Samobójstwa w latach 1985-2015 - analiza

Magdalena Bogacz

Spis treści

Wstęp

Wprowadzenie

Przedmiotem niniejszej analizy jest zbiór danych zawierający informacje na temat samobójstw popełnionych na całym świecie w latach 1985-2016. Dane pochodzą ze strony www.kaggle.com i są dostępne pod tym adresem.

Biblioteki

Podczas procesu analizy użyto następujących bibliotek:

```
library("tidyverse")
library("countrycode")
library(gridExtra)
library(grid)
library(broom)
library(knitr)
```

Czyszczenie danych - przegląd

Następnie przystąpiono do implementacji zbioru danych i przestudiowania ich struktury. W tym celu użyto trzech różnych funkcji - summary(), str() oraz glimpse(). Każda z nich podsumowuje dane w nieco inny sposób - funkcja summary() oferuje proste statystyczne podsumowanie każdej zmiennej, funkcja str() dostarcza informacji o strukturze tabeli a funkcja glimpse() oferuje wgląd w dane zawarte w poszczególnych zmiennych.

```
masterdata <- read_csv("master.csv")
summary(masterdata)
```

```
##    country               year          sex                age
##  Length:27820       Min.   :1985   Length:27820       Length:27820
##  Class :character   1st Qu.:1995   Class :character   Class :character
##  Mode  :character   Median :2002   Mode  :character   Mode  :character
##                     Mean   :2001
##                     3rd Qu.:2008
##                     Max.   :2016
##
##   suicides_no        population       suicides/100k pop country-year
##  Min.   :    0.0   Min.   :     278   Min.   :  0.00    Length:27820
##  1st Qu.:    3.0   1st Qu.:   97498   1st Qu.:  0.92    Class :character
##  Median :   25.0   Median :  430150   Median :  5.99    Mode  :character
##  Mean   :  242.6   Mean   : 1844794   Mean   : 12.82
##  3rd Qu.:  131.0   3rd Qu.: 1486143   3rd Qu.: 16.62
##  Max.   :22338.0   Max.   :43805214   Max.   :224.97
##
##   HDI for year   gdp_for_year ($)    gdp_per_capita ($)  generation
##  Min.   :0.483   Min.   :4.692e+07   Min.   :   251     Length:27820
##  1st Qu.:0.713   1st Qu.:8.985e+09   1st Qu.:  3447     Class :character
##  Median :0.779   Median :4.811e+10   Median :  9372     Mode  :character
##  Mean   :0.777   Mean   :4.456e+11   Mean   : 16866
```

```
##  3rd Qu.:0.855   3rd Qu.:2.602e+11   3rd Qu.: 24874
##  Max.   :0.944   Max.   :1.812e+13   Max.   :126352
##  NA's   :19456
```

```
str(masterdata)
```

```
## tibble [27,820 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ country           : chr [1:27820] "Albania" "Albania" "Albania" "Albania" ...
##  $ year              : num [1:27820] 1987 1987 1987 1987 1987 ...
##  $ sex               : chr [1:27820] "male" "male" "female" "male" ...
##  $ age               : chr [1:27820] "15-24 years" "35-54 years" "15-24 years" "75+ years" ...
##  $ suicides_no       : num [1:27820] 21 16 14 1 9 1 6 4 1 0 ...
##  $ population        : num [1:27820] 312900 308000 289700 21800 274300 ...
##  $ suicides/100k pop : num [1:27820] 6.71 5.19 4.83 4.59 3.28 2.81 2.15 1.56 0.73 0 ...
##  $ country-year      : chr [1:27820] "Albania1987" "Albania1987" "Albania1987" "Albania1987" ...
##  $ HDI for year      : num [1:27820] NA NA NA NA NA NA NA NA NA NA ...
##  $ gdp_for_year ($)  : num [1:27820] 2.16e+09 2.16e+09 2.16e+09 2.16e+09 2.16e+09 ...
##  $ gdp_per_capita ($): num [1:27820] 796 796 796 796 796 796 796 796 796 796 ...
##  $ generation        : chr [1:27820] "Generation X" "Silent" "Generation X" "G.I. Generation" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   country = col_character(),
##   ..   year = col_double(),
##   ..   sex = col_character(),
##   ..   age = col_character(),
##   ..   suicides_no = col_double(),
##   ..   population = col_double(),
##   ..   `suicides/100k pop` = col_double(),
##   ..   `country-year` = col_character(),
##   ..   `HDI for year` = col_double(),
##   ..   `gdp_for_year ($)` = col_number(),
##   ..   `gdp_per_capita ($)` = col_double(),
##   ..   generation = col_character()
##   .. )
```

```
glimpse(masterdata)
```

```
## Rows: 27,820
## Columns: 12
## $ country              <chr> "Albania", "Albania", "Albania", "Albania", "A...
## $ year                 <dbl> 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987...
## $ sex                  <chr> "male", "male", "female", "male", "male", "fem...
## $ age                  <chr> "15-24 years", "35-54 years", "15-24 years", "...
## $ suicides_no          <dbl> 21, 16, 14, 1, 9, 1, 6, 4, 1, 0, 0, 0, 2, 17, ...
## $ population           <dbl> 312900, 308000, 289700, 21800, 274300, 35600, ...
## $ `suicides/100k pop`  <dbl> 6.71, 5.19, 4.83, 4.59, 3.28, 2.81, 2.15, 1.56...
## $ `country-year`       <chr> "Albania1987", "Albania1987", "Albania1987", "...
## $ `HDI for year`       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ `gdp_for_year ($)`   <dbl> 2156624900, 2156624900, 2156624900, 2156624900...
## $ `gdp_per_capita ($)` <dbl> 796, 796, 796, 796, 796, 796, 796, 796, 796, 7...
## $ generation           <chr> "Generation X", "Silent", "Generation X", "G.I...
```

Poniższy kod zwraca informację o tym, czy w danej tabeli zostały określone poziomy (levels). Jeśli tak jest, kod zwróci ich nazwy, jeśli nie, zwróci komunikat 'NULL'.

```
levels(masterdata)
```

```
## NULL
```

Czyszczenie danych - zmienna 'generation'

Po wstępnym przeglądzie zdecydowano się na pominięcie w dalszej analizie zmiennej "generation" (pol. pokolenie). W każdej części świata poszczególne pokolenia są nieco inaczej definiowane - np. definicja tzw. millenialsów (pokolenia X) w Stanach Zjednoczonych i Polsce obejmuje różne grupy wiekowe.

```
masterdata %>%
  select(-generation) -> masterdata
```

Czyszczenie danych - wskaźnik HDI

Podczas wstępnej analizy danych szczególną uwagę zwróciła zmienna 'HDI'. Istnieje w niej wiele pustych rekordów, tak więc podczas dalszej pracy zmienna ta nie odzwierciedlałaby rzeczywistych wyników.

```
masterdata %>%
  select(`HDI for year`) %>%
  filter(!is.na(`HDI for year`)) %>%
  summarise(n())
```

n()
8364

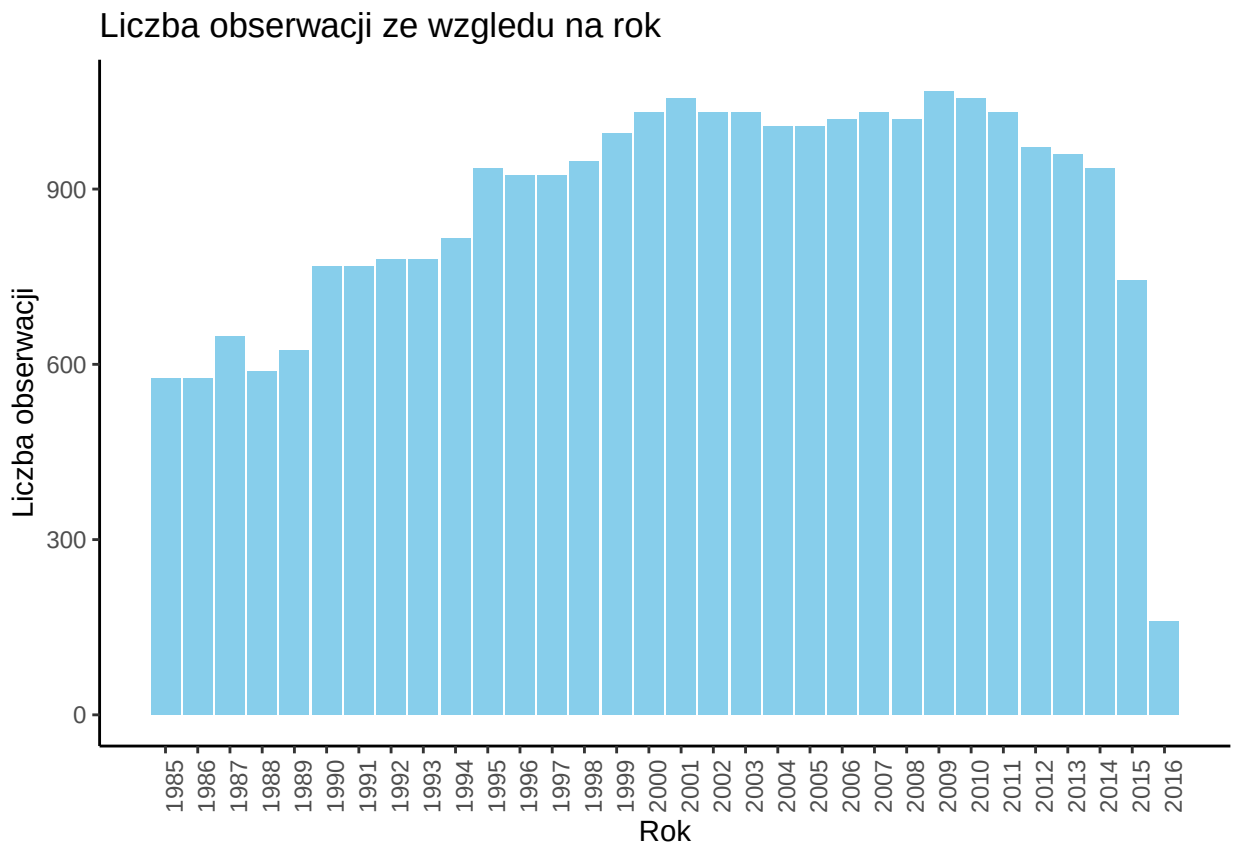
W ostatecznym rozrachunku zdecydowano się ją pominąć.

```
masterdata %>%
  select(-`HDI for year`) -> masterdata
```

Czyszczenie danych - zmienna 'year'

Poniższy wykres uwzględnia ilość obserwacji ze względu na rok.

```
masterdata %>%
  select(year) %>%
  group_by(year) %>%
  summarise(l_obserwacji=n()) %>%
  ggplot(aes( x = year, y=l_obserwacji)) +
  geom_col(fill="skyblue") +
  theme_classic()+
  labs(title="Liczba obserwacji ze względu na rok", fill= "Liczba obserwacji", x="Rok", y="Liczba obser
  theme(axis.text.x = element_text(angle = 90))+
  scale_x_continuous(breaks = seq(1985, 2016, 1))
```



Ze względu na znacznie niższą liczbę obserwacji w roku 2016, zdecydowano się na pominięcie tej zmiennej w dalszej analizie. Zbyt mała liczba rekordów sprawiłaby, że uzyskane informacje byłyby niemiarodajne.

```
masterdata %>%
  filter(!year==2016) -> masterdata
```

Czyszczenie danych - zmienna 'country'

Następnym krokiem było uzyskanie informacji na temat liczby obserwacji uzyskanych w poszczególnych krajach. Wyniki posortowana w kolejności od najniższego to najwyższego.

```
masterdata %>%
  select(country) %>%
  group_by(country) %>%
  summarise(l_obserwacji=n()) %>%
  arrange(l_obserwacji) %>%
  head(10)
```

country	l_obserwacji
Cabo Verde	12
Dominica	12
Macau	12
Bosnia and Herzegovina	24
Oman	36

country	l_obserwacji
Saint Kitts and Nevis	36
San Marino	36
Nicaragua	72
United Arab Emirates	72
Turkey	84

Podczas dalszej analizy wykluczono kraje, w których liczba obserwacji nie przekroczyła liczby 36 - w rezultacie nie wzięto pod uwagę 7 krajów.

```
masterdata %>%
filter(!country %in% c("Cabo Verde", "Dominica", "Macau", "Bosnia and Herzegovina", "Oman", "Saint Kitts
```

Sprawdzenie, czy występują rekordy puste:

```
is.null(masterdata)
```

```
## [1] FALSE
```

Edycja danych - zmienna 'years'

Usunięcie z wszystkich rekordów słowa "years" w celu zwiększenia przejrzystości danych:

```
masterdata$age <- gsub(" years", "", masterdata$age)
```

Edycja danych - zmienna 'country'

Następnie przystąpiono do stworzenia nowej kolumny o nazwie "continent". Jej celem jest przypisanie poszczególnych krajów do odpowiednich kontynentów. W tym celu użyto biblioteki "countrycode".

```
masterdata$continent <- countrycode(sourcevar = masterdata[["country"]],
                                    origin = "country.name",
                                    destination = "continent")
```

Edycja danych - zmienna 'gdp_for_year ($)'

Aby ułatwić wykorzystywanie zmiennej "gdp_for_year ($)", zmieniono jej nazwę na "gdp_for_year".

```
masterdata %>%
rename(gdp_for_year = `gdp_for_year ($)`,
         gdp_per_capita = `gdp_per_capita ($)`) -> masterdata
```

Edycja danych - zmiana typu na typ *factor*

Kolejnym krokiem była zmiana typu danych z *character* na *factor*. Celem tego zabiegu jest możliwość dalszego przetwarzania i analizy nie tylko danych liczbowych, ale także tekstowych.

```
masterdata %>%
   mutate_if(is.character,as.factor) -> masterdata
```

Szybki podgląd struktury tabeli, z którego wynika, że dane będące uprzednio typu *character* zmieniły swój typ na *factor*. Pojawiły się również informacje o poziomach (levels).

```
str(masterdata)
```

```
## tibble [27,492 x 11] (S3: tbl_df/tbl/data.frame)
##  $ country          : Factor w/ 93 levels "Albania","Antigua and Barbuda",..: 1 1 1 1 1 1 1 1 1 1 ..
##  $ year             : num [1:27492] 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","25-34",..: 1 3 1 6 2 6 3 2 5 4 ...
##  $ suicides_no      : num [1:27492] 21 16 14 1 9 1 6 4 1 0 ...
##  $ population       : num [1:27492] 312900 308000 289700 21800 274300 ...
##  $ suicides/100k pop: num [1:27492] 6.71 5.19 4.83 4.59 3.28 2.81 2.15 1.56 0.73 0 ...
##  $ country-year     : Factor w/ 2291 levels "Albania1987",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ gdp_for_year     : num [1:27492] 2.16e+09 2.16e+09 2.16e+09 2.16e+09 2.16e+09 ...
##  $ gdp_per_capita   : num [1:27492] 796 796 796 796 796 796 796 796 796 796 ...
##  $ continent        : Factor w/ 5 levels "Africa","Americas",..: 4 4 4 4 4 4 4 4 4 4 ...
```

```
is.factor(masterdata$country)
```

```
## [1] TRUE
```

Następnie przetłumaczono nazwy poziomów zmiennej 'continent' i 'sex' na ich polskie odpowiedniki. Zmieniono również typ zmiennej 'age' na *factor*, nadano jej odpowiednie poziomy i uporządkowano je w kolejności od najniższego do najwyższego wieku.

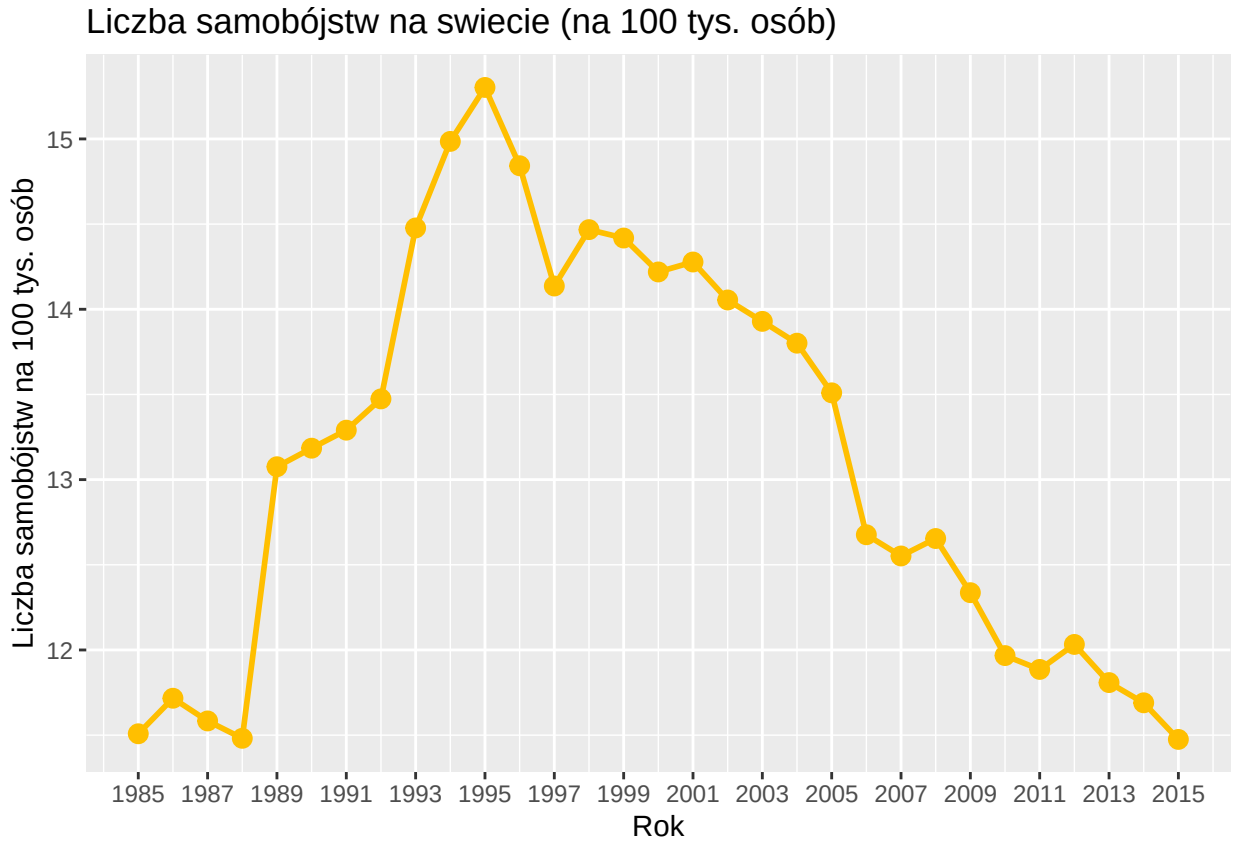
```
levels(masterdata$continent)[levels(masterdata$continent)=="Africa"] <- "Afryka"
levels(masterdata$continent)[levels(masterdata$continent)=="Americas"] <- "Ameryki"
levels(masterdata$continent)[levels(masterdata$continent)=="Asia"] <- "Azja"
levels(masterdata$continent)[levels(masterdata$continent)=="Europe"] <- "Europa"

levels(masterdata$sex)[levels(masterdata$sex)=="female"] <- "Kobiety"
levels(masterdata$sex)[levels(masterdata$sex)=="male"] <- "Mężczyźni"

masterdata$age <- factor(masterdata$age,
                         ordered = T,
                         levels = c('5-14', '15-24', '25-34','35-54','55-74', '75+'))
```

Analiza danych - podstawowa

Samobójstwa w skali globalnej

```
masterdata %>%
  group_by(year) %>%
  summarise(population = sum(population),
```

```
            suicides = sum(suicides_no),
            suicides_per_100k = (suicides / population) * 100000) %>%
  ggplot(aes(x = year, y = suicides_per_100k)) +
  geom_line(col = "#FFBF00", size = 1) +
  geom_point(col = "#FFBF00", size = 3) +
    labs(title = "Liczba samobójstw na świecie (na 100 tys. osób)",
       x = "Rok",
       y = "Liczba samobójstw na 100 tys. osób") +
  scale_x_continuous(breaks = seq(1985, 2015, 2)) +
  scale_y_continuous(breaks = seq(10, 20))
```

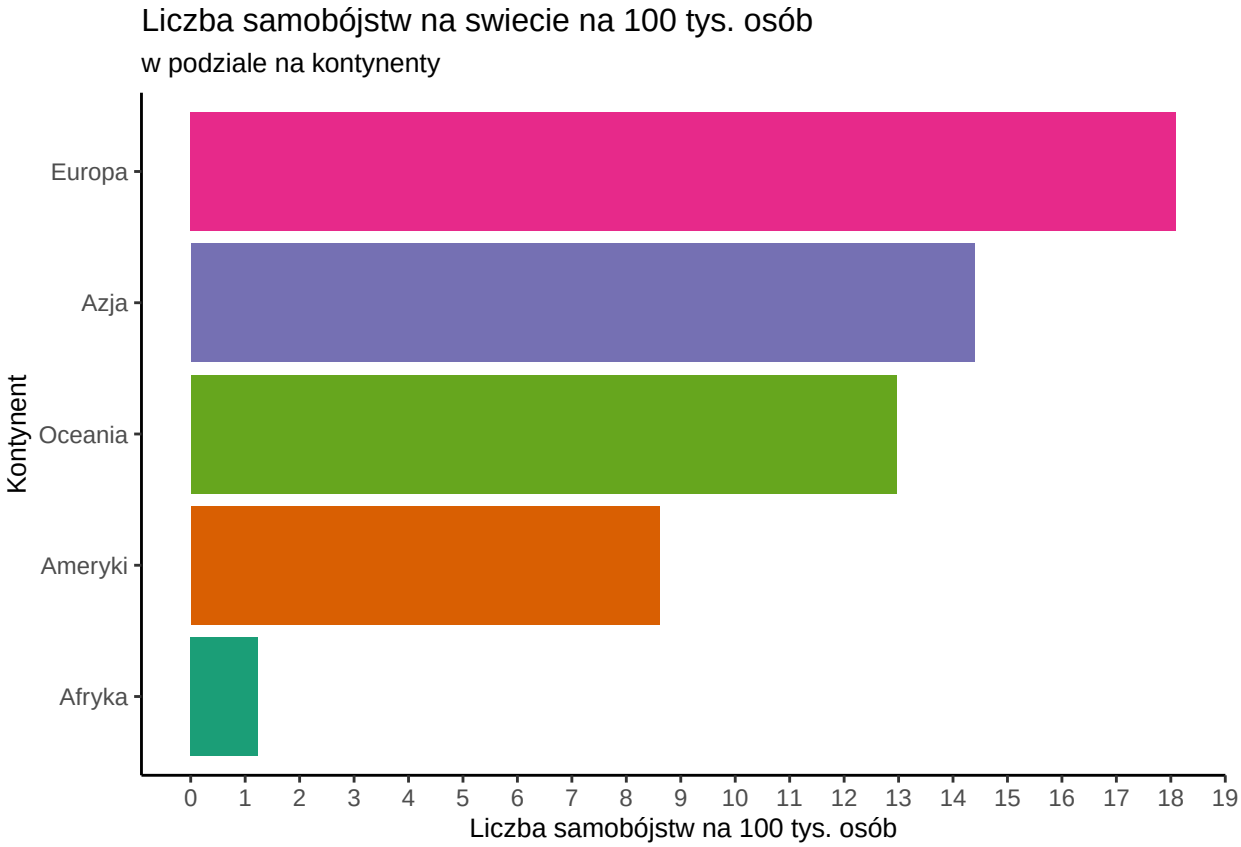


W latach 1989 - 1995 nastąpił wzrost liczby samobójstw, od tego czasu regularnie maleje ich ilość. W roku 2015 średnia liczba samobójstw w przeliczeniu na 100 tysięcy osób jest najniższa ze wszystkich lat, które są brane pod uwagę w tej analizie.

Podział na kontynenty

```
masterdata %>%
  group_by(continent) %>%
  summarise(suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000) %>%
  arrange(suicide_per_100k) -> continent

continent %>%
  ggplot(aes(x= reorder(continent,suicide_per_100k), y = suicide_per_100k, fill = continent)) +
```

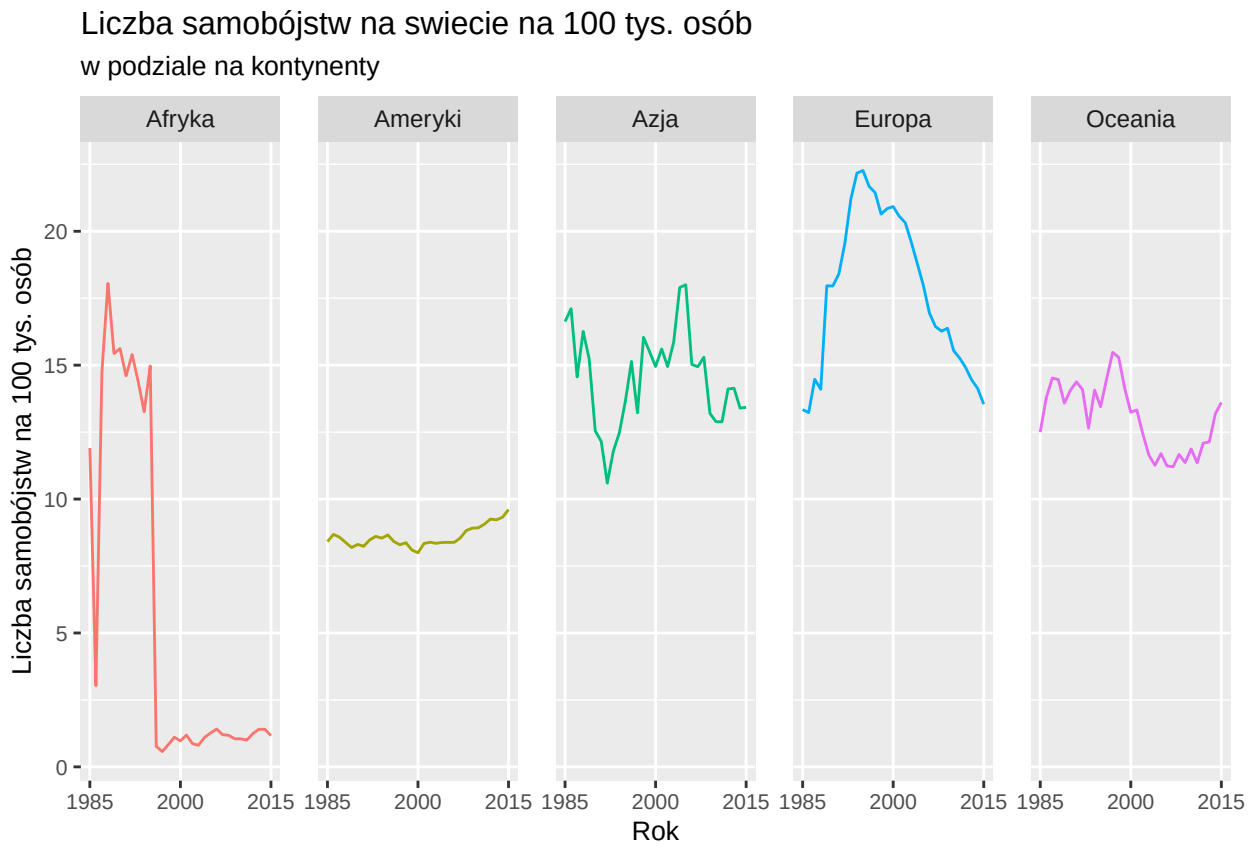
```
  geom_col() +
  labs(title = "Liczba samobójstw na świecie na 100 tys. osób",
       subtitle = "w podziale na kontynenty",
  x = "Kontynent",
  y = "Liczba samobójstw na 100 tys. osób") +
  scale_y_continuous(breaks = seq(0, 20, 1), minor_breaks = F)+
  scale_fill_brewer(palette = "Dark2")+
  theme_classic()+
  theme(legend.position = "none", title = element_text(size = 10))+
  coord_flip()
```



Najczęściej samobójstwo popełnia Europejczyk, a najrzadziej osoba zamieszkująca kontynent Afryki. W Europie dzieje się to średnio 14.5 razy częściej niż w Afryce. Tak duża różnica może wynikać z większej ilości danych na temat Europy niż Afryki zawartych w tej bazie.

```
continent_year <- masterdata %>%
  group_by(year, continent) %>%
  summarize(suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000)

continent_year %>%
  ggplot(aes(x = year, y = suicide_per_100k, color = continent)) +
  facet_grid(~continent, scales = "free_y") +
  geom_line() +
  labs(title = "Liczba samobójstw na świecie na 100 tys. osób",
       subtitle = "w podziale na kontynenty",
       x = "Rok",
```

```
      y = "Liczba samobójstw na 100 tys. osób") +
  theme(legend.position = "none",
        title = element_text(size = 10),
        panel.spacing = unit(1, "lines"),
        axis.text.x = element_text(size = 8),
        axis.text.y = element_text(size = 8)) +
  scale_x_continuous(breaks = seq(1985, 2015, 15), minor_breaks = F)
```



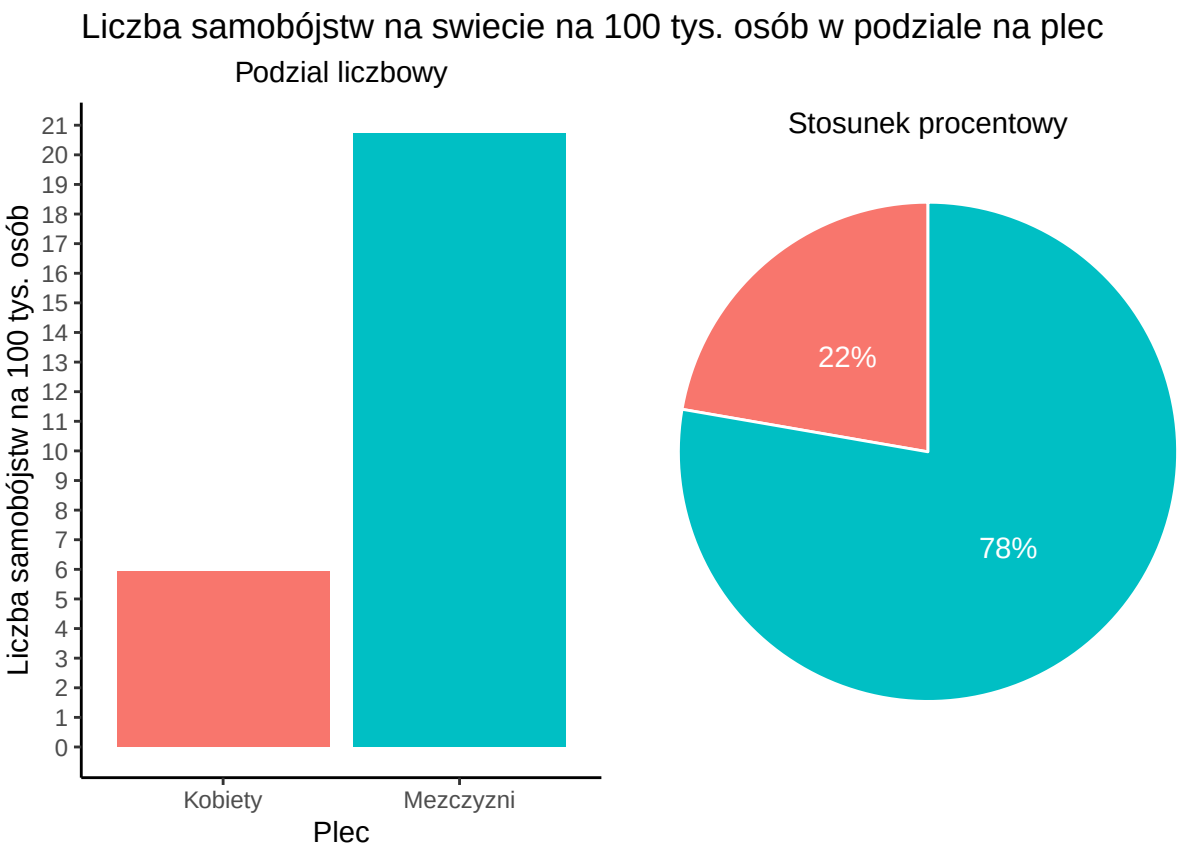
Na przestrzeni lat 1985 - 2015 Azja, Europa i Afryka wykazują tendencję spadkową liczby samobójstw na 100 tysięcy osób. Tymczasem w Amerykach oraz Oceanii widać wzrost.

W Afryce od 1996 roku średnia liczba samobójstw na 100 tysięcy osób wynosi mniej niż 1 osoba, co potwierdza, że zdarzenia te nie są wpisywane do bazy tak skrupulatnie jak w przypadku innych kontynentów.

Płeć

```
masterdata %>%
  group_by(sex) %>%
  summarise(suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000) -> suisex

suisex %>%
  ggplot(aes(x = sex, y = suicide_per_100k, fill = sex)) +
  geom_col() +
  labs(title = "Liczba samobójstw na świecie na 100 tys. osób w podziale na płeć",
       subtitle = "Podział liczbowy",
```

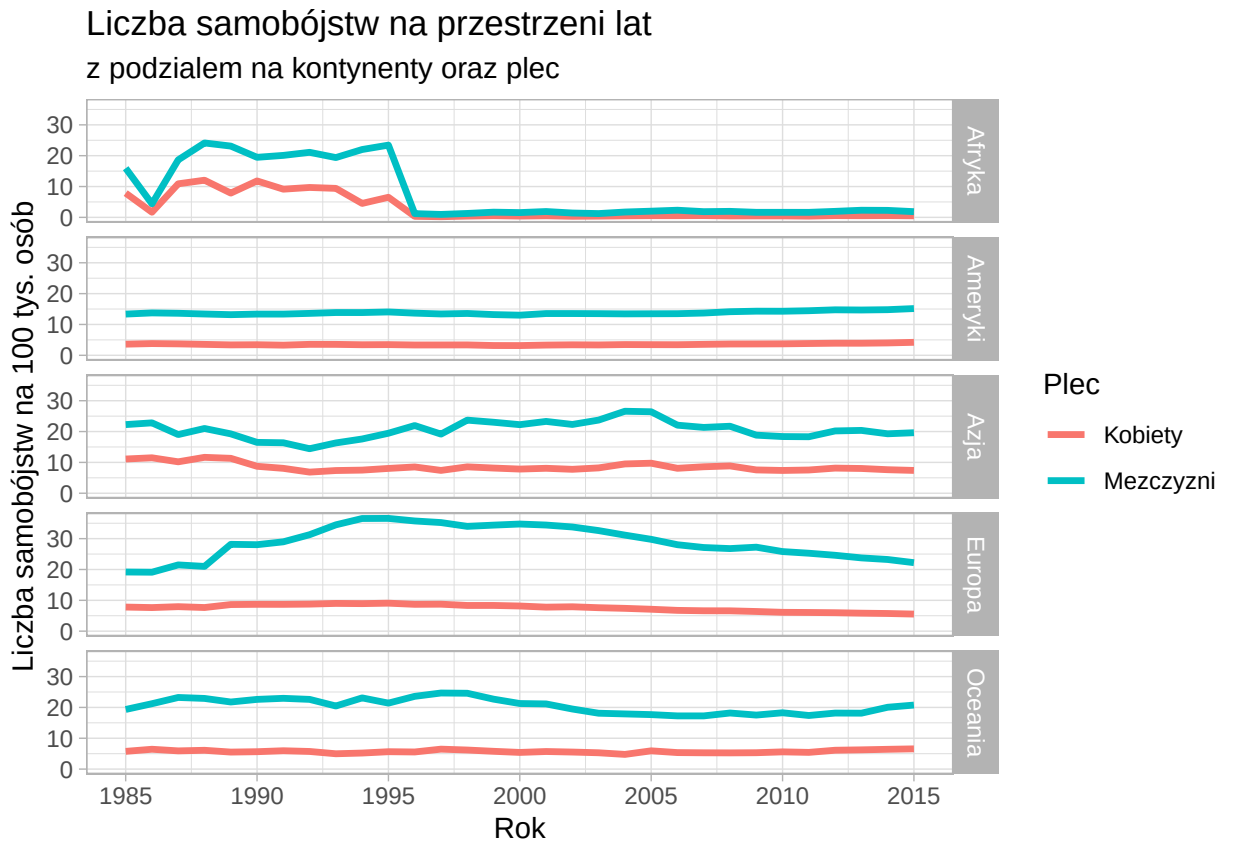
```
       x = "Płeć",
       y = "Liczba samobójstw na 100 tys. osób") +
  scale_y_continuous(breaks = seq(0, 25), minor_breaks = F)+
  theme_classic()+
  theme(legend.position = "none",plot.subtitle = element_text(hjust = 0.5)) -> sex_plot_bar
  #scale_x_discrete(labels = c("Kobiety", "Mężczyźni"))-> sex_plot_bar

suisex %>%
  ggplot(aes(x = "", y = suicide_per_100k, fill = sex)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0)+
  labs(subtitle = "Stosunek procentowy")+
  theme_void()+
  theme(legend.position="none",plot.subtitle = element_text(hjust = 0.5))+
  geom_text(aes(label = paste0(round(suicide_per_100k/sum(suicide_per_100k)*100), "%")), position = pos


grid.arrange(sex_plot_bar, sex_plot_pie, ncol = 2)
```



Samobójstwa zdecydowanie częściej popełniają mężczyźni niż kobiety. Globalnie wskaźnik samobójstw mężczyzn jest 3.7 razy wyższy niż u kobiet.

```
masterdata %>%
  group_by(year, continent, sex) %>%
  summarise(suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000) -> mix
```

```
mix %>%
  ggplot(aes(x = year, y = suicide_per_100k, col = continent, group=sex)) +
  facet_grid(continent~ .) +
  geom_line(aes(color=sex), size=1.2) +
  #geom_point(aes(color=sex))+
  labs(title = "Liczba samobójstw na przestrzeni lat",
       subtitle = "z podziałem na kontynenty oraz płeć",
       x = "Rok",
       y = "Liczba samobójstw na 100 tys. osób",
       color = "Płeć") +
   scale_x_continuous(breaks = seq(1985, 2015, 5))+
   theme_light()
```



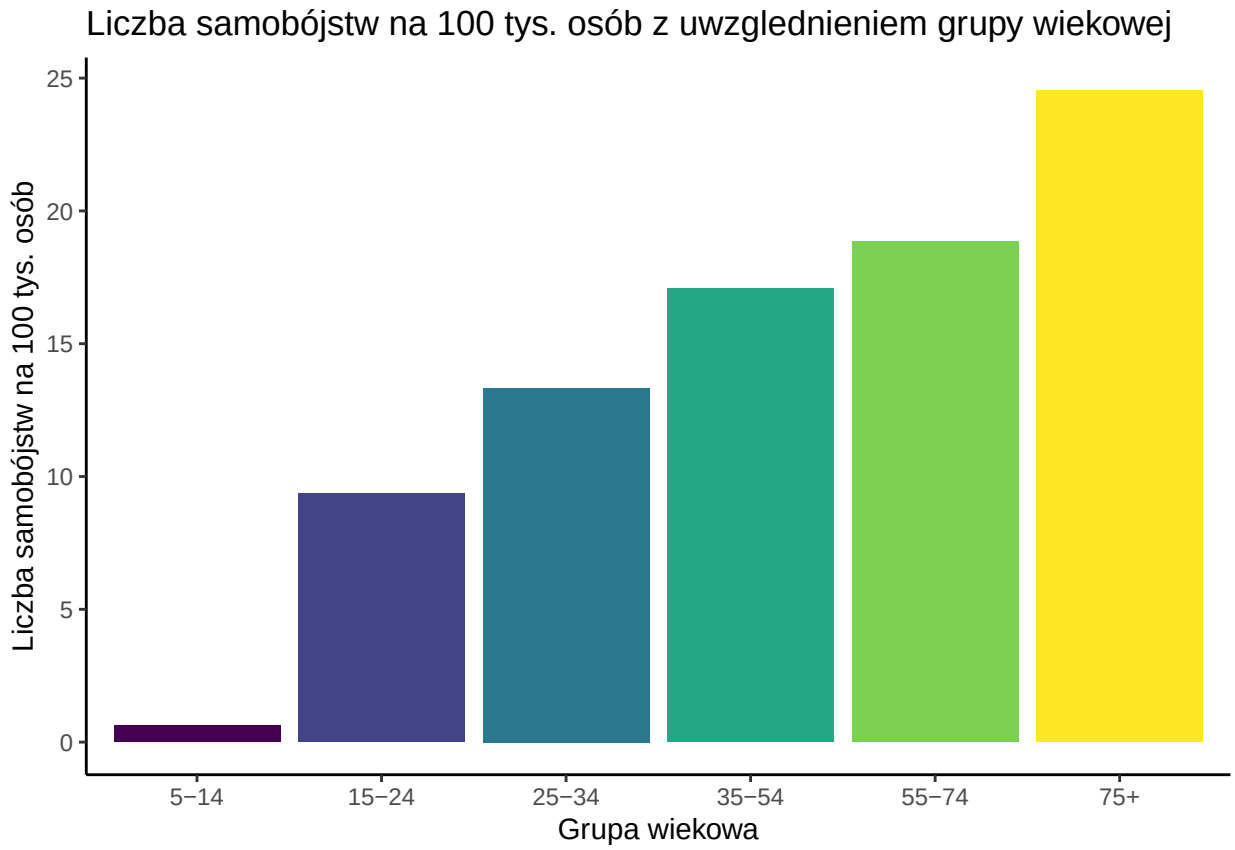

```
   #scale_color_discrete(labels = c("Kobiety", "Mężczyźni"))
```

Na wszystkich kontynentach oraz na przestrzeni wszystkich lat, które są brane pod uwagę podczas tej analizy, mężczyźni częściej popełniają samobójstwa. Z wykresu można odczytać, że mężczyźni z Europy to grupa, która najczęsciej popełniała samobójstwo.

Wiek

```
age<- masterdata %>%
  group_by(age) %>%
```

```
  summarize(suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000)

age %>%
  ggplot(aes(x = age, y = suicide_per_100k, fill = age)) +
  geom_col() +
  labs(title = "Liczba samobójstw na 100 tys. osób z uwzględnieniem grupy wiekowej",
       x = "Grupa wiekowa",
       y = "Liczba samobójstw na 100 tys. osób")+
  theme_classic()+
  theme(plot.subtitle = element_text(hjust = 0.5))+
  theme(legend.position="none")
```

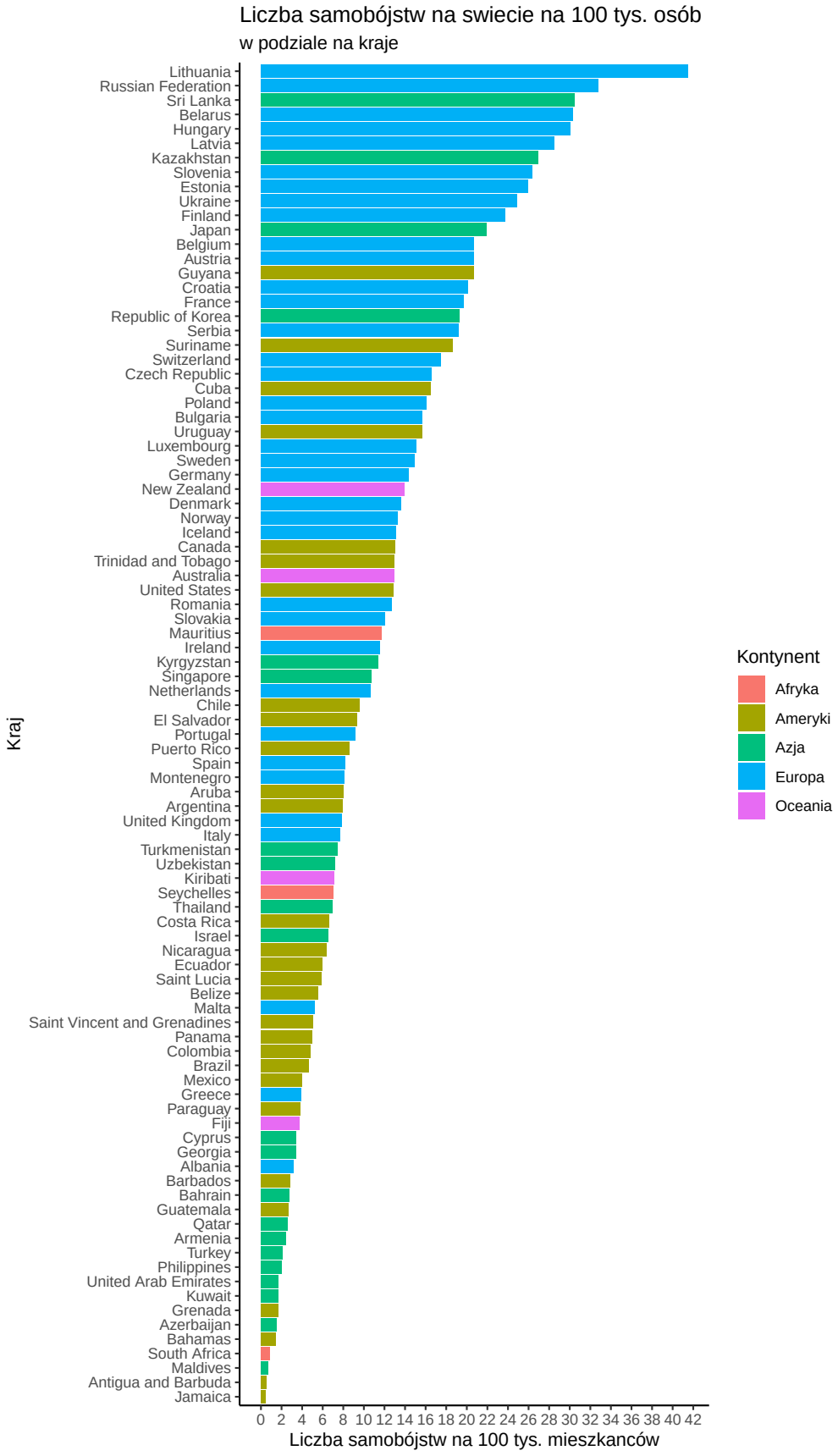


Wraz z wiekiem wzrasta ilość samobójstw. Osoby z grupy wiekowej 75+ popełniają samobójstwo 2.6 razy częściej niż osoby z przedziału wiekowego 15-24. Na 100 tysięcy osób średnio 24 osoby w wieku 75+ popełniają samobójstwo. Wskaźnik dla grupy 5-14 wynosi mniej niż 1 osoba na 100 tysięcy.

Kraj

```
masterdata %>%
  group_by(country, continent) %>%
  summarize(n = n(),
            suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000) %>%
  arrange(desc(suicide_per_100k)) -> country
```

```r
country %>%
  ggplot(aes(x = reorder(country, suicide_per_100k), y = suicide_per_100k, fill = continent)) +
  geom_bar(stat = "identity") +
  labs(title = "Liczba samobójstw na świecie na 100 tys. osób",
       subtitle = "w podziale na kraje",
       x = "Kraj",
       y = "Liczba samobójstw na 100 tys. mieszkańców",
       fill = "Kontynent") +
  coord_flip() +
  scale_y_continuous(breaks = seq(0, 45, 2)) +
  theme(legend.position = "bottom")+
  theme_classic()
```

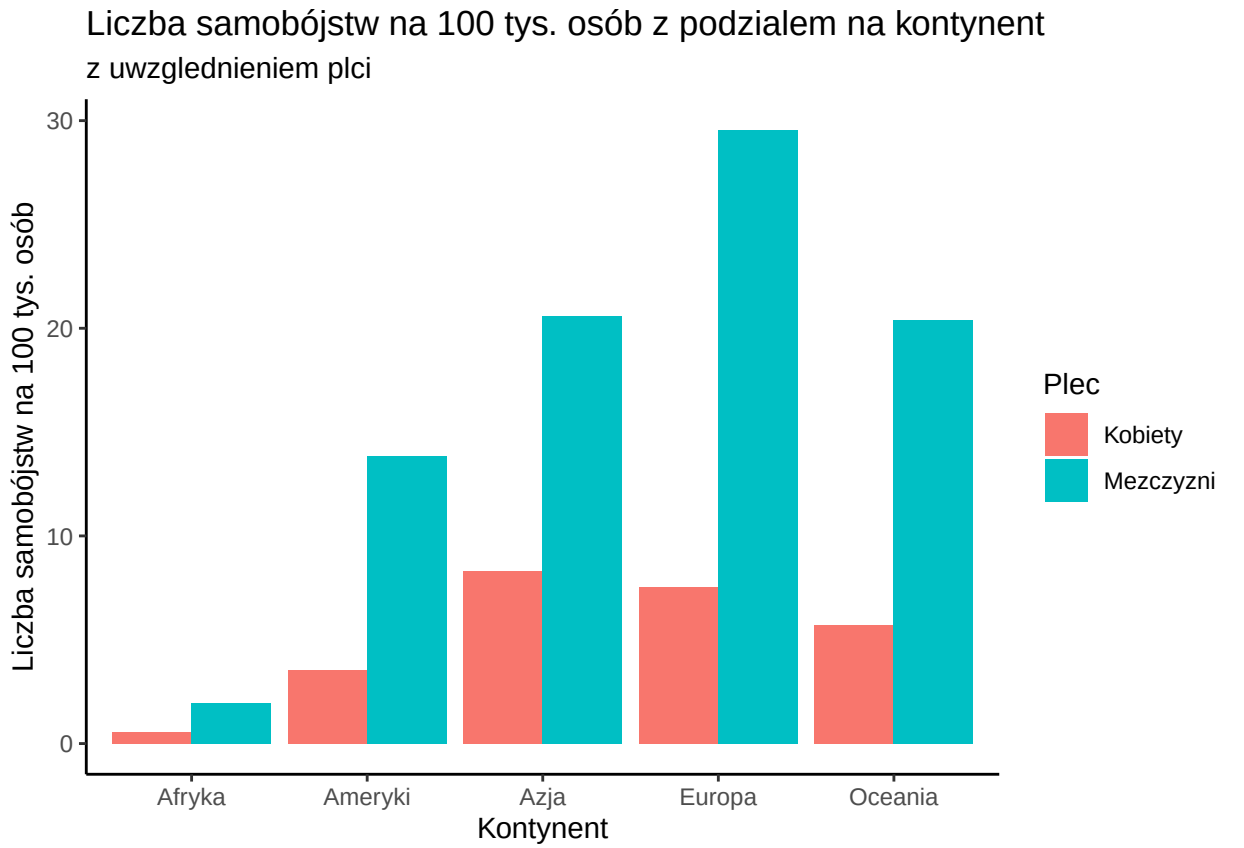

Liczba samobójstw na swiecie na 100 tys. osób
w podziale na kraje
Lithuania
Russian Federation
Sri Lanka
Belarus
Hungary
Latvia
Kazakhstan
Slovenia
Estonia
Ukraine
Finland
Japan
Belgium
Austria
Guyana
Croatia
France
Republic of Korea
Serbia
Suriname
Switzerland
Czech Republic
Cuba
Poland
Bulgaria
Uruguay
Luxembourg
Sweden
Germany
New Zealand
Denmark
Norway
Iceland
Canada
Trinidad and Tobago
Australia
United States
Romania
Slovakia
Mauritius
Ireland
Kyrgyzstan
Singapore
Netherlands
Chile
El Salvador
Portugal
Puerto Rico
Spain
Montenegro
Aruba
Argentina
United Kingdom
Italy
Turkmenistan
Uzbekistan
Kiribati
Seychelles
Thailand
Costa Rica
Israel
Nicaragua
Ecuador
Saint Lucia
Belize
Malta
Saint Vincent and Grenadines
Panama
Colombia
Brazil
Mexico
Greece
Paraguay
Fiji
Cyprus
Georgia
Albania
Barbados
Bahrain
Guatemala
Qatar
Armenia
Turkey
Philippines
United Arab Emirates
Kuwait
Grenada
Azerbaijan
Bahamas
South Africa
Maldives
Antigua and Barbuda
Jamaica
Kraj
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42
Liczba samobójstw na 100 tys. mieszkanców
Kontynent
Afryka
Ameryki
Azja
Europa
Oceania

Najwyższy wskaźnik samobójstw na 100 tysięcy osób występuje na Litwie, wynosząc 41 osób. Kraje europejskie znajdują się na wyższych pozycjach niż kraje z innych kontynentów.

Analiza danych - rozszerzona

Kontynent - płeć

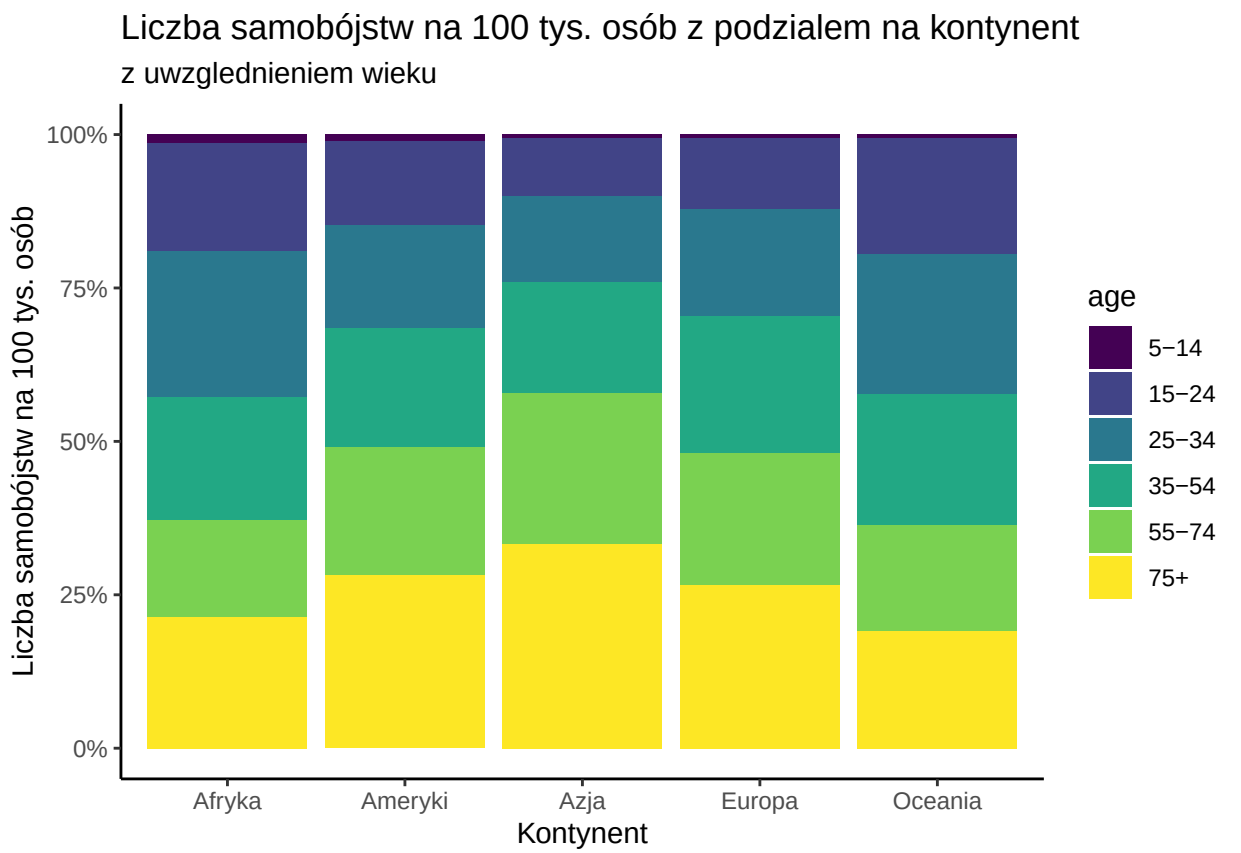
```
masterdata %>%
  group_by(continent, sex) %>%
  summarise(suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000) -> conti_sex

conti_sex %>%
  ggplot(aes(x = continent, y = suicide_per_100k, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Liczba samobójstw na 100 tys. osób z podziałem na kontynent",
       subtitle = "z uwzględnieniem płci",
       x = "Kontynent",
       y = "Liczba samobójstw na 100 tys. osób",
       fill="Płeć")+
  theme_classic()
```



Wykres wskazuje, że mężczyźni popełniają samobójstwa częściej niż kobiety bez względu na kontynent, na którym mieszkają. Wyraźnie widać, że najczęściej samobójstwa popełniają europejscy mężczyźni. Różnica płci pomiędzy liczbą samobójstw najbardziej zatarta jest w Afryce.

Kontynent - wiek

```
masterdata %>%
  group_by(continent, age) %>%
  summarise(suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000) -> conti_age


conti_age %>%
    filter(continent=="Afryka") %>%
    mutate(proportion = suicide_per_100k / sum(suicide_per_100k)) -> conti_age_Af

conti_age %>%
    filter(continent=="Ameryki") %>%
    mutate(proportion = suicide_per_100k / sum(suicide_per_100k)) -> conti_age_Am

conti_age %>%
    filter(continent=="Europa") %>%
    mutate(proportion = suicide_per_100k / sum(suicide_per_100k)) -> conti_age_Eu

conti_age %>%
    filter(continent=="Azja") %>%
    mutate(proportion = suicide_per_100k / sum(suicide_per_100k)) -> conti_age_As

conti_age %>%
    filter(continent=="Oceania") %>%
    mutate(proportion = suicide_per_100k / sum(suicide_per_100k)) -> conti_age_Oc


conti_age <- rbind(conti_age_Af,conti_age_Am, conti_age_As, conti_age_Eu, conti_age_Oc)


conti_age %>%
  ggplot(aes(x = continent, y = proportion, fill = age)) +
  geom_bar(stat = "identity") +
  labs(title = "Liczba samobójstw na 100 tys. osób z podziałem na kontynent",
       subtitle = "z uwzględnieniem wieku",
       x = "Kontynent",
       y = "Liczba samobójstw na 100 tys. osób")+
  scale_y_continuous(labels = scales::percent) +
  theme_classic()
```



Dla obu Ameryk, Azji oraz Europy wskaźnik samobójstw rośnie wraz z wiekiem, podczas gdy Oceania i Afryka najwyższy wskaźnik odnotowuje dla grupy wiekowej 25-34.

Badanie niezależności

H0: Nie istnieje zależność między liczbą samobójstw ma 100 tysięcy a PKB na mieszkańca
H1: Istnieje zależność między liczbą samobójstw a PKB na mieszkańca

```
masterdata %>%
  group_by(country, year, gdp_per_capita) %>%
  summarise(suicide_per_100k = (sum(suicides_no) / sum(population) * 100000))-> sui_gdp

shapiro.test(sui_gdp$gdp_per_capita)
```

```
Shapiro-Wilk normality test
```

data: sui_gdp$gdp_per_capita W = 0.78448, p-value < 2.2e-16

```
shapiro.test(sui_gdp$suicide_per_100k)
```

```
Shapiro-Wilk normality test
```

data: sui_gdp$suicide_per_100k W = 0.91896, p-value < 2.2e-16

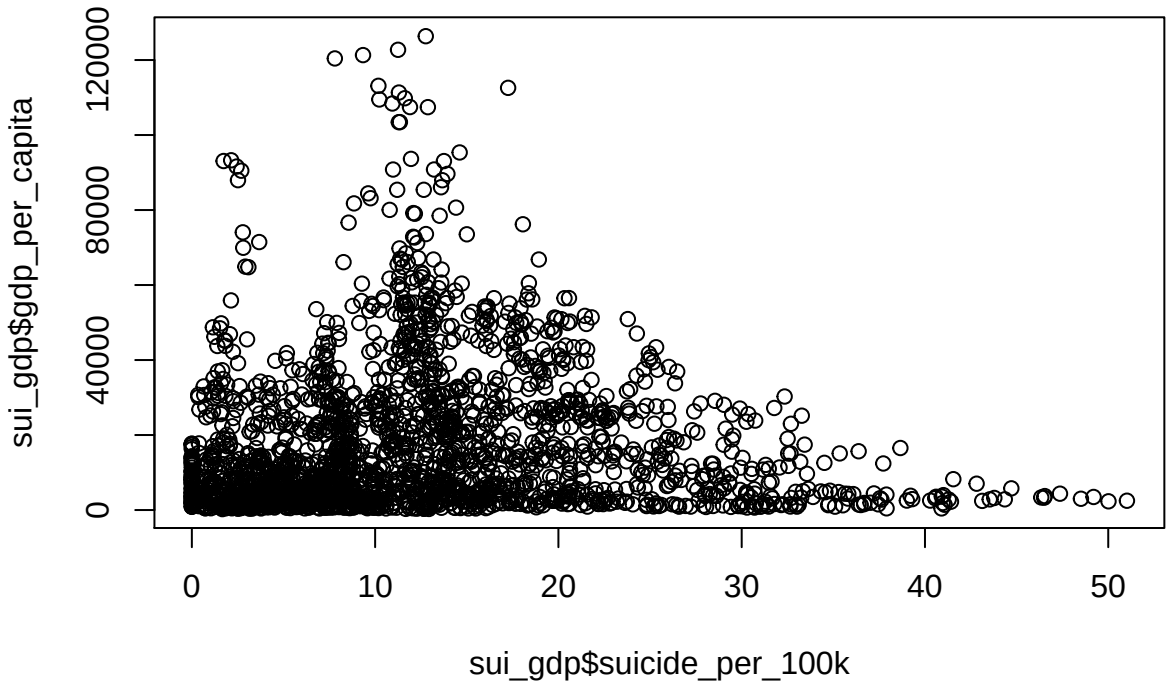
Zarówno zmienna opisująca PKB w przeliczeniu na 1 mieszkańca 'gdp_per_capita', jak i zmienna 'suicide_per_100k' określająca liczbę samobójstw na 100 tys. mieszkańców nie mają rozkładu normalnego, zatem do badania wykorzystano test Kendall'a.

```
cor.test( sui_gdp$suicide_per_100k,sui_gdp$gdp_per_capita, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  sui_gdp$suicide_per_100k and sui_gdp$gdp_per_capita
## z = 9.0418, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.1260757
```

Wartość *P-value* jest mniejsza od 0.05, ale wartość *tau* wynosi 0.12, więc nie istnieje zależność między zmiennymi.

```
plot(sui_gdp$suicide_per_100k,sui_gdp$gdp_per_capita)
```



Na powyższym wykresie również można zauważyć brak wyraźnego związku między zmiennymi, gdyż punkty są umieszczone w sposób dowolny.

Analiza historyczna w państwach europejskich

W ciągu ostatnich 30 lat Europa doświadczyłą znaczących zmian społeczno-gospodarczych. Biorąc pod uwagę, że użyty w tej analizie zbiór danych odnosi się niemal idealnie do tego okresu, naturalnym następnym krokiem wydała się analiza różnic pomiędzy poszczególnymi krajami europejskimi na przestrzeni lat.

Podział krajów europejskich

Kraje europejskie podzielono na dwie grupy. Pierwszą z nich stanowią kraje znajdujące się za tzw. żelazną kurtyną - a więc te pozostające po zakończeniu WWII pod wpływami dawnego Związku Radzieckiego. Drugą grupą stanowią kraje powiązane wpływami Aliantów. Ze względu na to, że dane obejmują okres 30 lat, a tylko 4 z nich Niemcy spędziły w stanie rozpadu, zostały one zakwalifikowane do tej właśnie grupy.

```
masterdata %>%
  filter((country %in% c("Poland",
                         "Czech Republic",
                         "Slovakia",
                         "Hungary",
                         "Romania",
                         "Bulgaria",
                         "Albania",
                         "Estonia",
                         "Latvia",
                         "Lithuania",
                         "Kazakhstan",
                         "Kyrgyzstan",
                         "Tajikistan",
                         "Turkmenistan",
                         "Uzbekistan",
                         "Belarus",
                         "Moldova",
                         "Ukraine",
                         "Russia",
                         "Armenia",
                         "Azerbaijan",
                         "Georgia",
                         "Bosnia and Herzegovina",
                         "Croatia",
                         "Macedonia",
                         "Montenegro",
                         "Serbia",
                         "Slovenia")) & continent=="Europa") -> eastern_bloc
```

```
`%notin%` <- Negate(`%in%`)

masterdata %>%
  filter(country %notin% c("Poland",
                         "Czech Republic",
                         "Slovakia",
                         "Hungary",
                         "Romania",
                         "Bulgaria",
```

```
                        "Albania",
                        "Estonia",
                        "Latvia",
                        "Lithuania",
                        "Kazakhstan",
                        "Kyrgyzstan",
                        "Tajikistan",
                        "Turkmenistan",
                        "Uzbekistan",
                        "Belarus",
                        "Moldova",
                        "Ukraine",
                        "Russia",
                        "Armenia",
                        "Azerbaijan",
                        "Georgia",
                        "Bosnia and Herzegovina",
                        "Croatia",
                        "Macedonia",
                        "Montenegro",
                        "Serbia",
                        "Slovenia")) -> non_eastern_bloc

non_eastern_bloc %>%
  filter(continent=="Europa") -> non_eastern_bloc
```

Kolejnym krokiem było wstępne uporządkowanie poszczególnych lat pod względem liczby samobójstw na 100 tys. mieszkańców.

```
masterdata %>%
  group_by(country,year) %>%
  summarise(suicide_per_100k = (sum(suicides_no) / sum(population) * 100000)) %>%
  arrange(desc(suicide_per_100k)) %>%
  head(5)
```

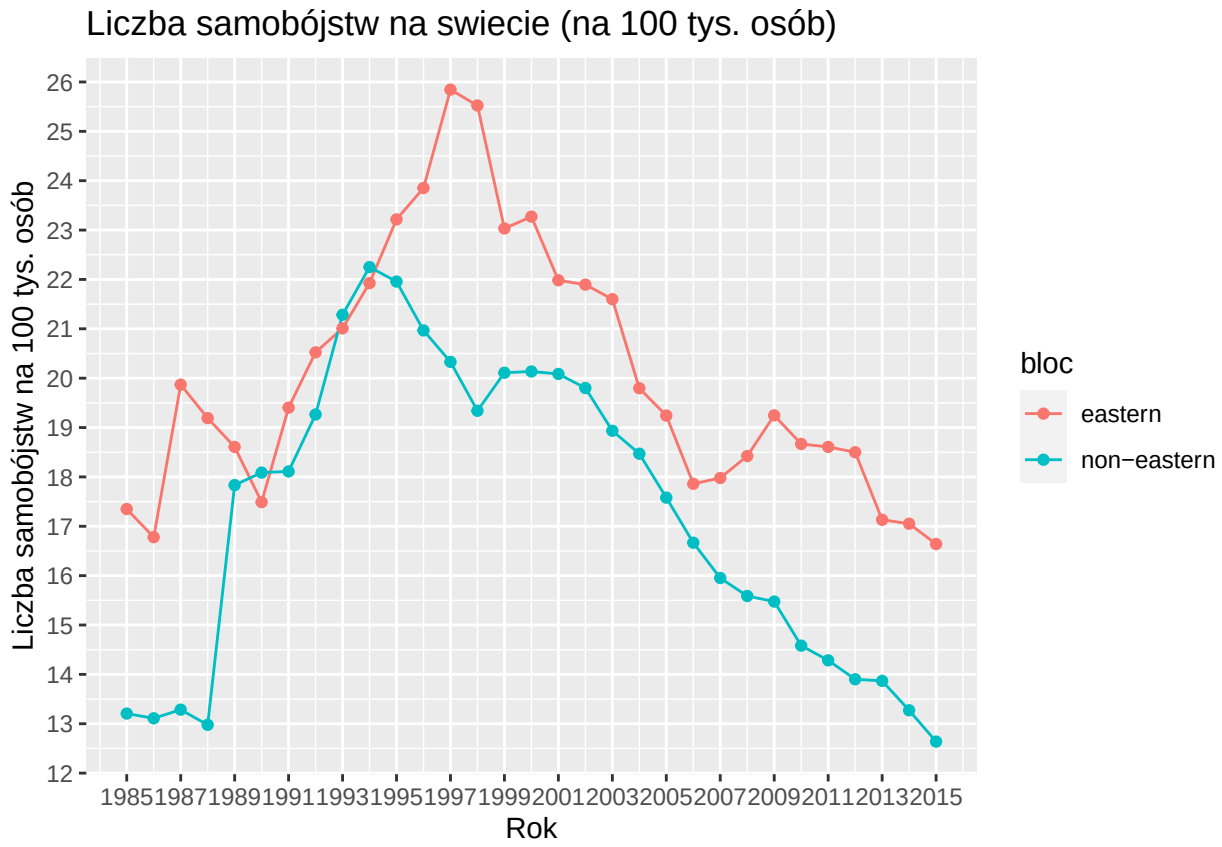
country	year	suicide_per_100k
Lithuania	1996	51.01976
Lithuania	1995	50.01256
Lithuania	2000	49.19875
Lithuania	1997	48.50974
Lithuania	2002	47.36618

Powyższa tabela pokazuje, że zdecydowanym liderem pod względem liczby samobójstw pozostaje Litwa. Dane z wszystkich 5 lat z największą liczbą samobójstw pochodzą właśnie z tego kraju.

Na przestrzeni lat

```
eastern_bloc %>%
  mutate(bloc="eastern") -> eastern_bloc
```

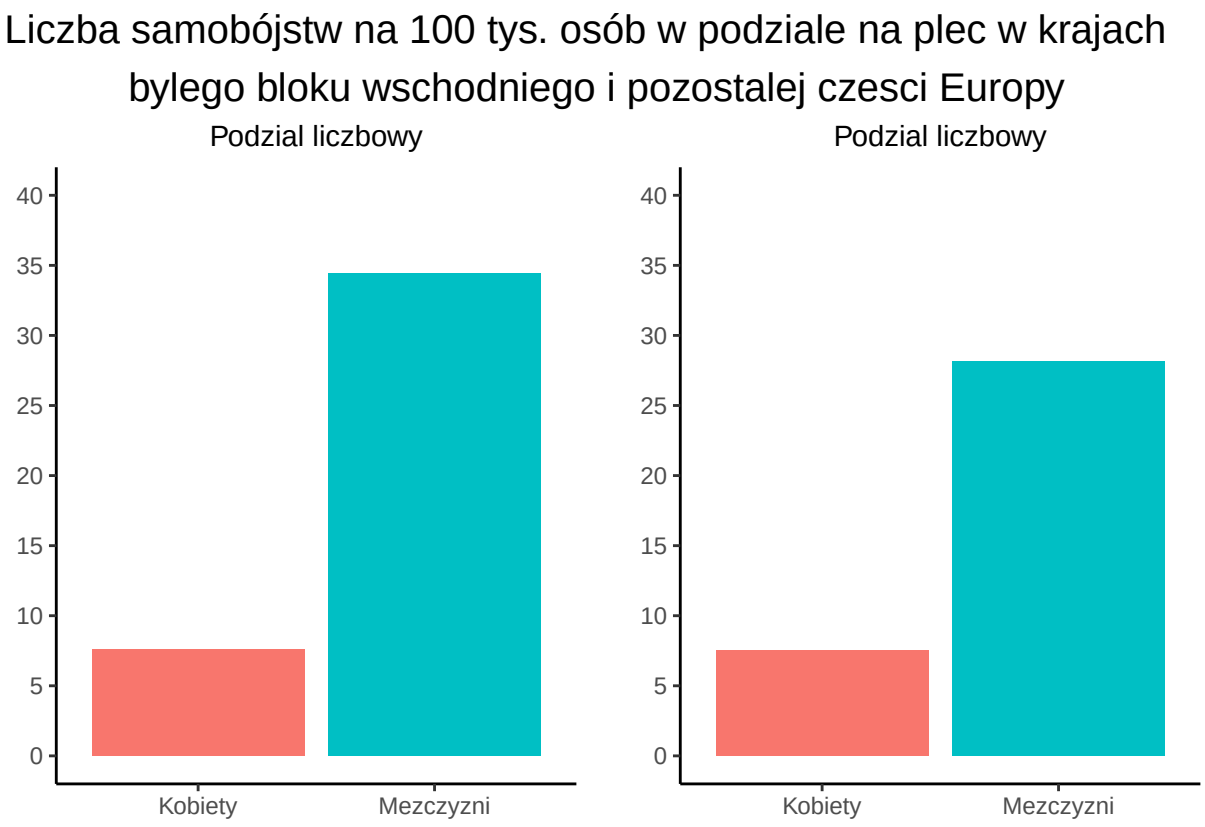
```r
non_eastern_bloc %>%
  mutate(bloc="non-eastern") -> non_eastern_bloc

rbind(eastern_bloc,non_eastern_bloc) -> blocs

blocs %>%
  group_by(year,bloc) %>%
  summarize(suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000) %>%
  ggplot(aes(x = year, y = suicide_per_100k, color=bloc)) +
  geom_line() +
  geom_point() +
  labs(title = "Liczba samobójstw na świecie (na 100 tys. osób)",
       x = "Rok",
       y = "Liczba samobójstw na 100 tys. osób",
       fill="Blok") +
  scale_x_continuous(breaks = seq(1985, 2015, 2)) +
  scale_y_continuous(breaks = seq(0, 30))
```



Powyższy wykres pokazuje, że na przestrzeni lat liczba samobójstw w krajach objętych wpływami byłego Związku Radzieckiego pozostawała wyższa niż w krajach zachodnich. Podczas gdy kraje zachodnie odnotowały spadek liczby samobójstw w latach 1994-1999, kraje zza tzw. żelaznej kurtyny odnotowały wtedy znaczący przyrost. Ponadto odnotowano stałą tendencję wzrostową liczby samobójstw w krajach wschodnich w latach następujących bezpośrednio po zmianach społecznych, które przyniósł ze sobą przełom lat 80. i 90.

Ze względu na płeć

```
eastern_bloc %>%
  group_by(sex) %>%
  summarise(suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000) %>%
  ggplot(aes(x = sex, y = suicide_per_100k, fill = sex)) +
  geom_col() +
  labs(subtitle = "Podział liczbowy",
       x = "",
       y = "") +
  scale_y_continuous(breaks = seq(0, 40, 5), limits=c(0,40))+
  theme_classic()+
  theme(legend.position = "none",plot.subtitle = element_text(hjust = 0.5)) -> eb_sex_bar

non_eastern_bloc %>%
  group_by(sex) %>%
  summarise(suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000) %>%
  ggplot(aes(x = sex, y = suicide_per_100k, fill = sex)) +
  geom_col() +
  labs(
       subtitle = "Podział liczbowy",
       x = "",
       y = "")+
  scale_y_continuous(breaks = seq(0, 40, 5), limits=c(0,40))+
  theme_classic()+
  theme(legend.position = "none",plot.subtitle = element_text(hjust = 0.5)) -> neb_sex_bar

grid.arrange(eb_sex_bar, neb_sex_bar, ncol = 2, top = textGrob("Liczba samobójstw na 100 tys. osób w po
```



Z powyższego wykresu wynika, iż mimo że liczba samobójstw wśród mężczyzn różni się w zależności od grupy państw, to liczba samobójstw wśród kobiet pozostaje na podobnym poziomie. Należy jednak zauważyć, że w obu przypadkach liczba samobójstw wśród mężczyzn znacznie przewyższa tę wśród kobiet, co wpisuje się w światowy trend.