W jakim województwie najlepiej założyć rodzinę?

Opracowała Julia Taborek

Spis treści

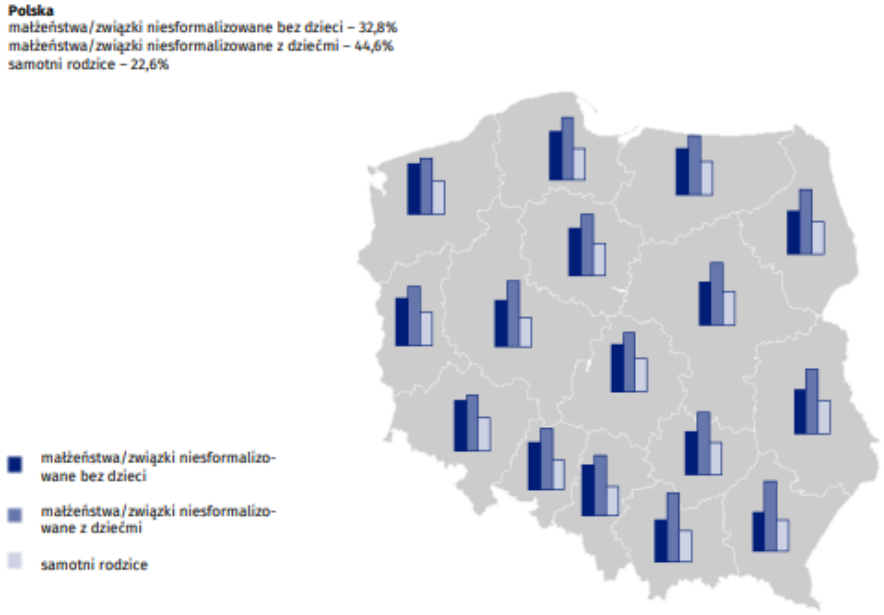
1. Wprowadzenie

Rodziny w Polsce na przestrzeni lat

Wraz z upływem lat w Polsce obserwuje się zmniejszanie się liczby rodzin. Według wyników Narodowego Spisu Powszechnego z 2021 roku liczba rodzin w Polsce liczyła 10 159,3 tysiąca. Porównując z wynikami z roku 2011 ich liczba zmniejszyła się o 813,2 tys. (7,4%). Redukcja dotyczyła głównie miast – tam odnotowano spadek o ponad 12% natomiast na wsi liczba rodzin wzrosła o 0,3%. Jest to skutek suburbanizacji (proces przemieszczania się ludności z obszarów miejskich do obszarów podmiejskich). W ciągu 10 lat zaobserwowano również spadek liczby małżeństw (w 2021 r. było ich ok. 7312,0 tys., tj. ponad 72% ogółu rodzin, natomiast w 2011 r. odpowiednio 8153,2 tys. co stanowiło 74,3% ogółu rodzin) spowodowany intensywnym wzrostem liczby par pozostających w związkach niesformalizowanych (liczba małżeństw zawartych w 2021 roku była o prawie 20% niższa niż liczba małżeństw zawarta w 2011r). Dodatkowo w okresie międzyspisowym zaobserwowano intensywny wzrost liczby rozwodów (odnotowano wzrost o ponad 800 tys. osób rozwiedzionych). Największe zmiany dotyczyły jednak małżeństw z dziećmi. Pomimo spadku o 1214,7 tys. (22,3%) w stosunku do roku 2011, grupa ta pozostała najczęstszym typem rodziny. Ponownie regres ten dotyczył głównie miast, gdzie rodziny z dziećmi reprezentują jedynie 37% struktury rodzin (na wsi stanowią prawie połowę). Opóźnił się również moment, w którym pary decydują się na pierwsze dziecko.

RODZINA

Dwie lub większa liczba osób, które są związane jako mąż i żona, wspólnie żyjący partnerzy (kohabitanci) - osoby płci przeciwnej lub jako rodzic i dziecko. Tak więc, rodzina obejmuje parę bez dzieci lub parę z jednym lub większą liczbą dzieci, albo też samotnego rodzica z jednym lub większą liczbą dzieci." (GUS)



Wykres 1. Rodziny według typów i województwa w 2021r. (w %)
Źródło: „Rodziny wyniki wstępne -NSP 2021" GUS

Sytuacja w województwach

W 2021r. największy odsetek rodzin z dziećmi odnotowano w województwach: podkarpackim (50,0%), małopolskim (48,7%) i wielkopolskim i (47,1%). Wskaźnik dla całego kraju spadł z poziomu 51,3% do 44,6%. W ciągu tych 10lat nie zmienili się liderzy. Najmniej rodzin z dziećmi zamieszkuje województwa dolnośląskie i zachodniopomorskie (po 40,3%) oraz lubuskie i łódzkie (42,3%).

Zmiany demograficzne

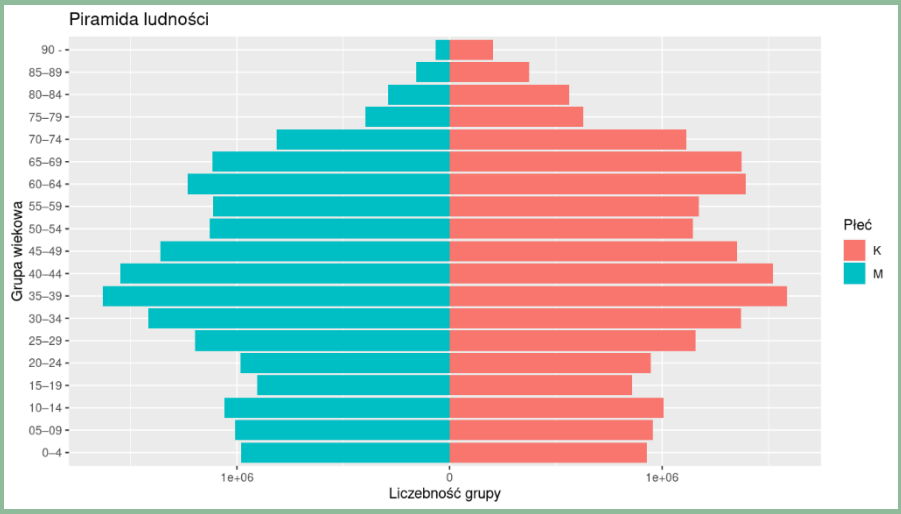
Wyniki spisu z 2021r. ukazały również spadek liczby ludności w Polsce o 476 tys. (1,2%) w odniesieniu do roku 2011. Natomiast na koniec 2022r. ludność Polski wynosiła o 141 tys. osób mniej niż na końcu roku poprzedniego. Największe różnice pomiędzy badaniami możemy zaobserwować w strukturze wieku ludności. Mediana wieku statystycznego mieszkańca Polski to 42 lata (w 2021r. – 41,9 jednak w 2022r. już 42,3). W czasie między spisami polskie

społeczeństwo postarzało się średnio o 4 lata. W skutek niekorzystnych trendów demograficznych zmniejszył się udział ludności w wieku przedprodukcyjnym. Głównym czynnikiem wpływającym na obniżenie liczby jednostek w tym wieku była niska liczba żywych urodzeń, która zmieniała się od 386 tys. w 2012 r. do 355 tys. w 2020 r. (najwięcej wynosiła w 2017r. – 401tys.). Liczba osób w wieku produkcyjnym jak również odsetek osób w tej grupie uległy zmianie na korzyść osób w wieku poprodukcyjnym. Procent osób w tej społeczności wzrósł z 16,9% do 22,3%. Oznacza to, że w ciągu dekady przybyło ponad 2 mln osób w grupie wiekowej 60 – 65 lat a średnio co piąty Polak ukończył już 60 lat.
W wyniku opisanych powyżej zmian zanotowano wzrost współczynnika obciążenia demograficznego w ekonomicznych grupach wieku[1].
W 2021r. na 100 osób w wieku produkcyjnym przypadało średnio 69 osób w wieku nieprodukcyjnym, prawie o 14 osób więcej niż dekadę wcześniej. W kolejnym roku liczba ta wzrosła o jedną jednostkę. Różnice pomiędzy wpisami ukazują szybki proces starzenia się ludności Polski wynikający z pozytywnego zjawiska jakim jest dłuższe trwanie w zdrowiu i życiu oraz spowodowany niskim poziomem dzietności. Ten proces najlepiej obrazują wskaźniki przedstawione w poniższej tabeli (*Tabela 1*).

Tabela 1. Wskaźniki obrazujące proces starzenia się społeczeństwa.

Rok	Mediana wieku	Wskaźnik starości demograficznej	Indeks starości	Współczynnik obciążenia demograficznego
2011	38	14	90	40
2021	42	19	119	52

Wraz z nasilaniem się tego procesu liczba osób młodych, zdolnych zastąpić osoby w wieku produkcyjnym będzie maleć, zaś liczba osób starszych potrzebujących pomocy i wsparcia będzie rosnąć. Problem ten nie dotyczy jedynie naszego kraju - na podstawie raportu ONZ możemy się spodziewać, iż liczba osób na świecie powyżej 65 roku życia podwoi się do 2050 roku. Jednakże Polska należy do krajów dla których ta zmiana będzie szczególnie drastyczna. Według Marka Okólskiego z jednego z najmłodszych krajów Unii Europejskiej w 2060r. Polska stanie się najstarszym.



Wykres 2. Piramida ludności Polski.
Źródło: Opracowanie własne wykonane w programie RStudio

Struktura ludności Polski ma typ regresywny – możemy to zaobserwować po wąskiej podstawie piramidy. Typ ten odzwierciedla społeczeństwo, w którym mamy do czynienia z malejącą liczbą urodzeń. Najliczniejszą grupę stanowią osoby w wieku 35-39 lat oraz 40-44 . Kolejną liczną grupę tworzą obywatele w wieku 60-64 oraz 65-69 lat.

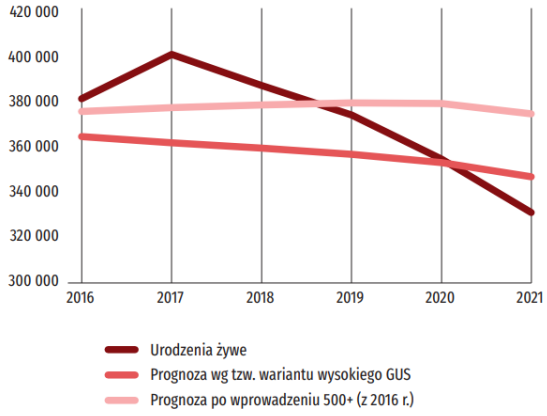
[1] Współczynnik obciążenia demograficznego w ekonomicznych grupach wieku - stosunek liczby osób w wieku nieprodukcyjnym (wiek 0-17 lat oraz 60 i więcej dla kobiet i 65 i więcej dla mężczyzn) do liczby osób w wieku produkcyjnym.

Przyczyny niskiej dzietności

Jednym z powodów starzenia się społeczeństwa jest niska dzietność w Polsce. Niezmiennie od lat 90 XX wieku utrzymuje się ona na małym poziomie, mimo to nieprzerwanie około połowa ankietowanych chciałaby mieć 2 dzieci. Jednak pośród rodzin z dziećmi przeważają te wychowujące jedynaków. Występuje zjawisko *Fertility Gap* – różnica pomiędzy liczbą dzieci jakie chciałaby mieć kobieta a współczynnikiem dzietności. Jako jeden z powodów niskiej dzietności wskazuję się fakt, że Polacy rzadko decydują się na 2 i 3 dziecko. W kolejnych pokoleniach rośnie procent osób nie posiadających potomstwa. Według opinii publicznej głównymi powodami tego zjawiska są czynniki ekonomiczne oraz poczucie niepewności. Najczęściej wskazywane są: brak stabilności finansowej i ograniczenia związane z sytuacją mieszkaniową. Wśród odpowiedzi pojawiało się również brak poczucia wsparcia ze strony państwa między innymi w zakresie edukacji i opieki medycznej.

Polityka rodzinna w Polsce

Od początku XXI w. w Polsce prowadzane są działania z zakresu polityki rodzinnej między innymi został wprowadzony urlop ojcowski, wydłużony urlop macierzyński oraz Karta Dużej Rodziny. W ostatnim latach najdroższym instrumentem polityki rodzinnej był program „Rodzina 500+". Program nie spełnił jednak głównego celu, który mu przypisywano – nie wpłynął trwale na wzrost dzietności. Szacowano, że projekt przyczyni się do sumarycznego wzrostu liczby urodzeń w ciągu dekady o 278tys. Jednak po 5 latach (w 2021r.) mogliśmy zauważyć, że liczba urodzeń spadała za wyjątkiem roku 2017r. Jednak warto zaznaczyć, że w ostatnich latach odnotowuje się wzrost udziału dzieci, które są trzecim i kolejnym potomkiem, podczas gdy obserwuje się przeciwny trend w przypadku dzieci urodzonych jako pierwsze i drugie. W grupach tych ostatnich, zdaje się, że świadczenia wychowawcze nie przyniosły oczekiwanego efektu, z wyjątkiem pojedynczego wzrostu w 2017 roku w przypadku drugiego potomstwa. Realizacja programu miała natomiast pozytywny wpływ na zmniejszanie poziomu ubóstwa - część funduszy wypłacanych w ramach programu trafia do rodzin będących w złej sytuacji finansowej.



Wykres 3. Liczba urodzeń żywych w latach 2016 – 2021
Źródło: „Skutki świadczenia 500+" Izabela Bień

Cel pracy

Skutki starzejącego się społeczeństwa:

- Wzrost obciążenia ekonomicznego ludności,
- Problemy emerytalne,
- Wzrost kosztów służby zdrowia,
- Niedobór siły roboczej,
- Niski popyt na innowacje,
- Spowolnienie wzrostu gospodarczego,
- Obniżenie tempa wzrostu PKB

Temat poprawienia dzietności i liczebności rodzin z dziećmi w Polsce jest szczególnie ważny ze względu na szybkość starzenia się społeczeństwa i konsekwencje gospodarcze tego procesu.

Programy rządowe nie przynoszą oczekiwanego rezultatu w tym temacie. Celem pracy jest znalezienie najlepszego województwa dla rodzin z dziećmi biorąc pod uwagę obawy, z którymi mierzą się Polacy przed podjęciem decyzji o założeniu rodziny.

Zmienne wykorzystane do swojej analizy (*Tabela 2)* zostały podzielone na kategorie odpowiadające podstawowym potrzebom dziecka: opiekę zdrowotną, edukację, kulturę czy bezpieczeństwo. Aby ocenić poziom służby zdrowia w poszczególnych

województwach wzięto pod uwagę zarówno umieralność okołoporodową, liczbę dzieci na jakie przypada 1 łóżko w szpitalu na oddziale pediatrycznym jak również porównano liczbę pracujących położnych przypadających na 10 000 kobiet w wieku produkcyjnym, aby sprawdzić jak wygląda ochrona zdrowia już podczas ciąży. W tej kategorii danych znalazły się tez informacje o zanieczyszczeniach, ponieważ mają one znaczący wpływ na zdrowie dzieci – powodują alergie, choroby skóry oraz osłabienie płuc. W kolejnej grupie zmiennych sprawdzono dostęp do edukacji od placówek przedszkolnych po szkoły średnie, ponieważ edukacja ma kluczowe znaczenie w rozwoju dziecka oraz nawiązywaniu pierwszych relacji przyjacielskich. Uwzględniono również poziom bezpieczeństwa w analizowanych regionach na podstawie liczby przestępstw stwierdzone przez policję przeciwko rodzinie i opiece oraz wskaźnikowi wykrywalności sprawców. Następna grupa zmiennych (*x13, x14*) odnosi się do kultury, która jest ważną częścią życia każdego człowieka, ponieważ dzięki niej rozwijamy się i poszerzamy swoje horyzonty. Kolejne dwie grupy zmiennych związane z gospodarstwem domowym i pieniędzmi odwołują się do obaw związanych z posiadaniem dziećmi przez obywateli: braku stabilności finansowej oraz sytuacji mieszkaniowej. W zbiorze zostały również umieszczone zmienne odpowiadające za przeciętne wynagrodzenia i wydatki jak również wskaźnik zagrożenia ubóstwem oraz odsetek osób, które w ankietach wskazały, że trudno wiążą koniec z końcem. Ostatnia kategoria zmiennych to transport, gdzie zawarto informację o wypadkach drogowych oraz autobusowej, która często jest kluczowa podczas dojazdu dzieci do szkół.

Źródłami danych był Bank Danych Lokalnych Głównego Urzędy Statystycznego oraz Europejskiego Badania Warunków Życia Ludności z 2021r. Większość rozpatrywanych danych pochodzi z roku 2022, dane z roku poprzedniego zostały oznaczone *. Dane zostały przekształcone, aby zamiast wielkości absolutnych otrzymać wskaźniki w tym samym rzędzie – zakresie 0-100.

(fot. shutterstock.com)

Zmienne wejściowe

Tabela 2. Zmienne wejściowe.

Województwo	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	x21	x22	x23	x24	x25	x26
Dolnośląskie	19,0	39	62	22,91	54,60	11,11	7	51,38	29,95	16,45	10,10	63,0	10	20	70,8	86,7	93,90	85,00	15,61	22,13	20,0	13,8	9,9	69,45	14,93	64,6
Kujawsko-pomorskie	20,2	50	69	25,79	45,84	11,28	10	53,62	35,03	20,38	11,60	77,2	11	19	71,7	87,5	94,00	86,20	12,52	19,32	37,0	14,1	20,4	58,89	16,10	38,3
Lubelskie	20,0	43	55	32,77	61,39	13,57	4	66,28	44,90	18,67	17,70	75,5	12	28	79,2	87,3	87,40	82,40	12,46	19,07	37,9	24,8	22,9	59,10	10,45	38,7
Lubuskie	20,1	35	52	20,61	58,57	11,56	4	49,49	35,51	18,57	19,30	67,8	8	24	73,4	79,8	95,20	86,20	13,00	21,27	28,4	12,2	12,8	60,14	12,05	50,8
Łódzkie	19,1	40	56	31,47	68,49	11,05	8	55,88	36,91	17,74	10,80	65,3	8	19	71,8	87,4	89,90	82,30	13,54	20,86	29,8	19,6	14,5	62,11	12,13	92,6
Małopolskie	21,2	28	49	26,36	59,32	11,75	7	70,98	44,07	16,13	7,60	74,8	13	21	80,6	88,7	95,20	86,20	10,85	20,39	31,0	10,5	14,9	68,25	4,14	64,9
Mazowieckie	21,2	31	52	27,35	58,41	11,68	5	67,05	34,64	16,84	13,90	62,7	5	17	73,9	85,4	93,80	88,40	15,08	24,50	31,4	12,0	20,7	79,13	19,38	52,5
Opolskie	18,3	24	66	23,61	55,18	8,90	11	53,16	42,55	18,57	8,70	68,8	20	32	82,0	89,1	95,90	85,00	13,67	16,91	25,6	8,9	10,1	61,34	15,37	46,3
Podkarpackie	20,7	39	59	31,74	61,85	9,27	5	72,29	51,13	17,36	8,80	68,1	16	32	84,5	88,2	92,90	84,60	10,05	17,04	34,2	14,3	17,6	56,63	19,10	51,0
Podlaskie	20,0	39	69	31,21	68,92	7,42	2	58,86	36,12	17,93	9,10	80,9	16	20	78,2	81,2	90,70	83,60	11,00	20,05	28,0	12,1	19,6	60,13	10,67	29,1
Pomorskie	21,7	46	58	21,83	48,55	10,14	3	66,49	32,65	16,54	13,00	65,2	12	13	73,2	88,9	95,80	88,60	14,69	18,36	24,4	15,8	13,1	66,97	15,83	67,5
Śląskie	19,2	48	59	27,15	61,06	12,81	29	45,92	32,97	17,55	15,60	80,0	8	17	72,7	87,0	95,70	84,50	14,21	21,80	23,3	17,3	11,2	67,28	7,89	45,5
Świętokrzyskie	18,7	37	49	31,48	53,64	8,73	10	67,05	46,09	20,71	17,40	75,5	9	21	77,6	82,1	88,20	82,90	10,77	18,65	34,3	11,7	13,4	57,83	7,47	54,9
Warmińsko-mazurskie	20,3	46	71	21,70	64,77	7,73	3	57,16	39,67	20,20	15,10	74,9	10	21	69,6	83,5	94,00	87,00	12,21	20,69	43,6	19,4	19,4	56,75	5,53	60,6
Wielkopolskie	21,5	35	59	28,64	54,99	12,19	5	57,25	37,18	17,09	12,60	72,2	9	18	81,9	84,5	95,50	87,10	12,48	18,81	30,2	18,6	16,2	60,20	11,77	65,8
Zachodniopomorskie	19,1	44	55,83	23,79	57,36	9,00	6	51,69	32,49	19,50	14,00	71,5	20	21	70,1	82,8	96,10	87,90	12,89	20,29	30,6	17,1	12,0	61,70	12,75	53,6

Ogólne

- *x1* – Procent ludności jaki stanowią dzieci i młodzież

Zdrowie

- *x2* - Zgony niemowląt na 10 000 żywych urodzeń
- *x3* - Umieralność okołoporodowa na 10 000 urodzeń żywych i martwych *
- *x4* - Liczba pracujących położnych przypadających na 10 000 kobiet w wieku produkcyjnym *
- *x5* - Liczba przychodni na 100 tys. ludności *
- *x6* - Liczba dzieci na jakie przypada 1 łóżko w szpitalu na oddziale pediatrycznym (w setkach)

Edukacja

- *x8* - Liczba placówek wychowania przedszkolnego na 100 tys. ludności
- *x9* - Liczba szkół podstawowych na 100 tys. ludności
- *x10* - Liczba szkół średnich na 100 tys. ludności

Bezpieczeństwo

- *x11* - Przestępstwa stwierdzone przez Policję przeciwko rodzinie i opiece na 10 000 mieszkańców
- *x12* - Wskaźnik wykrywalności sprawców przestępstw stwierdzonych przez policję

Kultura

- *x13* – Liczba domów kultury, centrum kultury, świetlic na 100 tys. ludności
- *x14* – Liczba bibliotek publicznych na 100 tys. ludności

Gospodarstwo domowe

- *x15* - Przeciętna powierzchnia użytkowa mieszkania w m^2
- *x16* - Gospodarstwa wyposażone w urządzenie z dostępem do Internetu w % ogółu gospodarstw *
- *x17* - Procent mieszkań wyposażonych w łazienkę
- *x18* - Procent mieszkań wyposażonych w centralne ogrzewanie
- *x19* - Przeciętne miesięczne wydatki na 1 osobę w gospodarstwach domowych w setkach
- *x20* - Przeciętny miesięczny dochód na 1 osobę w gospodarstwie w setkach
- *x21* - % gospodarstw domowych deklarujących brak możliwości realizacji potrzeby tygodniowego wypoczynku rodziny raz w roku *
- *x22* - % gospodarstw domowych, które określiły, że przy aktualnym dochodzie z trudnością "wiążą koniec z końcem"

Finanse

- *x23* - Wskaźnik zagrożenia ubóstwem po transferach społecznych w % *
- *x24* - Przeciętne miesięczne wynagrodzenie brutto w

Transport

- *x25* - Linie autobusowe w km na 100km^2
- *x26* - Wypadki drogowe na 100 tys. ludności

2. Analiza wstępna

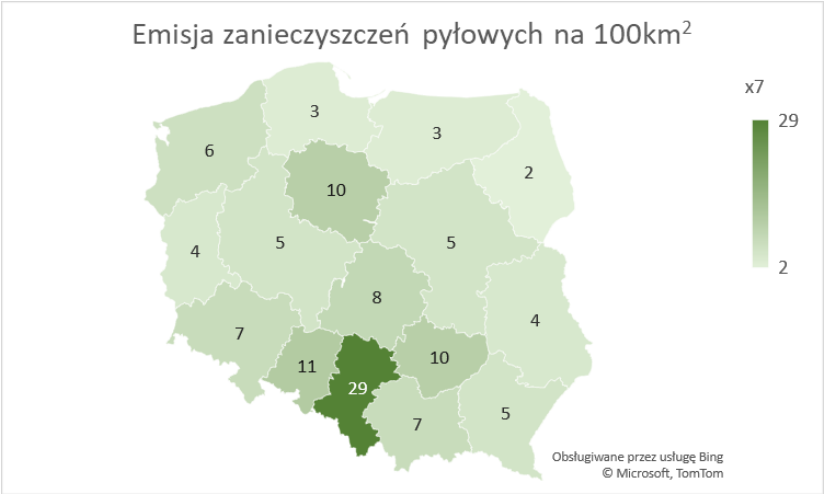
Podstawowe statystyki

Tabela 3. Podstawowe statystyki.

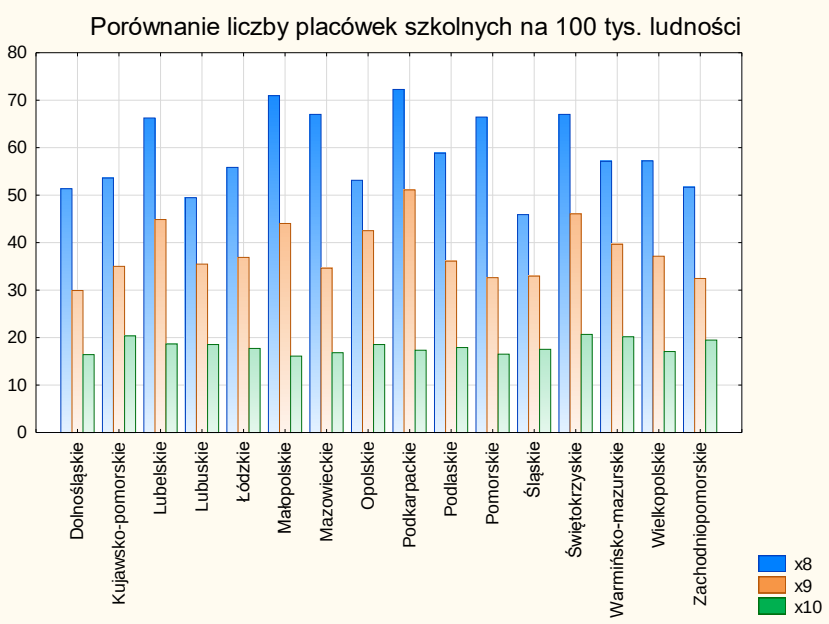
Statystyki	Zmienna																								
	x1	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*	*x12*	*x13*	*x14*	*x15*	*x16*	*x17*	*x18*	*x19*	*x20*	*x21*	*x22*	*x23*	*x24*	*x25*	*x26*
Średnia	20,02	39,00	58,80	26,78	58,31	10,51	7,44	59,03	38,24	18,14	12,83	71,46	11,69	21,44	75,70	85,63	93,39	85,49	12,82	20,01	30,61	15,14	15,53	62,87	12,22	54,79
Mediana	20,05	39,00	58,50	26,76	58,49	11,08	5,50	57,20	36,52	17,84	12,80	71,85	10,50	20,50	73,65	86,85	94,00	85,60	12,71	20,17	30,40	14,20	14,66	60,77	12,09	53,05
Minimum	18,31	24,00	49,00	20,61	45,84	7,42	2,00	45,92	29,95	16,13	7,60	62,70	5,00	13,00	69,60	79,80	87,40	82,30	10,05	16,91	20,00	8,90	9,92	56,63	4,14	29,10
Maksimum	21,67	50,00	71,00	32,77	68,92	13,58	29,00	72,29	51,13	20,71	19,30	80,90	20,00	32,00	84,50	89,10	96,10	88,60	15,61	24,50	43,60	24,80	22,88	79,13	19,38	92,60
Dolny	19,08	35,00	53,50	23,26	54,80	8,95	4,00	52,42	33,81	16,96	9,60	66,55	8,50	18,50	71,75	83,15	91,80	84,05	11,61	18,73	26,80	12,05	12,37	58,99	9,17	45,90
Górny	20,97	45,00	64,00	31,34	61,62	11,71	9,00	66,77	43,31	19,09	15,35	75,50	14,50	22,50	79,90	87,85	95,60	87,05	13,94	21,07	34,25	17,95	19,51	67,12	15,60	64,75
Odch.std	1,05	7,25	7,00	4,10	6,31	1,82	6,35	8,21	5,92	1,46	3,56	5,82	4,35	5,24	4,83	2,96	2,82	2,04	1,62	1,95	5,96	4,13	4,11	5,95	4,48	14,90
Wsp.zmn.	5,23	18,59	11,91	15,30	10,82	17,33	85,32	13,91	15,48	8,03	27,70	8,15	37,19	24,45	6,37	3,45	3,02	2,39	12,64	9,72	19,46	27,29	26,43	9,47	36,67	27,20
Skośność	0,08	-0,48	0,38	0,04	-0,14	-0,19	2,88	0,21	0,74	0,46	0,26	0,00	0,75	0,96	0,44	-0,65	-1,13	-0,10	0,00	0,45	0,35	0,71	0,30	1,49	-0,15	0,75
Gini	2,88	10,10	6,50	8,45	6,85	9,51	36,00	7,62	8,31	4,40	15,30	4,51	19,80	12,40	3,47	1,87	1,55	1,32	6,95	5,22	10,50	14,70	14,60	4,74	20,10	14,20

Powyżej (*Tabela 3*) zaprezentowano podstawowe statystyki zmiennych wykorzystanych do analizy. Można zaobserwować, że średnio dzieci i młodzież stanowią 20% ludności województw (*x1*). Najmniejszy odsetek reprezentują w województwie opolskim. Jednak województwa Polski są mało zróżnicowane pod tym względem, zmienna ta waha się od 18,311 do 21,67%.

Ważnym aspektem przy wyborze miejsca zamieszkania jest opieka zdrowotna w regionie. W 2021r. w Polsce średnio umierało 39 noworodków na 10 000 żywych urodzeń (w tym samym roku średnia w Europie wynosiła 32). W połowie województw na 10 000 kobiet w wieku produkcyjnym (*x4*) przypada od 23 do 31 położnych. Natomiast liczba przychodni na 100tys. ludności oscyluje pomiędzy 45,842 w województwie kujawsko-pomorskim a 68,920 w podlaskim. Największa liczba dzieci na jakie przypada jedno łóżko na oddziale pediatrycznym jest w województwie lubelskim. Najwyższa zmienność pomiędzy poszczególnymi regionami objawia się w przypadku emisji zanieczyszczeń pyłowych na 100km^2 (*Wykres 4*). Mediana tej zmiennej wynosi 5,5. Największą rozbieżność obserwuje się w przypadku województwa śląskiego, gdzie emisja zanieczyszczeń jest najwyższa. Mamy tutaj do czynienia z rozkładem prawoskośnym (prawostronnie asymetrycznym).



Wykres 4. Emisja zanieczyszczeń pyłowych na 100km^2.
Źródło: opracowanie własne w programie Excel



Wykres 5. Porównanie liczby placówek szkolnych.
Źródło: opracowanie własne wykonane w programie Statistica

Przyglądając się danym dotyczącym edukacji można dostrzec, że wraz ze wzrostem stopnia edukacji maleje średnia liczba placówek na 100 tys. ludności. Na wykresie (*Wykres 5*) możemy zauważyć, że ta zależność spełniona jest w każdym województwie.

Najwyższy wskaźnik przestępstw przeciwko rodzinie i opiece na 10 000 mieszkańców wynosi 19,3. Średnio wykrywalni są sprawcy 71,43% przestępstw stwierdzonych przez policję, przy czym w połowie województw wykrywa się między 66,55 a 75,5%.

Średnia liczba bibliotek publicznych na 100 tys. ludności jest prawie 2 razy większa od średniej liczby domów kultury, centrum kultury, świetlic na 100 tys. ludności. Co więcej średnio w województwach jest więcej bibliotek niż szkół średnich.

Województwa są mało zróżnicowane pod względem przeciętnej powierzchni użytkowej mieszkania oraz odsetku mieszkań wyposażonych w urządzenie z dostępem do Internetu czy wyposażonych w centralne ogrzewanie i łazienkę. W połowie województwa dostęp do łazienki ma od 91,8 do 95,6% mieszkań. Średni odsetek mieszkań z urządzeniami z dostępem do Internetu i wyposażonych w centralne ogrzewanie są zbliżone i oscylują około 85,5%.

Najwyższe przeciętne miesięczne wydatki na 1 osobę w gospodarstwie domowym są w województwie dolnośląskim i wynoszą 1561 zł podczas, gdy najwyższy przeciętny dochód na 1 osobę w gospodarstwie domowym jest w województwie mazowieckim i osiąga 2450 zł.

Przeciętnie prawie $\frac{1}{3}$ gospodarstw domowych nie może sobie pozwolić na tygodniowy wypoczynek rodziny raz w roku a 15% gospodarstw z trudnością wiąże koniec z końcem.

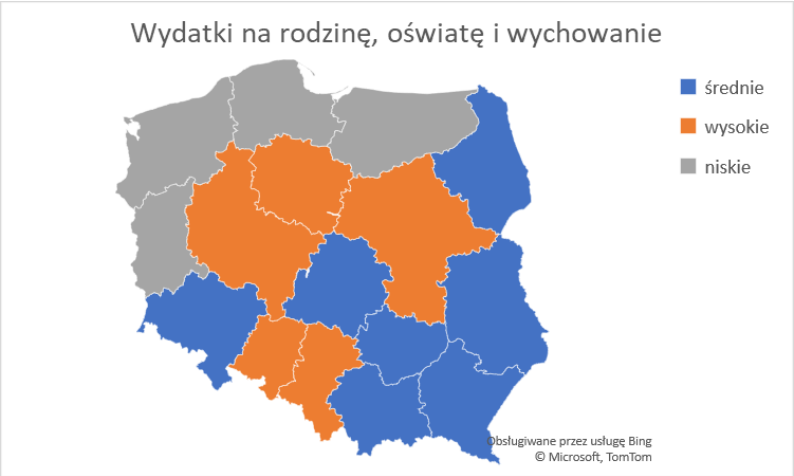
Średni wskaźnik zagrożenia ubóstwem po transferach społecznych wynosi 15,53% - oznacza to, że 15,53% społeczeństwa ma dochód ekwiwalentny poniżej progu zagrożenia ubóstwem. Najgorzej pod tym względem wypadło województwo lubelskie, gdzie odsetek ten wynosi prawie 23%. Przeciętne miesięczne wynagrodzenie brutto możemy zaobserwować w województwie mazowieckim. Wynosi ono 7913 zł i jest o ponad 25% wyższe niż średnie wynagrodzenie wszystkich województw.

Województwo o najlepszej dostępności komunikacji autobusowej charakteryzuje się posiadaniem 19,38 km linii autobusowych na każde 100 km^2, podczas gdy to o najgorszej dostępności posiada jedynie 4,14 km linii autobusowych na 100km^2. Przeciętnie w województwach zdarzają się 54,79 wypadki drogowe na 100 tys. ludności. Współczynniki zmienności oraz Giniego, które wynoszą mniej niż 10 zostały oznaczone w tabeli kolorem czerwonym i prawdopodobnie w późniejszym etapie zmienne odpowiadające tym współczynnikom zostaną usunięte ze zbioru, ponieważ przez swoją małą różnorodność słabo będą słabo różnicowały regiony. Na pomarańczowo zostały oznaczone współczynniki, które nie są większe od 10, ale drugi współczynnik odpowiadającej tej zmiennej jest powyżej tej wartości.

Województwa zostały podzielone na trzy kategorie w zależności od tego jaki procent wydatków w ciągu roku przeznaczają na oświatę i wychowanie oraz rodzinę. Średnio alokowały 5,27% wydatków na te sektory. Regiony, które przeznaczają powyżej 6% zostały określone jako województwa z wysokim poziomem wydatków, poniżej 4,5% z niskim a wszystkim województwom pomiędzy tymi wartościami została przypisana kategoria średnich wydatków (*Tabela 4*).

Tabela 4. Podział województw na kategorie w zależności od wydatków na rodzinę, oświatę i wychowanie.

Województwo	Procent wydatków	Kategoria
Dolnośląskie	5,09%	średnie
Kujawsko-Pomorskie	6,25%	wysokie
Lubelskie	5,34%	średnie
Lubuskie	3,43%	niskie
Łódzkie	4,77%	średnie
Małopolskie	4,99%	średnie
Mazowieckie	6,85%	wysokie
Opolskie	7,98%	wysokie
Podkarpackie	4,63%	średnie
Podlaskie	4,69%	średnie
Pomorskie	3,91%	niskie
Śląskie	6,68%	wysokie
Świętokrzyskie	4,95%	średnie
Warmińsko-Mazurskie	3,80%	niskie
Wielkopolskie	8,61%	wysokie
Zachodniopomorskie	2,30%	niskie
Średnia	**5,27%**	



Można zaobserwować (*wykres 6*), że północne województwa charakteryzują się niskim poziomem wydatków. Najmniejszy odsetek przeznaczany jest w województwie zachodniopomorskim. Natomiast województwa z wysokim poziomem wydatków skupiają się w centralnej części Polski oraz na Śląsku. Największy procent wydatków przeznaczany jest w województwie wielkopolskim i wynosi on 8,61%.

Wykres 6. Wydatki na rodzinę, oświatę i wychowanie.
Źródło: opracowanie własne w programie Excel

Co ciekawe w województwach ze średnim poziomem wydatków między innymi na oświatę jest średnio najwięcej placówek wychowania przedszkolnego (*x8*) oraz szkół podstawowych (*x9*), natomiast najmniej szkół średnich (*x10*). Jednak w przypadku ostatniej zmiennej nie ma znaczących różnic pomiędzy grupami.

Tabela 5. Porównanie ilości w szkół z zależności od poziomu wydatków.

X8			*X9*			*X10*		
Niskie	Średnie	Wysokie	Niskie	Średnie	Wysokie	Niskie	Średnie	Wysokie
56,206	63,248	55,399	35,079	41,309	36,474	18,703	17,856	18,086

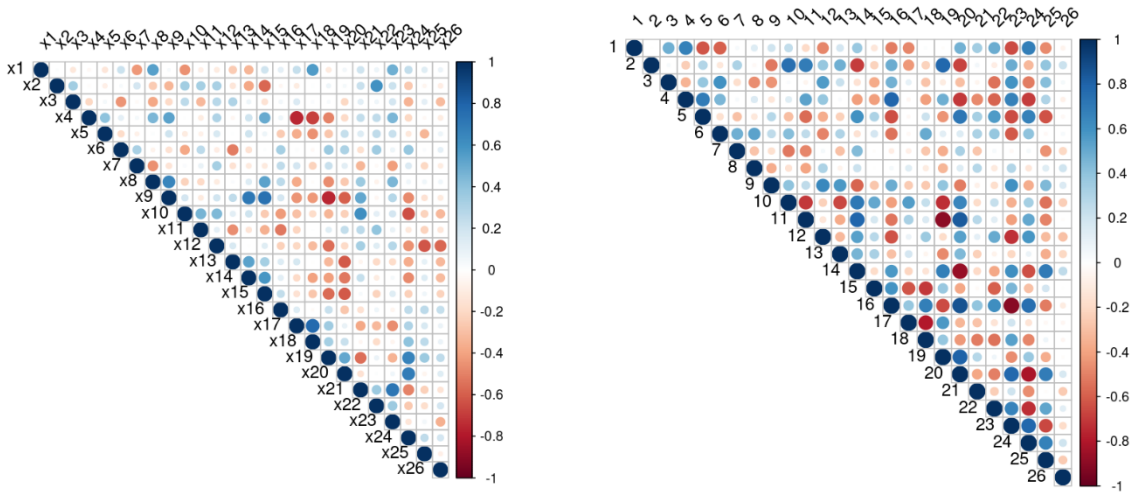
Korelacje zmiennych

Tabela 6. Korelacje zmiennych.

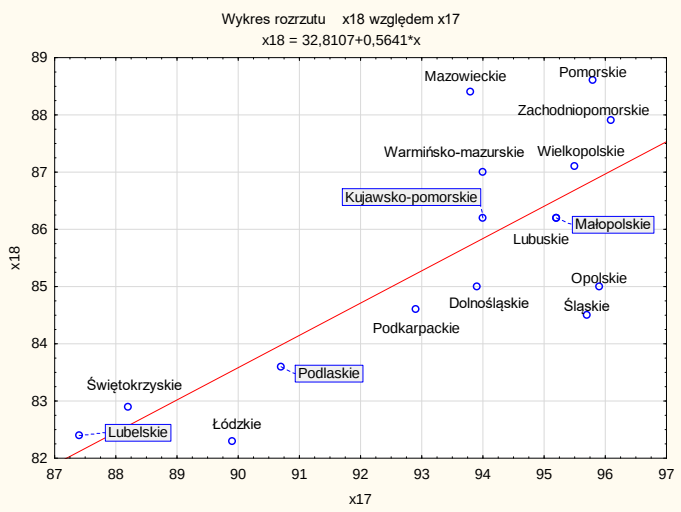
	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	x21	x22	x23	x24	x25	x26
x1	1,000	0,012	-0,121	-0,038	-0,134	0,219	-0,439	0,542	0,030	-0,455	-0,088	-0,108	-0,255	-0,333	0,206	0,118	0,238	0,567	-0,113	0,056	0,206	0,079	0,487	0,240	0,132	0,082
x2	0,012	1,000	0,343	-0,019	-0,089	0,032	0,187	-0,255	-0,335	0,345	0,333	0,328	-0,140	-0,388	-0,577	-0,043	-0,102	0,000	0,075	0,043	0,228	0,601	0,213	-0,270	-0,053	-0,100
x3	-0,121	0,343	1,000	-0,203	0,045	-0,444	-0,026	-0,389	-0,196	0,260	-0,329	0,279	0,304	0,070	-0,158	0,055	0,160	0,028	0,036	-0,206	0,121	0,062	0,185	-0,340	0,116	-0,318
x4	-0,038	-0,019	-0,203	1,000	0,404	0,115	0,046	0,455	0,530	0,003	-0,105	0,305	-0,047	0,203	0,509	0,076	-0,759	-0,688	-0,491	-0,196	0,244	0,259	0,452	-0,205	-0,035	-0,108
x5	-0,134	-0,089	0,045	0,404	1,000	-0,181	-0,071	0,019	0,203	-0,108	-0,076	0,169	0,006	0,204	0,074	-0,251	-0,367	-0,457	-0,271	0,233	0,147	0,259	0,223	-0,121	-0,333	0,075
x6	0,219	0,032	-0,444	0,115	-0,181	1,000	0,321	-0,065	-0,155	-0,387	0,265	-0,100	-0,513	-0,160	0,014	0,334	0,019	-0,038	0,361	0,315	-0,164	0,357	0,079	0,382	0,022	0,112
x7	-0,439	0,187	-0,026	0,046	-0,071	0,321	1,000	-0,452	-0,145	0,029	0,096	0,336	-0,159	-0,112	-0,116	0,253	0,189	-0,227	0,220	0,127	-0,333	-0,030	-0,415	0,160	-0,204	-0,101
x8	0,542	-0,255	-0,389	0,455	0,019	-0,065	-0,452	1,000	0,655	-0,229	-0,196	-0,109	0,021	0,146	0,538	0,269	-0,378	-0,016	-0,464	-0,261	0,364	-0,067	0,456	0,069	0,066	0,065
x9	0,030	-0,335	-0,196	0,530	0,203	-0,155	-0,145	0,655	1,000	0,181	-0,111	0,191	0,250	0,710	0,747	0,180	-0,441	-0,432	-0,755	-0,592	0,518	-0,047	0,289	-0,477	-0,151	-0,118
x10	-0,455	0,345	0,260	0,003	-0,108	-0,387	0,029	-0,229	0,181	1,000	0,460	0,441	0,128	0,200	-0,251	-0,427	-0,300	-0,187	-0,333	-0,234	0,616	0,119	0,184	-0,643	-0,221	-0,330
x11	-0,088	0,333	-0,329	-0,105	-0,076	0,265	0,096	-0,196	-0,111	0,460	1,000	0,115	-0,462	-0,141	-0,357	-0,523	-0,218	0,003	0,132	0,240	0,264	0,394	0,091	-0,130	-0,227	-0,139
x12	-0,108	0,328	0,279	0,305	0,169	-0,100	0,336	-0,109	0,191	0,441	0,115	1,000	0,166	-0,001	0,132	-0,232	-0,206	-0,313	-0,557	-0,197	0,338	0,133	0,289	-0,454	-0,625	-0,572
x13	-0,255	-0,140	0,304	-0,047	0,006	-0,513	-0,159	0,021	0,250	0,128	-0,462	0,166	1,000	0,536	0,339	0,142	0,173	-0,010	-0,330	-0,617	-0,054	-0,193	-0,183	-0,370	0,105	-0,325
x14	-0,333	-0,388	0,070	0,203	0,204	-0,160	-0,112	0,146	0,710	0,200	-0,141	-0,001	0,536	1,000	0,587	0,126	-0,193	-0,403	-0,419	-0,526	0,230	-0,102	0,042	-0,474	0,141	-0,320
x15	0,206	-0,577	-0,158	0,509	0,074	0,014	-0,116	0,538	0,747	-0,251	-0,357	0,132	0,339	0,587	1,000	0,240	-0,158	-0,273	-0,561	-0,611	0,039	-0,226	0,139	-0,233	0,061	-0,221
x16	0,118	-0,043	0,055	0,076	-0,251	0,334	0,253	0,269	0,180	-0,427	-0,523	-0,232	0,142	0,126	0,240	1,000	0,142	-0,024	0,224	-0,277	-0,159	0,069	-0,058	0,258	0,267	0,236
x17	0,238	-0,102	0,160	-0,759	-0,367	0,019	0,189	-0,378	-0,441	-0,300	-0,218	-0,206	0,173	-0,193	-0,158	0,142	1,000	0,779	0,357	0,092	-0,383	-0,327	-0,462	0,309	0,164	0,081
x18	0,567	0,000	0,028	-0,688	-0,457	-0,038	-0,227	-0,016	-0,432	-0,187	0,003	-0,313	-0,010	-0,403	-0,273	-0,024	0,779	1,000	0,333	0,225	-0,042	-0,186	-0,052	0,405	0,260	0,084
x19	-0,113	0,075	0,036	-0,491	-0,271	0,361	0,220	-0,464	-0,755	-0,333	0,132	-0,557	-0,330	-0,419	-0,561	0,224	0,357	0,333	1,000	0,518	-0,557	0,080	-0,348	0,662	0,369	0,257
x20	0,056	0,043	-0,206	-0,196	0,233	0,315	0,127	-0,261	-0,592	-0,234	0,240	-0,197	-0,617	-0,526	-0,611	-0,277	0,092	0,225	0,518	1,000	-0,157	0,013	0,062	0,698	-0,060	0,132
x21	0,206	0,228	0,121	0,244	0,147	-0,164	-0,333	0,364	0,518	0,616	0,264	0,338	-0,054	0,230	0,039	-0,159	-0,383	-0,042	-0,557	-0,157	1,000	0,363	0,729	-0,496	-0,240	-0,143
x22	0,079	0,601	0,062	0,259	0,259	0,357	-0,030	-0,067	-0,047	0,119	0,394	0,133	-0,193	-0,102	-0,226	0,069	-0,327	-0,186	0,080	0,013	0,363	1,000	0,362	-0,264	-0,180	0,186
x23	0,487	0,213	0,185	0,452	0,223	0,079	-0,415	0,456	0,289	0,184	0,091	0,289	-0,183	0,042	0,139	-0,058	-0,462	-0,052	-0,348	0,062	0,729	0,362	1,000	-0,137	0,059	-0,353
x24	0,240	-0,270	-0,340	-0,205	-0,121	0,382	0,160	0,069	-0,477	-0,643	-0,130	-0,454	-0,370	-0,474	-0,233	0,258	0,309	0,405	0,662	0,698	-0,496	-0,264	-0,137	1,000	0,250	0,184
x25	0,132	-0,053	0,116	-0,035	-0,333	0,022	-0,204	0,066	-0,151	-0,221	-0,227	-0,625	0,105	0,141	0,061	0,267	0,164	0,260	0,369	-0,060	-0,240	-0,180	0,059	0,250	1,000	-0,081
x26	0,082	-0,100	-0,318	-0,108	0,075	0,112	-0,101	0,065	-0,118	-0,330	-0,139	-0,572	-0,325	-0,320	-0,221	0,236	0,081	0,084	0,257	0,132	-0,143	0,186	-0,353	0,184	-0,081	1,000

Powyżej w tabeli przedstawione zostały korelacje zmiennych wejściowych. Kolorem czerwonym oznaczono korelacje powyżej 0,5 natomiast kolorem niebieskim wartości powyżej 0,45. Najsłabsze korelacje wykazuje zmienna *x3* (umieralność okołoporodowa na 10 000 urodzeń żywych i martwych). Jej najwyższa korelacja wynosi -0,444 ze zmienna *x6* (liczba dzieci na jakie przypada 1 łóżko w szpitalu na oddziale pediatrycznym (w setkach)). Zmienne związane z transportem (x25 i x26) również są słabo skorelowane z pozostałymi. Jedynie wyższą korelacje wykazują ze zmienna *x12* (Wskaźnik wykrywalności sprawców przestępstw stwierdzonych przez policję w 2022). Najwięcej umiarkowanych korelacji (powyżej 0,45) mają zmienne *x9*, *x15* i *x19*.

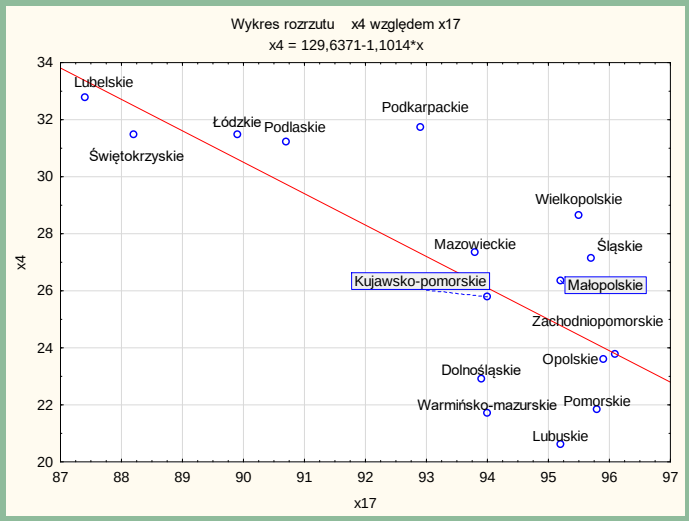
Macierz korelacji cząstkowej różni się kolorami (znakami) od macierzy po lewej, czyli występujące korelacje są pozorne. Będą wchodzić w związki z innymi cechami, więc jest potencjał do dobrego PCA.



Wykres 7. Korelacje całkowite oraz cząstkowe.
Źródło: opracowanie własne w programie RStudio.



Wykres 8. Wykres rozrzutu x18 względem x17.



Wykres 9. Wykres rozrzutu x4 względem x17.

Najwyższą dodatnią korelacje wykazują zmienne *x17* (Procent mieszkań wyposażonych w łazienkę) i *x18* (Procent mieszkań wyposażonych w ogrzewanie centralne). Oznacza to, że procent mieszkań wyposażonych w łazienkę rośnie wraz ze wzrostem procentu mieszkań wyposażonych w ogrzewanie centralne i jest z nim współzależny. Obydwie te zmienne związane są z komfortem i standardem życia. Korelacja ta wynosi 0,779, natomiast korelacja cząstkowa wynosi -0,792. Oznacza to, że współzależność badanych cech zmienia się przy świadomej eliminacji wpływu pozostałych cech. W związku z tym jest to relacja pozorna, ponieważ zależy od innych zmiennych. Na podstawie wykresu rozrzutu (Wykres 8. Wykres rozrzutu x18 względem x17.*Wykres 8*) można zaobserwować, że województwo mazowieckie oraz śląskie leżą najdalej od linii regresji. Przy czym dla województwa mazowieckiego wartość zmiennej *x18* jest znacząca wyższa niż szacowana przez regresję a dla województwa śląskiego niższa.

Najwyższą korelacje ujemną również ma zmienna *x17*. Jednak w tym przypadku względem zmiennej *x4* (Liczba pracujących położnych przypadających na 10 000 kobiet w wieku produkcyjnym). Osiąga ona wartość -0,759. Oznacza to, że wraz ze wzrostem procentu mieszkań wyposażonych w łazienkę maleje liczba pracujących położnych. Ponownie mamy do czynienia z korelacją pozorną, ponieważ korelacja cząstkowa wynosi -0.041. Od linii regresji znacząco odstaje województwo podkarpackie.

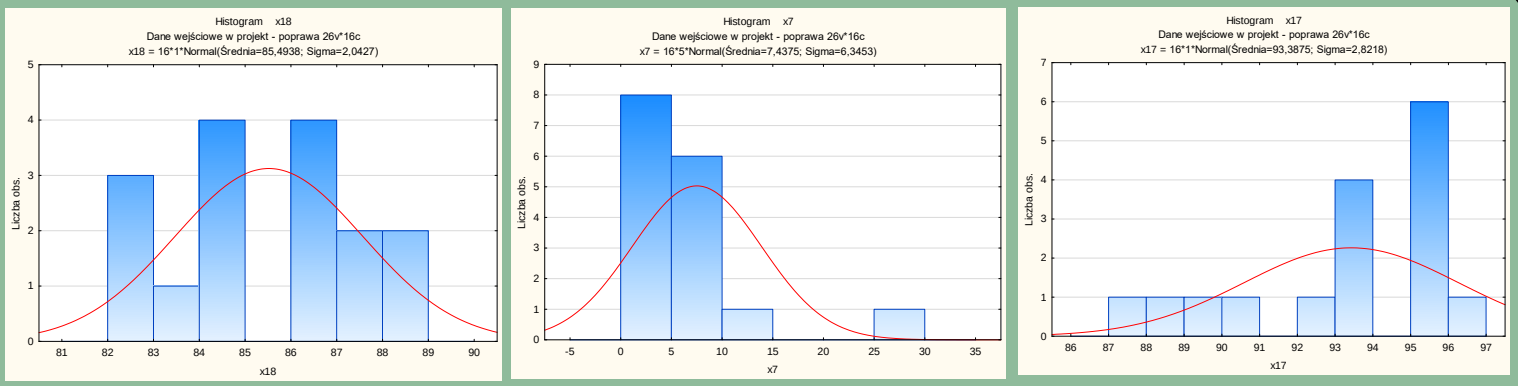
Korelacje powyżej 0,7 wykazują również zmienne:

- *x9* (Liczba szkół podstawowych na 100 tys. ludności) z *x14* (Liczba bibliotek publicznych na 100 tys. ludności),
- *x9* z *x15* (Przeciętna powierzchnia użytkowa mieszkania w m^2),
- *x9* z *x19* (Przeciętne miesięczne wydatki na 1 osobę w gospodarstwach domowych w setkach),
- *x21* (% gospodarstw domowych deklarujących brak możliwości realizacji potrzeby tygodniowego wypoczynku rodziny raz w roku *) z *x23* (Wskaźnik zagrożenia ubóstwem po transferach społecznych w % *)

Jednak w większości są to również korelacje pozorne.

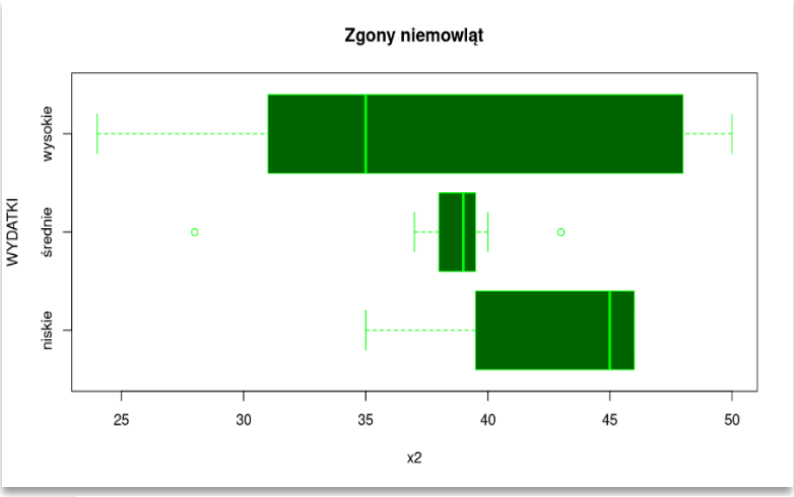
Wykresy zmiennych

Dane wejściowe zawierają zmienne zarówno z rozkładem bliskim symetrycznemu jak w przypadku zmiennej *x18* jak i zmienne z rozkładem asymetrycznym, czego przykładem są zmienne *x17* oraz *x7*. Na podstawie wykresu zmiennej *x7* można zaobserwować iż poziom zanieczyszczeń w większości województw oscyluje w przedziale 0-10. Natomiast w ponad połowie województw średnio w łazienkę wyposażonych jest co najmniej 93% gospodarstw domowych.

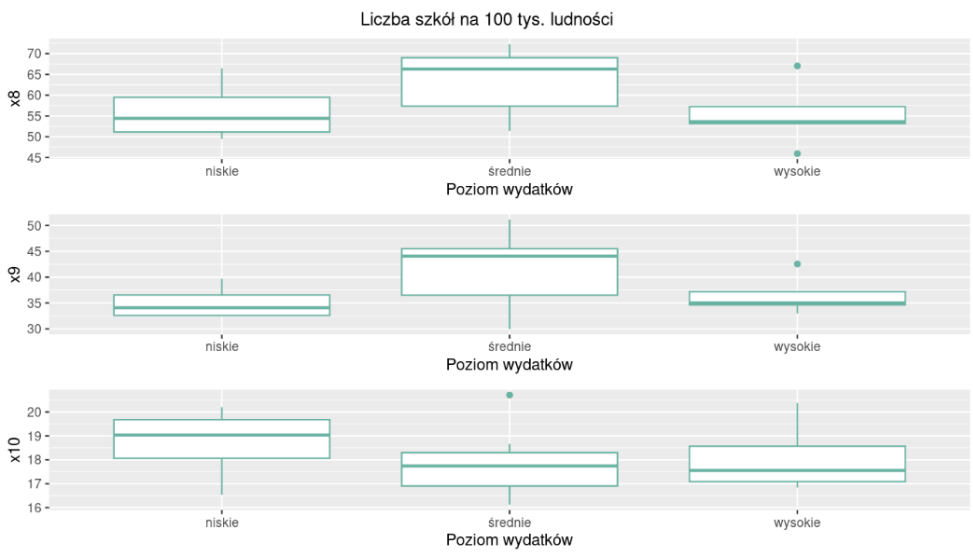


Wykres 10. Histogramy zmiennych.
Źródło: opracowanie własne w programie Statistica.

Stosując podział zmiennych na grupy w zależności od wydatków można zaobserwować, że średnia zgonów niemowląt jest najniższa dla województw o wysokich wydatkach natomiast najwyższa dla województw z niskim poziomem wydatków. Możliwe, że w województwach z wyższym poziomem wydatków między innymi na rodzinę i edukację jest lepszy dostęp do opieki zdrowotnej bądź większa dostępność edukacji zdrowotnej. Wartości dla województw z wysokim poziomem są najbardziej zróżnicowane a z średnim najmniej.



Wykres 11. Liczba zgonów niemowląt dla grup województw.
Źródło: opracowanie własne w programie RStudio.



Dodatkowo na *Wykres 12* można zaobserwować zależność ilości szkół od poziomu wydatków przedstawioną w *Tabela 5*.

Wykres 12. Porównanie ilości w szkół z zależności od poziomu wydatków.
Źródło: opracowanie własne w programie RStudio.

3. Analiza głównych składowych

3.1. Wybór zmiennych do analizy

Jako podstawę analizy głównych składowych wykorzystano macierz z danymi wejściowymi z pominięciem zmiennych, których współczynnik zmienności oraz Giniego przyjmują niskie wartości (obydwie mniejsze od 10). Wyjątek stanowił zmienna *x20* (przeciętny miesięczny dochód na 1 osobę w gospodarstwie w setkach), której współczynnik zmienności był bliski 10 (9,72) a korelacje z pozostałymi obserwacjami umiarkowane. Macierz danych zawiera obecnie 18 zmiennych (*Tabela 7).*

Tabela 7. Zmienne wejściowe do analizy głównych składowych.

Województwo	x2	x3	x4	x5	x6	x7	x8	x9	x11	x13	x14	x19	x20	x21	x22	x23	x25	x26
Dolnośląskie	39	62	22,91	54,60	11,11	7	51,38	29,95	10,10	10	20	15,61	22,13	20,0	13,8	9,9	14,93	64,6
Kujawsko-pomorskie	50	69	25,79	45,84	11,28	10	53,62	35,03	11,60	11	19	12,52	19,32	37,0	14,1	20,4	16,10	38,3
Lubelskie	43	55	32,77	61,39	13,57	4	66,28	44,90	17,70	12	28	12,46	19,07	37,9	24,8	22,9	10,45	38,7
Lubuskie	35	52	20,61	58,57	11,56	4	49,49	35,51	19,30	8	24	13,00	21,27	28,4	12,2	12,8	12,05	50,8
Łódzkie	40	56	31,47	68,49	11,05	8	55,88	36,91	10,80	8	19	13,54	20,86	29,8	19,6	14,5	12,13	92,6
Małopolskie	28	49	26,36	59,32	11,75	7	70,98	44,07	7,60	13	21	10,85	20,39	31,0	10,5	14,9	4,14	64,9
Mazowieckie	31	52	27,35	58,41	11,68	5	67,05	34,64	13,90	5	17	15,08	24,50	31,4	12,0	20,7	19,38	52,5
Opolskie	24	66	23,61	55,18	8,90	11	53,16	42,55	8,70	20	32	13,67	16,91	25,6	8,9	10,1	15,37	46,3
Podkarpackie	39	59	31,74	61,85	9,27	5	72,29	51,13	8,80	16	32	10,05	17,04	34,2	14,3	17,6	19,10	51,0
Podlaskie	39	69	31,21	68,92	7,42	2	58,86	36,12	9,10	16	20	11,00	20,05	28,0	12,1	19,6	10,67	29,1
Pomorskie	46	58	21,83	48,55	10,14	3	66,49	32,65	13,00	12	13	14,69	18,36	24,4	15,8	13,1	15,83	67,5
Śląskie	48	59	27,15	61,06	12,81	29	45,92	32,97	15,60	8	17	14,21	21,80	23,3	17,3	11,2	7,89	45,5
Świętokrzyskie	37	49	31,48	53,64	8,73	10	67,05	46,09	17,40	9	21	10,77	18,65	34,3	11,7	13,4	7,47	54,9
Warmińsko-mazurskie	46	71	21,70	64,77	7,73	3	57,16	39,67	15,10	10	21	12,21	20,69	43,6	19,4	19,4	5,53	60,6
Wielkopolskie	35	59	28,64	54,99	12,19	5	57,25	37,18	12,60	9	18	12,48	18,81	30,2	18,6	16,2	11,77	65,8
Zachodniopomorskie	44	55,83	23,79	57,36	9,00	6	51,69	32,49	14,00	20	21	12,89	20,29	30,6	17,1	12,0	12,75	53,6

Na tym etapie zostały usunięte zmienne:

- *x1* – Procent ludności jaki stanowią dzieci i młodzież
- *x10* – Liczba szkół średnich na 100 tys. ludności
- *x12* - Wskaźnik wykrywalności sprawców przestępstw stwierdzonych przez policję w 2022
- *x15* – Przeciętna powierzchnia użytkowa mieszkania w m^2
- *x16* – Gospodarstwa wyposażone w urządzenie z dostępem do Internetu w % ogółu gospodarstw
- *x17* – Procent mieszkań wyposażonych w łazienkę
- *x18* – Procent mieszkań wyposażonych w centralne ogrzewanie
- *x24* - Przeciętne miesięczne wynagrodzenie brutto w setkach

Punktem wyjścia analizy jest macierz korelacji:

Tabela 8. Korelacje zmiennych.

Zmienna	x2	x3	x4	x5	x6	x7	x8	x9	x11	x13	x14	x19	x20	x21	x22	x23	x25	x26
x2	1,000	0,343	-0,019	-0,089	0,032	0,187	-0,255	-0,335	0,333	-0,140	-0,388	0,075	0,043	0,228	0,601	0,213	-0,053	-0,100
x3	0,343	1,000	-0,203	0,045	-0,444	-0,026	-0,389	-0,196	-0,329	0,304	0,070	0,036	-0,206	0,121	0,062	0,185	0,116	-0,318
x4	-0,019	-0,203	1,000	0,404	0,115	0,046	0,455	0,530	-0,105	-0,047	0,203	-0,491	-0,196	0,244	0,259	0,452	-0,035	-0,108
x5	-0,089	0,045	0,404	1,000	-0,181	-0,071	0,019	0,203	-0,076	0,006	0,204	-0,271	0,233	0,147	0,259	0,223	-0,333	0,075
x6	0,032	-0,444	0,115	-0,181	1,000	0,321	-0,065	-0,155	0,265	-0,513	-0,160	0,361	0,315	-0,164	0,357	0,079	0,022	0,112
x7	0,187	-0,026	0,046	-0,071	0,321	1,000	-0,452	-0,145	0,096	-0,159	-0,112	0,220	0,127	-0,333	-0,030	-0,415	-0,204	-0,101
x8	-0,255	-0,389	0,455	0,019	-0,065	-0,452	1,000	0,655	-0,196	0,021	0,146	-0,464	-0,261	0,364	-0,067	0,456	0,066	0,065
x9	-0,335	-0,196	0,530	0,203	-0,155	-0,145	0,655	1,000	-0,111	0,250	0,710	-0,755	-0,592	0,518	-0,047	0,289	-0,151	-0,118
x11	0,333	-0,329	-0,105	-0,076	0,265	0,096	-0,196	-0,111	1,000	-0,462	-0,141	0,132	0,240	0,264	0,394	0,091	-0,227	-0,139
x13	-0,140	0,304	-0,047	0,006	-0,513	-0,159	0,021	0,250	-0,462	1,000	0,536	-0,330	-0,617	-0,054	-0,193	-0,183	0,105	-0,325
x14	-0,388	0,070	0,203	0,204	-0,160	-0,112	0,146	0,710	-0,141	0,536	1,000	-0,419	-0,526	0,230	-0,102	0,042	0,141	-0,320
x19	0,075	0,036	-0,491	-0,271	0,361	0,220	-0,464	-0,755	0,132	-0,330	-0,419	1,000	0,518	-0,557	0,080	-0,348	0,369	0,257
x20	0,043	-0,206	-0,196	0,233	0,315	0,127	-0,261	-0,592	0,240	-0,617	-0,526	0,518	1,000	-0,157	0,013	0,062	-0,060	0,132
x21	0,228	0,121	0,244	0,147	-0,164	-0,333	0,364	0,518	0,264	-0,054	0,230	-0,557	-0,157	1,000	0,363	0,729	-0,240	-0,143
x22	0,601	0,062	0,259	0,259	0,357	-0,030	-0,067	-0,047	0,394	-0,193	-0,102	0,080	0,013	0,363	1,000	0,362	-0,180	0,186
x23	0,213	0,185	0,452	0,223	0,079	-0,415	0,456	0,289	0,091	-0,183	0,042	-0,348	0,062	0,729	0,362	1,000	0,059	-0,353
x25	-0,053	0,116	-0,035	-0,333	0,022	-0,204	0,066	-0,151	-0,227	0,105	0,141	0,369	-0,060	-0,240	-0,180	0,059	1,000	-0,081
x26	-0,100	-0,318	-0,108	0,075	0,112	-0,101	0,065	-0,118	-0,139	-0,325	-0,320	0,257	0,132	-0,143	0,186	-0,353	-0,081	1,000

Ponownie kolorem czerwonym oznaczono korelacje powyżej 0,5 natomiast kolorem niebieskim powyżej 0,45. Można zauważyć, że zmienne *x9*, *x19* wykazują najwięcej wysokich korelacji z innymi. Natomiast zmienne *x3*, *x5*, *x25*, *x26* nie są silnie związane liniowo z pozostałymi.

Wyniki analizy głównych składowych dla 18 zmiennych wejściowych zostały przedstawione w poniższej tabeli (Tabela 9). Kolorem czerwonym zaznaczono wartości własne większe od 1. Zgodnie z kryterium

Kaisera, wykonując analizę w tym momencie otrzymano by 7 głównych składowych, które wyjaśniałyby 84,228% wariancji. Można zauważyć, że zaczynając od dziewiątej składowej składowe opisują co najwyżej 3% wariancji.

Tabela 9. Wartości własne i skumulowany % wariancji przy 18 zmiennych.

Nr składowej	Wartość własna	Skumulowany % wariancji
1	4,458	24,767
2	3,054	41,733
3	2,214	54,035
4	1,626	63,070
5	1,442	71,082
6	1,219	77,851
7	1,148	84,228
8	0,929	89,388
9	0,591	92,671
10	0,497	95,430
11	0,302	97,109
12	0,237	98,428
13	0,182	99,436
14	0,079	99,872
15	0,023	100,000

Tabela 10. Korelacje zmiennych z 7 składowymi.

Zmienna	1 składowa	2 składowa	3 składowa	4 składowa	5 składowa	6 składowa	7 składowa
x2	0,237	0,448	0,676	0,034	0,119	0,090	0,294
x3	-0,056	-0,256	0,842	-0,088	-0,172	-0,203	0,034
x4	-0,540	0,400	-0,221	0,151	0,056	-0,497	0,183
x5	-0,279	0,279	-0,011	0,333	-0,646	-0,425	-0,076
x6	0,393	0,445	-0,378	0,064	0,473	-0,259	0,122
x7	0,397	-0,029	-0,005	0,691	0,297	-0,235	0,002
x8	-0,637	0,197	-0,457	-0,404	0,020	0,084	0,084
x9	-0,889	0,066	-0,280	0,182	0,158	0,069	0,040
x11	0,254	0,584	0,074	0,181	0,330	0,395	-0,281
x13	-0,476	-0,644	0,308	0,124	0,024	0,002	0,161
x14	-0,669	-0,298	-0,017	0,230	0,269	-0,153	-0,050
x19	0,853	-0,100	-0,012	-0,233	0,087	-0,175	0,090
x20	0,628	0,359	-0,159	-0,126	-0,322	-0,239	-0,444
x21	-0,590	0,561	0,278	-0,129	-0,013	0,277	-0,124
x22	0,037	0,710	0,312	0,099	0,032	-0,059	0,509
x23	-0,473	0,623	0,249	-0,433	0,030	-0,244	-0,200
x25	0,079	-0,329	0,020	-0,643	0,377	-0,379	0,139
x26	0,285	0,086	-0,454	-0,081	-0,451	0,239	0,593

Na podstawie powyższej tabeli *(Tabela 10.* Korelacje zmiennych z 7 składowymi.) kolorem czerwonym zostały oznaczone najwyższe korelacje dla każdej zmiennej. Można zauważyć, że zmienne *x9* oraz *x19* mają największy wkład w budowanie 1 składowej. Zmienna *x26* w największym stopniu buduje 7 składową główną. Natomiast zmienne *x5* i *x6* najbardziej piątą. Zważając na to oraz niskie korelacje postanowiono usunąć *x5* i *x26* ze zbioru.

W wyniku tej czynności udało się zredukować liczbę składowych do 6 (*Tabela 11.*). Pierwsza składowa wyjaśnia obecnie ponad 27% wariancji, a więc więcej niż w poprzedniej analizie. Natomiast ostatnia składowa wyjaśnia jedynie 6% zmienności oraz jej wartość własna jest bliska 1. Procent wyjaśnianej wariancji przez sześć składowych jest bliski otrzymanemu za pomocą siedmiu składowych.

Tabela 11. Wartości własne i skumulowany % wariancji przy 16 zmiennych.

Nr składowej	Wartość własna	Skumulowany % wariancji
1	4,332	27,077
2	2,999	45,818
3	2,097	58,925
4	1,585	68,828
5	1,246	76,613
6	1,013	82,942
7	0,763	87,710
8	0,584	91,358
9	0,438	94,096
10	0,316	96,070
11	0,274	97,784
12	0,181	98,914
13	0,136	99,762
14	0,033	99,969
15	0,005	100,000

Korelacje zmiennych wejściowych z składowymi zaprezentowano w tabeli poniżej (*Tabela 12*). Kolorem niebieskim oznaczono drugą najwyższą korelacje, jeśli jest większa od 0,5 (w dalszej części będą wykorzystywane te same oznaczenia kolorystyczne w przypadku korelacji zmiennych ze składowymi).

Tabela 12. Korelacje zmiennych z 6 składowymi.

Zmienna	1 składowa	2 składowa	3 składowa	4 składowa	5 składowa	6 składowa
x2	0,265	0,468	0,667	-0,149	-0,151	-0,114
x3	-0,024	-0,249	0,851	0,095	-0,104	-0,176
x4	-0,517	0,393	-0,238	-0,103	-0,392	-0,437
x6	0,389	0,451	-0,461	-0,108	-0,459	0,094
x7	0,417	-0,045	-0,117	-0,688	-0,242	-0,307
x8	-0,669	0,230	-0,394	0,368	0,030	-0,128
x9	-0,899	0,092	-0,257	-0,226	-0,043	0,075
x11	0,279	0,601	-0,019	-0,263	0,207	0,546
x13	-0,477	-0,626	0,327	-0,149	-0,116	0,094
x14	-0,660	-0,283	-0,049	-0,238	-0,274	0,435
x19	0,842	-0,120	-0,033	0,246	-0,252	0,193
x20	0,661	0,304	-0,212	0,296	0,240	-0,128
x21	-0,574	0,597	0,309	0,052	0,229	0,157
x22	0,062	0,699	0,354	-0,170	-0,338	0,144
x23	-0,435	0,653	0,218	0,438	-0,092	-0,102
x25	0,055	-0,291	0,005	0,593	-0,599	0,220

Skład zmiennych, które w największym stopniu budują trzy pierwsze składowe nie zmienił się. Największy wpływ na budowanie ostatniej składowej ma teraz zmienna *x11*, jednak ma ona wyższą korelacje z 2 składową. W związku z tym postanowiłono jako następną wyeliminować *x25*, która buduje 5 składową oraz posiada niskie korelacje z innymi zmiennymi. Wskutek tego działania udało mi się zmniejszyć liczbę składowych do 5.

Tabela 13. Wartości własne i skumulowany % wariancji przy 15 zmiennych.

Nr składowej	Wartość własna	Skumulowany % wariancji
1	4,330	28,868
2	2,944	48,494
3	2,097	62,475
4	1,462	72,223
5	1,051	79,232

6	0,821	84,702
7	0,763	89,787
8	0,440	92,718
9	0,326	94,889
10	0,278	96,743
11	0,246	98,381
12	0,162	99,459
13	0,073	99,947
14	0,007	99,994
15	0,001	100,000

Procent zmienności wyjaśnianej przez 1. składową jest bliski 29. Ostatnia (piąta) składowa wyjaśnia natomiast 7% wariancji. Najwyższe korelacje ma ona ze zmiennymi *x4* oraz *x11*, jednak podobnie jak wcześniej zmienne te mają wysokie korelacje również z 2 pierwszymi składowymi. Zmienna *x7* natomiast korelacje powyżej 0,5 wykazuje jedynie z przedostatnią składową.

Tabela 15. Korelacje zmiennych z 5 składowymi.

Zmienna	1 składowa	2 składowa	3 składowa	4 składowa	5 składowa
x2	0,268	-0,467	-0,667	-0,225	-0,105
x3	-0,027	0,238	-0,851	0,062	-0,232
x4	-0,515	-0,404	0,238	-0,266	-0,521
x6	0,391	-0,461	0,462	-0,322	-0,138
x7	0,420	0,087	0,116	-0,701	-0,233
x8	-0,670	-0,259	0,395	0,331	-0,148
x9	-0,897	-0,085	0,257	-0,234	0,094
x11	0,285	-0,579	0,019	-0,212	0,655
x13	-0,481	0,625	-0,328	-0,162	0,020
x14	-0,664	0,270	0,048	-0,366	0,274
x19	0,837	0,088	0,033	0,087	-0,012
x20	0,664	-0,301	0,213	0,379	-0,047
x21	-0,568	-0,593	-0,308	0,116	0,256
x22	0,068	-0,697	-0,354	-0,324	-0,028
x23	-0,433	-0,691	-0,216	0,321	-0,194

Tabela 14. Korelacje 14 zmiennych.

Zmienna	x2	x3	x4	x6	x7	x8	x9	x11	x13	x14	x19	x20	x21	x22	x23
x2	1,000	0,343	-0,019	0,032	0,187	-0,255	-0,335	0,333	-0,140	-0,388	0,075	0,043	0,228	0,601	0,213
x3	0,343	1,000	-0,203	-0,444	-0,026	-0,389	-0,196	-0,329	0,304	0,070	0,036	-0,206	0,121	0,062	0,185
x4	-0,019	-0,203	1,000	0,115	0,046	0,455	0,530	-0,105	-0,047	0,203	-0,491	-0,196	0,244	0,259	0,452
x6	0,032	-0,444	0,115	1,000	0,321	-0,065	-0,155	0,265	-0,513	-0,160	0,361	0,315	-0,164	0,357	0,079
x7	0,187	-0,026	0,046	0,321	1,000	-0,452	-0,145	0,096	-0,159	-0,112	0,220	0,127	-0,333	-0,030	-0,415
x8	-0,255	-0,389	0,455	-0,065	-0,452	1,000	0,655	-0,196	0,021	0,146	-0,464	-0,261	0,364	-0,067	0,456
x9	-0,335	-0,196	0,530	-0,155	-0,145	0,655	1,000	-0,111	0,250	0,710	-0,755	-0,592	0,518	-0,047	0,289
x11	0,333	-0,329	-0,105	0,265	0,096	-0,196	-0,111	1,000	-0,462	-0,141	0,132	0,240	0,264	0,394	0,091
x13	-0,140	0,304	-0,047	-0,513	-0,159	0,021	0,250	-0,462	1,000	0,536	-0,330	-0,617	-0,054	-0,193	-0,183
x14	-0,388	0,070	0,203	-0,160	-0,112	0,146	0,710	-0,141	0,536	1,000	-0,419	-0,526	0,230	-0,102	0,042
x19	0,075	0,036	-0,491	0,361	0,220	-0,464	-0,755	0,132	-0,330	-0,419	1,000	0,518	-0,557	0,080	-0,348
x20	0,043	-0,206	-0,196	0,315	0,127	-0,261	-0,592	0,240	-0,617	-0,526	0,518	1,000	-0,157	0,013	0,062
x21	0,228	0,121	0,244	-0,164	-0,333	0,364	0,518	0,264	-0,054	0,230	-0,557	-0,157	1,000	0,363	0,729
x22	0,601	0,062	0,259	0,357	-0,030	-0,067	-0,047	0,394	-0,193	-0,102	0,080	0,013	0,363	1,000	0,362
x23	0,213	0,185	0,452	0,079	-0,415	0,456	0,289	0,091	-0,183	0,042	-0,348	0,062	0,729	0,362	1,000

Najniższe korelacje z innymi zmiennymi wykazuje zmienna *x3*. Jej najwyższa korelacja wynosi -0,444, pozostałe nie przekraczają 0,39. Zważając na to zdecydowano w następnym kroku ją wyeliminować. Pozwoliło to ograniczyć liczbę głównych składowych do czterech, ponieważ wartość własna piątej spadła do poziomu poniżej jeden (*Tabela 16*).

Tabela 16. Wartości własne i skumulowany % wariancji przy 14 zmiennych.

Nr składowej	Wartość własna	Skumulowany % wariancji
1	4,330	30,926
2	2,915	51,750
3	1,540	62,753
4	1,443	73,059
5	0,989	80,125
6	0,818	85,970
7	0,590	90,186
8	0,416	93,160
9	0,280	95,160
10	0,249	96,937
11	0,218	98,491
12	0,162	99,646
13	0,049	99,994
14	0,001	100,000

Procent zmienności wyjaśnianej przez 1 składową wzrósł do ponad 30. Natomiast łącznie 4 składowe wyjaśniają 73,059% wariancji.

Tabela 17. Korelacje zmiennych z 4 składowymi.

Zmienna	1 składowa	2 składowa	3 składowa	4 składowa
x2	0,271	-0,540	-0,652	-0,048
x4	-0,518	-0,376	0,237	0,423
x6	0,386	-0,404	0,364	0,529
x7	0,420	0,097	-0,116	0,732
x8	-0,673	-0,206	0,476	-0,150
x9	-0,899	-0,051	0,122	0,309
x11	0,282	-0,562	-0,197	0,110
x13	-0,478	0,588	-0,472	-0,045
x14	-0,663	0,275	-0,131	0,339
x19	0,837	0,086	0,100	-0,045
x20	0,662	-0,279	0,417	-0,218
x21	-0,568	-0,624	-0,186	-0,216
x22	0,068	-0,731	-0,396	0,177
x23	-0,433	-0,720	0,133	-0,269

Z powyższej tabeli (*Tabela 17*) można zauważyć, że ostatnią składową w największym stopniu budują zmienne *x6* oraz *x7*. Zważając na to, iż zmienna *x7* ma nieznacznie niższe korelacje z innymi, postanowiono w pierwszej kolejności ją wyeliminować. Ta czynność nie pozwoliła jednak zredukować liczby składowych (*Tabela 18*). Poprawił się jednak procent wariancji wyjaśnianej przez kolejne zmienne.

Tabela 18. Wartości własne i skumulowany % wariancji przy 13 zmiennych.

Nr składowej	Wartość własna	Skumulowany % wariancji
1	4,192	32,248
2	2,908	54,621
3	1,538	66,455
4	1,122	75,086
5	0,973	82,573
6	0,635	87,457
7	0,527	91,512
8	0,341	94,131
9	0,252	96,069
10	0,220	97,763
11	0,201	99,305

12	0,077	99,898
13	0,013	100,000

Tabela 19. Korelacje zmiennych z 4 składowymi (przy 13 zmiennych).

Zmienna	1 składowa	2 składowa	3 składowa	4 składowa
x2	0,269	-0,539	-0,651	-0,131
x4	-0,542	-0,401	0,292	0,141
x6	0,368	-0,408	0,425	0,585
x8	-0,646	-0,212	0,455	-0,227
x9	-0,921	-0,080	0,163	0,208
x11	0,289	-0,558	-0,183	0,268
x13	-0,489	0,577	-0,474	0,044
x14	-0,685	0,252	-0,088	0,486
x19	0,846	0,110	0,088	0,153
x20	0,682	-0,257	0,386	-0,257
x21	-0,543	-0,632	-0,208	-0,223
x22	0,080	-0,732	-0,373	0,332
x23	-0,392	-0,720	0,100	-0,297

Na podstawie korelacji zmiennych ze składowymi (*Tabela 199*) można zauważyć, iż ostatnią składową buduje głównie zmienna *x6*. Posiada ona również niskie korelacje z innymi (*Tabela 14*). Zważając na ten fakt została ona wykluczona z dalszej analizy.

Przez to działanie udało się zredukować liczbę głównych składowych do pożądanej – trzech.

Tabela 20. Wartości własne i skumulowany procent wariancji przy 12 zmiennych.

Nr składowej	Wartość własna	Skumulowany % wariancji
1	4,094	34,115
2	2,784	57,312
3	1,451	69,405
4	0,987	77,629
5	0,770	84,049
6	0,616	89,184
7	0,496	93,315
8	0,270	95,564
9	0,226	97,444
10	0,201	99,121
11	0,092	99,887
12	0,014	100,000

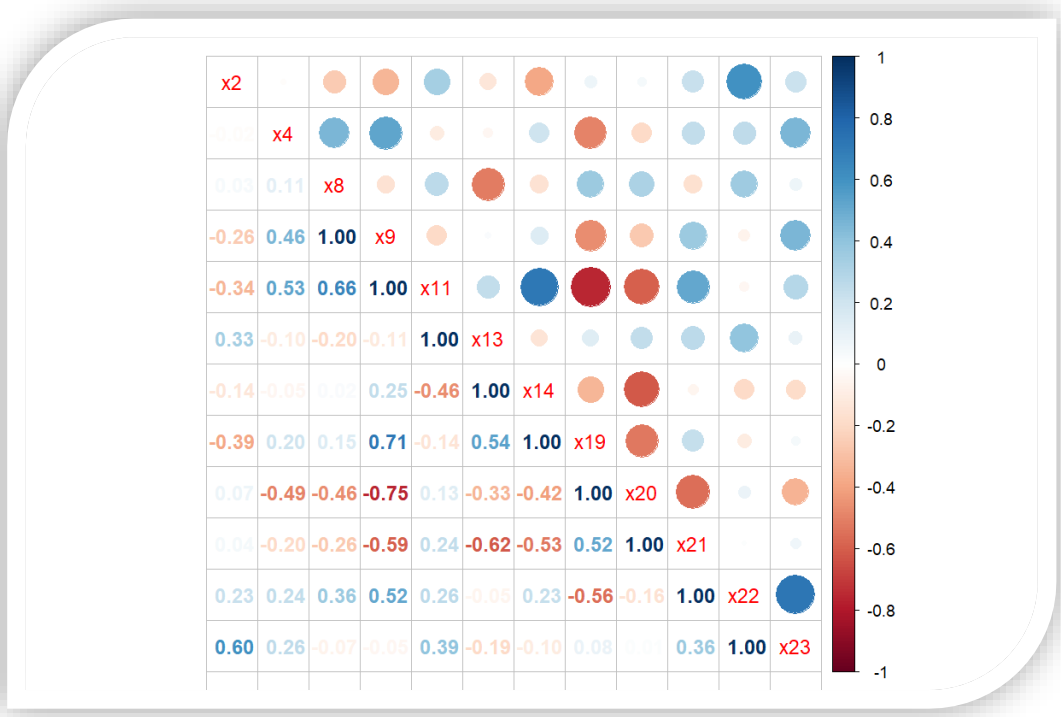
Z racji, że wyjaśniają one mniej niż 70% wariancji oraz w zmiennych poddanych analizie znajduje się nadal *x6*, która jest słabo współzależna od pozostałych postanowiono w ostatnim kroku ją wyeliminować.

3.2. Analiza

Ostatecznie do analizy zostało wykorzystanych 11 zmienne:

- *x2* - Zgony niemowląt na 10 000 żywych urodzeń
- *x4* - Liczba pracujących położnych przypadających na 10 000 kobiet w wieku produkcyjnym *
- *x8* - Liczba placówek wychowania przedszkolnego na 100 tys. ludności
- *x9* - Liczba szkół podstawowych na 100 tys. ludności
- *x13* – Liczba domów kultury, centrum kultury, świetlic na 100 tys. ludności
- *x14* – Liczba bibliotek publicznych na 100 tys. ludności

- *x19* - Przeciętne miesięczne wydatki na 1 osobę w gospodarstwach domowych w setkach
- *x20* - Przeciętny miesięczny dochód na 1 osobę w gospodarstwie w setkach
- *x21* - % gospodarstw domowych deklarujących brak możliwości realizacji potrzeby tygodniowego wypoczynku rodziny raz w roku *
- *x22* - % gospodarstw domowych, które określiły, że przy aktualnym dochodzie z trudnością "wiążą koniec z końcem"
- *x23* - Wskaźnik zagrożenia ubóstwem po transferach społecznych w % *



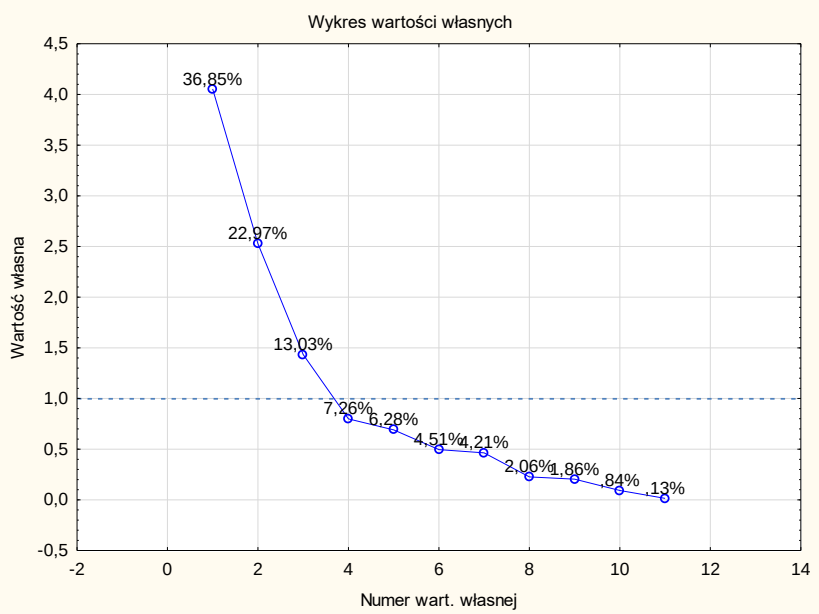
Wykres 13. Korelacje pozostawionych zmiennych.
Źródło: opracowanie własne w programie RStudio.

Pozostawiono zmienne z najwyższymi wartościami korelacji (*Wykres 13*). Na podstawie tych zmiennych zostały wyznaczone składowe(*Tabela* 2121)

Tabela 21. Wartości własne i skumulowany % wariancji przy 11 zmiennych.

Nr składowej	Wartość własna	Skumulowany % wariancji
1	4,054	36,851
2	2,527	59,823
3	1,434	72,857
4	0,799	80,121
5	0,691	86,400
6	0,496	90,908
7	0,463	95,120
8	0,226	97,178
9	0,204	99,035
10	0,092	99,872
11	0,014	100,000

Na podstawie kryterium Kaisera postanowiono ograniczyć liczbę głównych składowych do 3. Dzięki temu wartości własne odpowiadające wybranych składowym są nie mniejsze od 1 a dodatkowo wyjaśniają one względnie dużą część całkowitej zmienności – blisko 73%.



Na podstawie kryterium Cattella bazującego na wykresie osypiska (*Wykres 14*) powinno się szukać miejsca na wykresie, od którego na prawo występuje łagodny spadek wartości własnych. Zgodnie z tym kryterium właśnie byłoby wybranie 4 bądź 3 głównych składowych.

Wykres 14. Wykres osypiska.

3.3 Interpretacja kolejnych składowych

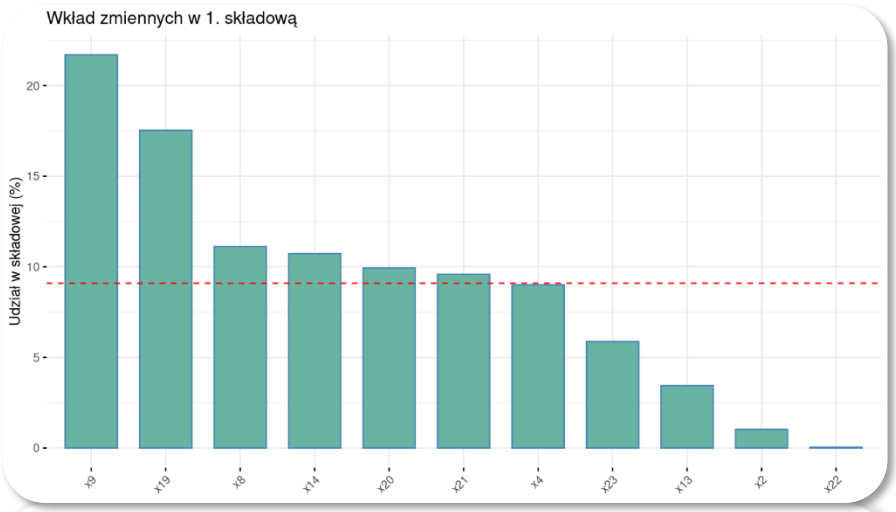
Tabela 22. Korelacje zmiennych z 3 głównymi składowymi.

Zmienna	1 składowa	2 składowa	3 składowa
x2	-0,204	0,639	-0,612
x4	0,604	0,315	0,188
x8	0,671	0,130	0,468
x9	0,938	-0,124	0,117
x13	0,374	-0,587	-0,519
x14	0,659	-0,445	-0,194
x19	-0,843	0,002	0,039
x20	-0,635	0,372	0,475
x21	0,623	0,554	-0,092
x22	0,044	0,709	-0,475
x23	0,488	0,703	0,149

Na podstawie powyższych korelacji (*Tabela 22*) można spróbować zinterpretować składowe. W przypadku pierwszej składowej najwyższe korelacje posiada ona kolejno z:

- *x9* - Liczba szkół podstawowych na 100 tys. ludności (0,938),
- *x19* - Przeciętne miesięczne wydatki na 1 osobę w gospodarstwach domowych w setkach (-0,843),
- *x8* - Liczba placówek wychowania przedszkolnego na 100 tys. ludności (0,671),
- *x14* – Liczba bibliotek publicznych na 100 tys. ludności (0,659),
- *x20* - Przeciętny miesięczny dochód na 1 osobę w gospodarstwie w setkach (-0,635),
- *x21* - % gospodarstw domowych deklarujących brak możliwości realizacji potrzeby tygodniowego wypoczynku rodziny raz w roku * (0,623),
- *x4* - Liczba pracujących położnych przypadających na 10 000 kobiet w wieku produkcyjnym * (0,604)

Składowa ta jest, więc głównie związana z pieniędzmi oraz edukacją. Korelacje *x19*,*x20*, *x2* mają przeciwne znaki od pozostałych. Oznacza to, że wzrost przeciętnych wydatków, dochodów oraz zgonów niemowląt związany jest ze spadkiem procentu gospodarstw, które nie mogą sobie pozwolić na tygodniowy wypoczynek rodziny raz w roku jak i zmniejszeniem się liczby szkół, przedszkoli, bibliotek oraz liczby położnych.



Wykres 15. Wkład zmiennych w budowanie 1 składowej.
Źródło: opracowanie własne w programie RStudio.

Obok (*Wykres 15*) przedstawiono procentowy wkład poszczególnych zmiennych w budowanie pierwszej składowej. Czerwoną linią oznaczono oczekiwany średni wkład – gdyby udział wszystkich zmiennych był jednolity, oczekiwany wkład wyniósłby 1/11≈9%. Zmienne o udziale większym niż wartość odcięta mogą być uważane za ważne w przypadku budowania tej składowej. W związku z tym pierwsza składowa (**Y₁**) jest **miarą finansowo – edukacyjną.** Uwzględnia ona czynniki opisane w pierwszym rozdziale, które były wskazane jako przyczyny niskiej dzietności – brak stabilności finansowej oraz słaby dostęp do edukacji, a więc odpowiada ona najważniejszym kwestiom przed zdecydowaniem się rodzin na dziecko.

Równanie pierwszej składowej:

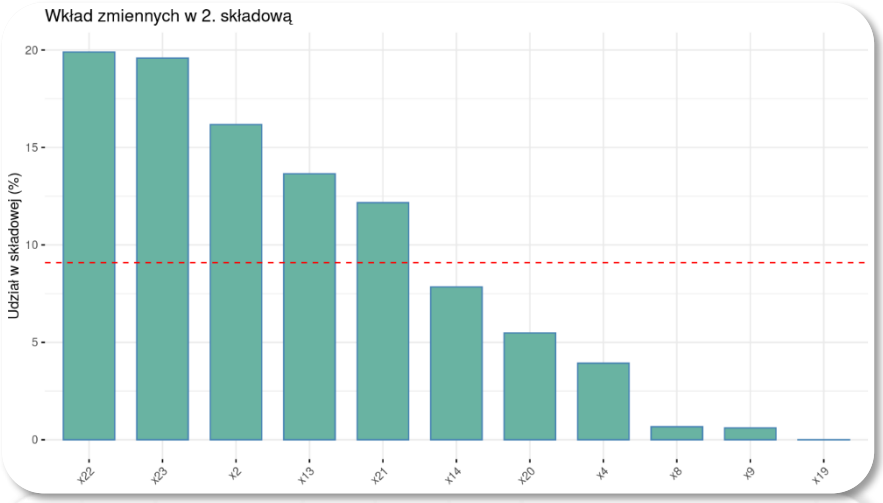
$$\boldsymbol{Y_1} = \underset{(-0{,}204)}{-0{,}101X_2} + \underset{(0{,}604)}{0{,}300X_4} + \underset{(0{,}671)}{0{,}333X_8} + \underset{(0{,}938)}{0{,}466X_9} + \underset{(0{,}374)}{0{,}186X_{13}} + \underset{(0{,}659)}{0{,}327X_{14}} - \underset{(-0{,}843)}{0{,}419X_{19}} - \underset{(-0{,}635)}{0{,}315X_{20}} + \underset{(0{,}623)}{0{,}310X_{21}} + \underset{(0{,}044)}{0{,}022X_{22}} + \underset{(0{,}488)}{0{,}242X_{23}}$$

Należy oczekiwać, że wyższe wartości pierwszej składowej odpowiadają większej liczbie placówek związanych z edukacją (*x8*,*x9*) i większą liczbą bibliotek .Jeśli rodzinie zależy na dobrym dostępie do edukacji i kultury lepszym wyborem byłyby województwa z wysoką wartością pierwszej składowej. Natomiast jeśli ważniejszy jest dla nich wyższy dochód i możliwość zapewnienia rodzinie tygodniowego wypoczynku przynajmniej raz w roku powinni rozważyć województwa z niższą wartością pierwszej składowej.

2

Druga składowa jest mniej ważna (wartość własna wynosi 2,527), lecz jest interpretowalna. Ma ona najwyższe korelacje z:

- *x22* - % gospodarstw domowych, które określiły, że przy aktualnym dochodzie z trudnością "wiążą koniec z końcem" (0,709),
- *x23* - Wskaźnik zagrożenia ubóstwem po transferach społecznych w % * (0,703),
- *x2* - Zgony niemowląt na 10 000 żywych urodzeń (0,639),
- *x13* – Liczba domów kultury, centrum kultury, świetlic na 100 tys. ludności (-0,587)



Wykres 16. Wkład zmiennych w budowanie 2. składowej. Źródło: opracowanie własne w RStudio.

Przy czym te dwie pierwsze zmienne budują w prawie 40 % składową. Według wykresu (*Wykres 16*) wkład pięciu zmiennych jest powyżej oczekiwanego. Piąta zmienną jest *x21*(% gospodarstw domowych deklarujących brak możliwości realizacji potrzeby tygodniowego wypoczynku rodziny raz w roku), której korelacja wynosi 0,554. Zważając na najwyższe korelacje $\mathbf{Y_2}$ można nazwać **miarą problemów finansowych.**

Równanie drugiej składowej:

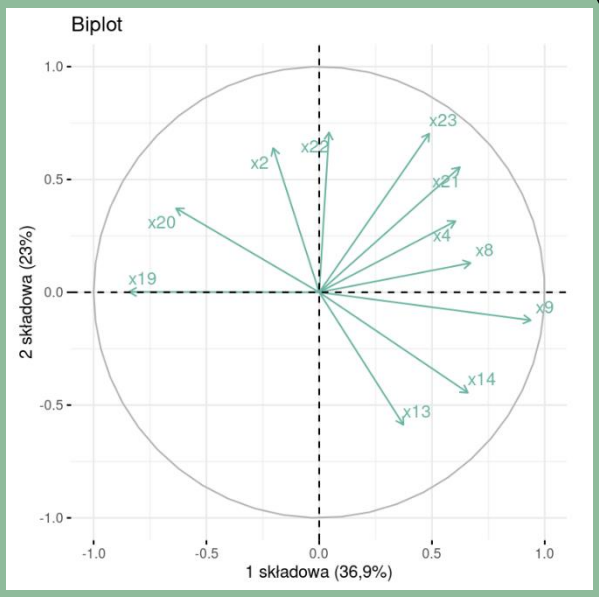
$$\boldsymbol{Y_2} = \underset{(0,639)}{0,402X_2} + \underset{(0,315)}{0,198X_4} + \underset{(0,130)}{0,082X_8} - \underset{(-0,124)}{0,078X_9} - \underset{(-0,587)}{0,369X_{13}} - \underset{(-0,445)}{0,280X_{14}} + \underset{(0,002)}{0,001X_{19}} + \underset{(0,372)}{0,234X_{20}} +$$

$$\underset{(0,554)}{0,349X_{21}} + \underset{(0,709)}{0,446X_{22}} + \underset{(0,703)}{0,443X_{23}}$$

Znaki przeciwne do pozostałych w tym równaniu posiadają zmienne:

- *x9* - Liczba szkół podstawowych na 100 tys. ludności,
- *x13* – Liczba domów kultury, centrum kultury, świetlic na 100 tys. ludności,
- *x14* – Liczba bibliotek publicznych na 100 tys. ludności,

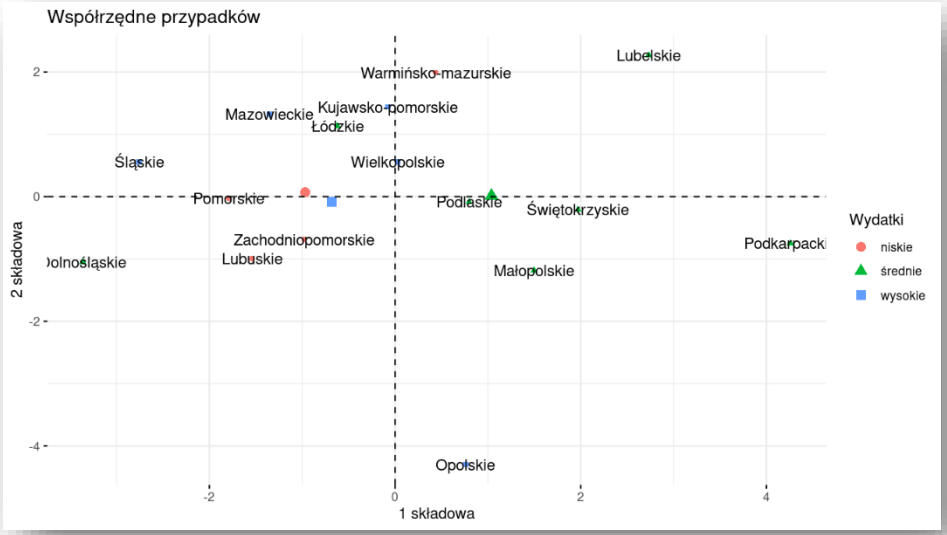
Są to obserwacje związane z rozwojem dzieci. Można zaobserwować, że w województwach z wyższym poziomem ubóstwa liczba powyższych placówek spada. W regionach tych lokalne władze mogą być zmuszone do skoncentrowania się na podstawowych usługach takich jak pomoc socjalna oraz opieka zdrowotna. Dodatkowo w takich obszarach występuje mniejsza świadomość korzyści z wizyt w bibliotece oraz czytania książek. Kultura czytelnicza może być słabo rozwinięta a dzieci mogą nie otrzymywać wystarczającej zachęty do czytania, zarówno w domu jak i w szkole. Podobne wnioski można wysnuć na temat domów kultury i braku świadomości kulturowej. Rodziny mogą nie zdawać sobie sprawy z korzyści płynących z uczestnictwa w zajęciach w takich miejscach. Kolejnym problemem jest brak środków na korzystanie przez dzieci w takich aktywnościach, w przypadku zajęć płatnych oraz trudności w dojeździe do takich miejsc. W związku z niskim zainteresowaniem takimi placówkami maleje ich ilość. Druga składowa ma również umiarkowanie wysoką korelacje z *x2* (zgonami niemowląt na 10 000 żywych urodzeń). Społeczności o niższym statusie ekonomicznym mogą doświadczać trudności w dostępie do podstawowych świadczeń zdrowotnych, odpowiedniej żywności, czy warunków mieszkaniowych, co może wpływać na zdrowie matki i noworodka.



Wykres 17. Dwuwymiarowy wykres 1 i 2 składowej.
Źródło: opracowanie własne w RStudio.

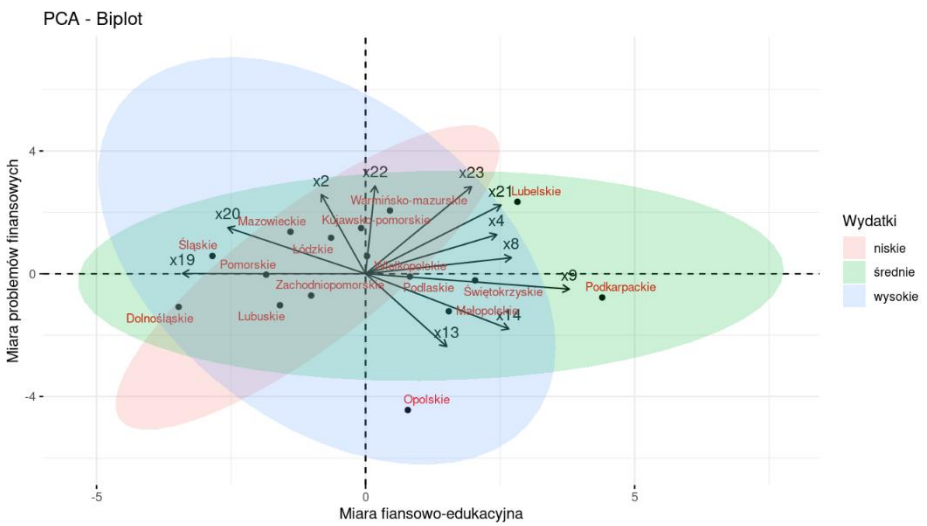
Wykres (Wykres 17. Dwuwymiarowy wykres 1 i 2 składowej.*Wykres 17*) prezentuje, iż zmienna *x19* buduje praktycznie tylko pierwszą składową (znajduje się ona praktycznie na osi X). Mały kąt nachylenia do osi X a w związku z tym duży wkład w budowanie tej składowej mają również zmienne *x9* oraz *x8,* przy czym ta pierwsza ma wyższą korelacje z składową co możemy zaobserwować po różnicy w długości odcinków odpowiadającym zmiennym. Natomiast za drugą składową w największym stopniu odpowiada *x22*. Znajduje się ona nad osią X, więc jej korelacja ze składową jest dodatnia. Zmienne *x14*, *x21* mają podobny wpływ na obydwie składowe – znajdują się między osiami, więc posiadają zbliżone wkład w obydwie składowe. Na podstawie wykresu można również zauważyć ujemną korelacje zmiennej *x14* z *x20,* ich odcinki tworzą przedłużenie.

Na kolejnym wykresie przedstawiono wartości pierwszej oraz drugiej składowej dla województw. Z wykresu można zaobserwować skoncentrowanie zmiennych wokół osi, jedynie województwo opolskie odstaje od reszty. Wyróżnia je wyraźnie niższa wartość drugiej składowej. Oznacza to, że jest w nim wysoka liczba placówek związanych z rozwojem dzieci oraz dosyć niski stan ubóstwa oraz wskaźnik problemów finansowych.



Wykres 18. Wartości 1. i 2. składowej dla województw.
Źródło: Opracowanie własne w programie RStudio.

Województwa wielkopolskie oraz kujawsko-pomorskie leżą na dodatniej części osi OY – ich pierwsza składowa jest bliska zeru. Może to oznaczać, że dane regiony nie są mocno związane z cechami, które reprezentuje ta składowa. Nieopisane znaczniki na wykresie to średnie wartości dla grup wydatków. Kategorie te różnią się średnią wartością pierwszej składowej, średnia wartość drugiej składowej jest natomiast zbliżona i oscyluje wokół zera. Średnio najwyższą wartość pierwszej składowej mają województwa z średnim poziomem wydatków na edukację, oświatę oraz rodzinę a najniższą – województwa z niskim poziomem. Średnio województwa z średnim poziomem wydatków mają więcej placówek edukacyjnych oraz bibliotek. Podobny wniosek można było zauważyć na *Wykres 12.*



Wykres 19. Wykres dwuwymiarowy dla 1. i 2. składowej z elipsami.
Źródło: opracowanie własne w programie RStudio.

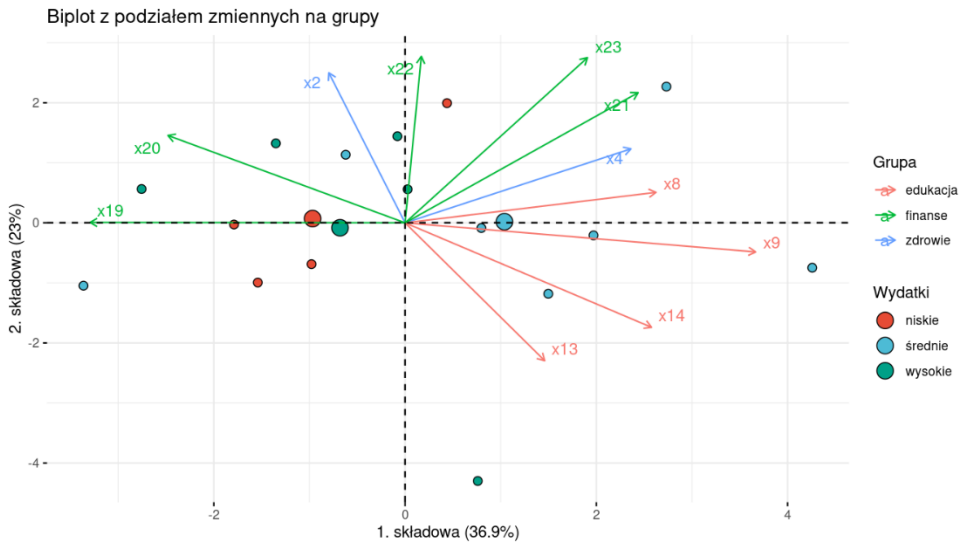
Można również zaobserwować, że poziomy wydatków nie rozdzielają województw w przypadku pierwszej oraz drugiej składowej. Większość obserwacji znajduje się na obszarach wyznaczonych przez elipsy kilku grup. Wyjątki stanowią województwa opolskie (wysoki poziom wydatków) oraz podkarpackie oraz lubelskie (średni poziom wydatków). Elipsa odpowiadająca średniemu poziomu jest zorientowana wokół osi X.

Natomiast pozostałe dwie leża pod kątem do osi Y?

Następnie zmienne wykorzystane do analizy głównych składowych zostały podzielone na 3 kategorie, łącząc z wejściowych kategorii edukację oraz kulturę:

Tabela 23. Podział zmiennych na trzy kategorie.

Zdrowie	Edukacja	Finanse
x2 - Zgony niemowląt na 10 000 żywych urodzeń	*x8* - Liczba placówek wychowania przedszkolnego na 100 tys. ludności	*x19* - Przeciętne miesięczne wydatki na 1 osobę w gospodarstwach domowych w setkach
x4 - Liczba pracujących położnych przypadających na 10 000 kobiet w wieku produkcyjnym *	*x9* - Liczba szkół podstawowych na 100 tys. ludności	*x20* - Przeciętny miesięczny dochód na 1 osobę w gospodarstwie w setkach
	x13 – Liczba domów kultury, centrum kultury, świetlic na 100 tys. ludności	*x21* - % gospodarstw domowych deklarujących brak możliwości realizacji potrzeby tygodniowego wypoczynku rodziny raz w roku *
	x14 – Liczba bibliotek publicznych na 100 tys. ludności	*x22* - % gospodarstw domowych, które określiły, że przy aktualnym dochodzie z trudnością "wiążą koniec z końcem"
		x23 - Wskaźnik zagrożenia ubóstwem po transferach społecznych w % *



Wykres 20. Wykres dwuwymiarowy dla 1. i 2. składowej z podziałem zmiennych na grupy.
Źródło: opracowanie własne w RStudio.

Dzięki powyższemu podziałowi oraz wykresowi (*Wykres 20*) można zauważyć, że zmienne związane z edukacją mają dodatnią korelację z pierwszą składową oraz ujemną korelacje ze drugą składową (za wyjątkiem *x8*). Natomiast zmienne związane z finansami oraz zdrowiem korelują dodatnio z drugą składową.

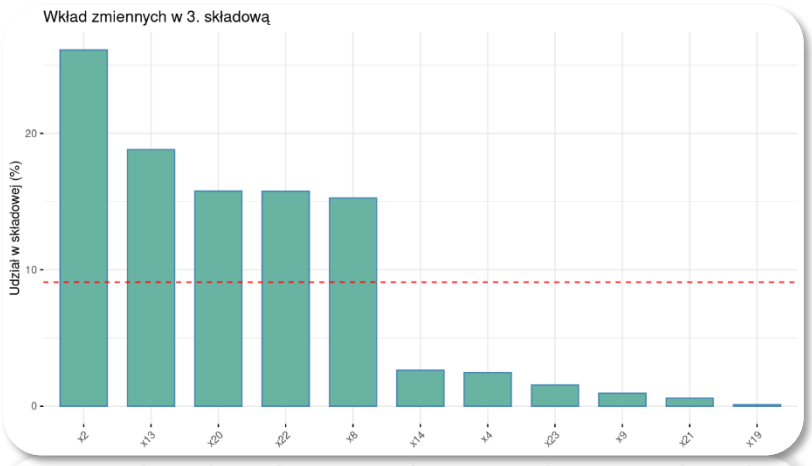
Równanie trzeciej składowej:

$$\boldsymbol{Y_3} = \underset{(-0{,}612)}{-0{,}511X_2} + \underset{(0{,}188)}{0{,}157X_4} + \underset{(0{,}468)}{0{,}391X_8} + \underset{(0{,}117)}{0{,}097X_9} - \underset{(-0{,}519)}{0{,}434X_{13}} - \underset{(-0{,}194)}{0{,}162X_{14}} + \underset{(0{,}039)}{0{,}033X_{19}} + \underset{(0{,}475)}{0{,}397X_{20}}$$

$$- \underset{(-0{,}092)}{0{,}077X_{21}} - \underset{(-0{,}475)}{0{,}397X_{22}} + \underset{(0{,}149)}{0{,}125X_{23}}$$

Ostatnia składowa posiada korelacje wyższe od 0,5 jedynie z:

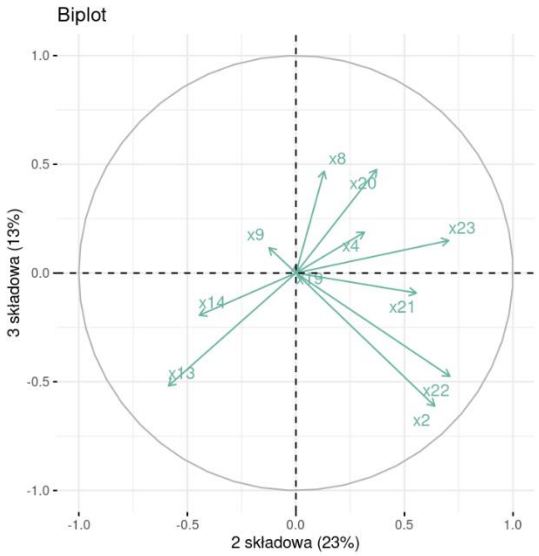
- *x2* - Zgony niemowląt na 10 000 żywych urodzeń (-0,639),
- *x13* – Liczba domów kultury, centrum kultury, świetlic na 100 tys. ludności (0,587)



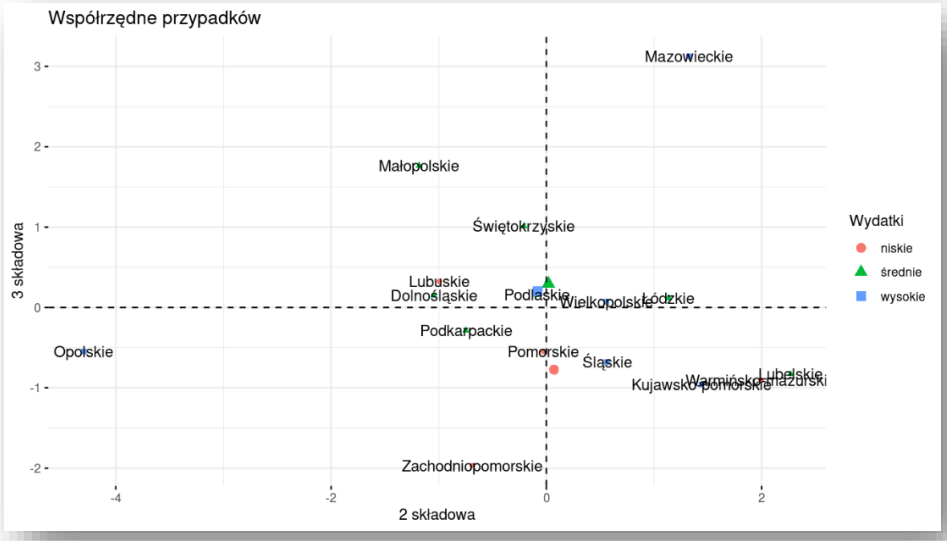
Wykres 21. Wkład zmiennych w budowanie 3. składowej.
Źródło: opracowanie własne w RStudio.

W przypadku trzeciej składowej pięć zmiennych ma udział powyżej 9%. Jednak w tym przypadku są to zmienne bardziej zróżnicowane, związane zarówno ze zdrowiem, kulturą jak i finansami, co utrudnia interpretacje. Jednak ponad 25% udziału w budowaniu tej składowej ma zmienna *x2*. W związku z tym postanowiono **Y3** nazwać **miarą opieki.**

Wykres (*Wykres 22*) pozwala rozpoznać, które zmienne tworzą drugą oraz trzecią składową. Mały kąt nachylenia z osią X oraz stosunkowo długi odcinek ma *x23*, więc można wywnioskować, że buduje ona drugą składową. Zmienna *x8* buduje w większym stopniu trzecią składową. Tymczasem zmienne *x22* oraz *x2* mają duży wpływ w budowanie obydwu składowych, przy czym ta pierwsza ma delikatnie większy wpływ na drugą składową a kolejna na trzecią. Natomiast po długości odcinka związanego z *x19* oraz wcześniejszej analizie właściwe jest stwierdzenie, że buduje ona pierwszą składową i praktycznie nie uczestniczy w tworzeniu kolejnych dwóch. Dodatkowo można zaobserwować ujemne skorelowanie *x9* z *x2*.

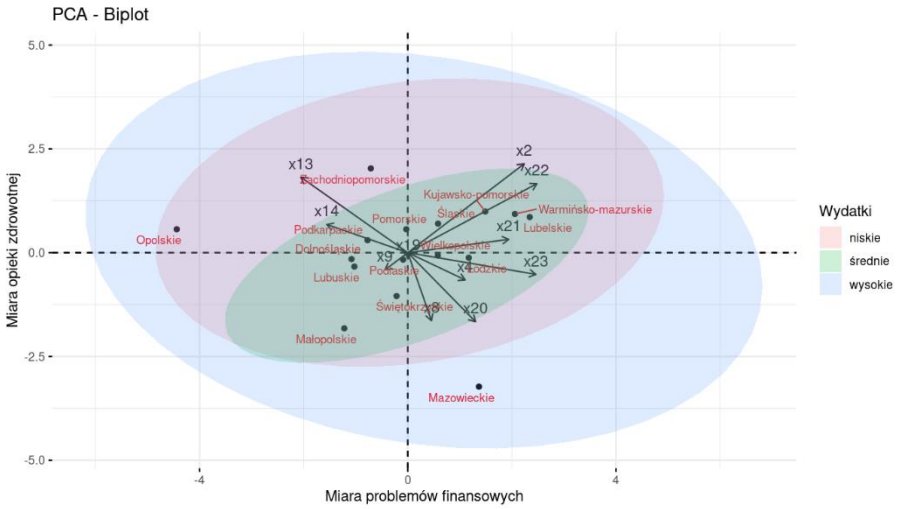


Wykres 22. Dwuwymiarowy wykres 2 i 3 składowej.
Źródło: opracowanie własne w programie RStudio.



Wykres 23. Wartości 2. i 3. składowej dla województw.
Źródło: opracowanie własne w RStudio.

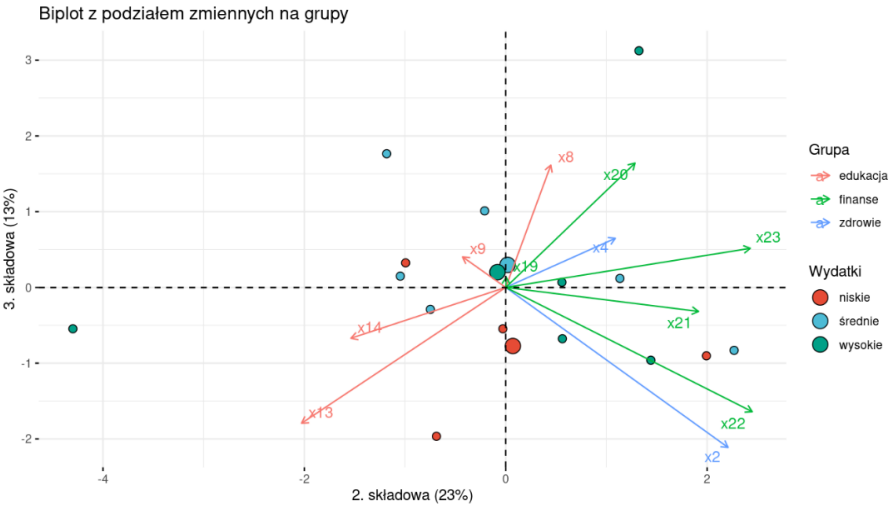
W przypadku trzeciej składowej zauważalnie wyższą wartość ma województwo mazowieckie a niską zachodniopomorskie. Średnio województwa z niskim poziomem wydatków cechują się niższą wartością tej składowej.



Ponownie wykres z elipsami dla wydatków jest słabo interpretowalny, ponieważ obszary nakładają się na siebie. Można jedynie zauważyć, że województwa ze średnimi wydatkami mają najmniej zróżnicowane wartości drugiej oraz trzeciej składowej a z wysoki najbardziej zróżnicowane.

Wykres 24. Wykres dwuwymiarowy dla 2. i 3. składowej z elipsami.
Źródło: opracowanie własne w programie RStudio.

W przypadku trzeciej składowej grupy zmiennych nie układają się po jednej stronie osi. Zarówno zmienne związane z edukacją jak i finansami czy zdrowiem są częściowo dodatnie a częściowo ujemnie skorelowane ze składową.



Wykres 25. Wykres dwuwymiarowy dla 2. i 3. składowej z podziałem zmiennych na grupy.
Źródło: opracowanie własne w programie RStudio.

4. Porządkowanie liniowe

4.1. Wstęp

Porządkowanie liniowe stanowi efektywne narzędzie analizy danych, które umożliwia uporządkowanie elementów w hierarchiczną strukturę, a więc ponumerowanie wszystkich obiektów zbiorów w taki sposób, że obiekt 1 jest najlepszy pod względem badanej cechy, natomiast ostatni najgorszy.
W pracy porządkowanie liniowe zostało użyte w kontekście klasyfikacji województw pod względem różnych cech społeczno-ekonomicznych tak, aby wskazać region najlepszy do życia rodzin z dziećmi.

Przed przystąpieniem do analizy zostały wybrane zmienne na podstawie, których będzie wykonywane porządkowanie. Podobnie jak w przypadku analizy głównych składowych wykluczono te, których zmienność jest poniżej 10 (*Tabela 3*), ponieważ nie różnicowałyby one dobrze zbioru. Tym razem nie pomijając zmiennej *x20*. Wyłączyło ze zbioru:

- *x1* – Procent ludności jaki stanowią dzieci i młodzież,
- *x10* – Liczba szkół średnich na 100 tys. ludności,
- *x12* - Wskaźnik wykrywalności sprawców przestępstw stwierdzonych przez policję w 2022,
- *x15* – Przeciętna powierzchnia użytkowa mieszkania w m^2
- *x16* – Gospodarstwa wyposażone w urządzenie z dostępem do Internetu w % ogółu gospodarstw
- *x17* – Procent mieszkań wyposażonych w łazienkę,
- *x18* – Procent mieszkań wyposażonych w centralne ogrzewanie,
- *x20* – przeciętny miesięczny dochód na 1 osobę w gospodarstwie w setkach,
- *x24* - Przeciętne miesięczne wynagrodzenie brutto w setkach.

Na tym etapie pozostało 17 zmiennych. Następnie przyjrzano się ponownie macierzy korelacji (*Tabela 24*).

Tabela 24. Macierz korelacji 17 zmiennych.

Zmienna	x2	x3	x4	x5	x6	x7	x8	x9	x11	x13	x14	x19	x21	x22	x23	x25	x26
x2	1,000	0,343	-0,019	-0,089	0,032	0,187	-0,255	-0,335	0,333	-0,140	-0,388	0,075	0,228	0,601	0,213	-0,053	-0,100
x3	0,343	1,000	-0,203	0,045	-0,444	-0,026	-0,389	-0,196	-0,329	0,304	0,070	0,036	0,121	0,062	0,185	0,116	-0,318
x4	-0,019	-0,203	1,000	0,404	0,115	0,046	0,455	0,530	-0,105	-0,047	0,203	-0,491	0,244	0,259	0,452	-0,035	-0,108
x5	-0,089	0,045	0,404	1,000	-0,181	-0,071	0,019	0,203	-0,076	0,006	0,204	-0,271	0,147	0,259	0,223	-0,333	0,075
x6	0,032	-0,444	0,115	-0,181	1,000	0,321	-0,065	-0,155	0,265	-0,513	-0,160	0,361	-0,164	0,357	0,079	0,022	0,112
x7	0,187	-0,026	0,046	-0,071	0,321	1,000	-0,452	-0,145	0,096	-0,159	-0,112	0,220	-0,333	-0,030	-0,415	-0,204	-0,101
x8	-0,255	-0,389	0,455	0,019	-0,065	-0,452	1,000	0,655	-0,196	0,021	0,146	-0,464	0,364	-0,067	0,456	0,066	0,065
x9	-0,335	-0,196	0,530	0,203	-0,155	-0,145	0,655	1,000	-0,111	0,250	0,710	-0,755	0,518	-0,047	0,289	-0,151	-0,118
x11	0,333	-0,329	-0,105	-0,076	0,265	0,096	-0,196	-0,111	1,000	-0,462	-0,141	0,132	0,264	0,394	0,091	-0,227	-0,139
x13	-0,140	0,304	-0,047	0,006	-0,513	-0,159	0,021	0,250	-0,462	1,000	0,536	-0,330	-0,054	-0,193	-0,183	0,105	-0,325
x14	-0,388	0,070	0,203	0,204	-0,160	-0,112	0,146	0,710	-0,141	0,536	1,000	-0,419	0,230	-0,102	0,042	0,141	-0,320
x19	0,075	0,036	-0,491	-0,271	0,361	0,220	-0,464	-0,755	0,132	-0,330	-0,419	1,000	-0,557	0,080	-0,348	0,369	0,257
x21	0,228	0,121	0,244	0,147	-0,164	-0,333	0,364	0,518	0,264	-0,054	0,230	-0,557	1,000	0,363	0,729	-0,240	-0,143
x22	0,601	0,062	0,259	0,259	0,357	-0,030	-0,067	-0,047	0,394	-0,193	-0,102	0,080	0,363	1,000	0,362	-0,180	0,186
x23	0,213	0,185	0,452	0,223	0,079	-0,415	0,456	0,289	0,091	-0,183	0,042	-0,348	0,729	0,362	1,000	0,059	-0,353
x25	-0,053	0,116	-0,035	-0,333	0,022	-0,204	0,066	-0,151	-0,227	0,105	0,141	0,369	-0,240	-0,180	0,059	1,000	-0,081
x26	-0,100	-0,318	-0,108	0,075	0,112	-0,101	0,065	-0,118	-0,139	-0,325	-0,320	0,257	-0,143	0,186	-0,353	-0,081	1,000

Z uwagi na to, że nie zawiera ona wysokich korelacji (powyżej 0,8), postanowiono nie eliminować już żadnej cechy.

Tabela 25. Zmienne wykorzystane do porządkowania liniowego.

Kujawsko-pomorskie	50	69	25,79	45,84	11,28	10	53,62	35,03	11,60	11	19	12,52	37,0	14,1	20,4	16,10	38,3
Lubelskie	43	55	32,77	61,39	13,57	4	66,28	44,90	17,70	12	28	12,46	37,9	24,8	22,9	10,45	38,7
Lubuskie	35	52	20,61	58,57	11,56	4	49,49	35,51	19,30	8	24	13,00	28,4	12,2	12,8	12,05	50,8
Łódzkie	40	56	31,47	68,49	11,05	8	55,88	36,91	10,80	8	19	13,54	29,8	19,6	14,5	12,13	92,6
Małopolskie	28	49	26,36	59,32	11,75	7	70,98	44,07	7,60	13	21	10,85	31,0	10,5	14,9	4,14	64,9
Mazowieckie	31	52	27,35	58,41	11,68	5	67,05	34,64	13,90	5	17	15,08	31,4	12,0	20,7	19,38	52,5
Opolskie	24	66	23,61	55,18	8,90	11	53,16	42,55	8,70	20	32	13,67	25,6	8,9	10,1	15,37	46,3
Podkarpackie	39	59	31,74	61,85	9,27	5	72,29	51,13	8,80	16	32	10,05	34,2	14,3	17,6	19,10	51,0
Podlaskie	39	69	31,21	68,92	7,42	2	58,86	36,12	9,10	16	20	11,00	28,0	12,1	19,6	10,67	29,1
Pomorskie	46	58	21,83	48,55	10,14	3	66,49	32,65	13,00	12	13	14,69	24,4	15,8	13,1	15,83	67,5
Śląskie	48	59	27,15	61,06	12,81	29	45,92	32,97	15,60	8	17	14,21	23,3	17,3	11,2	7,89	45,5
Świętokrzyskie	37	49	31,48	53,64	8,73	10	67,05	46,09	17,40	9	21	10,77	34,3	11,7	13,4	7,47	54,9
Warmińsko-mazurskie	46	71	21,70	64,77	7,73	3	57,16	39,67	15,10	10	21	12,21	43,6	19,4	19,4	5,53	60,6
Wielkopolskie	35	59	28,64	54,99	12,19	5	57,25	37,18	12,60	9	18	12,48	30,2	18,6	16,2	11,77	65,8
Zachodniopomorskie	44	55,83	23,79	57,36	9,00	6	51,69	32,49	14,00	20	21	12,89	30,6	17,1	12,0	12,75	53,6

Cechy pozostawione do analizy:

- *x2* - Zgony niemowląt na 10 000 żywych urodzeń
- *x3* - Umieralność okołoporodowa na 10 000 urodzeń żywych i martwych *
- *x4* - Liczba pracujących położnych przypadających na 10 000 kobiet w wieku produkcyjnym *
- *x5* - Liczba przychodni na 100 tys. ludności *
- *x6* - Liczba dzieci na jakie przypada 1 łóżko w szpitalu na oddziale pediatrycznym (w setkach)
- *x7* - Emisja zanieczyszczeń pyłowych na 100km^2
- *x8* - Liczba placówek wychowania przedszkolnego na 100 tys. ludności
- *x9* - Liczba szkół podstawowych na 100 tys. ludności
- *x11* - Przestępstwa stwierdzone przez Policję przeciwko rodzinie i opiece na 10 000 mieszkańców
- *x13* – Liczba domów kultury, centrum kultury, świetlic na 100 tys. ludności
- *x14* – Liczba bibliotek publicznych na 100 tys. ludności
- *x19* - Przeciętne miesięczne wydatki na 1 osobę w gospodarstwach domowych w setkach
- *x21* - % gospodarstw domowych deklarujących brak możliwości realizacji potrzeby tygodniowego wypoczynku rodziny raz w roku *
- *x22* - % gospodarstw domowych, które określiły, że przy aktualnym dochodzie z trudnością "wiążą koniec z końcem"
- *x23* - Wskaźnik zagrożenia ubóstwem po transferach społecznych w % *
- *x25* - Linie autobusowe w km na 100km^2
- *x26* - Wypadki drogowe na 100 tys. ludności

Przed przystąpieniem do analizy zmienne zostały podzielone na pozytywne (stymulanty), których wysokie wartości są korzystne z punktu widzenia analizowanego zjawiska oraz negatywne (destymulanty), których pożądane wartości są niskie. Rozpoczynając od pierwszych zmiennych *x2* oraz *x3* są destymulantami, ponieważ ich wysoka wartość może świadczyć o potencjalnych problemach opieki zdrowotnej nad niemowlętami w danym rejonie. Następna cecha jest natomiast stymulantą, ponieważ większa liczba położnych wiąże się z lepszą dostępnością opieki okołoporodowej oraz łatwiejszym dostępem do specjalistów. Podobnie w przypadku *x5* – bardziej pożądana jest większa ilość przychodni przypadająca na 100 tys. ludności. Ma to związek z łatwiejszym dostępem oraz prawdopodobnie mniejszymi kolejkami przed wizytami co w przypadku często chorujących dzieci jest kluczowe. Kolejna zmienna mówiąca o liczba dzieci na jakie przypada 1 łóżko w szpitalu na oddziale pediatrycznym jest negatywna. Jej wysoka wartość wskazywałaby na ograniczone zasoby i personel medyczny. Rezultatem tego jest długi czas oczekiwania na przyjęcie dziecka do szpitala i przymus przyjmowania wyłącznie osób ciężko chorych. Można było zaobserwować takie sytuacje w szpitalach, również na oddziałach pediatrycznych, podczas Covid-u. Zmienna dotycząca zanieczyszczeń również jest destymulantą. Przechodząc do kolejnej grupy cech – *x8* oraz *x9* oznaczono je jako pozytywne. Duża liczba szkół na 100 tys. a co za tym idzie łatwy dostęp do edukacji jest kluczowy dla rozwoju dzieci. Również biblioteki oraz centra kultury mają znaczący wpływ na postęp dziecka – te zmienne również są stymulantami. W przeciwieństwie do nich zmienna *x11* jest negatywna. Każdemu rodzicowi zależy, aby jego dziecko wychowywało się w bezpiecznej okolicy z jak najmniejszą ilościom przestępstw. Zmienne x19,x21,x22,x23 są również destymulantami. Sytuacje, w których wydatki przekraczają możliwości finansowe, mogą generować stres i niepewność związane z utrzymaniem standardu życia oraz spełnianiem podstawowych potrzeb. Wiąże się to z nierównościami społecznymi oraz mniejszymi możliwościami dziecka na rozwój. Przedostatnia zmienna mówiąca o długości linii autobusowych jest pozytywna. Duża jej wartość sprzyja mobilności i łatwemu dostępowi do różnych obszarów, wyrównują nierówności społeczne spowodowane przez brak samochodu. Co więcej mogą

przyczynić się do zmniejszenia zanieczyszczeń, ponieważ w przypadku rozwiniętej komunikacji część osób może zrezygnować z jazdy własnym pojazdem na rzecz tańszego autobusu. Ostatnia już zmienna jest destymulantą. Wzrost liczby wypadków samochodowych może zagrażać bezpieczeństwu publicznemu. Wypadki te mogą prowadzić do obrażeń, śmierci uczestników ruchu drogowego, co ma negatywny wpływ na dobrostan społeczeństwa i może powodować lęk rodziców o zdrowie ich dzieci.

Tabela 26. Podział zmiennych na stymulanty i destymulanty.

Stymulanty	Destymulanty
x4	x2
x5	x3
x8	x6
x9	x7
x13	x11
x14	x19
x25	x21
	x22
	x23
	x26

4.2. Metoda wzorca

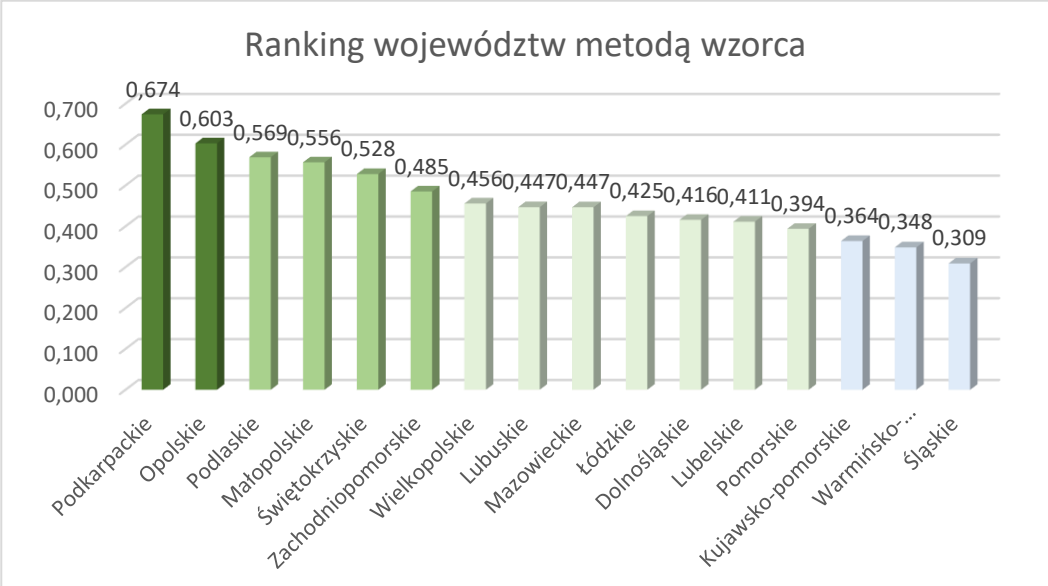
Jako pierwsza w pracy zostanie przedstawiona metoda wzorca rozwoju. Wykorzystuje ona ustandaryzowane wartości zmiennych i opiera się na skonstruowaniu obiektu idealnego – wzorca. Wyznacza się wzorzec pozytywny o najlepszych wartościach, a więc z najwyższymi wartościami pozytywnymi i najniższymi warstwami negatywnymi.

Tabela 27. Standaryzowane zmienne oraz wartości dla wzorca i antywzorca.

	D	D	S	S	D	D	S	S	D	S	S	D	D	D	D	S	D
Województwo	x2	x3	x4	x5	x6	x7	x8	x9	x11	x13	x14	x19	x21	x22	x23	x25	x26
Dolnośląskie	0,00	0,47	-0,97	-0,61	0,34	-0,07	-0,96	-1,45	-0,79	-0,40	-0,28	1,78	-1,84	-0,33	-1,41	0,62	0,68
Kujawsko-pomorskie	1,57	1,50	-0,25	-2,04	0,44	0,42	-0,68	-0,56	-0,36	-0,16	-0,48	-0,19	1,11	-0,26	1,22	0,89	-1,14
Lubelskie	0,57	-0,56	1,51	0,51	1,74	-0,56	0,91	1,16	1,41	0,07	1,29	-0,23	1,26	2,42	1,85	-0,41	-1,12
Lubuskie	-0,57	-1,00	-1,55	0,04	0,59	-0,56	-1,20	-0,48	1,88	-0,88	0,51	0,12	-0,38	-0,73	-0,69	-0,04	-0,28
Łódzkie	0,14	-0,41	1,18	1,67	0,31	0,09	-0,40	-0,23	-0,59	-0,88	-0,48	0,46	-0,14	1,12	-0,27	-0,02	2,62
Małopolskie	-1,57	-1,45	-0,10	0,17	0,70	-0,07	1,50	1,02	-1,52	0,31	-0,09	-1,25	0,07	-1,16	-0,17	-1,86	0,70
Mazowieckie	-1,14	-1,00	0,14	0,02	0,66	-0,40	1,01	-0,63	0,31	-1,59	-0,87	1,45	0,14	-0,78	1,30	1,65	-0,16
Opolskie	-2,14	1,06	-0,80	-0,51	-0,92	0,58	-0,74	0,75	-1,20	1,97	2,08	0,55	-0,87	-1,56	-1,37	0,73	-0,59
Podkarpackie	0,00	0,03	1,25	0,58	-0,70	-0,40	1,67	2,25	-1,17	1,02	2,08	-1,76	0,62	-0,21	0,51	1,58	-0,26
Podlaskie	0,00	1,50	1,12	1,74	-1,76	-0,89	-0,02	-0,37	-1,08	1,02	-0,28	-1,16	-0,45	-0,76	1,02	-0,36	-1,78
Pomorskie	1,00	-0,12	-1,25	-1,60	-0,21	-0,72	0,94	-0,98	0,05	0,07	-1,66	1,19	-1,08	0,17	-0,61	0,83	0,88
Śląskie	1,28	0,03	0,09	0,45	1,30	3,51	-1,65	-0,92	0,80	-0,88	-0,87	0,89	-1,27	0,54	-1,09	-1,00	-0,64
Świętokrzyskie	-0,28	-1,45	1,19	-0,76	-1,01	0,42	1,01	1,37	1,33	-0,64	-0,09	-1,31	0,64	-0,86	-0,54	-1,09	0,01
Warmińsko-mazurskie	1,00	1,80	-1,28	1,06	-1,58	-0,72	-0,24	0,25	0,66	-0,40	-0,09	-0,39	2,25	1,07	0,97	-1,54	0,40
Wielkopolskie	-0,57	0,03	0,47	-0,54	0,95	-0,40	-0,22	-0,18	-0,07	-0,64	-0,68	-0,21	-0,07	0,87	0,18	-0,10	0,76
Zachodniopomorskie	0,71	-0,44	-0,75	-0,16	-0,86	-0,23	-0,92	-1,00	0,34	1,97	-0,09	0,05	0,00	0,49	-0,90	0,12	-0,08
Wzorzec	-2,14	-1,45	1,51	1,74	-1,76	-0,89	1,67	2,25	-1,52	1,97	2,08	-1,76	-1,84	-1,56	-1,41	1,65	-1,78
Antywzorzec	1,57	1,80	-1,55	-2,04	1,74	3,51	-1,65	-1,45	1,88	-1,59	-1,66	1,78	2,25	2,42	1,85	-1,86	2,62

W powyższej tabeli przedstawiono wyniki oraz kolorem czerwonym oznaczono standaryzowane zmienne, które tworzą wzorzec. Można zaobserwować, iż najczęściej najlepsze wartości pochodzą od województw: opolskiego, podlaskiego oraz podkarpackiego. Przyglądając się poszczególnym grupom zmiennych wyznaczonym przez kolory komórek da się dostrzec, że niektóre kategorie mają swojego wyraźnego wzorca. W przypadku zmiennych dotyczących zdrowia ciężko wykryć jednoznacznie najlepsze województwo. Obiekt modelowy składa się z wartości pozyskanych od pięciu różnych regionów, jednak trzy z sześciu zmiennych zdominowało województwo podlaskie. W przypadku kolejnej grupy można jednoznacznie stwierdzić, że w województwie podkarpackim jest najlepszy dostęp do przedszkoli i szkół. Następna kategoria składa się jedynie z jednej cechy, więc w przypadku tej analizy możemy uznać województwo małopolskie za najbezpieczniejsze. Najlepsze wartości zarówno cechy *x13* jak i *x14* pokrywają się w dwóch województwach, lecz w przypadku obydwu zmiennych wartość ta jest z województwa opolskiego. Jest to region z najlepszym dostępem do kultury. Z kolei w kolejnych trzech grupach ciężko byłoby wyznaczyć najlepsze województwo.

W następnym etapie zostały obliczone odległości euklidesowe między każdą obserwacją i wzorcem oraz taksonomiczna miara rozwoju dla każdego obiektu.



Lider - województwo podkarpackie jest dosyć oddalony od wzorca, którego wartość miary taksonomicznej wynosi 1. Można powiedzieć, że z analizy wynika, iż w województwie podkarpackim rodzinom z dziećmi żyje się dwa razy lepiej niż w województwie śląskim, które uplasowało się na końcu rankingu.

Wykres 26. Ranking metodą wzorca.
Źródło: opracowanie własne w programie Excel.

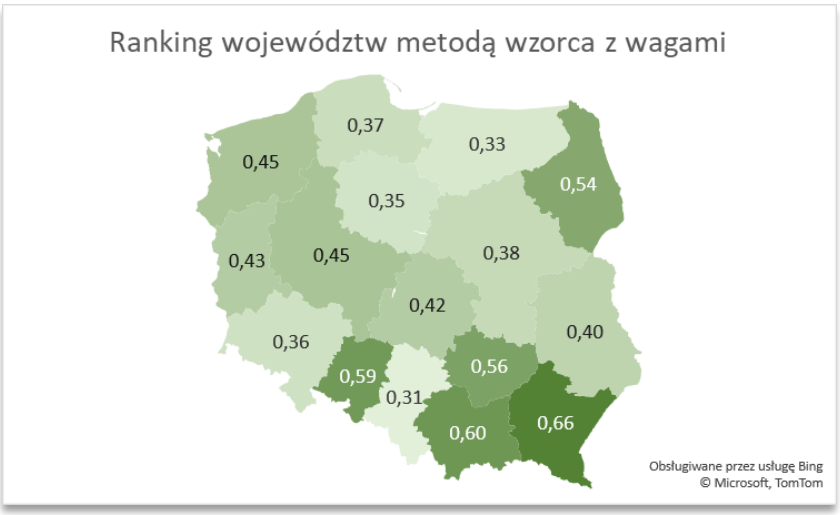
4.3. Metody wzorca z wagami

Kolejne dwie metody opierają się również na metodzie wzorca jednak wykorzystują ważenie zmiennych. Najpierw wagi zostały dobrane wprost proporcjonalnie do stopnia informacyjności cechy mierzonej współczynnikiem zmienności. Najwyższe wagi otrzymały zmienne *x13* oraz *x25*.

Wykres 27. Ranking metodą wzorca z wagami.
Źródło: opracowanie własne w programie Excel.

Można zauważyć, że najlepsze oraz najgorsze województwo pozostało bez zmian. Wrosła jednak różnica pomiędzy ich miarami. Pomiędzy pierwszym oraz drugim województwem w rankingu jest widoczna spora różnica w wielkości miary taksonomicznej. Natomiast różnice pomiędzy województwami z środka rankingu są niewielkie. Dodatkowo województwo śląskie wyraźnie odstaje nawet od pozostałych z końca hierarchii. Można, więc powiedzieć, że ta metoda daje nam wyraźną czwórkę najlepszych województw oraz jednoznacznie wskazuje to najgorsze.

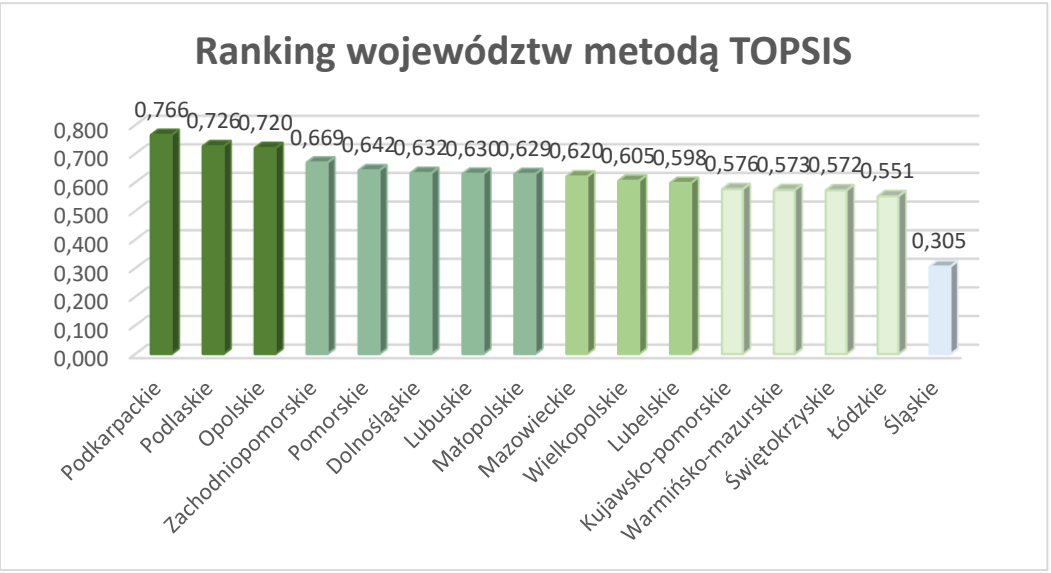
W kolejnej metodzie wagi zostały wyliczone na podstawie stopnia skorelowania z pozostałymi zmiennymi. Podczas analizy głównych składowych zauważono, że *x9* oraz *x19* mają najwięcej wysokich korelacji z pozostałymi i to właśnie tym zmiennym zostały przypisane teraz najwyższe wagi. Ponownie początek oraz koniec rankingu reprezentowany jest przez te same województwa. Jednak tym razem województwo śląskie nie jest zdecydowanie gorsze od pozostałych.



Wykres 28. Ranking metodą wzorca z wagami 2.
Źródło: opracowanie własne w programie Excel.

4.4. Pozostałe metody

Następna metoda – TOPSIS, wykorzystuje wzorzec pozytywny oraz wzorzec negatywny (antywzorzec). Zmienne normalizuje się za pomocą przekształcenia ilorazowego a ranking określa się na podstawie wartości zmiennej agregowanej liczonej przy pomocy odległości obiektów od wzorca i antywzorca.



Wykres 29. Ranking metodą TOPSIS.
Źródło: opracowanie własne w programie Excel.

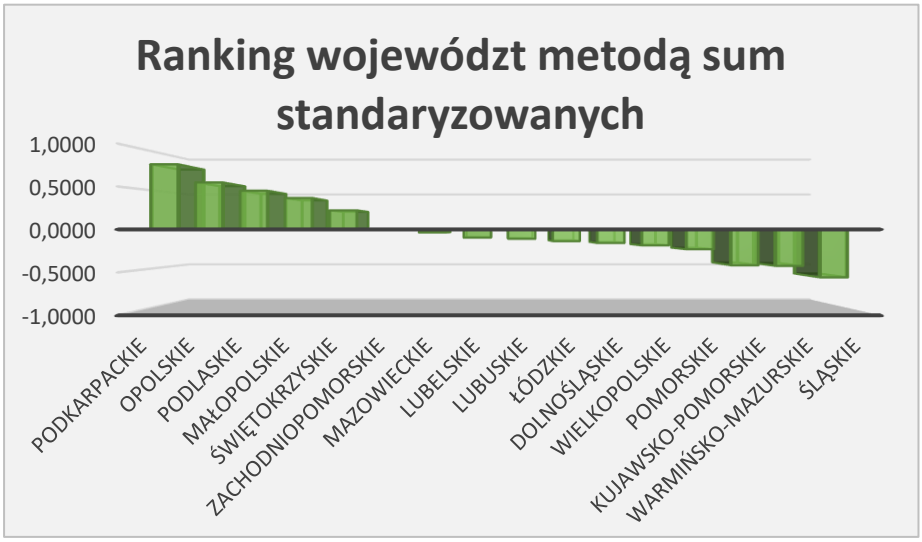
Niezmienne lider pozostał ten sam jednak bliżej mu do wzorca (jego wartość wynosi 1) niż w przypadku standardowej metody wzorca rozwoju. Ponadto województwa na „podium" oraz ostatnie miejsce również się nie zmieniły. Da się zauważyć, że pierwsze trzy województwa tworzą podgrupę, w której wartość zmiennej agregowanej są powyżej 0,7.

Metoda sumy rang opiera się na rangowaniu obiektów ze względu na każdą cechę po kolei (z uwzględnieniem jej charakteru) a następnie następuje sumowanie rang i ustalanie ich pozycji. W przypadku zmiennych ciągłych może być ona mniej dokładna od metod opartych na normalizacji zmiennych. Przyczyną tego jest fakt iż przyporządkowane obiektom liczby naturalne nie oddają rzeczywistych różnic pomiędzy obiektami. W wykonanej analizie w przypadku kilku obiektów z tą samą wartością zastosowano rangi wiązane. Pomimo, że zmienne wejściowe są ciągłe to wyniki (*Tabela 28*) są zbliżone do otrzymanych poprzednimi metodami.

Tabela 28. Ranking metodą sumy rang.

Ranking	Województwo
1	Podkarpackie
2	Podlaskie
3	Opolskie
4	Małopolskie
5	Świętokrzyskie
6	Lubelskie
7	Lubuskie
8	Zachodniopomorskie
9	Mazowieckie
10	Łódzkie
11	Dolnośląskie
12	Pomorskie
13	Wielkopolskie
14	Warmińsko-mazurskie
15	Kujawsko-pomorskie
16	Śląskie

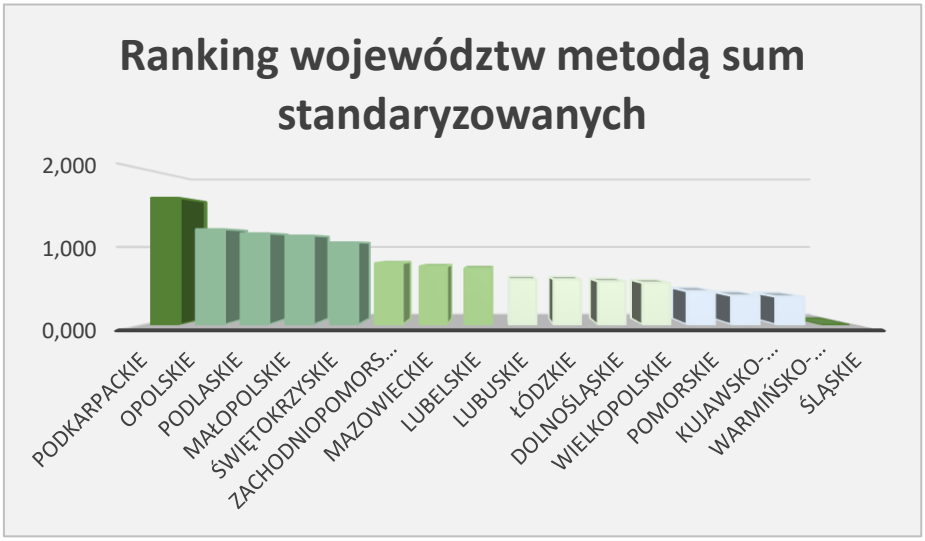
Przedostatnia metoda – metoda standaryzowanych sum opiera się na standaryzowanych zmiennych z uwzględnieniem ich charakteru. Następnie wyznacza się liniową funkcję porządkującą, nazywana „wielkością jednostki", będącą średnią arytmetyczną znormalizowanych wartości cech. Im wyższa jest jej wartość, tym wyższe miejsce w rankingu zajmuje obiekt. Miara ta przyjmuje wartość równą zero – neutralną, wyznaczającą średni poziom zjawiska złożonego.



Wykres 30. Ranking metodą sum standaryzowanych.
Źródło: opracowanie własne w programie Excel.

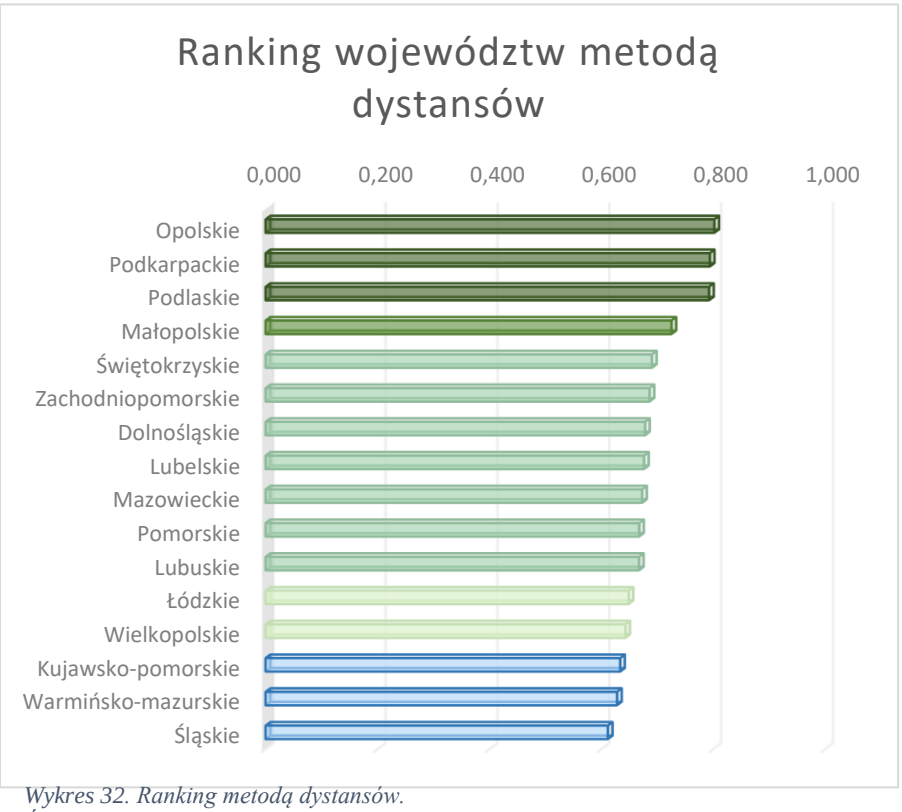
Można zaobserwować iż średni poziom wyznaczany jest przez szóste województwo w rankingu – zachodniopomorskie.
Aby ułatwić dalszą interpretację postanowiono znormalizować wielkość jednostki poprzez odjęcie od niej minimalnej wartości wskaźnika. Dzięki temu nastąpiło przesunięcie w skali wskaźnika, tak że rozpoczyna się on w zerze a nie ma ustalonej górnej granicy.

W wyniku tego łatwiej zauważyć, że lider oraz ostatnia pozycja znacząco różnią się od swoich sąsiadów. Pozostałe województwa można podzielić na cztery podgrupy regionów podobnych do siebie.



Wykres 31. Ranking metodą sum standaryzowanych po przekształceniu.
Źródło: opracowanie własne w programie Excel.

Ostatnia wykorzystana metoda porządkowania liniowego to metoda dystansów. Opiera się ona na odległości rzeczywistych obiektów od obiektu najlepszego pod względem danej cechy, zależnie od charakteru obiektu będzie to województwo z najwyższą lub najniższą wartością. Wskaźnik uzyskuje się poprzez uśrednienie wszystkich dystansów.



Wykres 32. Ranking metodą dystansów.
Źródło: opracowanie własne w programie Excel.

Zaszła zmiana na pozycji lidera. Województwo opolskie zajęło pierwsze miejsce w hierarchii, natomiast województwo podkarpackie spadło na drugie miejsce. Jednak różnice w wskaźniku pomiędzy regionami na „podium" są niewielkie. Największą różnicę można zaobserwować między trzecią a czwartą lokatą rankingu. Następne siedem pozycji zajmują województwa niewiele różniące się od siebie. Województwo śląskie ponownie znalazło się na końcu rankingu. Natomiast tym razem ma zbliżone wartości wskaźnika do województw zajmujących 14. i 15. miejsce.

4.5. Podsumowanie

Poniżej w tabeli zebrane zostały wyniki wszystkich wykorzystanych metod porządkowania liniowego:

Tabela 29. Wyniki porządkowania liniowego.

Województwo	Metoda wzorca bez wag	Metoda wzorca z waga 1	Metoda wzorca z waga 2	TOPSIS	Metoda sumy rang	Metoda sum standaryzowanych	Metoda dystansów
Podkarpackie	1	1	1	1	1	1	2
Opolskie	2	3	3	3	3	2	1
Podlaskie	3	2	5	2	2	3	3
Małopolskie	4	7	2	8	4	4	4
Świętokrzyskie	5	13	4	14	5	5	5
Zachodniopomorskie	6	4	7	4	8	6	6
Wielkopolskie	7	9	6	10	13	12	13
Lubuskie	8	8	8	7	7	9	11
Mazowieckie	9	10	11	9	9	7	9
Łódzkie	10	14	9	15	10	10	12
Dolnośląskie	11	5	13	6	11	11	7
Lubelskie	12	12	10	11	6	8	8
Pomorskie	13	6	12	5	12	13	10
Kujawsko-pomorskie	14	11	14	12	15	14	14
Warmińsko-mazurskie	15	15	15	13	14	15	15
Śląskie	16	16	16	16	16	16	16

Na podstawie analiz za najlepsze województwo dla rodzin z dziećmi można uznać podkarpackie. Mimo, iż w ostatniej metodzie uplasowało się na drugiej pozycji to różnica w metodzie dystansów do województwa opolskiego była niewielka a w pozostałych metodach często województwo podkarpackie było wyróżniającym się liderem. Wybierając miejsce do założenia rodziny warto też rozważyć województwa opolskie i podlaskie, które zajmowały głównie drugie i trzecie miejsca w rankingach. Dobrym wyborem może też być województwo małopolskie, które cztery razy zajęło 4. pozycję a raz zostało wiceliderem. Wszystkie te cztery województwa zostały wcześniej wyróżnione jako najlepsze ze względu na jakąś kategorię zmiennych. Środkowe pozycje w tabeli różnią się w zależności od wykorzystanej metody. Można jednak powiedzieć, iż województwo lubuskie plasuje się zawsze w okolicach ósmego miejsca, natomiast województwo mazowieckie w okolicach dziewiątego. Województwo pomorskie w dwóch porządkowaniach zajęło wyższe miejsce jednak częściej plasowało się za dziesiątym miejscem. Czternastą oraz piętnastą pozycję zajmują województwa kujawsko-pomorskie i warmińsko-mazurskie. Najgorszym województwem dla rodzin z dziećmi okazało się województwo śląskie, które przy pomocy każdej metody zajęło ostatnią pozycję w rankingu i często wyraźnie odbiegało od pozostałych. Na podstawie wcześniejszych obserwacji można było zauważyć, że województwo to szczególnie wyróżnia się dużym poziomem zanieczyszczeń w porównaniu do pozostałych.

5. Analiza skupień

5.1. Wprowadzenie

Celem analizy skupień jest wykrycie grup, czyli jednorodnych podzbiorów obiektów. Analiza bazuje na pomiarze stopnia podobieństwa lub zróżnicowania obiektów. O podobieństwie obiektów mówimy, gdy ich pewne własności są zbliżone. Do oceny podobieństwa lub zróżnicowania obiektów wykorzystywane są różne miary, których zadaniem jest wskazanie w jakim stopniu porównywane obiekty są podobne bądź niepodobne do siebie ze względu na wartości opisujących je zmiennych. Wynik grupowania zależy od tego jak zdefiniowana zostanie odległość między obserwacją a klasą oraz jak te odległość będziemy wyliczać. W pracy zostaną wykorzystane miary odległości (niepodobieństwa):

- **Metryka miejska (Manhattan):**

$$d_{rs} = \sum_{j=1}^{p} |x_{rj} - x_{sj}|$$

  Jest mało czujna na obserwacje odstające. Najważniejszym założeniem tej metryki jest wzajemne nieskorelowanie zmiennych.

- **Metryka euklidesowa:**

$$d_{rs} = \sqrt{\sum_{j=1}^{p} (x_{rj} - x_{sj})^2}$$

  Powszechnie znana jest jako odległość euklidesowa dwóch punktów w przestrzeni p-wielowymiarowej. Jest to najpopularniejsza odległość taksonomiczna. Można również posługiwać się kwadratem tej metryki – **kwadratową odległością euklidesową**:

$$(d_{rs})^2 = \sum_{j=1}^{p} (x_{rj} - x_{sj})^2$$

  Jest ona bezpiecznym wyborem, jednak podobnie jak metrykę miejską możemy ją stosować, gdy zmienne są nieskorelowane. Natomiast w praktyce to ograniczenie często nie jest uwzględniane.

- **Metryka maximum** lub **odległość Czebyszewa**:

$$d_{rs} = \max_j |x_{rj} - x_{sj}|$$

- **Metryka Canberra**:

$$d_{rs} = \sum_{j=1}^{p} \frac{|x_{rj} - x_{sj}|}{(x_{rj} + x_{sj})}$$

  Jest pewną odmianą metryki Manhattan. Przyjmuje wartości z przedziału [0,p] (p to liczba cech). Cechuje się dużą wrażliwością na małe zmiany wartości x_{rj} i x_{sj} bliskie 0.

Do grupowania zostaną wykorzystane metody hierarchiczne. Ich podstawą jest założenie, że wyjściowo każdy element zbioru uznawany jest za odrębną grupę jednoelementową. Wykorzystując miary odległości przeprowadza się sekwencyjne grupowanie obiektów w zależności od tego jak daleko są względem siebie. Pierwsze skupienie zawsze tworzy się z dwóch obiektów najbliżej leżących (o najmniejszej odległości). W następnych krokach dołączą się pojedynczy obiekt do już istniejącego skupienia bądź łączy skupienia, które są według najmniejszej odległości. Przed wykonaniem kolejnego etapu należy w określony sposób przeliczyć odległości między najnowszym skupieniem i pozostałymi, tak aby było ono w jednoznaczny sposób reprezentowane w macierzy odległości. Zawsze należy korzystać z uaktualnionej wersji macierzy.

Zatem każdy krok obejmuje łączenie dwóch najbliższych obiektów oraz przeliczenie odległości od nowego ugrupowania. Ciąg takich iteracji tworzy hierarchię klasyfikacji.
Grupowanie hierarchiczne nie jest procesem jednolitym. Istnieje kilka różnych wariantów (metod, kryteriów, strategii) przeliczania odległości. Najbardziej znanymi są:

- **Metoda najbliższego sąsiada**
  Polega na przeliczaniu odległości pomiędzy skupieniami według kryterium najmniejszej odległości – odległość między nowym skupieniem a skupieniem *i* wyznaczane są przez najbliższe obiekty w tym nowym skupieniu w stosunku do pozostałych skupień. Pierwsze skupienie tworzone jest przez obiektu o najmniejszej odległości. Następnie przed przystąpieniem do drugiego etapu należy określić odległości nowego skupienia od pozostałych. Jeżeli jako pierwsze połączyły się skupienia *r* z *s* to należy umieścić nowe odległości w wierszu *r* i kolumnie *r* (ponieważ *r* jest przed *s*), usuwając wiersz i kolumnę o numerze *s*. Zgodnie z kryterium będą to odległości:
  $$d_{ip} = \min(d_{ir}, d_{is}) \qquad (i = 1, \dots, n; i \neq r; i \neq s)$$
  Metoda ta poszukuje skupiska izolowane, ale nie zwraca uwagi na ich wewnętrzną spójność. Własność tej metody polegająca na jej zdolności do ujawniania długich, pierścieniowych grup znana jest jako tworzenie łańcuchów (chaining).
- **Metoda najdalszego sąsiada**
  O połączeniach między skupieniami decyduje odległość między najbardziej oddalonymi obiektami w skupieniach. Na każdym etapie łączy się skupienia o najmniejszej z największych odległości. W tej metodzie elementy macierzy odległości przekształca się zgodnie z kryterium:
  $$d_{ip} = \max(d_{ir}, d_{is})$$
  Poza tym procedura jest analogiczna do metody najbliższego sąsiada. Procedura ta skupia się na wewnątrzgrupowej spójności, ponieważ minimalizuje międzygrupowe odległości. Tworzy ona grupy spójne, ma jednak tendencję do tworzeniu kilku małych skupień. Wydłuża odległości w porównaniu do poprzedniej metody, jednak podobnie jak poprzednia jest czuła na występowania obserwacji oddalonych.
- **Metoda średniej grupowej**
  Jest to procedura przeliczania odległości bazująca na średniej arytmetycznej odległości między wszystkimi parami obiektów należących do dwóch porównywanych grup:
  $$d_{ip} = \frac{\sum_{j \in i} \sum_{k \in p} d_{jk}}{n_i n_p},$$
  gdzie:
  - d_{jk} – odległość między j-tym obiektem należącym do i-tego skupienia (j=1,..,n_i) a k-tym obiektem należącym do nowotworzonego skupienia p=rUs(k=1,…,np.), przy cyzm suma rozciąga się na wszystkie $n_i \cdot n_p$ par obiektów z dwóch skupień, przy czym $n_p = n_r + n_s$

  Łączone są ze sobą skupienia, dla których średnia odległość jest najmniejsza. Metoda ta nie zależy już w tak dużym stopniu od wartości skrajnych jak dwie poprzednie. Cechuje się ona łączeniem skupień o małych i raczej równych wariancjach. Daje ona strukturę pośrednią między metodą najbliższego i najdalszego sąsiada. Odległości między grupami są bardziej wyważone – mniejsze niż w metodzie najdalszego sąsiada, ale większe niż w metodzie najbliższego.
- **Metoda Warda**
  Jest ostatnią metodą jaka zostanie wykorzystana w pracy. Różni się ona od poprzednich metod sposobem wyliczania odległości jak również kryterium łączenia skupień. Stara się na każdym etapie zoptymalizować otrzymany podział poprzez łączenie dwóch elementów, używając w tym celu kryterium minimalnego wzrostu łącznej wewnątrzgrupowej sumy kwadratów odchyleń wszystkich zmiennych dla każdego obiektu od ich średnich grupowych. Metoda ta posiada cechę tworzenia grup

o tej samej w przybliżeniu liczebności. Jeżeli pojedynczy obiekt ma taką samą odległość od centroidów różnolicznych skupień to połączy się on ze skupieniem mniej licznym. Zapobiega to tworzeniu się łańcuchów. Na dalszych etapach, gdy korzystając z metody Warda na przeliczonych już odległościach poziomy łączenia skupień są większe niż w innych metodach pomijając metodę najdalszego sąsiada.

Jako efekt otrzymujemy ostateczną grupę obiektów oraz wykres hierarchicznego uporządkowania nazywany drzewkiem połączeń lub dendrogramem. Ukazuje on ważne cechy przeprowadzonego łączenia. Poziome linie na wykresie, które łączą skupienia nazywamy węzłami a ostępy pomiędzy dwoma następnymi węzłami - interwęzłami

5.2. Analiza

Tabela 30. Zmienne wykorzystane do analizy skupień.

Województwo	x2	x3	x4	x5	x6	x7	x8	x9	x11	x13	x14	x19	x21	x22	x23	x25	x26
Dolnośląskie	39	62	22,91	54,60	11,11	7	51,38	29,95	10,10	10	20	15,61	20,0	13,8	9,9	14,93	64,6
Kujawsko-pomorskie	50	69	25,79	45,84	11,28	10	53,62	35,03	11,60	11	19	12,52	37,0	14,1	20,4	16,10	38,3
Lubelskie	43	55	32,77	61,39	13,57	4	66,28	44,90	17,70	12	28	12,46	37,9	24,8	22,9	10,45	38,7
Lubuskie	35	52	20,61	58,57	11,56	4	49,49	35,51	19,30	8	24	13,00	28,4	12,2	12,8	12,05	50,8
Łódzkie	40	56	31,47	68,49	11,05	8	55,88	36,91	10,80	8	19	13,54	29,8	19,6	14,5	12,13	92,6
Małopolskie	28	49	26,36	59,32	11,75	7	70,98	44,07	7,60	13	21	10,85	31,0	10,5	14,9	4,14	64,9
Mazowieckie	31	52	27,35	58,41	11,68	5	67,05	34,64	13,90	5	17	15,08	31,4	12,0	20,7	19,38	52,5
Opolskie	24	66	23,61	55,18	8,90	11	53,16	42,55	8,70	20	32	13,67	25,6	8,9	10,1	15,37	46,3
Podkarpackie	39	59	31,74	61,85	9,27	5	72,29	51,13	8,80	16	32	10,05	34,2	14,3	17,6	19,10	51,0
Podlaskie	39	69	31,21	68,92	7,42	2	58,86	36,12	9,10	16	20	11,00	28,0	12,1	19,6	10,67	29,1
Pomorskie	46	58	21,83	48,55	10,14	3	66,49	32,65	13,00	12	13	14,69	24,4	15,8	13,1	15,83	67,5
Śląskie	48	59	27,15	61,06	12,81	29	45,92	32,97	15,60	8	17	14,21	23,3	17,3	11,2	7,89	45,5
Świętokrzyskie	37	49	31,48	53,64	8,73	10	67,05	46,09	17,40	9	21	10,77	34,3	11,7	13,4	7,47	54,9
Warmińsko-mazurskie	46	71	21,70	64,77	7,73	3	57,16	39,67	15,10	10	21	12,21	43,6	19,4	19,4	5,53	60,6
Wielkopolskie	35	59	28,64	54,99	12,19	5	57,25	37,18	12,60	9	18	12,48	30,2	18,6	16,2	11,77	65,8
Zachodniopomorskie	44	55,83	23,79	57,36	9,00	6	51,69	32,49	14,00	20	21	12,89	30,6	17,1	12,0	12,75	53,6

Podczas analizy zostaną uwzględnione te same zmienne co w przypadku porządkowania liniowego:

- *x2* - Zgony niemowląt na 10 000 żywych urodzeń
- *x3* - Umieralność okołoporodowa na 10 000 urodzeń żywych i martwych *
- *x4* - Liczba pracujących położnych przypadających na 10 000 kobiet w wieku produkcyjnym *
- *x5* - Liczba przychodni na 100 tys. ludności *
- *x6* - Liczba dzieci na jakie przypada 1 łóżko w szpitalu na oddziale pediatrycznym (w setkach)
- *x7* - Emisja zanieczyszczeń pyłowych na 100km^2
- *x8* - Liczba placówek wychowania przedszkolnego na 100 tys. ludności
- *x9* - Liczba szkół podstawowych na 100 tys. ludności
- *x11* - Przestępstwa stwierdzone przez Policję przeciwko rodzinie i opiece na 10 000 mieszkańców
- *x13* – Liczba domów kultury, centrum kultury, świetlic na 100 tys. ludności
- *x14* – Liczba bibliotek publicznych na 100 tys. ludności
- *x19* - Przeciętne miesięczne wydatki na 1 osobę w gospodarstwach domowych w setkach
- *x21* - % gospodarstw domowych deklarujących brak możliwości realizacji potrzeby tygodniowego wypoczynku rodziny raz w roku *
- *x22* - % gospodarstw domowych, które określiły, że przy aktualnym dochodzie z trudnością "wiążą koniec z końcem"
- *x23* - Wskaźnik zagrożenia ubóstwem po transferach społecznych w % *
- *x25* - Linie autobusowe w km na 100km^2
- *x26* - Wypadki drogowe na 100 tys. ludności

Celem analizy jest zidentyfikowanie podobieństw i różnic między poszczególnymi jednostkami terytorialnymi oraz grupowanie ich, aby uzyskać województwa podobne pod względem warunków do życia rodzin z dziećmi.

Jako etap wstępny przed przeprowadzeniem analizy zostaną ukazane macierze odległości.

Tabela 31. Macierz odległości euklidesowej.

	Dolnośląskie	Kujawsko-pomorskie	Lubelskie	Lubuskie	Łódzkie	Małopolskie	Mazowieckie	Opolskie	Podkarpackie	Podlaskie	Pomorskie	Śląskie	Świętokrzyskie	Warmińsko-mazurskie	Wielkopolskie	Zachodniopomorskie
Dolnośląskie	0.000	1400.129	2104.577	572.216	1324.508	1219.338	934.318	1064.759	1707.748	1952.748	460.284	1159.807	1217.071	1210.152	349.291	460.121
Kujawsko-pomorskie	1400.129	0.000	1107.479	1157.607	3918.930	2415.039	1316.430	1427.404	1543.648	1024.866	1492.312	1197.310	1428.446	1189.628	1324.985	834.464
Lubelskie	2104.577	1107.479	0.000	1137.561	3473.944	1538.636	973.888	1665.401	622.852	963.391	1934.271	1858.182	802.554	1182.085	1295.086	1076.583
Lubuskie	572.216	1157.607	1137.561	0.000	2220.565	1086.433	591.318	834.691	1327.259	1341.275	1060.811	1076.751	711.453	1109.844	554.099	350.656
Łódzkie	1324.508	3918.930	3473.944	2220.565	0.000	1546.136	2139.735	3270.680	2623.104	4423.828	1434.353	3027.705	2080.042	1726.625	965.676	1937.548
Małopolskie	1219.338	2415.039	1538.636	1086.433	1546.136	0.000	687.842	1379.482	899.017	2261.052	1037.375	2365.663	414.744	1448.202	592.990	1175.984
Mazowieckie	934.318	1316.430	973.888	591.318	2139.735	687.842	0.000	1258.083	871.687	1380.551	823.739	1749.085	522.096	1314.401	506.306	840.218
Opolskie	1064.759	1427.404	1665.401	834.691	3270.680	1379.482	1258.083	0.000	1084.725	1229.571	1888.662	1706.356	1285.657	1780.893	1167.740	942.275
Podkarpackie	1707.748	1543.648	622.852	1327.259	2623.104	899.017	871.687	1084.725	0.000	1311.243	1522.201	2393.177	668.800	1264.637	1076.363	1163.050
Podlaskie	1952.748	1024.866	963.391	1341.275	4423.828	2261.052	1380.551	1229.571	1311.243	0.000	2415.204	1716.791	1743.207	1567.130	1826.257	1209.975
Pomorskie	460.284	1492.312	1934.271	1060.811	1434.353	1037.375	823.739	1888.662	1522.201	2415.204	0.000	1889.231	964.169	1227.492	433.583	695.842
Śląskie	1159.807	1197.310	1858.182	1076.751	3027.705	2365.663	1749.085	1706.356	2393.177	1716.791	1889.231	0.000	1568.869	1806.067	1446.355	935.558
Świętokrzyskie	1217.071	1428.446	802.554	711.453	2080.042	414.744	522.096	1285.657	668.800	1743.207	964.169	1568.869	0.000	1200.555	571.279	817.220
Warmińsko-mazurskie	1210.152	1189.628	1182.085	1109.844	1726.625	1448.202	1314.401	1780.893	1264.637	1567.130	1227.492	1806.067	1200.555	0.000	711.609	818.203
Wielkopolskie	349.291	1324.985	1295.086	554.099	965.676	592.990	506.306	1167.740	1076.363	1826.257	433.583	1446.355	571.279	711.609	0.000	487.089
Zachodniopomorskie	460.121	834.464	1076.583	350.656	1937.548	1175.984	840.218	942.275	1163.050	1209.975	695.842	935.558	817.220	818.203	487.089	0.000

Można zaobserwować , że wyliczając odległości za pomocą kwadratowej odległości euklidesowej (*Tabela 31*) najmniejszą odległość (349,291) mają województwa wielkopolskie z dolnośląskim, niewiele większą (350,656) ma województwo lubuskie z zachodniopomorskim. Natomiast najbardziej oddalone są od siebie województwa podlaskie z łódzkim (4423,828). Odległość między nimi jest o ponad 500 większa od drugiej największej odległości wyznaczonej przez województwo łódzkie z kujawsko-pomorskim (3918,93). W przypadku odległości wyliczanych przy użyciu metryki Czebyszewa (***Błąd! Nie można odnaleźć źródła odwołania.***) województwa z największą odległością między sobą pozostają takie same, jednak tym razem najmniejsza odległość (10,00) jest między województwem małopolskim oraz świętokrzyskim. Odległość województw, które poprzednio były „najbliżej" siebie jest niewiele większa (10,20) od tych, które teraz mają najmniejszą odległość.

Tabela 32. Macierz odległości Czebyszewa.

	Dolnośląskie	Kujawsko-pomorskie	Lubelskie	Lubuskie	Łódzkie	Małopolskie	Mazowieckie	Opolskie	Podkarpackie	Podlaskie	Pomorskie	Śląskie	Świętokrzyskie	Warmińsko-mazurskie	Wielkopolskie	Zachodniopomorskie
Dolnośląskie	0.000	26.300	25.900	13.800	28.0	19.598	15.668	18.300	21.177	35.500	15.104	22.000	16.138	23.600	10.200	11.000
Kujawsko-pomorskie	26.300	0.000	15.551	17.000	54.3	26.600	19.000	26.000	18.675	23.077	29.200	19.000	20.000	22.300	27.500	15.300
Lubelskie	25.900	15.551	0.000	16.793	53.9	26.200	13.800	19.000	12.300	14.000	28.800	25.000	16.200	21.900	27.100	14.900
Lubuskie	13.800	17.000	16.793	0.000	41.8	21.492	17.562	14.000	22.800	21.700	16.997	25.000	17.563	19.000	15.000	12.000
Łódzkie	28.000	54.300	53.900	41.800	0.0	27.700	40.100	46.300	41.600	63.500	25.100	47.100	37.700	32.000	26.800	39.000
Małopolskie	19.598	26.600	26.200	21.492	27.070	0.000	15.233	18.600	14.953	35.800	18.000	25.063	10.000	22.000	13.735	19.295
Mazowieckie	15.668	19.000	13.800	17.562	40.1	15.233	0.000	15.000	16.486	23.400	15.000	24.000	11.903	19.000	13.300	15.365
Opolskie	18.300	26.000	19.000	14.000	46.3	18.600	15.000	0.000	19.131	17.200	22.000	24.000	17.000	22.000	19.500	20.000
Podkarpackie	21.177	18.675	12.300	22.800	41.6	14.953	16.486	19.131	0.000	21.900	19.000	26.371	11.623	15.135	15.043	20.603
Podlaskie	35.500	23.077	14.000	21.700	63.5	35.800	23.400	17.200	21.900	0.000	38.400	27.000	25.800	31.500	36.700	24.500
Pomorskie	15.104	29.200	28.800	16.997	25.010	18.000	15.000	22.000	19.000	38.400	0.000	26.000	13.438	19.200	11.000	14.801
Śląskie	22.000	19.000	25.000	25.000	47.1	25.063	24.000	24.000	26.371	27.000	26.000	0.000	21.134	26.000	24.000	23.000
Świętokrzyskie	16.138	20.000	16.200	17.563	37.7	10.000	11.903	17.000	11.623	25.800	13.438	21.134	0.000	22.000	10.900	15.366
Warmińsko-mazurskie	23.600	22.300	21.900	19.000	32.0	22.000	19.000	22.000	15.135	31.500	19.200	26.000	22.000	0.000	13.400	15.173
Wielkopolskie	10.200	27.500	27.100	15.000	26.080	13.735	13.300	19.500	15.043	36.700	11.000	24.000	10.900	13.400	0.000	12.200
Zachodniopomorskie	11.000	15.300	14.900	12.000	39.0	19.295	15.365	20.000	20.603	24.500	14.801	23.000	15.366	15.173	12.200	0.000

Zupełnie inne wyniki otrzymujemy wykorzystując metrykę Manhattan (*Tabela 33*). Największa wartość w macierzy odległości jest dla województw śląskiego i podkarpackiego. Natomiast najbliżej siebie znajdują się województwo lubelskie z zachodniopomorski – zajmowały one drugie miejsce w przypadku macierzy kwadratowej odległości euklidesowej.

Tabela 33. Macierz odległości Manhattan.

	Dolnośląskie	Kujawsko-pomorskie	Lubelskie	Lubuskie	Łódzkie	Małopolskie	Mazowieckie	Opolskie	Podkarpackie	Podlaskie	Pomorskie	Śląskie	Świętokrzyskie	Warmińsko-mazurskie	Wielkopolskie	Zachodniopomorskie
Dolnośląskie	0.000	101.940	155.939	78.521	98.647	108.091	102.022	95.600	129.973	122.706	67.139	100.997	112.858	111.729	63.593	67.473
Kujawsko-pomorskie	101.940	0.000	110.677	111.625	139.413	149.246	107.708	113.549	129.506	94.116	106.656	111.039	122.423	100.296	99.994	93.678
Lubelskie	155.939	110.677	0.000	110.205	135.598	121.268	105.955	148.253	86.782	109.188	135.032	134.996	91.228	107.888	108.747	110.677
Lubuskie	78.521	111.625	110.205	0.000	108.471	99.956	68.061	99.648	116.501	112.674	101.921	94.535	81.562	109.268	70.830	58.830
Łódzkie	98.647	139.413	135.598	108.471	0.000	116.530	116.699	156.716	134.181	120.771	111.040	130.279	117.749	114.046	65.325	98.916
Małopolskie	108.091	149.246	121.268	99.956	116.530	0.000	81.210	117.605	96.325	132.162	109.974	153.420	64.979	119.937	79.115	107.228
Mazowieckie	102.022	107.708	105.955	68.061	116.699	81.210	0.000	124.889	93.099	115.878	91.691	121.730	77.947	122.778	70.647	86.291
Opolskie	95.600	113.549	148.253	99.648	156.716	117.605	124.889	0.000	108.529	118.047	135.987	133.711	122.887	145.369	115.618	91.243
Podkarpackie	129.973	129.506	86.782	116.501	134.181	96.325	93.099	108.529	0.000	104.904	125.711	157.389	86.032	127.065	104.861	105.136
Podlaskie	122.706	94.116	109.188	112.674	120.771	132.162	115.878	118.047	104.904	0.000	138.504	137.606	123.488	103.179	104.805	107.179
Pomorskie	67.139	106.656	135.032	101.921	111.040	109.974	91.691	135.987	125.711	138.504	0.000	115.919	109.679	108.967	71.157	78.471
Śląskie	100.997	111.039	134.996	94.535	130.279	153.420	121.730	133.711	157.389	137.606	115.919	0.000	128.904	128.747	104.870	87.432
Świętokrzyskie	112.858	122.423	91.228	81.562	117.749	64.979	77.947	122.887	86.032	123.488	109.679	128.904	0.000	111.659	81.946	92.088
Warmińsko-mazurskie	111.729	100.296	107.888	109.268	114.046	119.937	122.778	145.369	127.065	103.179	108.967	128.747	111.659	0.000	84.352	92.348
Wielkopolskie	63.593	99.994	108.747	70.830	65.325	79.115	70.647	115.618	104.861	104.805	71.157	104.870	81.946	84.352	0.000	69.016
Zachodniopomorskie	67.473	93.678	110.677	58.830	98.916	107.228	86.291	91.243	105.136	107.179	78.471	87.432	92.088	92.348	69.016	0.000

Jedną z metod mierzenia dopasowania dendrogramu do macierzy odległości jest współczynnik korelacji kofenetycznej. Opiera się on na obliczeniu współczynnika korelacji według momentu iloczynowego między odpowiadającymi sobie odległościami w macierzy odległości D oraz macierzy kofenetycznej C. Ta druga, zwana też macierzą dendrogramu zawiera poziomy łączenia, na których para obiektów łączy się w tym samym skupieniu po raz pierwszy. Wysoka wartość korelacji sugeruje, że dendrogram dobrze odzwierciedla różnice pomiędzy obiektami.

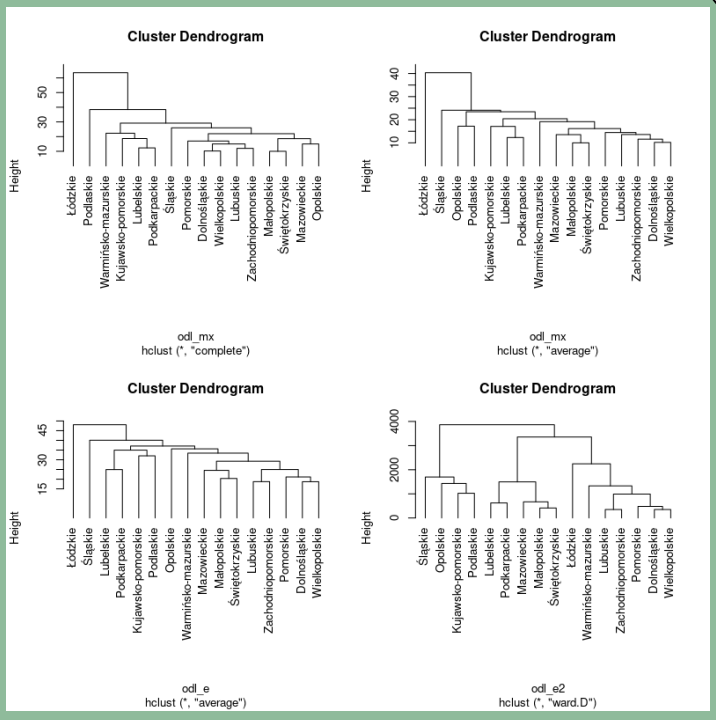
W poniższej tabeli przedstawiono wyniki współczynnika korelacji kofenetycznej dla wszystkich możliwych kombinacji metod oraz metryk opisanych w wprowadzeniu.

Tabela 34. Wartości współczynnika korelacji kofenetycznej dla różnych metryk oraz metod mierzenia odległości.

Metryka	Metoda	Współczynnik korelacji kofenetycznej
Kwadratowa odległość euklidesowa	Warda	0,461
	Najbliższego sąsiada	0,644
	Najdalszego sąsiada	0,651
	Średniej grupowej	0,733
Euklidesowa	Najbliższego sąsiada	0,711
	Najdalszego sąsiada	0,659
	Średniej grupowej	0,774
Manhattan	Najbliższego sąsiada	0,611
	Najdalszego sąsiada	0,489
	Średniej grupowej	0,703
Czebyszewa	Najbliższego sąsiada	0,755
	Najdalszego sąsiada	0,801
	Średniej grupowej	0,798
Canberra	Najbliższego sąsiada	0,667
	Najdalszego sąsiada	0,597
	Średniej grupowej	0,728

Najlepszy wynik został oznaczony ciemniejszym zielonym a dwa pozostałe z podium oraz wynik otrzymany za pomocą metody Warda jaśniejszym zielonym. Postanowiono przedstawić te cztery wyniki na dendrogramach:

- Wynik otrzymany za pomocą metryki Czebyszewa oraz metody najdalszego sąsiada (0,801),
- Wynik otrzymany za pomocą metryki Czebyszewa oraz metody średniej grupowej (0,798),
- Wynik otrzymany za pomocą metryki euklidesowej oraz metody średniej grupowej (0,774),
- Wynik otrzymany za pomocą kwadratowej odległości euklidesowej oraz metody Warda (0,461)

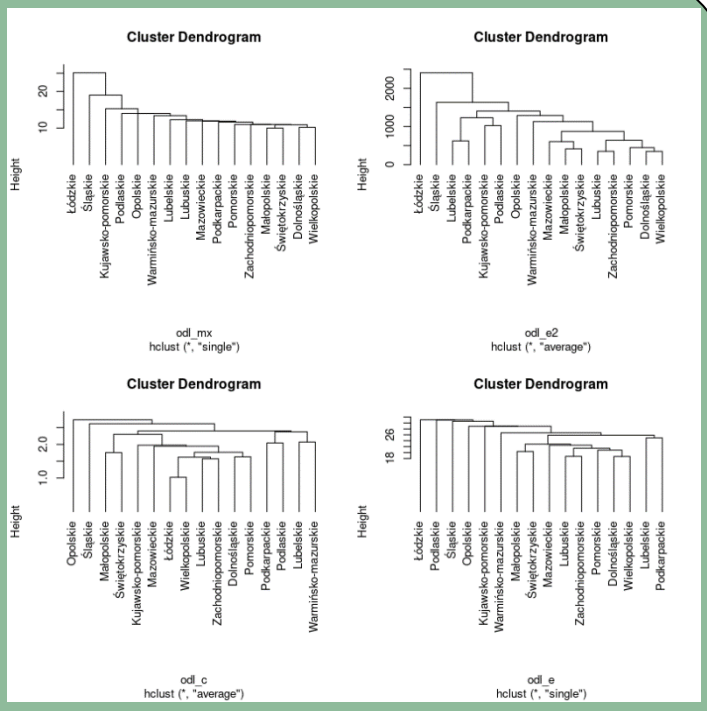


Wykres 33. Dendrogramy dla 3 najlepszych wyników korelacji oraz dla metody Warda. Źródło: opracowanie własne w programie RStudio.

Na podstawie dendrogramów (*Wykres 33.* Dendrogramy dla 3 najlepszych wyników korelacji oraz dla metody Warda. Źródło: opracowanie własne w programie RStudio.) można zauważyć, iż pomimo wysokich wartości współczynnika korelacji w przypadku trzech pierwszych wyników nie uzyskano struktury grupowej. Te drzewka połączeń nie są dobrymi podsumowaniami relacji, ponieważ tworzą się na nich łańcuchy. Jedynie w przypadku metody Warda (ostatni dendrogram), pomimo niskiego współczynnika korelacji kofenetycznej udało się uzyskać strukturę grupową. Można zaobserwować, że w przypadku trzech pierwszych metod województwo łódzkie dopiero na ostatnim etapie tworzy skupienie, odstaje ono od reszty. W przypadku drugiego oraz trzeciego dendrogramu województwo śląskie łączy się jako przedostatnie, natomiast na pierwszym dendrogramie jest to województwo podlaskie. W przypadku wszystkich metod na wczesnych etapach łączą się w skupiska województwo dolnośląskie z wielkopolskim, lubuskie z zachodniopomorskim oraz lubelskie z podkarpackim.

W związku z tym zdecydowano przyjrzeć się jeszcze czterem kolejnym wynikom z najwyższą korelacją:

- Wynik otrzymany za pomocą metryki Czebyszewa oraz metody najbliższego sąsiada (0,755),
- Wynik otrzymany za pomocą kwadratowej odległości euklidesowej oraz metody średniej grupowej (0,733),
- Wynik otrzymany za pomocą metryki Canberra oraz metody średniej grupowej (0,728),
- Wynik otrzymany za pomocą metryki euklidesowej oraz metody najbliższego sąsiada (0,711)



Wykres 34. Dendrogramy dla kolejnych 4 najlepszych wyników korelacji. Źródło: opracowanie własne w programie RStudio.

W przypadku czterech kolejnych wyników ponownie nie udało utworzyć się struktury grupowej. Najlepsza sytuacja jest przy zastosowaniu kwadratowej odległości Euklidesa z metodą średniej grupowej. Jednak najlepszą strukturę udało się uzyskać nadal metodą Warda. Najbardziej widoczną strukturę łańcuchową można natomiast zaobserwować na pierwszym oraz ostatnim dendrogramie wykorzystującym metodę najbliższego sąsiada. Ponownie województwo łódzkie często jako ostatnie dołącza do skupienia. Wyjątek stanowi jedynie trzeci dendrogram, gdzie zastosowano metrykę Canberra z metodą średniej grupowej, tam jako ostatnie grupuje się województwo opolskie.

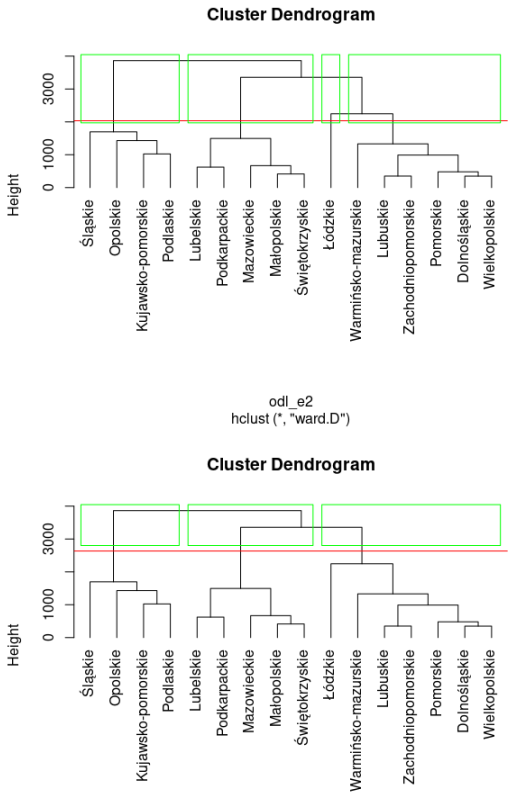
Pomimo niskiego współczynnika korelacji podjęto decyzję, aby opierać się na strukturach ukazanych na dendrogramach i wybrać wynik otrzymany metodą Warda do dalszej analizy.

Następnym krokiem po wybraniu metryki oraz metody mierzenia odległości jest analiza dendrogramu oraz wybór liczby skupień. Jest to ważny element procesu grupowania. W trakcie każdego etapu hierarchicznego grupowania, określana jest ilość wyróżnionych skupień, liczba ta stopniowo maleje w miarę wzrostu poziomu hierarchii. Kluczowe jest ustalenie reguły przerywania grupowania, która ułatwi wybór jednego z podziałów. Jedną z reguł jest kryterium zaproponowane przez R.Mojene bazujące na względnej wysokości różnych poziomów łączenia. Polega ono na wyborze liczby skupień odpowiadającej temu poziomowi dendrogramu, dla którego spełniona jest nierówność:

$$h_{n-k-i} > \bar{h} + a \cdot \hat{s}_h$$

Gdzie:

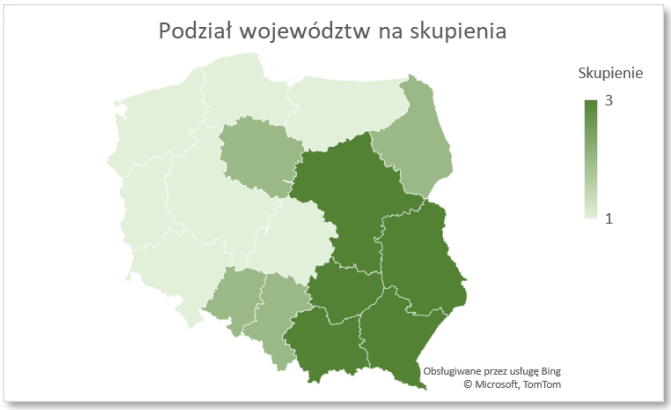
- $h_1, h_2, \ldots, h_{n-1}$- poziomy łączenia odpowiadające etapom z n,n-1,...,1 skupieniami,
- $\bar{h}$ - średnia z wartości $h_1, h_2, \ldots, h_{n-1}$,
- $\hat{s}_h$ - odchylenie standardowe w zbiorze wartości $h_1, h_2, \ldots, h_{n-1}$,
- a – stała



Wykres 35. Zastosowanie kryterium Mojeny na dendrogramach.
Źródło: opracowanie własne w programie RStudio.

Badacze sugerują różne wartości stałej a. Postanowiono sprawdzić, w którym miejscu zostanie przerwane grupowanie przy zastosowaniu stałej $a = 0{,}7$ oraz $a = 1{,}25$.
Za pomocą stałej równej 0,7 otrzymano cztery grupy w tym jedną zawierającą jednostkę izolowaną natomiast za pomocą wyższej wartości stałej otrzymano trzy skupiska. Z racji na niewielką odległość linii odcięcia przy bardziej rygorystycznej metodzie od przyłączenia województwa łódzkiego do skupiska zdecydowano się wybrać podział na 3 klasy.
Jako pierwsze skupienie utworzyły województwo dolnośląskie z województwem wielkopolskim, co zgadza się z macierzą kwadratowej odległości euklidesowej przedstawionej na poprzednich stronach. Szybko również połączyły się województwa lubuskie z zachodniopomorskim oraz małopolskie z świętokrzyskim. Jako ostatnie z jednoelementowego skupienia łączy się województwo łódzkie.

Północnozachodnie województwa utworzyły pierwsze skupienie, natomiast południowo-wschodnie skupienie trzecie. Drugie skupienie tworzą województwa kujawsko-pomorskie, podlaskie, śląskie oraz opolskie.



Wykres 36. Podział województwa na skupienia.
Źródło: opracowanie własne w programie Excel.

Wykres 37. Profile skupień.
Źródło: opracowanie własne w programie RStudio.

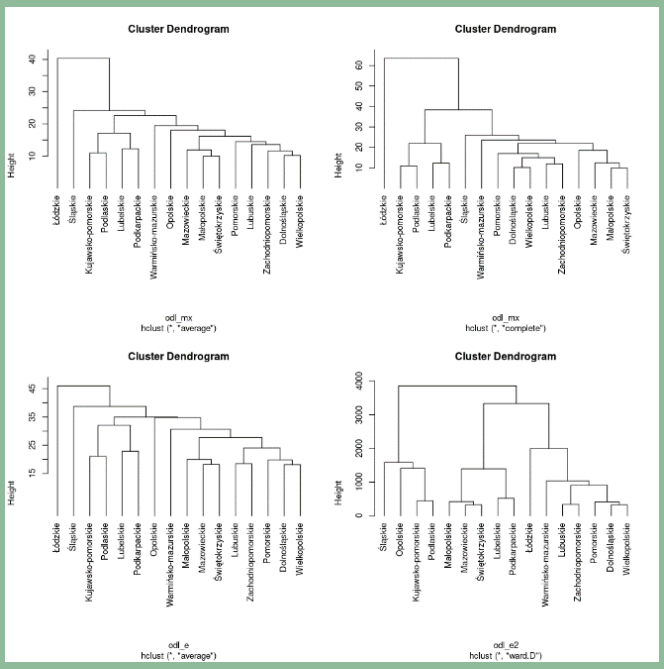
Dla utworzonych skupień obliczono średnie wartości każdej zmiennej, aby porównać ich profile. Na podstawie wykresu można zauważyć, że niektóre średnie mniej lub bardziej różnią się

między skupieniami. Profile zmiennych x19, x25, x5 oraz x6 są do siebie zbliżone. W związku z tym, że cechy te nie różnicują województw zdecydowano wyeliminować je z analizy.
Po wyeliminowaniu wcześniej wymienionych zmiennym ponownie obliczono współczynniki korelacji kofenetycznej.

Tabela 35. Wartości współczynnika korelacji kofenetycznej dla różnych metryk oraz metod mierzenia odległości po wyeliminowaniu 4 zmiennych.

Metryka	Metoda	Współczynnik korelacji kofenetycznej
Kwadratowa odległość euklidesowa	Warda	0,466
	Najbliższego sąsiada	0,578
	Najdalszego sąsiada	0,375
	Średniej grupowej	0,724
Euklidesowa	Najbliższego sąsiada	0,648
	Najdalszego sąsiada	0,436
	Średniej grupowej	0,764
Manhattan	Najbliższego sąsiada	0,601
	Najdalszego sąsiada	0,620
	Średniej grupowej	0,696
Czebyszewa	Najbliższego sąsiada	0,757
	Najdalszego sąsiada	0,764
	Średniej grupowej	0,802
Canberra	Najbliższego sąsiada	0,667
	Najdalszego sąsiada	0,380
	Średniej grupowej	0,727

Zastosowano te same oznaczenia kolorystyczne co poprzednio.
Do dalszej analizy wybrano wyniki z trzema najwyższymi korelacjami oraz wynik otrzymany metodą Warda:



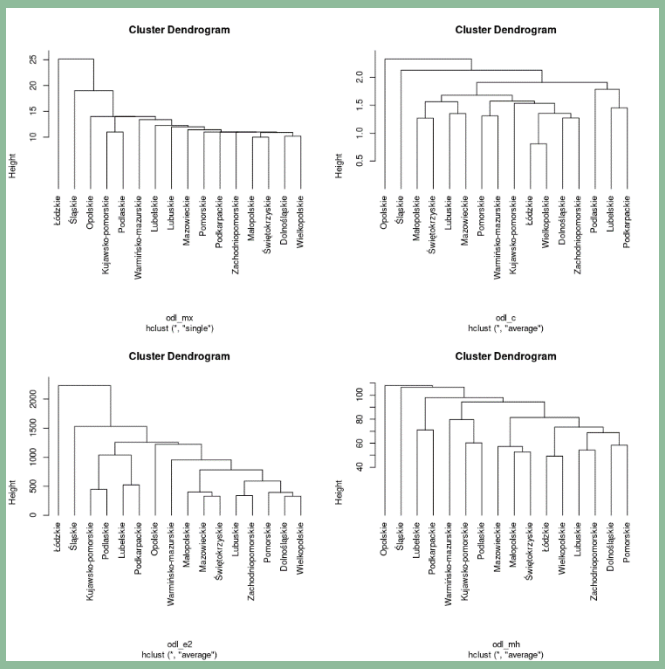
Wykres 38. Dendrogramy dla 3 najlepszych wyników korelacji oraz dla metody Warda po usunięciu zmiennych.
Źródło: opracowanie własne w programie RStudio.

- Wynik otrzymany za pomocą metryki Czebyszewa oraz metody średniej grupowej (0,802),
- Wynik otrzymany za pomocą metryki Czebyszewa oraz metody najdalszego sąsiada (0,764),
- Wynik otrzymany za pomocą metryki euklidesowej oraz metody średniej grupowej (0,764),
- Wynik otrzymany za pomocą kwadratowej odległości euklidesowej oraz metody Warda (0,466).

Sytuacja na dendrogramach jest podobna jak w przypadku analizy z zastosowaniem wszystkich zmiennych. Ponownie najlepsza strukturę grupową udało się uzyskać za pomocą Warda, mimo niższego współczynnika korelacji kofenetycznej.

Znowu województwo łódzkie tworzy jako ostatnie skupienie.
Analogicznie jak wcześniej przeanalizowano cztery kolejne wyniki z najwyższym współczynnikiem korelacji:

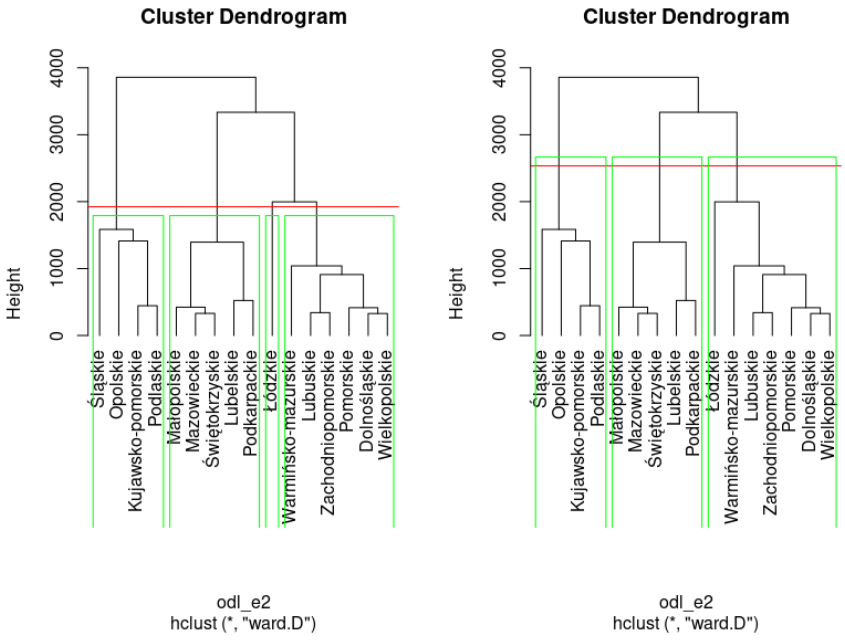
- Wynik otrzymany za pomocą metryki Czebyszewa oraz metody najbliższego sąsiada (0,757),
- Wynik otrzymany za pomocą metryki Canberra oraz metody średniej grupowej (0,727),
- Wynik otrzymany za pomocą kwadratowej odległości euklidesowej oraz metody średniej grupowej (0,724),
- Wynik otrzymany za pomocą metryki Manhattan oraz metody średniej grupowej (0,696)



Wykres 39. Dendrogramy dla kolejnych 4 najlepszych wyników korelacji po usunięciu zmiennych.
Źródło: opracowanie własne w programie RStudio.

Sytuacja jest zbliżona do tej przed usunięciem zmiennych. Ponownie można dostrzec strukturę łańcuchową na dendrogramach. Jest ona szczególnie widoczna na pierwszym – wykorzystującym metodę najbliższego sąsiada. Metoda średniej grupowej z odległością Euklidesa (trzecie drzewko) ma najlepsza strukturę z tych przedstawionych na wykresie, ale nadal gorszą niż Ward. Do dalszej analizy ponownie zostanie wykorzystana metoda Warda.

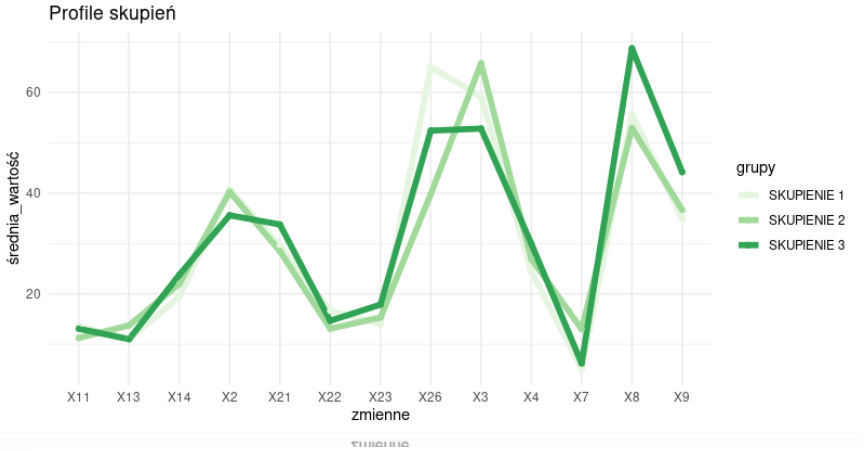
Następnie zostały porównane podziały otrzymane za pomocą różnych wartości stałej w kryterium Mojeny. Przy użyciu stałej równej 0,7 województwo łódzkie tworzy oddzielne skupisko. Ponownie postanowiono wybrać wyższą wartość stałej, której użycie tworzy trzy skupiska. Pomimo wyeliminowania trzech zmiennych przynależność województw do skupień pozostała bez zmian. Aktualna pozostaje mapa przedstawiona na *Wykres 36.*



Wykres 40. Zastosowanie kryterium Mojeny na dendrogramach.
Źródło: opracowanie własne w programie RStudio.

Tabela 36. Średnie wartości dla skupień.

	skupienie_1	skupienie_2	skupienie_3	średnie
x2	40.714	40.250	35.600	39.000
x3	59.118	65.750	52.800	58.802
x4	24.423	26.941	29.940	26.776
x7	5.143	13.000	6.200	7.438
x8	55.619	52.889	68.733	59.034
x9	34.909	36.667	44.164	38.241
x11	13.557	11.250	13.080	12.831
x13	11.000	13.750	11.000	11.688
x14	19.429	22.000	23.800	21.438
x21	29.571	28.475	33.760	30.606
x22	16.643	13.100	14.660	15.137
x23	13.981	15.316	17.874	15.531
x26	65.071	39.800	52.400	54.794



Wykres 41. Profile skupień po wyeliminowaniu 4 zmiennych. Źródło: opracowanie własne w programie RStudio.

Średnie wartości zmiennych dla profili zostały przedstawione na wykresie oraz w tabeli. Na podstawie wykresu można zaobserwować, że największe różnice między skupieniami są widoczne w przypadku zmiennych *x3,x8* oraz *x26*. Dodatkowo w tabeli zostały uwzględnione średnie wartości cechy. Biorąc pod uwagę charakter zmiennych, kolorem czerwonym zostały oznaczone najgorsze wartości spośród wszystkich skupień (dla stymulant najniższe, dla destymulant najwyższe), natomiast kolorem zielonym najlepsze. Można zaobserwować, że pierwsze skupienie charakteryzuje się największą ilością najgorszych wartości cech. Zmienne *x2,x3,x4,x7* odpowiadają za kwestie związane ze zdrowiem. Pierwsza grupa województw ma najgorsze średnie wartości wszystkich tych zmiennych za wyjątkiem *x3* – umieralności okołoporodowej. Charakteryzuje się ona również największą liczbą przestępstw przeciwko rodzinie i opiece (*x11*) oraz największą liczbą wypadków samochodowych na 100tys. ludności (*x26*). Wartość tej drugiej zmiennej w pierwszym skupieniu jest o ponad 10 większa od średniej dla wszystkich województw. Dodatkowo województwa tworzące to skupienie charakteryzują się słabym dostępem do miejsc kultury (*13,x14*). Jedynie wskaźnik zagrożenia ubóstwem (*x23*) oraz emisji zanieczyszczeń (*x7*) jest korzystniejszy niż w przypadku pozostałych grup. Województwa z pierwszego skupienia można, więc uznać za najgorsze do życia rodzin z dziećmi. W przypadku dwóch kolejnych skupień sytuacja nie jest już tak jednoznaczna. Zarówno drugie jak i trzecie ugrupowanie posiada podobną liczbę najlepszych średnich wartości. W przypadku drugiego skupienia wyróżnia się ono małą liczbą wypadków drogowych i przestępstw oraz największą liczbą domów kultury. Ponadto województwa z drugiej grupy mają średnio niższy poziom problemów finansowych (*x21,x22*). Negatywną stroną tych regionów jest wysoka umieralność okołoporodowa, duża emisja zanieczyszczeń pyłowych (*x7*) oraz średnio najmniejsza liczba placówek przedszkolnych (*x8*). Ostatnie skupienie wyróżnia się natomiast najlepszą dostępnością przedszkoli oraz szkół podstawowych. Średnie wartości dla tego skupienia są znacznie wyższe niż średnia dla wszystkich województw i znaczącą różnią się od najgorszej wartości dla skupień. Dodatkowo cechuje się najlepszą opieką zdrowotną oraz największą ilościom bibliotek. Województwa z tej grupy średnio najsłabiej wypadają w kwestiach finansowych oraz dostępności domów kultury. Z uwagi na lepszą opiekę zdrowotną, która jest kluczowa szczególnie w pierwszych latach oraz lepszą dostępność placówek przedszkolnych oraz szkół podstawowych województwa z trzeciego skupienia można uznać za najlepsze do założenia rodziny. W późniejszych latach życia dziecka również dobrym wyborem byłyby

województwa z drugiego skupienia cechujące się bezpieczeństwem oraz uśredniając lepszą sytuacją finansową osób w nich mieszkających.
Dodatkowo można zauważyć, że wszystkie województwa, które charakteryzowały się niskim poziomem wydatków na oświatę, edukację i rodziny trafiły do pierwszego skupienia, charakteryzującego się słabą dostępnościom miejsc kultury oraz szkół podstawowych. Jednak w skupieniu tym znalazły się też województwa o średnim i wysokim poziomie wydatków. W przypadku drugiej grupy województw większość regionów posiada wysoki poziom wydatków a przypadku trzeciego średni. Województwo podkarpackie, które zostało ocenione jako najlepsze województwo do założenia rodziny znalazło się w ostatnim skupieniu – wyróżniającym się największą liczbą najlepszych średnich wartości cech. Dwa kolejne województwa z podium wchodzą w skład skupienia drugiego. Województwo, które zostało uznane za najgorsze dla rodzin również tworzy to skupienie. Charakteryzowało się ono najwyższą wartością zanieczyszczeń co jest widoczne przy średniej dla skupienia.

Tabela 37. Porównanie grup utworzonych przez skupienia oraz poziom wydatków.

Województwo	Skupienie	Wydatki
Dolnośląskie	1	średnie
Lubuskie	1	niskie
Łódzkie	1	średnie
Pomorskie	1	niskie
Warmińsko-mazurskie	1	niskie
Wielkopolskie	1	wysokie
Zachodniopomorskie	1	niskie
Kujawsko-pomorskie	2	wysokie
Opolskie	2	wysokie
Podlaskie	2	średnie
Śląskie	2	wysokie
Lubelskie	3	średnie
Małopolskie	3	średnie
Mazowieckie	3	wysokie
Podkarpackie	3	średnie
Świętokrzyskie	3	średnie

6. Podsumowanie

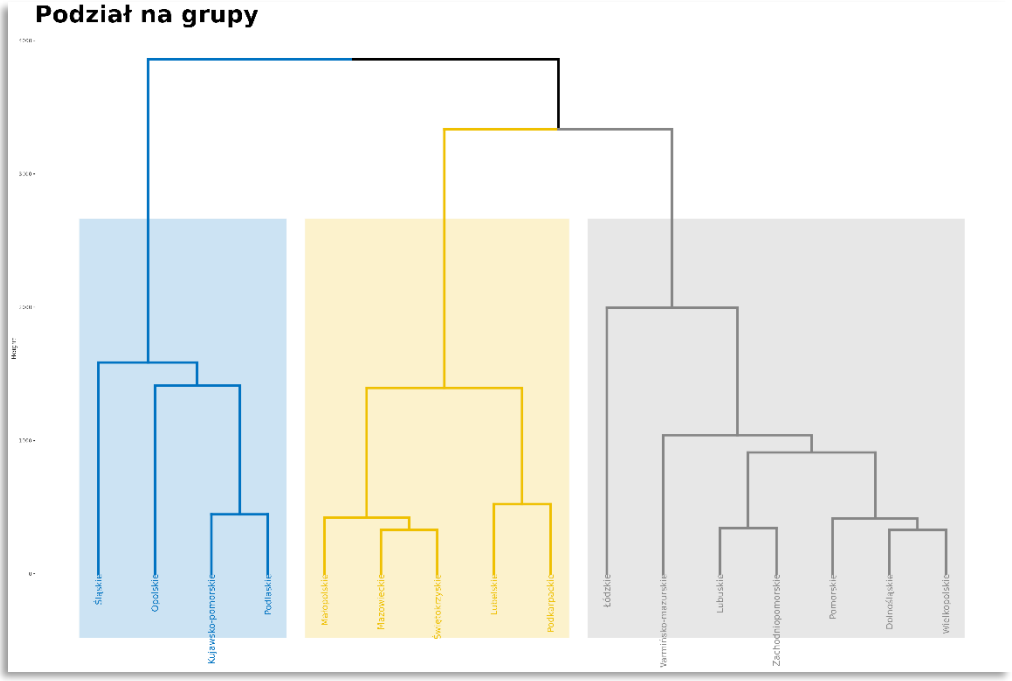
Celem pracy było znalezienie najlepszego województwa do założenia rodziny. Aby to dokonać do analizy zostały wybrane zmienne odpowiedzialne za podstawowe potrzeby dziecka związane z opieką zdrowotną, edukacją, kulturą oraz uwzględnione czynniki ważne dla przyszłych rodziców przed podjęciem decyzji o powiększeniu rodziny takie jak stabilizacja finansowa oraz warunki mieszkalne w danym rejonie czy poziom bezpieczeństwa.

Jako początkowy etap została wykonana analiza wstępna uwzględniająca badanie statystyk, korelacji oraz rozkładów zmiennych. Zauważono, że najwyższe korelacje występujące w zbiorze są pozorne. Podczas tej części obiekty zostały również podzielone na kategorie w zależności od tego jaki procent wydatków przeznaczają na oświatę, edukację oraz rodzinę. Udało się zaobserwować zależność, iż województwa z średnim poziomem wydatków posiadają średnio więcej placówek przedszkolnych oraz szkół podstawowych na 100 tys. ludności a wraz ze wzrostem poziomu wydatków maleje średnia liczba zgonów niemowląt.

Następnie przed przystąpieniem do analizy głównych składowych zostały usunięte zmienne o niskim współczynniku zmienności oraz stopniowo zostały wyeliminowane zmienne wykazujące niską korelacje z innymi. Jako wynik udało się otrzymać trzy główne składowe wyjaśniające ponad 70 % wariancji zbioru. Pierwsza składowa została oznaczona jako miara edukacyjno-finansowa, natomiast druga jako miara problemów finansowych. Odpowiadają, więc one na najważniejsze obawy związane z założeniem dzieci – stabilizację finansową oraz dostęp do edukacji. Na podstawie wykresów udało się zauważyć, że województwo opolskie charakteryzuje się wyraźnie niższą wartością drugiej składowej w stosunku do pozostałych regionów – wyróżnia się liczbą bibliotek i miejsc kultury. Można było również zauważyć, że zmienne związane z edukacją mają dodatnią korelację z pierwszą składową oraz ujemną korelacje ze drugą składową (za wyjątkiem *x8*). Natomiast zmienne związane z finansami oraz zdrowiem korelują dodatnio z drugą składową. Trzecia składowa, nazwana miara opieki, okazała się cięższa w interpretacji.

Kolejny rozdział pracy został poświęcony porządkowaniu liniowemu. Celem tej analizy było odpowiedzenie na pytanie jakie województwo jest najlepsze do założenia rodziny. Ocena regionów została przeprowadzona z zastosowaniem różnych metod dających bardziej lub mniej zbliżone wyniki. W wyniku analizy udało się wskazać jednoznacznie lidera (podkarpackie) oraz trzy kolejne w kolejności województwa. Wspólną cechą pierwszej czwórki rankingu było uzyskanie najlepszego rezultatu w danej grupie zmiennych. W przypadku województwa podkarpackiego był to dostęp do edukacji, opolskiego – do kultury, podkarpackiego opieki zdrowotnej natomiast małopolskie okazało się najbardziej bezpiecznym województwem. Porządkowanie wskazało również jednoznacznie najgorsze województwo do założenia rodziny – śląskie. W niektórych metodach wyraźnie odstawało ono od innych. Jego największą wadą jest wysoki wskaźnik zanieczyszczeń pyłowych. Słabo również wypadły województwa kujawsko-pomorskie oraz warmińsko-mazurskie, uplasowały się one na 14. oraz 15. miejscu.

Ostatnia część została poświęcona analizie skupień. Jej zadaniem było znalezienie województw podobnych do siebie pod pewnymi względem. Przed ostatecznym pogrupowaniem województw dokonano analizy współczynnika korelacji kofenetycznej oraz dendrogramów w celu wyboru najlepszej metryki oraz metody liczenia odległości i wyeliminowano zmienne słabo różnicujące zbiór. Okazało się, że mimo wysokich współczynników wyniki nie miały struktury grupowej. W związku z tym postanowiono opierać się na metodzie Warda. Udało się uzyskać trzy skupienia województw.



Wykres 42. Otrzymany podział województw na skupienia.
Źródło: opracowanie własne w programie RStudio.

Pierwsze od lewej skupienie charakteryzuje się wysokim poziomem bezpieczeństwa oraz dobrą sytuacją finansową mieszkańców. Jego wadą jest najwyższy wskaźnik zanieczyszczeń pyłowych. Drugie od lewej skupienie zawierające województwo podkarpackie wypadło najlepiej. Województwa w nim zawarte mają średnio najlepszą opiekę zdrowotną oraz najwięcej placówek przedszkolnych i szkół podstawowych. Grupa ostatnich województw wypadła najgorzej. Średnie wskaźniki bezpieczeństwa jak również placówek kultury osiągnęły w niej najgorsze wyniki.

W skutek przeprowadzonych analiz udało się odpowiedzieć na tytułowe pytanie. Najbardziej rodzinnym województwem okazało się województwo podkarpackie, najgorzej wypadło województwo śląskie. Analiza głównych składowych pozwoliła stworzyć miary odpowiadające najważniejszym potrzebom dziecka oraz obawom przyszłych rodziców. Natomiast analiza skupień wskazała grupy województw podobnych do siebie.

7. Spis tabel

8. Spis wykresów

9. Bibliografia

1. „Narodowy Spis Powszechny Ludności i Mieszkań 2021. Starzenie się ludności Polski w świetle wyników narodowego spisu powszechnego ludności i mieszkań 2021" Główny Urząd Statystyczny, Warszawa, 2023r.
2. „World Population Ageing 2020 Highlights" United Nations, Departament of Economic and Social Affairs, 2020r.
3. „Wyzwania demograficzne Europy i Polski" Marek Okólski, Szkoła Wyższa Psychologii Społecznej, 2010r.
4. „Czy zwiększenie dzietności w Polsce jest możliwe?" Małgorzata Sikorska, Instytut badań strukturalnych, 2021r.
5. „Skutki „świadczenia 500+"", Izabela Bień, Biuro Analiz Sejmowych, 2022r.
6. „Dochody i warunki życia ludności Polski – raport z badania EU-SILC 2021" Główny Urząd Statystyczny, Warszawa, 2023r.
7. „Demografia Liczba, rozmieszczenie i struktura ludności" Materiały dydaktyczne, Opracowano na podst. J. Holzer, Demografia, Warszawa 2003
8. „Warunki życia rodzin w Polsce" Główny Urząd Statystyczny, Warszawa, 2014r.
9. „Starzenie się społeczeństwa polskiego i jego skutki" Biuro Analiz I Dokumentacji, Zespół Analiz i Opracowań Tematycznych, 2011r.
10. „Regionalne zróżnicowanie procesu starzenia się ludności Polski w latach 1990-2015 oraz w perspektywie do 2040 roku" Joanna Stańczak, Dorota Szałtys
11. „Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne" Andrzej Balicki, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk, 2009r.
12. „Detailed study of Principal Component Analysis" (Chapter 4 Detailed study of Principal Component Analysis | A Machine Learning Compilation (f0nzie.github.io))