Я буду використовувати такі бібліотеки R в своєму аналізі:

- library(magrittr)
- library(plyr)
- library(dplyr)
- library(ggplot2)
- library(grid)
- library(gridExtra)

Завантаження даних

Першим кроком у процесі аналізу наборів даних є їх завантаження. Отже, ініціалізуємо змінну df:

```
df <- read.csv(r"{C:\Users\DELL\Documents\data\master.csv}")
```

Описова статистика, Feature Engineering

Оглянемо 5 перших рядків датафрейму.

```
> head(df, 5)
  country year    sex         age suicides_no population suicides.100k.pop country.year
1 Albania 1987   male 15-24 years          21     312900              6.71  Albania1987
2 Albania 1987   male 35-54 years          16     308000              5.19  Albania1987
3 Albania 1987 female 15-24 years          14     289700              4.83  Albania1987
4 Albania 1987   male   75+ years           1      21800              4.59  Albania1987
5 Albania 1987   male 25-34 years           9     274300              3.28  Albania1987
  HDI.for.year gdp_for_year.... gdp_per_capita....      generation
1           NA    2,156,624,900               796    Generation X
2           NA    2,156,624,900               796          Silent
3           NA    2,156,624,900               796    Generation X
4           NA    2,156,624,900               796 G.I. Generation
5           NA    2,156,624,900               796         Boomers
>
```

Яка розмірність даних? Типи змінних?

```
> dim(df)
[1] 27820    12
> str(df)
'data.frame':   27820 obs. of  12 variables:
 $ country           : chr  "Albania" "Albania" "Albania" "Albania" ...
 $ year              : int  1987 1987 1987 1987 1987 1987 1987 1987 1987 1987 ...
 $ sex               : chr  "male" "male" "female" "male" ...
 $ age               : chr  "15-24 years" "35-54 years" "15-24 years" "75+ years" ...
 $ suicides_no       : int  21 16 14 1 9 1 6 4 1 0 ...
 $ population        : int  312900 308000 289700 21800 274300 35600 278800 257200 137500 31
1000 ...
 $ 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      : chr  "Albania1987" "Albania1987" "Albania1987" "Albania1987" ...
 $ HDI.for.year      : num  NA NA NA NA NA NA NA NA NA NA ...
 $ gdp_for_year....  : chr  "2,156,624,900" "2,156,624,900" "2,156,624,900" "2,156,624,900"
 ...
 $ gdp_per_capita....: int  796 796 796 796 796 796 796 796 796 796 ...
 $ generation        : chr  "Generation X" "Silent" "Generation X" "G.I. Generation" ...
>
```

Чи є в нас пропущені значення?

```
> sum(!complete.cases(df)) # кількість рядків з пропущеними значеннями
[1] 19456
```

Так, ми маємо дуже багато «неповних» рядків.

Який розподіл пропущених значень відносно змінних (колонок)?

```
> colSums(is.na(df))
           country               year                sex                age
                 0                  0                  0                  0
       suicides_no         population  suicides.100k.pop       country.year
                 0                  0                  0                  0
      HDI.for.year   gdp_for_year.... gdp_per_capita....         generation
             19456                  0                  0                  0
```

Бачимо, що «найбільше проблем» ми маємо з колонкою *HDI.for.year*. Я її видалю.

```
> df$HDI.for.year <- NULL
> names(df)
 [1] "country"            "year"               "sex"                "age"
 [5] "suicides_no"        "population"         "suicides.100k.pop"  "country.year"
 [9] "gdp_for_year...."   "gdp_per_capita...." "generation"
> sum(!complete.cases(df)) # кількість рядків з пропущеними значеннями
[1] 0
```

Тепер у нас 0 пропущених значень. Чудово.

Повернімося до типів даних. Я помітив, що деякі колонки, які мають бути int, мають тип char, і тому подібне. Також треба сконвертувати колонки типу char у factor.

```r
str(df)
# 'data.frame': 27820 obs. of  11 variables:
#   $ country           : chr  "Albania" "Albania" "Albania" "Albania" ...
# $ year              : int  1987 1987 1987 1987 1987 1987 1987 1987 1987 1987 ...
# $ sex               : chr  "male" "male" "female" "male" ...
# $ age               : chr  "15-24 years" "35-54 years" "15-24 years" "75+ years" ...
# $ 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      : chr  "Albania1987" "Albania1987" "Albania1987" "Albania1987" ...
# $ gdp_for_year....  : chr  "2,156,624,900" "2,156,624,900" "2,156,624,900" "2,156,624,900" ...
# $ gdp_per_capita....: int  796 796 796 796 796 796 796 796 796 796 ...
# $ generation        : chr  "Generation X" "Silent" "Generation X" "G.I. Generation" ...

factor_cols <- c('country', 'sex',  'age',  'country.year', 'generation')

df$gdp_for_year.... = as.numeric(gsub(",", "", df$gdp_for_year....))

df[factor_cols] <- lapply(df[factor_cols], factor)  ## as.factor() could also be used
```

```r
str(df)
# 'data.frame': 27820 obs. of  12 variables:
#   $ 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 1987 ...
# $ 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 ...
# $ gdp_for_year....  : num  2.16e+09 2.16e+09 2.16e+09 2.16e+09 2.16e+09 ...
# $ 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 ...
# $ gdp_for_year      : num  2.16e+09 2.16e+09 2.16e+09 2.16e+09 2.16e+09 ...

numericVars <- select_if(df, is.numeric)

factorVars <- select_if(df, is.factor)

cat("This dataset has ", length(numericVars), "numeric Variables and ",
    length(factorVars),"factor Variables")
# This dataset has  7 numeric Variables and  5 factor Variables
```

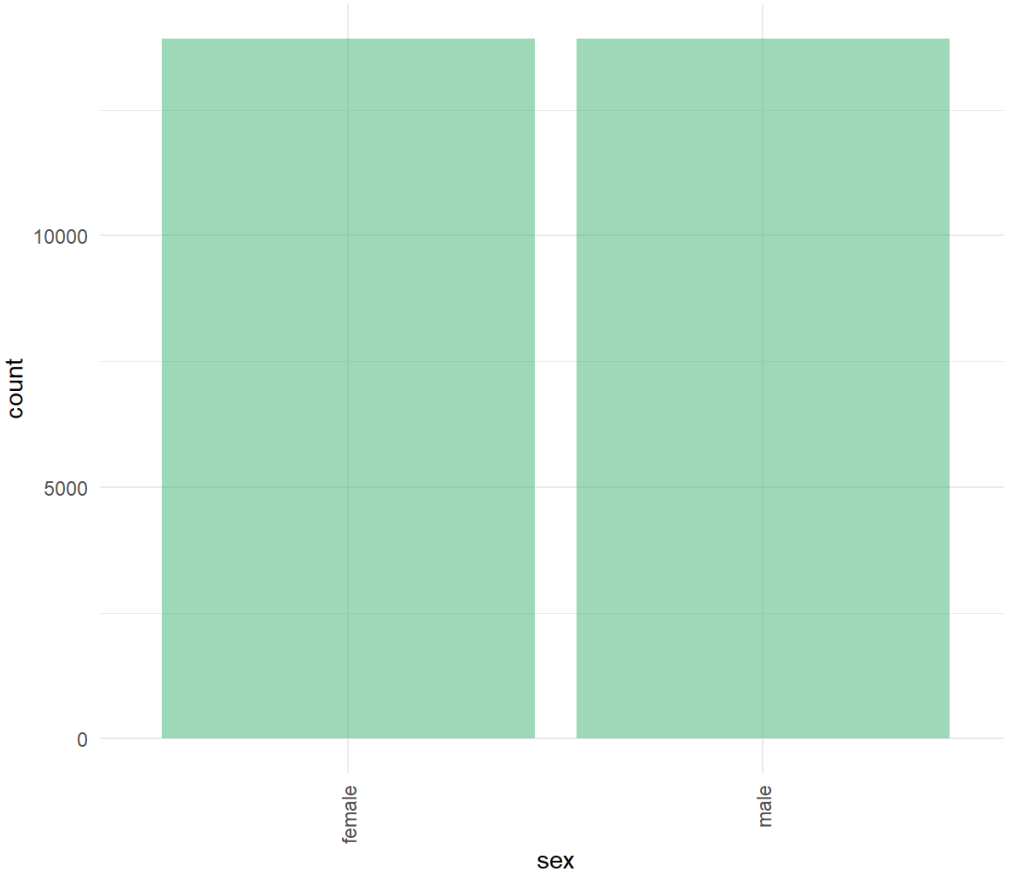
Нарешті все готово для аналізу.

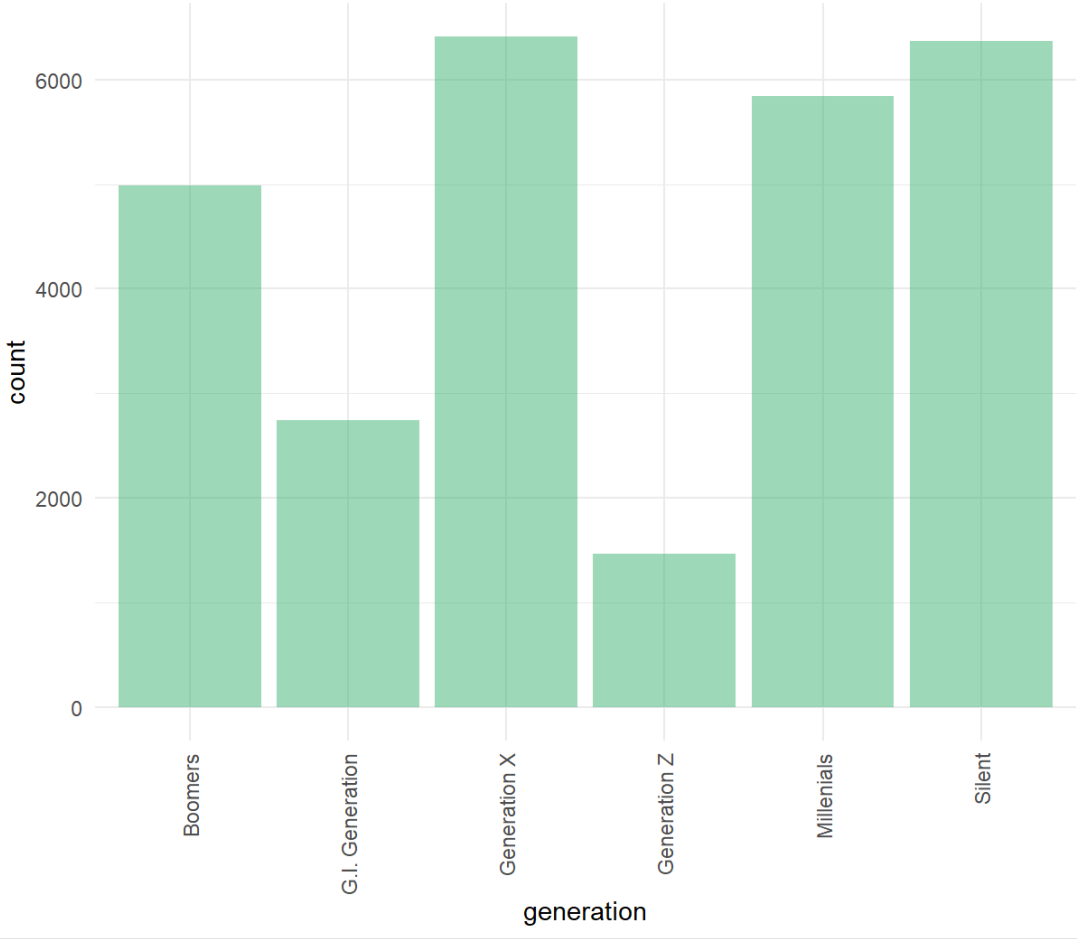
Однофакторні сюжети

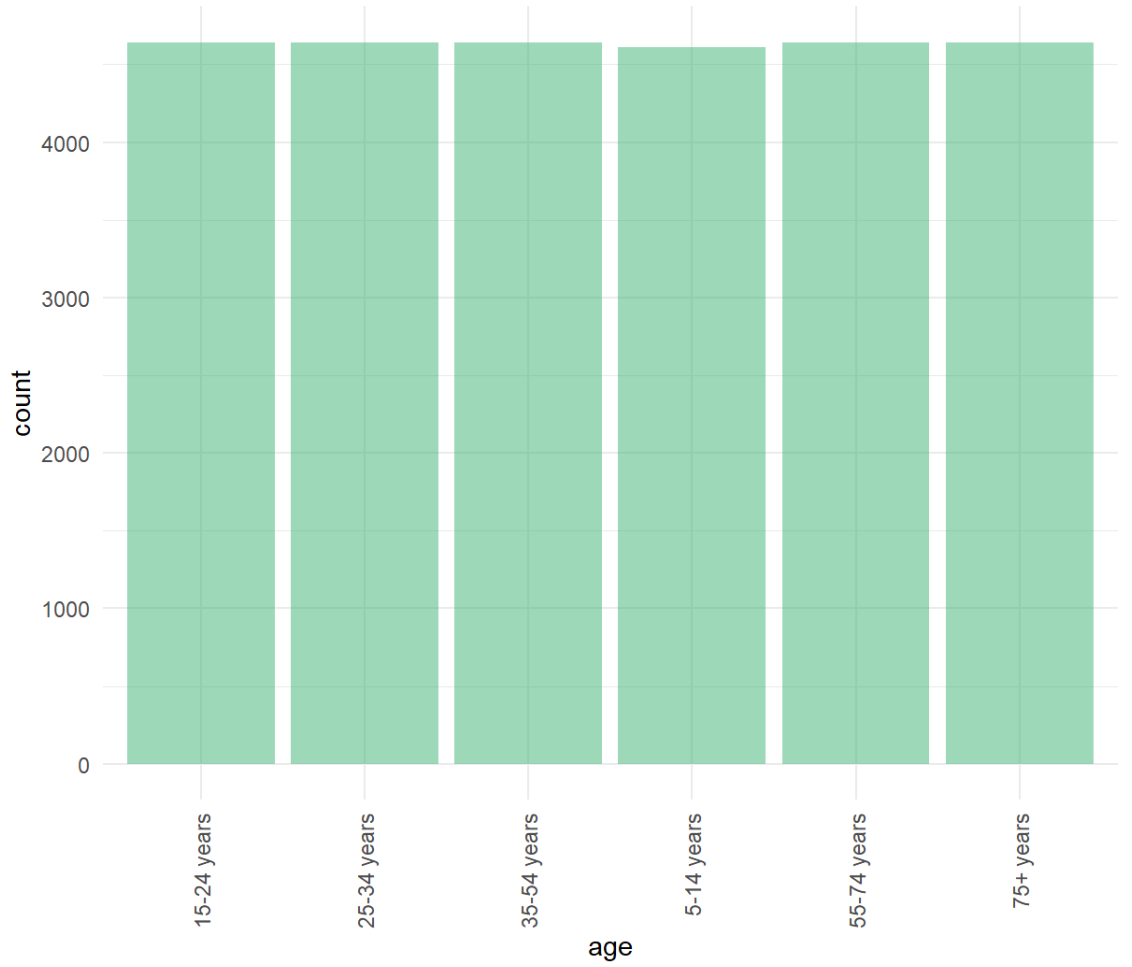
У цьому розділі я подивлюся на розподіл значень для кожної змінної в наборі даних, створюючи різні графіки за допомогою бібліотеки ggplot2. Я намагаюся з'ясувати, чи є більше даних для очищення, включаючи викиди чи сторонні значення. Це також може допомогти мені почати ідентифікувати будь-які зв'язки між змінними, які варто дослідити далі. (код див. у файлі з .R розширенням.)

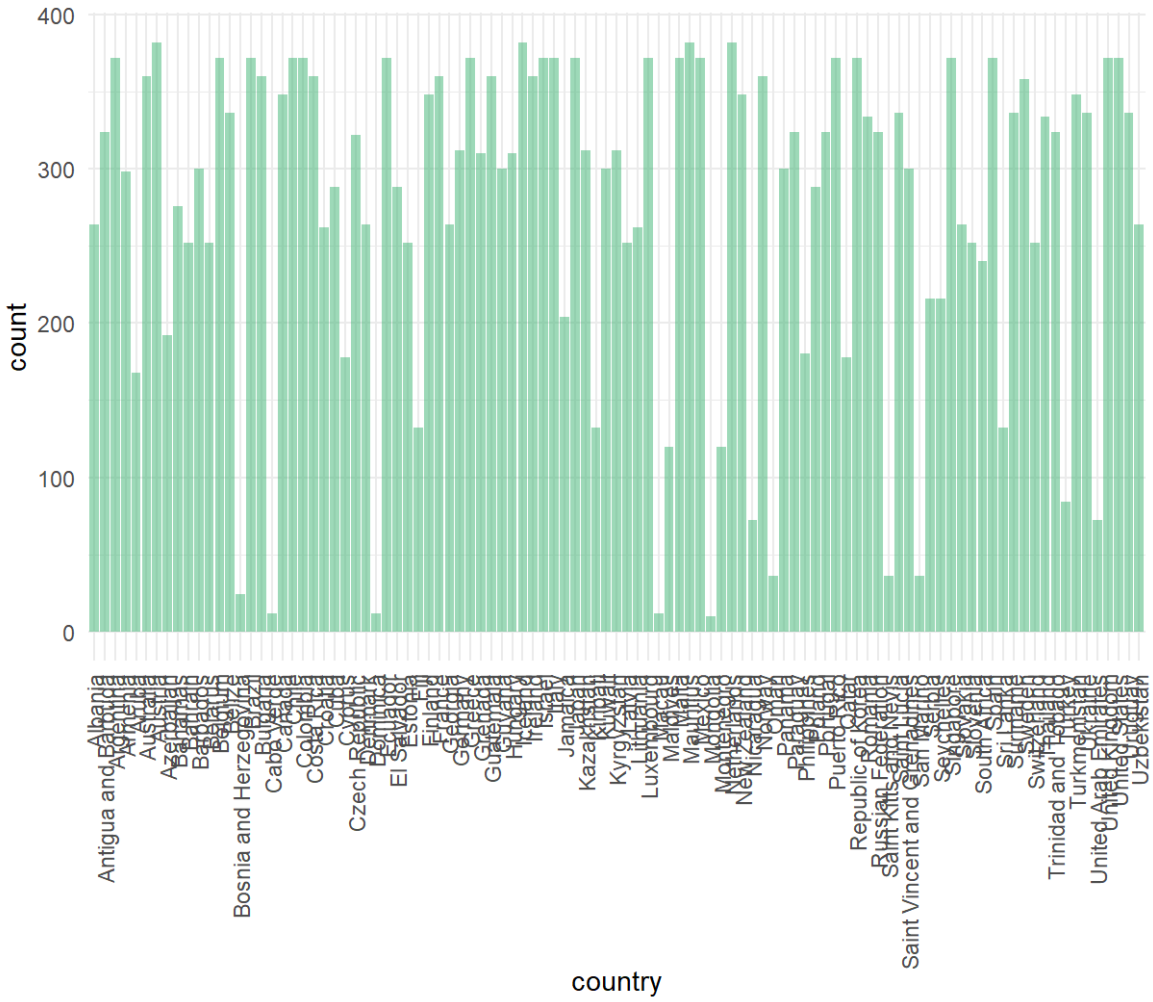
```
> summary(df)
        country           year          sex
 Austria    :  382   Min.   :1985   female:13910
 Iceland    :  382   1st Qu.:1995   male  :13910
 Mauritius  :  382   Median :2002
 Netherlands:  382   Mean   :2001
 Argentina  :  372   3rd Qu.:2008
 Belgium    :  372   Max.   :2016
 (Other)    :25548
          age         suicides_no        population
 15-24 years:4642   Min.   :    0.0   Min.   :     278
 25-34 years:4642   1st Qu.:    3.0   1st Qu.:   97498
 35-54 years:4642   Median :   25.0   Median :  430150
 5-14 years :4610   Mean   :  242.6   Mean   : 1844794
 55-74 years:4642   3rd Qu.:  131.0   3rd Qu.: 1486143
 75+ years  :4642   Max.   :22338.0   Max.   :43805214

 suicides.100k.pop      country.year   gdp_for_year....
 Min.   :  0.00    Albania1987:   12   Min.   :4.692e+07
 1st Qu.:  0.92    Albania1988:   12   1st Qu.:8.985e+09
 Median :  5.99    Albania1989:   12   Median :4.811e+10
 Mean   : 12.82    Albania1992:   12   Mean   :4.456e+11
 3rd Qu.: 16.62    Albania1993:   12   3rd Qu.:2.602e+11
 Max.   :224.97    Albania1994:   12   Max.   :1.812e+13
                   (Other)    :27748
 gdp_per_capita....          generation    gdp_for_year
 Min.   :   251     Boomers        :4990   Min.   :4.692e+07
 1st Qu.:  3447     G.I. Generation:2744   1st Qu.:8.985e+09
 Median :  9372     Generation X   :6408   Median :4.811e+10
 Mean   : 16866     Generation Z   :1470   Mean   :4.456e+11
 3rd Qu.: 24874     Millenials     :5844   3rd Qu.:2.602e+11
 Max.   :126352     Silent         :6364   Max.   :1.812e+13
```

1. Факторні дані

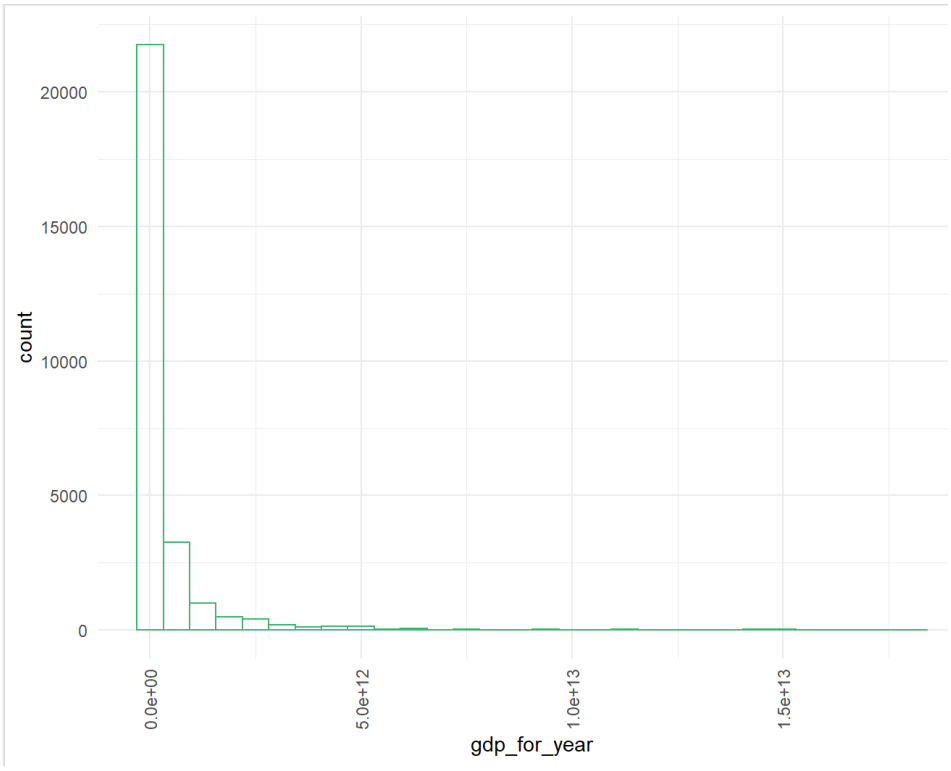


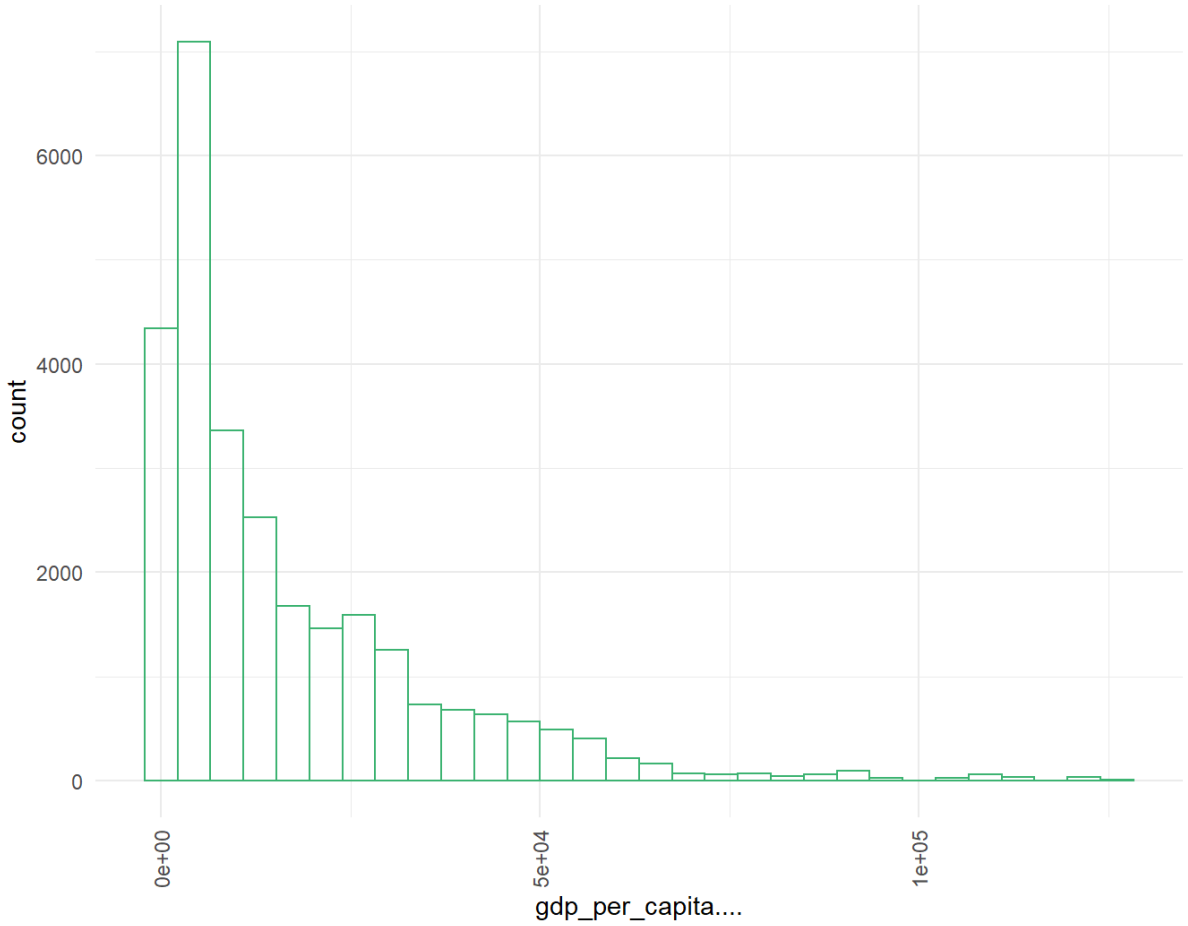


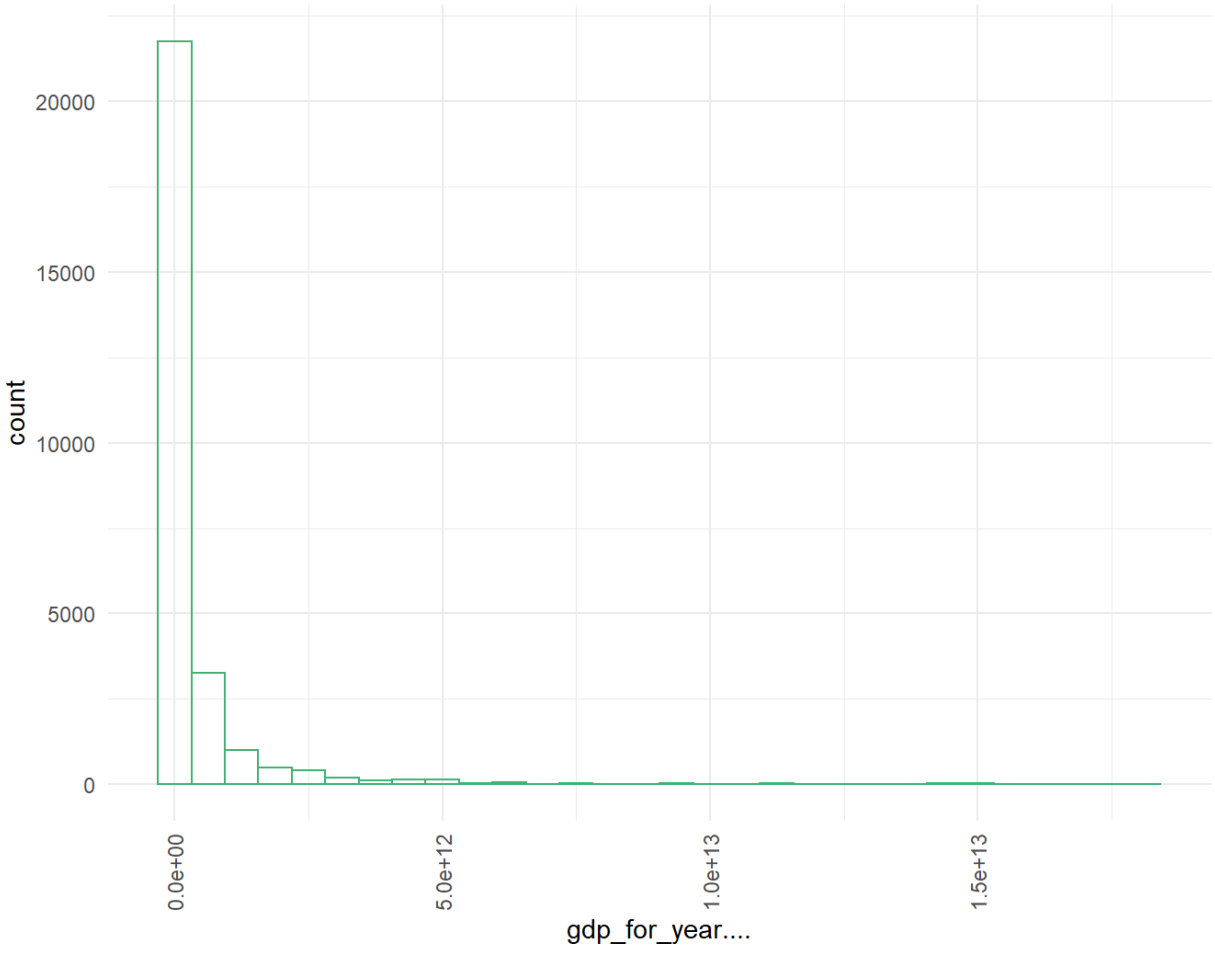

count
4000
3000
2000
1000
0
15-24 years
25-34 years
35-54 years
5-14 years
55-74 years
75+ years
age

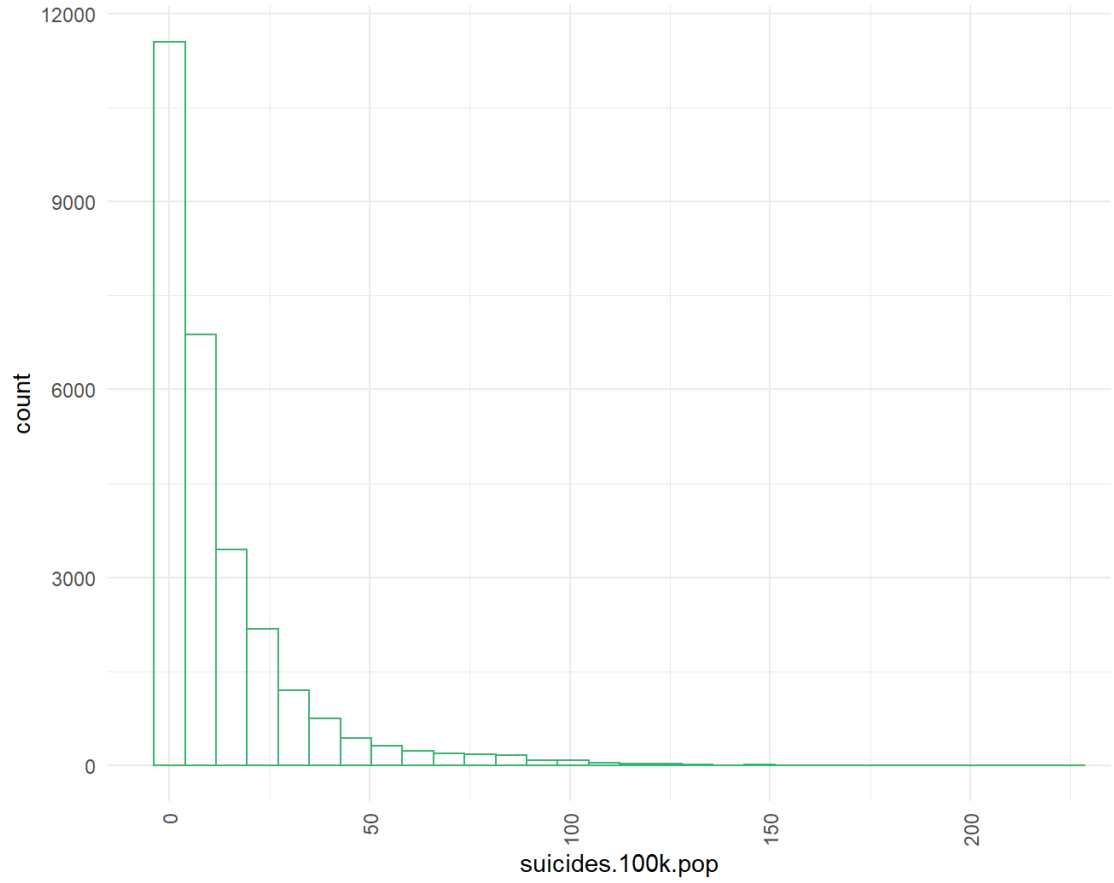


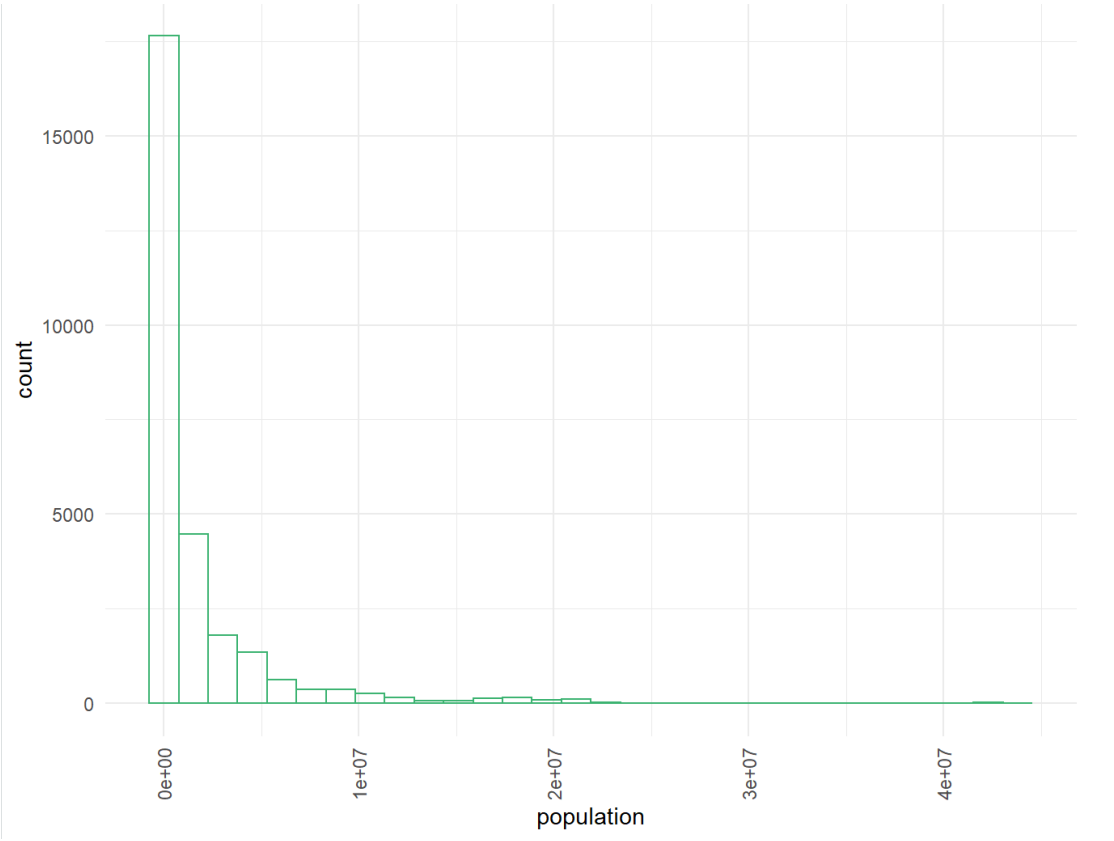
2. Числові змінні

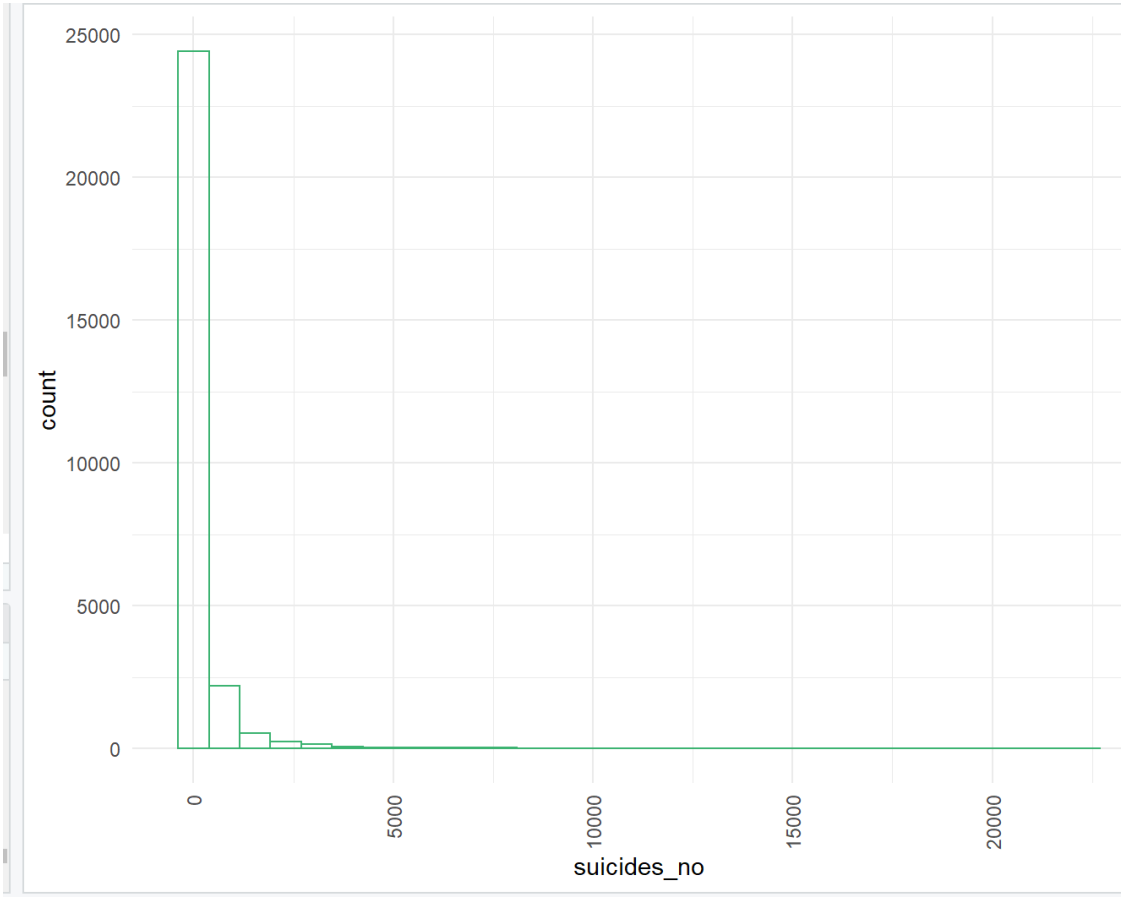


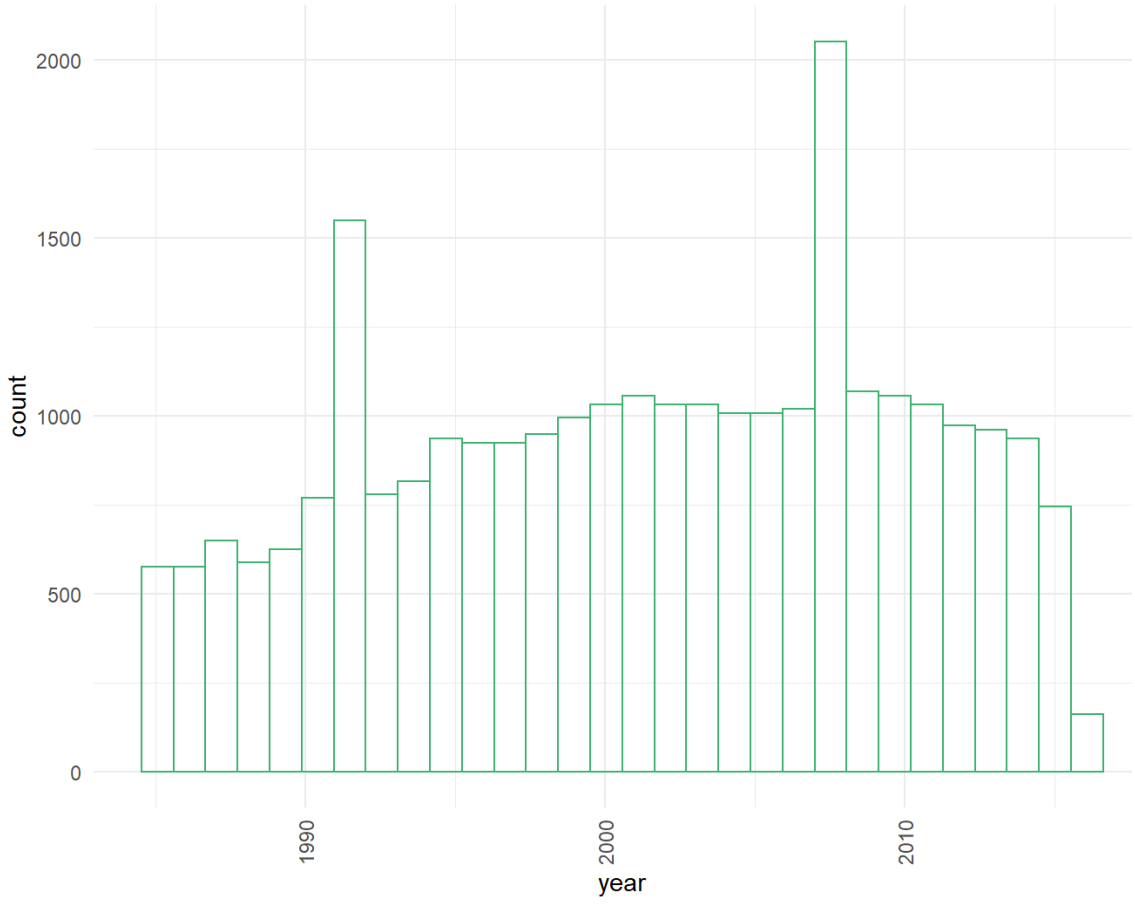

6000
4000
2000
0
count
0e+00
5e+04
1e+05
gdp_per_capita....


20000
15000
10000
5000
0
count
0.0e+00
5.0e+12
1.0e+13
1.5e+13
gdp_for_year....

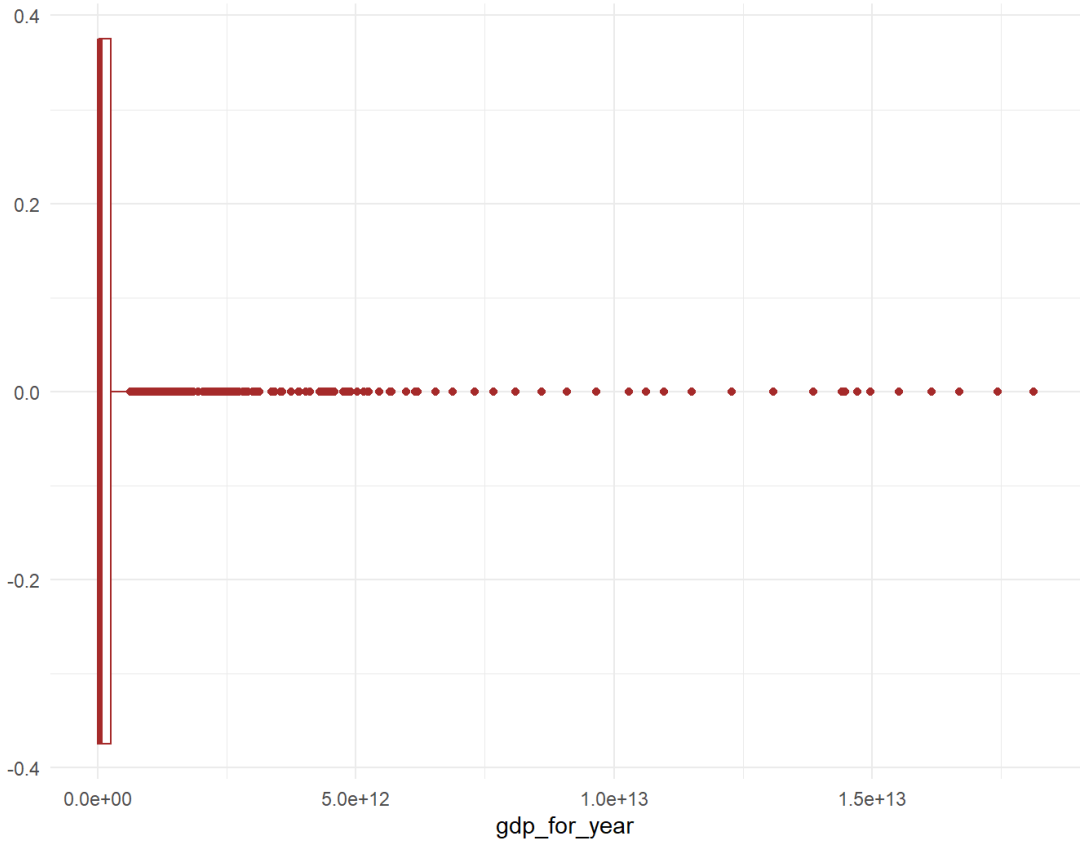

12000
9000
6000
3000
0
count
0
50
100
150
200
suicides.100k.pop

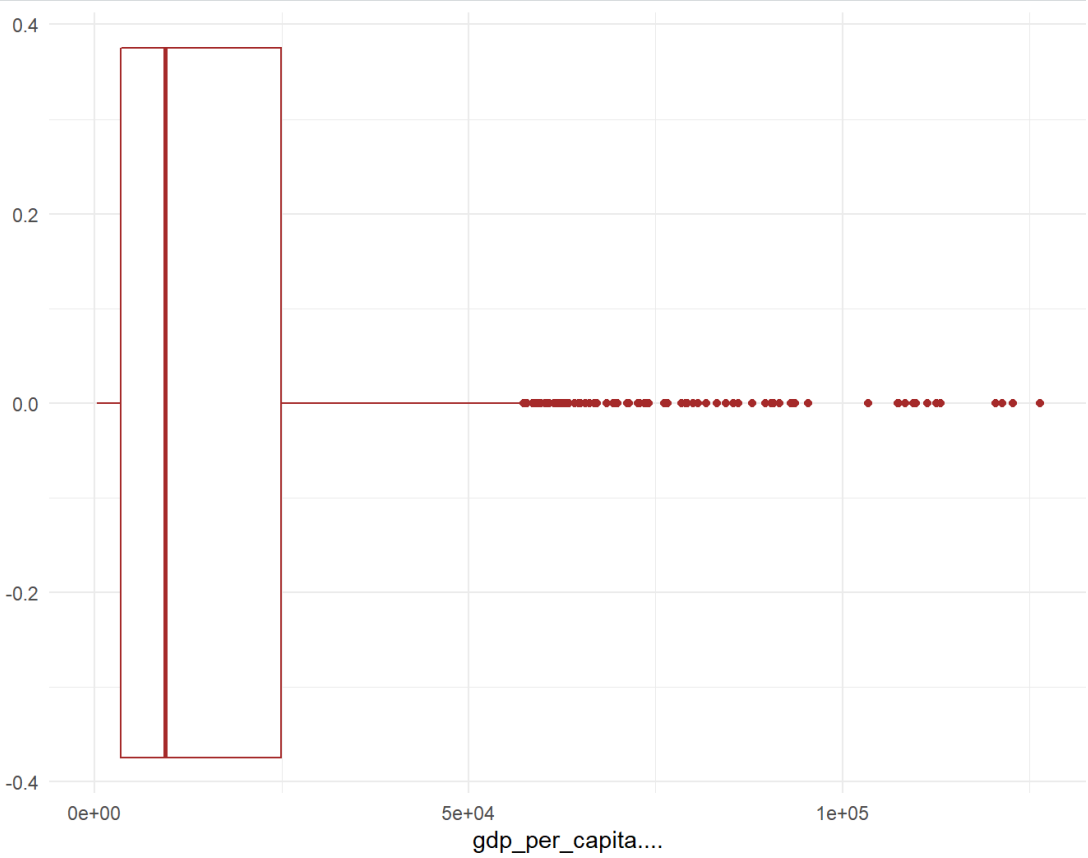

15000
10000
5000
0
count
0e+00
1e+07
2e+07
3e+07
4e+07
population

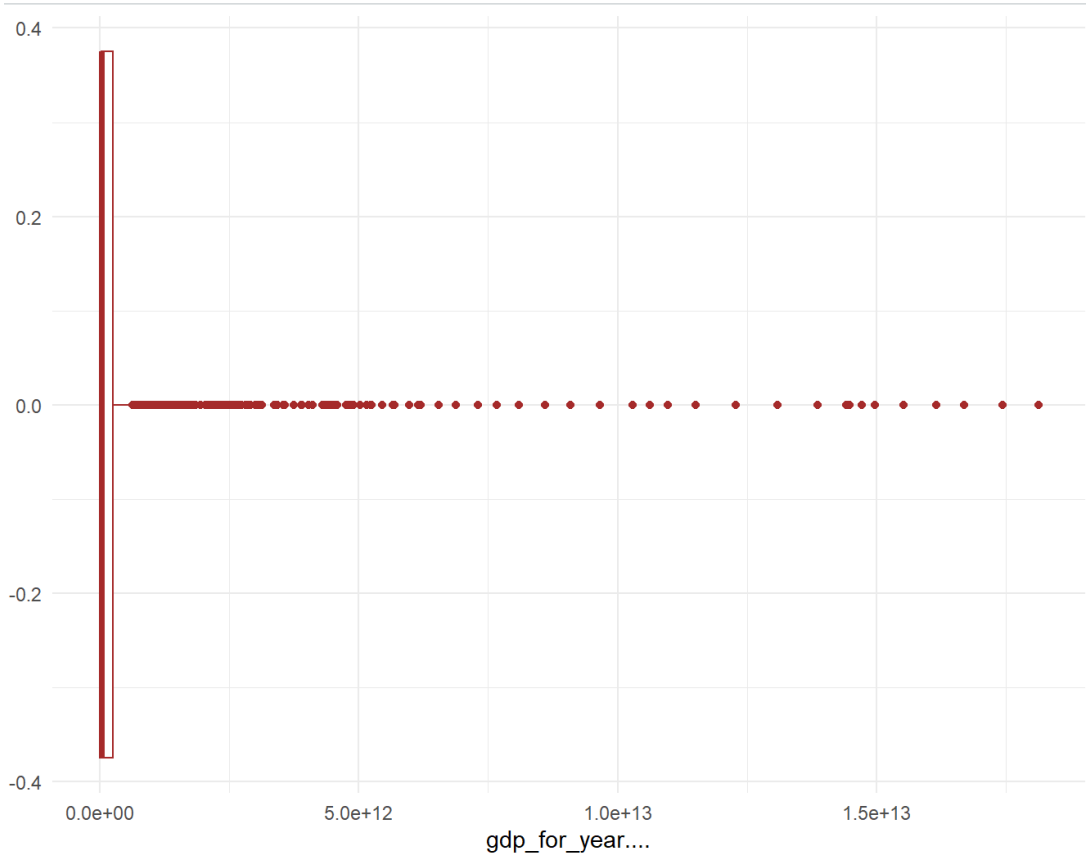

25000
20000
15000
10000
5000
0
count
0
5000
10000
15000
20000
suicides_no

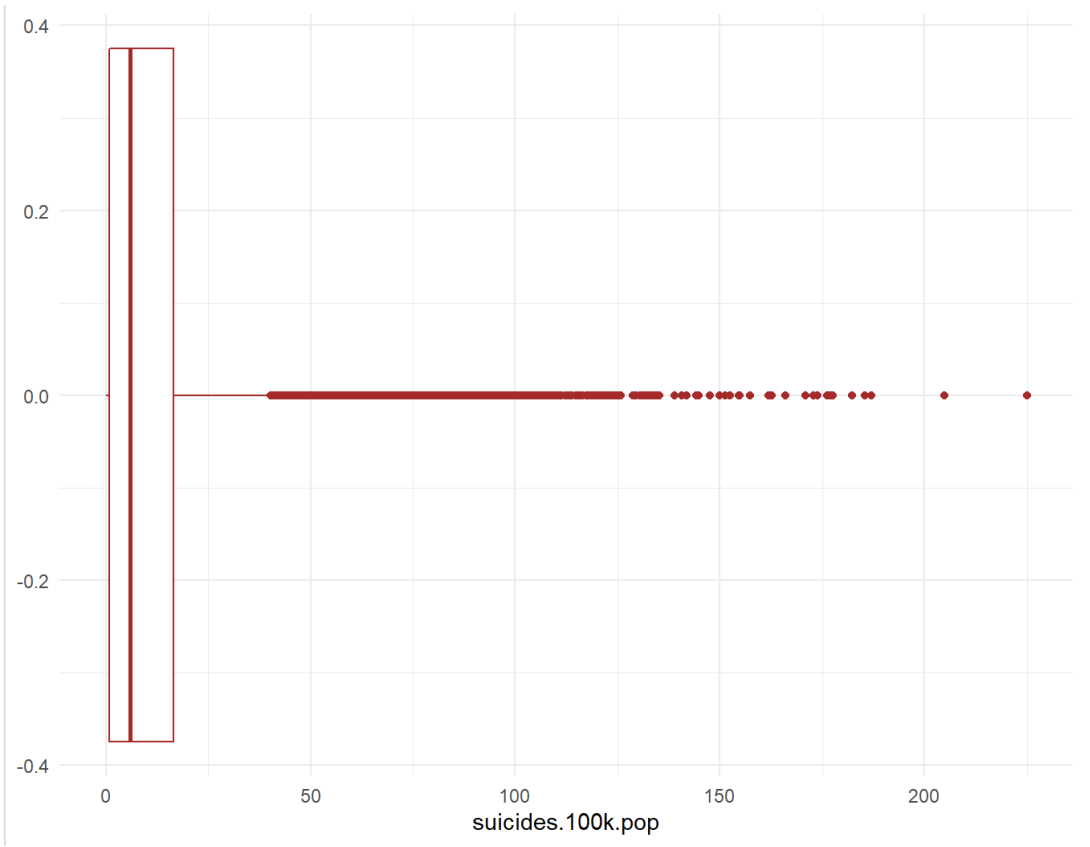

2000
1500
1000
500
0
count
1990
2000
2010
year

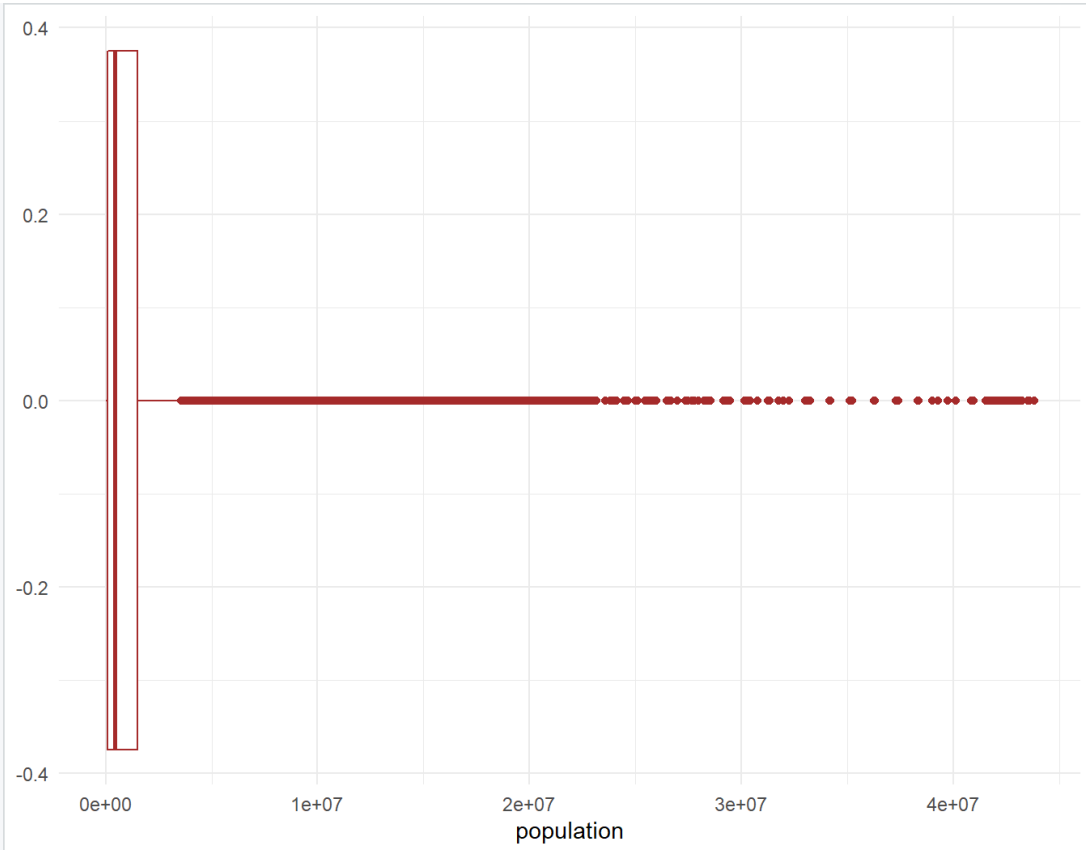
3. Боксплоти для числових змінних.

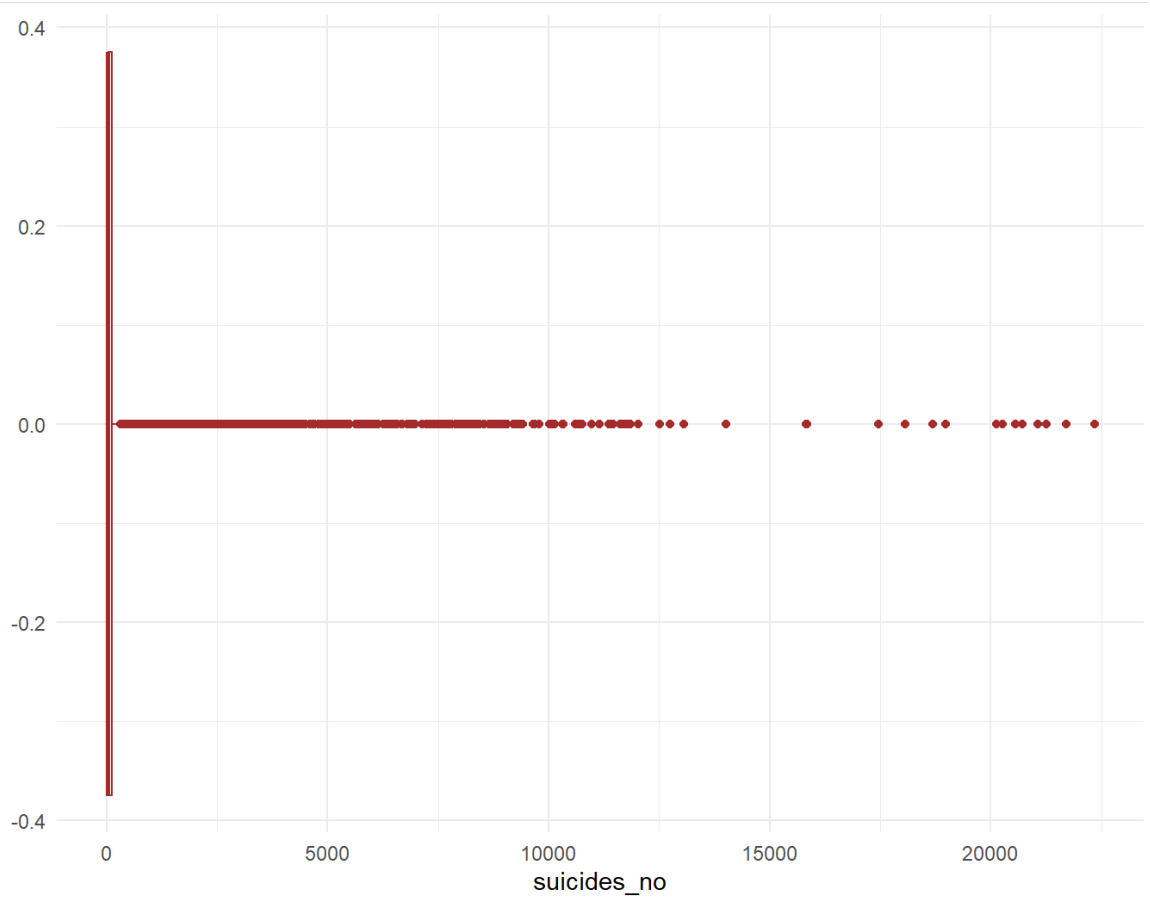


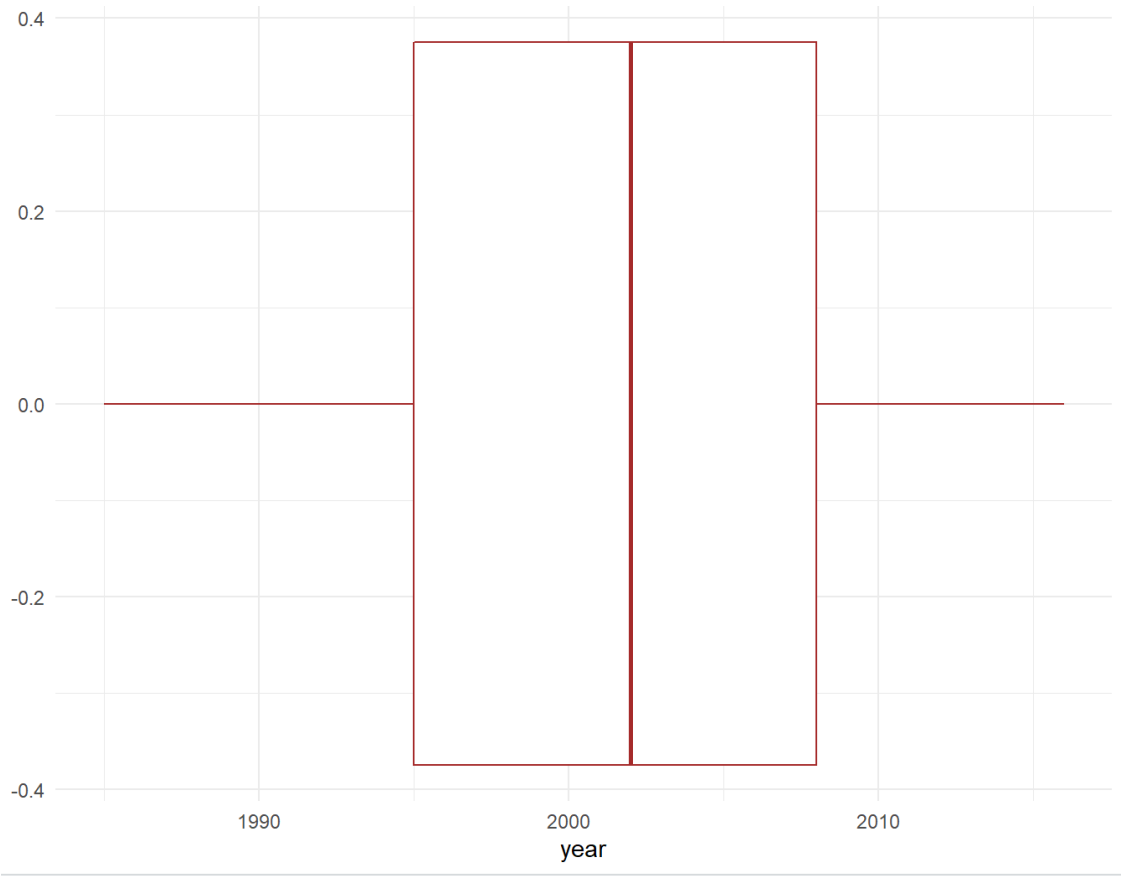



0.4
0.2
0.0
-0.2
-0.4
0.0e+00
5.0e+12
1.0e+13
1.5e+13
gdp_for_year....


0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
200
suicides.100k.pop


0.4
0.2
0.0
-0.2
-0.4
0e+00
1e+07
2e+07
3e+07
4e+07
population


0.4
0.2
0.0
-0.2
-0.4
0
5000
10000
15000
20000
suicides_no

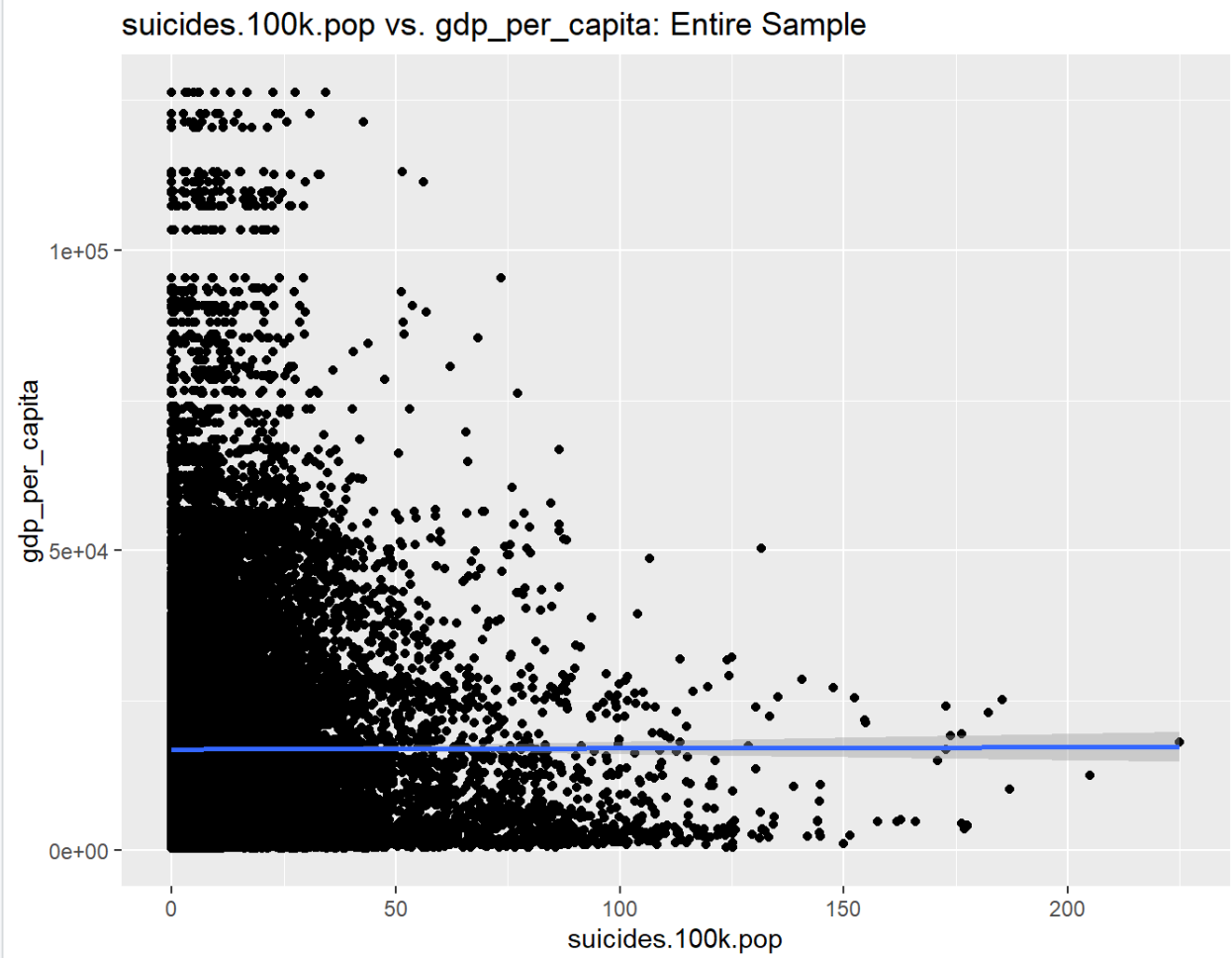


Як видно з гістограм і ящиків з вусами – наші змінні містять дуже багато аномальних викидів.

Двомірні ділянки

У цьому розділі я буду використовувати більше методів створення діаграм ggplot2, щоб уявити, як одна змінна впливає на іншу. Я почну з того, як вага впливає на MPG, зробивши діаграму розсіювання, накладену лінійною лінією, яка найкраще підходить.

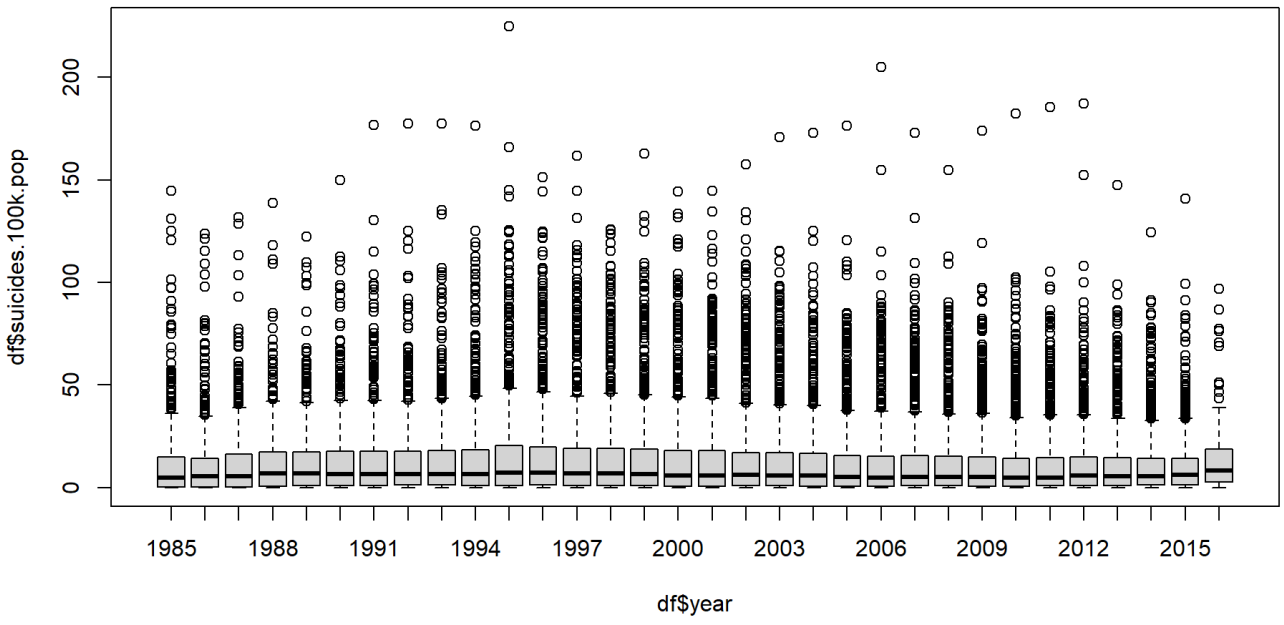
```
> ggplot(data = df, aes(x = suicides.100k.pop, y = gdp_per_capit
a....)) +
+   geom_point() +
+   geom_smooth(method='lm') +
+   xlab('suicides.100k.pop') +
+   ylab('gdp_per_capita') +
+   ggtitle('suicides.100k.pop vs. gdp_per_capita: Entire Sample')
`geom_smooth()` using formula 'y ~ x'
>
```



Як бачимо, чим менший показник ВВП на душу населення, тим вищий показник самогубств на 100 000. Авжеж є вийнятки з цього правила, але основна тенденція видна неозброєним оком.

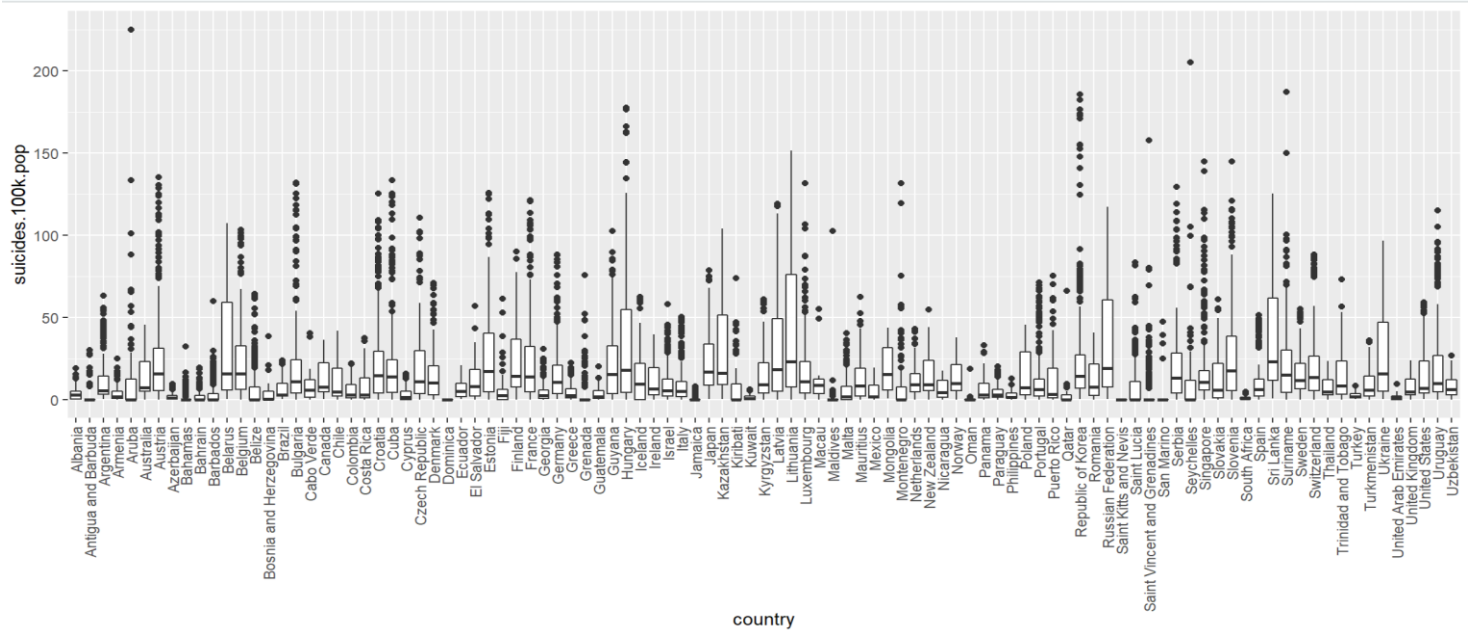
```
> fit = lm(suicides.100k.pop ~ gdp_per_capita...., data=df)
> summary(fit)

Call:
lm(formula = suicides.100k.pop ~ gdp_per_capita...., data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-13.012 -11.894  -6.827   3.802 212.152

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.279e+01  1.524e-01  83.888   <2e-16 ***
gdp_per_capita.... 1.792e-06  6.019e-06   0.298    0.766
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.96 on 27818 degrees of freedom
Multiple R-squared:  3.187e-06, Adjusted R-squared:  -3.276e-05
F-statistic: 0.08865 on 1 and 27818 DF,  p-value: 0.7659
```

На наступному графіку за допомогою боксплотів погляньмо на динаміку суіцидів за роками.



Як бачимо, цифри середнього повільно ростуть, але загалом все лишається на своїх місцях.

Смерті на 100 000 населення по всім рокам відносно різних країн.



Високими середніми показниками вирізняються Угорщина, Білорусь, росія, Литва, Шрі-Ланка, Україна, тощо. Також у всіх країн бували періоди аномального зростання цього показника.

Найвиразніші аномальні викиди бачимо на боксплотах таких країн – Республіка Корея, Словенія, Австрія, Куба, Хорватія, тощо.

Висновки

Було проведено детальний аналіз даних, вирішено проблему з пропущеними значеннями, приведенням змінних до вірних типів, статистичний і графічний аналіз всіх колонок поодинці, а також візуальний аналіз деяких найважливіших і найцікавіших комбінацій пар змінних.

Було встановлено, що в різних країнах показники самогубств дуже різняться, в деякі періоди історії були різкі «скачки» цього показнику, які вилилися в аномалії на представлених графіках.

Також було показано, що такий фундаментальний економічний показник як ВВП на душу населення, високо корелює з показником самогубств у всіх країнах – чим нижчий достаток громадян, тим вища кількість самогубств на 100 000 населення.