



Gdzie najlepiej urodzić dziecko?

Analiza wielowymiarowa województw na podstawie danych z 2023 roku

SPIS TREŚCI

1. Wprowadzenie
2. Analiza wstępna
3. Analiza kanoniczna
4. Analiza głównych składowych (PCA)
5. Porządkowanie liniowe
6. Analiza skupień
7. Podsumowanie
8. Bibliografia



Wprowadzenie

W Polsce, podobnie jak w wielu innych krajach europejskich, liczba urodzeń sukcesywnie spada, co budzi obawy o przyszłość rynku pracy i systemu zabezpieczeń społecznych. Rządy poszczególnych państw wdrażają różnorodne strategie prorodzinne, obejmujące świadczenia finansowe, ułatwienia w dostępie do opieki nad dziećmi oraz wsparcie dla młodych rodziców na rynku pracy. Ich skuteczność jest jednak zróżnicowana, a decyzja o posiadaniu dzieci pozostaje w dużej mierze uwarunkowana czynnikami ekonomicznymi, społecznymi i kulturowymi. Kluczowe znaczenie ma m.in. jakość opieki medycznej, dostępność infrastruktury wspierającej rodziny, bezpieczeństwo publiczne oraz koszty życia w poszczególnych regionach.

Niniejszy raport stanowi próbę wielowymiarowej analizy warunków narodzin i wychowania dziecka w różnych województwach Polski. Badanie opiera się na danych Głównego Urzędu Statystycznego za rok 2023 i uwzględnia szeroki zestaw zmiennych. Aby uzyskać kompleksowy obraz warunków sprzyjających rodzicielstwu, uwzględniono następujące obszary:

- **Dostępność opieki zdrowotnej** – określająca jakość i dostępność podstawowej oraz specjalistycznej opieki medycznej dla noworodków i matek.
- **Infrastruktura szpitalna** – oceniająca jakość opieki świadczonej matkom i dzieciom w okresie okołoporodowym i wczesnym dzieciństwie.
- **Usługi społeczne** – wskazujące na możliwości wsparcia rodzin w opiece nad najmłodszymi dziećmi.
- **Transport publiczny i infrastruktura miejska** – wpływające na komfort życia rodzin oraz codzienną mobilność.
- **Koszty życia** – istotne dla młodych rodziców przy wyborze miejsca zamieszkania.
- **Bezpieczeństwo** – mające kluczowe znaczenie dla stabilności i komfortu życia rodzin.
- **Edukacja i kultura** – określające dostępność usług edukacyjnych i kulturalnych wspierających rozwój dziecka.

Przeprowadzona analiza pozwala wskazać województwa o najlepszych warunkach do narodzin i ukazać stopień zróżnicowania tych warunków pomiędzy poszczególnymi regionami Polski. Wnioski z raportu mogą stanowić cenne źródło informacji zarówno dla rodzin planujących powiększenie rodziny, jak i dla instytucji publicznych odpowiedzialnych za kształtowanie polityki prorodzinnej i rozwoju regionalnego. W kontekście wyzwań demograficznych Europy poprawa warunków analizowanych czynników mieć istotne znaczenie w przeciwdziałaniu spadkowi urodzeń i długofalowym skutkom starzenia się społeczeństwa.

Analiza wstępna

Zmienne zostały podzielone na 4 kategorie:

- Zdrowie i opieka medyczna: X1, X2, X5, X7, X8, X9, X12, X18
- Infrastruktura i dostępność usług X3, X6, X13, X16, X17
- Społeczeństwo i demografia X10, X11, X15,
- Ekonomia i wydatki X4, X14



Tabela 1: Przedstawiająca podstawowe statystyki

Zmienna	X1	X2	X3	X4	X5	X6	X7	X8	X9
Średnia	1,83	14,06	6,31	3,76	2,44	2,52	13,63	3,04	5,66
Odchylenie standardowe	0,47	2,6	2,97	0,95	0,53	0,52	1,36	0,44	0,7
Skośność	0,33	0,32	0,75	1,14	0,13	0,69	0,6	0,77	-0,06
Wartość minimum	1,13	9,75	2,66	2,77	1,63	1,93	12,06	2,46	4,41
Wartość maximum	2,79	19,02	11,96	6,07	3,31	3,53	16,03	4,08	6,9
Współczynnik zmienności	25,41	18,46	47,03	25,28	21,8	20,52	9,97	14,51	12,34
Kwartył 1	1,35	11,41	3,1	2,96	1,84	1,95	12,27	2,61	4,74
Kwartył 3	2,13	15,38	6,86	3,97	2,88	2,79	14,46	3,31	6,11
Gini	13,88	10,07	24,44	12,78	12,03	10,89	5,29	7,7	6,72

Zmienna	X10	X11	X12	X13	X14	X15	X16	X17	X18
Średnia	2,02	2,86	10,23	11,71	1,6	2,49	8,36	2,56	12,12
Odchylenie standardowe	0,42	1,34	2,06	4,2	0,18	0,47	0,86	1,55	4,2
Skośność	-0,1	0,36	-0,04	0,6	-0,21	0,45	0,42	1,31	0,82
Wartość minimum	1,16	0,92	6,81	5,61	1,32	1,77	7,17	1	4,05
Wartość maximum	2,75	5,27	13,44	19,75	1,85	3,53	9,94	7	23,34
Współczynnik zmienności	20,52	46,81	20,16	35,87	11,19	18,91	10,29	60,4	34,65
Kwartył 1	1,64	1,44	7,65	7,8	1,33	1,94	7,4	1	9,06
Kwartył 3	2,22	3,88	11,64	13,27	1,72	2,87	8,75	3	13,52
Gini	11,13	25,43	11,17	19,23	6,15	10,17	5,61	29,73	17,22

Źródło: Opracowanie własne. Obliczenia wykonane w programie R, prezentacja danych w programie Excel

Tabela 2: Macierz początkowa

Województwo	X1	X2	X3	X4	X5	X6	X7	X8	X9
dolnośląskie	2,27	14,48	6,09	4,2	2,71	2,33	13,05	2,99	4,41
kujawsko-pomorskie	2,04	13,18	6,62	3,01	2,15	2,45	13,12	2,65	4,95
lubelskie	1,53	14,77	5,2	3,84	2,24	2,24	13	3,33	5,78
lubuskie	1,13	11,49	2,66	2,91	2,46	2,36	12,95	2,65	6,58
łódzkie	2,37	15,92	7,02	3,33	2,62	2,84	13,25	3,71	5,34
małopolskie	1,94	14,75	11,24	5,5	2,01	2,77	16,03	2,99	5,49
mazowieckie	2,79	19,02	6,81	6,07	2,11	1,98	15,63	4,08	5,78
opolskie	1,36	11,32	4,94	2,77	3,1	3,1	12,76	2,75	5,37
podkarpackie	1,46	9,75	6,02	3,35	1,98	2,61	14,07	2,46	6,05
podlaskie	2,35	16,43	3,58	3,61	3,07	1,93	14,16	3,22	6,5
pomorskie	2,08	13,69	5,3	4,73	1,7	1,95	15,76	2,87	4,53
śląskie	1,81	18,54	11,96	3,04	3,31	2,38	12,32	3,36	5,4
świętokrzyskie	1,33	11,55	11,71	3,21	2,05	3,51	12,06	2,72	5,66
warmińsko-mazurskie	1,49	12,45	2,99	3,03	3,09	3,53	12,21	2,57	6,9
wielkopolskie	1,91	12,5	5,68	3,9	1,63	1,95	15,35	2,93	5,57
zachodniopomorskie	1,45	15,2	3,21	3,62	2,82	2,33	12,43	3,3	6,31

Województwo	X10	X11	X12	X13	X14	X15	X16	X17	X18
dolnośląskie	2,75	2,16	11,58	9,66	1,44	3,53	7,54	3	9,21
kujawsko-pomorskie	2,17	1,96	6,81	10,92	1,7	2,45	8,72	7	10,39
lubelskie	1,65	1,95	11,81	11,78	1,6	2	9,3	3	13,23
lubuskie	2,53	4,65	8,28	7,79	1,6	2,92	7,38	3	4,05
łódzkie	2,11	3,8	13,4	7,92	1,62	2,27	8,85	3	11,45
małopolskie	2,09	1,88	10,93	12,57	1,85	2,47	7,17	4	10,34
mazowieckie	2,21	0,92	13,44	5,61	1,7	2,88	9,93	4	12,06
opolskie	1,65	5,27	7,71	19,75	1,63	3,1	7,79	2	14,23
podkarpackie	1,16	2,78	8,57	16,8	1,78	2,07	7,43	1	23,34
podlaskie	1,62	2,67	12,19	15,37	1,34	2,36	7,99	2	17,34
pomorskie	2,27	2,81	10,85	12,5	1,77	2,47	8,48	2	13,28
śląskie	2,6	1,01	10,13	7,8	1,46	2,47	7,75	1	10,14
świętokrzyskie	1,73	4,88	9,41	9,58	1,56	1,77	8,3	1	10,4
warmińsko-mazurskie	1,87	4,12	7,59	11,12	1,32	1,89	9,94	2	11,03
wielkopolskie	1,84	1,9	9,5	8,77	1,32	2,31	8,66	1	14,51
zachodniopomorskie	2,13	2,94	11,53	19,37	1,85	2,86	8,58	2	8,91

Źródło: Opracowanie własne w programie Excel

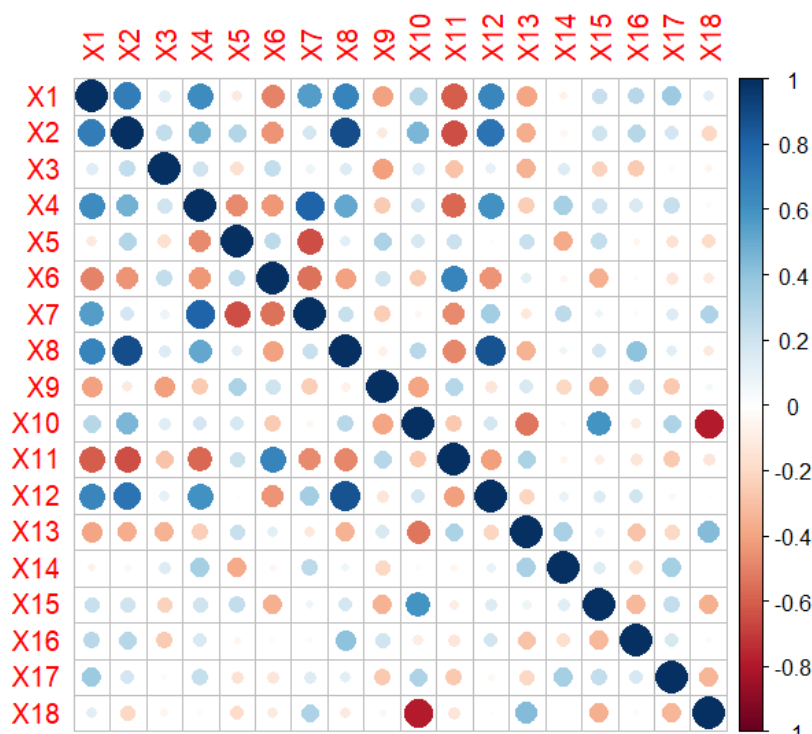
Zmienne i ich objaśnienia:

- X1** – liczba lekarzy pediatrów na 100 tys. osób
- X2** – liczba lekarzy anestezjologii i intensywnej terapii na 100 tys. osób
- X3** – liczba przystanków autobusowych (z trolejbusowymi i tramwajowymi) na 10 km²
- X4** – średnia cena mieszkań sprzedanych w ramach transakcji rynkowych 100 tys. zł
- X5** – liczba szpitali ogólnych na 100 tys. osób
- X6** – liczba domów stacjonarnej pomocy społecznej na 100 tys. osób
- X7** – liczba porodów na 1000 kobiet
- X8** – liczba lekarzy położnictwa i ginekologii na 10 000 kobiet
- X9** – liczba stanowisk porodowych w szpitalach ogólnych na 100 tys. kobiet
- X10** – liczba przestępstw stwierdzonych przez policję na 100 osób
- X11** – liczba wypadków drogowych na 100 tys. mieszkańców
- X12** – liczba lekarzy dentystów na 10 tys. osób
- X13** – liczba ośrodków kultury (centrum, domów, ośrodków kultury, klubów i świetlic)
- X14** – przeciętne miesięczne wydatki na 1 osobę (w tys. zł)
- X15** – odsetek dzieci 0-3 lata w żłobkach i innych formach opieki (na 10)
- X16** – liczba liceów ogólnokształcących na 100 tys. osób
- X17** – tereny rekreacji i wypoczynku jako % całego terenu
- X18** – wskaźnik urodzeń martwych na 1000 urodzeń żywych

Analiza kanoniczna

Pierwsza wykonana analiza ma na celu sprawdzić czy zmienne z jednej mogą być użyte, aby wyjaśnić jak największy zakres zmienności zmiennych z innej grupy.

Wykres 1: Macierz korelacji między zmiennymi



Źródło: Opracowanie własne w programie R

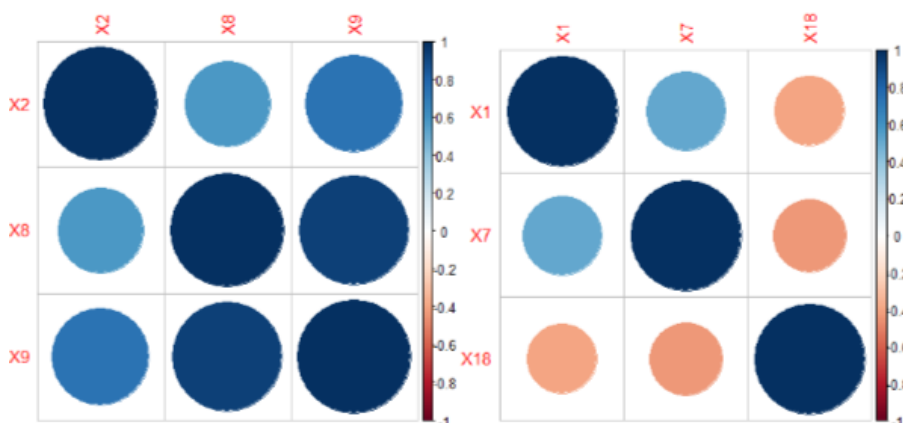
Na podstawie obserwacji zależności między zmiennymi w macierzy korelacji, do analizy kanonicznej zostały wybrane poniższe dwa zbiory:

1. Opieka zdrowotna: X2, X8, X9
2. Rodzące się dzieci: X1, X7, X18

Następnie utworzono dla zbioru X i dla zbioru Y:

- korelogramy

Wykres 2: Korelogramy dla zbioru X i Y



Źródło: Opracowanie własne w programie R

- Macierz współkorelacji między zbiorami X i Y:

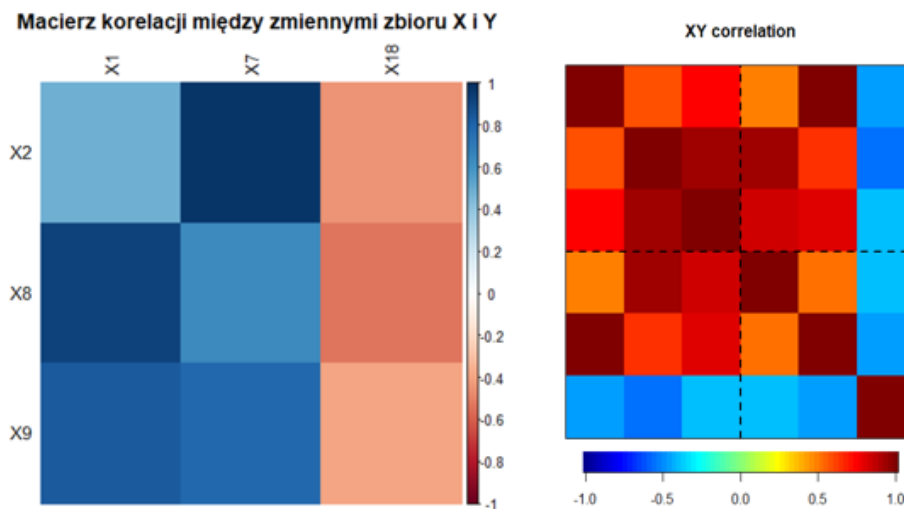
Tabela 3: Macierz współkorelacji między zbiorami X i Y

	X1	X7	X18
X2	0.483	0.976	-0.447
X8	0.921	0.639	-0.537
X9	0.830	0.786	-0.397

Źródło: Opracowanie własne w programie R

- Macierze korelacji między zmiennymi zbioru X i Y:

Wykres 3, wykres 4: Macierze korelacji między zmiennymi zbioru X i Y



Źródło: Opracowanie własne w programie R.

Następnie przeprowadzono **analizę kanoniczną**. Poniżej przedstawiono wyniki:

Analiza kanoniczna

```
$scores$corr.X.xscores
      [,1]      [,2]      [,3]
X2 -0.9213043  0.3803677 -0.08073884
X8 -0.8441693 -0.4704451  0.25702072
X9 -0.8971234 -0.1493440  0.41577145

$scores$corr.Y.xscores
      [,1]      [,2]      [,3]
X1 -0.7497477 -0.5029803  0.198766358
X7 -0.9316044  0.3084921 -0.007823298
X18 0.5830003  0.3254719  0.419814278

$scores$corr.X.yscores
      [,1]      [,2]      [,3]
X2 -0.9133987  0.3432785 -0.04681824
X8 -0.8369256 -0.4245725  0.14903926
X9 -0.8894254 -0.1347816  0.24109446

$scores$corr.Y.yscores
      [,1]      [,2]      [,3]
X1 -0.7562368 -0.5573245  0.34277593
X7 -0.9396675  0.3418229 -0.01349141
X18 0.5880462  0.3606373  0.72397680
```

List of length 5

0.991 0.902 0.580

List of length 3

-0.0419 -0.0504 0.1194 0.0321 -0.1572 0.6090 -0.0788 -0.1652 1.5067 ...

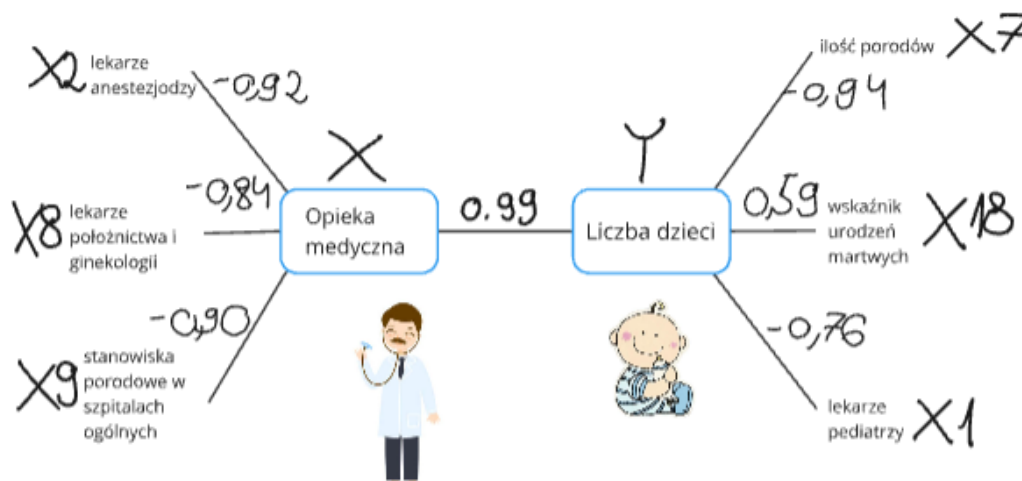
-0.0373 -0.0402 0.0348 -0.0994 0.0572 0.1054 0.0798 0.0042 0.2482 ...

List of length 6

Źródło: Opracowanie własne w programie R

Ilustracja 1: zależności dla 1. zmiennej kanonicznej

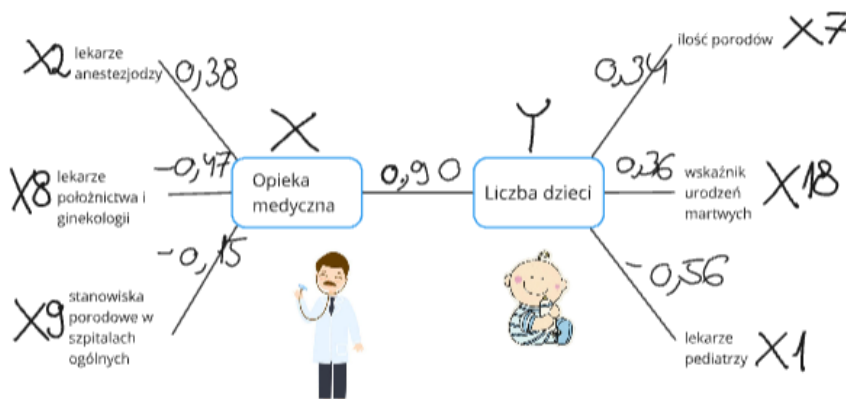
Dla 1 zmiennej kanonicznej:



Źródło: Opracowanie własne w programie Miro

Ilustracja 2: zależności dla 2. zmiennej kanonicznej

Dla 2 zmiennej kanonicznej:



Źródło: Opracowanie własne w programie Miro

Wszystkie zmienne w zbiorze X w podobnym stopniu wpływają na wyniki w zbiorze Y . Może to wskazywać na szeroki i systematyczny związek między tymi zestawami zmiennych. Poza ogólną zależnością istnieje specyficzny związek między każdą z tych zmiennych. **Przeprowadzono test Lambda-Wilksa, który wykazał istotność statystyczną dla 2 pierwszych zmiennych kanonicznych:**

Wynik testu Lambda-Wilksa

```
Canonical correlation analysis of:
  3 X variables: X2, X8, X9
with 3 Y variables: X1, X7, X18

CanR CanRSQ Eigen percent cum scree
1 0.9914 0.9829 57.5205 92.1542 92.15 *****
2 0.9025 0.8145 4.3905 7.0341 99.19 **
3 0.5799 0.3363 0.5066 0.8116 100.00

Test of H0: The canonical correlations in the
current row and all that follow are zero

CanR LR test stat approx F numDF denDF Pr(> F)
1 0.99142 0.00210 31.5274 9 24.488 2.581e-11 ***
2 0.90249 0.12313 10.1739 4 22.000 8.085e-05 ***
3 0.57987 0.66375 6.0792 1 12.000 0.02973 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Raw canonical coefficients

X variables:
Xcan1 Xcan2 Xcan3
X2 -0.041862 0.03206 -0.078812
X8 -0.050417 -0.15719 -0.165249
X9 0.119374 0.60896 1.506711

Y variables:
Ycan1 Ycan2 Ycan3
X1 -0.037302 -0.099393 0.0798090
X7 -0.040245 0.057191 0.0041997
X18 0.034790 0.105434 0.2481820
Ycan1 Ycan2 Ycan3
X1 -0.03730189 -0.09939251 0.079809020
X7 -0.04024474 0.05719128 0.004199668
X18 0.03479045 0.10543407 0.248181965
```

Źródło: Opracowanie własne w programie R

Na końcu przeprowadzono próbę redundancji zmiennych:

Redundancies for the OPIEKA ZDROWOTNA variables & total X canonical redundancy

Xcan1	Xcan2	Xcan3	total X Y
0.77527	0.10542	0.02751	0.90821

Redundancies for the URODZENIA variables & total Y canonical redundancy

Ycan1	Ycan2	Ycan3	total Y X
0.58997	0.15136	0.07194	0.81327

Źródło: Opracowanie własne w programie R.

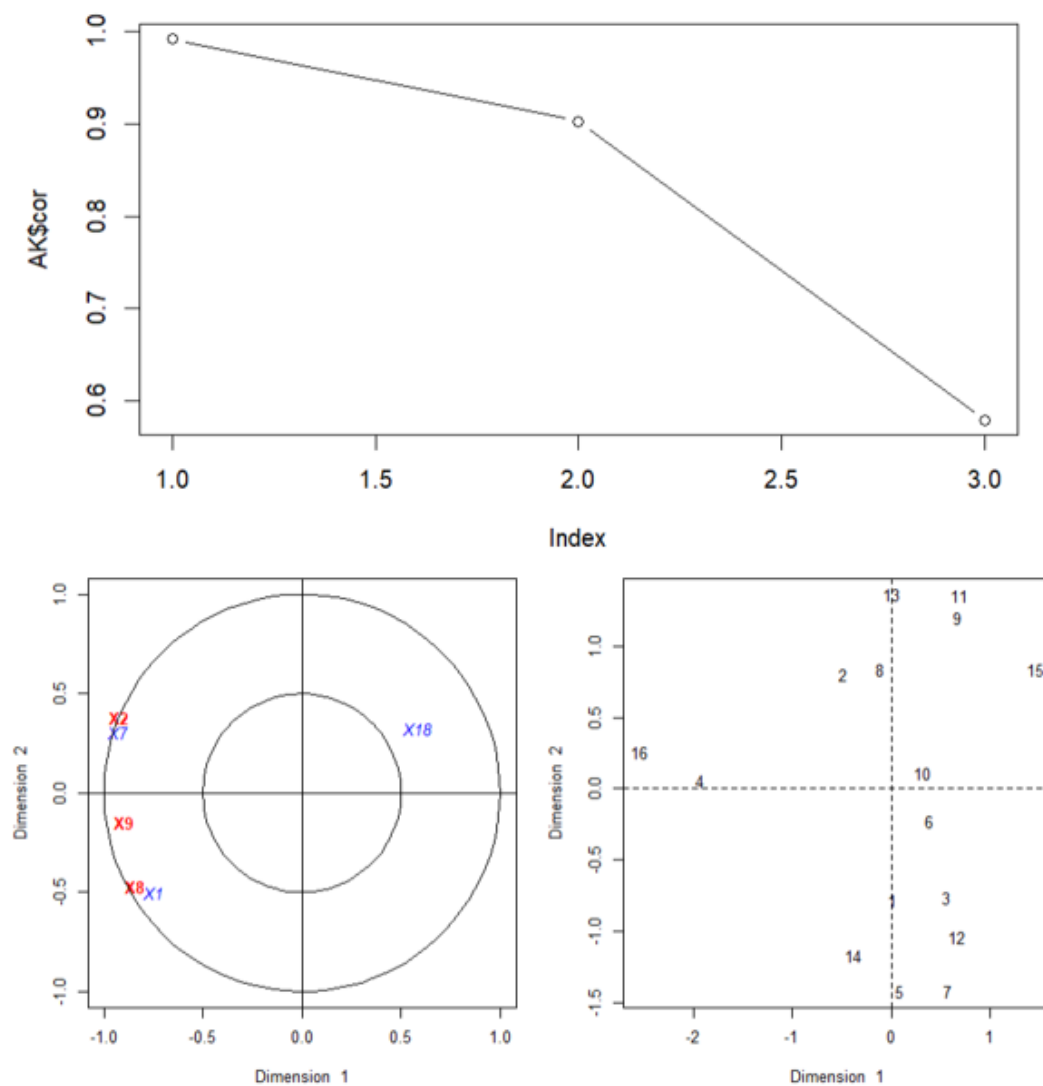
Interpretacja dla zmiennych zbioru X:

Pierwsza zmienna kanoniczna wyjaśnia 77,5 % wariancji zmiennych X. Druga zmienna wyjaśnia 10,5 % wariancji zmiennych Y. Razem wyjaśniają 88 % wariancji zmiennych Y.

Interpretacja dla zmiennych zbioru Y:

Pierwsza zmienna kanoniczna wyjaśnia 59 % wariancji zmiennych Y. Druga zmienna wyjaśnia 15,14 % wariancji zmiennych X. Razem wyjaśniają 74 % wariancji zmiennych X.

Wykres 5: Wizualizacja wyniku analizy kanonicznej



Źródło: Opracowanie własne w programie R.

Analiza głównych składowych (PCA):

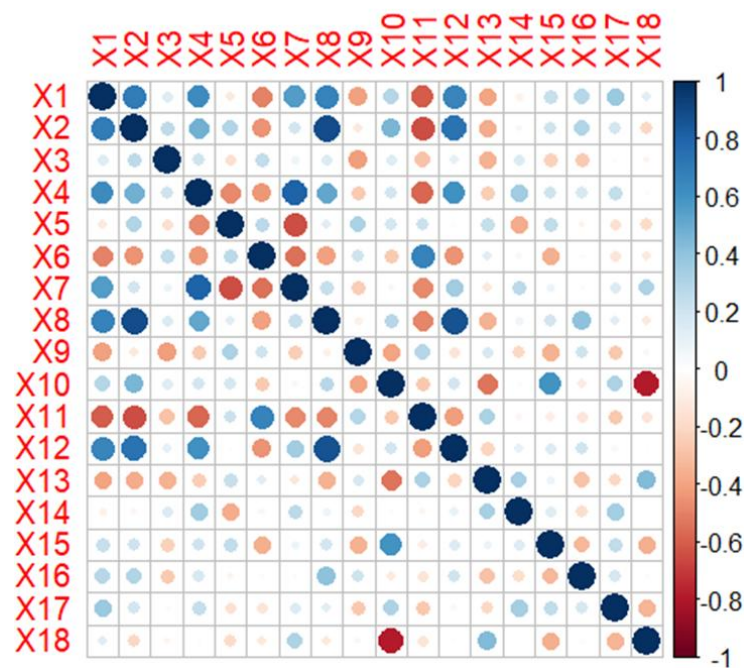
Celem tej analizy jest eliminacja zmiennych, które wnoszą zbyt mało informacji. Szukany jest jak najmniejszy zestaw danych, który jak najdokładniej mógłby opisać zebrane. By wykonać analizę głównych składowych, przeanalizowano poniższą macierz korelacji. **Zielonym** kolorem oznaczono wysokie współczynniki korelacji. Obrano to za kryterium wyboru zmiennych do **PCA**.

Tabela 4: Macierz korelacji zmiennych wyjściowych

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18
X1	1.000	0.695	0.134	0.628	-0.111	-0.495	0.550	0.665	-0.406	0.283	-0.607	0.650	-0.393	-0.070	0.229	0.275	0.367	0.123
X2	0.695	1.000	0.250	0.477	0.292	-0.445	0.182	0.883	-0.112	0.458	-0.649	0.734	-0.368	-0.049	0.205	0.290	0.185	-0.204
X3	0.134	0.250	1.000	0.207	-0.165	0.244	0.077	0.147	-0.413	0.136	-0.283	0.103	-0.341	0.142	-0.225	-0.257	-0.029	-0.058
X4	0.628	0.477	0.207	1.000	-0.477	-0.432	0.802	0.513	-0.260	0.186	-0.579	0.607	-0.248	0.340	0.206	0.169	0.234	0.027
X5	-0.111	0.292	-0.165	-0.477	1.000	0.262	0.643	0.123	0.311	0.176	0.215	-0.019	0.226	-0.370	0.247	-0.053	-0.155	-0.187
X6	-0.495	-0.445	0.244	-0.432	0.262	1.000	-0.543	-0.409	0.209	-0.255	0.663	-0.445	0.121	-0.045	-0.354	0.027	-0.131	-0.116
X7	0.550	0.182	0.077	0.802	0.643	-0.543	1.000	0.222	-0.249	-0.044	-0.477	0.341	-0.123	0.267	0.071	-0.028	0.148	0.308
X8	0.665	0.883	0.147	0.513	0.123	-0.409	0.222	1.000	-0.076	0.270	-0.486	0.861	-0.350	0.068	0.187	0.404	0.125	-0.115
X9	-0.406	-0.112	-0.413	-0.260	0.311	0.209	-0.249	-0.076	1.000	-0.392	0.288	-0.135	0.164	-0.203	-0.349	0.200	-0.262	0.041
X10	0.283	0.458	0.136	0.186	0.176	-0.255	-0.044	0.270	-0.392	1.000	-0.261	0.188	-0.531	-0.029	0.597	-0.102	0.306	-0.788
X11	-0.607	-0.649	-0.283	-0.579	0.215	0.663	-0.477	-0.486	0.288	-0.261	1.000	-0.412	0.313	-0.052	-0.091	-0.132	-0.268	-0.133
X12	0.650	0.734	0.103	0.607	-0.019	-0.445	0.341	0.861	-0.135	0.188	-0.412	1.000	-0.220	0.097	0.148	0.193	-0.039	0.011
X13	-0.393	-0.368	-0.341	-0.248	0.226	0.121	-0.123	-0.350	0.164	-0.531	0.313	-0.220	1.000	0.325	0.075	-0.283	-0.207	0.435
X14	-0.070	-0.049	0.142	0.340	-0.370	-0.045	0.267	0.068	-0.203	-0.029	-0.052	0.097	0.325	1.000	0.139	-0.180	0.331	-0.006
X15	0.229	0.205	-0.225	0.206	0.247	-0.354	0.071	0.187	-0.349	0.597	-0.091	0.148	0.075	0.139	1.000	-0.329	0.243	-0.358
X16	0.275	0.290	-0.257	0.169	-0.053	0.027	-0.028	0.404	0.200	-0.102	-0.132	0.193	-0.283	-0.180	-0.329	1.000	0.173	-0.026
X17	0.367	0.185	-0.029	0.234	-0.155	-0.131	0.148	0.125	-0.262	0.306	-0.268	-0.039	-0.207	0.331	0.243	0.173	1.000	-0.336
X18	0.123	-0.204	-0.058	0.027	-0.187	-0.116	0.308	-0.115	0.041	-0.788	-0.133	0.011	0.435	-0.006	-0.358	-0.026	-0.336	1.000

Źródło: Opracowanie własne w programie R.

Wykres 6: Macierz korelacji zmiennych wyjściowych



Źródło: Opracowanie własne w programie R

Zmienne, które zostały uznane za nieistotne, czyli **X3, X9, X13, X14, X15, X16** oraz **X17** zostały wykluczone z analizy. Istotne zmienne przedstawiono w poniższej tabeli.

Tabela 5: Macierz zmiennych do analizy PCA

X1	X2	X4	X5	X6	X7	X8	X10	X11	X12	X18
2,27	14,48	4,2	2,71	2,33	13,05	2,99	2,75	2,16	11,58	9,21
2,04	13,18	3,01	2,15	2,45	13,12	2,65	2,17	1,96	6,81	10,39
1,53	14,77	3,84	2,24	2,24	13,00	3,33	1,65	1,95	11,81	13,23
1,13	11,49	2,91	2,46	2,36	12,95	2,65	2,53	4,65	8,28	4,05
2,37	15,92	3,33	2,62	2,84	13,25	3,71	2,11	3,8	13,4	11,45
1,94	14,75	5,5	2,01	2,77	16,03	2,99	2,09	1,88	10,93	10,34
2,79	19,02	6,07	2,11	1,98	15,63	4,08	2,21	0,92	13,44	12,06
1,36	11,32	2,77	3,1	3,1	12,76	2,75	1,65	5,27	7,71	14,23
1,46	9,75	3,35	1,98	2,61	14,07	2,46	1,16	2,78	8,57	23,34
2,35	16,43	3,61	3,07	1,93	14,16	3,22	1,62	2,67	12,19	17,34
2,08	13,69	4,73	1,7	1,95	15,76	2,87	2,27	2,81	10,85	13,28
1,81	18,54	3,04	3,31	2,38	12,32	3,36	2,6	1,01	10,13	10,14
1,33	11,55	3,21	2,05	3,51	12,06	2,72	1,73	4,88	9,41	10,4
1,49	12,45	3,03	3,09	3,53	12,21	2,57	1,87	4,12	7,59	11,03
1,91	12,5	3,9	1,63	1,95	15,35	2,93	1,84	1,9	9,5	14,51
1,45	15,2	3,62	2,82	2,33	12,43	3,3	2,13	2,94	11,53	8,91

Źródło: Opracowanie własne w programie Excel

Przeprowadzono analizę głównych składowych i przeanalizowano stopień, w jakim każda kolejna składowa wyjaśnia wariancję.

Tabela 6. Wartości własne kolejnych głównych składowych

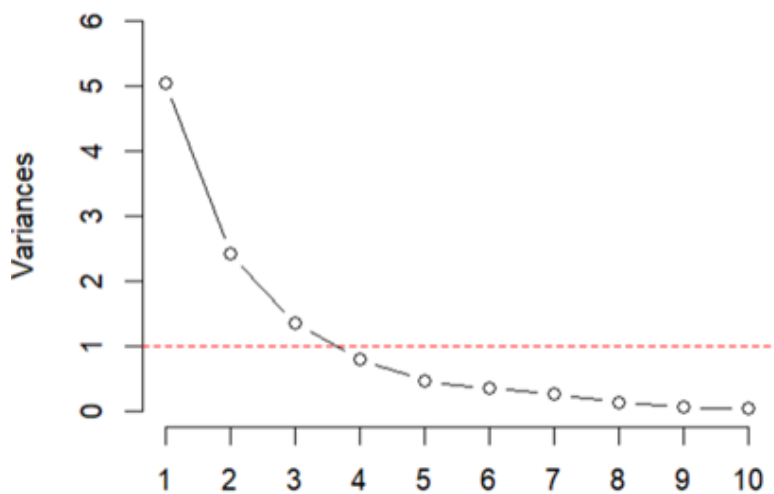
Rząd głównej składowej	Wartość własna	% wyj. wariancji	% wyj. Skumulowanej wariancji
1	5,045968	45,87243	45,87243
2	2,415743	21,9613	67,83373
3	1,348811	12,26191	80,09565
4	0,812997	7,39088	87,48653
5	0,463629	4,21481	91,70134
6	0,366909	3,335534	95,03687

7	0,262308	2,384622	97,42149
8	0,146346	1,330417	98,75191
9	0,062476	0,567966	99,31988
10	0,049425	0,449316	99,76919
11	0,025389	0,230806	100

Źródło: Opracowanie własne, obliczenia wykonane w programie R, prezentacja w programie Excel

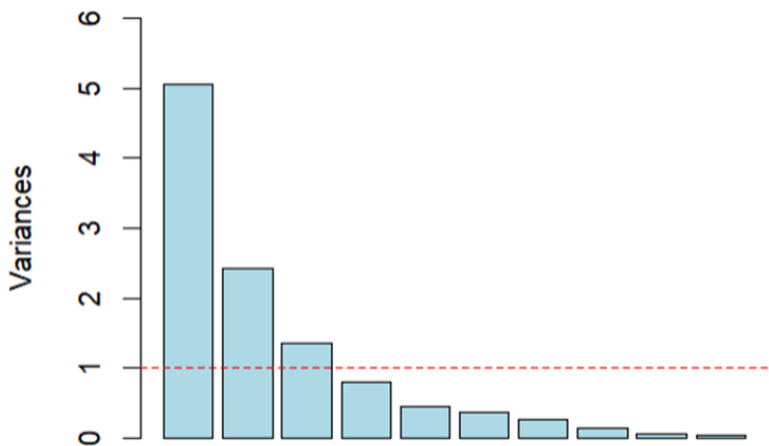
Wstępnie zastosowano **regułę Kaisera** i zdecydowano się na wybór głównych składowych do **3 rzędu**. By potwierdzić decyzję zastosowano wizualną ocenę wykresów zgodnie z **regułą Catella**.

Wykres 7: Wykres osypiska



Źródło: Opracowanie własne w programie R

Wykres 8: Wykres słupkowy



Źródło: Opracowanie własne w programie R

Na podstawie tych dwóch reguł podjęto decyzję o ograniczeniu głównych składowych do 3.

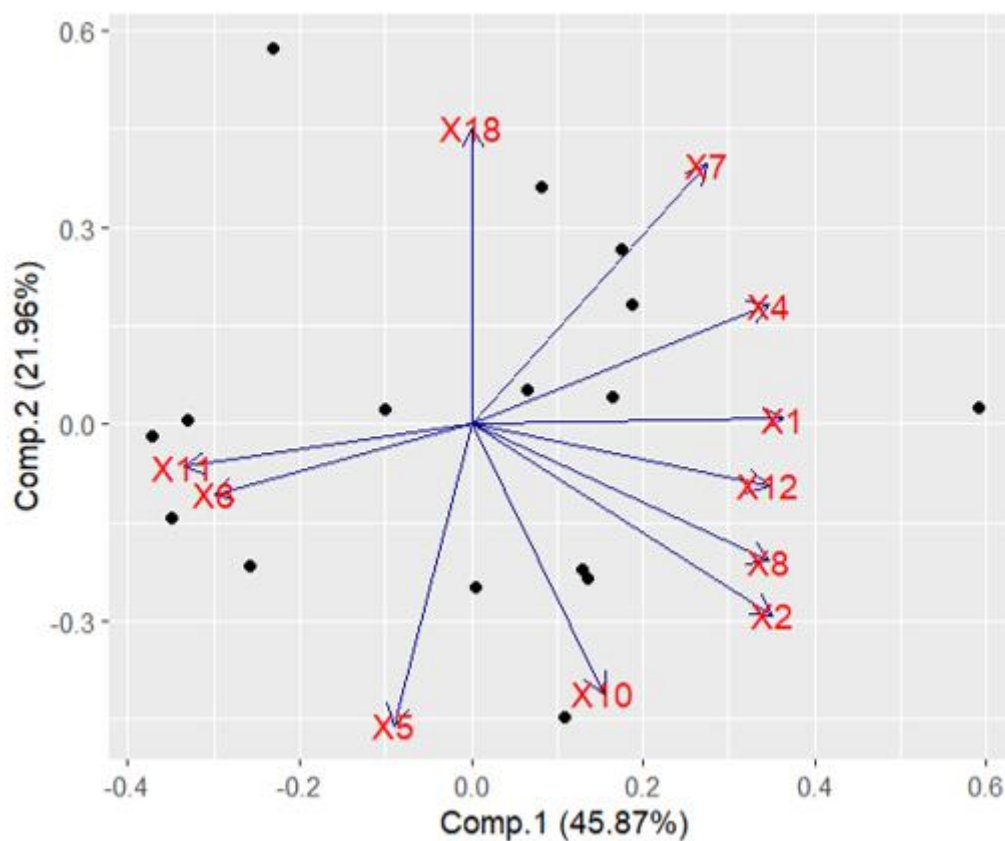
Tabela 7: Wektory własne macierzy korelacji:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
X1	-0,37864	-0,00865	-0,10423	0,018075	-0,4439	-0,49125	0,531134	-0,04258
X2	-0,36554	0,303175	-0,1895	-0,06297	-0,12479	0,187123	-0,09279	-0,27822
X4	-0,35978	-0,18881	0,191331	0,353972	-0,2136	0,166058	-0,44298	0,218738
X5	0,094298	0,476629	-0,41151	-0,20709	-0,24891	-0,2729	-0,56475	0,06076
X6	0,312293	0,111047	-0,11287	0,562806	-0,58658	0,295163	0,133873	0,084638
X7	-0,28557	-0,41085	0,206854	0,106276	-0,14444	-0,29611	-0,3846	-0,32095
X8	-0,3609	0,216996	-0,27027	0,230212	0,234261	0,167938	0,13424	-0,44711
X10	-0,15981	0,426861	0,514615	-0,12584	-0,10996	-0,22866	0,067954	0,3365
X11	0,347102	0,067353	-0,03932	0,507402	0,255982	-0,59206	-0,06903	-0,18301
X12	-0,35963	0,097188	-0,25219	0,356486	0,409334	-0,05706	0,020209	0,574304
X18	-0,00163	-0,46987	-0,53733	-0,21473	-0,12169	-0,11193	0,021423	0,287417

Źródło: Opracowanie własne, obliczenia wykonane w programie R, prezentacja w programie Excel

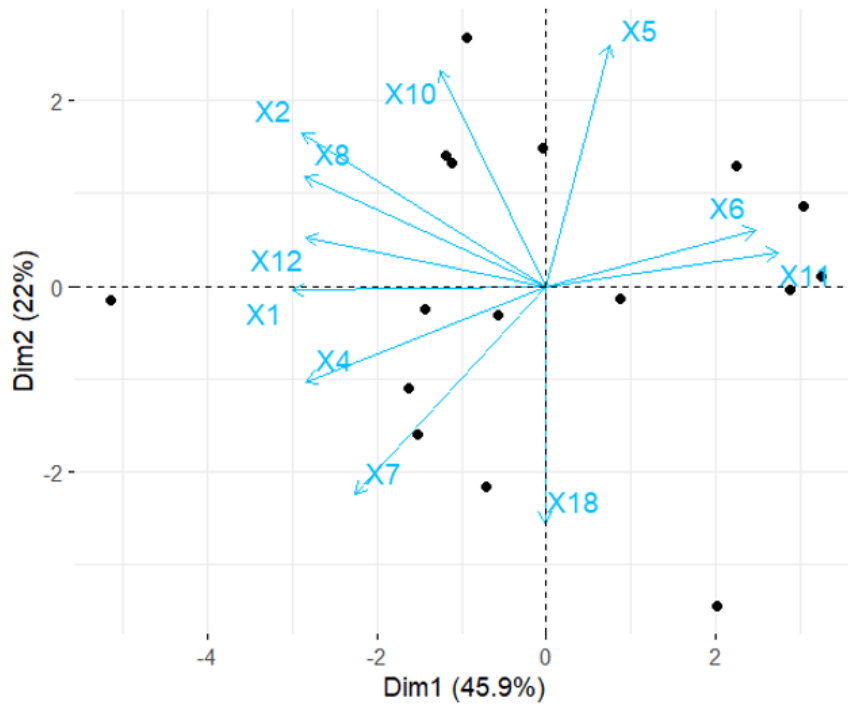
Poniżej przedstawiono wykresy opisujące wkład poszczególnych zmiennych w 1 i 2 składową, wykres ładunków czynnikowych oraz rzut województw na osie **PC1** i **PC2**.

Wykres 9: wykres Bigplot dla PC1



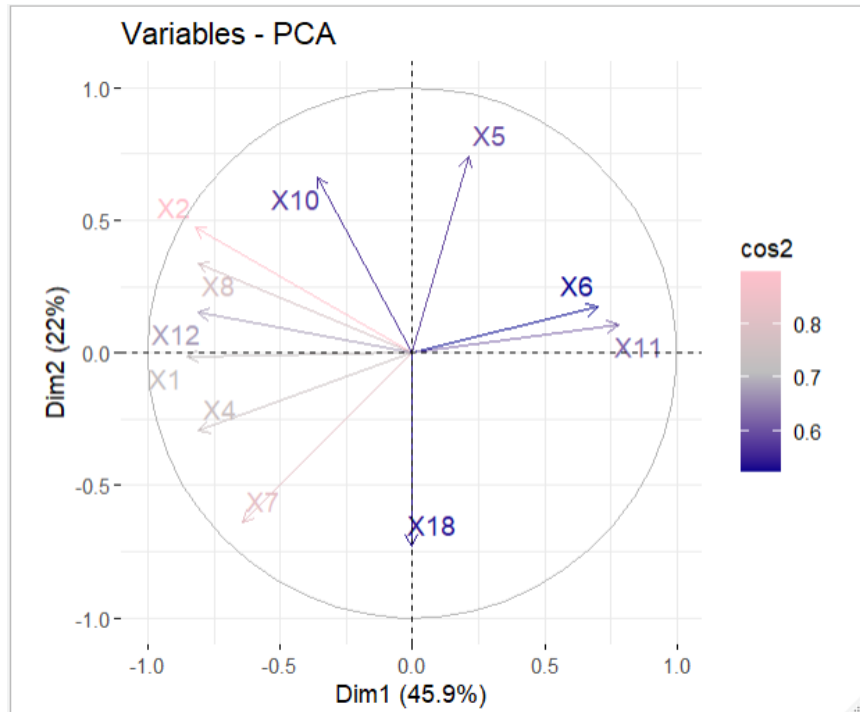
Źródło: Opracowanie własne w programie R

Wykres 9: wykres Bigplot dla PC1



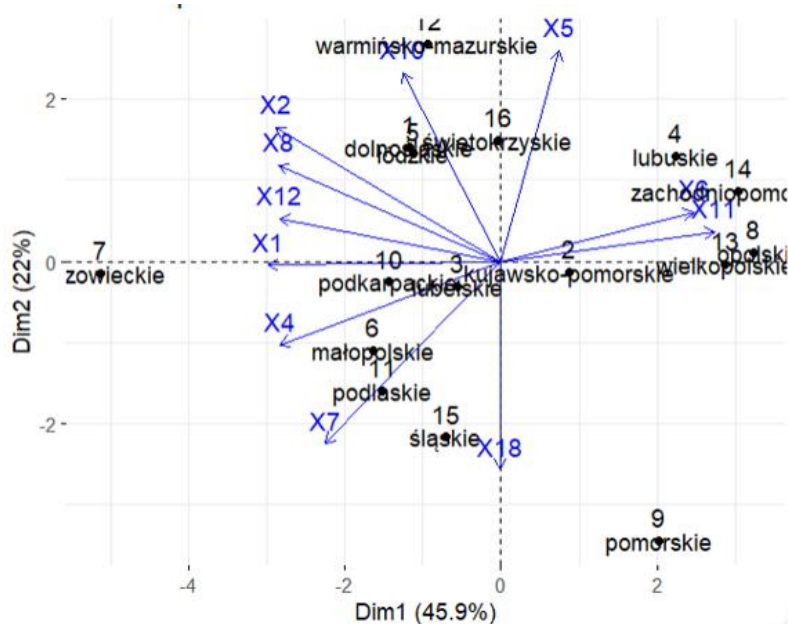
Źródło: Opracowanie własne w progamie R

Wykres 10: wykres ładunków czynnikowych



Źródło: Opracowanie własne w progamie R

Wykres 10: wykres ładunków czynnikowych



Źródło: Opracowanie własne w programie R

Wnioski:

- Dłuższy wektor zmiennej odpowiada za jej większy udział w zmienności danej składowej. Z wykresu wynika, że wszystkie zmienne mają podobny udział w **PC1** i **PC2**. Są to zmienne zarówno opisujące kategorie zdrowie i opieka medyczna, społeczeństwo i demografia jak i infrastruktura i dostępność usług. Największy udział mają zmienne: **X5**, **X7** i **X2**.
- Zmienna **X18** praktycznie nie jest skorelowana ze zmienną **X1** i bardzo słabo skorelowana ze zmienną **X11**.
- Silna dodatnia korelacja występuje między zmiennymi **X2**, **X8**, **X12** i **X1** oraz silna dodatnia korelacja między **X6**, a **X11**.
- Ujemna korelacja występuje między zmiennymi **X5**, a **X7** i **X4** oraz ujemna korelacja między **X6**, a **X7** i **X4**.

Równania głównych składowych

Poniżej przedstawiono równania dla pierwszych trzech głównych składowych oraz tabele ładunków.

$$Y_1 = -0.38X_1 - 0.37X_2 - 0.36X_4 + 0.09X_5 + 0.31X_6 - 0.29X_7 - 0.36X_8 - 0.16X_{10} - 0.35X_{11} - 0.36X_{12} - 0.001X_{18}$$

$$Y_2 = -0.009X_1 + 0.30X_2 - 0.19X_4 + 0.48X_5 + 0.11X_6 - 0.41X_7 + 0.21X_8 + 0.42X_{10} + 0.07X_{11} + 0.10X_{12} - 0.47X_{18}$$

$$Y_3 = -0.10X_1 - 0.19X_2 + 0.19X_4 - 0.41X_5 - 0.11X_6 + 0.21X_7 - 0.27X_8 + 0.51X_{10} - 0.04X_{11} - 0.25X_{12} - 0.53X_{18}$$

Tabela 8: Ładunki zmiennych wyjściowych w analizie głównych składowych

	Ładunek PC1	Ładunek PC2	Ładunek PC3
X1	-0,85	-0,01	-0,12
X2	-0,82	-0,47	-0,22
X4	-0,81	-0,29	0,22
X5	0,21	0,74	-0,48
X6	0,7	0,17	-0,13
X7	-0,64	-0,64	0,24
X8	-0,81	-0,34	-0,31
X10	-0,36	0,66	0,59
X11	0,78	0,1	-0,04
X12	-0,81	0,15	-0,29
X18	-0,004	-0,73	-0,62

Źródło: Opracowanie własne, obliczenia wykonane w programie R, prezentacja w programie Excel

Wnioski z analizy PCA:

- Pierwsza składowa jest indeksem przedstawiającym zależności pomiędzy większością zmiennych, które wpływają na to, w którym województwie warto założyć rodzinę,
- Druga składowa przedstawia relacje zmiennych **X2**, **X5**, **X7**, **X10** i **X18**. Zmienne pochodzą z kategorii *Zdrowie i opieka medyczna* (**X2**, **X5**, **X7** i **X18**) oraz *Spółeczeństwo i demografia* (**X10**),
- Trzecia składowa przedstawia relacje zmiennych **X5**, **X10** i **X18**, czyli zależność pomiędzy liczbą szpitali ogólnych, a liczbą urodzeń martwych oraz liczbą przestępstw.

Porządkowanie liniowe

Celem porządkowania liniowego jest ustalenie hierarchii województw. Do wykonania porządkowania liniowego wykorzystane zostaną wszystkie czynniki.

W projekcie użyto następujących metod porządkowania liniowego:

- TOPSIS oraz TOPSIS z wagami
- Hellwig (metoda wzorca) oraz metoda wzorca z wagami
- COPRAS oraz COPRAS z wagami
- metoda sumy rang
- metoda dystansów

Pierwszym etapem porządkowania liniowego jest określenie charakteru zmiennych.

Tabela 9: Charakter zmiennych S/D

	Nazwa zmiennej	Charakter zmiennej
X1	liczba lekarzy pediatrów na 100 tys. osób	S
X2	liczba lekarzy anestezjologii i intensywnej terapii na 100 tys. osób	S
X3	liczba przystanków autobusowych (z trolejbusowymi i tramwajowymi) na 10 km ²	S
X4	średnia cena mieszkań sprzedanych w ramach transakcji rynkowych 100 tys. zł	D
X5	liczba szpitali ogólnych na 100 tys. osób	S
X6	ilość domów stacjonarnej pomocy społecznej na 100 tys. osób	S
X7	ilość porodów na 1000 kobiet	S
X8	ilość lekarzy położnictwa i ginekologii na 10 000 kobiet	S
X9	stanowiska porodowe w szpitalach ogólnych na 100 tys. kobiet	S
X10	ilość przestępstw stwierdzonych przez policję na 100 osób	D
X11	ilość wypadków drogowych na 100 tys. mieszkańców	D
X12	ilość lekarzy dentystów na 10 tys. osób	S
X13	liczba ośrodków kultury (centrum, domów, ośrodków kultury, klubów i świetlic)	S
X14	przeciętne miesięczne wydatki na 1 osobę (w tys. zł)	D
X15	odsetek dzieci 0-3 lata w żłobkach i innych formach opieki (na 10)	S
X16	liczba liceów ogólnokształcących na 100 tys. osób	S
X17	tereny rekreacji i wypoczynku jako % całego terenu	S
X18	wskaźnik urodzeń martwych na 1000 urodzeń żywych	D

Źródło: Opracowanie własne, obliczenia wykonane w programie R, prezentacja w programie Excel

Oznaczono je w następujący sposób:

- **S** – stymulanty – zmienne, dla których korzystne są wysokie wartości
- **D** – destymulanty – zmienne, dla których korzystne są niskie wartości

Poniżej przedstawiono wyniki dla sześciu przeprowadzonych metod porządkowania liniowego.

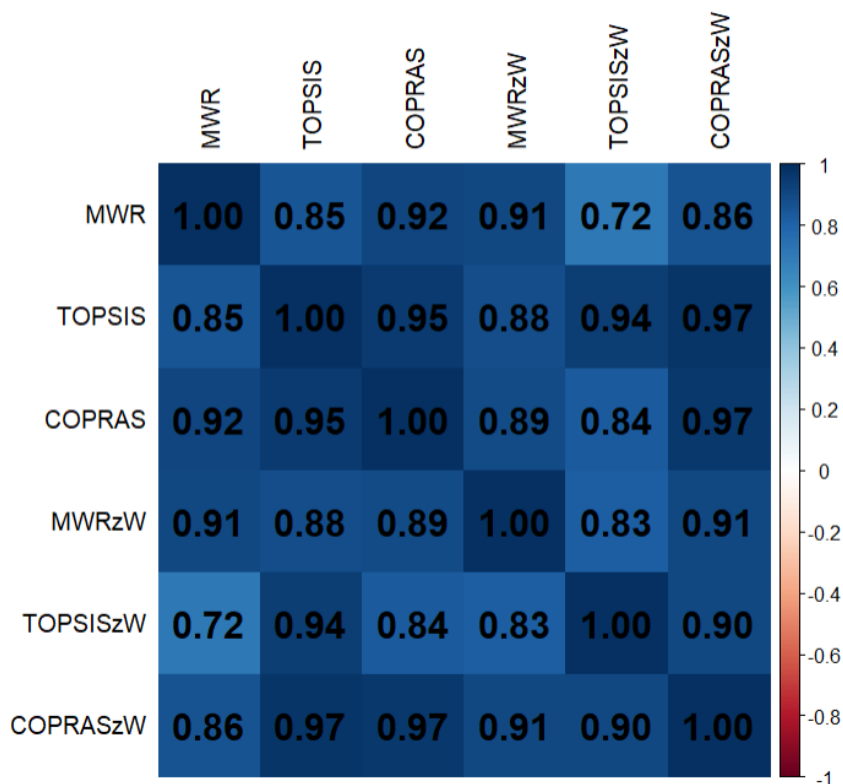
Tabela 10: wyniki porządkowania liniowego

Województwo	MWR	TOPSIS	COPRAS	MWRzW	TOPSISzW	COPRASzW
małopolskie	1	1	1	1	2	1
mazowieckie	2	3	2	3	3	3
dolnośląskie	3	4	4	4	4	5
łódzkie	4	8	6	6	7	7
zachodniopomorskie	5	6	7	7	8	8
podlaskie	6	9	8	9	12	9
kujawsko-pomorskie	7	2	3	2	1	2
lubelskie	8	7	9	5	6	6
śląskie	9	5	5	10	5	4
pomorskie	10	11	11	8	11	10
opolskie	11	12	10	11	14	11
warmińsko-mazurskie	12	15	13	12	16	14
wielkopolskie	13	14	14	14	13	13
świętokrzyskie	14	10	12	15	10	12
lubuskie	15	13	16	13	9	15
podkarpackie	16	16	15	16	15	16

Źródło: Opracowanie własne w programie Excel

Poniżej przedstawiono zgodność poszczególnych metod porządkowania.

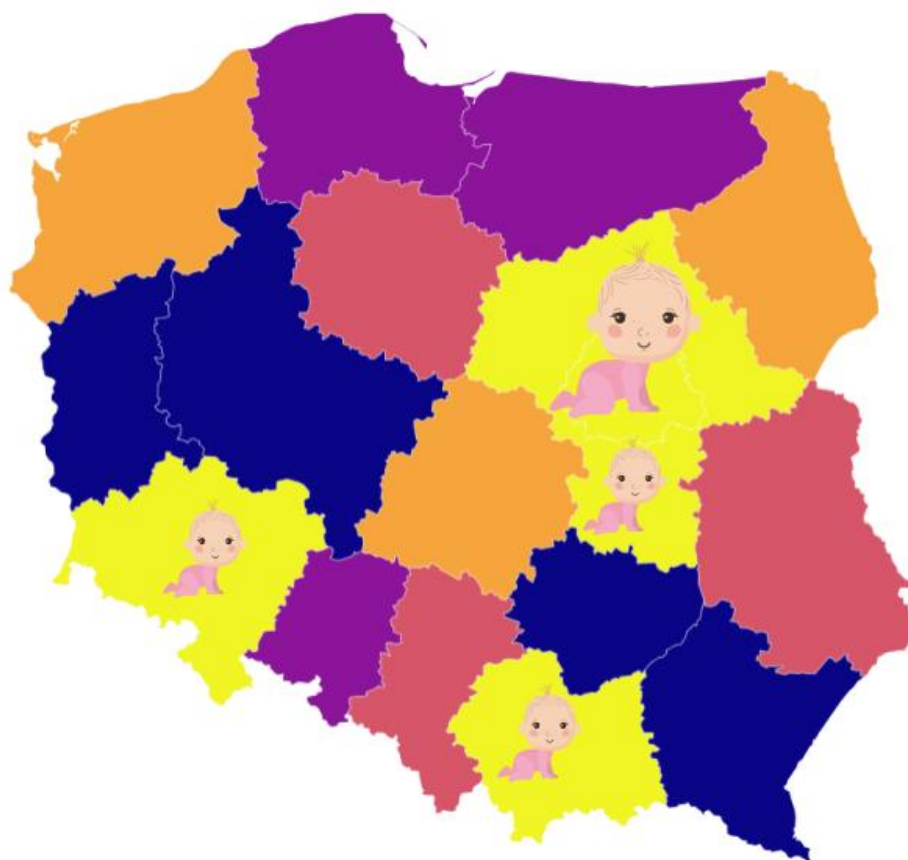
Wykres 11: Analiza zgodności metod porządkowania, macierz korelacji



Źródło: Opracowanie własne w programie R

Wyniki zilustrowano na mapie Polski. Oceny dokonano na podstawie metody wytypowanej jako najlepsza – **metody wzorca rozwoju**.

Ilustracja 1: Ranking województw według metody wzorca



Źródło: Opracowanie własne za pomocą oprogramowania na stronie: <https://app.datawrapper.de/>

Rankingi według pozostałych metod zostały przedstawione poniżej.

Ilustracja 2: Ranking województw wg metody sumy rang



Źródło: Opracowanie własne w programie Canva

Ilustracja 3: Ranking województw wg metody dystansów



Źródło: Opracowanie własne w programie Canva

Podsumowanie metod porządkowania liniowego

Wyniki porządkowania liniowego wskazują, że województwa, w których najlepiej założyć rodzinę to:

- Małopolskie
- Mazowieckie
- Śląskie
- Dolnośląskie
- Kujawsko-pomorskie

Natomiast województwa, w których najmniej opłaca się założyć rodzinę to:

- Podkarpackie
- Lubuskie
- Wielkopolskie
- Świętokrzyskie
- Warmińsko-Mazurskie

Analiza Skupień

Do analizy skupień wybrano jedynie te dane, które brano pod uwagę w metodach porządkowania – wykluczono zmienne o niskim współczynniku zmienności. Kierowano się założeniem, że zmienne o niskiej zmienności w mniejszym stopniu wyjaśniają różnice między obiektami. Za graniczną zmienność przyjęto wartość 0,15 podobnie jak w metodach porządkowania liniowego. Analiza skupień będzie dotyczyć zmiennych przedstawionych w **tabeli 11**.

Tabela 11: Zmienne zastosowane w analizie skupień

	Nazwa zmiennej
X1	liczba lekarzy pediatrów na 100 tys. osób
X2	liczba lekarzy anestezjologii i intensywnej terapii na 100 tys. osób
X3	liczba przystanków autobusowych (z trolejbusowymi i tramwajowymi) na 10 km ²
X4	średnia cena mieszkań sprzedanych w ramach transakcji rynkowych 100 tys. zł
X5	liczba szpitali ogólnych na 100 tys. osób
X6	ilość domów stacjonarnej pomocy społecznej na 100 tys. osób
X10	ilość przestępstw stwierdzonych przez policję na 100 osób
X11	ilość wypadków drogowych na 100 tys. mieszkańców
X12	ilość lekarzy dentystów na 10 tys. osób
X13	liczba ośrodków kultury (centrum, domów, ośrodków kultury, klubów i świetlic)
X15	odsetek dzieci 0-3 lata w żłobkach i innych formach opieki (na 10)
X17	tereny rekreacji i wypoczynku jako % całego terenu (na 10)
X18	wskaźnik urodzeń martwych na 1000 urodzeń żywych

Źródło: Opracowanie własne w programie Excel

W analizie skupień rozważono standaryzację danych oraz ważoną standaryzację danych, w której wagą była zmienność. Pierwsza z tych metod okazała się być mniej efektywna w grupowaniu zmiennych, druga natomiast pomimo większej efektywności spowodowała zjawisko dużej gęstości zmiennych, co na tyle mocno niwelowało zróżnicowanie, że pomiędzy grupami było ono niewielkie. Postanowiono więc wykonać analizę skupień na danych wyrażanych w jednostkach wyjściowych.

Tabela 12: wyjściowe dane do analizy skupień

	X1	X2	X3	X4	X5	X6	X10	X11	X12	X13	X15	X17	X18
dolnośląskie	2,27	14,48	6,09	4,2	2,71	2,33	2,75	2,16	11,58	9,66	3,53	3	9,21
kujawsko-pomorskie	2,04	13,18	6,62	3,01	2,15	2,45	2,17	1,96	6,81	10,92	2,45	7	10,39
lubelskie	1,53	14,77	5,2	3,84	2,24	2,24	1,65	1,95	11,81	11,78	2	3	13,23
lubuskie	1,13	11,49	2,66	2,91	2,46	2,36	2,53	4,65	8,28	7,79	2,92	3	4,05
łódzkie	2,37	15,92	7,02	3,33	2,62	2,84	2,11	3,8	13,4	7,92	2,27	3	11,45
małopolskie	1,94	14,75	11,24	5,5	2,01	2,77	2,09	1,88	10,93	12,57	2,47	4	10,34
mazowieckie	2,79	19,02	6,81	6,07	2,11	1,98	2,21	0,92	13,44	5,61	2,88	4	12,06
opolskie	1,36	11,32	4,94	2,77	3,1	3,1	1,65	5,27	7,71	19,75	3,1	2	14,23
podkarpackie	1,46	9,75	6,02	3,35	1,98	2,61	1,16	2,78	8,57	16,8	2,07	1	23,34
podlaskie	2,35	16,43	3,58	3,61	3,07	1,93	1,62	2,67	12,19	15,37	2,36	2	17,34
pomorskie	2,08	13,69	5,3	4,73	1,7	1,95	2,27	2,81	10,85	12,5	2,47	2	13,28
śląskie	1,81	18,54	11,96	3,04	3,31	2,38	2,6	1,01	10,13	7,8	2,47	1	10,14
świętokrzyskie	1,33	11,55	11,71	3,21	2,05	3,51	1,73	4,88	9,41	9,58	1,77	1	10,4
warmińsko-mazurskie	1,49	12,45	2,99	3,03	3,09	3,53	1,87	4,12	7,59	11,12	1,89	2	11,03
wielkopolskie	1,91	12,5	5,68	3,9	1,63	1,95	1,84	1,9	9,5	8,77	2,31	1	14,51
zachodniopomorskie	1,45	15,2	3,21	3,62	2,82	2,33	2,13	2,94	11,53	19,37	2,86	2	8,91

Źródło: Opracowanie własne w programie Excel

By najskuteczniej wybrać metodę grupowania, stworzono macierz korelacji kofenetycznych w następujący sposób:

Dla każdej z poniżej wymienionych **miar odległości**:

- Kwadratowa odległość euklidesowa (*euclidean_sq*)
- Metryka euklidesowa (*euclidean*)
- Metryka miejska (*manhattan*)
- Metryka dominacji (*maximum*)

Została policzona korelacja kofenetyczna dla poszczególnych **aglomeracyjnych metod grupowania hierarchicznego**:

- Metoda najbliższego sąsiada (*single*)
- Metoda najdalszego sąsiada (*complete*)
- Metoda średniej grupowej (*average*)
- Metoda centroidalna (*centroid*)
- Metoda mediany (*median*)
- Metoda Warda (*ward.D*)

Poniżej przedstawiono macierz korelacji kofenetycznych. Kolorem zielonym oznaczono najwyższe wartości.

Tabela 13: macierz korelacji kofenetycznych miar odległości i metod grupowania hierarchicznego

	single	complete	average	centroid	median	ward.D
euclidean_sq	0,7199968	0,6399444	0,7478304	0,7266184	0,6570264	0,398413
euclidean	0,7590878	0,6991816	0,7927856	NA	NA	NA
manhattan	0,6334105	0,563745	0,6716131	NA	NA	NA
maximum	0,7611599	0,6610395	0,7996515	NA	NA	NA

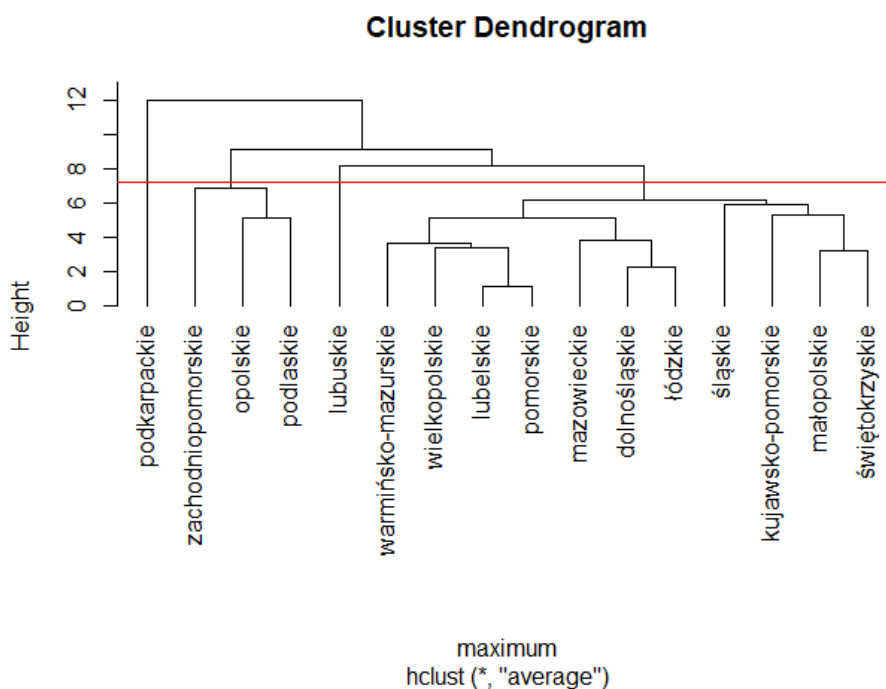
Źródło: Opracowanie własne, obliczenia wykonane w programie R, prezentacja w programie Excel

Najbardziej skuteczne okazały się kombinacje:

1. metryki dominacji z metodą średniej grupowej,
2. metryki euklidesowej z metodą średniej grupowej,
3. metryki dominacji z metodą najbliższego sąsiada.

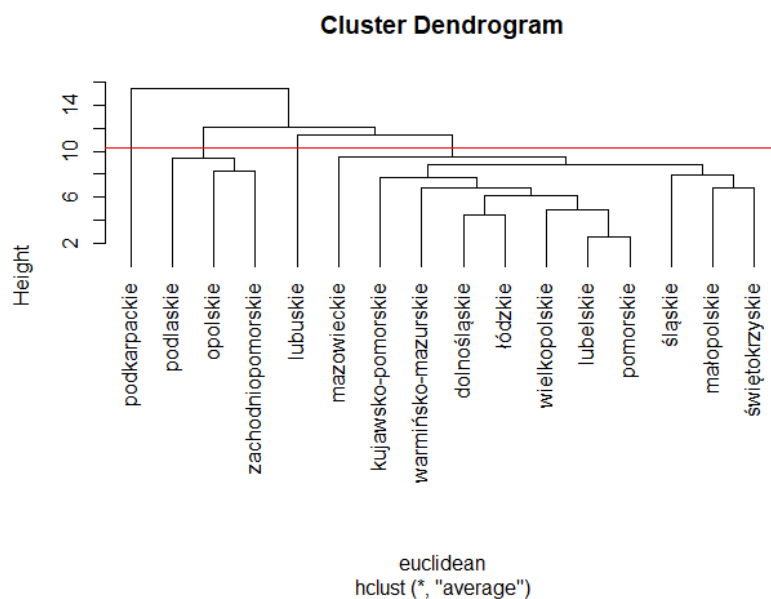
Poniżej przedstawiono dendrogramy dla każdej z metod z zastosowaniem **kryterium Mojeny**, by zobrazować sugerowany podział na grupy:

Wykres 12: Dendrogram dla kombinacji 1.



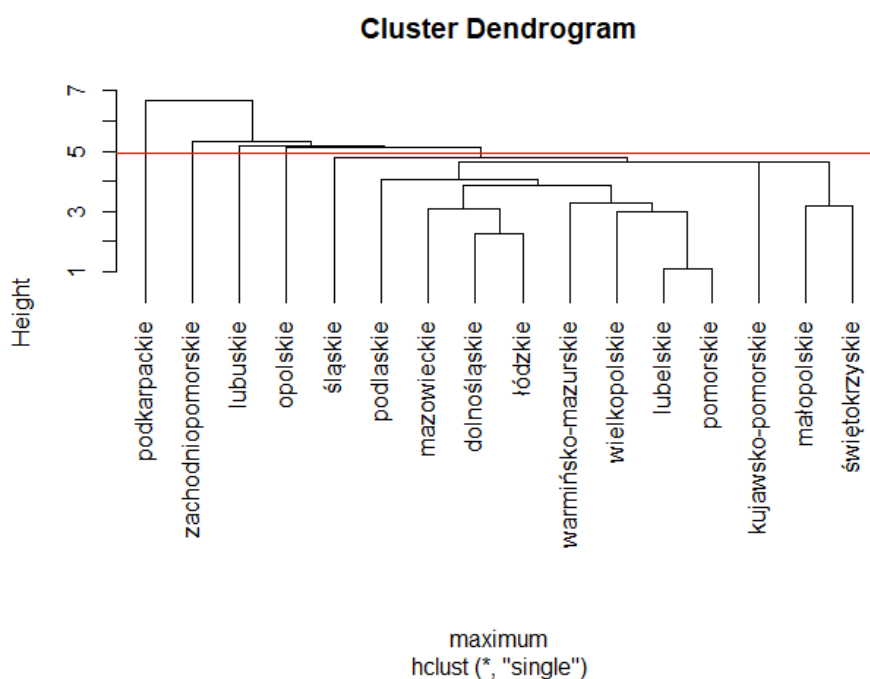
Źródło: Opracowanie własne w programie R

Wykres 13: Dendrogram dla kombinacji 2.



Źródło: Opracowanie własne w programie R

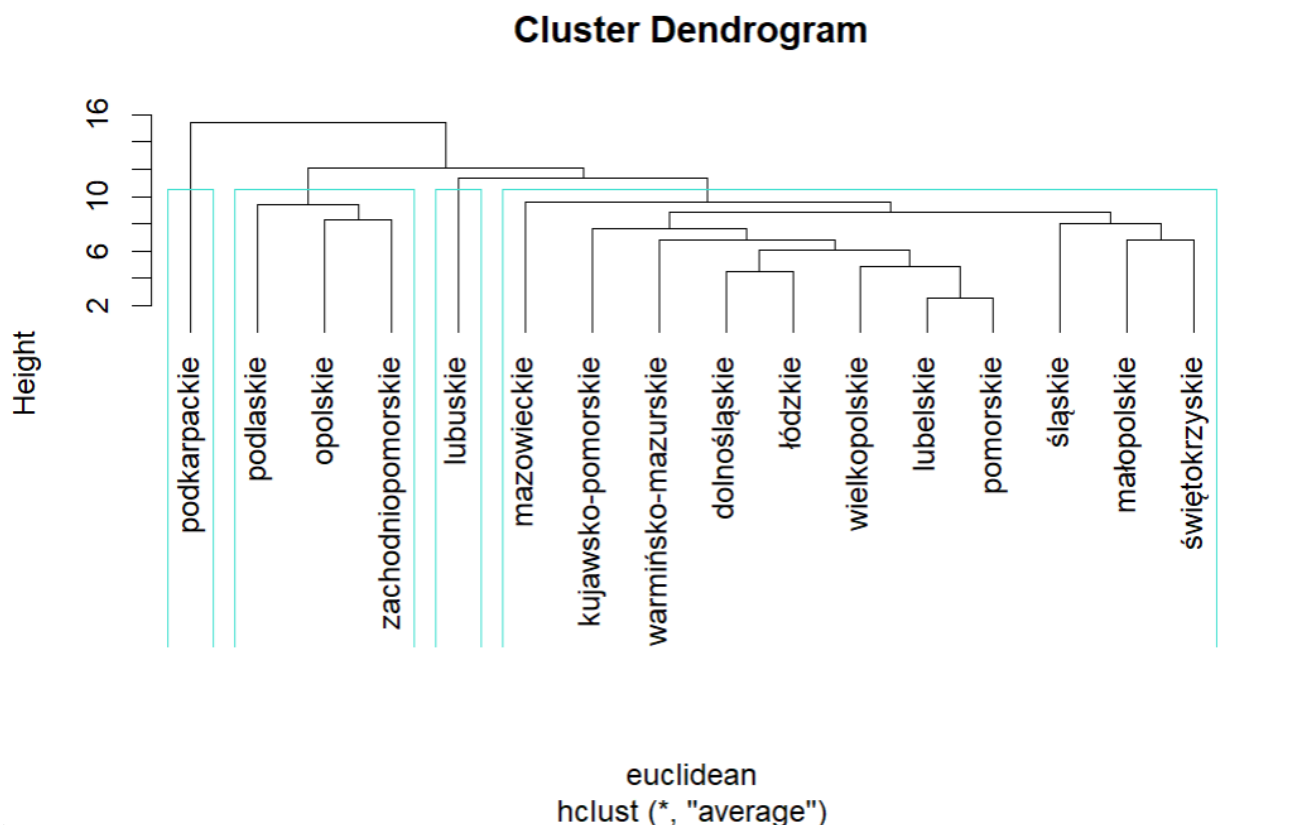
Wykres 14: Dendrogram dla kombinacji 3.



Źródło: Opracowanie własne w programie R

Zastosowano kryterium Mojeny o współczynniku $\alpha = 0.7$ i wyznaczoną w ten sposób dla każdego dendrogramu wartość oznaczono czerwoną kreską. Między 1 a 2 sposobem grupowania widać bardzo duże podobieństwo, jednak **dendrogram 2** dla metryki dominacji z metodą średniej grupowej odrobinę wyraźniej oddziela grupy. Zdecydowano się na sposób grupowania przedstawiony na **wykresie 15**.

Wykres 15: grupowanie województw zgodnie z metodą metryki euklidesowej z metodą średniej grupowej, $k = 4$



Źródło: Opracowanie własne w programie R

Utworzono 4 grupy województw:

- A.** Mazowieckie, kujawsko-pomorskie, warmińsko-mazurskie, dolnośląskie, łódzkie, wielkopolskie, lubelskie, pomorskie, śląskie, małopolskie, świętokrzyskie
- B.** Lubuskie
- C.** Podlaskie, opolskie, zachodniopomorskie
- D.** Podkarpackie

Tworząc w ten sposób **Skupienie A**, **Skupienie B**, **Skupienie C** i **Skupienie D**.

W macierzy przedstawionej w **tabeli 14** przedstawiono średnie wartości zmiennych w każdym ze skupień. **Na ilustracji 4** zobrazowano podział na mapie polski, a **wykresy 16 i 17** przedstawiają odpowiednio porównanie średnich wartości zmiennych pomiędzy skupieniami oraz porównanie wartości zmiennych między skupieniami i średnią globalną każdej zmiennej.

Tabela 14: Macierz średnich wartości zmiennych w każdym ze skupień

	skupienie_a	skupienie_b	skupienie_c	skupienie_d
X1	1.96	1.13	1.72	1.46
X2	14.62	11.49	14.32	9.75
X3	7.33	2.66	3.91	6.02
X4	3.99	2.91	3.33	3.35
X5	2.33	2.46	3	1.98
X6	2.54	2.36	2.45	2.61
X10	2.12	2.53	1.8	1.16
X11	2.49	4.65	3.63	2.78
X12	10.5	8.28	10.48	8.57
X13	9.84	7.79	18.16	16.8
X15	2.41	2.92	2.77	2.07
X17	2.82	3	2	1
X18	11.46	4.05	13.49	23.34

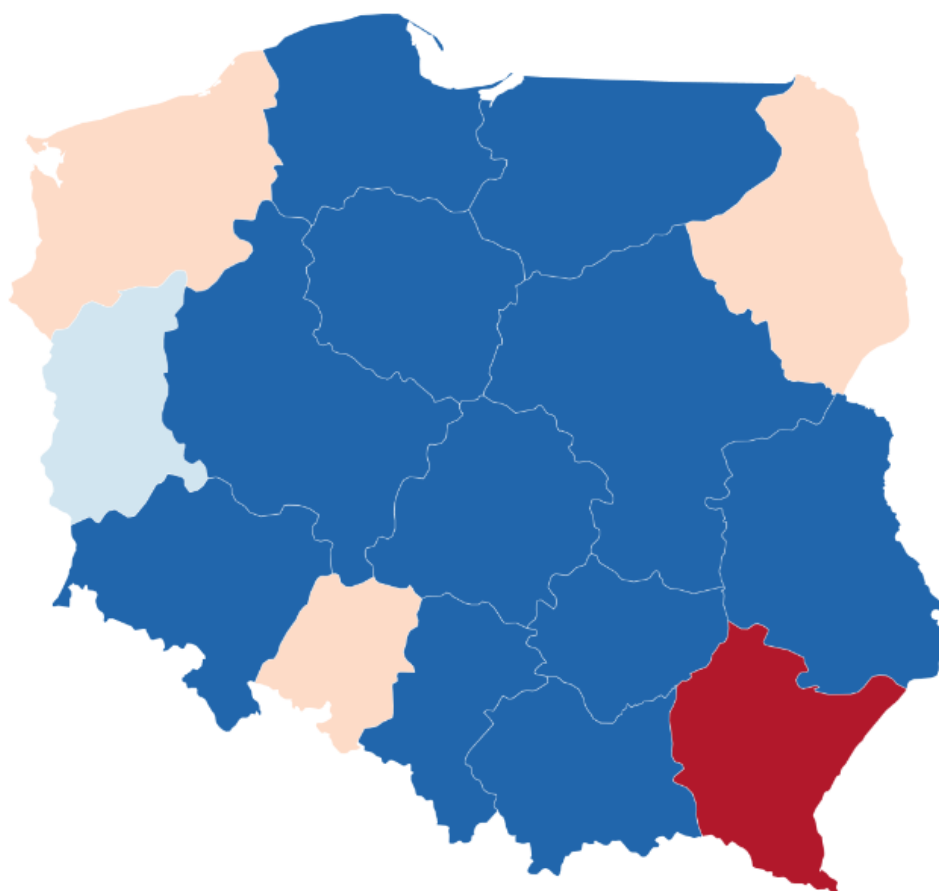
Źródło: Opracowanie własne, obliczenia wykonane w programie R, prezentacja w programie Excel

Ilustracja 4: mapa Polski z podziałem na skupienia

Mapa Polski z podziałem na skupienia

Legenda

■ Skupienie A ■ Skupienie B ■ Skupienie C ■ Skupienie D

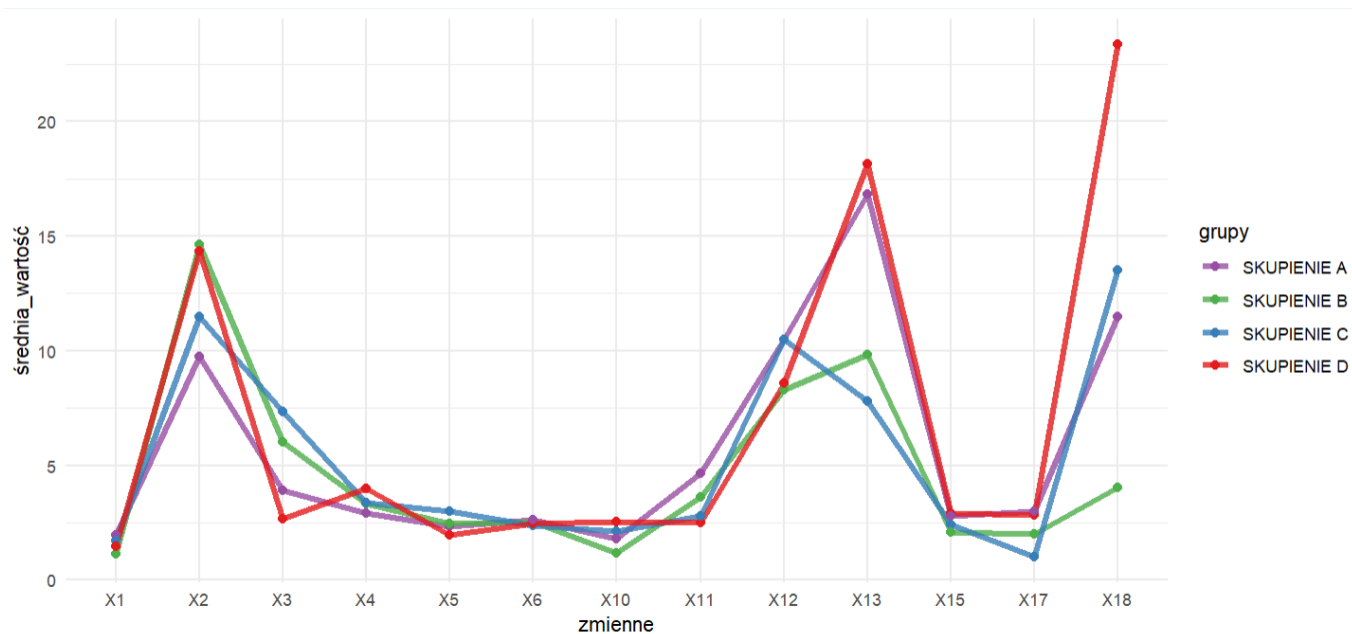


Get the data • Created with Datawrapper



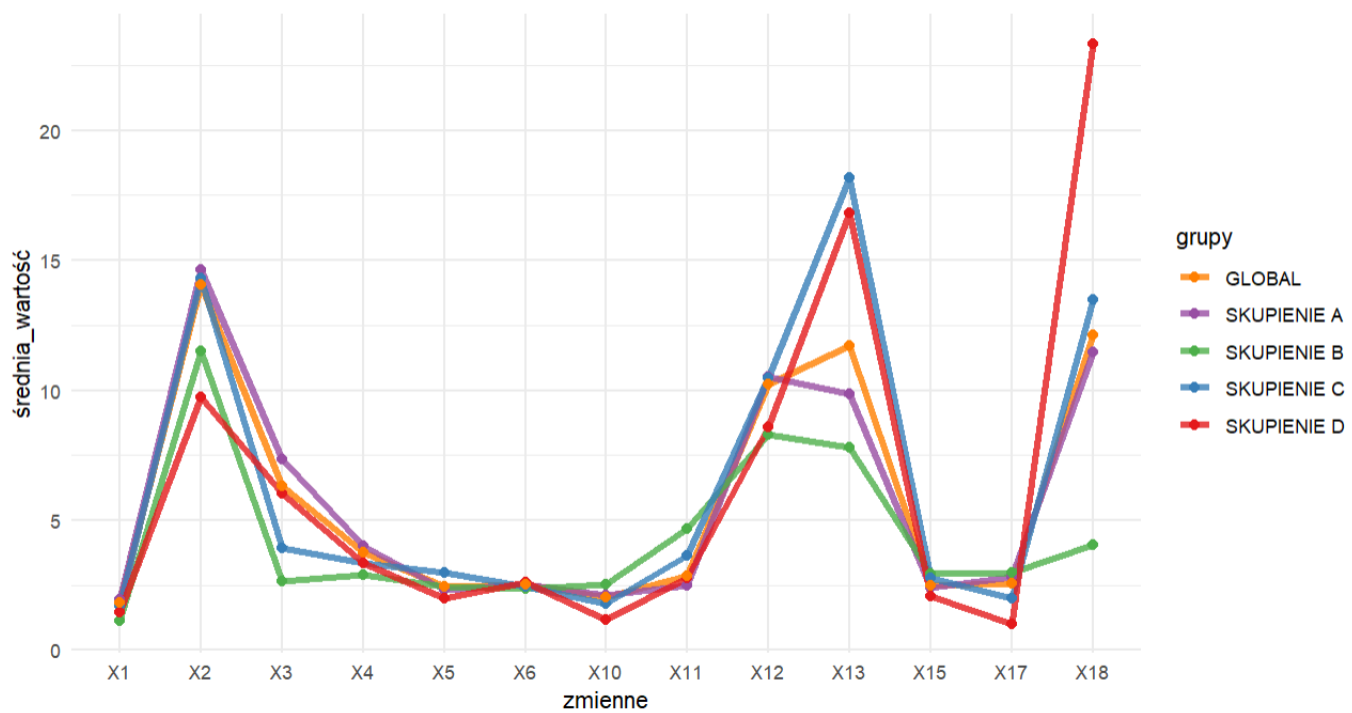
Źródło: Opracowanie własne za pomocą oprogramowania na stronie: <https://app.datawrapper.de/>

Wykres 16: porównanie średnich wartości zmiennych w każdym ze skupień



Źródło: Opracowanie własne w progamie R

Wykres 16: porównanie średnich wartości zmiennych w każdym ze skupień ze średnia globalną



Źródło: Opracowanie własne w progamie R

Wnioski z analizy skupień

Z wykresów można wywnioskować, że najbardziej skupienia różnicują zmienne **X2**, **X3**, **X12**, **X13** i **X18**.

- **X2** - liczba lekarzy anestezjologii i intensywnej terapii na 100 tys. osób
- **X3** - liczba przystanków autobusowych (z trolejbusowymi i tramwajowymi) na 10 km²
- **X12** - ilość lekarzy dentyistów na 10 tys. osób
- **X13** - liczba ośrodków kultury (centrum, domów, ośrodków kultury, klubów i świetlic) na 100 tys. osób
- **X18** - wskaźnik urodzeń martwych na 1000 urodzeń żywych

Na podstawie obserwacji:

- Ilustracja przedstawiająca podział województw na skupienia w dobry sposób pokrywa się z wykresami.
- Woj. Podkarpackie nie różniłoby się tak znacznie, gdyby nie bardzo wysoka statystyka urodzeń martwych na 1000 urodzeń żywych w tym województwie.
- **Skupienie A** reprezentuje najlepsze województwa pod względem cechy badanej

Podsumowanie

Raport przedstawia analizę województw w Polsce z perspektywy młodej osoby zastanawiającej się nad wyborem najlepszego miejsca do założenia rodziny. W analizie uwzględniono kluczowe czynniki wpływające na jakość życia rodzin z dziećmi, w tym dostęp do wysokiej jakości opieki zdrowotnej, rozwiniętą infrastrukturę edukacyjną i rekreacyjną, bezpieczeństwo oraz aspekty ekonomiczne, takie jak średnie ceny mieszkań.

W celu przeprowadzenia kompleksowej oceny zastosowano zaawansowane metody statystyczne, takie jak analiza kanoniczna, analiza głównych składowych (PCA), metody porządkowania liniowego oraz analiza skupień.

- **Analiza kanoniczna** wykazała silne zależności pomiędzy dostępem do opieki medycznej a liczbą dzieci w danym województwie.
- **Analiza głównych składowych (PCA)** pozwoliła na redukcję liczby zmiennych do trzech głównych składowych, które najlepiej opisują badane zależności.
- **Metody porządkowania liniowego** umożliwiły stworzenie rankingu województw pod kątem ich atrakcyjności dla rodzin, natomiast analiza skupień pomogła zidentyfikować grupy województw o podobnych charakterystykach.
- **Analiza skupień** wyróżniła grupę województw, która jest najlepsza pod względem warunków do zakładania rodziny. Jednocześnie pokazała, które cechy najbardziej różnicują polskie województwa

Efektem przeprowadzonych przez nas badań była odpowiedź na pytanie: „W jakim województwie najlepiej założyć rodzinę?”. Wyniki wskazują, że najlepszym województwem do założenia rodziny jest **województwo małopolskie**, które wyróżnia się **najlepszą opieką medyczną, wysoką jakością życia** oraz **dobrze rozwiniętą infrastrukturą**. Wśród najbardziej atrakcyjnych regionów znalazły się również **województwa mazowieckie, śląskie i dolnośląskie**, które także oferują szeroki dostęp do **usług medycznych, edukacyjnych i rekreacyjnych**. Wśród najgorszych regionów pod względem cechy badanej znalazły się województwa **podkarpackie, zachodniopomorskie, podlaskie i opolskie**.

Przeprowadzone badania dostarczają cennych informacji dla osób planujących osiedlenie się w Polsce oraz mogą stanowić podstawę do dalszych analiz związanych z polityką prorodziną i rozwojem regionalnym.

Bibliografia:

A. Balicki, 2013, „Statystyczna analiza wielowymiarowa i jej zastosowania społecznoekonomiczne”.

J.F. Hair Jr., W.C. Black, B.J. Babin, R.E. Anderson, 2019, „Multivariate Data Analysis”, Eighth Edition, Cengage Learning