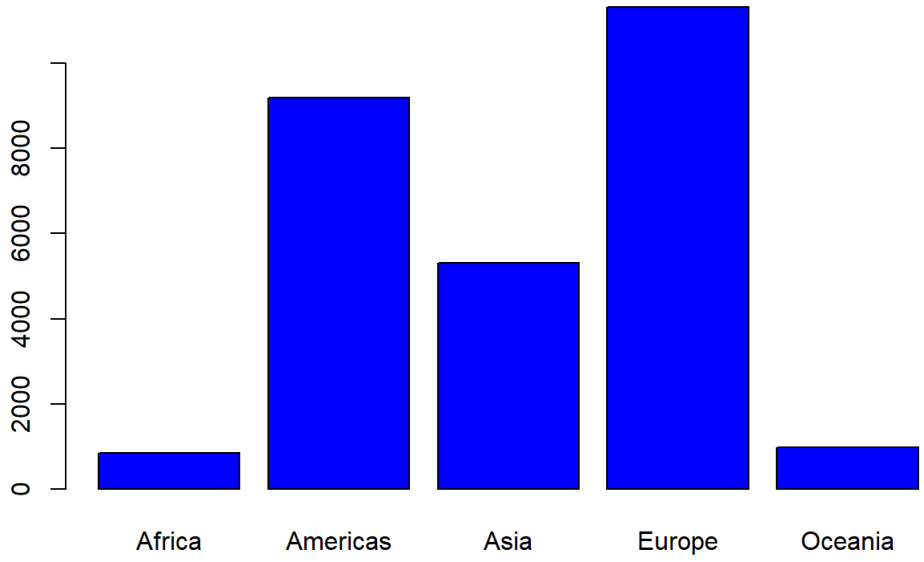
```
barplot(table(data$age), col="blue")
```



This is really **weird** for me. I would not divide data in these categories. I would make one category **for each age**. But I guess they chosed these categories because that makes data **equally distributed**.

Continent

```
barplot(table(data$continent), col="blue")
```



I wanted to have a look on **countries** at first. But there were too many of them and it was **difficult** to read.

So when we switch to continents. We can see most data comes from **Europe and America**.

So it is possible that the models and prediction will be more accurate for **them** and not so good for **other continents** like Africa or Oceania.

Analyze correlations

Now we analyzed key variables and it is time to move on some **'high-level'** charts like interesting corelations.

Suicide ratio

What about overall **suicide rate**? Is it increasing, decreasing or stagnating?

```
#libraries
library(magrittr)
library(dplyr)
library(ggplot2)


data %>%
  #sort it from the lowest year to the highest
  group_by(year) %>%
  # ?? summ
  summarize(population = sum(population),
            suicides = sum(suicides_no),
            suicides_per_100k = (suicides / population) * 100000) %>%
  #put name on axes
  ggplot(aes(x = year, y = suicides_per_100k)) +
  geom_line(col = "deepskyblue3", size = 1) +
  geom_point(col = "deepskyblue3", size = 2) +
  labs(title = "Global Suicides (per 100k)",
       subtitle = "Trend over time, 1985 - 2015.",
       x = "Year",
       y = "Suicides per 100k") +
  scale_x_continuous(breaks = seq(1985, 2015, 2)) +
  scale_y_continuous(breaks = seq(10, 20))
```

We can see that overall trend is that suicide rate is **decreasing overtime**. And this greatly shows how statistics **can be wrong**, because our world is too complex to be predict by these models! (For now)

If we see this **model**. Everyone would predict that in couple of years the suicide rate will be even lower. But **COVID-19** hit. And we can see that suicides are massively increasing [1] [2].

To have a **correct** prediction we would have to develop model which can predict, when war **strikes or pandemic strikes** etc. And that is impossible in **near future**, in my opinion.

Man vs female

This statistics should be pretty accurate because there is equal number man and woman in dataset.

Continents

There we will not get accurate results because, as I showed few sections before, we have far far more data from Europe/America than from other continents. So there can be huge innaccuracy.

Age

In plot few sections before we saw that this variable is equal among all categories so the results will be accurate too.

---

1. Gunnell D, Appleby L, Arensman E, et al., COVID-19 Suicide Prevention Research Collaboration. Suicide risk and prevention during the COVID-19 pandemic. Lancet Psychiatry2020;7:468-71. doi:10.1016/S2215-0366(20)30171-1 pmid:32330430↩
2. Reger MA, Stanley IH, Joiner TE. Suicide mortality and coronavirus disease 2019: a perfect storm?JAMA Psychiatry2020. [Epub ahead of print.] doi:10.1001/jamapsychiatry.2020.1060 pmid:32275300↩