Projekt z porządkowania liniowego i analizy skupień

Jakub Zięba

Spis treści

[Cel pracy 1](#_Toc153483211)

[Opis i analiza wybranych zmiennych 1](#_Toc153483212)

[Wybrane techniki porządkowania liniowego 8](#_Toc153483213)

[Metoda standaryzowanych sum 9](#_Toc153483214)

[Metoda sumy rang 10](#_Toc153483215)

[Metoda Hellwiga 11](#_Toc153483216)

[Metoda TOPSIS 12](#_Toc153483217)

[Analiza wyników porządkowania liniowego 14](#_Toc153483218)

[Analiza skupień 15](#_Toc153483219)

[Grupowanie podziałowe 15](#_Toc153483220)

[Określanie liczby grup – metoda łokcia 15](#_Toc153483221)

[Określanie liczby grup – metoda profilu 16](#_Toc153483222)

[Metoda k-średnich 17](#_Toc153483223)

[Metoda k-medoid 17](#_Toc153483224)

[Grupowanie hierarchiczne 17](#_Toc153483225)

[Analiza wyników analizy skupień 18](#_Toc153483226)

[Metoda k-średnich 18](#_Toc153483227)

[Metoda k-medoid 20](#_Toc153483228)

[Grupowanie hierarchiczne 23](#_Toc153483229)

# Cel pracy

W trakcie wykonywania projektu postaram się przyjrzeć różnym województwom w Polsce pod względem zmiennych związanych z szeroko pojętym dostępem do medycyny oraz jej poziomem. Celem projektu jest wyłonienie najlepszych województw pod względem poziomu i dostępu do medycyny na podstawie danych zebranych przez Główny Urząd Statystyczny.

# Opis i analiza wybranych zmiennych

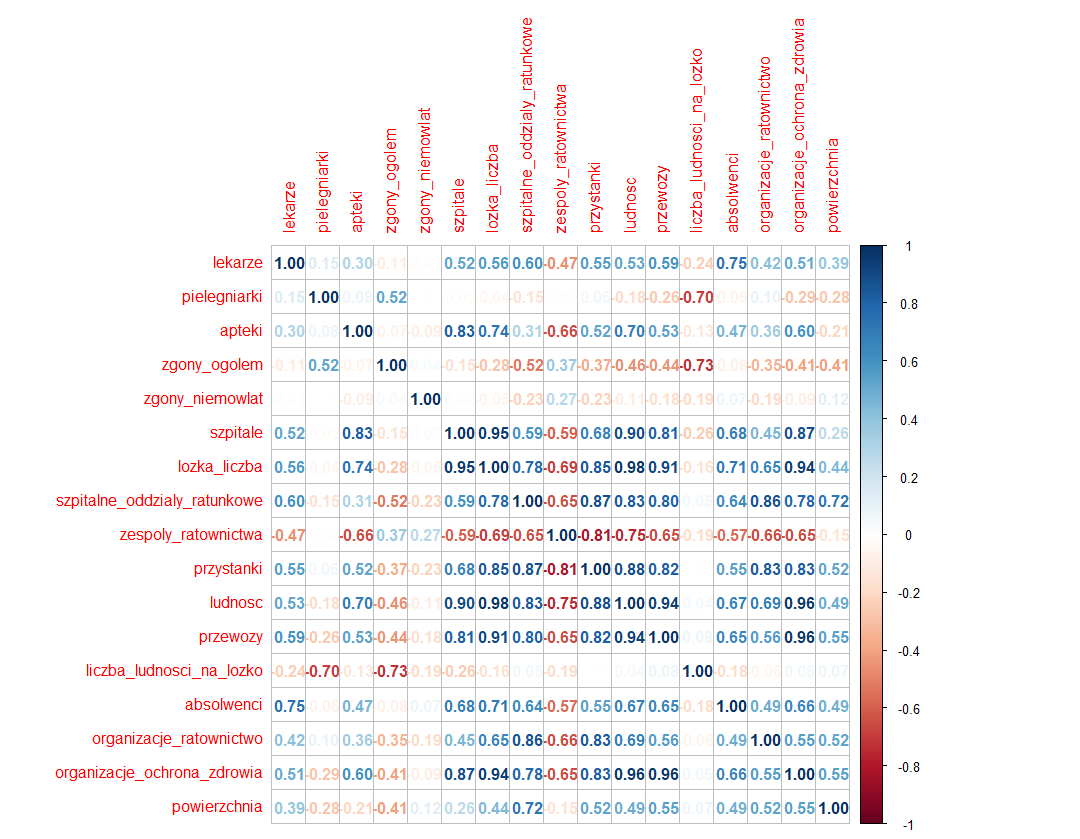
Badanie nad danymi rozpoczynam od wybrania dość dużej próbki danych. W tym przypadku jest to 17 zmiennych:

1. Lekarze, wyrażona w liczbie osób;
2. Pielęgniarki, wyrażona w liczbie osób;
3. Apteki, wyrażona w liczbie takich obiektów;
4. Zgony ogółem, wyrażona w liczbie zgonów na tysiąc urodzeń żywych;
5. Zgony niemowląt, wyrażona w liczbie zgonów niemowląt na tysiąc urodzeń żywych;
6. Szpitale, wyrażona w liczbie obiektów;
7. Łóżka szpitalne, wyrażona w liczbie łóżek;
8. Szpitalne oddziały ratunkowe, wyrażona w liczbie oddziałów;
9. Zespoły ratownictwa medycznego, wyrażona w liczbie takich oddziałów;
10. Przystanki, wyrażona w liczbie przystanków komunikacji miejskiej;
11. Ludność, wyrażona w tys. osób;
12. Przewozy pasażerskie, wyrażona w mln osób;
13. Liczba ludności na łóżko szpitalne, wyrażona w liczbie osób na jedno łóżko szpitalne w szpitalach ogólnych;
14. Absolwenci uczelni medycznych, wyrażona w liczbie osób;
15. Organizacje non profit ratownictwa, wyrażona jako procentowy udział organizacji non profit w ogóle organizacji;
16. Organizacje non profit ochrona zdrowia, wyrażona jako procentowy udział organizacji non profit w ogóle organizacji;
17. Powierzchnia, wyrażona w km2;

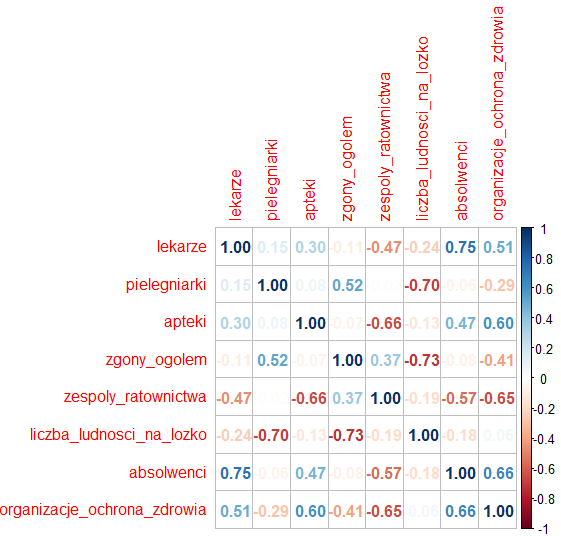
Po przejrzeniu danych można zauważyć duże rozbieżności jeśli chodzi o rzędy wielkości. Z tego powodu, oraz z uwagi na chęć wykluczenia sytuacji, w której najlepiej wypadają województwa zwyczajnie największe, dane poddałem obróbce. Obrobione zmienne prezentują się następująco:

1. Lekarze, wyrażona w liczbie osób na tys. ludności;
2. Pielęgniarki, wyrażona w liczbie osób na tys. ludności;
3. Apteki, wyrażona w liczbie obiektów na km2;
4. Zgony ogółem, wyrażona w liczbie zgonów na tysiąc urodzeń żywych;
5. Zgony niemowląt, wyrażona w liczbie zgonów niemowląt na tysiąc urodzeń żywych;
6. Szpitale, wyrażona w liczbie obiektów;
7. Łóżka szpitalne, wyrażona w liczbie łóżek;
8. Szpitalne oddziały ratunkowe, wyrażona w liczbie oddziałów;
9. Zespoły ratownictwa medycznego, wyrażona w liczbie zespołów na tys. ludności;
10. Przystanki, wyrażona w liczbie przystanków komunikacji miejskiej;
11. Ludność, wyrażona w tys. osób;
12. Przewozy pasażerskie, wyrażona w mln osób;
13. Liczba ludności na łóżko szpitalne, wyrażona w liczbie osób na jedno łóżko szpitalne w szpitalach ogólnych;
14. Absolwenci uczelni medycznych, wyrażona w liczbie osób;
15. Organizacje non profit ratownictwa, wyrażona jako procentowy udział organizacji non profit w ogóle organizacji;
16. Organizacje non profit ochrona zdrowia, wyrażona jako procentowy udział organizacji non profit w ogóle organizacji;
17. Powierzchnia, wyrażona w km2;

Następnie zajmuję się wyborem odpowiednich zmiennych. Dokonuję tego najpierw przez zbadanie współczynników korelacji między zmiennymi.



Aby wyeliminować zmienne bardzo mocno skorelowane, zdecydowałem się nie uwzględniać w zbiorze zmiennych odpowiedzialnych za liczbę szpitali, liczbę łóżek, liczbę szpitalnych oddziałów ratunkowych, liczbę przewozów pasażerskich, odsetek organizacji non profit związanych z ratownictwem, liczbę przystanków oraz liczbę zgonów niemowląt. Nie uwzględniam również ludności ani powierzchni, jako że te zmienne zostały wcześniej wykorzystane do wyliczenia innych zmiennych.



Po zmianie zbioru zmiennych macierz korelacji prezentuje się dużo lepiej, nie występują już wartośći korelacji na poziomie wyższym niż 0,9. Dla wybranych zmiennych weryfikuję również współczynniki zmienności.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

Każdy ze współczynników zmienności jest na poziomie co najmniej 0.1, zatem możemy przyjąć taki zbiór danych do dalszej pracy.

Następnie skupimy się na podstawowych statystykach opisowych wybranego zbioru zmiennych, w stosunku do całości kraju.

Obraz zawierający tekst, zrzut ekranu, czarne, Czcionka

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

W przypadku zmiennej lekarze, średnia w całym kraju wynosi ok. 3,9 lekarza na tysiąc ludności, przy odchyleniu standardowym na poziomie ok. 0,77. Jest to dość wysokie odchylenie standardowe względem średniej. Mediana natomiast jest na poziomie lekko wyższym od średniej, co może być dobrym znakiem jeśli chodzi o dostęp do lekarzy w kraju. Wartość minimalną zmienna przyjmuje dla województwa Lubuskiego, a maksymalną dla Mazowieckiego.

Jeśli chodzi o pielęgniarki, to tutaj średnia kształtuje się na poziomie 8.17, natomiast odchylenie standardowe na poziomie 1.12, nie jest to więc aż tak duża wartość względem średniej. Podobnie jak w przypadku zmiennej powyżej, średnia i mediana są do siebie mocno zbliżone. Wartość minimalną zmienna przyjmuje w przypadku województwa Wielkopolskiego, a maksymalną dla województwa Świętokrzyskiego.

W przypadku aptek, średnia i mediana również są do siebie zbliżone. Odchylenie standardowe na poziomie 0.02 jest też bardzo bliskie średniej, która wynosi około 0,04. Najmniej aptek na kilometr kwadratowy możemy znaleźć w województwie Warmińsko-Mazurskim, najwięcej natomiast w województwie Śląskim.

Dla zgonów ogółem na tysiąc urodzeń żywych wartość minimalną znajdujemy w województwie Pomorskim, od którego bardzo blisko plasuje się województwo Małopolskie. Wartość maksymalną natomiast znajdujemy przy województwie Świętokrzyskim.

Jeśli chodzi o zespoły ratownictwa medycznego oraz ludność na łóżko szpitalne, to ich współczynniki zmienności są na podobnym poziomie, odpowiednio 15% i 12%, co nie różni się znacząco od 13% współczynnika zmienności dla pielęgniarek.

Prawdopodobnie najciekawszą zmienną będą tu absolwenci uniwersytetów medycznych. Jako że nie w każdym województwie mamy uczelnie medyczną, zarówno odchylenie standardowe jak i minimum zachowują się dość ekstremalnie. Odchylenie standardowe minimalnie przekracza średnią, a minimum jest równe zero i obejmuje wiele województw. Wartość maksymalna w wysokości prawie 2300 absolwentów znalazła się w województwie Śląskim.

W przypadku ostatniej zmiennej, odsetka organizacji non profit wśród organizacji ochrony zdrowia współczynnik zmienności również ma wysoką wartość. Rozpiętość danych również jest bardzo wysoka, jako że wartość minimalna wynosi jedynie 1.6%, natomiast maksymalna aż 18%.

Obraz zawierający linia, diagram, Wykres, tekst

Opis wygenerowany automatycznie

Obraz zawierający linia, tekst, Wykres, diagram

Opis wygenerowany automatycznie

Wykresy gęstości przedstawione powyżej pozwalają nam zobaczyć podstawową strukturę rozkładu danych. W większości przypadków wartości są raczej dość mocno skupione wokół średniej, jednak w żadnym z nich nie można mówić o rozkładzie stricte normalnym. Zdarzają się również rozkłady o grubych ogonach, szczególnie jeśli chodzi o prawe ogony. W przypadku lewych ogonów taka sytuacja nie występuję.

Obraz zawierający tekst, diagram, Plan, Równolegle

Opis wygenerowany automatycznie

Jeśli chodzi o wartości odstające, to sprawdzając pojedyncze zmienne, znajdujemy takie wartości w przypadku zmiennych dotyczących organizacji non profit, zespołów ratownictwa, ludności na łóżko szpitalne oraz aptek. Nie jesteśmy natomiast w stanie stwierdzić na tej podstawie, czy któryś z obiektów jest outlierem w przypadku analizy wielu zmiennych jednocześnie.

# Wybrane techniki porządkowania liniowego

Pierwszym krokiem w porządkowaniu liniowym jest zamiana zmiennych na stymulanty. Zmienne określiłem jako:

1. Lekarze – stymulanta;
2. Pielęgniarki – stymulanta;
3. Apteki – stymulanta;
4. Zgony ogółem – destymulanta;
5. Zespoły ratownictwa medycznego – stymulanta;
6. Ludność na łóżko szpitalne – destymulanta;
7. Absolwenci uczelni medycznych – stymulanta;
8. Organizacje non profit dot. ochrony zdrowia – nominanta na poziomie 9%.

W takim wypadku musimy zamienić zmienne 5, 6 i 8 na stymulanty. Robimy to mnożąc destymulanty przez -1 oraz stosując następujący wzór dla nominanty:

Obraz zawierający tekst, Czcionka, pismo odręczne, biały

Opis wygenerowany automatycznie

Następnym krokiem jest standaryzacja danych. Również wykonuje się ją dla większości metod, jednak dla metody standaryzowanych sum i Hellwiga stosuje się inny wzór niż dla metody TOPSIS, dla metody sumy rang standaryzację pomijamy.

W każdym przypadku używamy wersji metody z uwzględnieniem następujących wag:

* Lekarze – 0,25;
* Pielęgniarki – 0,1;
* Apteki – 0,1;
* Zgony ogółem – 0,2;
* Zespoły ratownictwa medycznego – 0,075;
* Ludność na łóżko szpitalne – 0,075;
* Absolwenci uczelni medycznych – 0,15;
* Organizacje non profit dot. ochrony zdrowia – 0,05.

Największą uwagę będziemy w takim razie przywiązywać do zmiennych dotyczących lekarzy, zgonów i absolwentów uczelni medycznych. Najmniej ważne będą natomiast dane dotyczące organizacji non profit.

## Metoda standaryzowanych sum

Wykonujemy standaryzację według wzoru: Obraz zawierający Czcionka, biały, diagram, symbol

Opis wygenerowany automatycznie

Dane po standaryzacji wyglądają następująco:

Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

Następnie wyliczamy i standaryzujemy syntetyczną miarę z uwzględnieniem wag, która pozwoli nam wybrać najlepsze z województw. Robimy to według wzoru:

Obraz zawierający Czcionka, biały, tekst, typografia

Opis wygenerowany automatycznieObraz zawierający tekst, Czcionka, biały, linia

Opis wygenerowany automatycznie

Wynikiem tych działań jest następująca tablica danych:

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatyczniePozostało nam jedynie posortować województwa według miary wyliczonej w kolumnie S2:

## Metoda sumy rang

W metodzie sumy rang, po zamianie zmiennych na stymulanty, musimy nadać obiektom rangi w obrębie poszczególnych zmiennych. Rangi te są rosnące, tj. najmniejsza ranga oznacza najlepszą zmienną.

Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

Następnie uwzględniając wagi sumujemy rangi dla poszczególnych województw. Tak otrzymaną tabelę sortujemy według ww. sum. Najmniejsza suma oznacza obiekt najlepszy.

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

## Metoda Hellwiga

W metodzie Hellwiga, po zamianie zmiennych na stymulanty, dokonujemy standaryzacji:

Obraz zawierający Czcionka, biały, diagram, symbol

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

Następnie tworzymy wzorzec, tzn. najlepszy obiekt, posiadający maksymalną wartość w przypadku każdej zmiennej:

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznie

Teraz obliczamy odległości od poszczególnych badanych obiektów do wyżej wyliczonego wzorca oraz „możliwie daleką” odległość, której użyjemy do wyliczenia miar dla obiektów.

Obraz zawierający pismo odręczne, Czcionka, kaligrafia, biały

Opis wygenerowany automatycznieObraz zawierający Czcionka, biały, typografia, kaligrafia

Opis wygenerowany automatycznie

Kiedy już wyliczymy odległości oraz odległość d0 na ich podstawie, wykorzystujemy następujący wzór aby wyliczyć kolejną miarę dla obiektów:

Obraz zawierający Czcionka, biały, numer, design

Opis wygenerowany automatycznie

Na podstawie wyliczonej miary sortujemy nasze województwa:

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

## Metoda TOPSIS

W przypadku metody TOPSIS, również standaryzujemy dane zamienione na symulanty, jednak robimy to według innego wzoru:

Obraz zawierający Czcionka, biały, diagram, pismo odręczne

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, menu, Czcionka

Opis wygenerowany automatycznie

Następnie, podobnie jak w metodzie Hellwiga, wyznaczamy wzorzec, czyli obiekt składający się z wartości najlepszych. Jednak dodatkowo tworzymy też antywzorzec, czyli obiekt, który charakteryzują wartości najgorsze.

Obraz zawierający tekst, Czcionka, biały, design

Opis wygenerowany automatycznie

Następnie na podstawie wzorca i antywzorca, obliczamy odległości obiektów od ww. tworów.

Obraz zawierający Czcionka, pismo odręczne, tekst, biały

Opis wygenerowany automatycznie

Wykorzystując wyżej wyliczone odległości, wyliczamy miarę współczynnika rankingowego, według którego sortujemy nasze województwa w poszukiwaniu najlepszego:

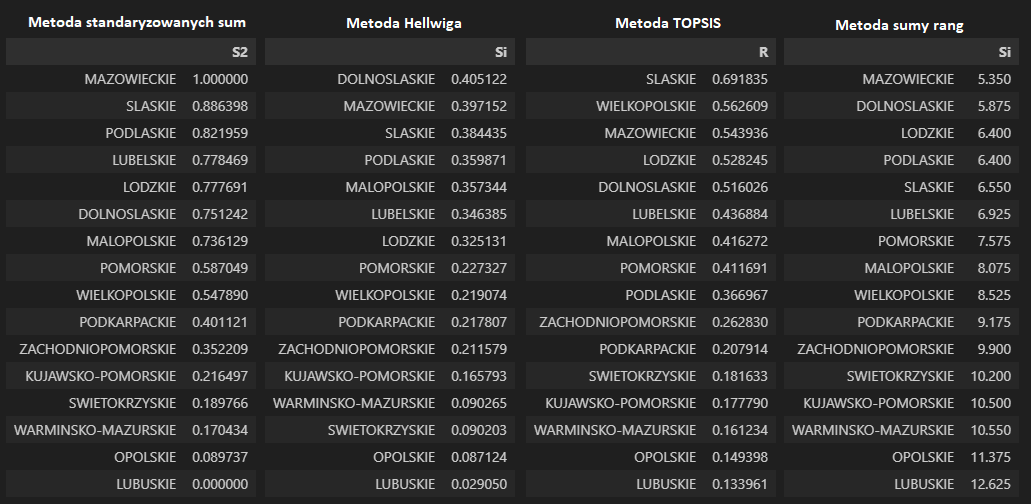
Obraz zawierający Czcionka, diagram, biały, design

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

# Analiza wyników porządkowania liniowego



Wyniki różnych metod porządkowania liniowego różnią się między sobą, zauważamy jednak pewne zależności, które powtarzają się w różnych metodach. Jest to szczególnie widoczne na końcach rankingów.

Województwa Lubuskie i Opolskie zawsze plasują się na samym końcu rankingu. Mają one najniższe wartości jeśli chodzi o liczbę lekarzy, wysokie liczby zgonów, dość wysoką liczbę ludności na łóżko, są również bardzo daleko od nominanty w przypadku organizacji. Są to też województwa, które nie posiadają uniwersytetu medycznego.

Następne od końca są województwa Warmińsko-Mazurskie, Świętokrzyskie i Kujawsko-Pomorskie, które zamieniają się między sobą ale zajmują zawsze od trzeciego do piątego z ostatnich miejsc.

U góry rankingu również mamy kilka stałych obiektów. Są to województwa Mazowieckie i Śląskie, które zawsze pojawiają się w top 5, czy województwa Podlaskie i Łódzkie, które w większości metod również plasują się wysoko.

Szczególnie wyróżnia się województwo Mazowieckie, dwukrotnie będąc na pierwszym miejscu. Wartości dla tego województwa w przypadku każdej zmiennej są dość wysokie, lub niskie w przypadku destymulant. Prawdopodobnie największym minusem województwa Mazowieckiego jest bardzo wysoki odsetek organizacji non profit, ponad dwukrotnie wyższy od pożądanej wartości tej nominanty.

Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

# Analiza skupień

Aby zająć się analizą skupień, dane muszą spełniać warunki dotyczące współczynników korelacji między zmiennymi oraz współczynników zmienności. Współczynniki korelacji nie powinny przekraczać na moduł wartości 0.9, natomiast współczynniki zmienności powinny mieć wartość przynajmniej 0,1. Zostało to natomiast sprawdzone już we wcześniejszej części projektu i wybrane zmienne spełniają te założenia.

## Grupowanie podziałowe

W przypadku grupowanie podziałowego, musimy najpierw określić na jaką liczbę grup chcemy podzielić nasze obiekty. Aby tego dokonać można skorzystać m.in. z metody łokcia oraz metody profilu.

### Określanie liczby grup – metoda łokcia

Metoda łokcia pozwala dobrać liczbę grup w taki sposób, żeby zminimalizować zmienność wewnątrz grup. W metodzie łokcia wynik interpretujemy poprzez analizę wykresu. Szukamy takiego punktu na osi X, aby zmiana wartości na osi Y była niewielka, czyli aby linia łącząca punkty była spłaszczona.

Obraz zawierający linia, Wykres, diagram, Równolegle

Opis wygenerowany automatycznie

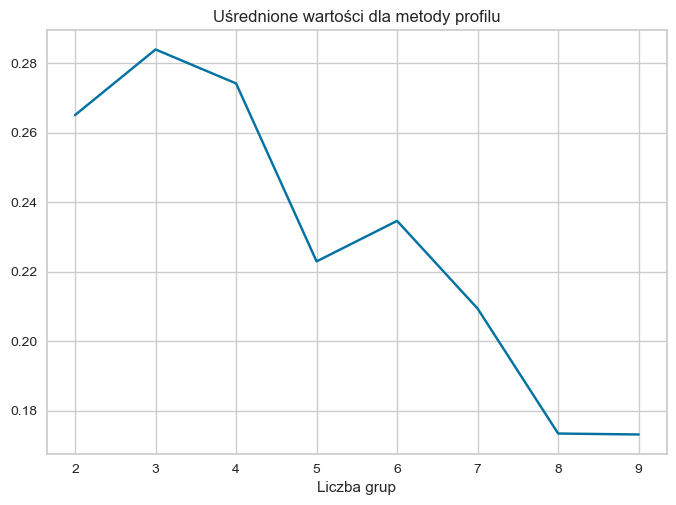
Na przedstawionym wykresie widzimy, że zmiana wartości na osi Y jest duża pomiędzy 1 a 2 na osi X. Również pomiędzy 2 a 3 wykres dość stromo maleje. Natomiast pomiędzy 3 a 4 jest już bardziej wypłaszczony, tak jak między 4 a 5 i dalej. Szukamy jednak liczby możliwie małej, ponieważ dla dużej liczby grup zmienność wewnątrzgrupowa zawsze będzie maleć, ale zbyt małe grupy nie pozwolą nam de facto stworzyć sensownych grup z podobnych sobie obiektów. Z tego powodu na podstawie metody łokcia wybrałbym podział na 3 lub 4 grupy.

### Określanie liczby grup – metoda profilu

W metodzie profilu poza zmiennością wewnątrzgrupową uwzględniana jest również odległość między grupami. Dla każdego obiektu podlegającego podziałowi obliczany jest współczynnik, pozwalający ocenić czy obiekt został dobrze przypisany. Współczynnik ten przyjmuje wartości w zakresie od -1 do 1 i jego interpretacja jest następująca:

* wartości bliskie -1 oznaczają błędne przypisanie, obiekt lepiej pasuje do innej grupy;
* wartości bliskie 0 oznaczają, że obiekt jest na granicy dwóch grup;
* wartości bliskie 1 oznaczają poprawne przypisanie do grupy.

Oblicza się również uśrednione wartości współczynników dla danej grupy, co pozwala ocenić jakość dopasowania w całej grupie.



W przypadku metody profilu chcemy maksymalizować wartość współczynników, dlatego szukamy maksimum funkcji. Na wykresie widać, że maksimum otrzymujemy przy liczbie grup równej 3.

Biorąc pod uwagę obie metody można dojść do wniosku, że dobrym wyborem może być podzielenie zbioru na 3 grupy.

### Metoda k-średnich

Najpierw skorzystam z metody k-średnich. Tę metodę można w Pythonie znaleźć w pakiecie scikit learn. Aby zwizualizować wyniki trzeba przekształcić dane z wielowymiarowych do dwuwymiarowych, tak aby dało się je przedstawić na wykresie. Odpowiednie funkcje również znajdujemy w pakiecie scikit learn.

### Metoda k-medoid

Kolejną metodą jest metoda k-medoid. Można ją znaleźć w pakiecie scikit learn extra. Podobnie jak w poprzedniej metodzie, do graficznego przedstawienia podziału na grupy potrzebujemy zeskalować dane do dwóch wymiarów.

## Grupowanie hierarchiczne

W przypadku grupowania hierarchicznego nie musimy określać liczby klastrów, na jaką chcemy podzielić zbiór obiektów na samym początku. Odpowiedni algorytm zrobi to za nas w trakcie podziału. Pierwszym krokiem jest tutaj określenie odległości między obiektami a następnie dokonanie grupowania na ich podstawie, z użyciem wybranej metody. Liczbę grup należy jednak wybrać później. Można tego dokonać przez analizę i odpowiednie ucięcie dendrogramu czy przez analizę indeksów jakości klasyfikacji.

# Analiza wyników analizy skupień

## Metoda k-średnich

Po przeprowadzeniu grupowania metodą k-średnich otrzymujemy następujący podział:

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

Obraz zawierający tekst, diagram, linia, zrzut ekranu

Opis wygenerowany automatycznie

Po sprowadzeniu obiektów do postaci dwuwymiarowej można zobaczyć podział na wykresie.

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

Grupa pierwsza charakteryzuje się najwyższą liczbą lekarzy i pielęgniarek na tysiąc ludności. O ile w przypadku pielęgniarek różnica pomiędzy grupą 0 i 1 nie jest duża, to w przypadku lekarzy jest to różnica znacząca. Również w przypadku aptek grupa 0 wypada najlepiej, natomiast druga w kolejności jest tutaj grupa 2.

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

Jeśli chodzi o zgony ogółem na tysiąc urodzeń żywych to najlepiej wypada tutaj grupa 2, kolejna jest grupa 0. Grupa 0 plasuje się również po środku jeśli chodzi o zespoły ratownictwa medycznego, pierwsza jest grupa 1. W przypadku kolejnej zmiennej, tj. ludności na łóżko szpitalne, wracamy do kolejności znanej nam z liczby pielęgniarek.

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

W przypadku liczby absolwentów grupa 0 ma ich średnio najwięcej i mocno deklasuje dwie pozostałe grupy, co pozwala przypuszczać że jest to grupa z największą liczbą województw posiadających na swoim terenie uniwersytety medyczne. Natomiast żadna z grup nie ma tutaj wartości 0, co oznacza że w każdej grupie znalazł się przynajmniej jeden uniwersytet. Jeśli chodzi o organizacje non profit, to grupy 0 i 2 znajdują się dość blisko nominanty równej 9%, natomiast grupa 1 mocno odstaje w negatywną stronę.

Podsumowując, grupa 0 to grupa w gruncie rzeczy wyglądająca najlepiej na tle pozostałych grup. Patrząc na średnie wartości oraz mediany, grupa ta plasuje się zwykle na pierwszym miejscu według różnych zmiennych, choć zdarzają się miejsca drugie.

Dwie pozostałe grupy wymieniają się pozycjami jeśli chodzi o takie uporządkowanie i trzeba porównywać każdą zmienną osobno, ponieważ inaczej ciężko jest dojść do jednego konkretnego wniosku.

## Metoda k-medoid

W przypadku wykorzystania metody k-medoid wyniki prezentują się następująco:

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

Obraz zawierający tekst, diagram, linia, zrzut ekranu

Opis wygenerowany automatycznie

Na pierwszy rzut oka widać tutaj pewne różnice, tak w składach grup, jak i w ich liczebności. Warto więc zauważyć, że nie można porównywać grup z różnych metod grupowania ze sobą, właśnie z powodu różnych obiektów wchodzących w ich skład jak i ich różnej liczebności.

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

Biorąc pod uwagę średnią liczbę lekarzy, pomiędzy grupami 0 i 2 nie ma wielkich różnic. Stają się one bardziej widoczne w przypadku porównywania median. Na pierwszym miejscu plasuje się grupa 0. Podobnie sytuacja wygląda w przypadku liczby pielęgniarek. Grupa 0 ma również średnio najwięcej aptek na km2 ze wszystkich grup.

Obraz zawierający tekst, zrzut ekranu, Czcionka, menu

Opis wygenerowany automatycznie

Analizując zgony, najlepiej wypada grupa 2, natomiast najgorzej grupa 0. Najwięcej zespołów ratownictwa ma natomiast grupa 1. Jeśli chodzi o liczbę ludności na łóżko szpitalne, tutaj ponownie najlepiej wypada grupa 0.

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

W przypadku liczby absolwentów, widać tutaj mniejsze różnice pomiędzy grupami niż otrzymaliśmy przy metodzie k-średnich. Mimo występującej mocno odstającej grupy 1, dwie pozostałe grupy są do siebie mocniej zbliżone. Na pierwszym miejscu plasuje się grupa 0. Jeśli chodzi o organizacje non profit w sferze ochrony zdrowia, to najbliżej nominanty znajduje się grupa 2, natomiast najbardziej oddalona jest grupa 1.

W tej metodzie również grupa 0 wypada ogólnie bardzo dobrze. Jest to prawdopodobnie związane z dużym podobieństwem grup 0 z obu metod. W metodzie k-metoid grupa 0 jest natomiast mniej liczna i brakuje jej jednego członka względem k-średnich, jednak poza tym jej skład się nie różni. Z tego powodu, podobnie jak poprzednio, pozostałe grupy wymieniają się jeśli chodzi o ich jakość względem poszczególnych zmiennych.

## Grupowanie hierarchiczne

Obraz zawierający tekst, diagram, linia, Równolegle

Opis wygenerowany automatycznie

Wynikiem grupowania hierarchicznego jest natomiast powyższy dendrogram, który postanowiłem uciąć na poziomie 6, tworząc 3 grupy. Dzięki takiemu podziałowi otrzymujemy grupy o sensownej liczebności, nie dostajemy grup o bardzo małej liczbie obiektów. Otrzymane grupy dokładnie odpowiadają grupom wydzielonym z pomocą metody k-średnich.