Statystyczna Analiza danych – projekt 1

## Paweł Brodziak

Poniższa praca ma na celu sklasyfikowanie i zgrupowanie powiatów województwa Mazowieckiego z wykorzystaniem metod analizy skupisk. Wykorzystane zostały metody grupowania podziałowego oraz hierarchicznego oraz został wykonany ranking z użyciem metod porządkownia liniowego. Cała analiza została wykonana przy użycie języka Python.

### Opis danych

Dane pochodzą z 2020 roku i dotyczą powiatów województwa Mazowieckiego z wyłączeniem miast na prawach powiatów.

W przygotowanych danych możemy znaleźć:

* Dochody budżetów powiatów na 1 mieszkańca, podane w zł – oznaczone jako ‘DOCH’
* Gęstość zaludnienia, podana jako ludność na 1 km2 – oznaczona jako ‘LUDN’
* Liczba osób w wieku produkcyjnym – oznaczona jako ‘PROD’
* Przeciętne miesięczne wynagrodzenia brutto – oznaczone jako ‘WYN’
* Liczba osób bezrobotnych pozostających bez pracy dłużej niż rok – oznaczona jako ‘BEZR’
* Liczba mieszkań oddanych do użytkowania – oznaczona jako ‘MIESZ’
* Powierzchnia parków spacerowo-wypoczynkowych, w ha – oznaczona jako ‘ZIEL’
* Długość ścieżek rowerowych, podana w km – oznaczona jako ‘ROW’

Na ryc.1 przedstawiono podstawowe statystyki dla każdego atrybutu.

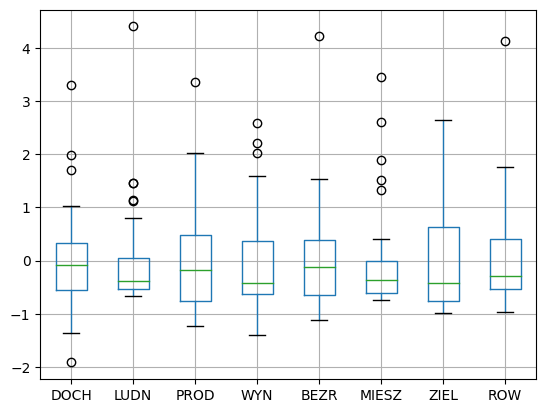
Możemy zauważyć wysokie wartości odchylenia standardowego dla gęstości zaludnienia (najpewniej wynikająca z wysokiej wartości maksymalnej), dla liczby mieszkań oddanych do użytku oraz ścieżek rowerowych.

Obraz zawierający tekst, tablica wyników

Opis wygenerowany automatycznie

Ryc. Podstawowe statystyki każdego atrybutu dla badanych powiatów

Dla wszystkich atrybutów wartości maksymalne są znacznie bardziej oddalone od wartości średnich co możemy zweryfikować spoglądając na wykresy pudełkowe badanych atrybutów dla znormalizowanych danych (Ryc.2).



Ryc. Wykres pudełkowy odchyleń badanych atrybutów

Wykres pudełkowy potwierdza teorię o wielu wysokich wartościach odstających, znacznie wyższych niż średnia dla prawie wszystkich atrybutów.

Praktycznie nie występują obiekty o wartościach odstających znacznie niższych niż średnia.

Na rycinie 3 przedstawiono macierz korelacji badanych atrybutów.

Można zauważyć wyraźną silnie dodatnią korelację między ilością oddanych mieszkań a liczbą osób w wieku produkcyjnym oraz między oddanych mieszkań a gęstością zaludnienia. Gęstość zaludnienia jest silnie dodatnie skorelowana z niemal wszystkimi pozostałymi cechami oprócz średnich dochodów, które to są negatywnie skorelowane z wszystkimi pozostałymi atrybutami .

Obraz zawierający tekst, tablica wyników, zrzut ekranu, kilka

Opis wygenerowany automatycznie

Ryc. Macierz korelacji wszystkich atrybutów

### Porządkowanie liniowe

W celu stworzenia rankingu powiatów z użyciem porządkowania liniowego, została użyta metoda Hellwiga.

Bezrobocie zostało uznano za destymulantę, a gęstość zaludnienia za nominantę i została zamieniona na stymulantę.

Po wyznaczeniu odchyleń i odległości obiektów od wzorca otrzymano listę współczynników Hellwiga które pozwoliły na ułożenie rankingu powiatów widocznego w Tab.1 (wersja skrócona) – pełna wersja dostępna jest w załączniku csv oraz w kodzie źródłowym analizy.

Tabela . Pierwsze 20 powiatów rankingu wynikającego z porządkowania liniowego

Obraz zawierający tekst, tabliczka, tablica wyników

Opis wygenerowany automatycznie

Jak można zauważyć w powyższej tabeli, powiaty Warszawski Zachodni, Pruszkowski i Piaseczyński są najbliższe wzorca obliczonego na podstawie przyjętych stymulantów – cech pożądanych. Wskaźnik ten jednak nie jest wysoki, co oznacza, że żaden powiat nie spełnia wszystkich założeń w wysokim stopniu.

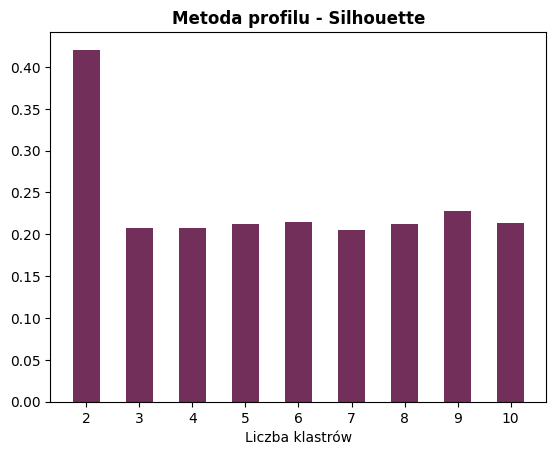
### Analiza skupień

W celu łatwiejszego zwizualizowania skupień cechy powiatów zostały zdekomponowane do 2 atrybutów przy użycie metody PCA (Principal Component Analysis).

Nowe atrybuty tłumaczą wariancję składowych wartości łącznie w 64.84% co jest satysfakcjonującym wynikiem przy redukcji atrybutów z 8 do 2.

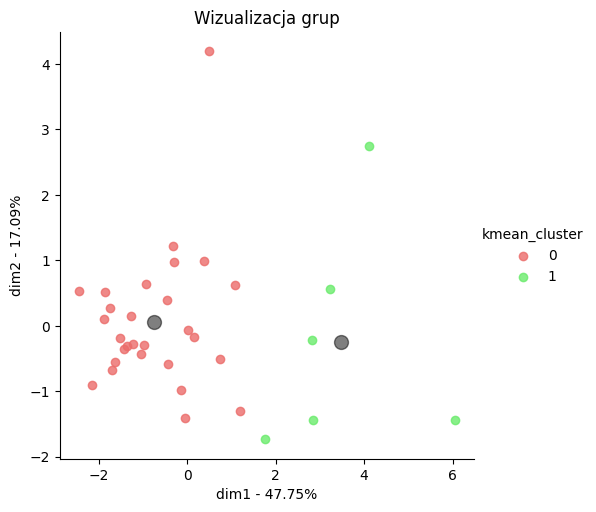
Skupienia zostały zidentyfikowane przy użyciu grupowania przedziałowego - metody k-średnich oraz przy użyciu grupowania hierarchicznego – metody Warda.

Do znalezienia optymalnej liczby klastrów (grup) dla obu przypadków została wykorzystana metoda Silhoutte – metoda profilu, która oblicza współczynnik dla potencjalnej ilości grup. Jak widoczne na poniższym wykresie (ryc.4), optymalną wartością w tym przypadku było 2. Na tej podstawie obie metody grupowania wskazały 2 grupy obiektów.

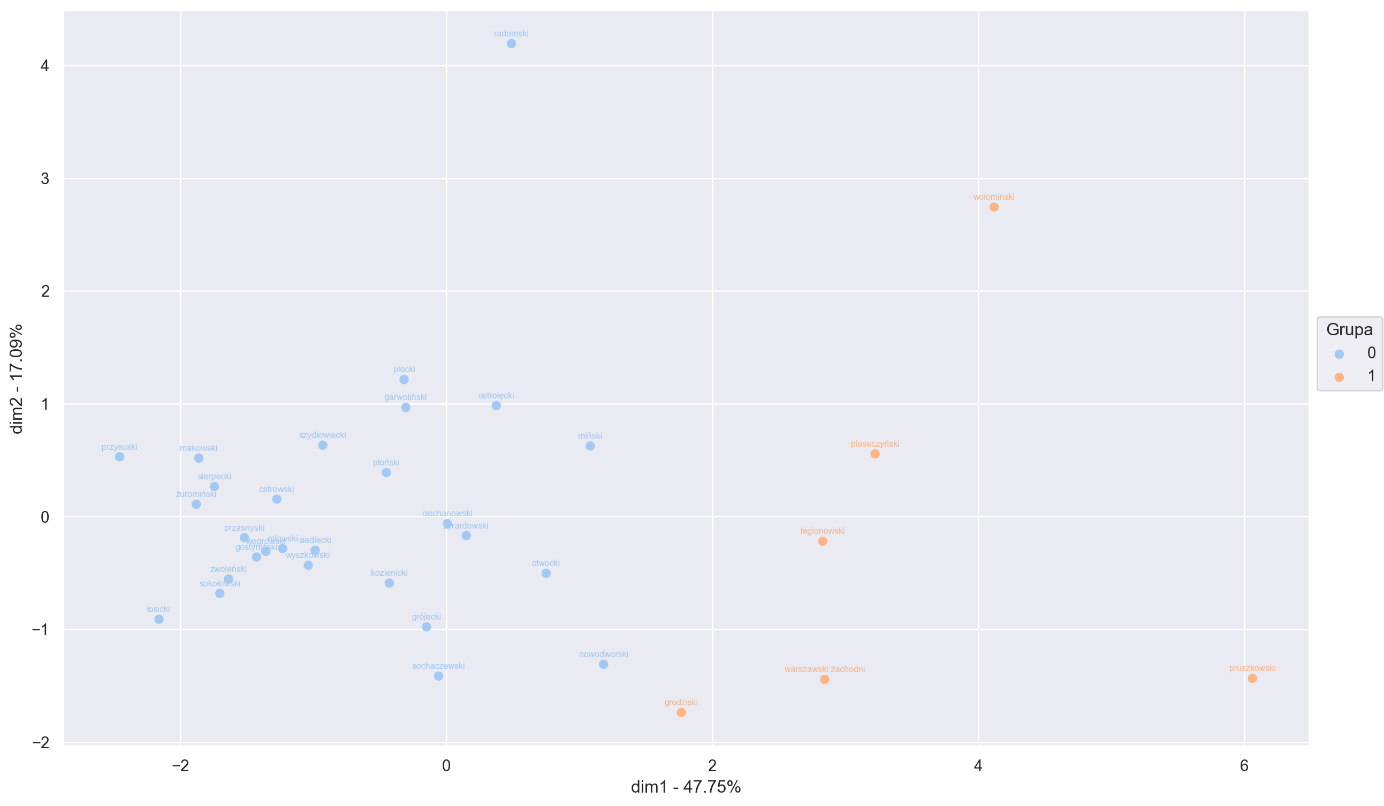


Ryc. Optymalna liczba klastrów (grup) wg metody silhouette

Poniżej załączono wskazane grupy i ich centra wskazane przez metodę k-średnich (Ryc.5) oraz te same grupy z podpisami nazw powiatów w celu ułatwienia identyfikacji (Ryc.6).

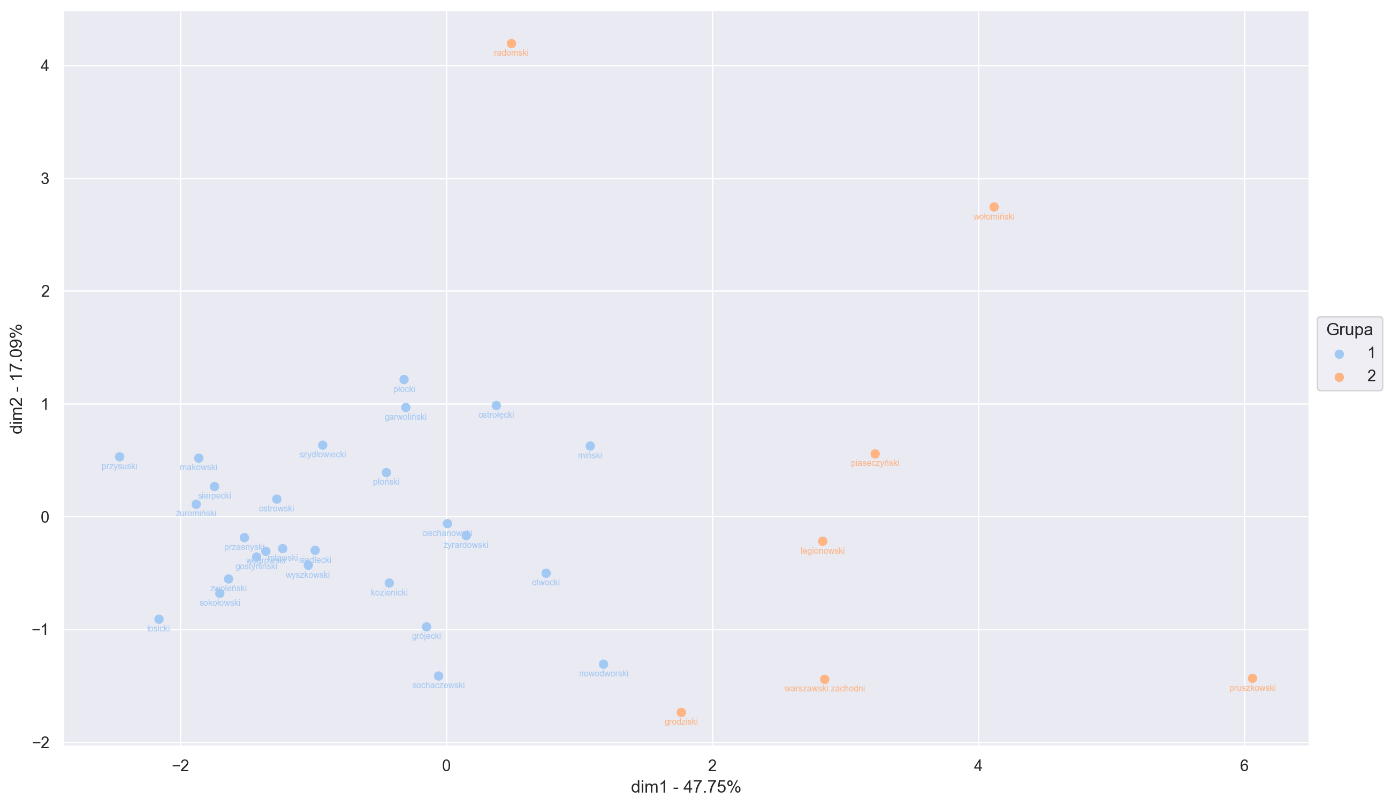


Ryc. Grupy powiatów wskazane przez metodę k-średnich wraz z centrami grup



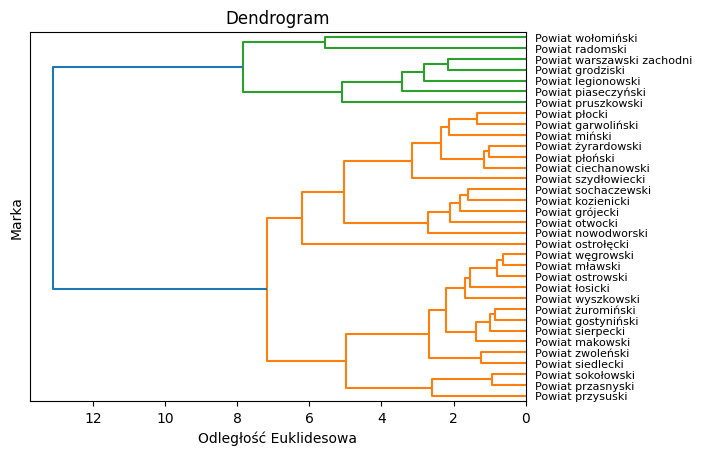
Ryc. Grupy wskazane przez metodę k-średnich wraz z podpisami

Metoda Warda, podzieliła powiaty na takie same grupy wykorzystując metody grupowania podziałowego (Ryc.7)



Ryc. Klastry wskazane przez metodę Warda

Metoda Warda do klasyfikacji klastrów używa Dendrogramu widocznego poniżej (Ryc.8).



Ryc. Dendrogram używany przez metodę Warda

### Podsumowanie