Analiza czynników wpływających na oczekiwaną długość życia w wybranych krajach

Regresja liniowa

Agnieszka Olechnowicz (211023), Patrycja Pobuta (211028),

Patrycja Król (230077), Radosław Polak (229500)

Spis treści

1. Streszczenie

Celem pracy jest analiza czynników wpływających na oczekiwaną długość życia w 130 krajach. W badaniu zastosowano regresję liniową i metodę Lasso do doboru istotnych zmiennych. Wyniki wskazują, że najważniejszymi czynnikami determinującymi długość życia są wskaźnik rozwoju społecznego, umieralność dorosłych, liczba zgonów z powodu HIV/AIDS oraz wydatki na zdrowie. Model uzyskał wysoką wartość R^2, a przeprowadzone testy potwierdziły poprawność jego założeń. Badanie podkreśla znaczenie polityki zdrowotnej i społecznej w kontekście wydłużania życia populacji.

2. Słowa kluczowe

oczekiwana długość życia, regresja liniowa, regresja lasso, badanie założeń

3. Wprowadzenie

Oczekiwana długość życia to ważny miernik wykorzystywany w demografii. Określa on średnią długość życia w danej populacji. Na jego podstawie można pośrednio wnioskować o jakości życia, poziomie opieki zdrowotnej, warunkach ekonomicznych i społecznych oraz stylu życia mieszkańców danego kraju lub regionu. Wyższe wartości oczekiwanej długości życia mogą wskazywać m.in. na lepszy dostęp do opieki medycznej, wysoki rozwój czy poziom życia, a także bezpieczeństwo na danym terenie. Zestawienie wartości dla różnych krajów pozwala porównać przeciętną długość życia w innych częściach świata oraz badanie czynników, które wpływają na ich odmienność. Dogłębna analiza tego zagadnienia może dostarczyć informacji, co można zrobić, aby wydłużyć średnią długość życia mieszkańców regionów o niskich wartościach.

4. Cel i zakres badania

Celem badania jest analiza czynników wpływających na oczekiwaną długość życia na podstawie danych dotyczących 130 krajów świata, dla których dostępne były rozpatrywane zmienne. Należą do nich zarówno kraje rozwinięte, jak i rozwijające się, co również zostało uwzględnione jako potencjalny czynnik wpływający na długość życia. Ponadto w analizie wzięto pod uwagę także m. in. wskaźniki umieralności dzieci i dorosłych na 1000 mieszkańców, udział niemowląt zaszczepionych na poszczególne choroby, Produkt Krajowy Brutto, wydatki na zdrowie danego kraju, Wskaźnik Rozwoju Społecznego czy spożycie alkoholu na osobę.

Analizowane dane pochodzą z The Global Health Observatory (GHO), repozytorium Światowej Organizacji Zdrowia (WHO) i dotyczą 2013 roku. Sama baza danych z wybranymi zmiennymi, która została użyta w badaniu, pochodzi z platformy Kaggle[1].

5. Przegląd literatury

Analiza średniej długości życia jest popularnym tematem badań, rozpatrywanym zarówno na przestrzeni lat, jak i w kontekście czynników na nią wpływających. Poniżej przytoczono kilka prac, w których do badania zależności między oczekiwaną długością życia a innymi czynnikami zastosowano regresję liniową.

Chen, Z., Ma, Y., Hua, J., Wang, Y., & Guo, H. (2021)[2] stworzyli dwa modele liniowe, rozdzielając kraje rozwinięte i rozwijające się na dwie oddzielne grupy, w których znalazło się po 10 wybranych przez autorów państw. Jako zmienne objaśniające wybrano 9 potencjalnych czynników: PKB na osobę, wskaźnik urbanizacji, wydatki na opiekę zdrowotną oraz szkolnictwo, udział terenów leśnych, współczynnik Giniego, średnie roczne narażenie na $PM_{2,5}$, emisja CO_2 oraz zużycie nawozów. Przy pomocy zbudowanych modeli wykazano, że w krajach rozwiniętych na oczekiwaną długość życia największy pozytywny wpływ ma PKB na mieszkańca, a największy negatywny zużycie nawozów. Wśród krajów rozwijających się stopień urbanizacji ma największy pozytywny wpływ na długość życia, podczas gdy współczynnik Giniego ma największy negatywny wpływ.

Jafrin, N., Masud, M.M., Seif, A.N.M., Mahi, M., Khanam, M. (2021)[3] analizowali czynniki wpływające na oczekiwaną długość życia na podstawie pięciu wybranych krajów SAARC (Bangladesz, Indie, Pakistan, Nepal i Sri Lanka) w latach od 2000 do 2016. Wśród rozpatrywanych zmiennych zalazły się u nich procentowy wzrost PKB, akumulacja brutto (% PKB), wydatki na zdrowie, średnia liczba lat nauki, współczynnik dzietności, wskaźnik urbanizacji, emisja CO_2, udział ludności korzystającej z przynajmniej podstawowych usług sanitarne oraz używających Internetu, a także liczba abonamentów komórkowych na 100 osób. Wyniki ich badania sugerują, że spadek całkowitego współczynnika dzietności, populacji miejskiej i emisji CO_2 doprowadzi do wzrostu oczekiwanej długości życia. Zwracają również uwagę na nietypowy negatywny i istotny statystycznie wpływ wydatków na zdrowie uzyskany po oszacowaniu modelu regresji.

Amos, B.K., & Smirnov, I. (2022)[4] w swojej pracy użyli natomiast takiego samego zestawu zmiennych jak w tym badaniu, jednak dane obejmowały lata od 2000 do 2015 i dotyczyły 193 krajów (zastosowano uzupełnienie braków danych). Skupili się na przedstawieniu korelacji pojedynczych zmiennych z objaśnianą oczekiwaną długością życia, a następnie zbudowali regresję liniową, grzbietową oraz lasso dla wszystkich zmiennych. Najlepsze wyniki uzyskali dla regresji liniowej, która wykazała wszystkie zmienne istotne, z wyjątkiem Alcohol, Hepatitis_B, Measles, Population, Slimness.1.19 i Slimness.5.9. Wykazali także, że średnia długość życia rośnie z biegiem lat i jest średnio dłuższa w krajach rozwiniętych niż w krajach rozwijających się.

6. Przedstawienie zmiennych

Oznaczenie oraz opis zmiennych dostępnych w wykorzystywanym zbiorze danych przedstawiony został w poniższej tabeli:

	Zmienna	Opis
X1	Status	status rozwoju kraju (Developed = 1, Developing = 0)
X2	Adult_Mortality	wskaźniki umieralności dorosłych obu płci (prawdopodobieństwo śmierci od 15 do 60 lat na 1000 mieszkańców)
X3	infant_deaths	liczba śmierci niemowląt na 1000 mieszkańców
X4	Alcohol	spożycie alkoholu na osobę (w litrach czystego alkoholu)
X5	percentage_expenditure	wydatki na zdrowie jako procent Produktu Krajowego Brutto na mieszkańca (%)
X6	Hepatitis_B	zasięg szczepień przeciwko wirusowemu zapaleniu wątroby typu B (HepB) wśród dzieci w wieku 1 roku (%)
X7	Measles	liczba zgłoszonych przypadków odry na 1000 mieszkańców
X8	BMI	średni wskaźnik masy ciała całej populacji
X9	under_five_deaths	liczba śmierci dzieci do 5 roku życia na 1000 mieszkańców
X10	Polio	zasięg szczepień przeciwko polio (Pol3) wśród dzieci w wieku 1 roku (%)
X11	Total_expenditure	wydatki sektora instytucji rządowych i samorządowych na zdrowie jako odsetek całkowitych wydatków rządowych (%)
X12	Diphtheria	zasięg szczepień dzieci w wieku 1 roku na toksoid błoniczy, tężcowi i krztusiec (DTP3) (%)
X13	HIV_AIDS	zgony na 1000 żywych urodzeń HIV/AIDS
X14	GDP	Produkt Krajowy Brutto na mieszkańca (w USD)
X15	Population	populacja kraju
X16	thinness_1_19_years	częstość występowania szczupłości wśród dzieci i młodzieży w wieku od 10 do 19 lat (%)
X17	thinness_5_9_years	częstość występowania szczupłości wśród dzieci w wieku od 5 do 9 lat (%)
X18	Income_composition_of_resources	Wskaźnik Rozwoju Społecznego pod względem struktury dochodów zasobów (wskaźnik od 0 do 1)
X19	Schooling	średnia liczba lat nauki
Y	Life_expectancy	oczekiwana długość życia w latach

Tabela 1. Opis wykorzystanych zmiennych.

7. Wstępna analiza danych

7.1. Statystyki opisowe

Na początek zostały obliczone statystyki opisowe dla wszystkich zmiennych. Poniżej zamieszczona tabela przedstawia, jak kształtowały się wartości średnie, odchylenie standardowe, mediana, wartość minimalna i maksymalna, zakres, skośność, kurtoza, błąd standardowy oraz kwartyl 1 i 3.

	mean	sd	median	min	max	range	skew	kurtosis	se	Q0.25	Q0.75
X2	155.50	108.36	140.00	6.00	518.00	512.00	0.77	0.34	9.50	68.00	227.00
X3	29.72	104.27	3.00	0.00	1000.00	1000.00	7.35	62.05	9.15	0.00	20.75
X4	3.84	4.15	2.42	0.01	15.04	15.03	0.73	-0.66	0.36	0.01	7.23
X5	762.52	2005.94	173.75	1.00	15516.00	15515.00	5.36	31.91	175.93	40.38	687.25
X6	81.76	24.32	92.00	6.00	99.00	93.00	-2.12	3.71	2.13	78.00	96.00
X7	1370.48	5526.32	11.00	0.00	52852.00	52852.00	7.31	62.11	484.69	0.00	234.25
X8	40.07	20.32	45.00	2.10	76.70	74.60	-0.20	-1.30	1.78	22.35	58.80
X9	40.22	139.83	3.00	0.00	1300.00	1300.00	7.11	57.20	12.26	1.00	27.00
X10	83.41	23.23	93.50	7.00	99.00	92.00	-2.31	4.79	2.04	82.00	97.00
X11	6.27	2.48	6.00	1.00	12.00	11.00	0.14	-0.30	0.22	4.50	8.00
X12	86.67	17.35	93.00	8.00	99.00	91.00	-2.85	9.46	1.52	83.00	97.00
X13	0.90	1.78	0.10	0.10	9.80	9.70	3.00	9.76	0.16	0.10	0.60
X14	6381.75	13988.27	1864.50	14.00	113752.00	113738.00	4.99	30.56	1226.85	561.00	6000.25
X15	13268546.37	28875910.91	1682963.50	393.00	181712595.00	181712202.00	3.76	15.90	2532584.92	349385.50	12734345.50
X16	4.75	4.51	3.25	0.10	26.80	26.70	1.95	5.11	0.40	1.50	6.80
X17	4.72	4.41	3.25	0.10	27.50	27.40	1.99	5.88	0.39	1.50	6.68
X18	0.67	0.15	0.70	0.34	0.93	0.59	-0.24	-0.97	0.01	0.54	0.77
X19	12.61	2.72	12.70	5.30	20.30	15.00	-0.06	-0.04	0.24	10.72	14.40
Y	70.39	8.28	71.50	49.90	87.00	37.10	-0.37	-0.34	0.73	65.25	75.52

Tabela 2. Statystyki opisowe analizowanych zmiennych.

	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	Y
wsp_zm	69.686	350.908	108.056	263.066	29.742	403.239	50.716	347.646	27.848	39.6	20.019	197.415	219.192	217.627	95.026	93.345	22.807	21.551	11.77

Tabela 3. Współczynniki zmienności dla poszczególnych zmiennych.

Wszystkie zmienne charakteryzują się dużą zmiennością, więc na tym etapie nie ma podstaw do odrzucenia którejkolwiek, jeśli chcielibyśmy wyeliminować zmienne o znikomym wpływie na wartość wyjściową Y.

Interpretacja statystyk opisowych:

X2 – Umieralność dorosłych (Adult Mortality)

Średnia wartość umieralności dorosłych (tj. prawdopodobieństwo śmierci osób w wieku od 15 do 60 lat) wynosi 155,50 na 1000 mieszkańców, przy odchyleniu standardowym równym 108,36. Mediana wskaźnika wynosi 140, podczas gdy zakres zmienności to od 18 do 518 (rozpiętość 500). Rozkład cechuje się lekką dodatnią asymetrią (skośność 0,77), co oznacza, że występują nieco częściej wyższe wartości. Kurtoza wynosi 0,34, co wskazuje na nieznaczne spłaszczenie rozkładu w porównaniu do rozkładu normalnego. Percentyle 25. i 75. wynoszą odpowiednio 68 i 227.

X3 – Zgony niemowląt (Infant Deaths)

Średnia liczba zgonów niemowląt na 1000 urodzeń wynosi 29,72, z odchyleniem standardowym równym 104,27. Mediana wynosi 3, a wartości wahają się od 0 do 1000 (rozpiętość 1000). Rozkład jest mocno asymetryczny dodatnio (skośność 7,35), co świadczy o rzadkim występowaniu bardzo wysokich wartości. Kurtoza wynosi 62,05, wskazując na znaczną koncentrację wartości wokół minimum oraz obecność dużych odstających wartości. Percentyle 25. i 75. to odpowiednio 0 oraz 20,75.

X4 – Spożycie alkoholu (Alcohol)

Średnie spożycie alkoholu na osobę wynosi 3,84 litra rocznie, przy odchyleniu standardowym 4,15. Mediana wynosi 2,42 litra, a wartości wahają się od 0 do 12,50. Rozkład cechuje się niewielką dodatnią asymetrią (skośność 0,77) i ma kurtozę -0,24, co wskazuje na rozkład zbliżony do normalnego. Percentyle 25. i 75. wynoszą odpowiednio 0,79 i 6,03 litra.

X5 – Wydatki na zdrowie jako procent PKB (Percentage Expenditure)

Średnie wydatki na zdrowie wynoszą 762,52% PKB per capita, przy bardzo wysokim odchyleniu standardowym 2005,94. Mediana wynosi 173,75%, a wartości wahają się od 0 do 15 515%. Rozkład jest silnie asymetryczny dodatnio (skośność 5,36) z kurtozą 31,91, co sugeruje obecność wyjątkowo wysokich wartości w niektórych krajach. Percentyle 25. i 75. wynoszą odpowiednio 40,38 i 687,25%.

X6 – Szczepienia przeciw WZW B (Hepatitis B)

Średni zasięg szczepień przeciw wirusowemu zapaleniu wątroby typu B wynosi 80,04%, przy odchyleniu standardowym 22,52. Mediana wynosi 90%, a wartości wahają się od 0 do 100%. Rozkład jest lekko asymetryczny ujemnie (skośność -1,11), co wskazuje na większą koncentrację wartości blisko maksymalnych. Percentyle 25. i 75. wynoszą odpowiednio 73% oraz 96%.

X7 – Liczba zachorowań na odrę (Measles)

Średnia liczba zgłoszonych przypadków odry wynosi 1370,48 na 1000 mieszkańców, przy bardzo wysokim odchyleniu standardowym 5526,32. Mediana wynosi 11, a wartości wahają się od 0 do 52 852, co wskazuje na dużą zmienność i sporadyczność występowania wyjątkowo wysokich wartości. Rozkład jest silnie asymetryczny dodatnio (skośność -7,31), z kurtozą 62,11. Percentyle 25. i 75. wynoszą odpowiednio 0 i 234,25.

X8 – Średni wskaźnik masy ciała (BMI)

Średnia wartość wskaźnika masy ciała populacji wynosi 40,07, z odchyleniem standardowym 32,34. Mediana wynosi 32, a wartości wahają się od 11 do 60. Rozkład jest zbliżony do symetrycznego (skośność 0,15), z kurtozą -1,18. Percentyle 25. i 75. wynoszą odpowiednio 28 i 58,80.

X9 – Zgony dzieci poniżej 5 roku życia (Under Five Deaths)

Średnia liczba zgonów dzieci poniżej 5 roku życia wynosi 40,22 na 1000 narodzin, z odchyleniem standardowym 139,83. Mediana to 3, a wartości wahają się od 0 do 1300. Rozkład jest mocno asymetryczny dodatnio (skośność 7,11), co sugeruje obecność nielicznych wyjątkowo wysokich wartości. Kurtoza wynosi 57,20. Percentyle 25. i 75. wynoszą odpowiednio 0 oraz 27.

X10 – Szczepienia przeciwko polio (Polio)

Średni zasięg szczepień przeciwko polio wynosi 81,43%, z odchyleniem standardowym 22,35. Mediana to 93%, a wartości wahają się od 0 do 100%. Rozkład jest asymetryczny ujemnie (skośność -1,22) z kurtozą 0,40. Percentyle 25. i 75. wynoszą odpowiednio 71% oraz 97%.

X11 – Wydatki rządowe na zdrowie (Total Expenditure)

Średnie wydatki rządowe na zdrowie wynoszą 6,27% całkowitych wydatków, z odchyleniem standardowym 2,48. Mediana wynosi 5,90%, a wartości wahają się od 0 do 17,6%. Rozkład

cechuje się lekką dodatnią asymetrią (skośność 0,68) i niewielką koncentracją wokół wartości średnich (kurtoza 0,42). Percentyle 25. i 75. wynoszą odpowiednio 5,02% oraz 8,02%.

X12 – Szczepienia przeciw błonicy, tężcowi i krztuścowi (Diphtheria)

Średni zasięg szczepień wynosi 82,24%, z odchyleniem standardowym 21,07. Mediana wynosi 93%, a wartości wahają się od 0 do 100%. Rozkład jest lekko asymetryczny ujemnie (skośność -1,21), co wskazuje na większe wartości blisko maksimum. Percentyle 25. i 75. wynoszą odpowiednio 72% oraz 96%.

X13 – Zgony z powodu HIV/AIDS

Średnia liczba zgonów z powodu HIV/AIDS wynosi 0,90 na 1000 osób, z odchyleniem standardowym 1,78. Mediana to 0,10, a wartości wahają się od 0 do 9,76. Rozkład jest silnie asymetryczny dodatnio (skośność 3,12), co wskazuje na rzadkie, lecz znaczące przypadki wysokich wartości. Kurtoza wynosi 9,76.

X14 – Produkt Krajowy Brutto (GDP)

Średni PKB na mieszkańca wynosi 6381,75 USD, przy odchyleniu standardowym 13 988,27. Mediana wynosi 1864 USD, a wartości wahają się od 228 do 113 752 USD. Rozkład jest mocno asymetryczny dodatnio (skośność 15,90), co oznacza, że w nielicznych przypadkach PKB jest wyjątkowo wysokie. Kurtoza wynosi 253,26.

X15 – Populacja (Population)

Średnia liczba ludności wynosi 132,68 mln, z odchyleniem standardowym 387,89 mln. Mediana wynosi 19,10 mln, a wartości wahają się od 393 tys. do 1,81 mld. Rozkład jest silnie asymetryczny dodatnio (skośność 5,11).

X16 – Częstość szczupłości wśród młodzieży (Thinness 10-19 Years)

Średnia wynosi 4,75%, z odchyleniem standardowym 4,51. Mediana to 3,75%, a wartości wahają się od 0 do 26,80%. Rozkład jest asymetryczny dodatnio (skośność 1,95), co wskazuje na rzadkość wysokich wartości.

X17 – Częstość szczupłości wśród dzieci (Thinness 5-9 Years)

Średnia wynosi 4,72%, z odchyleniem standardowym 4,41. Mediana to 3,75%, a wartości wahają się od 0 do 27,50%. Rozkład jest asymetryczny dodatnio (skośność 2,05), co wskazuje na rzadkość wysokich wartości.

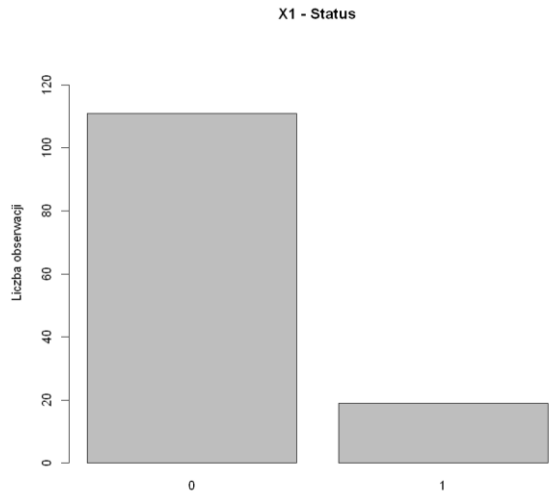
X18 – Wskaźnik Rozwoju Społecznego pod względem struktury dochodów zasobów (Income composition of resource)

Średnia wynosi 0,67, z odchyleniem standardowym 0,15. Mediana to 0,7, a wartości wahają się od 0,34 do 0,93. Rozkład jest asymetryczny ujemnie (skośność -0,24).

X19 – Średnia liczba lat nauki (Schooling)

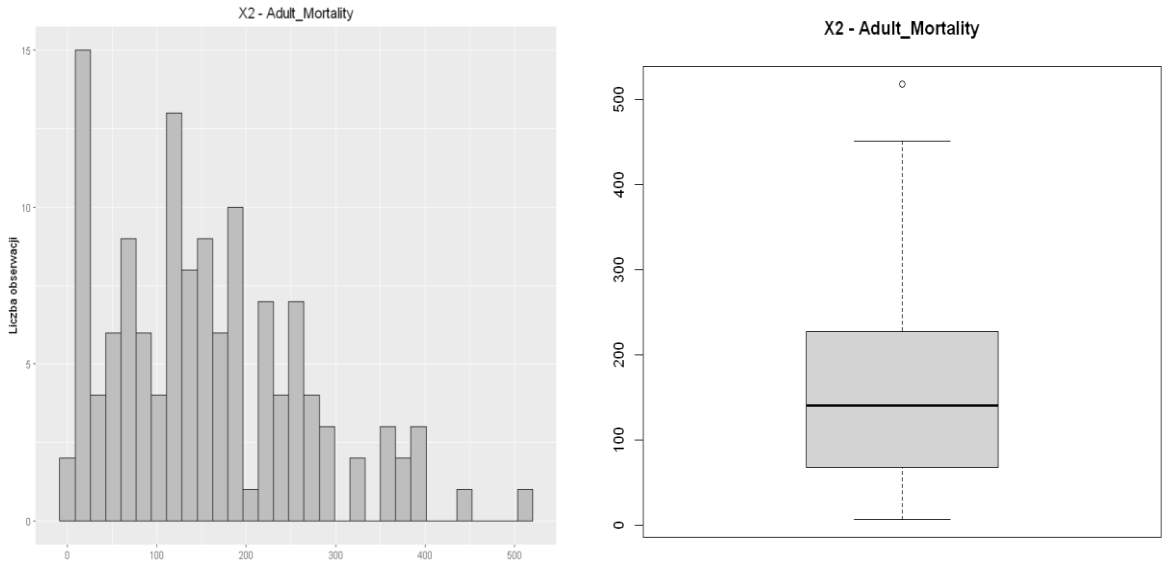
Średnia wynosi 12,61 lat, z odchyleniem standardowym 4,72. Mediana to 12 lat, a wartości wahają się od 0 do 20. Rozkład jest asymetryczny ujemnie (skośność -0,97).

7.2. Wizualizacja danych



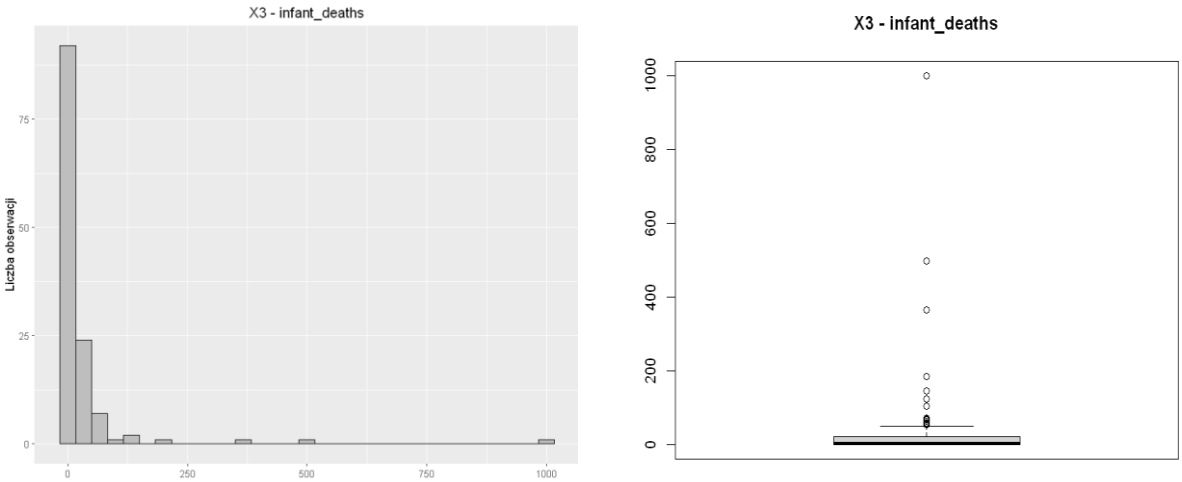
Wykres. 1. Rozkład zmiennej Status.

Jak widać po rozkładzie dla zmiennej X1 dość niewielka część państw (około 20) jest uznawana za państwa rozwinięte. Pozostałe państwa zostały sklasyfikowane jako rozwijające się (około 110). Część państw nie została uwzględniona w analizie, ze względu na występujące braki w danych.



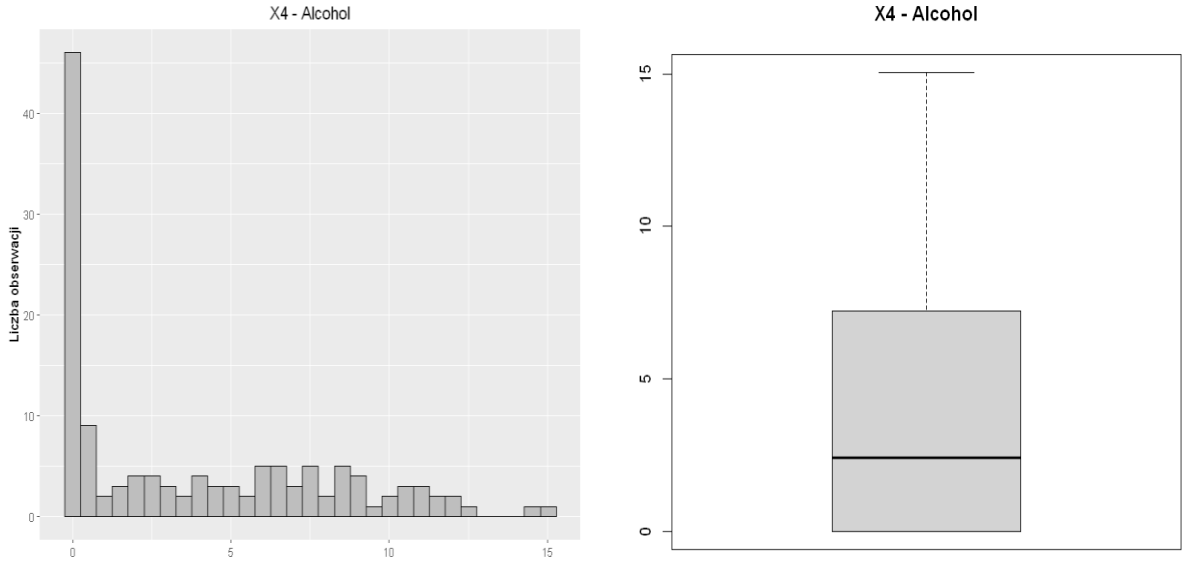
Wykres. 2. Histogram i boxplot dla zmiennej Adult Mortality.

Rozkład dla zmiennej X2 jest prawostronnie asymetryczny. Większość wartości mieści się w przedziale 0-300 oraz występuje wyraźny pik dla wartości około 12. Na podstawie wykresu pudełkowego widać, że występuje jedna obserwacja odstająca równa 500 co widać na rozkładzie jak i na wykresie pudełkowym. Mediana na poziomie około 150.



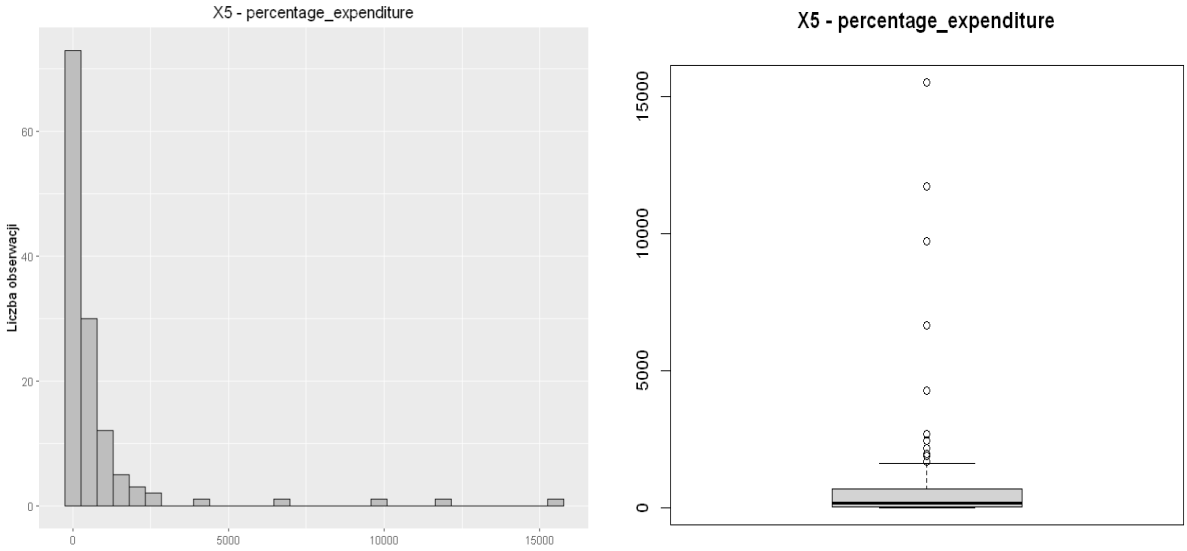
Wykres. 3. Histogram i boxplot dla zmiennej Infant deaths.

Rozkład zmiennej X3 jest silnie prawostronnie asymetryczny. Znaczna większość wartości jest silnie skupiona wokół średniej i znajduje się w przedziale 0-100, z czego duża część jest bliska zeru o czym świadczy najwyższy pik widoczny na rozkładzie. Występują wartości odstających powyżej wartości mniej więcej na poziomie 50, które są widoczne na wykresie pudełkowym.



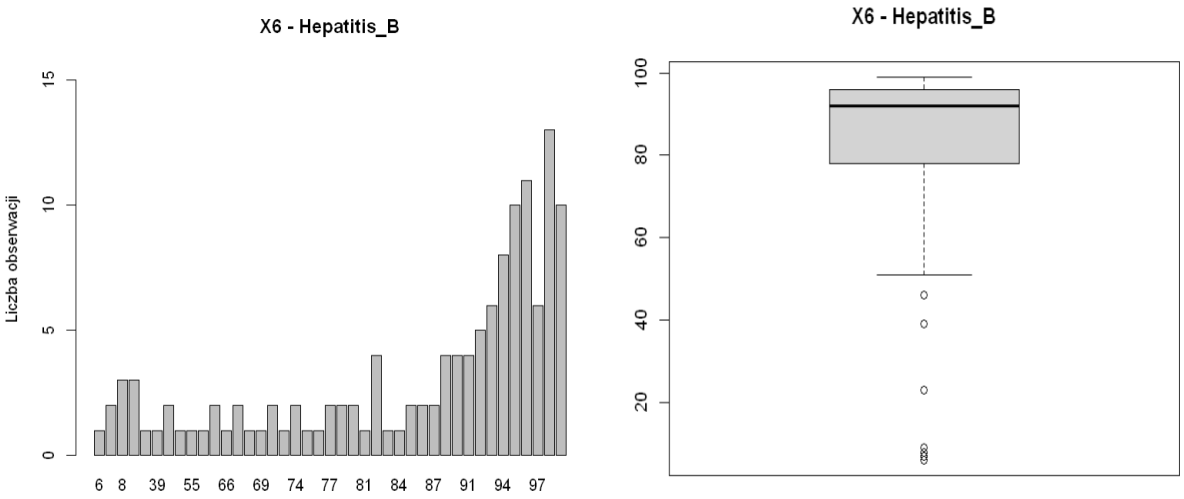
Wykres. 4. Histogram i boxplot dla zmiennej Alcohol.

Rozkład zmiennej X4 jest prawostronnie asymetryczny, dla wartości zero jest duży pik. Ponadto występuje dużo wartości w prawym ogonie rozkładu. Większość obserwacji znajduje się w przedziale 0-10. Mediana wynosi około 2,5. Nie występują natomiast wartości odstające.



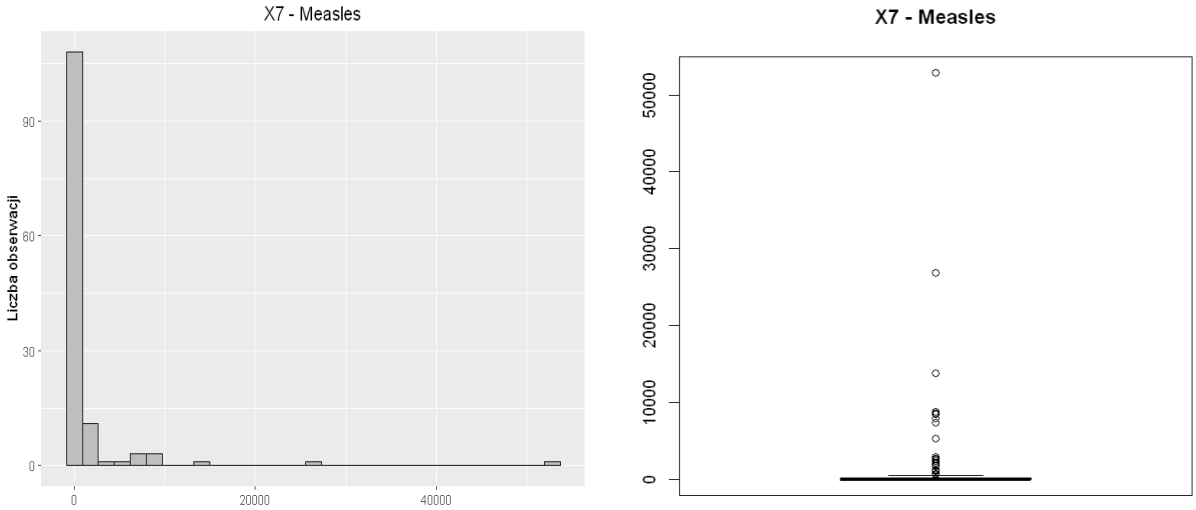
Wykres. 5. Histogram i boxplot dla zmiennej percentage expenditure.

Rozkład zmiennej X5 również wykazuje silną prawostronną asymetryczność. Większość obserwacji odpowiada wartościom z przedziału około 0-1700. Widać największy pik dla wartości około zera. Powyżej wartości około 1600 występuje kilkanaście obserwacji odstających co widać na wykresie pudełkowym.



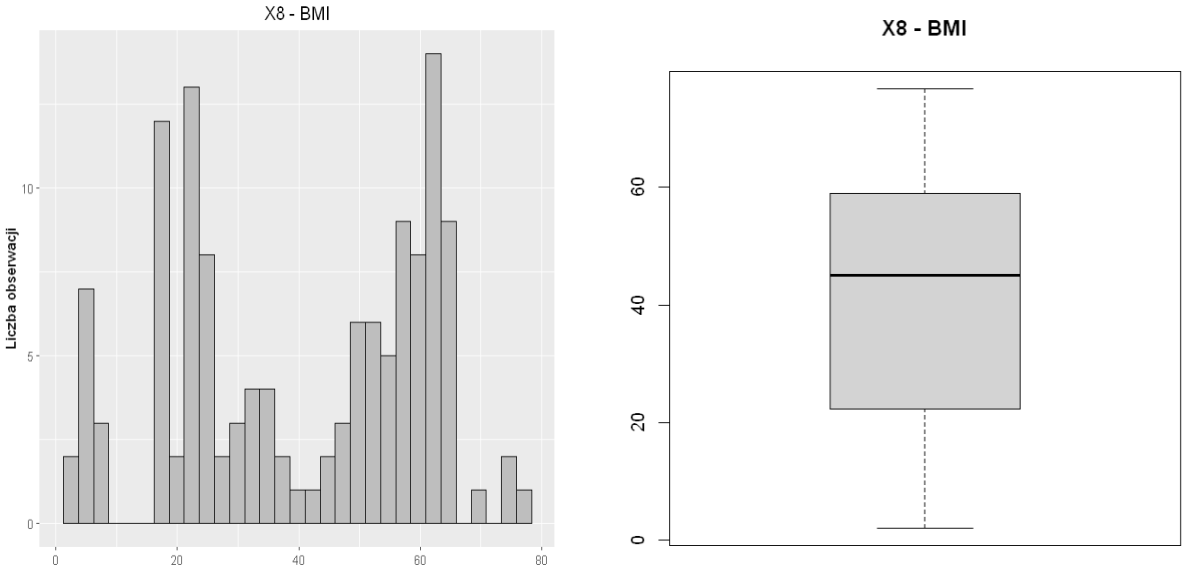
Wykres. 6. Rozkład i boxplot dla zmiennej Hepatitis_B.

Rozkład zmiennej X6 jest silnie lewostronnie asymetryczny. Większość obserwacji znajduje się w przedziale około 77-100, dla pozostałych obserwacji wartości są z przedziału 6-84. Obserwacje o wartościach poniżej 50 uznano za odstające co obrazuje wykres pudełkowy. Mediana ma wartość na poziomie około 92.



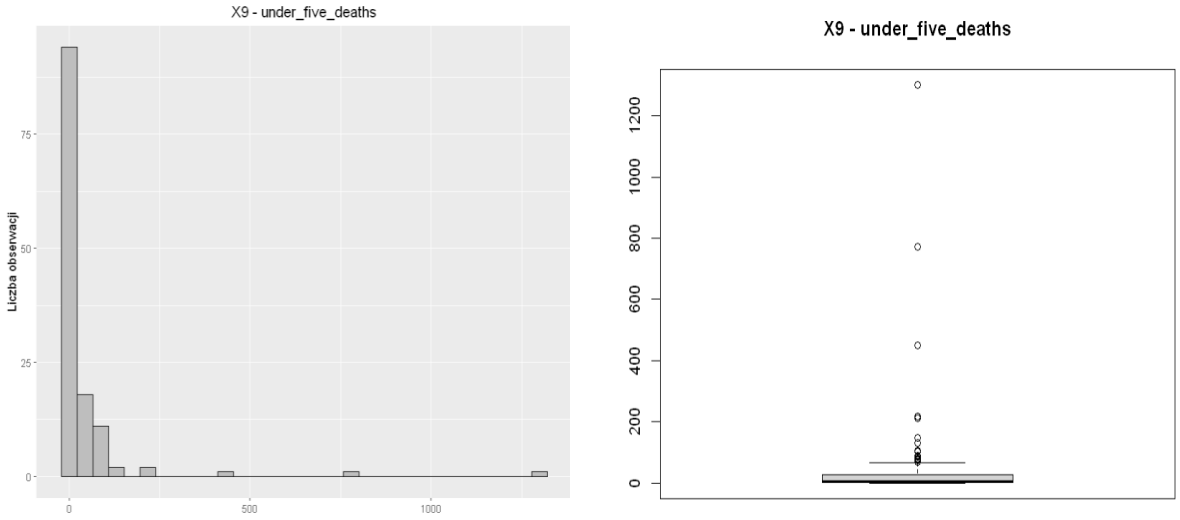
Wykres. 7. Histogram i boxplot dla zmiennej Measles.

Rozkład zmiennej X7 jest prawostronnie asymetryczny, przy czym znacząca większość obserwacji charakteryzuje się wartościami bliskimi 0 w zestawieniu z całkowitym zakresem przyjmowanych wartości. Powyżej wartości około 1000 obserwacje uznawane są za odstające i jest ich dosyć dużo.



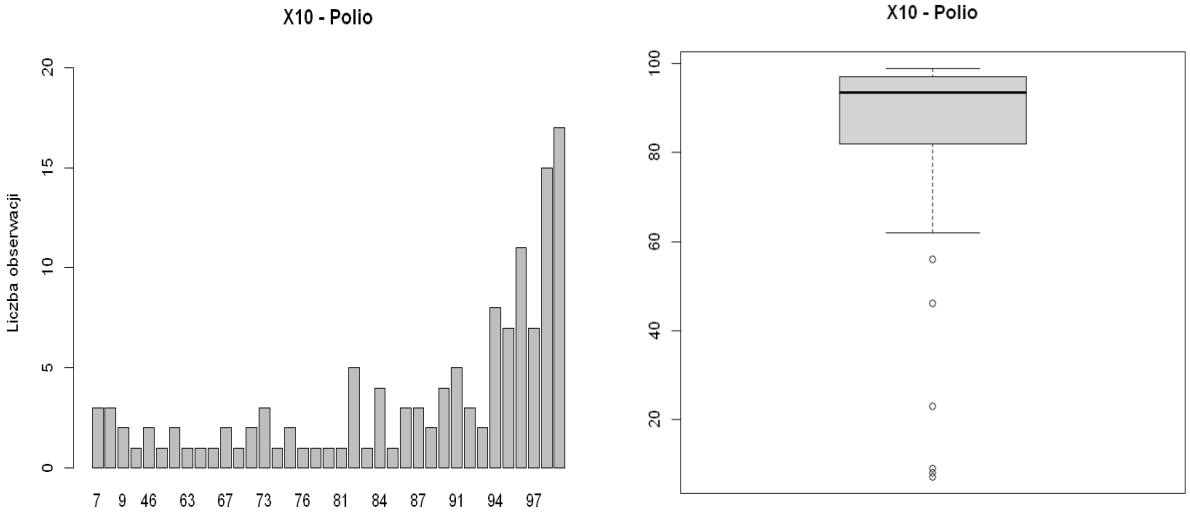
Wykres. 8. Histogram i boxplot dla zmiennej BMI.

Rozkład zmiennej X8 charakteryzuje się lekką lewostronną asymetrycznością z tym, że w lewym ogonie rozkładu występują 2 piki (dla wartości około 16 i 24) porównywalne z najwyższym pikiem (wartość około 63). Większość obserwacji zawiera się w szacunkowym przedziale między 16 a 66. Nie występują obserwacje odstające, z kolei mediana wynosi blisko 45.



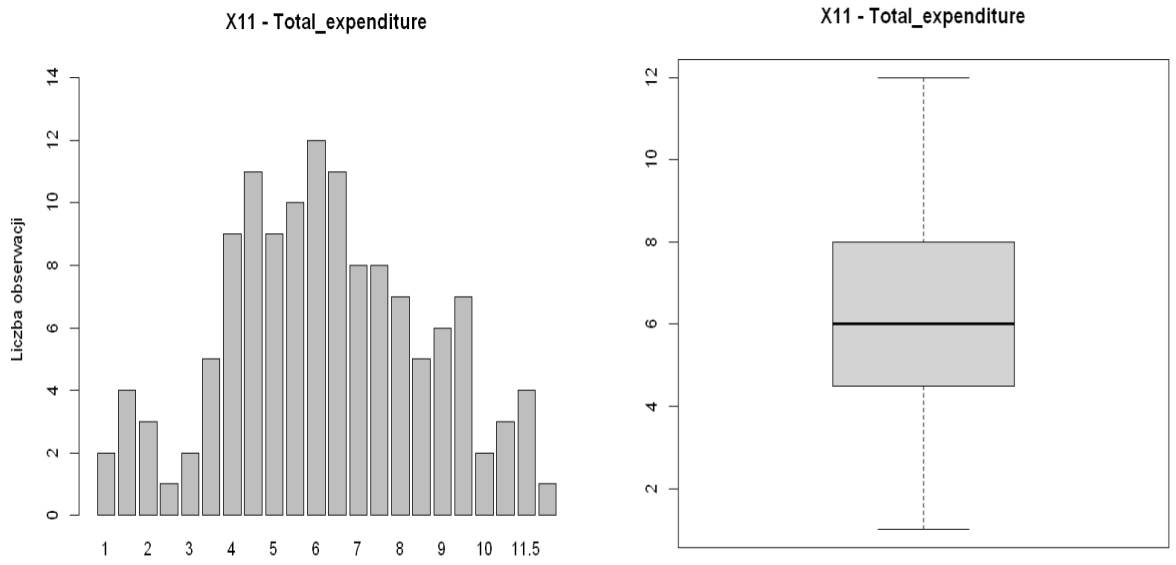
Wykres. 9. Histogram i boxplot dla zmiennej under five deaths.

Rozkład zmiennej X9 jest prawostronnie asymetryczny. Już na tym etapie można zauważyć silną korelację między tą zmienną a wcześniej omówioną zmienną X3 (Infant deaths), ponieważ ich rozkłady i wykresy pudełkowe są do siebie bardzo zbliżone.



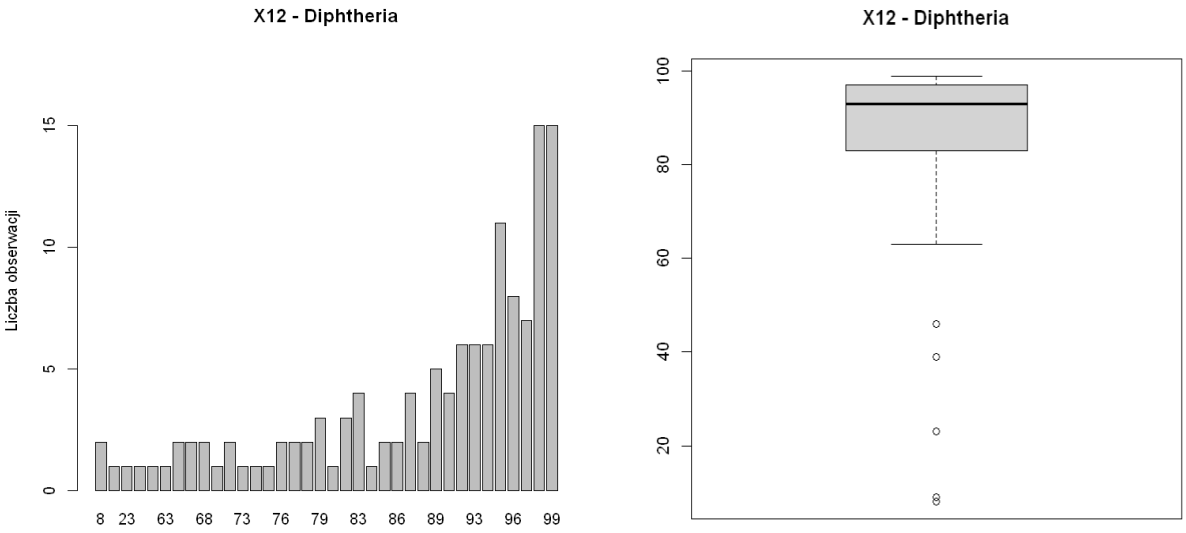
Wykres. 10. Rozkład i boxplot dla zmiennej Polio.

Rozkład zmiennej X10 jest silnie lewostronnie asymetryczny. Większość wartości zawiera się w przedziale około 84-100. Na wykresie pudełkowym widać, że poniżej wartości 60 występuje kilka obserwacji odstających, mediana około 93.



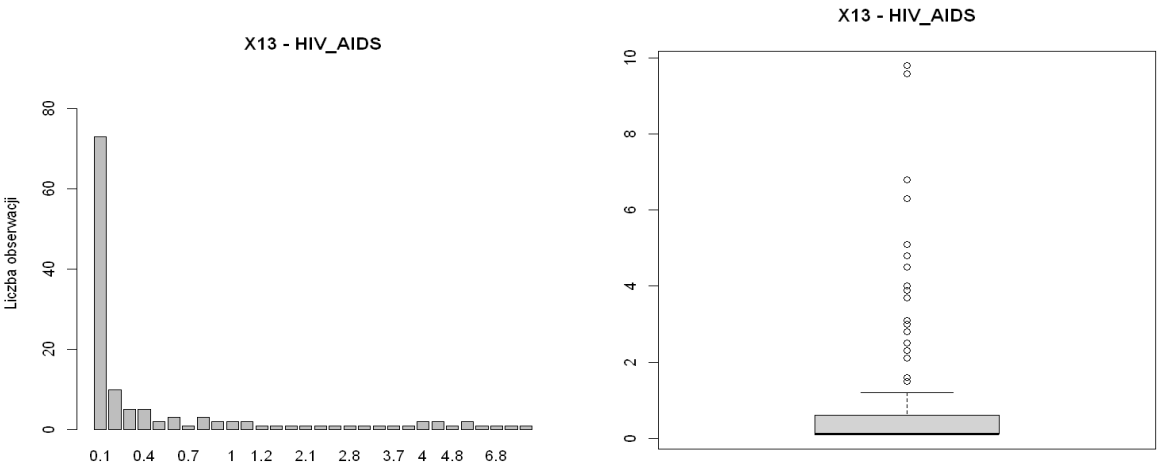
Wykres. 11. Rozkład i boxplot dla zmiennej Total expenditure.

Rozkład zmiennej X11 jest mocno zbliżony do symetrycznego. Największy pik pokrywa się z wartością mediany równą 6. Zarówno na rozkładzie jak i na wykresie pudełkowym widać, że większość obserwacji odpowiada wartościom z przedziału 3-10, brak obserwacji odstających.



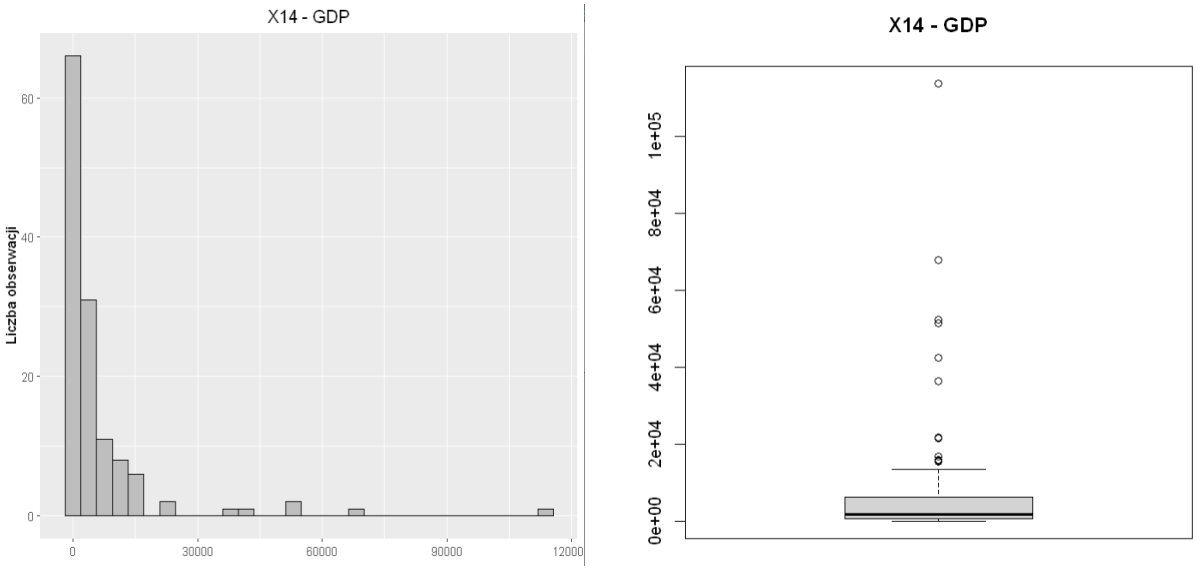
Wykres. 12. Rozkład i boxplot dla zmiennej Diphtheria.

Rozkład zmiennej X12 charakteryzuje silna lewostronna asymetryczność z pikami dla wartości około 99. Większość obserwacji jest z przedziału 79-99. Na wykresie pudełkowym widać kilka obserwacji odstających występujących dla wartości poniżej 60.



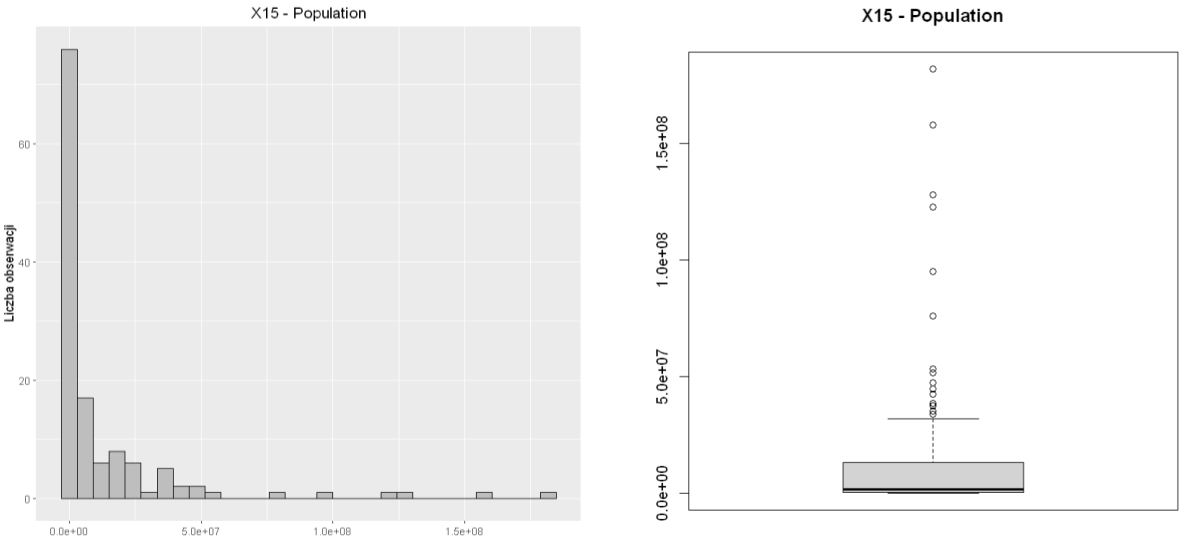
Wykres. 13. Rozkład i boxplot dla zmiennej HIV_AIDS.

Rozkład dla zmiennej X13 jest silnie prawostronnie asymetryczny z dużym pikiem (dla wartości 0,1). Większość obserwacji osiąga wartości z przedziału 0,1-0,4. Mniej więcej powyżej wartości 1,4 występuje kilkanaście obserwacji odstających, mediana jest bliska 0.



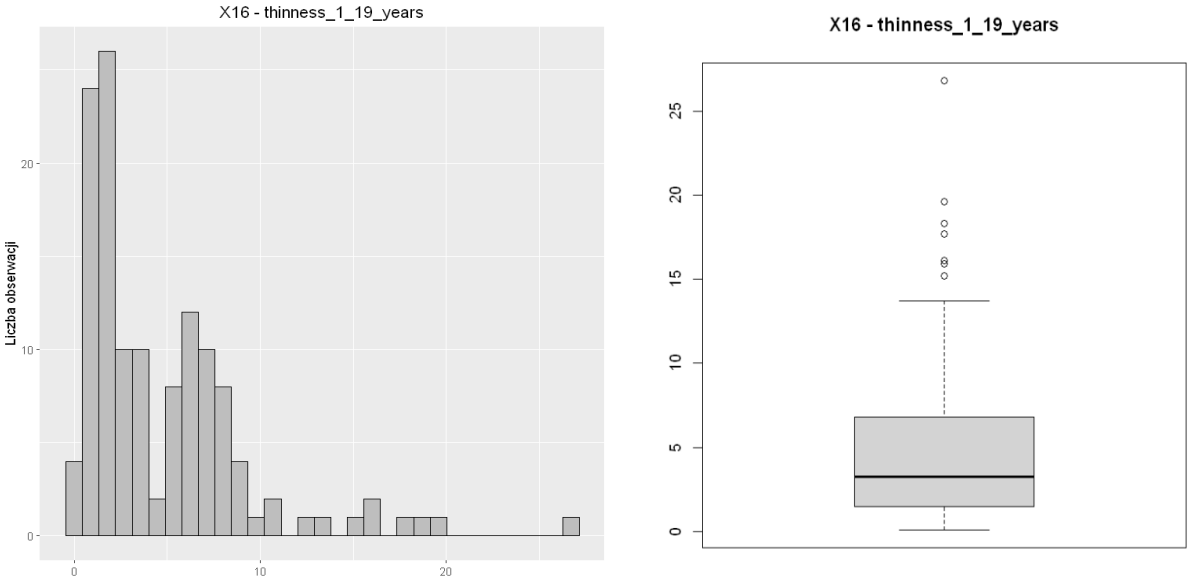
Wykres. 14. Histogram i boxplot dla zmiennej GDP.

Rozkład w przypadku zmiennej X14 także jest silnie prawostronnie asymetryczny, bardzo znacząca większość obserwacji osiąga wartości na przedziale 0-15000. Na podstawie wykresu pudełkowego ciężko oszacować medianę na ze względu na bardzo szeroki zakres wartości, natomiast jej dokładna wartość jest zawarta w statystykach opisowych. Od wartości powyżej około 14000 występują obserwacje odstające.



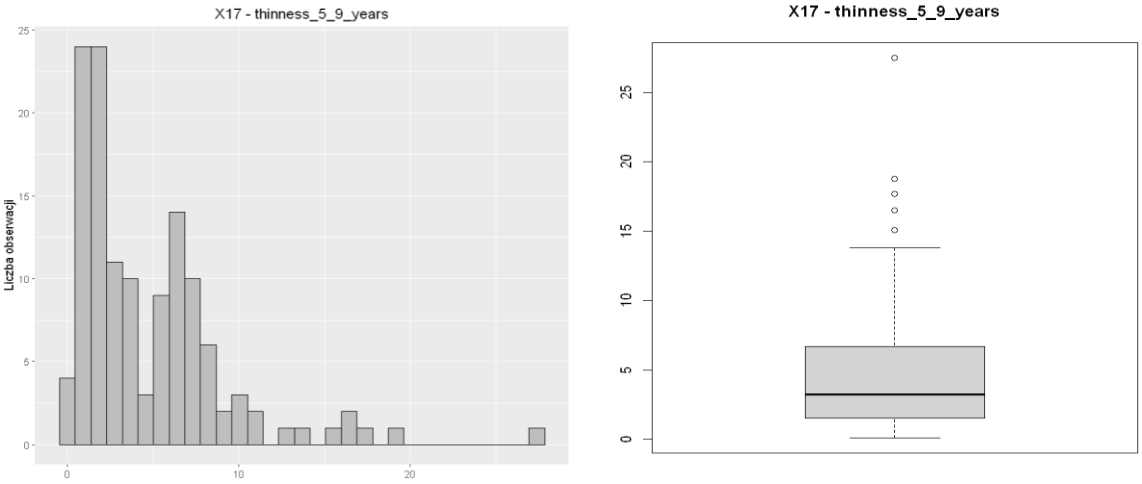
Wykres. 15. Histogram i boxplot dla zmiennej Population.

Rozkład zmiennej X15 również charakteryzuje prawostronna asymetryczność. Większość obserwacji stanowią wartości z przedziału 0-3,8e+07. Tutaj też ciężko odczytać medianę z wykresu pudełkowego ze względu na szeroki zakres. Powyżej około 3,8e+07 są obserwacje odstające.



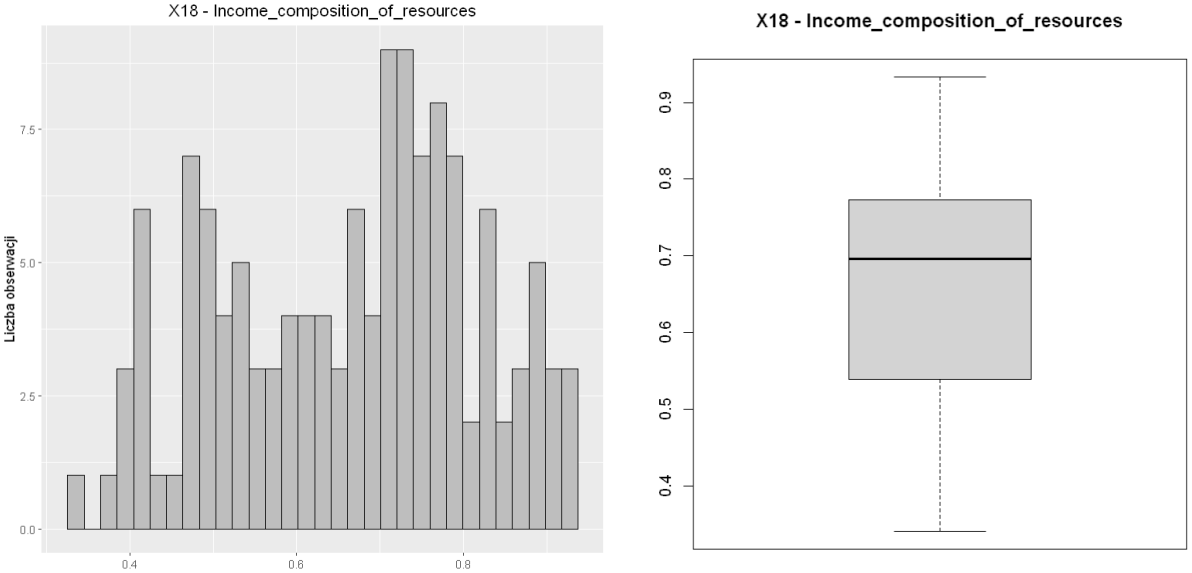
Wykres. 16. Histogram i boxplot dla zmiennej Population.

Rozkład zmiennej X16 ma prawostronną asymetrię z dwoma pikami (wartości między 1,5 i 2,5). Większość obserwacji zawiera się w przedziale 0-10. Powyżej wartości 14 jest kilka obserwacji odstających co widać na wykresie pudełkowym. Mediana o wartości około 3.



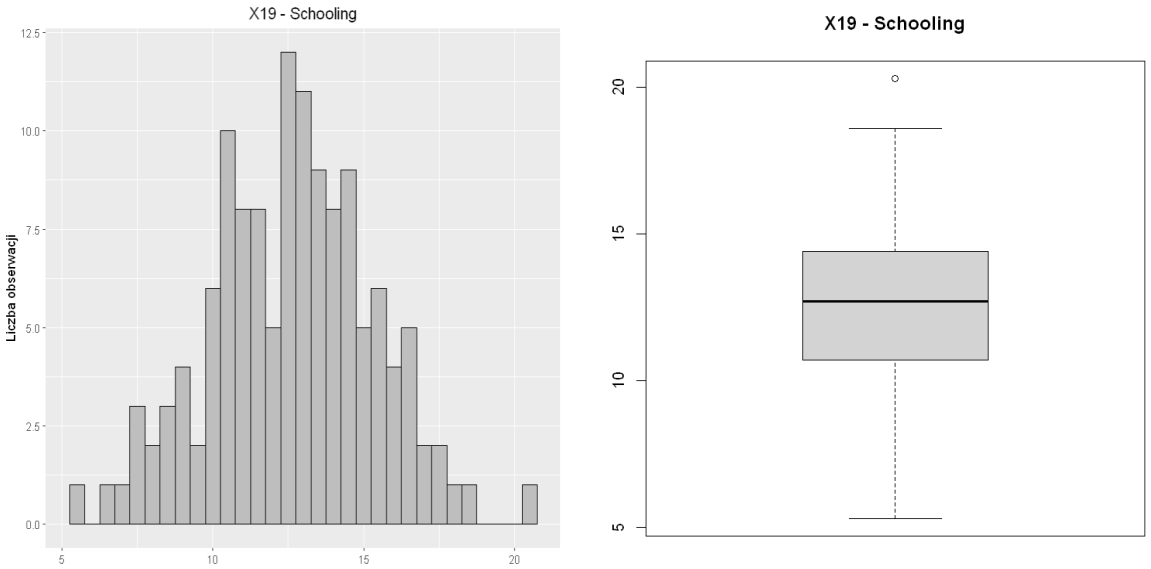
Wykres. 17. Histogram i boxplot dla zmiennej Population.

Rozkład oraz wykres pudełkowy dla zmiennej X17 są mocno zbliżone do zmiennej X16, co najprawdopodobniej świadczy o wysokiej korelacji między tymi zmiennymi.



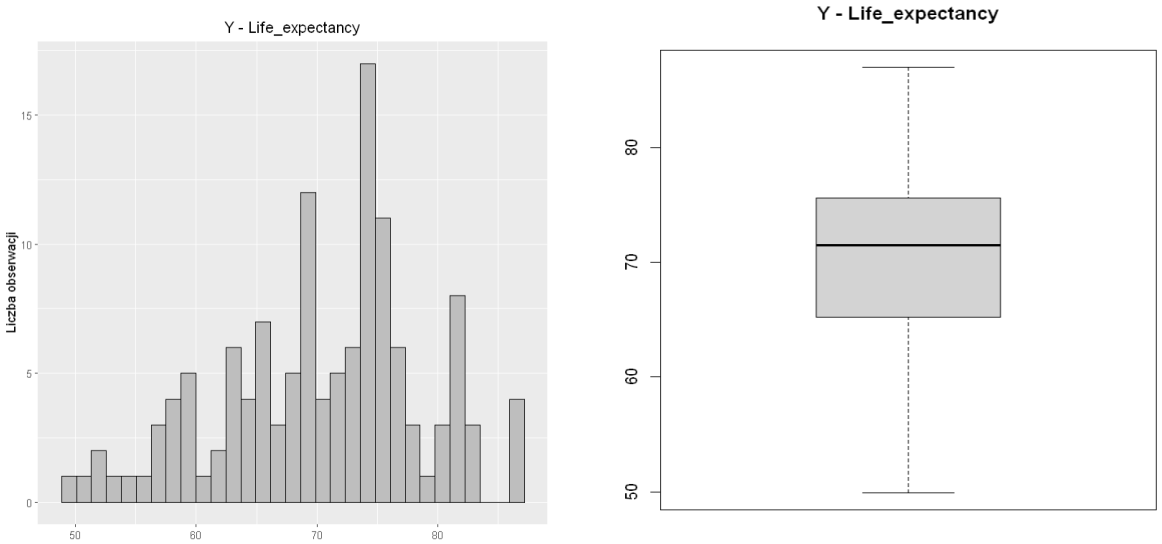
Wykres. 18. Histogram i boxplot dla zmiennej Population.

Rozkład zmiennej X18 charakteryzuje lewostronna asymetryczność, najwyższe piki osiągane dla wartości między 0,7-0,74. Większość obserwacji zawiera się w przedziale około 0,48-0,9. Mediana o wartości na poziomie 0,7 na podstawie wykresu pudełkowego oraz brak obserwacji odstających.



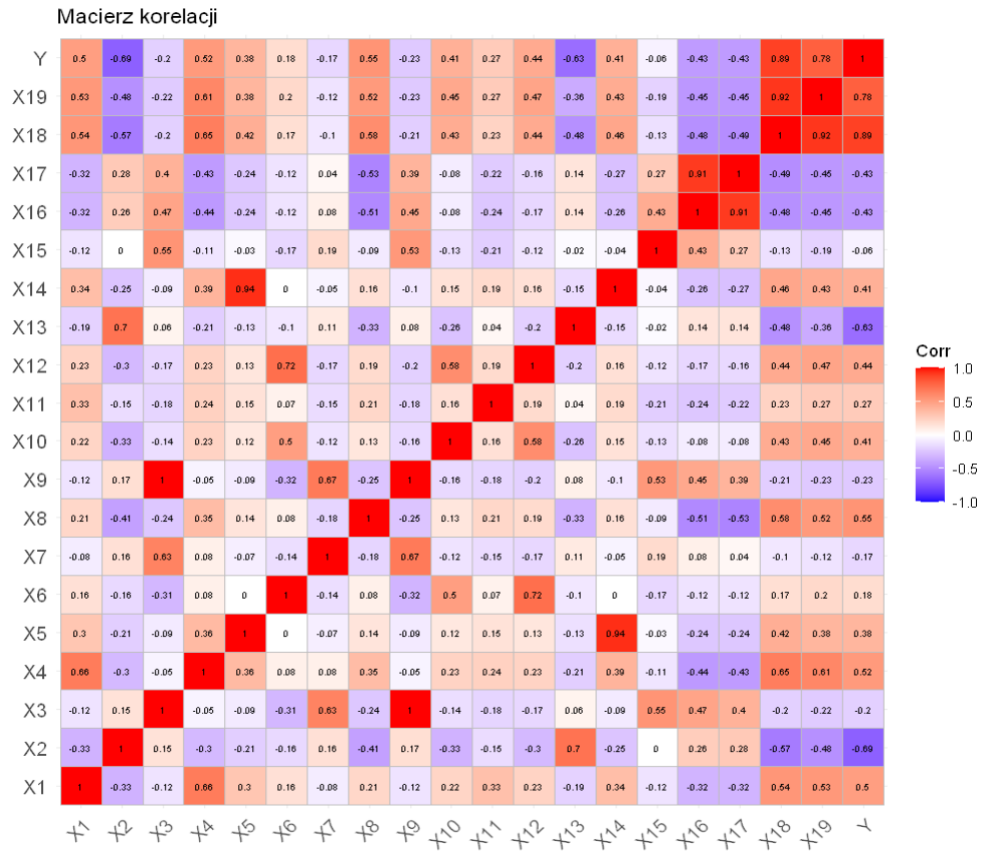
Wykres. 19. Histogram i boxplot dla zmiennej Schooling.

Tutaj dla zmiennej X19 można jej rozkład uznać za symetryczny i mocno zbliżony do normalnego. Najwyższy pik jest osiągany dla wartości około 12,5, która jest zbliżona do mediany, którą można odczytać z wykresu pudełkowego. Jak widać występuje jedna obserwacja odstająca powyżej wartości 20.



Wykres. 20. Histogram i boxplot dla zmiennej Life expectancy.

Rozkład zmiennej wyjściowej Y wykazuje lewostronną asymetryczność. Najwyższy pik osiągany dla wartości około 75. Większość obserwacji jest dla wartości z przedziału około 57-82. Na podstawie wykresu pudełkowego mediana w przybliżeniu wynosi 72 oraz nie występują obserwacje odstające.



Wykres 21. Macierz korelacji między zmiennymi.

Dla analizowanych danych wyznaczono również macierz korelacji między zmiennymi objaśniającym, a także zmienną objaśnianą oraz przedstawiono te wyniki na wykresie. Pojawiły się pary zmiennych silnie skorelowanych ze sobą, co uniemożliwiłoby ich użycie w regresji liniowej, jednak na tym etapie nie odrzucono żadnej z nich, ponieważ dobór zmiennych w pierwszej kolejności odbył się poprzez zastosowanie regresji Lasso dla pełnego zbioru cech, a dopiero następnie ewentualnym odrzuceniu zmiennych o silnej korelacji z innymi.

7.3. Braki danych

Pełna baza zawierała dane dotyczące 193 krajów, jednak pojawiały się wśród nich puste wartości, dlatego usunięto rekordy, w których one występowały. Po tym zabiegu pozostało 130 rekordów i to właśnie na podstawie tych państw przeprowadzono analizę.

7.4. Obserwacje odstające

Obserwacje odstające zostały zastąpione wartościami wąsów boxplotów odpowiednio wąsem górnym lub dolnym.

8. Opis metod wykorzystanych w pracy

8.1. Regresja liniowa

Regresja liniowa jest jedną z najprostszych i najczęściej stosowanych metod analizy statystycznej, używaną do modelowania relacji pomiędzy zmienną zależną (objaśnianą) a jedną lub większą liczbą zmiennych niezależnych (objaśniających). Celem regresji liniowej jest określenie, w jaki sposób zmienne objaśniające wpływają na zmienną objaśnianą oraz prognozowanie wartości zmiennej zależnej na podstawie znanych wartości zmiennych niezależnych.

Założenia regresji liniowej

Aby wyniki regresji liniowej były wiarygodne, muszą być spełnione następujące założenia:

1. Liniowość relacji: Relacja między zmienną zależną a każdą zmienną niezależną powinna być liniowa.
2. Homoskedastyczność: Rozproszenie reszt (odchyłek od linii regresji) powinno być stałe dla wszystkich wartości zmiennych niezależnych.
3. Normalność reszt: Reszty powinny mieć rozkład normalny.
4. Brak autokorelacji: Reszty nie powinny być skorelowane ze sobą.
5. Brak współliniowości: Zmienne niezależne nie powinny być silnie skorelowane między sobą.

Równanie regresji liniowej:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + + \beta_n X_n + \varepsilon$$

Regresja liniowa prosta:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Estymacja parametrów modelu

Współczynniki β_0, β_1, ..., β_n są szacowane przy użyciu metody najmniejszych kwadratów (OLS – Ordinary Least Squares), która minimalizuje sumę kwadratów reszt:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Zalety regresji liniowej

- Prostota i intuicyjność modelu.
- Łatwość interpretacji współczynników regresji.
- Możliwość szybkiej identyfikacji zależności między zmiennymi.

Wady regresji liniowej

- Wrażliwość na odstające wartości (outliers).
- Ograniczenia związane z założeniami liniowości i homoskedastyczności.
- Trudności w interpretacji przy współwystępowaniu wielokolinearności.

9. Wyniki przeprowadzonych badań

9.1. Dobór zmiennych regresją Lasso

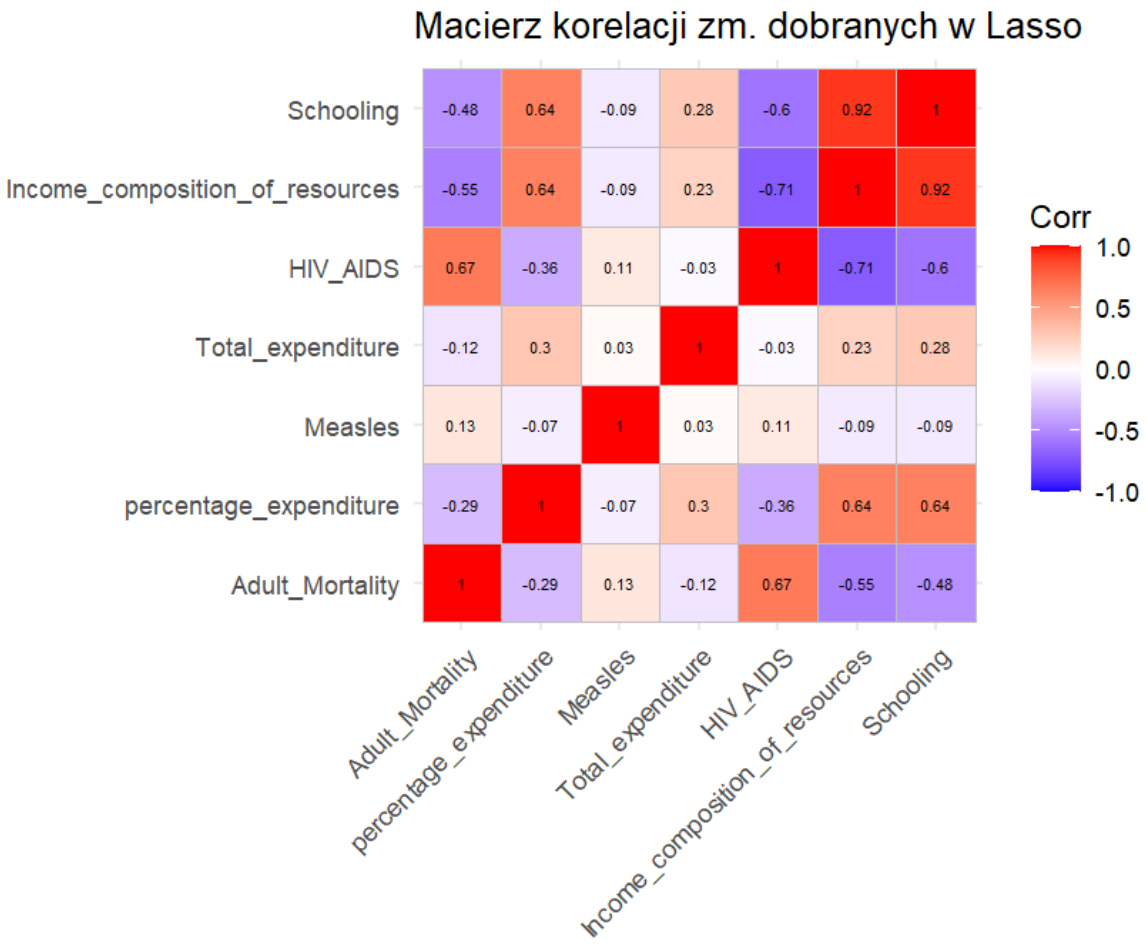
Obserwacje podzielono na zbiór treningowy i testowy (liczność analizowanego zbioru to 130 obserwacji). Doboru zmiennych do regresji liniowej dokonano na podstawie regresji Lasso. Zbudowano dla tej metody model z wszystkimi potencjalnymi zmiennymi objaśniającymi, a następnie wybrano te, których oszacowane współczynniki różniły się od zera. Wyniki tego działania przedstawiono poniżej.

Zmienne dobrane w metodzie Lasso (na podstawie danych treningowych):

```
20 x 1 sparse Matrix of class "dgCMatrix"
                                        s0
(Intercept)                     49.4307465162
Status                           .
Adult_Mortality                 -0.0147215856
infant_deaths                    .
Alcohol                          .
percentage_expenditure           0.0002073562
Hepatitis_B                      .
Measles                         -0.0005675578
BMI                              .
under_five_deaths                .
Polio                            .
Total_expenditure                0.1498303077
Diphtheria                       .
HIV_AIDS                        -2.8278067145
GDP                              .
Population                       .
thinness_1_19_years              .
thinness_5_9_years               .
Income_composition_of_resources 34.2159373386
Schooling                        0.0486402219
```

Tabela 4. Oszacowania współczynników w metodzie Lasso.

W dalszej części badania wykorzystano zatem następujące zmienne: Adult_Mortality, percentage_expenditure, Measles, Total_expenditure, HIV_AIDS, Income_composition_of_resources oraz Schooling. Sprawdzono jeszcze korelację między tymi zmiennymi, aby zapewnić, że odpowiednie zmienne mogą zostać użyte w regresji liniowej.



Wykres 22. Macierz korelacji między zmiennymi wybranymi w Lasso.

Ponieważ między zmiennymi Schooling oraz Income_composition_of_resources wystąpiła bardzo wysoka korelacja dodatnia, usunięto zmienną Schooling. Pozostałe zmienne wykazały się niskimi lub umiarkowanymi zależnościami, więc zostały uwzględnione w modelu regresji liniowej.

9.2. Regresja liniowa

Zwykła regresja liniowa dla dobranych zmiennych (na podstawie danych treningowych):

Wykres 23. Regresja liniowa dla dobranych zmiennych (dopasowanie do danych uczących).

```
Call:
lm(formula = Life_expectancy ~ ., data = train[, -7])

Residuals:
    Min      1Q  Median      3Q     Max
-6.7438 -1.7003  0.0399  1.4529  8.8616

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    49.5397054  2.3680385  20.920  < 2e-16 ***
Adult_Mortality                -0.0155441  0.0035792  -4.343  3.4e-05 ***
percentage_expenditure          0.0004615  0.0006777   0.681  0.49747
Measles                        -0.0016245  0.0012442  -1.306  0.19472
Total_expenditure               0.2438923  0.1170180   2.084  0.03971 *
HIV_AIDS                       -3.1862644  1.0362281  -3.075  0.00272 **
Income_composition_of_resources 34.5586524  3.2382448  10.672  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.772 on 99 degrees of freedom
Multiple R-squared:  0.8886,    Adjusted R-squared:  0.8819
F-statistic: 131.7 on 6 and 99 DF,  p-value: < 2.2e-16
```

Tabela 5. Wyniki oszacowania regresji liniowej.

Reszty:

- Min: -6.7438, Max: 8.8616 – wartości te wskazują na zakres reszt (różnice między wartościami przewidywanymi a rzeczywistymi).
- Mediana: 0.0399 – środkowa wartość reszt, co sugeruje, że model nie ma dużych błędów systematycznych.

Zmienne:

Spośród badanych zmiennych istotne okazały się:

- Stała (nie interpretujemy jej)
- Adult Mortality
- Total expenditure
- HIV AIDS
- Income composition of resources

R kwadrat – model tłumaczy 88.86% zmienności długości życia, co oznacza bardzo dobre dopasowanie, Im wyższy R^2, tym lepsze dopasowanie modelu do danych. Wartość bliska 1 oznacza, że model świetnie przewiduje zmienną zależną.

(Skorygowany R^2) = 0.8819, Wartość 0.8819 oznacza, że po uwzględnieniu liczby zmiennych w modelu jego wyjaśnialność nadal jest bardzo wysoka. Ponieważ skorygowany R^2 jest tylko nieznacznie niższy od R^2, sugeruje to, że zmienne w modelu są dobrze dobrane i nie dodano zbędnych predyktorów, które nie poprawiają jakości modelu.

F-statistic = 131.7, p < 2.2e-16, Bardzo wysoka wartość F (**131.7**) oraz ekstremalnie niski p-value (< 2.2e-16) wskazują, że przynajmniej jedna zmienna niezależna istotnie wpływa na długość życia. Niski p-value sugeruje, że model jako całość jest statystycznie istotny i nie powstał przypadkowo. Test statystyki F ocenia, czy cały model (czyli wszystkie zmienne niezależne razem) istotnie wpływa na zmienną zależną.

*** p < 0.001 (bardzo istotne)
** p < 0.01 (istotne)
* p < 0.05 (umiarkowanie istotne)
brak gwiazdek p > 0.05 (nieistotne)

<u>Interpretacja istotnych zmiennych.</u>
Adult Mortality - jeżeli liczba zgonów między 15 a 60 rokiem życie wzrośnie o jedną jednostkę, to oczekiwana długość życia spadnie o 0.01554 lat, przy założeniu ceteris paribus.

Total_expenditure - jeśli liczba wydatków sektora instytucji rządowych i samorządowych na zdrowie jako odsetek całkowitych wydatków rządowych wzrośnie o jedną jednostkę to oczekiwana długość życia wzrośnie o 0.2438 lat, przy założeniu ceteris paribus.

HIV_AIDS - jeśli liczba zgonów na HIV/AIDS wzrośnie o jedną jednostkę to szacowana długość życia spadnie o 3.1862 lat, przy założeniu ceteris paribus.

Income composition of resources - jeśli liczba wskaźnika rozwoju społecznego pod względem struktury dochodów zasobów wzrośnie o jedną jednostkę, to oczekiwana długość życia wzrośnie o 34.5586 lata, przy założeniu ceteris paribus.

Z wyżej zinterpretowanych zmiennych istotne są w stopniu:

Income_composition_of_resources 34.5586524

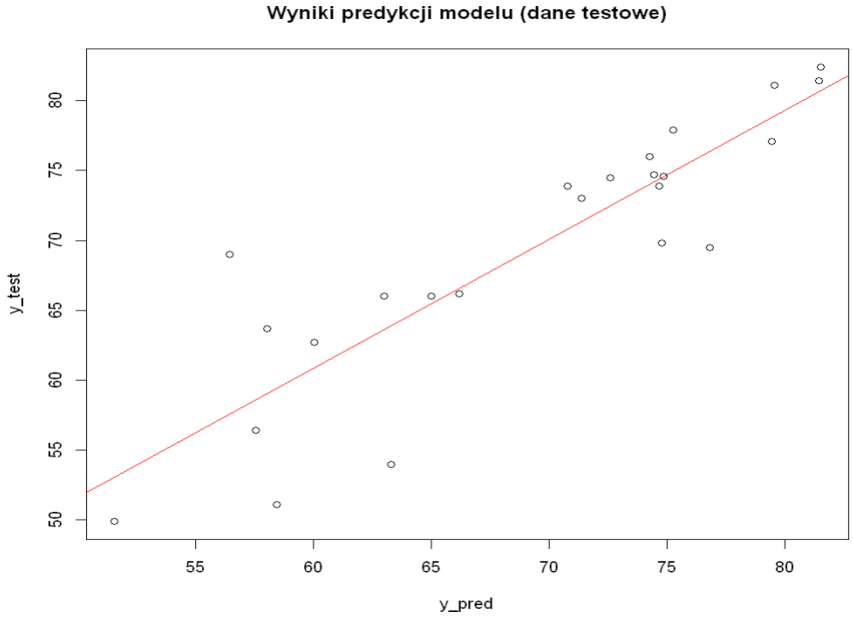
Adult_Mortality -0.0155441

HIV_AIDS -3.1862644

Total_expenditure 0.2438923

Według przeprowadzonej analizy największy wpływ na oczekiwaną długość życia ma Wskaźnik Rozwoju Społecznego w danym kraju. Im wyższa jest jego wartość, tym dłuższego średniego życia mieszkańców można się spodziewać. Drugim w kolejności wpływającym czynnikiem jest liczba zgonów z powodu HIV/AIDS. Ta zależność jest ujemna – im większa umieralność, tym niższa średnia długość życia.

Przetestowanie powstałego modelu liniowego na danych testowych:



Wykres 24. Przetestowanie modelu dla danych testowych.

Wartości numerycznych mierników jakości predykcji, błąd średniokwadratowy (MSE) i średni błąd bezwzględny (MAE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_{i,\,test} - y_{i,\,pred}\right)^2 = 19{,}39$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{i,\,test} - y_{i,\,pred}\right| = 3{,}08$$

Osiągnięta wartość błędu średniokwadratowego świadczy o tym, że występowały pewne znaczące odchylenia między predykcjami a wartościami rzeczywistymi, które zostały bardziej ukarane ze względu na występowanie kwadratu. Wartość tego błędu jest znacznie wyższa od

średniego błędu bezwzględnego co potwierdza tę zależność. Natomiast jeśli chodzi o interpretację średniego błędu bezwzględnego to można powiedzieć, że model na podstawie danych wejściowych średnio określa długość życia z błędem około 3 lat. Z drugiej strony jednak analizowany zbiór był dość nieliczny (130 obserwacji), gdzie zbiór treningowy stanowił 80% obserwacji, więc dokładność predykcji mogła być nieco ograniczona. Być może przy większej liczbie obserwacji otrzymane wyniki byłyby dokładniejsze.

9.3. Badanie założeń

Na regresję liniową nałożony jest szereg założeń, które pozwalają zidentyfikować poprawność zastosowanego modelu. Oszacowany model liniowy poddano ich badaniom, a wyniki przedstawione zostały poniżej.

9.3.1. Brak współliniowości

Współliniowość to sytuacja, w której zmienne objaśniające w modelu regresji są silnie skorelowane. Zostało to zweryfikowane na podstawie macierzy korelacji, więc można podejrzewać, że założenie to jest spełnione. Aby potwierdzić tą hipotezę wyliczono dodatkowo współczynniki wariancji inflacji (Variance Inflation Factor; VIF). Wartości VIF > 5 oznaczają wysoką współliniowość, która wskazuje konieczność zmiany modelu.

Adult_Mortality	percentage_expenditure
1.897863	1.792546
Measles	Total_expenditure
1.023186	1.142391
HIV_AIDS	Income_composition_of_resources
2.757438	3.137654

Tabela 6. Współczynniki VIF

Ponieważ dla wszystkich zmiennych uzyskane wartości są mniejsze od 5, stwierdzono brak współliniowości stale zakłócającej wyniki.

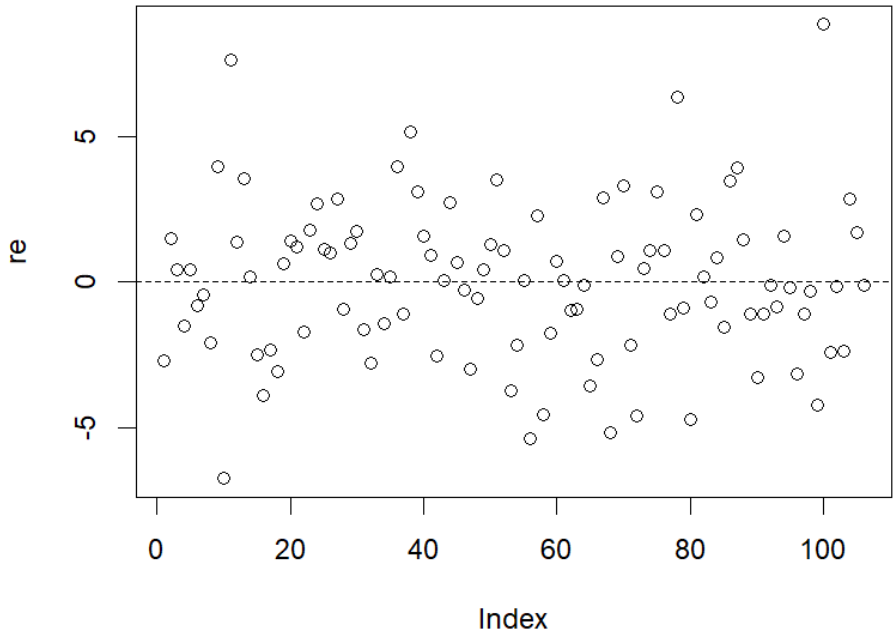
9.3.2. Wartości oczekiwane skł. los. są równe zero

Wartości oczekiwane składników losowych powinny być równe zero. Oznacza to, że zakłócenia, które reprezentują składniki losowe mają tendencje do wzajemnej redukcji.

$$E(\varepsilon_i) = -4.497355e^{-17} \approx 0$$

Średnia reszt, które w modelu reprezentują składnik losowy ma wartość bardzo niską, zbliżoną do zera, dlatego przyjęto, że założenie jest spełnione.

9.3.3. Losowość reszt modelu



Wykres 25. Wykres reszt uzyskanych z modelu

Analizując wykres reszt można zauważyć, że są one rozrzucone dość losowo, co może sugerować, że mają właśnie taki charakter. Aby to potwierdzić przeprowadzony został test serii, który polega na podzieleniu ciągu reszt na serie o takich samych znakach oraz obliczeniu statystyki na podstawie uzyskanych liczb serii. Wyniki tego testu w programie R zostały zaprezentowane poniżej.

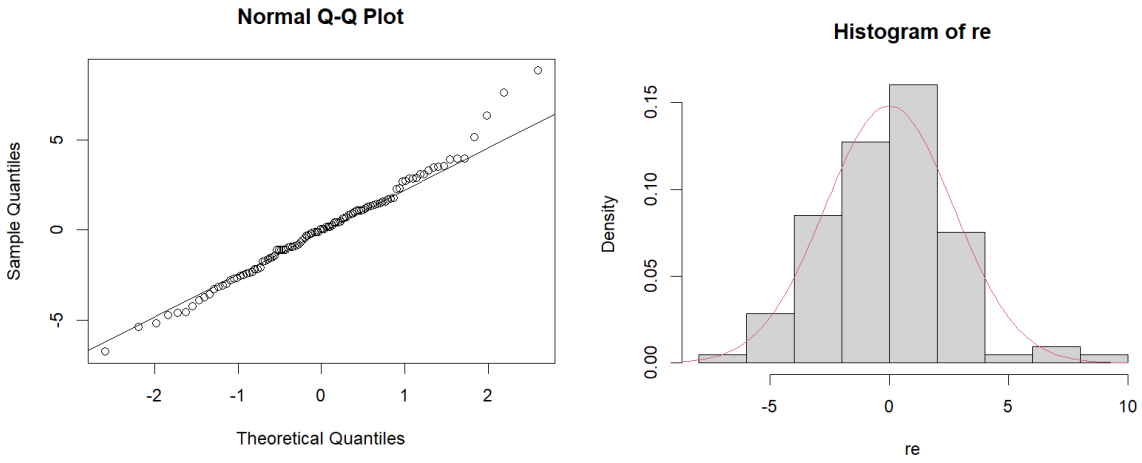
```
        Runs Test

data:  factor(sign(re))
Standard Normal = -0.58209, p-value = 0.5605
alternative hypothesis: two.sided
```

Hipoteza zerowa H_0 w tym teście głosi, że rozpatrywany ciąg reszt jest losowy. Ponieważ p-value > 0,05, to stwierdzono brak podstaw do odrzucenia H0, a więc założenie jest spełnione.

9.3.4. Normalność składnika losowego

Brak normalności rozkładu składnika losowego nie wpływa na utratę własności nieobciążoności, efektywności i zgodności estymatora MNK, jednak założenie to wpływa na interpretację istotności parametrów modelu. Jeśli reszty modelu nie mają rozkładu normalnego to nie należy przeprowadzać wnioskowania o istotności parametrów strukturalnych modelu na podstawie testu t-Studenta lub F-Snedecora.

Przed przystąpieniem do badania normalności składnika losowego testem statystycznym, reszty z modelu przedstawiono na wykresie kwantyl-kwantyl (Q-Q plot) oraz histogramie.



Wykres 26. Q-Q plot i histogram dla reszt modelu

Na wykresie kwantyl-kwantyl reszty w większości pokrywają się z linią, co może sugerować, że ich rozkład jest zbliżony do normalnego. Mniej oczywiste wnioski płyną z histogramu – większość słupków układa się w sposób bliski do rozkładu normalnego, jednak widać wyraźne odstępstwa.

W celu uzyskania odpowiedzi na to pytanie wykonano test normalności Shapiro-Wilka.

```
        Shapiro-Wilk normality test

data:  re
W = 0.98596, p-value = 0.331
```

Wartość p-value na poziomie 0,331 > 0,05 wskazuje na brak podstaw do odrzucenia hipotezy zerowej mówiącej o normalność rozkładu.

9.3.5. Homoskedastyczność

Model regresji liniowej ściśle wymaga, aby wariancje składnika losowego były stałe. Taką własność określa się jako homoskedastyczność. Założenie to zostało przebadane testem Breusch'a-Pagan'a, gdzie hipoteza zerowa mówi, że wariancja jest stała.

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.2711131, Df = 1, p = 0.60259
```

Uzyskane wyniki wskazują na brak podstaw do odrzucenia H_0., a zatem założenie stałej wariancji składnika losowego jest spełnione.

10. Podsumowanie i wnioski

Praca przedstawia analizę czynników wpływających na oczekiwaną długość życia w różnych krajach na podstawie regresji liniowej. Do analizy danych zastosowano regresję liniową oraz metodę Lasso w celu doboru istotnych zmiennych. Wyniki wskazują, że największy wpływ na długość życia mają: wskaźnik rozwoju społecznego, umieralność dorosłych, liczba zgonów z

powodu HIV/AIDS oraz wydatki na zdrowie. Model uzyskał wysoką wartość współczynika determinacji (0,8886), co świadczy o jego dobrej jakości dopasowania. Przeprowadzono także weryfikację założeń regresji, które zostały spełnione. Wnioski wskazują, że poprawa jakości życia i systemu opieki zdrowotnej, zwiększenie wydatków na zdrowie oraz redukcja umieralności dorosłych mogą istotnie przyczynić się do wydłużenia życia w populacji.

11. Bibliografia

[1] https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data

[2] Chen, Z., Ma, Y., Hua, J., Wang, Y., & Guo, H. (2021). Impacts from Economic Development and Environmental Factors on Life Expectancy: A Comparative Study Based on Data from Both Developed and Developing Countries from 2004 to 2016. *International Journal of Environmental Research and Public Health*, 18(16), 8559. https://doi.org/10.3390/ijerph18168559

[2] Jafrin, N., Masud, M.M., Seif, A.N.M., Mahi, M., & Khanam, M. (2021). A panel data estimation of the determinants of life expectancy in selected SAARC countries. *Operations Research and Decisions,* No. 4. DOI: 10.37190/ord210404

[3] Amos, B.K., & Smirnov, I. (2022). Determinants factors in Predicting Life Expectancy Using Machine Learning. Advanced Engineering Research (Rostov-on-Don), 22 (4), 373-383. doi: 10.23947/2687-1653-2022-22-4-373-383

[4] Wątroba, J. (2011). Prosto o doładowaniu prostych, czyli analiza regresji liniowej w praktyce https://media.statsoft.pl/_old_dnn/downloads/analiza_regresji_liniowej_w_praktyce.pdf