2025/26

Segmentacja miast europejskich na podstawie cech społeczno-ekonomicznych i demograficznych przy użyciu K-means

PROJEKT ZALICZENIOWY Z PRZEDMIOTU „UCZENIE MASZYNOWE”

MICHAŁ WROŃSKI

Cel projektu

Celem projektu jest zbadanie, czy wśród europejskich miast można wyróżnić wyraźne grupy (klastry) o podobnych cechach demograficznych, społeczno-ekonomicznych i infrastrukturalnych. Analiza ta ma na celu sprawdzenie, czy zastosowanie algorytmów uczenia maszynowego, w szczególności K-means, pozwala na identyfikację miast o podobnym profilu, np. pod względem:

- roli turystycznej (miasta przyciągające dużą liczbę turystów),
- wielkości i znaczenia metropolitalnego,
- struktury demograficznej (np. miasta o wysokiej liczbie imigrantów),
- poziomu rozwoju infrastruktury i usług (np. liczba samochodów, dostęp do edukacji wyższej).

Hipoteza

Hipotezą badawczą projektu jest stwierdzenie, że europejskie miasta można podzielić na kilka wyraźnych grup (klastrów) na podstawie cech społeczno-ekonomicznych i demograficznych. Każdy klaster powinien odpowiadać pewnemu charakterystycznemu typowi miasta, np.:

- Duże metropolie – miasta o wysokiej populacji, rozwiniętej infrastrukturze transportowej i wysokim poziomie edukacji.
- Miasta turystyczne – mniejsze lub średnie miasta o dużej liczbie przyjezdnych i wysokiej liczbie miejsc pracy w usługach.
- Miasta z silną imigracją – miasta z wyższym udziałem obcokrajowców i specyficznymi cechami demograficznymi.

Powody przeprowadzenia takiej analizy

Analiza grupowania miast może mieć wiele praktycznych zastosowań:

- Planowanie polityki miejskiej i regionalnej – identyfikacja typów miast pozwala lepiej dopasować strategie rozwoju, np. infrastrukturę transportową czy politykę mieszkaniową.
- Analiza trendów demograficznych i społecznych – klastrowanie pozwala zauważyć wspólne cechy miast o podobnych wyzwaniach, np. starzenie się populacji czy wysoki udział studentów.
- Wsparcie decyzji biznesowych – przedsiębiorstwa mogą wykorzystywać informacje o typach miast do planowania ekspansji, lokalizacji nowych usług czy produktów.
- Badania naukowe i porównawcze – grupowanie miast pozwala prowadzić bardziej spójne analizy porównawcze między krajami i regionami.

Źródło danych

Dane wykorzystane w projekcie pochodzą z Eurostatu, głównego źródła statystyk europejskich. Ze względu na to, że dostępne dane pochodzą z różnych lat, dla każdej zmiennej wybierano najświeższe dostępne informacje, aby analiza była jak najbardziej aktualna. Dane są dostępne pod adresem:

https://ec.europa.eu/eurostat/databrowser/explore/all/general?sort=category&lang=en&subtheme=urb.urb_cgc&display=list [DANE]

https://ec.europa.eu/eurostat/cache/metadata/en/urb_esms.htm [OPIS / METADANE]

Struktura i przygotowanie danych

Zbiór danych składa się z kilkunastu osobnych tabel, obejmujących różne aspekty społeczno-ekonomiczne i demograficzne miast europejskich. Aby umożliwić ich analizę w ramach jednego modelu, tabele zostały połączone w jeden spójny zbiór danych według unikalnych identyfikatorów miast.

W pierwotnym zbiorze znajduje się bardzo wiele zmiennych, jednak dla celów projektu wybrano arbitralnie te, które najlepiej reprezentują interesujące nas cechy miast, takie jak:

- średnia powierzchnia mieszkalna na mieszkańca,
- wskaźniki demograficzne (np. współczynnik urodzeń, udział osób starszych),
- wskaźniki społeczne (np. odsetek studentów w szkolnictwie wyższym, stopa bezrobocia),
- wskaźniki infrastruktury i transportu (np. liczba samochodów na 1000 mieszkańców, liczba ofiar śmiertelnych na drogach),
- udział cudzoziemców w populacji.

Wybór zmiennych był podyktowany chęcią uchwycenia zjawisk, które mogą różnicować miasta pod kątem charakteru społeczno-ekonomicznego i funkcjonalnego.

Charakterystyka zbioru

Ostateczny zbiór danych obejmuje 926 europejskich miast. Najmniejsze z miast w zbiorze liczy około 40 000 mieszkańców, co zapewnia, że analiza koncentruje się na miastach o istotnym znaczeniu regionalnym lub lokalnym.

Dzięki takiemu przygotowaniu możliwe było zastosowanie algorytmu K-means do grupowania miast w klastry na podstawie podobieństwa wybranych cech. Zbiór jest wystarczająco różnorodny, aby wykryć grupy miast o zróżnicowanym charakterze, np. metropolie, miasta turystyczne czy o dużym udziale cudzoziemców.

Etap I - Wstępne przetwarzanie danych

W pierwszym etapie projektu zajęto się przygotowaniem danych do dalszej analizy i modelowania. Dane źródłowe pochodzą z Eurostatu i obejmują szereg wskaźników dotyczących europejskich miast, takich jak populacja, wskaźniki demograficzne, infrastruktura transportowa, edukacja, gospodarka czy turystyka. Ze względu na zróżnicowane lata pozyskiwania danych, dla każdego wskaźnika i miasta wybrano zawsze najnowsze dostępne pomiary.

1. Wczytanie i połączenie danych

Dane były dostępne w postaci kilkunastu plików CSV. Wszystkie pliki zostały wczytane i połączone w jeden spójny DataFrame, dzięki czemu możliwe było dalsze przetwarzanie w jednolitej strukturze. Dodatkowo ustawiono pełną widoczność kolumn w Pandas, aby ułatwić eksplorację szerokiego zbioru danych.

2. Selekcja i oczyszczanie danych

- Usunięto obserwacje z brakującymi wartościami w kolumnie OBS_VALUE.
- Wykluczono obserwacje oznaczone flagą 'b' (break in time series), które najczęściej zawierały błędne lub niejednorodne jednostki.
- W przypadku danych z różnych lat, dla każdego miasta i wskaźnika zachowano wyłącznie najnowszą wartość, aby dane były aktualne i porównywalne.

3. Struktura danych i pivotowanie

Dane w oryginale miały postać “długą” (każdy wskaźnik w osobnym wierszu). Aby ułatwić analizę i przygotować je pod modelowanie, dane przekształcono do postaci szerokiej (wide format), gdzie każdy wskaźnik stał się osobną kolumną, a wiersze odpowiadały miastom. Dodatkowo zachowano informacje o kraju (Country_code) dla każdego miasta.

4. Nazwy kolumn i konwersja typów

- Wszystkie kolumny otrzymały czytelne, krótkie i spójne nazwy (np. Population, Share_foreigners, Infant_mortality_rate, Avg_living_area_m2_per_person).
- Wszystkie kolumny numeryczne zostały jawnie skonwertowane na typy numeryczne, co umożliwia bezproblemowe wykonywanie operacji matematycznych i statystycznych.

5. Korekta i przeliczanie wskaźników

- Niektóre wskaźniki wymagały przeskalowania. Przykładem jest Infant_mortality_rate, gdzie analiza wartości odstających sugerowała błędne jednostki (dane wprowadzone w przeliczeniu na 1 mln zamiast 1 tys. urodzeń).

- Wskaźniki zależne od populacji, takie jak liczba zgonów z przyczyn krążeniowo-oddechowych (Deaths_under_65_circulatory_respiratory) czy odpady komunalne (Municipal_waste_1000t), zostały przeliczone na wartości względne, np. na 100 tys. mieszkańców lub na mieszkańca.

6. Poprawa spójności nazw

- Usunięto dopisek (greater city) z nazw miast, aby ujednolicić nazewnictwo.
- Kody krajów zostały zredukowane do dwóch liter, zgodnie z międzynarodowym standardem ISO.

7. Zapis przygotowanego zbioru

Ostatecznie wstępnie przetworzony i oczyszczony zbiór danych zapisano do pliku CSV (preprocessed_data.csv), który stanowi bazę do dalszej analizy, eksploracji statystycznej i modelowania. Zawiera on wszystkie wybrane wskaźniki dla 926 miast europejskich, od najmniejszych miast liczących ok. 40 tys. mieszkańców, po największe metropolie.

Podsumowanie etapu I:

Etap wstępnego przetwarzania danych zapewnił spójny, kompletny i gotowy do analizy zbiór danych. Zostały uwzględnione aspekty jakości danych, różnorodność wskaźników i przeliczanie zmiennych na miary względne, co jest niezbędne do rzetelnej segmentacji miast w dalszych etapach projektu.

Etap II – Eksploracja danych, imputacja braków i standaryzacja

Po zakończeniu wstępnego przetwarzania danych kolejnym krokiem była szczegółowa eksploracja zbioru, ocena jakości zmiennych oraz przygotowanie danych do modelowania metodami opartymi na odległościach. Etap ten obejmował analizę braków danych, selekcję zmiennych, imputację brakujących wartości oraz standaryzację cech numerycznych.

1. Eksploracyjna analiza danych (EDA)

Na początku przeprowadzono analizę statystyk opisowych wszystkich zmiennych numerycznych (średnia, odchylenie standardowe, wartości minimalne i maksymalne), co pozwoliło zidentyfikować zmienne o silnej skośności, dużym zakresie wartości oraz potencjalnych obserwacjach odstających.

Szczególną uwagę poświęcono **strukturze braków danych**. W tym celu miasta zostały podzielone na przedziały wielkości populacji, a następnie obliczono procent braków dla każdej zmiennej w poszczególnych grupach. Wizualizacja w postaci mapy cieplnej pozwoliła stwierdzić, że:

- braki danych nie są równomiernie rozłożone,
- mniejsze miasta częściej nie raportują bardziej złożonych wskaźników infrastrukturalnych i sektorowych,
- część zmiennych cechuje się bardzo wysokim odsetkiem braków (70–80%), co wyklucza ich sensowną imputację.

Dodatkowo przeanalizowano rozkłady poszczególnych zmiennych przy użyciu histogramów, co umożliwiło ocenę ich stabilności, symetrii oraz potencjalnych problemów z imputacją.

2. Selekcja zmiennych na podstawie braków danych

Na podstawie analizy braków usunięto wszystkie zmienne, dla których odsetek brakujących wartości przekraczał 50%. Decyzja ta miała na celu ograniczenie ryzyka wprowadzania sztucznej struktury do danych poprzez agresywną imputację zmiennych o niskiej jakości informacyjnej.

3. Ocena przydatności zmiennych do imputacji

Kolejnym krokiem była koncepcyjna ocena zmiennych pod kątem możliwości wiarygodnej imputacji. Uwzględniono:

- charakter zmiennej (strukturalna vs. sektorowa),
- stabilność rozkładu,
- zależność od wielkości miasta,
- możliwość przewidywania wartości na podstawie innych cech.

Zmienne zakwalifikowane do imputacji

Do imputacji pozostawiono głównie:

- zmienne demograficzne,
- wskaźniki zdrowotne,
- dane dotyczące rynku pracy i edukacji.

Są to cechy relatywnie stabilne w czasie, silnie skorelowane z innymi zmiennymi i opisujące fundamentalne właściwości miast.

Zmienne odrzucone z dalszej analizy

Z analizy wykluczono m.in.:

- zmienne absolutne silnie zależne od skali miasta (np. odpady komunalne),
- zmienne infrastrukturalne i sektorowe (np. turystyka, liczba łóżek szpitalnych),
- zmienne o niejednoznacznym znaczeniu z punktu widzenia braków danych.

W ich przypadku imputacja mogłaby prowadzić do zafałszowania rzeczywistych różnic między miastami.

4. Imputacja brakujących wartości

Zastosowano **zróżnicowane metody imputacji**, dobrane indywidualnie do charakteru zmiennych:

- **Średnia** – dla zmiennych z bardzo niewielką liczbą braków i stabilnym rozkładem.
- **Mediana** – dla zmiennych podatnych na wartości odstające.
- **KNN Imputer** – dla zmiennych silnie zależnych od podobieństwa miast (np. bezrobocie, udział cudzoziemców).
- **MICE (Iterative Imputer)** – dla zmiennych o bardziej złożonych zależnościach wielowymiarowych.

Dla zmiennych o silnej skośności (np. populacja, liczba samochodów na 1000 mieszkańców) zastosowano **transformację logarytmiczną** przed imputacją metodą MICE, co pozwoliło ustabilizować proces estymacji i ograniczyć wpływ ekstremalnych wartości.

Po imputacji dokonano cofnięcia transformacji oraz przywrócenia danych do pierwotnej skali.

5. Ocena wpływu imputacji

W celu oceny jakości imputacji porównano statystyki opisowe zmiennych przed i po uzupełnieniu braków. Analiza wykazała, że:

- zmiany średnich wartości w większości przypadków nie przekraczały 1%,
- obserwowano umiarkowany spadek odchylenia standardowego, typowy dla imputacji,
- wartości minimalne i maksymalne pozostały niezmienione.

Oznacza to, że imputacja nie wprowadziła istotnych zniekształceń w strukturze danych i zachowała ich interpretowalność.

6. Analiza korelacji i redukcja redundancji

Na kolejnym etapie przeanalizowano macierz korelacji pomiędzy zmiennymi. Choć w zbiorze nie występowały bardzo silne korelacje ($|r| > 0.8$), zidentyfikowano blok umiarkowanie skorelowanych zmiennych demograficznych ($r \approx 0.5–0.7$).

Aby zapobiec nadmiernemu wpływowi jednego wymiaru (struktury wieku) na wynik klasteryzacji, część zmiennych demograficznych została usunięta, pozostawiając jedynie najbardziej reprezentatywne cechy. Dodatkowo wykluczono zmienne wykazujące wysoką korelację bez jednoznacznej interpretacji przyczynowej.

7. Standaryzacja danych

Ponieważ w dalszej części projektu zastosowano algorytm K-means, który opiera się na odległościach euklidesowych, wszystkie wybrane zmienne numeryczne zostały standaryzowane (średnia = 0, odchylenie standardowe = 1).

Równolegle zapisano:

- wersję niestandaryzowaną danych – wykorzystywaną do interpretacji klastrów,
- wersję standaryzowaną – używaną bezpośrednio w modelu klasteryzacyjnym.

Podsumowanie etapu II

Etap eksploracji, imputacji i standaryzacji pozwolił uzyskać spójny, kompletny i zbalansowany zbiór danych, w którym każda zmienna wnosi odrębną informację o charakterze miasta. Dzięki świadomej selekcji cech oraz odpowiednio dobranym metodom imputacji przygotowany zbiór danych jest dobrze dostosowany do dalszej analizy klasteryzacyjnej i umożliwia interpretowalne porównywanie europejskich miast.

ETAP III – Modelowanie i klasteryzacja miast

1. Cel etapu

Celem etapu modelowania było przeprowadzenie klasteryzacji europejskich miast w oparciu o przygotowany zbiór cech społeczno-demograficznych i ekonomicznych oraz ocena, czy możliwe jest wydzielenie względnie jednorodnych grup miast o podobnym profilu. Zastosowano algorytm K-means, który jest jedną z najczęściej wykorzystywanych metod klasteryzacji w analizie danych ilościowych.

Ze względu na charakter algorytmu K-means kluczowym problemem metodologicznym jest dobór liczby klastrów k, która nie jest znana a priori i w istotny sposób wpływa na interpretowalność wyników.

2. Przygotowanie danych do modelowania

Do klasteryzacji wykorzystano dane ustandaryzowane, co jest niezbędne w przypadku algorytmów opartych na odległościach euklidesowych. Z macierzy cech usunięto kolumny identyfikacyjne (nazwę miasta oraz kod kraju), pozostawiając wyłącznie zmienne numeryczne opisujące charakterystykę miast.

Równolegle zachowano wersję danych niestandaryzowanych, która została wykorzystana wyłącznie na etapie interpretacji klastrów (obliczanie średnich wartości zmiennych).

3. Dobór liczby klastrów – metoda łokcia

W pierwszym kroku przeanalizowano zależność wartości funkcji celu algorytmu K-means (inertia) od liczby klastrów kw przedziale od 2 do 10. Metoda łokcia polega na identyfikacji punktu, dla którego dalsze zwiększanie liczby klastrów przestaje przynosić istotną redukcję błędu.

Uzyskany wykres charakteryzuje się **monotonicznym, niemal liniowym spadkiem wartości inertia**, bez wyraźnego punktu załamania. Oznacza to, że:

- zwiększanie liczby klastrów systematycznie poprawia dopasowanie modelu,
- brak jest jednej wartości k, która w sposób jednoznaczny mogłaby zostać uznana za optymalną.

Taki rezultat sugeruje, że dane nie zawierają naturalnie wyraźnie rozdzielonych, kulistych skupień, co jest typowe dla danych społeczno-demograficznych opisujących miasta.

4. Analiza silhouette score

W celu uzupełnienia analizy obliczono wartości silhouette score dla tych samych wartości k. Miara silhouette jednocześnie ocenia spójność klastrów oraz stopień ich separacji, przyjmując wartości z przedziału $\langle -1, 1 \rangle$.

Uzyskane wartości silhouette score były **relatywnie niskie (około 0.12–0.15)** i nie wykazywały wyraźnego globalnego maksimum. Wskazuje to na:

- słabą separację klastrów,
- ciągły charakter przestrzeni cech,

- brak jednoznacznej, „naturalnej” liczby klastrów.

Najwyższą wartość silhouette score uzyskano dla $k = 10$, jednak różnice pomiędzy kolejnymi wartościami $k$były niewielkie. Porównywalne wyniki uzyskano również dla $k = 2$, $k = 6$oraz $k = 8$, co sugeruje istnienie kilku potencjalnie sensownych poziomów segmentacji.

5. Wizualizacja klastrów w przestrzeni PCA

Aby lepiej zrozumieć strukturę danych i sposób działania algorytmu K-means, przeprowadzono redukcję wymiarowości za pomocą analizy głównych składowych (PCA) do dwóch wymiarów. Następnie na tej samej projekcji wizualizowano wyniki klasteryzacji dla różnych wartości k.

Uzyskane wykresy potwierdziły wcześniejsze wnioski:

- punkty tworzą zwartą, ciągłą chmurę,
- brak jest wyraźnych granic pomiędzy potencjalnymi klastrami,
- zwiększanie liczby klastrów prowadzi głównie do dalszego dzielenia tej samej struktury, a nie do ujawnienia nowych, naturalnych skupień.

Wizualnie najbardziej czytelny podział uzyskano dla $k = 3$. Przy tej liczbie klastrów grupy są względnie równomierne, a ich interpretacja jest prostsza niż w przypadku większych wartości k. Dla $k = 4$ oraz wyższych obserwowano tworzenie się klastrów skupiających pojedyncze obserwacje odstające, co utrudniało analizę.

6. Finalny model K-means (k = 3)

Na podstawie łącznej analizy:

- metody łokcia,
- silhouette score,
- wizualizacji PCA,

do dalszej analizy wybrano model K-means z **trzema klastrami**. Choć wybór ten ma charakter częściowo arbitralny, jest on dobrze uzasadniony z punktu widzenia interpretowalności i czytelności wyników.

Po dopasowaniu modelu każdemu miastu przypisano etykietę klastra, która została dołączona zarówno do zbioru danych ustandaryzowanych, jak i niestandaryzowanych.

7. Charakterystyka klastrów

W celu interpretacji klastrów obliczono średnie wartości zmiennych dla każdego klastra na podstawie danych niestandaryzowanych. Pozwoliło to opisać typowe cechy miast należących do poszczególnych grup w rzeczywistych jednostkach (np. procenty, wartości bezwzględne).

Dodatkowo przeanalizowano rozkład klastrów w podziale na kraje. Dla państw posiadających dużą liczbę miast w zbiorze danych przedstawiono procentowy udział miast w poszczególnych klastrach.

Analiza ta pozwoliła zidentyfikować kraje, w których dominują określone typy miast, oraz kraje o bardziej zróżnicowanej strukturze urbanistycznej.

8. Największe miasta w klastrach

Na zakończenie, w celu ułatwienia interpretacji klastrów, dla każdego z nich wyodrębniono listę 30 największych miast pod względem liczby ludności. Zestawienie to pozwala intuicyjnie zrozumieć charakter poszczególnych grup poprzez odniesienie do znanych ośrodków miejskich.

Wnioski końcowe

1. Ogólna ocena wyników klasteryzacji

Przeprowadzona analiza potwierdziła, że europejskie miasta nie tworzą jednoznacznie rozdzielonych, naturalnych skupień, lecz raczej ciągłe spektrum cech społeczno-demograficznych i ekonomicznych. Pomimo tego, zastosowanie algorytmu K-means z liczbą klastrów $k = 3$pozwoliło na wydzielenie trzech względnie spójnych grup miast, które różnią się zarówno strukturą demograficzną, jak i profilem społeczno-ekonomicznym.

Uzyskane klastry nie powinny być traktowane jako „twarde" kategorie, lecz raczej jako **uproszczone typologie**, ułatwiające porównywanie miast i identyfikację dominujących wzorców urbanistycznych w Europie.

2. Charakterystyka klastrów

Cluster 0 – duże metropolie i miasta globalne

Miasta należące do klastra 0 charakteryzują się:

- najwyższym udziałem cudzoziemców (średnio 16,6%),
- relatywnie wysokim współczynnikiem urodzeń,
- największą liczbą studentów szkolnictwa wyższego,
- największą średnią populacją (ok. 384 tys. mieszkańców),
- najniższą stopą bezrobocia spośród analizowanych klastrów.

Jednocześnie miasta te cechują się:

- niższą liczbą samochodów na 1000 mieszkańców,
- niską śmiertelnością wypadków drogowych,
- umiarkowanym wskaźnikiem starzenia się populacji.

Lista największych miast w tym klastrze (m.in. Paryż, Londyn, Berlin, Madryt, Barcelona, Mediolan, Amsterdam) jednoznacznie wskazuje, że są to duże metropolie pełniące funkcje międzynarodowych centrów gospodarczych, akademickich i migracyjnych. Klastr ten można interpretować jako grupę miast globalnych i silnych metropolii, charakteryzujących się wysoką atrakcyjnością migracyjną i relatywnie młodą strukturą demograficzną.

Cluster 1 – miasta zachodnioeuropejskie o stabilnym, starzejącym się profilu

Cluster 1 wyróżnia się:

- największą powierzchnią mieszkalną na osobę,
- najwyższym wskaźnikiem zależności osób starszych,

- najniższym współczynnikiem urodzeń,
- najwyższą liczbą samochodów na 1000 mieszkańców,
- najwyższą stopą bezrobocia wśród klastrów.

Jednocześnie udział cudzoziemców i studentów jest tu wyraźnie niższy niż w klastrze 0.

Struktura miast (m.in. Rzym, Turyn, Palermo, Bologna, Florencja, Genua, ale także wiele średnich miast Francji i Niemiec) sugeruje, że jest to klaster obejmujący dojrzałe miasta Europy Zachodniej i Południowej, często o długiej historii urbanistycznej, stabilnej, lecz starzejącej się populacji i mniejszej dynamice demograficznej.

Można je interpretować jako miasta o wysokim standardzie życia, lecz ograniczonym napływie młodej ludności i migrantów.

Cluster 2 – miasta Europy Środkowo-Wschodniej i peryferyjne o mniejszej skali

Cluster 2 charakteryzuje się:

- najmniejszą powierzchnią mieszkalną na osobę,
- najniższym udziałem cudzoziemców,
- najmniejszym udziałem studentów,
- umiarkowanym poziomem bezrobocia,
- średnią wielkością populacji (ok. 175 tys.).

Lista miast (m.in. Warszawa, Budapeszt, Praga, Kraków, Wrocław, Sofia, Wilno, Zagrzeb) wskazuje jednoznacznie na dominację miast Europy Środkowo-Wschodniej oraz wybranych regionów peryferyjnych Europy Zachodniej.

Miasta te cechują się niższą internacjonalizacją, mniejszą atrakcyjnością migracyjną oraz ograniczonym rynkiem akademickim w porównaniu z dużymi metropoliami. Jednocześnie wiele z nich pełni funkcje krajowych lub regionalnych centrów administracyjnych i gospodarczych.

3. Zróżnicowanie klastrów pomiędzy krajami

Analiza rozkładu klastrów w podziale na kraje ujawnia wyraźne różnice regionalne:

- Polska, Rumunia i Turcja są niemal w całości reprezentowane przez cluster 2, co odzwierciedla wspólny profil miast Europy Środkowo-Wschodniej.
- Włochy są silnie skoncentrowane w klastrze 1, co wskazuje na dominację miast o stabilnej, starzejącej się strukturze demograficznej.
- Francja i Niemcy charakteryzują się dużym zróżnicowaniem – obecne są zarówno miasta globalne (cluster 0), jak i miasta o bardziej tradycyjnym profilu (cluster 1).
- Wielka Brytania wykazuje znaczący udział klastrów 0 i 2, co podkreśla silne kontrasty pomiędzy metropoliami a miastami regionalnymi.

4. Odpowiedź na pytanie badawcze

Postawione na początku pytanie badawcze – *czy możliwe jest wydzielenie sensownych grup europejskich miast na podstawie cech społeczno-demograficznych* – można uznać za **częściowo potwierdzone**.

Choć dane nie wykazują naturalnej, silnej struktury klastrów, możliwe jest wydzielenie interpretowalnych typów miast, które różnią się:

- skalą,
- poziomem internacjonalizacji,
- strukturą demograficzną,
- funkcją społeczną i gospodarczą.

5. Ograniczenia analizy

Należy podkreślić kilka istotnych ograniczeń:

- dane pochodzą z różnych lat,
- wybór zmiennych miał charakter arbitralny,
- algorytm K-means wymusza kulistą strukturę klastrów,
- brak danych przestrzennych ogranicza analizę kontekstu geograficznego.

6. Możliwe kierunki dalszych badań

W przyszłości analiza mogłaby zostać rozszerzona o:

- inne algorytmy klasteryzacji (DBSCAN, HDBSCAN, GMM),
- uwzględnienie danych przestrzennych,
- analizę zmian w czasie,
- pogłębioną analizę pojedynczych krajów.