POLITECHNIKA
RZESZOWSKA
im. IGNACEGO ŁUKASIEWICZA

Wydział Matematyki i Fizyki Stosowanej

Inżynieria i analiza danych

Projekt z Statystyczna Analiza danych

Przegląd wskaźników samobójstw od 1985 do 2016 r.

Daniel Krzysik 166667

Rzeszów 2022

Spis treści

1. Opis użytych danych

Życie w dzisiejszych czasach jest ciężkie. Ludzie, którzy nie radzą sobie z życiem lub łatwo z niego rezygnują, mogą podjąć łatwą, ale irracjonalną decyzję o zakończeniu własnego życia. Każdy ma swoje własne problemy. Zależy to od tego, jak silni są ludzie, a wsparcie emocjonalne ze strony rodzinny i przyjaciół również odgrywa ważną rolę w zapobieganiu podejmowania takich złych decyzji. Nasz zbiór danych nie zawierał zmiennych związanych z emocjami. Jednak te czynniki również odgrywają rolę, a ludzie z pewnością podejmą decyzję w zależności od nich.

Zbiór danych dla tego projektu zawiera 13 zmiennych: kraj, rok, płeć, wiek, liczba samobójstw, populacja, samobójstwa na 100 tys. mieszkańców, kraj-rok, HDI, PKB na rok, PKB na mieszkańca oraz pokolenie. Dane są rejestrowane dla 101 krajów od 1985 do 2016 roku.

Dane nie zawierają jednak informacji na temat religii/kultury i prawa w poszczególnych krajach, które mogłyby być czynnikami wpływającymi na wskaźnik samobójstw. W wielu krajach zachowania samobójcze są potępiane przez społeczeństwo lub nawet niezgodne z prawem ze względów religijnych/kulturowych. Samobójstwo wspomagane przez lekarza może mieć wpływ na wskaźniki samobójstw w krajach, w których jest ono legalne. Biorąc pod uwagę zakres tego zbioru danych, projekt koncentruje się na ustaleniu, czy płeć, wiek lub PKB mają wpływ na wskaźnik samobójstw i jak wpływają na ten wskaźnik.

Link do pobrania danych:

https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016

Większość danych wykorzystanych w tej analizie pochodzi ze Światowej Organizacji Zdrowia.

2. Wczytanie danych

Wskaźnik rozwoju społecznego został usunięty ze zbioru danych dotyczących samobójstw, ponieważ brakuje w nim wielu danych. Również pomijamy kolumnę kraj-rok, gdyż jest ona nam niepotrzebna oraz ustaliliśmy polskie nazwy kolumn.

```
setwd("E:/STUDIA/SEMESTR IV/Statystyczna analiza danych/Projekt")
getwd()

#wczytujemy plik
(dane <- read.csv("master.csv", sep = ","))
#usuwamy kolumne country-year
(dane <- dane[,-8])
#usuwamy kolumne HDI z powodu braku danych
(dane <- dane[,-8])

#nadajemy nowe nazwy kolumna
(
  names(dane) <-
    c(
      "Kraj",
      "Rok",
      "Plec",
      "Wiek",
      "LiczbaSamobojstw",
      "Populacja",
      "SamobojstwaNa100k",
      "PKBnaROK",
      "PKBnaMieszkanca",
      "Pokolenie"
    )
)
```

```
[1] "Kraj"              "Rok"              "Plec"             "Wiek"              "LiczbaSamobojstw"
[6] "Populacja"         "SamobojstwaNa100k" "PKBnaROK"         "PKBnaMieszkanca"   "Pokolenie"
```

Kontynenty zostały dodane do zbioru danych za pomocą pakietów '**countrycode**' oraz '**tidyverse'**, następnie rozdzielono kontynent Ameryki na Amerykę Południową oraz Amerykę Północną.

```
#dzieki pakietowi countrycode przyporzadkowywuje kazdemu kraju odpowiedni kontynent
(
  dane$Kontynent <- countrycode(
    sourcevar = dane$Kraj,
    origin = "country.name",
    destination = "continent"
  )
)

#kraje Ameryki Poludniowej
(
  AmerykaPoludniowa <-
    c(
      'Argentina',
      'Brazil',
      'Chile',
      'Colombia',
      'Ecuador',
      'Guyana',
      'Paraguay',
      'Suriname',
      'Uruguay'
    )
)

#Segregacja krajow Ameryki na Ameryke Poludniowa i Ameryke Polnocna
#South America przypisujemy kazdemu krajowi ktory znalazl sie w "AmerykaPoludniowa"
(dane$Kontynent[dane$Kraj %in% AmerykaPoludniowa] <-
    'South America')
#North America przypisujemy reszte krajow
(dane$Kontynent[dane$Kontynent == 'Americas'] <- 'North America')
```

Wybrano losowy wiersz aby sprawdzić czy poprawnie wczytano kontynenty.

```
> unique(dane[, 11]) #sprawdzamy czy prawidlowo dodalismy kontynenty
[1] "Europe"        "North America" "South America" "Asia"          "Oceania"       "Africa"
> dane[6007,]
        Kraj  Rok   Plec       Wiek LiczbaSamobojstw Populacja SamobojstwaNa100k        PKBnaROK
6007 Colombia 2015 female 5-14 years              44   3904896             1.13 291,519,591,533
     PKBnaMieszkanca    Pokolenie     Kontynent
6007            6552 Generation Z South America
```

Pogrupowano wszystkie obserwacje według krajów, a następnie z uporządkowanej listy liczby obserwacji dla każdego kraju, a następnie odrzucono pierwsze 11 krajów, które miały zbyt mało danych (mniej niż 100 obserwacji).

```
(grupowanieKrajow <- count(dane,Kraj))

(sortowanieKrajow <- arrange(grupowanieKrajow, n))

#pominiecie pierwszych 11 krajów ze zbyt małą liczbą danych (mniej niż 100 obserwacji)
(dane <- dane %>%
  filter(!(Kraj %in% head(sortowanieKrajow$Kraj, 11))))
```

Usunięte kraje, które miały mniej niż 100 obserwacji.

	Kraj	n
1	Mongolia	10
2	Cabo Verde	12
3	Dominica	12
4	Macau	12
5	Bosnia and Herzegovina	24
6	Oman	36
7	Saint Kitts and Nevis	36
8	San Marino	36
9	Nicaragua	72
10	United Arab Emirates	72
11	Turkey	84

3. Podstawowe parametry

- Średnia arytmetyczna

```
> mean(dane$SamobojstwaNa100k)
[1] 12.94974
```

- Wartość minimalna

```
> min(dane$SamobojstwaNa100k)
[1] 0
```

- Wartość maksymalna

```
> max(dane$SamobojstwaNa100k)
[1] 224.97
```

- Przedział zmienności (min, max)

```
> range(dane$SamobojstwaNa100k)
[1]   0.00 224.97
```

- Rozstęp

```
> max(dane$SamobojstwaNa100k) - min(dane$SamobojstwaNa100k)
[1] 224.97
```

- Wariancja

```
> var(dane$SamobojstwaNa100k)
[1] 362.7561
```

- Moment centralny rzędu drugiego

```
> moment(dane$SamobojstwaNa100k, order = 2, central = TRUE)
[1] 362.7428
```

- Odchylenie standardowe

```
> sd(dane$SamobojstwaNa100k)
[1] 19.04616
```

- Długość

```
> length(dane$SamobojstwaNa100k)
[1] 27414
```

- Mediana

```
> median(dane$SamobojstwaNa100k)
[1] 6.11
```

- Kwantyle

```
> quantile(dane$SamobojstwaNa100k, seq(0,1, by = .1))
    0%    10%    20%    30%    40%    50%    60%    70%    80%    90%   100%
  0.00   0.00   0.43   1.68   3.65   6.11   9.23  13.76  20.73  33.50 224.97
```

- Odchylenie przeciętne od mediany

```
> mad(dane$SamobojstwaNa100k)
[1] 8.74734
```

- Rozstęp ćwiartkowy

```
> IQR(dane$SamobojstwaNa100k)
[1] 15.8875
```

- Błąd standardowy

```
> sd(dane$SamobojstwaNa100k)/sqrt(length(dane$SamobojstwaNa100k))
[1] 0.1150327
```

- Współczynnik zmienności

```
> sd(dane$SamobojstwaNa100k)/mean(dane$SamobojstwaNa100k)
[1] 1.470776
```

4. Globalna analiza - graficzne prezentacje danych

4.1 Globalna analiza

Linią przerywaną zaznaczono średni globalny wskaźnik samobójstw w latach 1985-2016: 13,27 zgonów (na 100 tys.).

```
> (srednia <- (sum(as.numeric(dane$LiczbaSamobojstw)) / sum(as.numeric(dane$Populacja))) * 100000)
[1] 13.2701
```

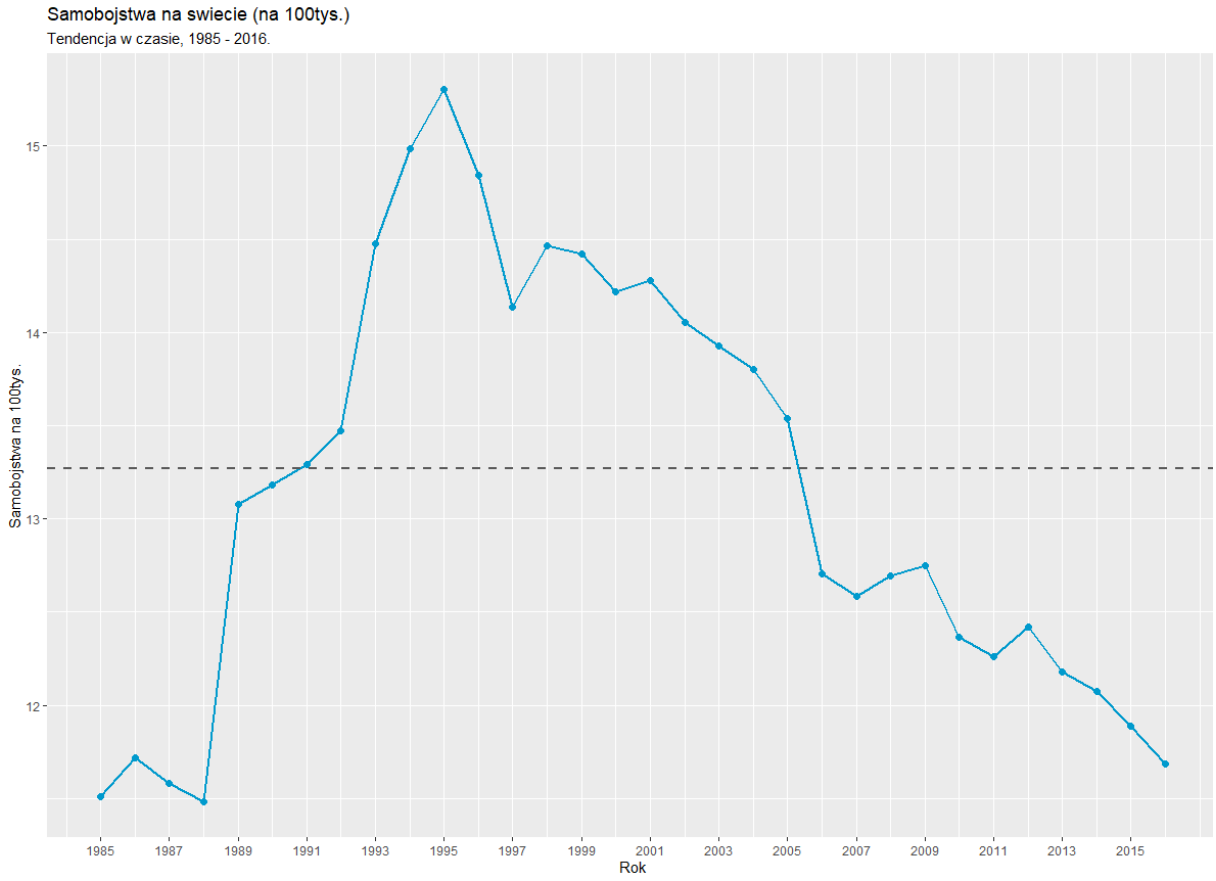
Współczynnik liczby samobójstw w poszczególnych latach, wyznaczamy dane, które są nam niezbędne do narysowania wykresu.

```
(globalna_analiza <- dane %>%
  group_by(Rok) %>%
  summarize(populacja = sum(Populacja),
            samobojstwa = sum(LiczbaSamobojstw),
            samobojstwa100k = (samobojstwa / populacja) * 100000)
)
```

```
# A tibble: 32 x 4
     Rok  populacja samobojstwa samobojstwa100k
   <int>      <int>       <int>           <dbl>
 1  1985 1008533686      116063            11.5
 2  1986 1029909613      120670            11.7
 3  1987 1095029726      126842            11.6
 4  1988 1054094424      121026            11.5
 5  1989 1225514347      160244            13.1
 6  1990 1466581000      193361            13.2
 7  1991 1489949284      198020            13.3
 8  1992 1569500347      211473            13.5
 9  1993 1530416654      221565            14.5
10  1994 1548402830      232036            15.0
# ... with 22 more rows
```

Kod do wykresu:

```r
ggplot(globalna_analiza, aes(x = Rok, y = samobojstwa100k)) +
geom_line(col = "deepskyblue3", size = 1) +
geom_point(col = "deepskyblue3", size = 2) +
geom_hline(yintercept = srednia, linetype = 2, color = "grey35", size = 1) +
labs(title = "Samobojstwa na swiecie (na 100tys.)",
     subtitle = "Tendencja w czasie, 1985 - 2016.",
     x = "Rok",
     y = "Samobojstwa na 100tys.") +
scale_x_continuous(breaks = seq(1985, 2015, 2)) +
scale_y_continuous(breaks = seq(10, 20))
```



Wnioski:

- Szczytowy wskaźnik samobójstw wyniósł 15,3 zgonów na 100 tys. mieszkańców w 1995 r.
- Stale spadał, do 11,8 na 100 tys. mieszkańców w 2015 roku (spadek o ~25%),
- Wskaźniki dopiero teraz wracają do poziomu sprzed lat 90,

4.2 Względem płci

Współczynnik liczby samobójstw w poszczególnych latach względem płci. Zaczynamy od wyznaczenia potrzebnej grupy danych

```r
(
  wykres_plci <- dane %>%
    select(Rok, Plec, LiczbaSamobojstw, Populacja) %>%
    group_by(Rok, Plec) %>%
    summarise(Lsam = round((
      sum(LiczbaSamobojstw) / sum(Populacja)
    ) * 100000, 2))
)
```

```
# A tibble: 64 x 3
# Groups:   Rok [32]
     Rok Plec    Lsam
   <int> <chr>  <dbl>
 1  1985 female  6.33
 2  1985 male   16.9
 3  1986 female  6.45
 4  1986 male   17.2
 5  1987 female  6.26
 6  1987 male   17.1
 7  1988 female  6.13
 8  1988 male   17.1
 9  1989 female  6.57
10  1989 male   20.0
# ... with 54 more rows
```

Kod do stworzenia wykresu:

```
(
  hc2 <- highchart() %>%
    hc_add_series(wykres_plci, hcaes(
      x = Rok, y = Lsam, group = Plec
    ), type = "scatter") %>% #wyswietlamy dane
    hc_tooltip(
      headerFormat = "",
      pointFormat = paste("Samobojstw: <b>{point.y}</b> <br> Rok: <b>{point.x}</b>")
    ) %>% #css
    hc_title(
      text = "<i>Samobojstwa przez plec</i>",
      style = list(color = "white", useHTML = TRUE)
    ) %>%  #tytul
    hc_subtitle(text = "1985-2015") %>% #podtytul
    hc_yAxis(title = list(text = "Samobojstwa na 100k ludzi"), #opis oxi y
             plotLines = list(list(
               color = "black", width = 1, dashStyle = "Dash",
               value = mean(srednia),
               label = list(text = "Srednia = 13.27",
               style = list(color = "black", fontSize = 11))))
             ) %>%
    hc_xAxis(title = list(text = "Rok"))
)
```

Wnioski:

- 1. Wskaźnik samobójstw wśród mężczyzn jest ~3,5x wyższy,
- 2. Zarówno wskaźnik samobójstw mężczyzn, jak i kobiet osiągnął najwyższy poziom w 1995 roku i od tego czasu spada,
- 3. Stosunek ten, wynoszący 3,5 : 1 (mężczyźni : kobiety), utrzymuje się na względnie stałym poziomie od połowy lat 90.
- 4. W latach 80. stosunek ten wynosił zaledwie 2,7 : 1 (mężczyzna : kobieta)

4.3 Względem wieku

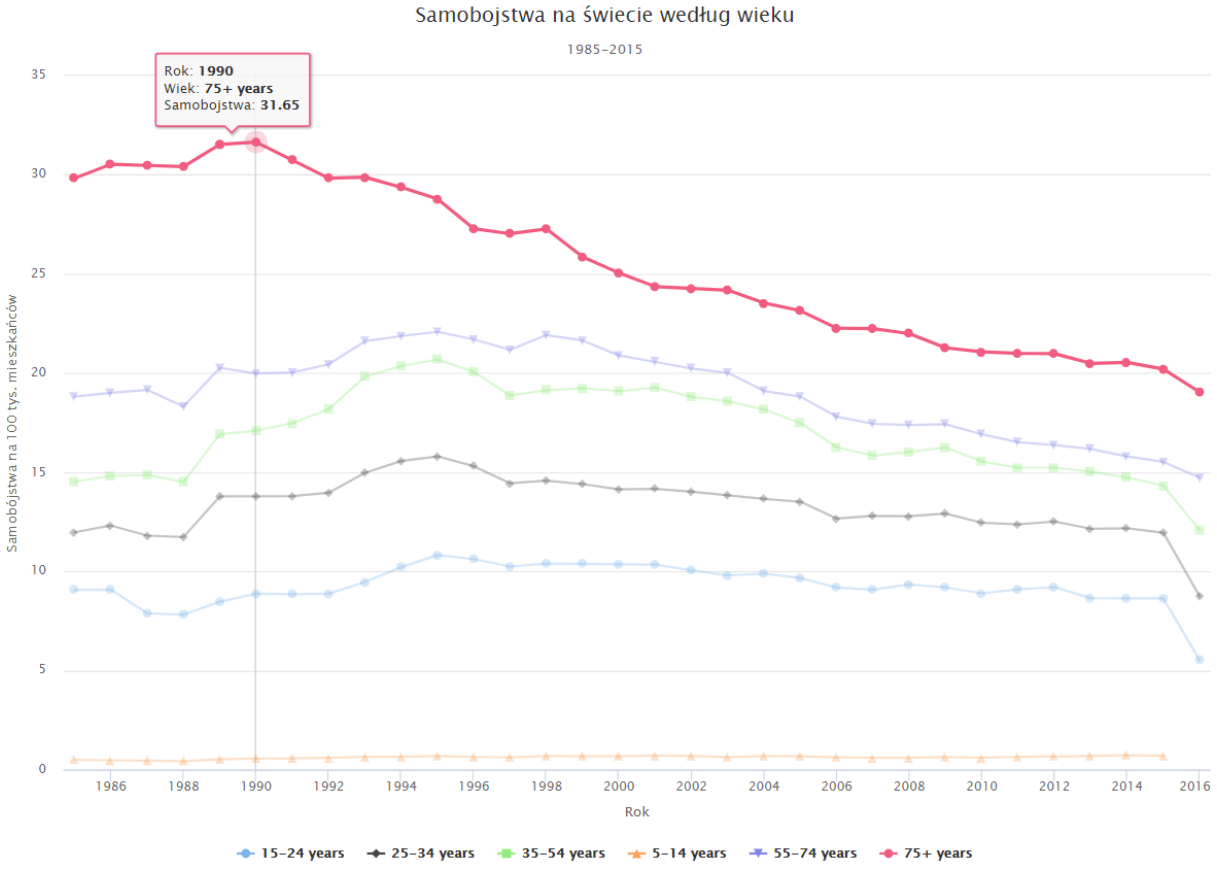
Współczynnik liczby samobójstw według wieku. Zaczynamy od wyznaczenia potrzebnej grupy danych

```
(
    wykres_wiek <- dane %>%
    select(Rok, Wiek, LiczbaSamobojstw, Populacja) %>%
    group_by(Rok, Wiek) %>%
    summarise(Lsam = round((sum(LiczbaSamobojstw) / sum(Populacja)) * 100000, 2))
)
```

```
# A tibble: 191 x 3
# Groups:   Rok [32]
     Rok Wiek          Lsam
   <int> <chr>        <dbl>
 1  1985 15-24 years  9.07
 2  1985 25-34 years 12.0
 3  1985 35-54 years 14.5
 4  1985 5-14 years   0.49
 5  1985 55-74 years 18.8
 6  1985 75+ years   29.8
 7  1986 15-24 years  9.07
 8  1986 25-34 years 12.3
 9  1986 35-54 years 14.8
10  1986 5-14 years   0.47
# ... with 181 more rows
```

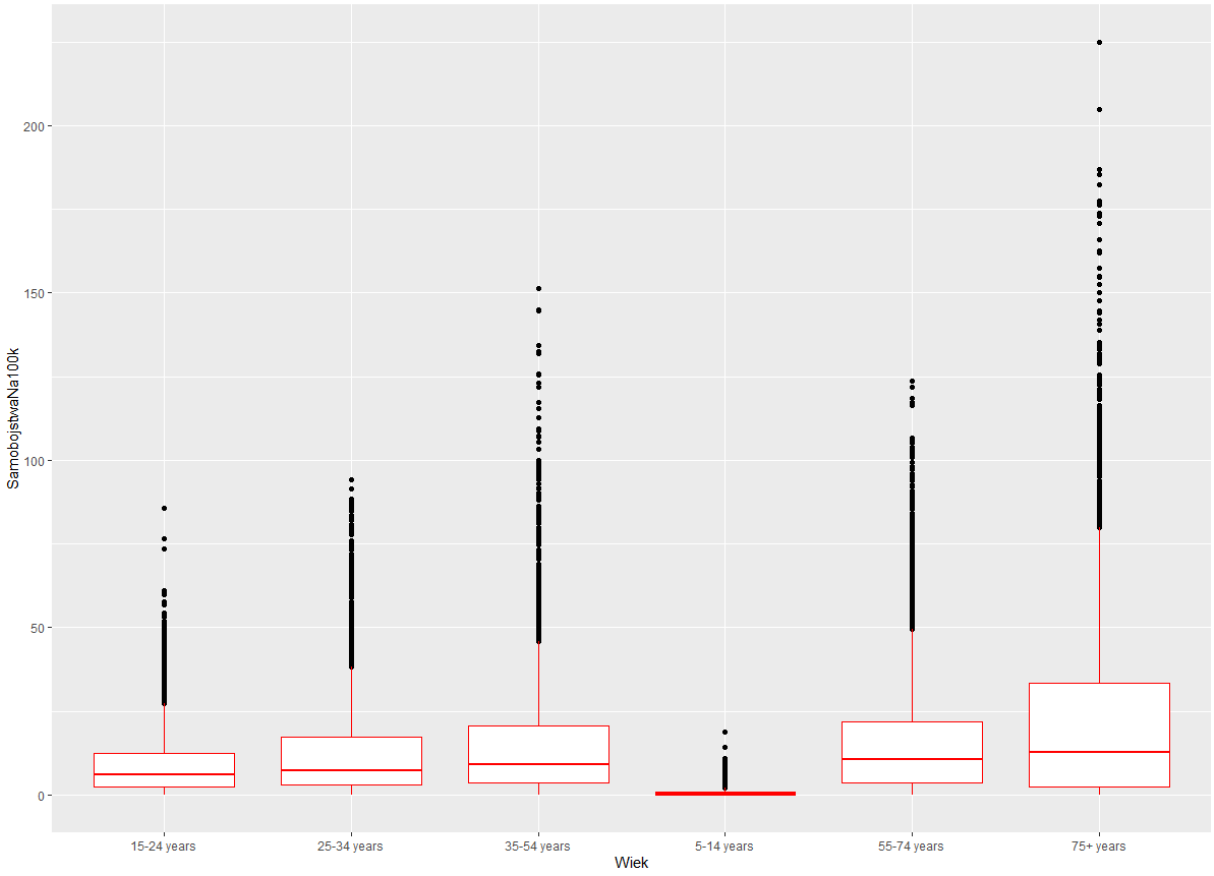
Kod do stworzenia wykresu:

```
(
  highchart() %>%
  hc_add_series(wykres_wiek, hcaes(x = Rok, y = Lsam, group = Wiek), type = "line") %>%
  hc_tooltip(crosshairs = TRUE, borderWidth = 1.5, headerFormat = "",
             pointFormat = paste("Rok: <b>{point.x}</b> <br>","Wiek: <b>{point.Wiek}</b><br>",
                                 "Samobojstwa: <b>{point.y}</b>")) %>%
  hc_title(text = "Samobojstwa na świecie według wieku") %>%
  hc_subtitle(text = "1985-2015") %>%
  hc_xAxis(title = list(text = "Rok")) %>%
  hc_yAxis(title = list(text = "Samobójstwa na 100 tys. mieszkańców"))
)
```



Dodatkowy wykres:

```
(
  ggplot(dane, aes(x= Wiek ,y=SamobojstwaNa100k)) +
  geom_boxplot(color="red" , outlier.color="black")
)
```



Z wykresu pudełkowego widać, że lewy bok jest wyznaczony przez pierwszy kwartyl, zaś prawy bok przez trzeci kwartyl. Szerokość pudełka odpowiada wartości rozstępu ćwiartkowego. Wewnątrz pudełka znajduje się pionowa linia, określająca wartość mediany. Rysunek uzupełniamy po prawej i lewej stronie odcinkami. Lewy koniec odcinka wyznacza najmniejsza wartość w zbiorze, natomiast prawy koniec prawego odcinka to wartość największa.

Wnioski:

- W skali globalnej prawdopodobieństwo popełnienia samobójstwa wzrasta wraz z wiekiem
- 2. Od 1995 roku wskaźnik samobójstw wśród osób w wieku >= 15 lat maleje liniowo.
- 3. Wskaźnik samobójstw w kategorii "5-14" pozostaje mniej więcej statyczny i niewielki (< 1 na 100 tys. rocznie).

4.4 Względem kontynentu

Zaczynamy od utworzenia grupy danych:

```
(g5 <- dane %>%
    group_by(Kontynent) %>%
    summarise(LS = round((sum(LiczbaSamobojstw) / sum(Populacja)) * 100000, 2)) %>%
    arrange(LS))
```

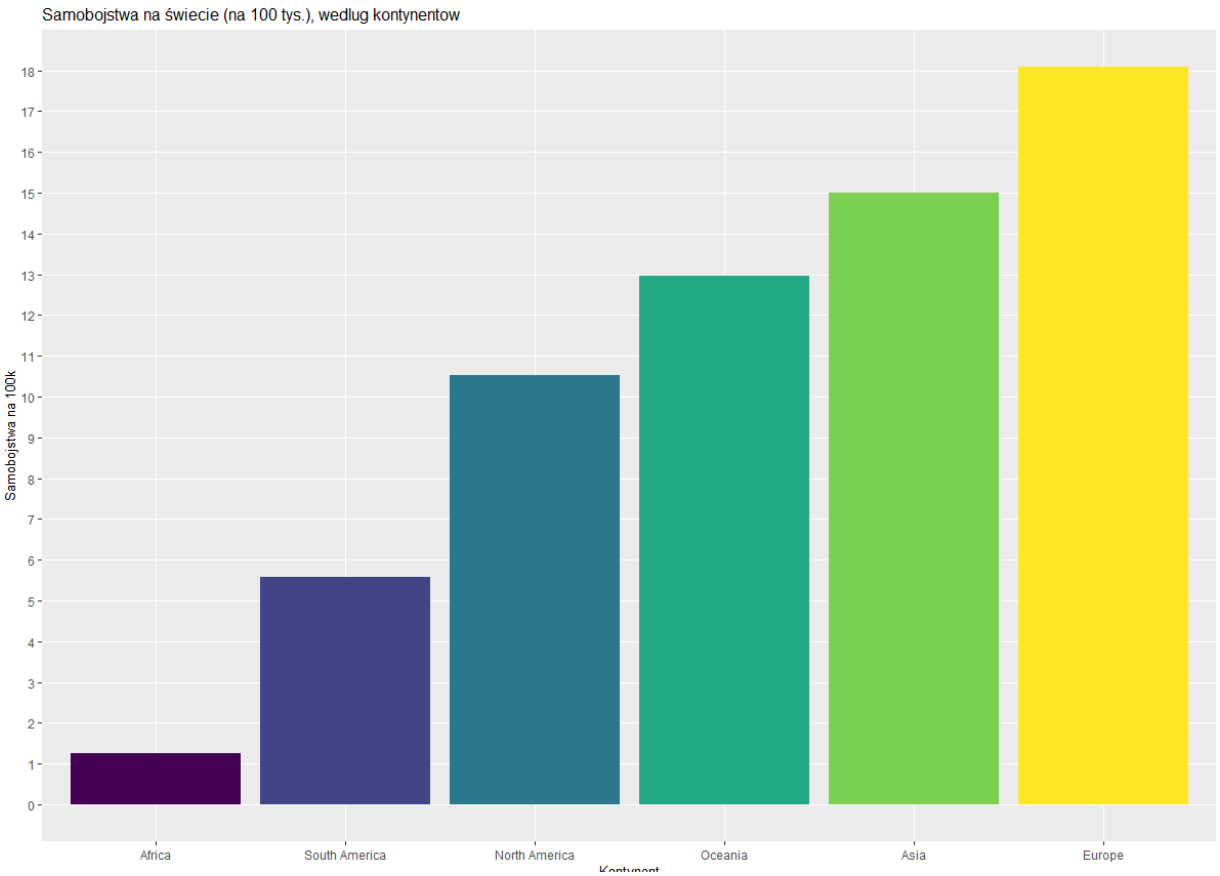
```
# A tibble: 6 x 2
  Kontynent        LS
  <chr>         <dbl>
1 Africa         1.25
2 South America  5.58
3 North America 10.5
4 Oceania       13.0
5 Asia          15.0
6 Europe        18.1
```

Chcemy aby na naszym wykresie kontynenty były posortowane od najniższego do najwyższego współczynnika samobójstw.

```
> (g5$Kontynent <- factor(g5$Kontynent, ordered = T, levels = g5$Kontynent)) #sortowaie
[1] Africa        South America North America Oceania       Asia          Europe
Levels: Africa < South America < North America < Oceania < Asia < Europe
```

Kod wykresu:

```
(
  Kontynent_plot <- ggplot(g5, aes(x = Kontynent, y = LS, fill = Kontynent)) +
  geom_bar(stat = "identity") +
  labs(title = "Samobojstwa na świecie (na 100 tys.), wedlug kontynentow",
       x = "Kontynent",
       y = "Samobojstwa na 100k",
       fill = "Kontynent") +
  theme(legend.position = "none", title = element_text(size = 10)) +
  scale_y_continuous(breaks = seq(0, 20, 1), minor_breaks = F)
)
```



Utwórzmy jeszcze pomocniczy wykres kontynenty na przestrzeni lat. Zacznijmy od wyznaczenia grupy danych:

```
(g5a <- dane %>%
  group_by(Rok, Kontynent) %>%
  summarise(LS = round((sum(LiczbaSamobojstw) / sum(Populacja)) * 100000, 2))
)
```

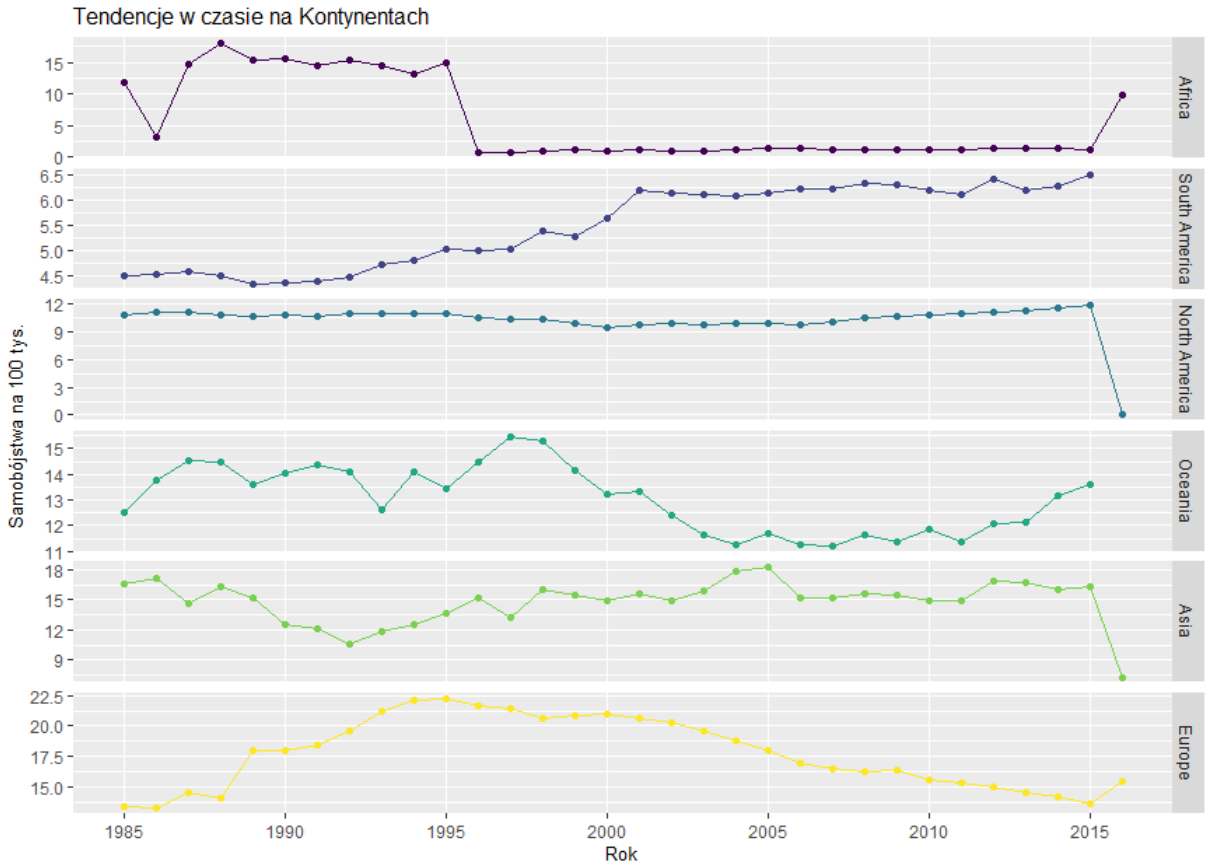
```
# A tibble: 190 x 3
# Groups:   Rok [32]
     Rok Kontynent        LS
   <int> <chr>         <dbl>
 1  1985 Africa         11.9
 2  1985 Asia           16.6
 3  1985 Europe         13.3
 4  1985 North America  10.8
 5  1985 Oceania        12.5
 6  1985 South America   4.49
 7  1986 Africa          3.04
 8  1986 Asia           17.1
 9  1986 Europe         13.2
10  1986 North America  11.1
# ... with 180 more rows
```

Ponownie posortujemy nasze dane.

```
> (g5a$Kontynent <- factor(g5a$Kontynent, ordered = T, levels = g5$Kontynent)) #sortowaie
  [1] Africa        Asia          Europe        North America Oceania       South America Africa
  [8] Asia          Europe        North America Oceania       South America Africa        Asia
 [15] Europe        North America Oceania       South America Africa        Asia          Europe
 [22] North America Oceania       South America Africa        Asia          Europe        North America
 [29] Oceania       South America Africa        Asia          Europe        North America Oceania
 [36] South America Africa        Asia          Europe        North America Oceania       South America
 [43] Africa        Asia          Europe        North America Oceania       South America Africa
 [50] Asia          Europe        North America Oceania       South America Africa        Asia
 [57] Europe        North America Oceania       South America Africa        Asia          Europe
 [64] North America Oceania       South America Africa        Asia          Europe        North America
 [71] Oceania       South America Africa        Asia          Europe        North America Oceania
 [78] South America Africa        Asia          Europe        North America Oceania       South America
 [85] Africa        Asia          Europe        North America Oceania       South America Africa
 [92] Asia          Europe        North America Oceania       South America Africa        Asia
 [99] Europe        North America Oceania       South America Africa        Asia          Europe
[106] North America Oceania       South America Africa        Asia          Europe        North America
[113] Oceania       South America Africa        Asia          Europe        North America Oceania
[120] South America Africa        Asia          Europe        North America Oceania       South America
[127] Africa        Asia          Europe        North America Oceania       South America Africa
[134] Asia          Europe        North America Oceania       South America Africa        Asia
[141] Europe        North America Oceania       South America Africa        Asia          Europe
[148] North America Oceania       South America Africa        Asia          Europe        North America
[155] Oceania       South America Africa        Asia          Europe        North America Oceania
[162] South America Africa        Asia          Europe        North America Oceania       South America
[169] Africa        Asia          Europe        North America Oceania       South America Africa
[176] Asia          Europe        North America Oceania       South America Africa        Asia
[183] Europe        North America Oceania       South America Africa        Asia          Europe
[190] North America
Levels: Africa < South America < North America < Oceania < Asia < Europe
```

Kod do stworzenia wykresu:

```r
(Kontynent_czas_plot <- ggplot(g5a, aes(x = Rok, y = LS, col = factor(Kontynent))) +
  facet_grid(Kontynent ~ ., scales = "free_y") +
  geom_line() +
  geom_point() +
  labs(title = "Tendencje w czasie na Kontynentach",
       x = "Rok",
       y = "Samobójstwa na 100 tys.",
       color = "Kontynent") +
  theme(legend.position = "none", title = element_text(size = 10)) +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F))
grid.arrange(Kontynent_plot, Kontynent_czas_plot, ncol = 2)
```



Wnioski:

- 1. Wskaźnik europejski jest ogólnie najwyższy, ale od 1995 r. stale spada o ~40%.
- 2. Wskaźnik europejski za rok 2015 jest podobny do wskaźnika dla Azji i Oceanii.
- 3. Linia trendu dla Afryki jest spowodowana niską jakością danych - tylko 3 kraje dostarczyły dane.

4.5 Według państwa

Współczynnik samobójstw na 100tys. mieszkańców w poszczególnych krajach. Zacznijmy od wyznaczenia grupy danych.

```r
(
  g7 <- dane %>%
    group_by(Kraj, Kontynent) %>%
    summarise(LS = round((sum(LiczbaSamobojstw) / sum(Populacja)) * 100000, 2)) %>%
    arrange(desc(LS))
)
```

```
# A tibble: 90 x 3
# Groups:   Kraj [90]
   Kraj                Kontynent    LS
   <chr>               <chr>     <dbl>
 1 Lithuania           Europe     41.2
 2 Russian Federation  Europe     32.8
 3 Sri Lanka           Asia       30.5
 4 Belarus             Europe     30.3
 5 Hungary             Europe     29.7
 6 Latvia              Europe     28.5
 7 Kazakhstan          Asia       26.9
 8 Slovenia            Europe     26.4
 9 Estonia             Europe     26.0
10 Ukraine             Europe     24.9
# ... with 80 more rows
```

Chcemy aby nasz wykres był posegregowany od najniższego współczynnika do najwyższego:

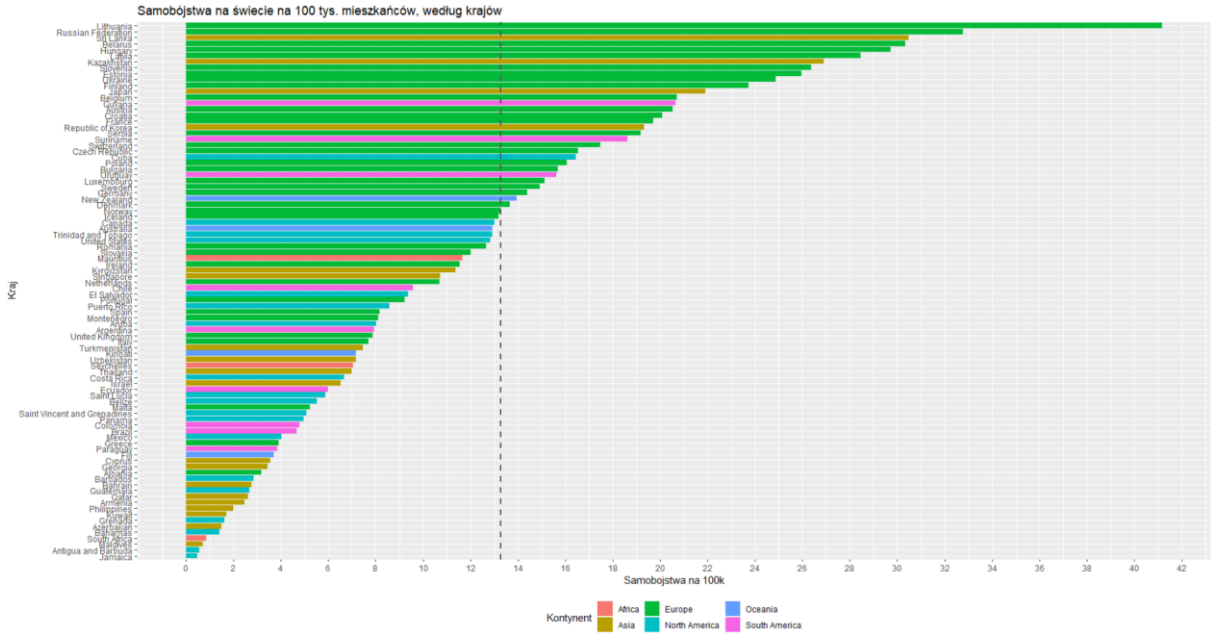
```
> (g7$Kraj <- factor(g7$Kraj, ordered = T, levels = rev(g7$Kraj)))
 [1] Lithuania                Russian Federation       Sri Lanka
 [4] Belarus                  Hungary                  Latvia
 [7] Kazakhstan               Slovenia                 Estonia
[10] Ukraine                  Finland                  Japan
[13] Belgium                  Guyana                   Austria
[16] Croatia                  France                   Republic of Korea
[19] Serbia                   Suriname                 Switzerland
[22] Czech Republic           Cuba                     Poland
[25] Bulgaria                 Uruguay                  Luxembourg
[28] Sweden                   Germany                  New Zealand
[31] Denmark                  Norway                   Iceland
[34] Canada                   Australia                Trinidad and Tobago
[37] United States            Romania                  Slovakia
[40] Mauritius                Ireland                  Kyrgyzstan
[43] Singapore                Netherlands              Chile
[46] El Salvador              Portugal                 Puerto Rico
[49] Spain                    Montenegro               Aruba
[52] Argentina                United Kingdom           Italy
[55] Turkmenistan             Kiribati                 Uzbekistan
[58] Seychelles               Thailand                 Costa Rica
[61] Israel                   Ecuador                  Saint Lucia
[64] Belize                   Malta                    Saint Vincent and Grenadines
[67] Panama                   Colombia                 Brazil
[70] Mexico                   Greece                   Paraguay
[73] Fiji                     Cyprus                   Georgia
[76] Albania                  Barbados                 Bahrain
[79] Guatemala                Qatar                    Armenia
[82] Philippines              Kuwait                   Grenada
[85] Azerbaijan               Bahamas                  South Africa
[88] Maldives                 Antigua and Barbuda      Jamaica
90 Levels: Jamaica < Antigua and Barbuda < Maldives < South Africa < Bahamas < Azerbaijan < ... < Lithuania
```

Również chcemy na naszym wykresie umieścić średni globalny wskaźnik samobójstw w latach 1985-2016: 13,27 zgonów (na 100 tys.).

```
> (srednia <- (sum(as.numeric(dane$LiczbaSamobojstw)) / sum(as.numeric(dane$Populacja))) * 100000)
[1] 13.2701
```

Kod do stworzenia wykresu:

```
(
  ggplot(g7, aes(x = Kraj, y = LS, fill = Kontynent)) +
  geom_bar(stat = "identity") +
  geom_hline(yintercept = srednia, linetype = 2, color = "grey35", size = 1) +
  labs(title = "Samobójstwa na świecie na 100 tys. mieszkańców, według krajów",
       x = "Kraj",
       y = "Samobojstwa na 100k",
       fill = "Kontynent") +
  coord_flip() +
  scale_y_continuous(breaks = seq(0, 45, 2)) +
  theme(legend.position = "bottom")
)
```



Wnioski:

- 1. Na Litwie wskaźnik ten jest zdecydowanie najwyższy: > 41 samobójstw na 100 tys. mieszkańców,
- 2. W Polsce wskaźnik ten wynosi >16, ponad średnią,
- 3. Znaczna nadreprezentacja krajów europejskich o wysokich wskaźnikach, niewiele o niskich.

4.6 Czy wraz z bogaceniem się kraju spada liczba samobójstw?

To zależy od kraju - w prawie każdym kraju istnieje wysoka korelacja między Rok, a PKBnaMieszkanca tzn. w miarę upływu czasu PKB rośnie liniowo.

```
(KrajRokPKB <- dane %>%
    group_by(Kraj, Rok) %>%
    summarize(PKBnaMieszkanca = mean(PKBnaMieszkanca)))
```

```
# Groups:   Kraj [90]
   Kraj       Rok PKBnaMieszkanca
   <chr>    <int>           <dbl>
 1 Albania   1987             796
 2 Albania   1988             769
 3 Albania   1989             833
 4 Albania   1992             251
 5 Albania   1993             437
 6 Albania   1994             697
 7 Albania   1995             835
 8 Albania   1996            1127
 9 Albania   1997             793
10 Albania   1998             899
# ... with 2,277 more rows
```

```
(KrajRokPKBKorelacja <- KrajRokPKB %>%
    ungroup() %>%
    group_by(Kraj) %>%
    summarize(RokPKBKorelacja = cor(Rok, PKBnaMieszkanca)))
```

```
# A tibble: 90 x 2
   Kraj                 RokPKBKorelacja
   <chr>                          <dbl>
 1 Albania                        0.882
 2 Antigua and Barbuda            0.944
 3 Argentina                      0.738
 4 Armenia                        0.919
 5 Aruba                          0.914
 6 Australia                      0.905
 7 Austria                        0.938
 8 Azerbaijan                     0.427
 9 Bahamas                        0.843
10 Bahrain                        0.928
# ... with 80 more rows
```

```
> (mean(as.numeric(KrajRokPKBKorelacja$RokPKBKorelacja)))
[1] 0.8926335
```

Obliczyłem korelacje Pearsona między "rokiem" a "PKBnaMieszkanca" w każdym kraju, a następnie podsumowałem wyniki:

Średnia korelacja wyniosła 0.8926335, co oznacza, że są one wysoce dodatnio skorelowane.

W większości krajów wraz ze wzrostem PKB wzrasta również liczba samobójstw. Jednak zdarzają się wyjątki.

4.7 Czy kraje bogatsze mają wyższy wskaźnik samobójstw?

Zamiast przyglądać się trendom w poszczególnych krajach, biorę każdy kraj i obliczam jego średni PKB (na mieszkańca) we wszystkich latach, dla których dostępne są dane. następnie mierze, jak to się ma do współczynnika samobójstw w danych kraju we wszystkich latach.

```
(KrajSredniePKB <- dane %>%
  group_by(Kraj, Kontynent) %>%
  summarize(LS = (sum(as.numeric(LiczbaSamobojstw)) / sum(as.numeric(Populacja))) * 100000,
            PKBnaMieszkanca = mean(PKBnaMieszkanca)))
```

Funkcją służącą do budowy modelu liniowego w R jest funkcja lm

```
model1 <- lm(LS ~ PKBnaMieszkanca, data = KrajSredniePKB)
```

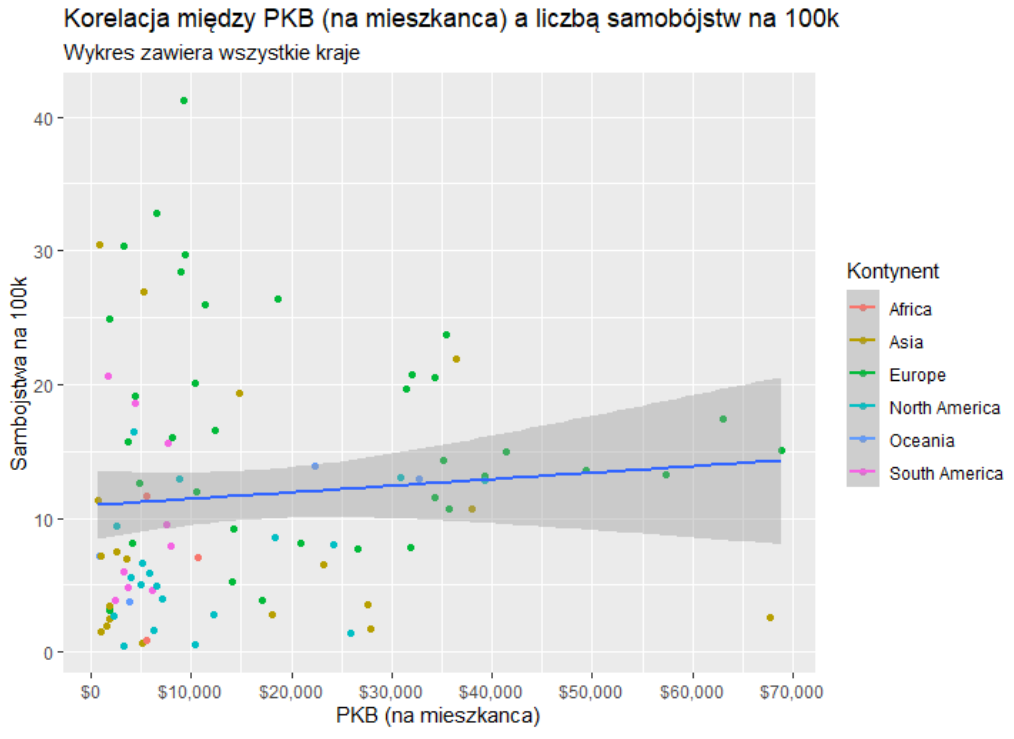
Aby zobaczyć wszystkie parametry tak powstałego modelu należy wykorzystać funkcję summary():

W wyniku której uzyskamy następujące podsumowanie parametrów modelu:

```
> summary(model1)

Call:
lm(formula = LS ~ PKBnaMieszkanca, data = KrajSredniePKB)

Residuals:
    Min      1Q  Median      3Q     Max
-11.633  -6.400  -2.049   4.678  29.783

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.095e+01  1.301e+00   8.412 6.66e-13 ***
PKBnaMieszkanca 4.889e-05  5.709e-05   0.856    0.394
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.794 on 88 degrees of freedom
Multiple R-squared:  0.008264,  Adjusted R-squared:  -0.003005
F-statistic: 0.7333 on 1 and 88 DF,  p-value: 0.3941
```

Zaznaczono na niebiesko otrzymaliśmy wartości dla współczynników równania liniowego, na zielono błąd standardowy, pomarańczowo zaznaczona jest wartość statystyki t oraz na czerwono wynik testu istotności dla danego współczynnika.

```
(ggplot(KrajSredniePKB, aes(x = PKBnaMieszkanca, y = LS, col = Kontynent)) +
  geom_point() +
  geom_smooth(method = "lm", aes(group = 1)) +
  scale_x_continuous(labels=scales::dollar_format(prefix="$"), breaks = seq(0, 70000, 10000)) +
  labs(title = "Korelacja między PKB (na mieszkanca) a liczbą samobójstw na 100k",
       subtitle = "Wykres zawiera wszystkie kraje",
       x = "PKB (na mieszkanca)",
       y = "Sambojstwa na 100k",
       col = "Kontynent") )
```



Istnieje słaba, ale istotna dodatnia zależność liniowa - bogatsze kraje wiążą się z wyższymi wskaźnikami samobójstw, ale jest to zależność słaba, co widać na powyższym wykresie.

5. Porównanie Wielkiej Brytanii, Irlandii, Ameryki, Francji, Danii oraz Polski.

Warto byłoby porównać kilka krajów, które ludzie mogą uważać za podobne do Wielkiej Brytanii (pod względem kulturowym, prawnym, ekonomicznym).

5.1 Ogólne porównanie

Wybór krajów, których użyjemy w poniższych testach.

```
(zestawDanych <- dane %>%
  filter(Kraj %in% c("United Kingdom", "Ireland","United States", "France", "Denmark", "Poland")))
```

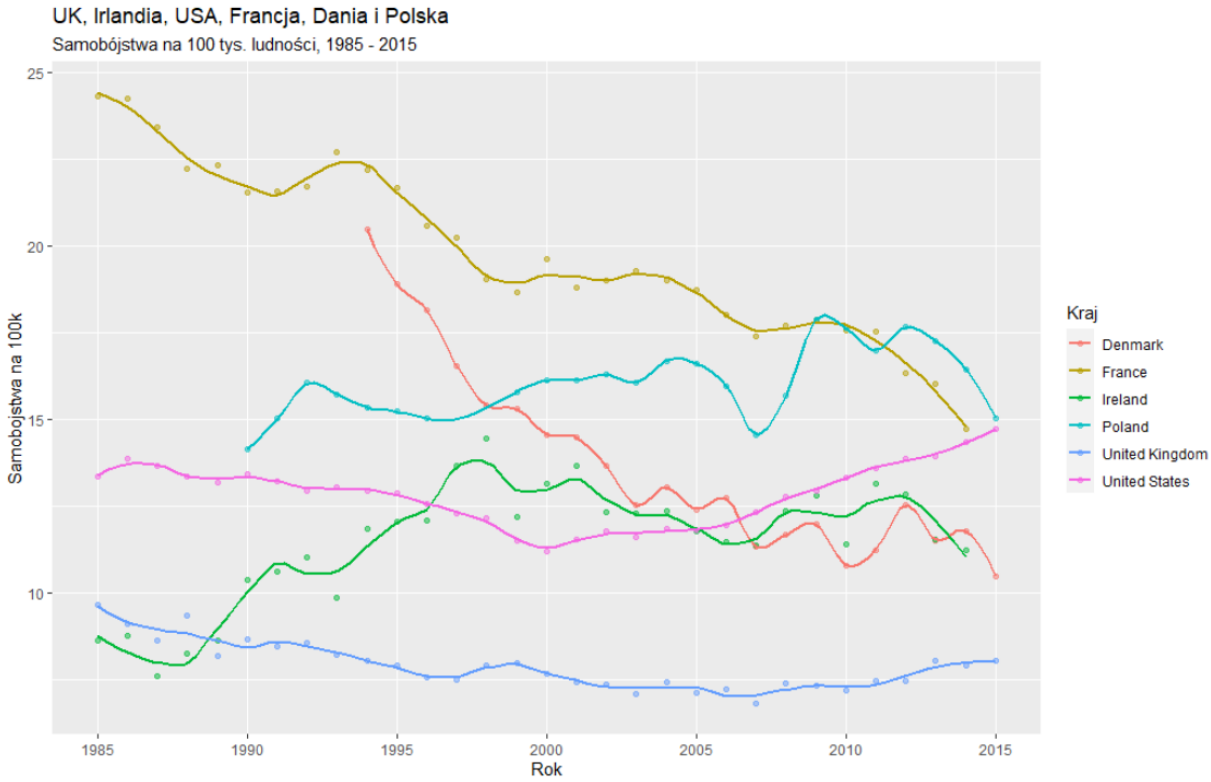
Grupowanie danych:

```
(g10 <- zestawDanych %>%
  group_by(Kraj, Rok) %>%
  summarize(LS = (sum(as.numeric(LiczbaSamobojstw)) / sum(as.numeric(Populacja))) * 100000))
```

```
# Groups:   Kraj [6]
   Kraj      Rok    LS
   <chr>   <int> <dbl>
 1 Denmark  1994  20.5
 2 Denmark  1995  18.9
 3 Denmark  1996  18.1
 4 Denmark  1997  16.5
 5 Denmark  1998  15.4
 6 Denmark  1999  15.3
 7 Denmark  2000  14.6
 8 Denmark  2001  14.5
 9 Denmark  2002  13.6
10 Denmark  2003  12.5
# ... with 158 more rows
```

Kod do tworzenia wykresu:

```
(ggplot(g10, aes(x = Rok, y = LS, col = Kraj)) +
  geom_point(alpha = 0.5) +
  geom_smooth(se = F, span = 0.2) +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F) +
  labs(title = "UK, Irlandia, USA, Francja, Dania i Polska",
       subtitle = "Samobójstwa na 100 tys. ludności, 1985 - 2015",
       x = "Rok",
       y = "Samobojstwa na 100k",
       col = "Kraj"))
```



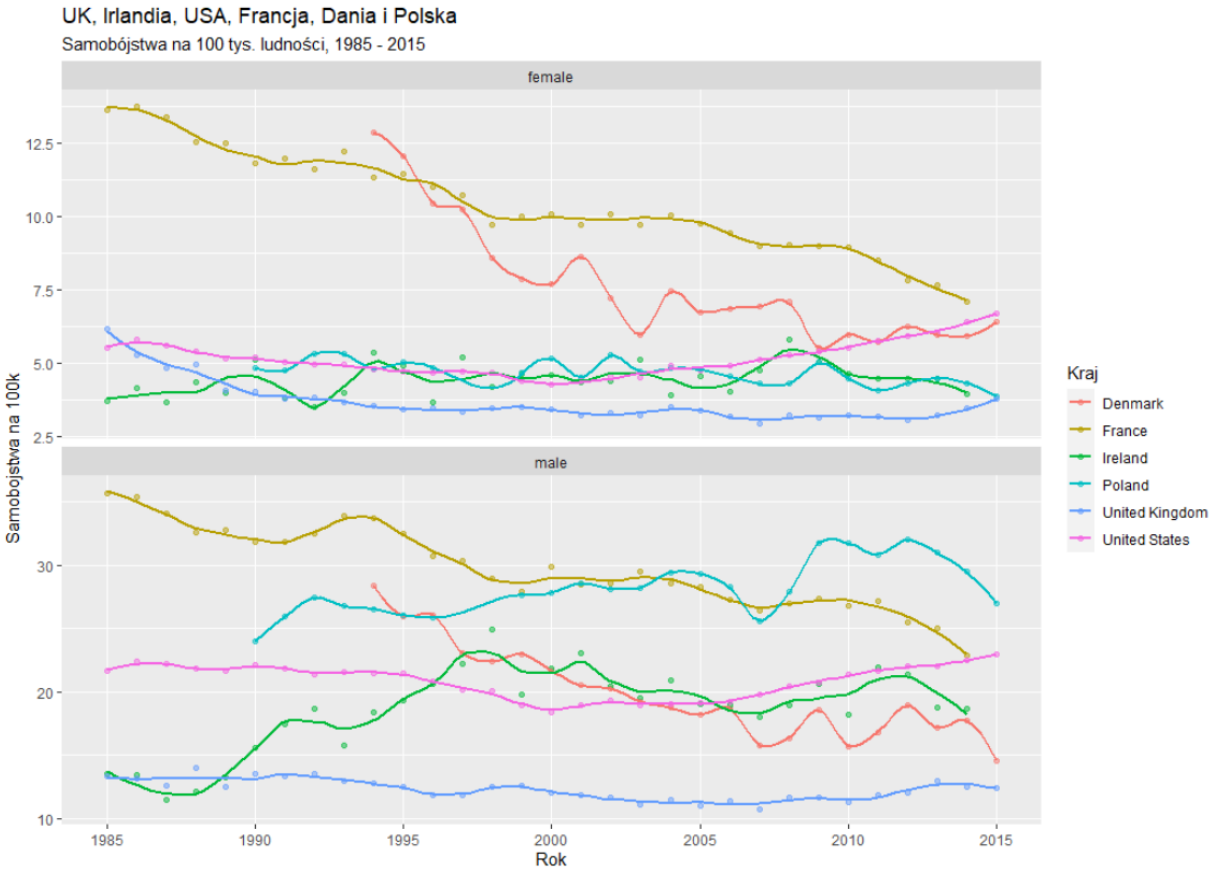
Wnioski:

- 1. Wskaźnik samobójstw w Wielkiej Brytanii jest niezmiennie najniższy od 1990 roku, a od około 1995 roku utrzymuje się na dość statycznym poziomie.
- 2. Francja miała najwyższy wskaźnik, ale obecnie jest on mniej więcej równy amerykańskiemu i polskiemu
- 3. Stany Zjednoczone wykazują najbardziej niepokojącą tendencję - od 2000 roku wskaźnik wzrósł liniowo o ~1/3.

5.2 Względem płci na przestrzeni lat

Kod do stworzenia wykresu:

```
(zestawDanych %>%
  group_by(Kraj, Plec, Rok) %>%
  summarize(LS = (sum(as.numeric(LiczbaSamobojstw)) / sum(as.numeric(Populacja))) * 100000) %>%
  ggplot(aes(x = Rok, y = LS, col = Kraj)) +
  geom_point(alpha = 0.5) +
  geom_smooth(se = F, span = 0.2) +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F) +
  facet_wrap(~ Plec, scales = "free_y", nrow = 2) +
  labs(title = "UK, Irlandia, USA, Francja, Dania i Polska",
       subtitle = "Samobójstwa na 100 tys. ludności, 1985 - 2015",
       x = "Rok",
       y = "Samobojstwa na 100k",
       col = "Kraj"))
```



Wnioski:

- 1. Odmienne linie trendu dla mężczyzn i kobiet w Irlandii oraz Polsce - w 1990 roku wzrasta wskaźnik dla mężczyzn, ale nie można tego samego zaobserwować dla kobiet
- 2. W przypadku mężczyzn i kobiet we Francji odnotowano spadek wskaźnika do poziomu zbliżonego do amerykańskiego.

6. Hipotezy

6.1 Testowanie różnicy średnich między samobójstwami mężczyzn i kobiet/100 tys. mieszkańców

Zweryfikujemy hipotezę, że średnia liczba samobójstw kobiet jest równa średniej liczbie samobójstw wśród mężczyzn.

Hipoteza zerowa: H_0: $\mu = \mu_0$

Hipoteza alternatywna: H_1: $\mu \neq \mu_0$

Kod:

Pobieramy dane potrzebne do wykonania hipotezy.

```
(
  samobojstwa_plec <- dane %>%
  group_by(Plec,Rok) %>%
  summarize(LS = (sum(as.numeric(LiczbaSamobojstw)) / sum(as.numeric(Populacja))) * 100000)
)
```

Współczynnik samobójstw względem kobiet.

```
(
  samobojstwa_kobiety <- samobojstwa_plec %>%
  filter(Plec=='female')
)
```

Współczynnik samobójstw względem mężczyzn.

```
(
  samobojstwa_mezczyzni <- samobojstwa_plec %>%
  filter(Plec=='male')
)
```

Przeprowadzamy test.

```
(test <- t.test(x = samobojstwa_mezczyzni$LS, y = samobojstwa_kobiety$LS))
```

Uzyskane wyniki:

```
        Welch Two Sample t-test

data:  samobojstwa_mezczyzni$LS and samobojstwa_kobiety$LS
t = 37.551, df = 33.421, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 13.88859 15.47897
sample estimates:
mean of x mean of y
20.689090  6.005311
```

Wartość p w teście hipotezy wskazuje istotność wyników, przy czym niskie wartości oznaczają odrzucenie hipotezy zerowej, a wysokie wartości - brak odrzucenia hipotezy zerowej. Wnioskujemy stąd, że możemy odrzucić hipotezę o równości średniej liczby samobójstw. Na 95% średnia liczby samobójstw wśród mężczyzn jest większa od średniej liczby samobójstw wśród kobiet, różnica na 95% jest w przedziale (13.89; 15,48).

Odrzucamy hipotezę zerową, gdy p-wartość $\leq \alpha$. W naszym przypadku p-value ≤ 0.05, hipoteza odrzucona.

Istnieje istotna różnica między średnimi wartością samobójstw mężczyzn na 100tys. mieszkańców i samobójstw kobiet 100 tys. mieszkańców. Średnia samobójstw mężczyzn 100tys. ludności jest istotnie większa niż samobójstw kobiet.

Drugie założenie mówi, że wariancja między dwiema grupami jest taka sama. Można to sprawdzić za pomocą testu F:

```
> var.test(LS ~ Plec, data = samobojstwa_plec)

        F test to compare two variances

data:  LS by Plec
F = 0.039103, num df = 31, denom df = 31, p-value = 1.511e-14
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.01908797 0.08010625
sample estimates:
ratio of variances
        0.03910327
```

Wartość p jest mniejsza niż 0.05, zatem wariancja danych dotyczących kobiet i mężczyzn jest istotnie różna od siebie, co oznacza, że to założenie również nie zostało spełnione.

6.2 Zależność między wiekiem, a liczbą samobójstw na 100tys. mieszkańców.

Zweryfikujemy hipotezę, że wraz z wiekiem wzrasta liczba samobójstw na 100tys. mieszkańców.

Hipoteza zerowa: H_0: $\mu = \mu_0$

Hipoteza alternatywna: H_1: $\mu \neq \mu_0$

Kod:

```
(
  KrajWiek <- dane %>%
    group_by(Rok,Kraj,Wiek) %>%
    summarize(LS = (sum(as.numeric(LiczbaSamobojstw)) / sum(as.numeric(Populacja))) * 100000)
)

(modelWiek <- lm(LS~Wiek,data = KrajWiek))

summary(modelWiek)
```

Otrzymujemy wynik:

```
Call:
lm(formula = LS ~ Wiek, data = KrajWiek)

Residuals:
    Min      1Q  Median      3Q     Max
-20.185  -6.986  -0.627   3.233 100.394

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.1124     0.2441  37.329   <2e-16 ***
Wiek25-34 years   3.2235     0.3452   9.337   <2e-16 ***
Wiek35-54 years   5.8684     0.3452  16.999   <2e-16 ***
Wiek5-14 years   -8.4852     0.3458 -24.539   <2e-16 ***
Wiek55-74 years   6.3344     0.3452  18.349   <2e-16 ***
Wiek75+ years    11.0726     0.3452  32.074   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.67 on 13701 degrees of freedom
Multiple R-squared:  0.2158,    Adjusted R-squared:  0.2155
F-statistic: 753.9 on 5 and 13701 DF,  p-value: < 2.2e-16
```

Wartość p-value $\leq$ 0.05, zatem odrzucamy hipotezę zerową. Istnieje dodatnia zależność między wiekiem, a liczbą samobójstw, wraz ze wzrostem wieku wzrasta liczba samobójstw na 100tys. mieszkańców.

6.3 PKB na mieszkańca względem liczby samobójstw.

Zbadam wpływ średniego Produktu Krajowego Brutto (PKB) na liczbę samobójstw. PKB (mierzony w dolarach amerykańskich) jest miarą produktywności gospodarczej danego kraju. Jest to wartość wszystkich dóbr i usług wytworzonych przez dany kraj w określonym czasie. PKB na mieszkańca to po prostu PKB podzielony przez liczbę ludności i jest miarą tego, jak dużą wartość produkcji gospodarczej można przypisać przeciętnemu obywatelowi. Oznacza to, że średni PKB dla kobiety i mężczyzny będzie taki sam, ale jak widzieliśmy ich współczynnik samobójstw jest różny, więc użyteczne będzie

przeanalizowanie, czy zmiana średniego PKB na mieszkańca może lepiej przewidzieć współczynnik samobójstw dla kobiety niż dla mężczyzny w różnych krajach.

Potrzebne dane z których skorzystamy:

```
(
  plecPKB <- dane %>%
  group_by(Kraj, Plec) %>%
  summarise(CLS = mean(sum(as.numeric(LiczbaSamobojstw))),
                    PKB = mean(as.numeric(PKBnaMieszkanca)))
)
```

Za pomocą prostej linii regresji ustalimy, czy PKB na osobę jest istotnym czynnikiem samobójstw.

6.3.1 Dla kobiet

Utworzymy zbiór zawierające dane tylko dla kobiet.

```
(kobietyPKB <- plecPKB[which(plecPKB$Plec == 'female'),])

(
  kobietyPKBlm = lm(CLS ~ PKB, data = kobietyPKB)
)

summary(kobietyPKBlm)
```

Uzyskany wynik:

```
> summary(kobietyPKBlm)

Call:
lm(formula = CLS ~ PKB, data = kobietyPKB)

Residuals:
   Min     1Q Median     3Q    Max
-40112 -14902 -10579  -1470 225324

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.022e+04  6.296e+03   1.624    0.108
PKB         4.419e-01  2.762e-01   1.600    0.113

Residual standard error: 42550 on 88 degrees of freedom
Multiple R-squared:  0.02825,   Adjusted R-squared:  0.01721
F-statistic: 2.558 on 1 and 88 DF,  p-value: 0.1133
```

Wartość p-value ≥ 0.05, zatem nie ma podstaw do odrzucenia hipotezy zerowej.

6.3.2 Dla mężczyzn

Utworzymy zbiór zawierające dane tylko dla mężczyzn.

```
(mezczyzniPKB <- plecPKB[which(plecPKB$Plec == 'male'),])

(
  mezczyzniPKBlm = lm(CLS ~ PKB, data = mezczyzniPKB)
)

summary(mezczyzniPKBlm)
```

Uzyskany wynik:

```
> summary(mezczyzniPKBlm)

Call:
lm(formula = CLS ~ PKB, data = mezczyzniPKB)

Residuals:
    Min      1Q  Median      3Q     Max
-108534  -50320  -40840  -15340  947278

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.166e+04  2.223e+04   1.874   0.0642 .
PKB         9.926e-01  9.754e-01   1.018   0.3116
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 150300 on 88 degrees of freedom
Multiple R-squared:  0.01163,   Adjusted R-squared:  0.0003993
F-statistic: 1.036 on 1 and 88 DF,  p-value: 0.3116
```

Wartość p-value ≥ 0.05, zatem nie ma podstaw do odrzucenia hipotezy zerowej.

Zatem istnieje słaba, ale dodatnia zależność liniowa - bogatsze kraje wiążą się z wyższymi wskaźnikami samobójstw, ale jest to zależność słaba.

7. Opis użytych pakietów

- library(tidyverse) – pakiet do selekcji danych (%>%)
- library(plotly) - pakiet do tworzenia wykresów plot
- library(countrycode) – pakiet do utworzenia kolumny ‘kontynenty’
- library(highcharter) – pakiet do tworzenia wykresów highcarter
- library(moments) – pakiet do opisu statystyk np. kurtozę

8. Podsumowanie

Ogólnie rzecz biorąc, wskaźniki samobójstw są znacznie wyższe wśród mężczyzn. Wskaźnik samobójstw mężczyzn jest 3-4 razy wyższy niż wskaźnik samobójstw kobiet. Zjawisko to występuje we wszystkich 90 krajach. Wskaźnik samobójstw zarówno wśród mężczyzn, jak i kobiet był najwyższy w 1995 roku i od tego czasu spada. W roku 2015 wskaźniki samobójstw spadły mniej więcej do tego samego poziomu, co w latach 1988-1991. W skali globalnej wskaźniki samobójstw rosną wraz z wiekiem, przy czym grupa wiekowa 5-14 lat (najmłodsza grupa wiekowa w danych) ma najniższe wskaźniki samobójstw, a grupa wiekowa 75+ - najwyższe.