

Project.R

zb117

2020-02-20

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.0 --
## v ggplot2 3.2.1      v purrr   0.3.3
## v tibble  2.1.3      v dplyr   0.8.3
## v tidyr   1.0.2      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.4.0
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
```

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
## The following object is masked from 'package:dplyr':
##
##   group_rows
```

```
library(treemap)
```

```
library(ggthemes)
```

```
library(highcharter)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method          from
##   as.zoo.data.frame zoo
## Highcharts (www.highcharts.com) is a Highsoft software product which is
## not free for commercial and Governmental use
```

```
library(summarytools)
```

```
## Registered S3 method overwritten by 'pryr':
##   method      from
##   print.bytes Rcpp
```

```

## For best results, restart R session and update pander using devtools:: or
remotes::install_github('rapporter/pander')

##

## Attaching package: 'summarytools'

## The following object is masked from 'package:tibble':

##
##      view

library(corrplot)

## corrplot 0.84 loaded

library(formattable)
library(ggcorrplot)

# Loading the packages
options(warn = -1)

packags <- c("tidyverse", "knitr", "kableExtra", "ggthemes", "treemap", "highc
harter", "summarytools", "ggcorrplot", "knitr", "formattable")

purrr::walk(packags, library, character.only = T, quietly = T)

# Importing dataset

data <- read.csv("/Users/zb117/Desktop/Suicide Rate Analysis.csv")

# Data summary

# We are using str() & head() function to inspect and have a brief overview o
f the dataset.

str(data)

## 'data.frame':    27820 obs. of  12 variables:
##  $ i..country      : Factor w/ 101 levels "Albania","Antigua and Barbuda
",...: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year            : int   1987 1987 1987 1987 1987 1987 1987 1987 1987 1
987 ...
##  $ sex             : Factor w/ 2 levels "female","male": 2 2 1 2 2 1 1 1
2 1 ...

```

```
## $ age          : Factor w/ 6 levels "15-24 years",...: 1 3 1 6 2 6 3
2 5 4 ...
## $ suicides_no   : int   21 16 14 1 9 1 6 4 1 0 ...
## $ population    : int   312900 308000 289700 21800 274300 35600 278800
257200 137500 311000 ...
## $ suicides.100k.pop : num   6.71 5.19 4.83 4.59 3.28 2.81 2.15 1.56 0.73 0
...
## $ country.year   : Factor w/ 2321 levels "Albania1987",...: 1 1 1 1 1 1
1 1 1 1 ...
## $ HDI.for.year   : num    NA NA NA NA NA NA NA NA NA NA ...
## $ gdp_for_year.... : Factor w/ 2321 levels "1,002,219,052,968",...: 727 7
27 727 727 727 727 727 727 727 727 ...
## $ gdp_per_capita....: int    796 796 796 796 796 796 796 796 796 796 ...
## $ generation     : Factor w/ 6 levels "Boomers","G.I. Generation",...:
3 6 3 2 1 2 6 1 2 3 ...
```

summary(data)

##	i..country	year	sex	age
##	Austria	: 382	Min.	:1985 female:13910 15-24 years:4642
##	Iceland	: 382	1st Qu.:	1995 male :13910 25-34 years:4642
##	Mauritius	: 382	Median	:2002 35-54 years:4642
##	Netherlands:	382	Mean	:2001 5-14 years :4610
##	Argentina	: 372	3rd Qu.:	2008 55-74 years:4642
##	Belgium	: 372	Max.	:2016 75+ years :4642
##	(Other)	:25548		
##	suicides_no	population	suicides.100k.pop	country.year
##	Min.	: 0.0	Min.	: 278 Min. : 0.00 Albania1987: 12
##	1st Qu.:	3.0	1st Qu.:	97498 1st Qu.: 0.92 Albania1988: 12
##	Median	: 25.0	Median	: 430150 Median : 5.99 Albania1989: 12
##	Mean	: 242.6	Mean	: 1844794 Mean : 12.82 Albania1992: 12
##	3rd Qu.:	131.0	3rd Qu.:	1486143 3rd Qu.: 16.62 Albania1993: 12
##	Max.	:22338.0	Max.	:43805214 Max. :224.97 Albania1994: 12
##				(Other) :27748
##	HDI.for.year	gdp_for_year....	gdp_per_capita....	
##	Min.	:0.483	1,002,219,052,968:	12 Min. : 251
##	1st Qu.:	0.713	1,011,797,457,139:	12 1st Qu.: 3447
##	Median	:0.779	1,016,418,229 :	12 Median : 9372
##	Mean	:0.777	1,018,847,043,277:	12 Mean : 16866

```
## 3rd Qu.:0.855 1,022,191,296 : 12 3rd Qu.: 24874
## Max. :0.944 1,023,196,003,075: 12 Max. :126352
## NA's :19456 (Other) :27748
## generation
## Boomers :4990
## G.I. Generation:2744
## Generation X :6408
## Generation Z :1470
## Millenials :5844
## Silent :6364
##
```

```
head(data)
```

```
## i..country year sex age suicides_no population suicides.100k.
pop
## 1 Albania 1987 male 15-24 years 21 312900 6
.71
## 2 Albania 1987 male 35-54 years 16 308000 5
.19
## 3 Albania 1987 female 15-24 years 14 289700 4
.83
## 4 Albania 1987 male 75+ years 1 21800 4
.59
## 5 Albania 1987 male 25-34 years 9 274300 3
.28
## 6 Albania 1987 female 75+ years 1 35600 2
.81

## country.year HDI.for.year gdp_for_year.... gdp_per_capita.... gener
ation
## 1 Albania1987 NA 2,156,624,900 796 Generat
ion X
## 2 Albania1987 NA 2,156,624,900 796 S
ilent
## 3 Albania1987 NA 2,156,624,900 796 Generat
ion X
## 4 Albania1987 NA 2,156,624,900 796 G.I. Gener
ation
## 5 Albania1987 NA 2,156,624,900 796 Bo
omers
## 6 Albania1987 NA 2,156,624,900 796 G.I. Gener
ation
```

```

# No of Columns in the dataset.
length(data[, -1])

## [1] 11

# Cleaning Data
# droppig NA values from suicide nos collumn.
clean_data <- data %>%
filter(suicides_no != "NA" & suicides_no!=0)

# Checking for the missing values in each collumns.
colSums(is.na(clean_data))

##           i..country           year           sex           ag
e
##           0           0           0
0

##           suicides_no           population  suicides.100k.pop           country.yea
r
##           0           0           0
0

##           HDI.for.year  gdp_for_year....  gdp_per_capita....           generatio
n
##           16332           0           0
0

# Cleaning HDI collumn.
clean_data$HDI.for.year <- NULL

# Changing collumn name.
colnames(data)[colnames(data) == "i..country"] <- "country"

# Data exploration

#Nearly 70% of the data is missing.
sum(is.na(data$HDI.for.year))/length(data$HDI.for.year) * 100

## [1] 69.9353

# Qualitative Variable frequencies

# No of occurences of each generation in the dataset.
data %>% group_by(generation) %>%

```

```
summarize(nb = n()) %>% kable () %>%
kable_styling(bootstrap_options = "striped", full_width = F)
```

generation	nb
Boomers	4990
G.I. Generation	2744
Generation X	6408
Generation Z	1470
Millenials	5844
Silent	6364

```
# X generation and silent are the most popular.
# Generation Z is the smallest group.
hcbars(x = data$generation, name = "Génération") %>%
hc_add_theme(hc_theme_economist())
```

GénérationBoomersG.I. GenerationGeneration XGeneration ZMillenialsSilent01k2k3k4k5k6k7k

```
# By Age Groups
# Age groups are equally distributed.
hcbars(x = data$age, name = " ge") %>%
hc_add_theme(hc_theme_economist())
```

ge15-24 years25-34 years35-54 years5-14 years55-74 years75+ years01k2k3k4k5k

```
# By Sex
# Both are equally distributed
hcbars(x = data$sex, name = "Sexe") %>%
hc_add_theme(hc_theme_economist())
```

Sexefemalemale02.5k5k7.5k10k12.5k15k

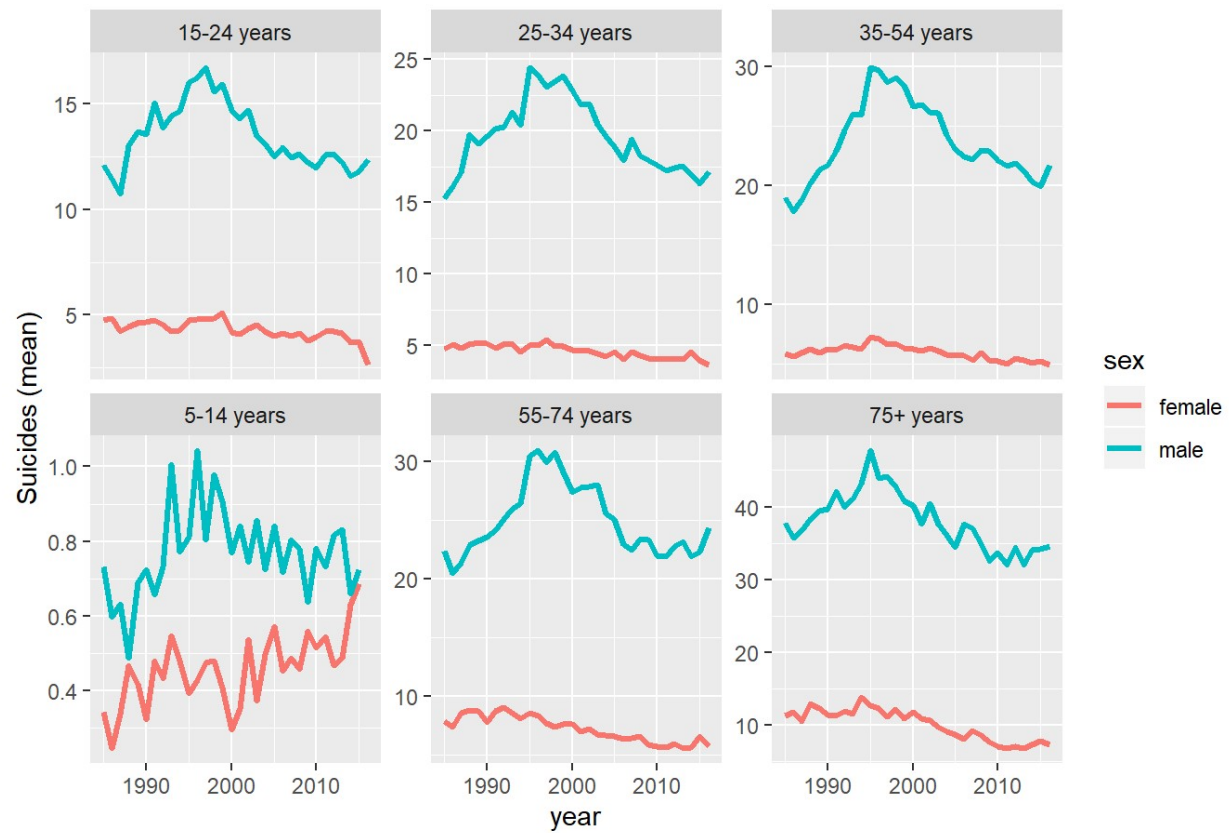
```
# By year
hbar(x = as.character(data$year), name = "Years") %>%
hc_add_theme(hc_theme_economist())
```

Years1985198619871988198919901991199219931994199519961997199819992000200120022003200420052006200720082009201020112012201320142015201620040060080010001200

```
# Suicide rates by Sex and Age group
# For all age groups suicide rate is higher for men than women.
# This means 'sex' variable differentiates the population of dataset.

data %>% group_by(year, sex, age) %>%
summarize(moy_suicide = mean(suicides.100k.pop)) %>%
ggplot(aes(x= year, y= moy_suicide)) +
geom_line(aes(color = sex), size=1.1) + facet_wrap(~age, scale = "free_y") +
ylab("Suicides (mean)") + ggtitle("Evolution of suicide rate per sex and age
categories")
```

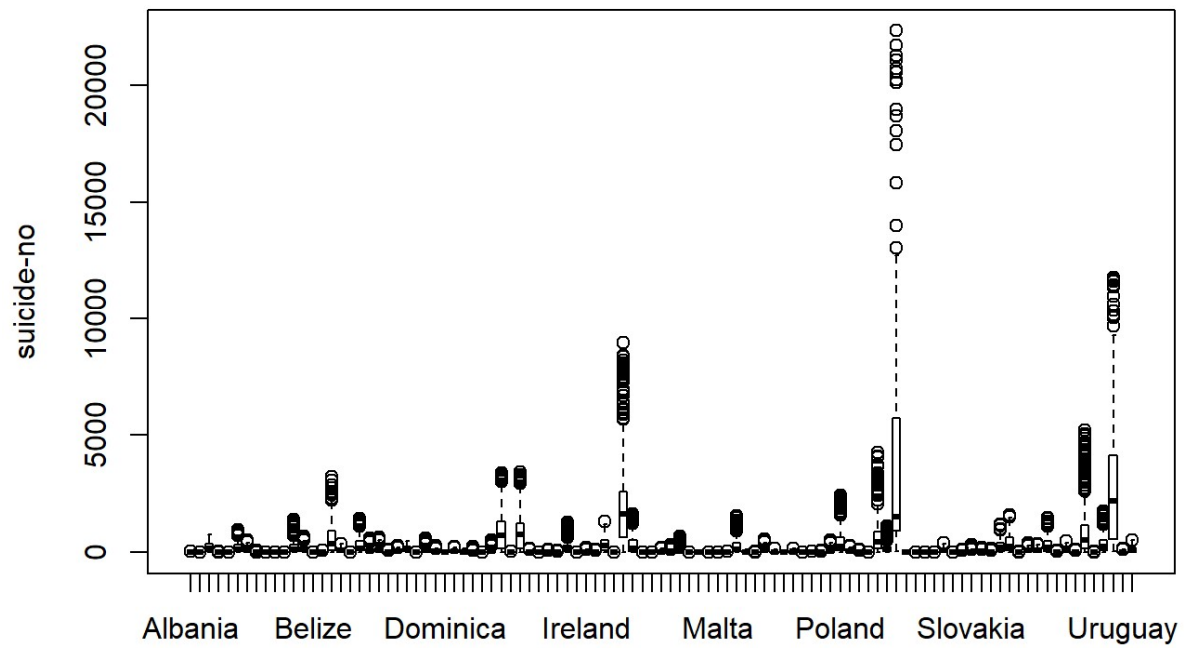
Evolution of suicide rate per sex and age categories



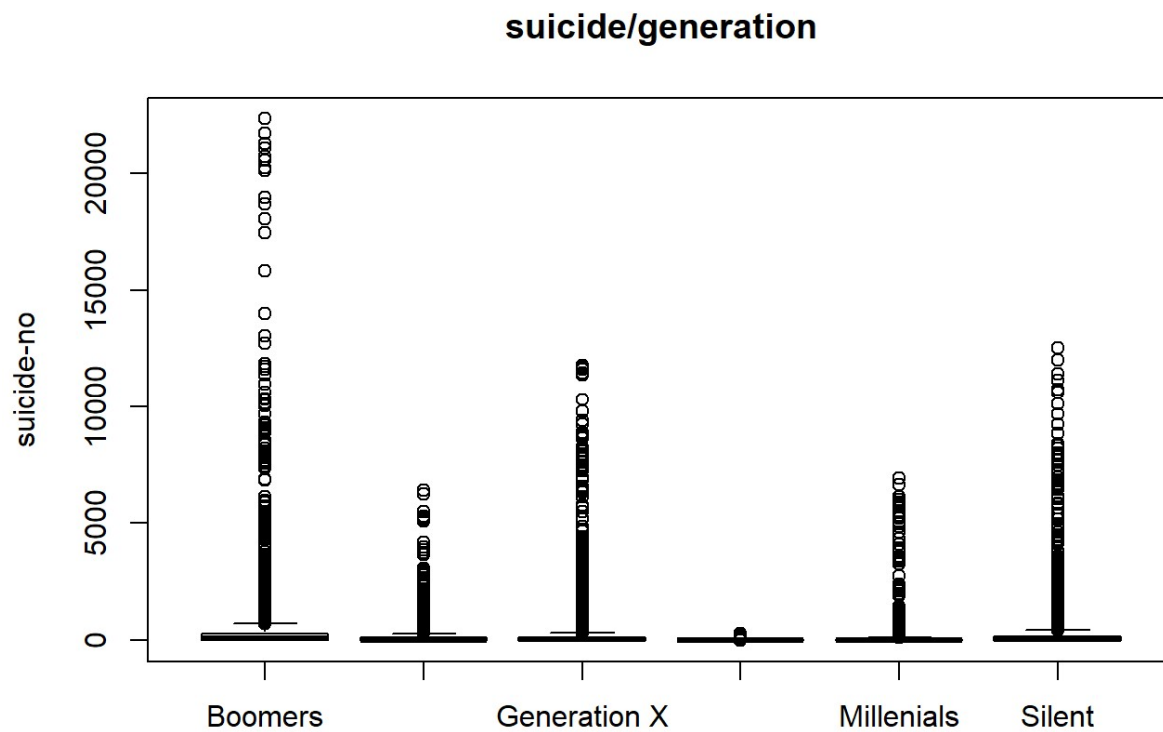
```
#Visualization
```

```
plot(data[,1], data[,5], main = "suicide/country", xlab="", ylab="suicide-no")
```


suicide/country



```
plot(data[,12], data[,5], main = "suicide/generation", xlab="", ylab="suicide-  
no")
```



```
# Tests
```

```
# 1. Analysis Of Variance
```

```
temp <- data %>% group_by(year,sex, age) %>%
summarize(moy_suicide = mean(suicides.100k.pop))
fit <- aov(formula = moy_suicide~age+sex, data = temp)
summary(fit)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## age         5  19222    3844   171.4 <2e-16 ***
## sex         1  20754   20754   925.1 <2e-16 ***
## Residuals   375    8413     22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

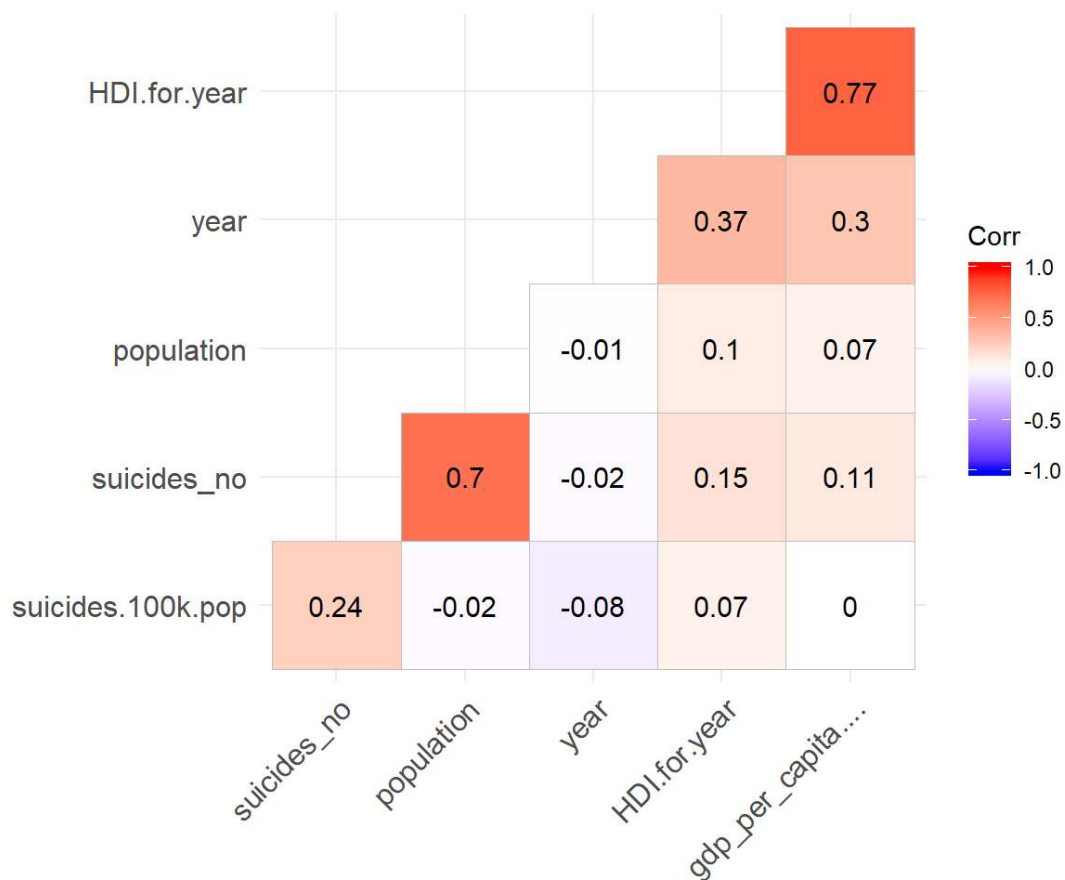
```
# There's a Statistical significance between the two groups men and women.
```

```
# This difference is stronger than the age group.
```

```
# (Look at the F value, by default R doesn't print numbers < to 2e-16).

# 2. Co-relation between pairs of quantitative variables.

options(repr.plot.height = 4, repr.plot.res = 180, repr.plot.width = 6)
data[,sapply(data, is.numeric)] %>%
cor(use = "complete.obs") %>% ggcorrplot(hc.order = TRUE, type = "lower", lab
= TRUE)
```



```
# Human development index correlates positively with the per capita GDP (0.77),
# which means that these two variables tend to go in the same direction.

# The number of suicide is positively correlated with the population (country).
```

```
# The reason for this correlation is that it is a time series: the number of
suicide increases with the growth of the population which itself increases wi
th time,
```

```
# the same goes for the GDP, which increases every year.
```

```
# 3. T-test
```

```
data_1<-transform(data, age = as.numeric(age))
```

```
t.test(data_1$age,data_1$sucides_no, var.equal = TRUE, paired=FALSE)
```

```
##
```

```
## One Sample t-test
```

```
##
```

```
## data: data_1$age
```

```
## t = 341.58, df = 27819, p-value < 2.2e-16
```

```
## alternative hypothesis: true mean is not equal to 0
```

```
## 95 percent confidence interval:
```

```
## 3.479345 3.519505
```

```
## sample estimates:
```

```
## mean of x
```

```
## 3.499425
```

```
t.test(data_1$gdp_per_capita,data_1$sucides_no, var.equal = TRUE, paired=FALSE)
```

```
##
```

```
## One Sample t-test
```

```
##
```

```
## data: data_1$gdp_per_capita
```

```
## t = 148.95, df = 27819, p-value < 2.2e-16
```

```
## alternative hypothesis: true mean is not equal to 0
```

```
## 95 percent confidence interval:
```

```
## 16644.51 17088.42
```

```
## sample estimates:
```

```
## mean of x
```

```
## 16866.46
```