

Długość życia w poszczególnych państwach Europy (2017-2021)

Przemysław Popowski

26 styczeń 2024r.

Punkt 1 - Opis projektu

Tematem projektu jest podstawowa analiza danych średniej długości życia Europejczyków w latach 2017-2021. W dokumentacji opisałem sposób zbierania danych do bazy danych, czyszczenia i przygotowania ich do potrzeb badań, pokazałem różne wizualizacje oraz przeprowadziłem ich analizę. Główną myślą, która kierowała mną do takiego tematu, była chęć zobaczenia wpływu pandemii COVID-19 na długość życia w Europie.

Punkt 2 - Zebranie danych do bazy danych

Dane pozyskałem z poniższego źródła:

https://ec.europa.eu/eurostat/databrowser/view/demo_r_mlifexp/default/table?lang=en&category=demo.demomreg

Obejmują one średnią długość życia w 37 państwach Europy w latach 2017-2021.

Punkt 2.1 - Struktura bazy danych

Panstwo <chr>	Wszyscy2021 <chr>	Mezcyżni2021 <chr>	Kobiety2021 <chr>	Wszyscy2020 <chr>	Mezcyżni2020 <chr>	Kobiety2020 <chr>
Belgium	81,9	79,4	84,3	80,8	78,5	83,0
Bulgaria	71,4	68,0	75,1	73,6	70,0	77,5
Czechia	77,2	74,1	80,5	78,2	75,2	81,3
Denmark	81,5	79,6	83,3	81,6	79,7	83,6
Germany	80,8	78,4	83,3	81,1	78,7	83,5
Estonia	77,2	72,7	81,4	78,9	74,4	83,0
Ireland	82,4	80,5	84,3	82,6	80,8	84,4
Greece	80,2	77,4	82,9	81,4	78,8	83,9
Spain	83,3	80,4	86,2	82,4	79,6	85,2
France	82,4	79,3	85,5	82,3	79,2	85,3

1-10 of 37 rows | 1-7 of 16 columns

Previous 1 2 3 4 Next

Punkt 2.2 - Czyszczenie danych

Podane na stronie dane były już wstępnie przygotowane do dalszej analizy, nastąpiły jednak wyjątki w postaci dwóch państw:

- W przypadku Wielkiej Brytanii badania obowiązują tylko do 2018 roku.
- Natomiast w Turcji, posiadamy statystyki tylko do 2019 roku.

Czyszczenie danych będzie polegało na usunięciu znaków “:”, które oznaczają brak danych i zastąpienie ich wartościami “NA”. Dodatkowo musimy zamienić znaki “,” oddzielające liczbę całości od ułamka na “.”. Na koniec zostanie zmienić typ danych w bazie z “char” na “double” dla liczb.

```
clear_data <- data %>% mutate_if(is.character, ~ ifelse(is.na(.), NA, gsub(",", ".", gsub(":", "null", .))))
clear_data <- clear_data %>% mutate(across(-1,
  as.numeric))
```

Punkt 2.3 - Struktura bazy danych po dokonanych czyszczeniu

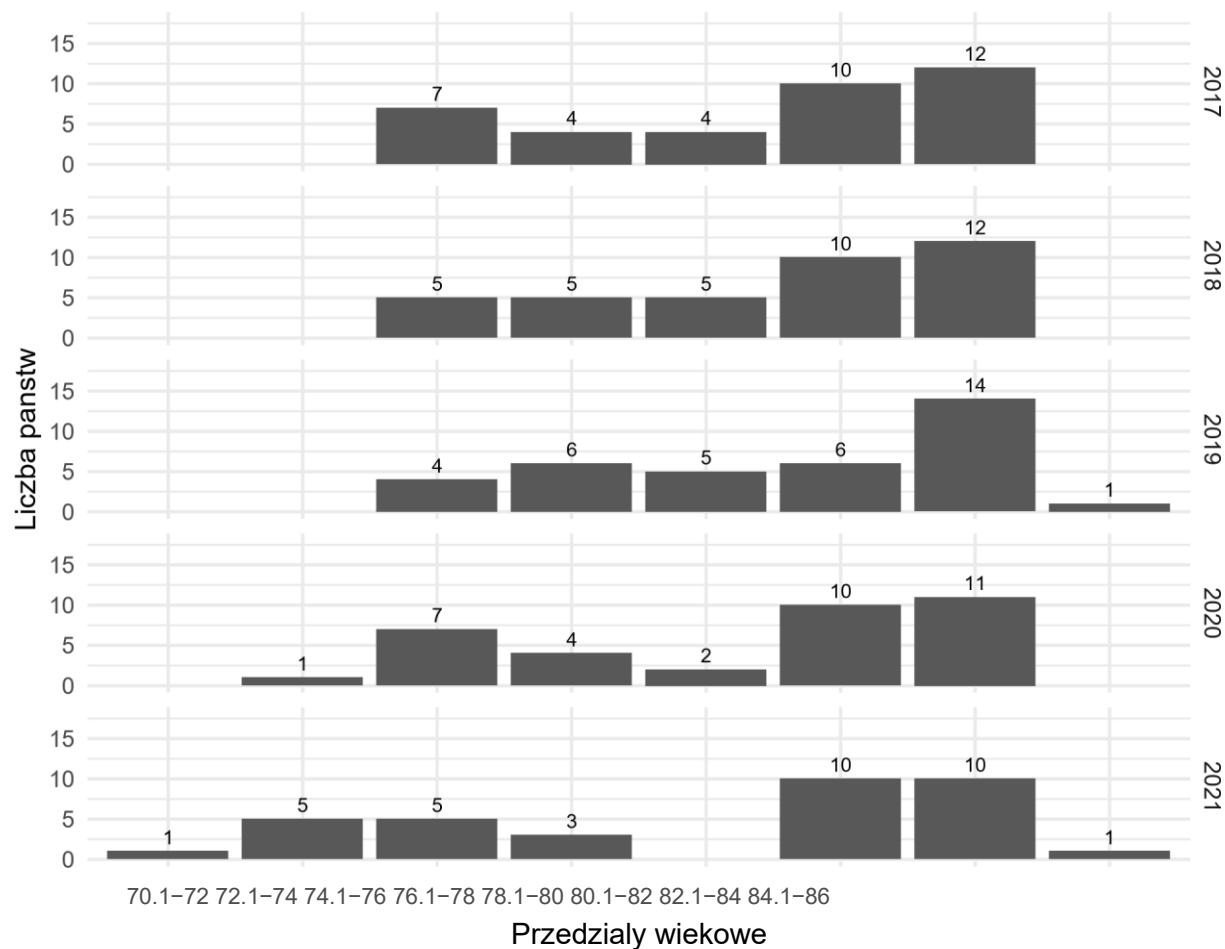
Panstwo <chr>	Wszyscy2021 <dbl>	Meczczyzni2021 <dbl>	Kobiety2021 <dbl>	Wszyscy2020 <dbl>	Meczczyzni2020 <dbl>	Kobiety2020 <dbl>
Belgium	81.9	79.4	84.3	80.8	78.5	83.0
Bulgaria	71.4	68.0	75.1	73.6	70.0	77.5
Czechia	77.2	74.1	80.5	78.2	75.2	81.3
Denmark	81.5	79.6	83.3	81.6	79.7	83.6
Germany	80.8	78.4	83.3	81.1	78.7	83.5
Estonia	77.2	72.7	81.4	78.9	74.4	83.0
Ireland	82.4	80.5	84.3	82.6	80.8	84.4
Greece	80.2	77.4	82.9	81.4	78.8	83.9
Spain	83.3	80.4	86.2	82.4	79.6	85.2
France	82.4	79.3	85.5	82.3	79.2	85.3

1-10 of 37 rows | 1-7 of 16 columns

Previous 1 2 3 4 Next

Punkt 3 - Analiza eksploracyjna

Punkt 3.1 - Przeanalizujemy przedziały wiekowe średniej długości życia obu płci na liczbę państw z podziałem na rok



Na pierwszy rzut oka widać, że na przestrzeni lat 2019-2021 pojawiło się więcej wartości w słupkach reprezentujących przedziały wiekowe poniżej 78 roku życia.

Zbadajmy jeszcze wskaźniki:

Wartość najmniejsza, dolny kwartył, mediana, średnia, górny kwartył, wartość największa:

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	71.4	74.6	81.3	79.1	82.5	84.4

Wariancja:

[1] 16.38649

Odchylenie standardowe:

[1] 4.048022

Odchylenie przeciętne:

```
## [1] 2.81694
```

Zakres:

```
## [1] 71.4 84.4
```

Rozstęp kwartyłowy:

```
## [1] 7.9
```

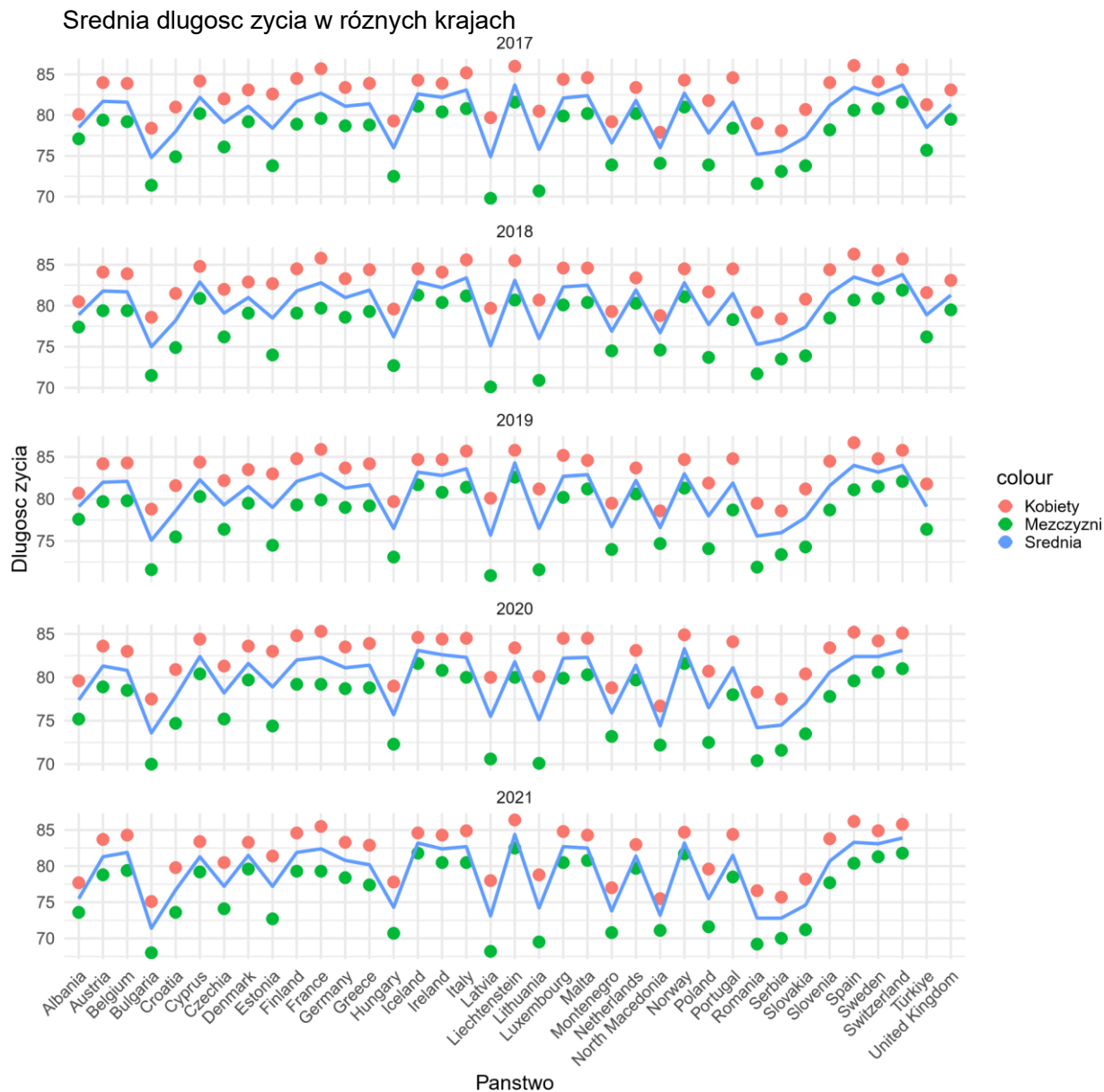
Moda:

```
## [1] 72.8 75.5 77.2 81.3 81.5 81.9 82.4 82.7 83.2
```

```
## attr("freq")
```

```
## [1] 10
```

Punkt 3.2 - Liczba wystąpień poszczególnych średnich długości życia z podziałem na rok, płeć i państwo



Po dokładnym przeanalizowaniu tych wykresów możemy zaobserwować spadki średniej długości życia w niektórych państwach Europy. Najbardziej widoczne są zmiany na wykresie z roku 2020. Pandemia najmocniej poruszyła Liechtenstein, Północną Macedonię oraz Litwę.

Na sam koniec zbadajmy wskaźniki:

Wartość najmniejsza, dolny kwartyl, mediana, średnia, górny kwartyl, wartość największa:

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	68.00	73.90	78.80	77.16	80.30	82.60

Wariancja:

```
## [1] 14.66578
```

Odchylenie standardowe:

```
## [1] 3.829593
```

Odchylenie przeciętne:

```
## [1] 3.40998
```

Zakres:

```
## [1] 68.0 82.6
```

Rozstęp kwartyłowy:

```
## [1] 6.4
```

Moda:

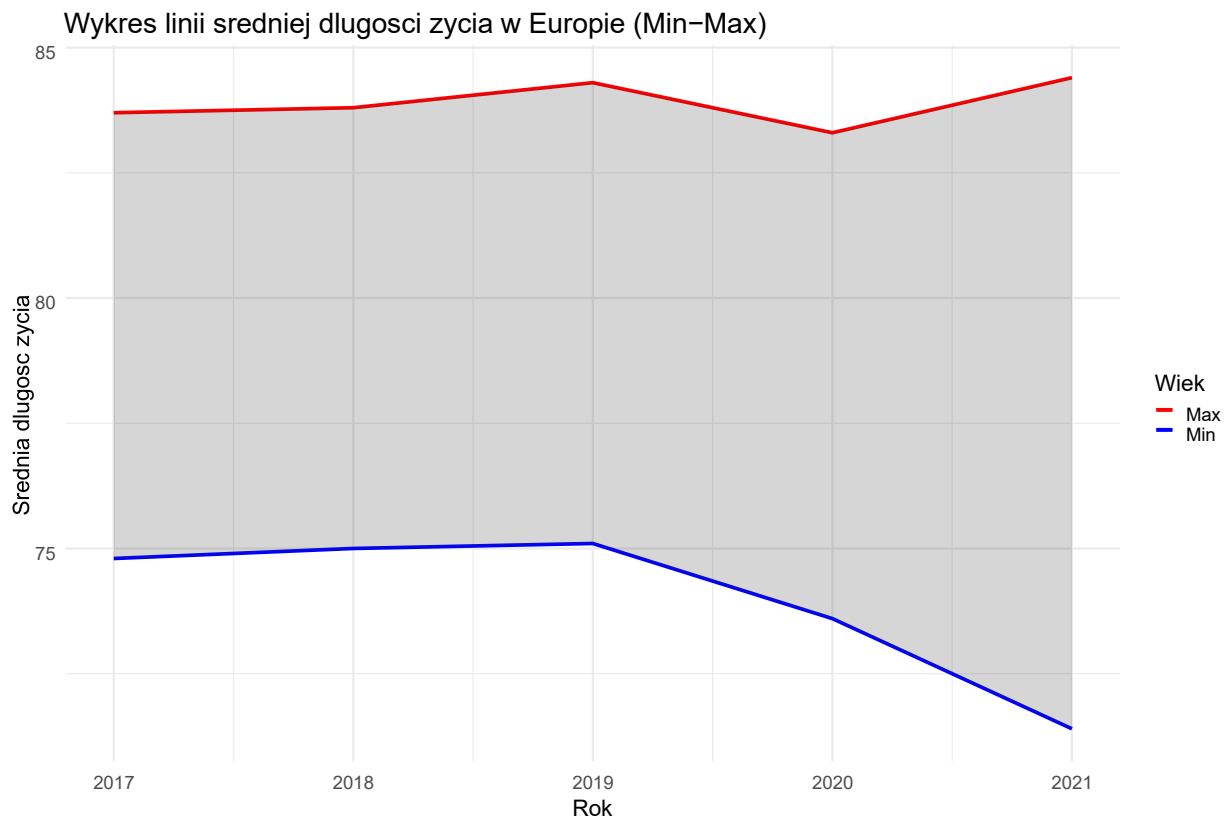
```
## [1] 79.2
```

```
## attr("freq")
```

```
## [1] 6
```

Punkt 4 - Analiza zależności zmiennych

Punkt 4.1 - Zmiana średniej długości życia na przestrzeni lat 2017-2021

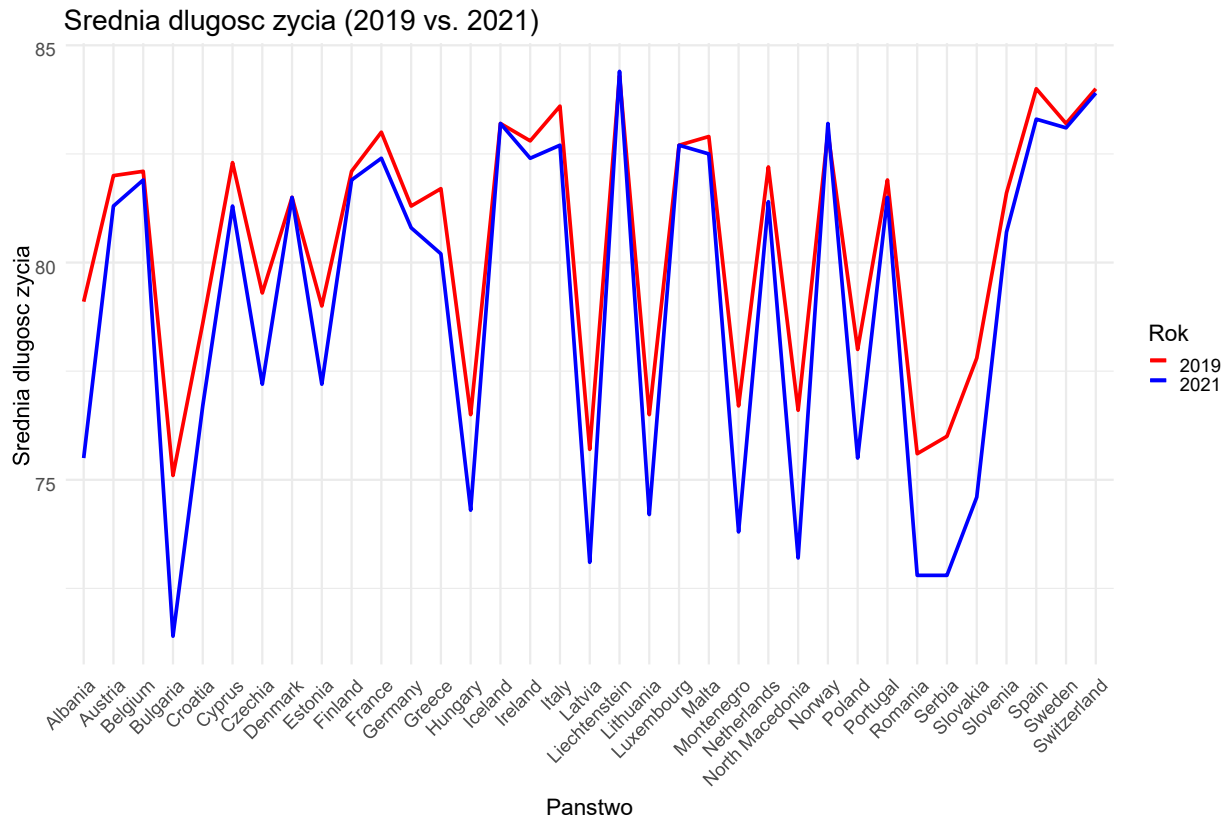


Wniosek:

Zauważyć możemy duży spadek wartości maksymalnej w 2020 roku, która odpowiada nam za największą średnią długość życia w Europie, oraz progresyjny spadek wartości minimalnej od roku 2019, odpowiadającej najmniejszej długości życia od roku 2019.

Myśląc o podanych latach, nasuwa nam się od razu pandemia COVID-19. To ona musiała mieć wpływ na tak gwałtowne różnice w średnim wieku ludności Europy.

Punkt 4.2 - Wpływ epidemii COVID-19 na średnią długość życia w poszczególnych państwach (różnica między 2021 a 2019 rokiem)



Wniosek:

Na tym wykresie możemy idealnie zaobserwować prawdziwe skutki epidemii. Bardzo mocno dotknęła długość życia obywateli Bułgarii, Rumunii, Serbii oraz Słowacji. Potwierdziło się moje założenie, że w ciągu ostatnich lat, największy wpływ na regresję średniej długości życia w Europie wpłynął COVID-19.

Różnice w wymienionych wyżej państwach (w latach):

Bułgaria

[1] -3.7

Rumunia

[1] -2.8

Serbia

[1] -3.2

Słowacja

[1] -3.2

Test t-studenta dla par zależnych (obserwacja tych samych państw w latach 2019 i 2021):

##

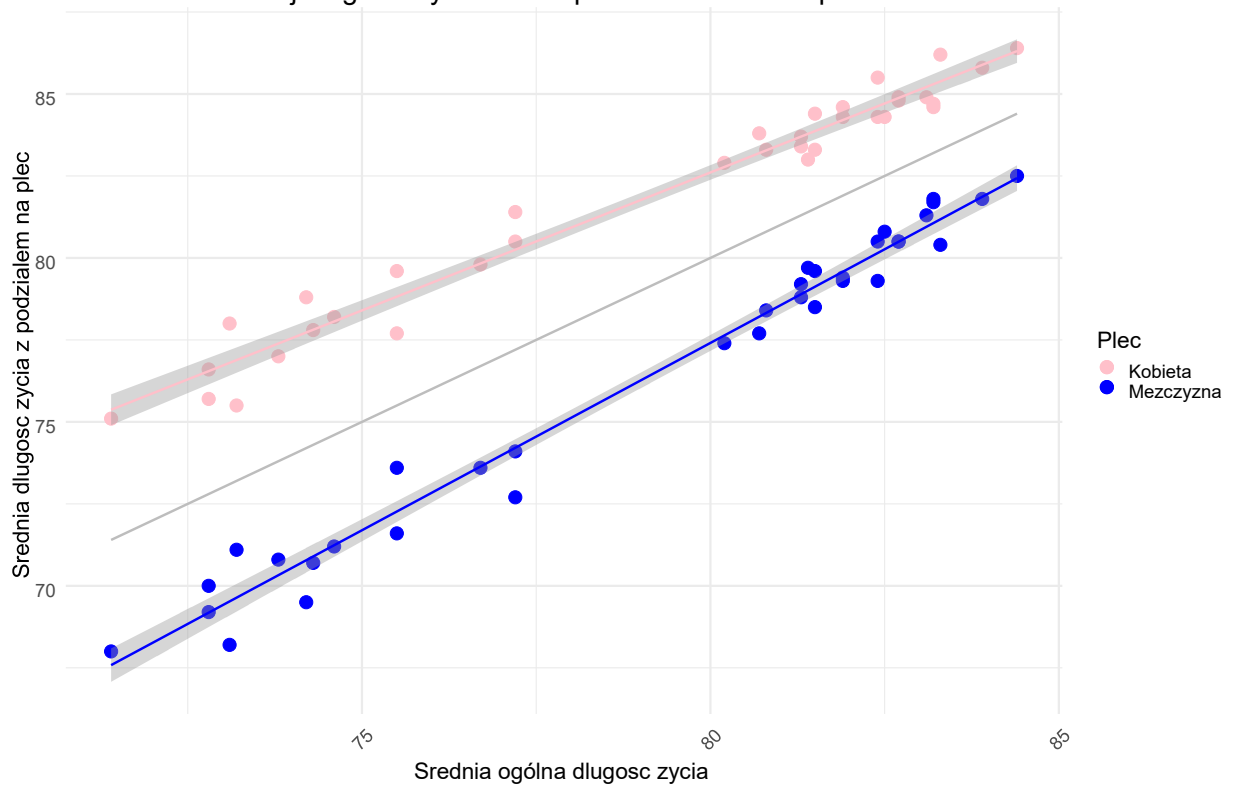
Paired t-test


```
##
## data: data_covid$DlugoscZycia2021 and data_covid$DlugoscZycia2019
## t = -6.3948, df = 34, p-value = 2.658e-07
## alternative hypothesis: true mean difference is not equal to 0 ## 95 percent
confidence interval: ## -1.7809092 -0.9219479 ## sample estimates:
## mean difference
## -1.351429
```

Dzięki temu testowi możemy zauważyć, że średnia różnica długości życia w Europie między rokiem 2019 a 2021 wyniosła około -1.35 roku.

Punkt 4.3 - Zależność średniej długości życia w Europie w 2021 roku od płci

Zależność średniej długości życia w Europie w 2021 roku od płci



Przeprowadzimy test t-studenta dla par zależnych między długością życia mężczyzn a ogólną średnią długością życia.

```
##
## Paired t-test
##
## data: data_gender2021$Mężczyzni and data_gender2021$Ogólnie
## t = -17.999, df = 34, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0 ## 95 percent
confidence interval: ## -3.027111 -2.412889 ## sample estimates:
## mean difference
## -2.72
```

Jak możemy zauważyć, średnia różnica długości życia mężczyzn różni się od średniej długości życia w Europie o - 2.72 roku.

A teraz test t-studenta dla par zależnych między długością życia kobiet a ogólną średnią długością życia.

```
##
## Paired t-test
##
## data: data_gender2021$Kobiety and data_gender2021$Ogolnie
## t = 18.052, df = 34, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0 ## 95 percent
confidence interval: ## 2.439145 3.057998 ## sample estimates:
## mean difference
##          2.748571
```

Jak możemy zauważyć, średnia różnica długości życia kobiet różni się od średniej długości życia w Europie o około +2.75 roku.

Aby potwierdzić jeszcze liniową zależność wykresów, wykonałem testy korelacji:

- Mężczyzn:

```
##
## Pearson's product-moment correlation
##
## data: data_gender2021$Ogolnie and data_gender2021$Meczczyzni
## t = 39.686, df = 33, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0 ## 95 percent
confidence interval: ## 0.9794811 0.9948285 ## sample estimates:
##          cor
## 0.9896856
```

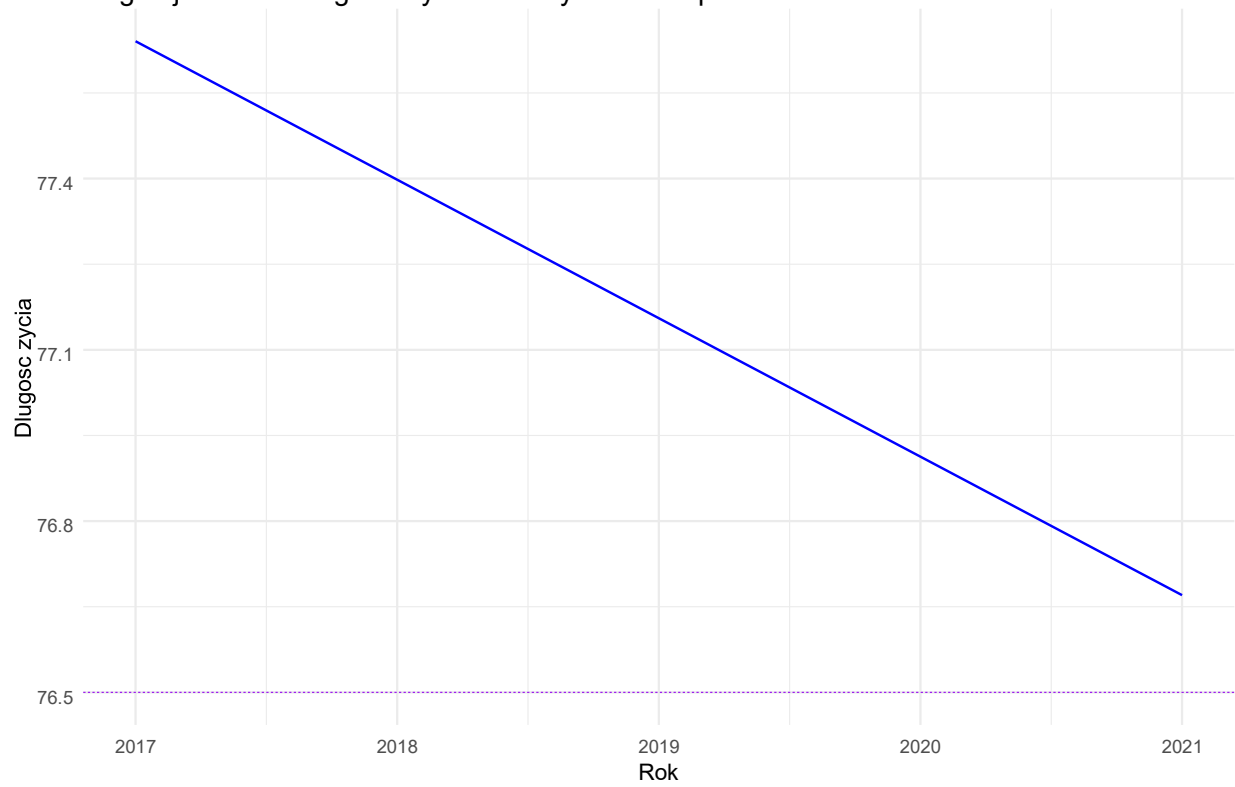
- Kobiet:

```
##
## Pearson's product-moment correlation
##
## data: data_gender2021$Ogolnie and data_gender2021$Kobiety
## t = 31.706, df = 33, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval: ##
0.9682195 0.9919559 ## sample estimates:
##          cor
## 0.9839794
```

Wyniki testu potwierdzają, że zależność między dotychczasową średnią długością życia mężczyzn, a poszczególnymi punktami jest wręcz idealnie liniowa. Oszacowany współczynnik korelacji wynosi prawie 0.99, a 95% przedział ufności - od 0.979 do 0.994. Natomiast w przypadku kobiet współczynnik osiągnął wartość lekko ponad 0.98, przy również 95% przedziale ufności będącym od 0.968 do 0.991, zatem również jest to prawie liniowa zależność.

Punkt 4.4 - Regresja liniowa średniej długości życia mężczyzn w Europie w latach 2017-2021

Regresja liniowa długości życia mężczyzn w Europie

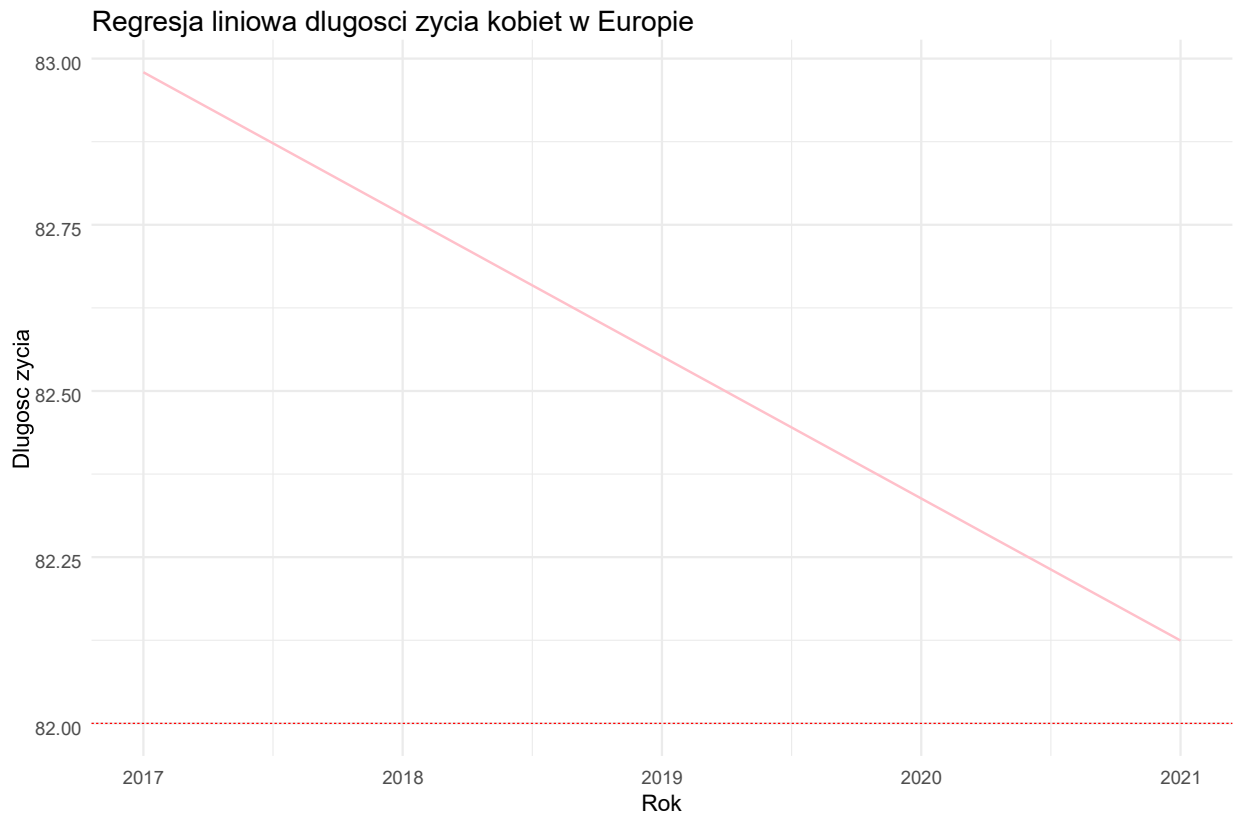


```
## as.numeric(Rok)
```

```
## -0.2425714
```

Średni spadek długości życia mężczyzn na rok. Przy takiej wartości możemy zauważyć, że mniej więcej na początku 2022 roku, średnia długość życia mężczyzn w Europie wyniosła by już mniej niż 76.5 roku, a w 2024 roku byłoby to niecałe 76 lat. (Jestem szczerze ciekaw ile aktualnie wynosi, gdyż robię ten projekt w 2024 roku :D)

Punkt 4.5 - Regresja liniowa średniej długości życia kobiet w Europie w latach 2017-2021



```
## as.numeric(Rok)
##      -0.2137143
```

Średni spadek długości życia kobiet na rok. Przy takiej wartości możemy zauważyć, że mniej więcej na początku kwietnia 2022 roku, średnia długość życia kobiet w Europie wyniosła by już mniej niż 82 lata.

Punkt 5 - Podsumowanie

Możemy dojść do konkluzji, że faktycznie pandemia COVID-19 miała duży wpływ na spadek średniej długości życia w Europie. Niestety dane, które są aktualnie dostępne, nie są wystarczające, aby stwierdzić, czy regresja będzie się utrzymywała przez następne lata.

Dzięki użyciu środowiska R oraz nabytej wiedzy podczas całego przedmiotu "Rachunek prawdopodobieństwa i statystyka", byłem w stanie odpowiedzieć na kilka nurtujących mnie od liceum pytań związanych z długością życia w Europie. Zyskałem także podstawową wiedzę z praktycznego użytkowania języka R do rozwiązywania problemów statystycznych.

Wykorzystane biblioteki: DBI - łączenie się z bazą danych, knitr - dynamiczne generowanie raportów, rmarkdown - stworzenie całego notebooka, dplyr i tidyr - ułatwienie modyfikacji danych oraz pomoc w zachowaniu czystego i przejrzystego kodu, ggplot2 - rysowanie wykresów, DescTools - wykorzystany do wyliczania mody