Joanna Matuszak 255762, Joanna Wojciechowicz 255747

Sprawozdanie 1

3 kwietnia 2022

Spis treści

1. Wstęp

Analizowane dane dotyczą ilości samobójstw w zależności od roku, płci i wieku w wybranych państwach świata. Zbiór danych pochodzi ze strony Kaggle `https://www.kaggle.com/szamil/who-suicide-statistics` i został zebrany przez Światową Organizację Zdrowia WHO.

Celem analizy będzie odpowiedź na pytanie „Kto częściej popełnia samobójstwa – kobiety czy mężczyźni?". Przeprowadzimy analizę w zależności od kraju i grupy wiekowej. Na początku skupimy się na analizie całościowej - weźmiemy pod uwagę cały zbiór danych, a następnie na reprezentatywnej grupie państw Europy Zachodniej.

Aby to zrobić dokonamy wstępnej obróbki danych, w której przeanalizujemy braki danych oraz błędne wartości w zbiorze. Następnie, wykorzystując różnego rodzaju wykresy, takie jak wykres słupkowy oraz wykres liniowy, wyciągniemy wnioski, które pozwolą odpowiedzieć na postawione pytanie badawcze.

2. Wstępna obróbka danych

Zbiór danych, którego kilka początkowych wierszy zostało przedstawionych w Tabeli 2.1, zawiera zarówno zmienne kategoryczne jak i numeryczne.

	country	year	sex	age	suicides_no	population
29881	Poland	1983	female	15-24 years	120	2695500
29882	Poland	1983	female	25-34 years	175	3246500
29883	Poland	1983	female	35-54 years	246	4280400
29884	Poland	1983	female	5-14 years	8	2821000
29885	Poland	1983	female	55-74 years	200	3142500
29886	Poland	1983	female	75+ years	56	910600
29887	Poland	1983	male	15-24 years	535	2838800
29888	Poland	1983	male	25-34 years	1059	3336300
29889	Poland	1983	male	35-54 years	1331	4146300
29890	Poland	1983	male	5-14 years	45	2949800
29891	Poland	1983	male	55-74 years	650	2362300

Tabela 2.1. Prezentacja danych

Mamy 43776 obserwacji sześciu zmiennych. Każdy wiersz zawiera informacje o liczbie samobójstw i rozmiarze populacji w danym kraju dla danego roku, płci i kategorii wiekowej.

Wśród zmiennych kategorycznych mamy:

— country - zmienna opisująca, z którego kraju spośród 141 zbadanych pochodzą dane,
— sex - płeć:
  — female - kobiety,
  — male - mężczyźni,
— age - grupa wiekowa:
  — 5-14 years,
  — 15-24 years,
  — 25-34 years,
  — 35-54 years,
  — 55-74 years,
  — 75+ years.

Natomiast wśród zmiennych numerycznych wyróżniamy:

— year - rok, z którego pochodzą dane (lata od 1979 do 2016),
— suicides_no - liczba samobójstw (od 0 do 22338),
— population - liczba osobników w populacji (od 259 do 43805214).

2.1. Braki danych

W naszym zbiorze danych występuje 2256 braków danych dotyczących zmiennej suicides_no oraz 5460 dotyczących zmiennej population. Oprócz tego, nie wszystkie kraje posiadają pełne dane od 1979 do 2016 roku.

W dalszej analizie pomijamy rekordy, w których pojawia się brak danych. Decyzja ta jest uwarunkowana specyfiką danych. Na liczbę samobójstw w danym roku wpływ mogło mieć wiele czynników, dlatego ryzkownym byłoby zastępowanie braków inną wartością, na przykład średnią. W naszej opinii pominięcie wierszy z brakami danych to w tym przypadku sensowne rozwiązanie.

2.2. Błędne dane

Możemy zauważyć, że nasz zbiór danych zawiera podejrzane obserwacje, dla których przy dość dużej populacji danego kraju zmienna suicides_no tylko w niektórych latach przyjmuje wartość 0. Możliwe, że wartości te zostały błędnie wpisane i powinny być oznaczone jako NA. Niestety nie jesteśmy w stanie jednoznacznie stwerdzić, które z nich powinniśmy traktować jako brak danych, a które jako informatywne dane. Wobec tego nie wprowadzamy żadnych zmian i zakładamy, że obecne w danych zera zostały wpisany intencjonalnie.

3. Analiza danych

3.1. Analiza całego zbioru danych

Analizę rozpoczniemy od wizualizacji proporcji dotyczących płci z wykorzystaniem wszystkich danych. Z wykresu 3.1 na stronie 6 możemy odczytać, że problem samobójstw na przestrzeni prawie czterech dekad dotyczył głównie mężczyzn.

płeć	Europa	Azja	Afryka	Ameryka Płn.	Ameryka Płd.	Australia i Oceania
♀	0.24	0.25	0.24	0.21	0.22	0.23
♂	0.76	0.75	0.76	0.79	0.78	0.77

Tabela 3.1. Procentowy udział kobiet i mężczyzn w łącznej liczbie popełnionych samobójstw w zależności od kontynentu

Gdy rozpatrzymy problem w zależności od kontynentu dla badanych państw (tabela 3.1) okazuje się, że tendencja ta jest zachowana.

Analizując wykres 3.2 na stronie 6 możemy zauważyć, że w grupie wiekowej $35-54$ lata ilość samobójstw była największa. Musimy jednak pamiętać, że podział na przedziały wiekowe nie jest równomierny - występują przedziały o różnych długościach. Niemniej jednak, w każdej z rozważanych grup przeważają mężczyźni.

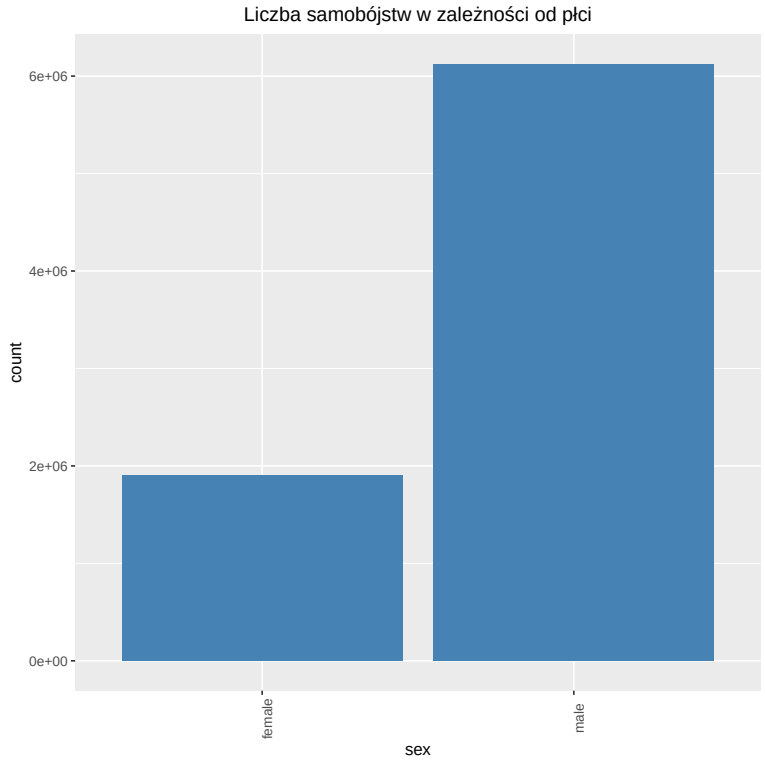
Gdy przeanalizujemy, jak zmienia się stosunek liczby samobójstw do całej populacji na przestrzeni lat (wykres 3.3 na stronie 7) okazuje się, że od mniej więcej roku 1995 pojawia się minimalna tendencja spadkowa. Możemy jednak uznać, że proporcja ta jest w miarę jednostajna.

3.2. Analiza dla wybranych państw

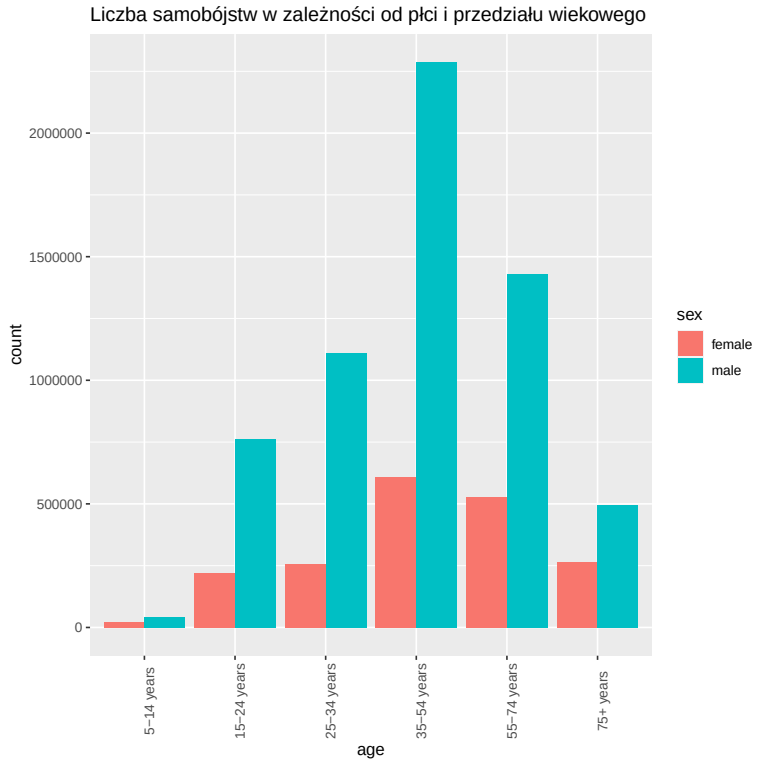
W dalszej analizie skupimy się na państwach Europy Zachodniej, spośród których wybierzemy te najbardziej reprezentatywne - nieposiadające braków danych w latach od 1979 do 2014 (pomijamy lata 2015 i 2016 z racji tego, że dane dla tych lat są niekompletne). Będą to Belgia, Francja, Irlandia, Włochy, Luksemburg, Malta oraz Holandia. Dzięki temu będziemy pracować na ciągłych danych i otrzymamy bardziej miarodajne wyniki.

Rozpoczniemy od zwizualizowania różnic między łączną liczbą samobójstw w wybranych państwach w latach 1979-2014.

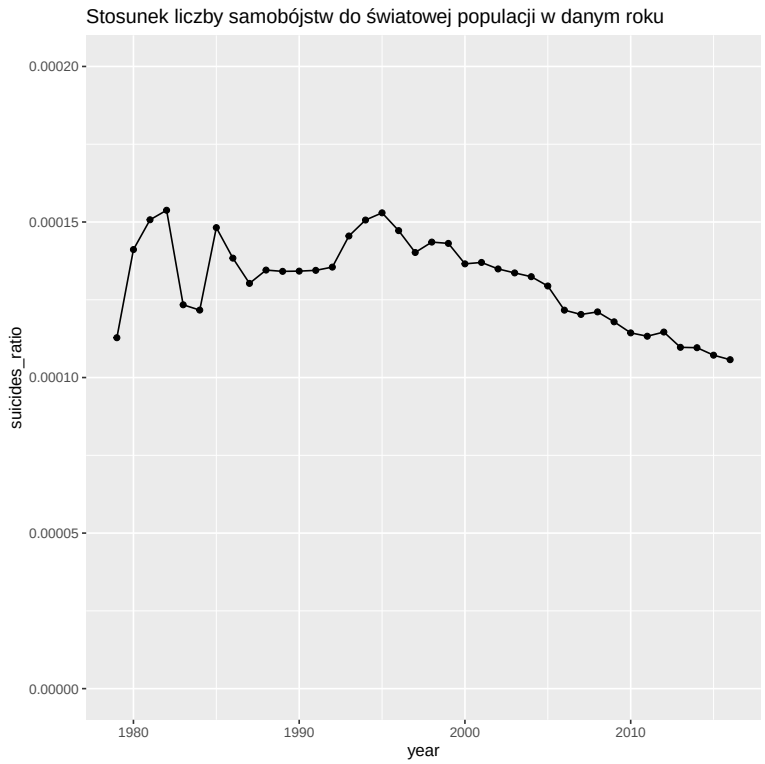
Z wykresu 3.4 na stronie 8 możemy odczytać, że w latach 1979-2014 najwięcej samobójstw zostało popełnionych we Francji. Problem jest także



Rysunek 3.1. Liczba samobójstw w zależności od płci



Rysunek 3.2. Liczba samobójstw w zależności od płci i przedziału wiekowego



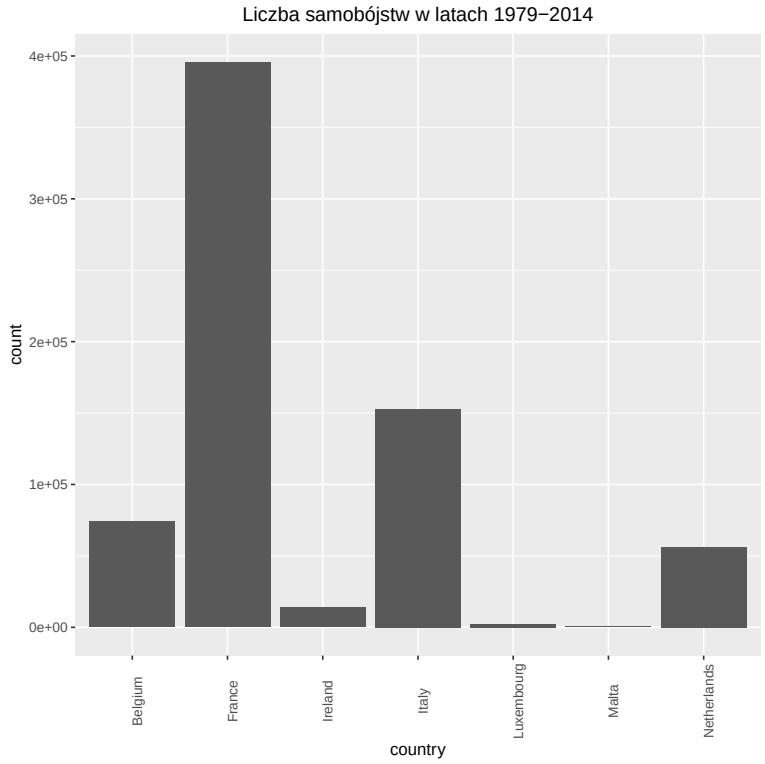
Rysunek 3.3. Stosunek liczby samobójstw do światowej populacji w danym roku

znaczący we Włoszech oraz Belgii i Holandii. W pozostałych analizowanych krajach liczba samobójstw była znacząco mniejsza.

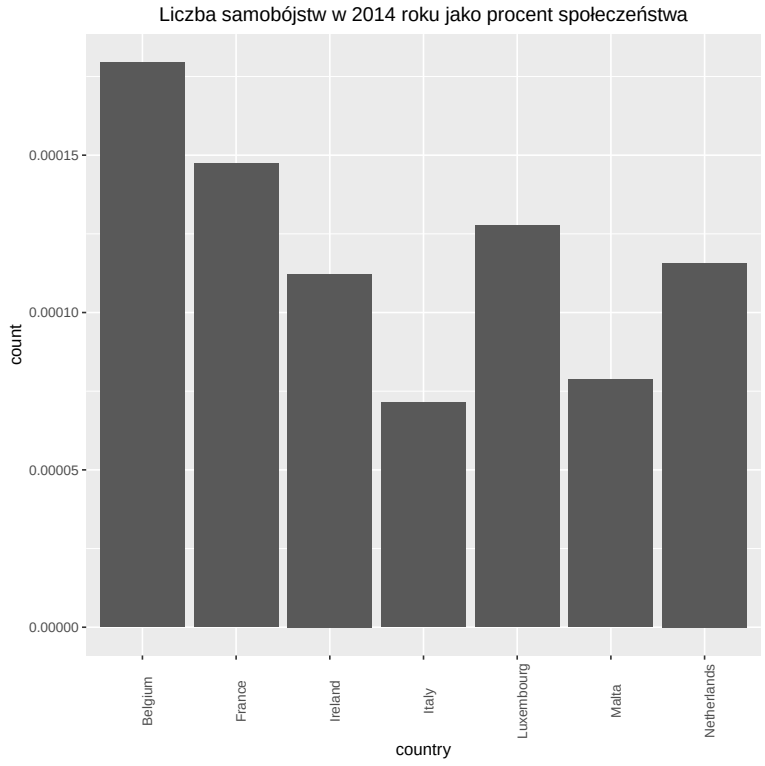
Aby uwzględnić różnice w populacjach danych państw, przedstawimy liczbę samobójstw jako procent społeczeństwa danego kraju w roku 2014. Patrząc na unormowany wykres 3.5 na stronie 8 okazuje się, że różnice te nie są aż tak drastyczne. Możemy wyciągnąć zupełnie inne wnioski - biorąc pod uwagę liczbę mieszkańców danego kraju, najlepiej wypadają Włochy i Malta, a najgorzej Belgia, Francja i Luksemburg.

Jak widzimy na wykresie 3.6 na stronie 9, problem samobójstw w latach 1979-2013 dotyczy głównie mężczyzn, bez względu na rozpatrywany kraj.

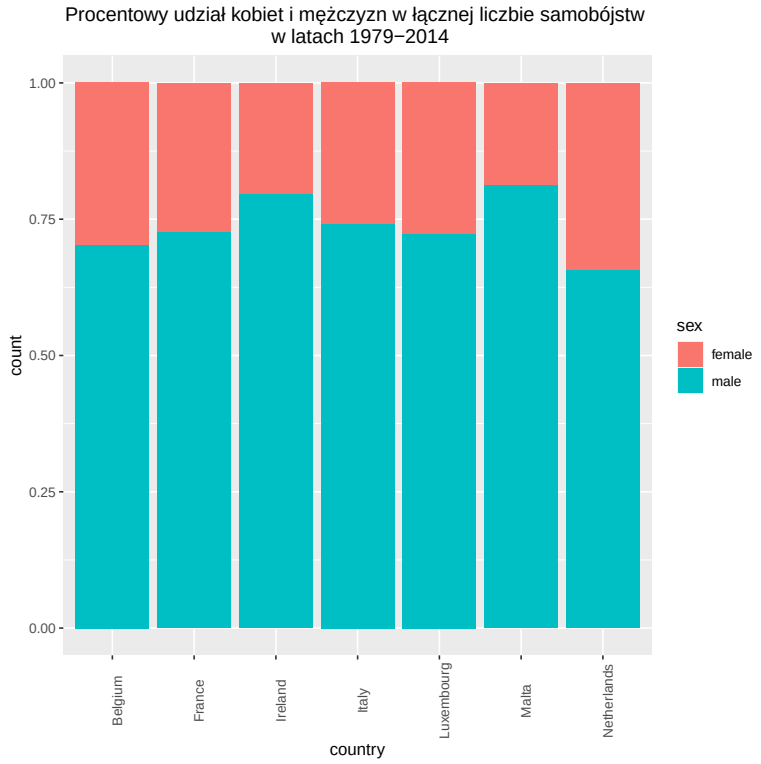
Okazuje się, że na przestrzeni lat tendencja przewagi liczby mężczyzn wśród samobójstw pogłębia się (wykres 3.7 na stronie 9) - stanowią coraz większy procent wśród samobójców rozpatrywanych krajów Europy Zachodniej.



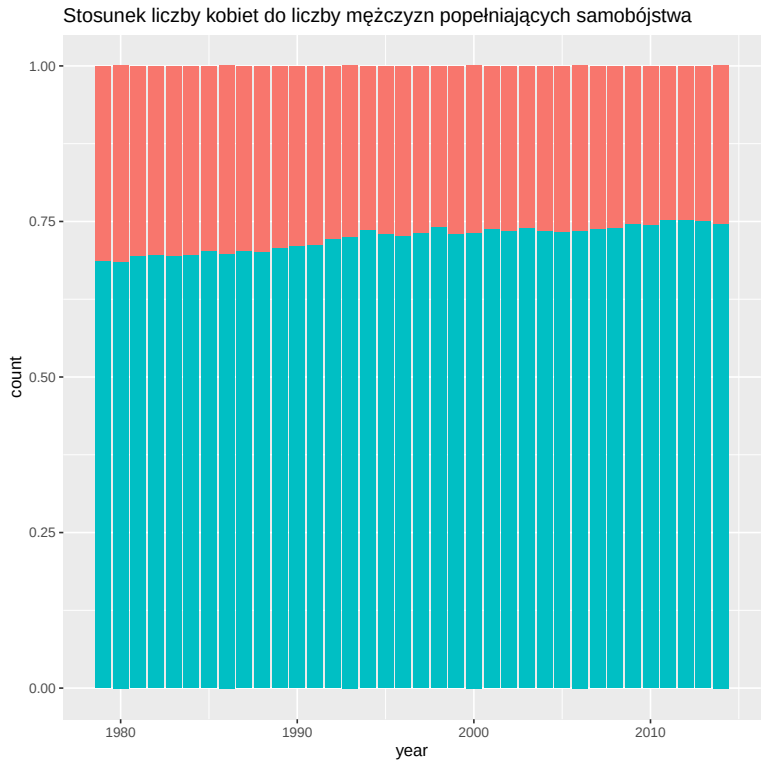
Rysunek 3.4. Liczba samobójstw w latach 1979-2014



Rysunek 3.5. Liczba samobójstw w 2014 roku jako procent społeczeństwa



Rysunek 3.6. Procentowy udział kobiet i mężczyzn w łącznej liczbie samobójstw w latach 1979-2014



Rysunek 3.7. Stosunek liczby kobiet do liczby mężczyzn popełniających samobójstwa

4. Podsumowanie

Analiza całościowa danych po obsłudze braków danych oraz analiza grupy krajów reprezentatywnych (pod względem zachowanej ciągłości danych) pozwoliła na wyciągnięcie następujących wniosków:

— znacząco więcej samobójców jest mężczyznami,
— tendencja do przewagi mężczyzn nie zależy znacząco od kontynentu ani grupy wiekowej,
— we wszystkich przeanalizowanych krajach przewaga mężczyzn nad kobietami wśród samobójców zostaje zachowana,
— w rozważanych państwach Europy Zachodniej udział procentowy mężczyzn w samobójstwach nieznacznie rośnie na przestrzeni lat.

Nawiązując do pytania badawczego, mężczyźni częściej popełniają samobójstwa, a tendencja ta nie zależy znacząco od pozostałych analizowanych zmiennych.