ANALIZA EKSPLORACYJNA DANYCH

Oczekiwana długość życia - Dane WHO

Projekt: Metody Eksploracji Danych

Autorzy:

Paweł Myszka
Stanisław Dutkiewicz
Łukasz Jugo

Grudzień 2025

Spis treści

1 Opis danych w strukturze **3**
1.1 Wymiary zbioru danych . . . 3
1.2 Zmienne w zbiorze danych . . . 3
1.3 Zmienna docelowa . . . 3
1.4 Rozkład brakujących wartości . . . 3
1.5 Analiza . . . 4

2 Analiza zależności - korelacje zmiennych **4**
2.1 Wprowadzenie . . . 4
2.2 Top 10 predyktorów . . . 5
2.3 Interpretacja wyników . . . 5
2.3.1 Zmienne o silnej korelacji . . . 5
2.3.2 Zmienne o umiarkowanej korelacji . . . 6
2.4 Podsumowanie . . . 6

3 Zakresy i stopień zmienności zmiennych **6**
3.1 Zmienne o ekstremalnej zmienności (CV $>$ 200%) . . . 7
3.2 Zmienne o wysokiej zmienności (100% $<$ CV $<$ 200%) . . . 7
3.3 Zmienne o umiarkowanej zmienności (50% $<$ CV $<$ 100%) . . . 7
3.4 Zmienne o niskiej zmienności (CV $<$ 50%) . . . 8
3.5 Charakterystyka zmiennej docelowej . . . 8

4 Braki danych- stopień wypełnienia **8**
4.1 Analiza brakujących danych . . . 8
4.2 Kompletne obserwacje . . . 9

5 Rozkłady zmiennych i normalność rozkładu **9**
5.1 Histogram zmiennej docelowej . . . 9
5.2 Normalność rozkładu – Q-Q plot . . . 9
5.3 Box plot – zmienne z największą zmiennością . . . 9

6 Hipoteza badawcza **10**

7 Wnioski **10**

1 Opis danych w strukturze

1.1 Wymiary zbioru danych

Zbiór danych WHO dotyczący oczekiwanej długości życia zawiera:

- **Liczba wierszy:** 2938 obserwacji
- **Liczba kolumn:** 22 zmienne
- **Zmienne numeryczne:** 20
- **Zmienne kategorialne:** 2 (Country, Status)

1.2 Zmienne w zbiorze danych

Lp.	Nazwa zmiennej	Typ danych
1	Country	object
2	Year	int64
3	Status	object
4	Life expectancy	float64
5	Adult Mortality	float64
6	infant deaths	int64
7	Alcohol	float64
8	percentage expenditure	float64
9	Hepatitis B	float64
10	Measles	int64
11	BMI	float64
12	under-five deaths	int64
13	Polio	float64
14	Total expenditure	float64
15	Diphtheria	float64
16	HIV/AIDS	float64
17	GDP	float64
18	Population	float64
19	thinness 1-19 years	float64
20	thinness 5-9 years	float64
21	Income composition of resources	float64
22	Schooling	float64

Tabela 1: Zmienne w zbiorze danych WHO Life Expectancy

1.3 Zmienna docelowa

Life expectancy (Oczekiwana długość życia w latach) jest zmienną objaśnianą w analizie. Pozostałe 19 zmiennych numerycznych pełnią rolę zmiennych objaśniających.

1.4 Rozkład brakujących wartości

Tak przygotowane dane wymagają obróbki braków wartości przed modelowaniem.

Zmienna	Non-Null Count	% Kompletności
Country	2938	100.0%
Year	2938	100.0%
Status	2938	100.0%
infant deaths	2938	100.0%
Measles	2938	100.0%
percentage expenditure	2938	100.0%
under-five deaths	2938	100.0%
HIV/AIDS	2938	100.0%
Life expectancy	2928	99.7%
Adult Mortality	2928	99.7%
Alcohol	2744	93.4%
Hepatitis B	2385	81.2%
BMI	2904	98.8%
GDP	2490	84.8%
Population	2286	77.8%
Total expenditure	2712	92.3%

Tabela 2: Kompletność danych dla głównych zmiennych

1.5 Analiza

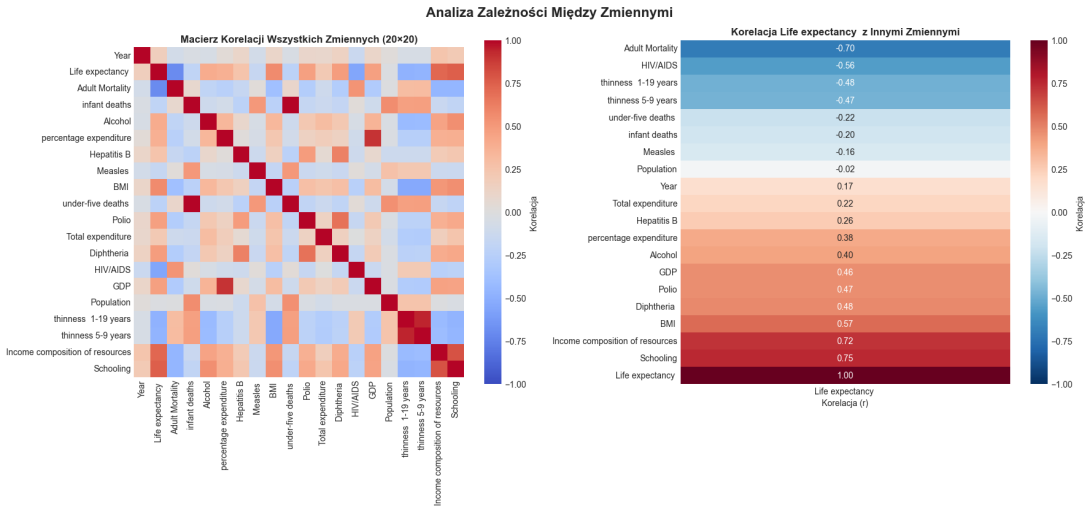
Zbiór obejmuje dane z 161 krajów za lata 2000-2015. Struktura jest regularna - każda obserwacja reprezentuje kraj-rok. Zmienne obejmują wskaźniki zdrowotne (śmiertelność, choroby zakaźne, HIV/AIDS), ekonomiczne (PKB, ekspozycja zdrowotna, skład dochodów) i społeczne (szkolnictwo, alkohol).

2 Analiza zależności - korelacje zmiennych

2.1 Wprowadzenie

W ramach tej analizy obliczono dwie macierze korelacji Pearsona:

1. **Macierz korelacji wszystkich zmiennych** — Macierz o wymiarach (20×20) prezentująca powiązania między wszystkimi 20 zmiennymi numerycznymi.
2. **Macierz korelacji zmiennej docelowej ze zmiennymi objaśniającymi** — Wektor korelacji (20×1) pokazujący siłę i kierunek powiązań każdej zmiennej ze zmienną docelową (Life expectancy). Macierz ta jest podstawą do identyfikacji zmiennych o największym wpływie na oczekiwaną długość życia.



Rysunek 1: Macierz korelacji zmiennych i korelacje ze zmienną docelową Life expectancy

2.2 Top 10 predyktorów

Tabela 3 prezentuje dziesięć zmiennych o największej bezwzględnej wartości korelacji ze zmienną docelową, uporządkowanych malejąco.

Tabela 3: Top 10 predyktorów zmiennej Life expectancy wg siły korelacji

Zmienna	Korelacja (r)	Siła	Kierunek
Schooling	+0,752	SILNA	dodatnia
Income composition of resources	+0,725	SILNA	dodatnia
Adult Mortality	-0,696	UMIARKOWANA	ujemna
BMI	+0,568	UMIARKOWANA	dodatnia
HIV/AIDS	-0,557	UMIARKOWANA	ujemna
Diphtheria	+0,479	UMIARKOWANA	dodatnia
thinness 1-19 years	-0,477	UMIARKOWANA	ujemna
thinness 5-9 years	-0,472	UMIARKOWANA	ujemna
Polio	+0,466	UMIARKOWANA	dodatnia
GDP	+0,461	UMIARKOWANA	dodatnia

2.3 Interpretacja wyników

2.3.1 Zmienne o silnej korelacji

Dwie zmienne wykazują silne powiązanie ze zmienną docelową ($|r| > 0,70$):

- **Schooling** ($r = +0,752$) — Lata edukacji przypadające na mieszkańca wykazują najsilniejszą dodatnią korelację z oczekiwaną długością życia. Wyższe poziomy wykształcenia społeczeństwa są powiązane ze zwiększoną średnią długością życia, co sugeruje, że edukacja jest jednym z najważniejszych determinantów zdrowia populacji.

- **Income composition of resources** ($r = +0,725$) — Skład dochodów kraju (proporcja dochodu wydawanego na jedzenie, mieszkanie itp.) wykazuje drugie miejsce

pod względem siły korelacji. Lepsze zasoby ekonomiczne i ich właściwy przydział są istotnie powiązane z wyższą długością życia.

2.3.2 Zmienne o umiarkowanej korelacji

Osiem zmiennych wykazuje umiarkowaną korelację ($0,40 < |r| < 0,70$) ze zmienną docelową:

- **Adult Mortality** ($r = -0,696$) — Ujemna korelacja wskazuje, że wyższa śmiertelność dorosłych jest związana z niższą oczekiwaną długością życia.
- **BMI** ($r = +0,568$) — Średni indeks masy ciała populacji wykazuje dodatnią korelację, sugerując, że populacje z wyższym średnim BMI mają dłuższą średnią długość życia.
- **HIV/AIDS** ($r = -0,557$) — Ujemna korelacja odzwierciedla fakt, że wyższe rozpowszechnienie HIV/AIDS negatywnie wpływa na oczekiwaną długość życia.
- **Wskaźniki wyszczepienia** (Diphtheria, Polio) — Dodatnie korelacje wskaźników wyszczepienia ($r \approx +0,47/+0,48$) odzwierciedlają dostęp do opieki medycznej i profilaktyki zdrowotnej.
- **Niedowaga wśród dzieci** (thinness 1-19, thinness 5-9 years) — Obie zmienne wykazują ujemne korelacje ($r \approx -0,47/-0,48$), wskazując na związek niedowagi z niższą oczekiwaną długością życia.
- **GDP** ($r = +0,461$) — Produkt krajowy brutto kraju wykazuje dodatnią korelację, co sugeruje, że bogatsza gospodarki osiągają wyższe średnie długości życia.

2.4 Podsumowanie

Przeprowadzona analiza korelacji ujawnia, że najważniejszymi determinantami oczekiwanej długości życia są:

1. **Czynniki społeczno-ekonomiczne** — edukacja i dochody populacji wykazują najsilniejsze powiązania ze zmienną docelową
2. **Czynniki zdrowotne** — dostęp do opieki medycznej (wyszczepienia), rozpowszechnienie chorób zakaźnych (HIV/AIDS), oraz wskaźniki zdrowia populacji
3. **Zasobność ekonomiczna** — PKB kraju wpływa istotnie na możliwość zapewnienia opieki zdrowotnej

3 Zakresy i stopień zmienności zmiennych

W ramach tego etapu obliczono podstawowe statystyki opisowe dla każdej zmiennej, w tym miary tendencji centralnej (średnia, mediana) oraz miary rozproszenia (odchylenie standardowe, współczynnik zmienności). Współczynnik zmienności (CV%) wyrażony wzorem:

$$CV\% = \frac{\sigma}{\mu} \times 100$$

gdzie σ oznacza odchylenie standardowe, a μ średnią, pozwala na porównanie zmienności między zmiennymi o różnych skalach pomiarowych.

3.1 Zmienne o ekstremalnej zmienności (CV > 200%)

Cztery zmienne wykazują niezwykle wysoką zmienność, co wskazuje na ekstremalne różnice między krajami:

- **Population** (CV = 478,40%) — Liczba ludności kraju wykazuje ogromny rozrzut od 34 do 1,293,859,294 osób. Ta ekstrema zmienność odzwierciedla fundamentalnie różne rozmiary populacji krajów w zbiorze danych.
- **Measles** (CV = 473,93%) — Liczba zarejestrowanych przypadków odry charakteryzuje się ekstremalnymi wartościami, wahając się od 0 do 212,183 przypadków. Takie rozprzestrzenie wskazuje na ogromne różnice w rozpowszechnieniu tej choroby zakaźnej.
- **infant deaths** (CV = 389,15%) — Liczba zgonów niemowląt (min: 0, max: 1,800) wykazuje znaczące dysproporcje między krajami, odzwierciedlające różnice w dostępie do opieki zdrowotnej i poziomie rozwoju społeczno-gospodarczego.
- **under-five deaths** (CV = 381,69%) — Liczba zgonów dzieci poniżej 5 lat wykazuje podobnie wysoką zmienność, wskazując na ściśle powiązane czynniki zdrowotne z wcześniejszą zmienną.

3.2 Zmienne o wysokiej zmienności (100% < CV < 200%)

Trzy zmienne wykazują wysoką zmienność w przedziale 100–200%:

- **HIV/AIDS** (CV = 291,47%) — Rozpowszechnienie wirusa HIV wykazuje zmienność od 0,10 do 50,60, z największym rozprzestrzenieniem w krajach afrykańskich.
- **percentage expenditure** (CV = 269,27%) — Wydatki na opiekę zdrowotną jako procent PKB wahają się między 0,00 a 19,479,91, odzwierciedlając ogromne różnice w priorytetach budżetowych krajów.
- **GDP** (CV = 190,70%) — Produkt krajowy brutto krajów zawiera się w przedziale od 1,68 do 119,172,74, co wskazuje na znaczne różnice ekonomiczne między krajami.

3.3 Zmienne o umiarkowanej zmienności (50% < CV < 100%)

Osiem zmiennych charakteryzuje się umiarkowaną zmiennością:

- **Wskaźniki niedowagi** — Zmienne *thinness 5-9 years* (CV = 92,58%) i *thinness 1-19 years* (CV = 91,33%) odzwierciedlają znaczące różnice w zdrowotnym stanie odżywienia dzieci i młodzieży między krajami.
- **Alcohol** (CV = 88,04%) — Spożycie alkoholu na mieszkańca wykazuje znaczną zmienność między krajami, uwarunkowaną czynnikami kulturowymi i ekonomicznymi.

- **Adult Mortality** (CV = 75,42%) — Śmiertelność dorosłych standaryzowana na liczbę mieszkańców wykazuje umiarkowaną zmienność, odzwierciedlającą różne warunki zdrowotne między krajami.
- **BMI** (CV = 52,31%) — Średni indeks masy ciała populacji zawiera się w przedziale 1,00 do 87,30, wskazując na znaczące różnice w stanie odżywienia populacji.

3.4 Zmienne o niskiej zmienności (CV $<$ 50%)

Pozostałe zmienne (Hepatitis B, Polio, Diphtheria, Total expenditure, Income composition, Schooling) wykazują relatywnie niską zmienność (CV $<$ 50%), co wskazuje na bardziej skoncentrowane i stabilne wartości wokół średniej.

3.5 Charakterystyka zmiennej docelowej

Zmienną docelową jest **Life expectancy** (oczekiwana długość życia), która wykazuje bardzo **stabilną zmienność**:

Tabela 4: Statystyki opisowe zmiennej docelowej

Statystyka	**Wartość**
Minimum	36,30 lat
Maksimum	89,00 lat
Średnia	69,22 lat
Mediana	72,10 lat
Odchylenie standardowe	9,52 lat
Współczynnik zmienności (CV%)	13,76%

Niska zmienność zmiennej docelowej (CV = 13,76%) oznacza, że oczekiwana długość życia w analizowanym zbiorze jest stosunkowo jednorodna i skoncentrowana wokół średniej. Jest to pozytywna cecha dla celów modelowania predykcyjnego.

4 Braki danych- stopień wypełnienia

4.1 Analiza brakujących danych

Zmienne z najwyższym odsetkiem brakujących obserwacji to:

Tabela 5: Zmienne z największym udziałem brakujących danych

Zmienna	**Liczba brakujących**	**Procent**
Population	652	22,19%
Hepatitis B	553	18,82%
GDP	448	15,25%
Total expenditure	226	7,69%
Alcohol	194	6,60%
Income composition of resources	167	5,68%
Schooling	163	5,55%

Siedem pozostałych zmiennych z brakującymi danymi (BMI, thinness, Polio, Diphtheria, Life expectancy, Adult Mortality) zawiera mniej niż 2% brakujących wartości.

4.2 Kompletne obserwacje

Osiem zmiennych (36,36% wszystkich zmiennych) zawiera pełne dane bez brakujących wartości: Country, Year, Status, infant deaths, percentage expenditure, Measles, under-five deaths, HIV/AIDS.

5 Rozkłady zmiennych i normalność rozkładu

5.1 Histogram zmiennej docelowej

Histogram Life expectancy ujawnia rozkład zbliżony do normalnego, z dominantą wokół 72 lat. Średnia wartość (zaznaczona czerwoną linią przerywaną) pokrywa się w przybliżeniu z medianą, co wskazuje na symetryczność rozkładu. Obserwacje rozprzestrzeniają się w zakresie 36.30–89.00 lat, z rzadkimi przypadkami na krańcach.

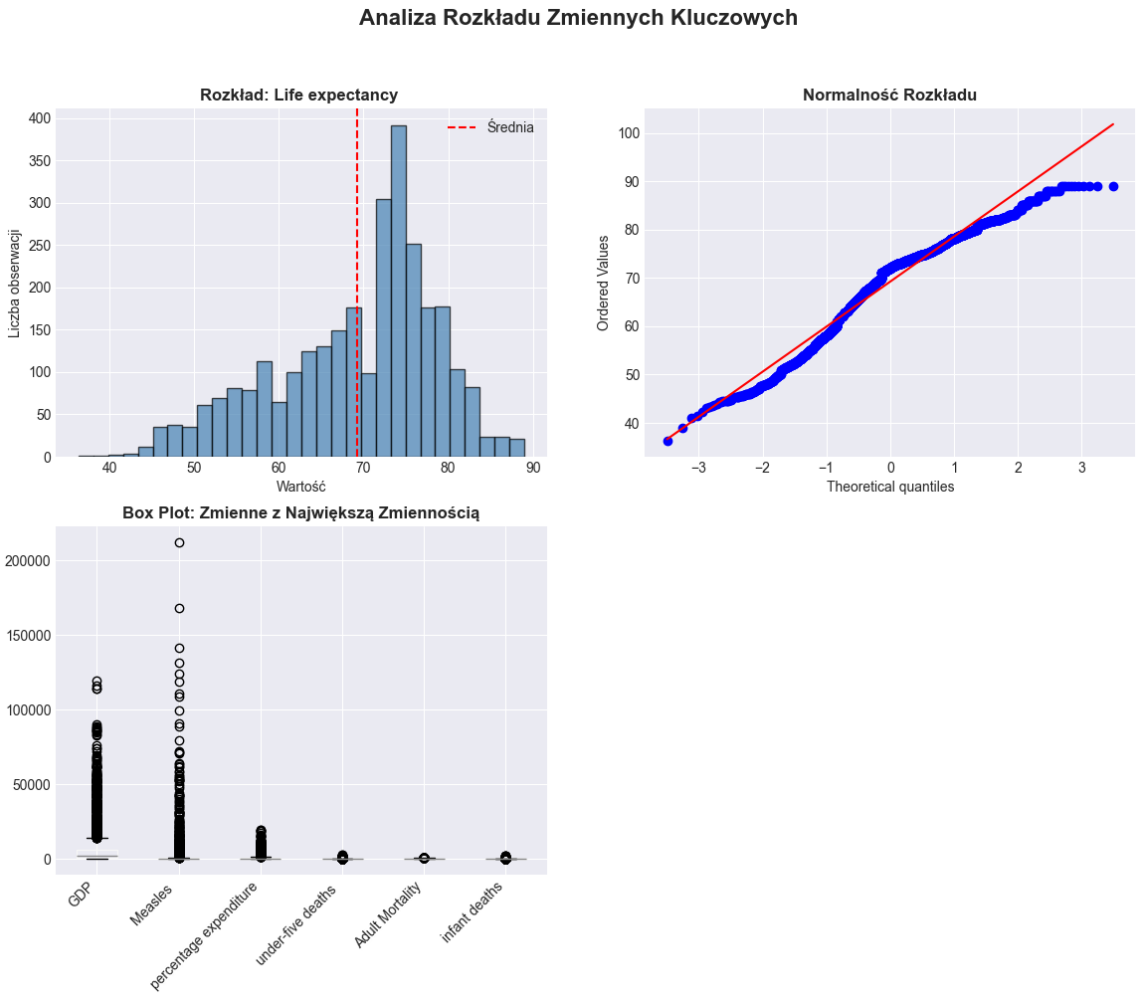
5.2 Normalność rozkładu – Q-Q plot

Q-Q plot (Quantile-Quantile plot) porównuje kwantyle obserwowanych danych z kwantylami rozkładu normalnego. Punkty leżą blisko linii diagonalnej, szczególnie w środkowej części rozkładu, co potwierdza **normalność rozkładu zmiennej docelowej**. Lekkie odchylenia obserwowane na krańcach rozkładu (zarówno na górze jak i na dole) są typowe dla rzeczywistych danych i nie stanowią naruszenia założenia normalności.

5.3 Box plot – zmienne z największą zmiennością

Box plot przedstawia sześć zmiennych o największej zmienności (z wyłączeniem Population):

- **Zmienne z ekstremalnymi outlierami**: GDP, Measles i percentage expenditure wykazują znaczące wartości odstające (outliers) powyżej górnego wąsa. Obserwacje te są konsekwencją ekstremalnej zmienności (CV $>$ 200%) tych zmiennych.
- **Zmienne ze skompaktowanym rozkładem**: BMI, Diphtheria i Under-five deaths wykazują bardziej zwarte rozkłady z mniejszą liczbą wartości odstających.
- **Asymetria rozkładów**: Mediana (linia w środku pudełka) niekoniecznie dzieli pudełko na równe części, co wskazuje na asymetryczne rozkłady w niektórych zmiennych.



Rysunek 2: Analiza rozkładu zmiennych kluczowych: (a) histogram Life expectancy z zaznaczoną średnią, (b) Q-Q plot dla oceny normalności rozkładu, (c) box plot sześciu zmiennych o największej zmienności.

6 Hipoteza badawcza

Oczekiwana długość życia w poszczególnych krajach jest silnie determinowana przez wskaźniki ekonomiczne (GDP), dostęp do opieki medycznej, oraz warunki sanitarne i edukacyjne. Zmienne o najsilniejszej korelacji będą stanowić predyktory modelu regresji.

7 Wnioski

Przeprowadzona analiza eksploracyjna danych WHO dotyczących oczekiwanej długości życia pozwoliła na identyfikację kluczowych czynników wpływających na zdrowie populacji. Najsilniejsze zależności ze zmienną docelową wykazały zmienne społeczno-ekonomiczne, w szczególności poziom edukacji oraz skład dochodów, co potwierdza ich fundamentalną rolę w kształtowaniu długości życia.

Istotny wpływ mają również czynniki zdrowotne, takie jak śmiertelność dorosłych, rozpowszechnienie HIV/AIDS oraz poziom wyszczepienia, które w sposób bezpośredni odzwierciedlają jakość systemu opieki zdrowotnej. Analiza zmienności ujawniła znaczne

dysproporcje między krajami, zwłaszcza w zakresie zmiennych demograficznych i ekonomicznych, przy jednocześnie stabilnym rozkładzie zmiennej docelowej.

Uzyskane wyniki stanowią solidną podstawę do dalszego modelowania predykcyjnego oraz budowy modeli regresyjnych w kolejnych etapach projektu.