Klasteryzacja mieszkańców Stanów Zjednoczonych

Jakub Piwko, Malwina Wojewoda

Projekt w ramach przedmiotu
Wstęp do uczenia maszynowego
semestr letni 2021/2022

Plan prezentacji

- Opis zadania
- Dane
- EDA
- Preprocessing
- Modele
- Wybór najlepszego modelu
- Charakteryzacja klastrów

Opis zadania

Celem zadania jest znalezienie grup mieszkańców Stanów Zjednoczonych o podobnych charakterystykach i umieszczenie ich w klastrach.

Dane

- Dane pochodzące ze spisu powszechnego przeprowadzonego w Stanach Zjednoczonych w 1990r.
- Losowa jednoprocentowa próbka całej populacji Stanów Zjednoczonych.
- Zbiór danych w dwóch wersjach: oryginalnej i przetworzonej

dTravtime	iVietnam	dWeek89	iWork89	iWorklwk	iWWII
5	0	2	1	1	0
1	0	2	1	1	0
2	0	2	1	1	0
1	0	1	1	1	0
0	0	0	2	2	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	2	1	1	0
0	0	1	1	2	0
0	0	1	1	2	0
4	0	2	1	1	0
2	0	2	1	1	0
0	0	0	0	2	0
0	0	0	0	0	0
3	0	1	1	1	0

Dane

Zbiór po przetworzeniu:

- odrzucenie nic nie wnoszących kolumn,
- skategoryzowanie zmiennych różnymi metodami,
- prefiksy określające czy kolumna jest zmodyfikowana czy nie,
- 2458255 wierszy i 69 kolumn.

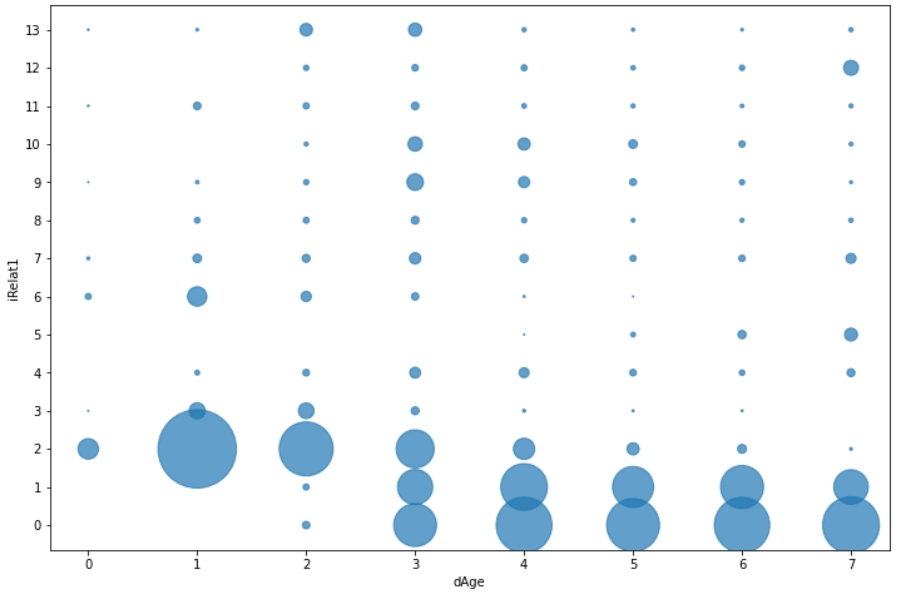
Niezmodyfikowana

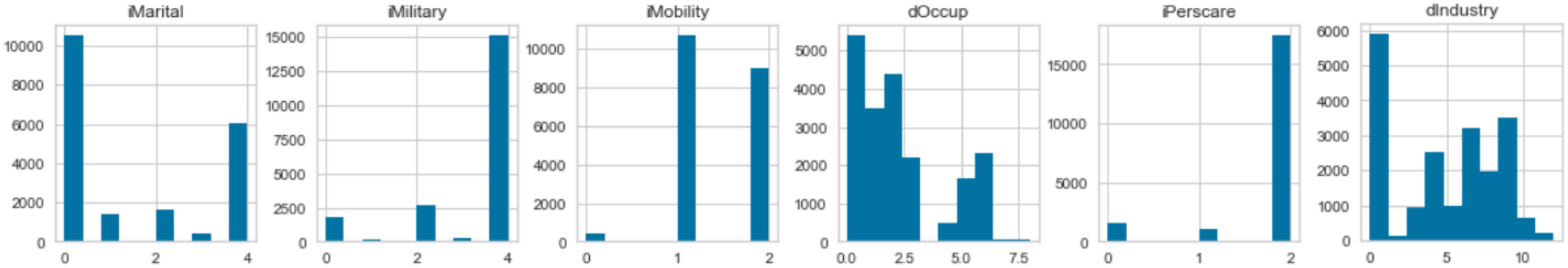
Nowa zmienna

iClass	dDepart
5	3
7	5
7	4
1	3
0	0

EDA

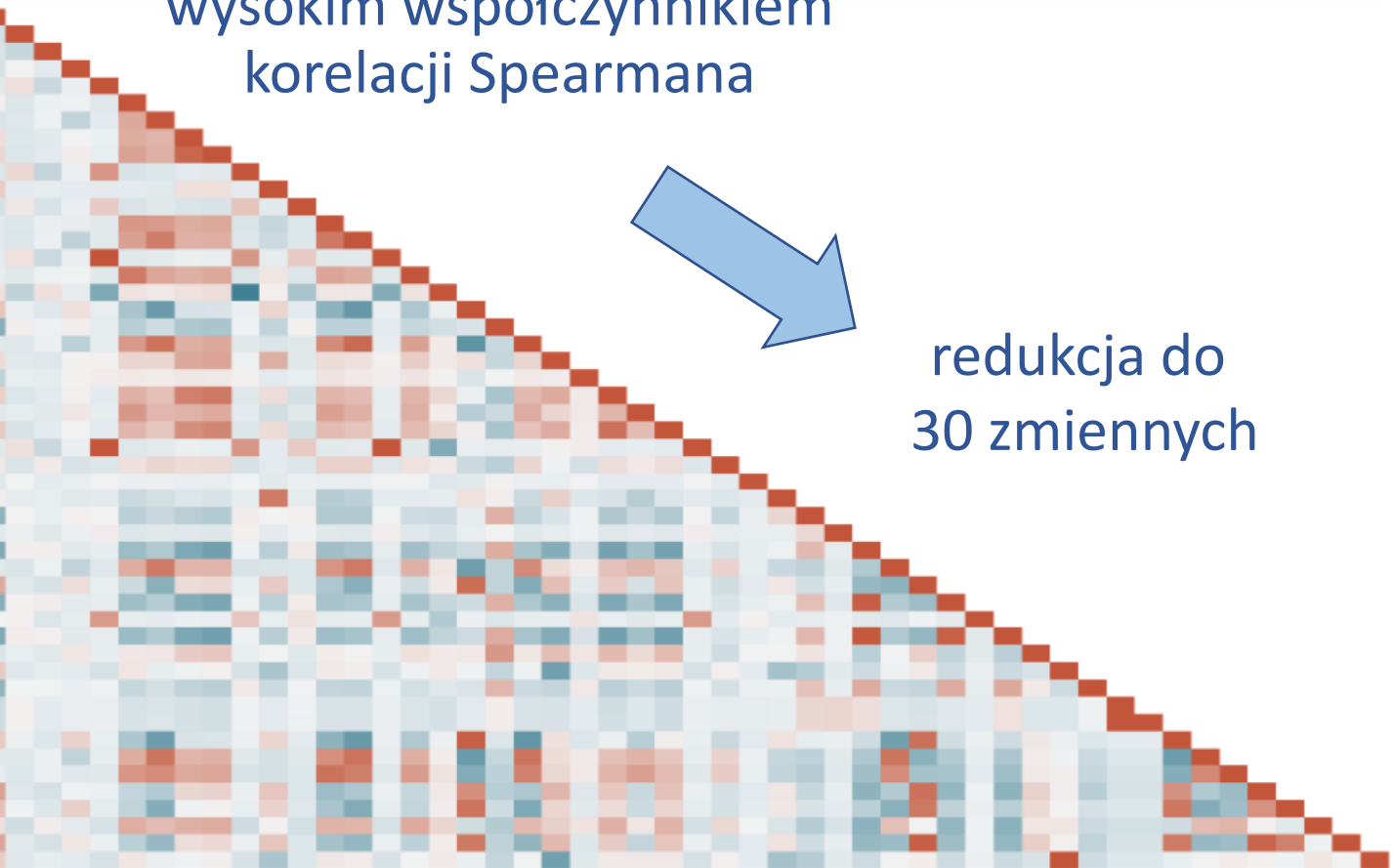
- Analiza braków danych
- Badanie rozkładów zmiennych
- Wykresy bąbelkowe pokazujące zależności między zmiennymi





Redukcja zmiennych

usunięcie zmiennych z wysokim współczynnikiem korelacji Spearmana



redukcja do 30 zmiennych

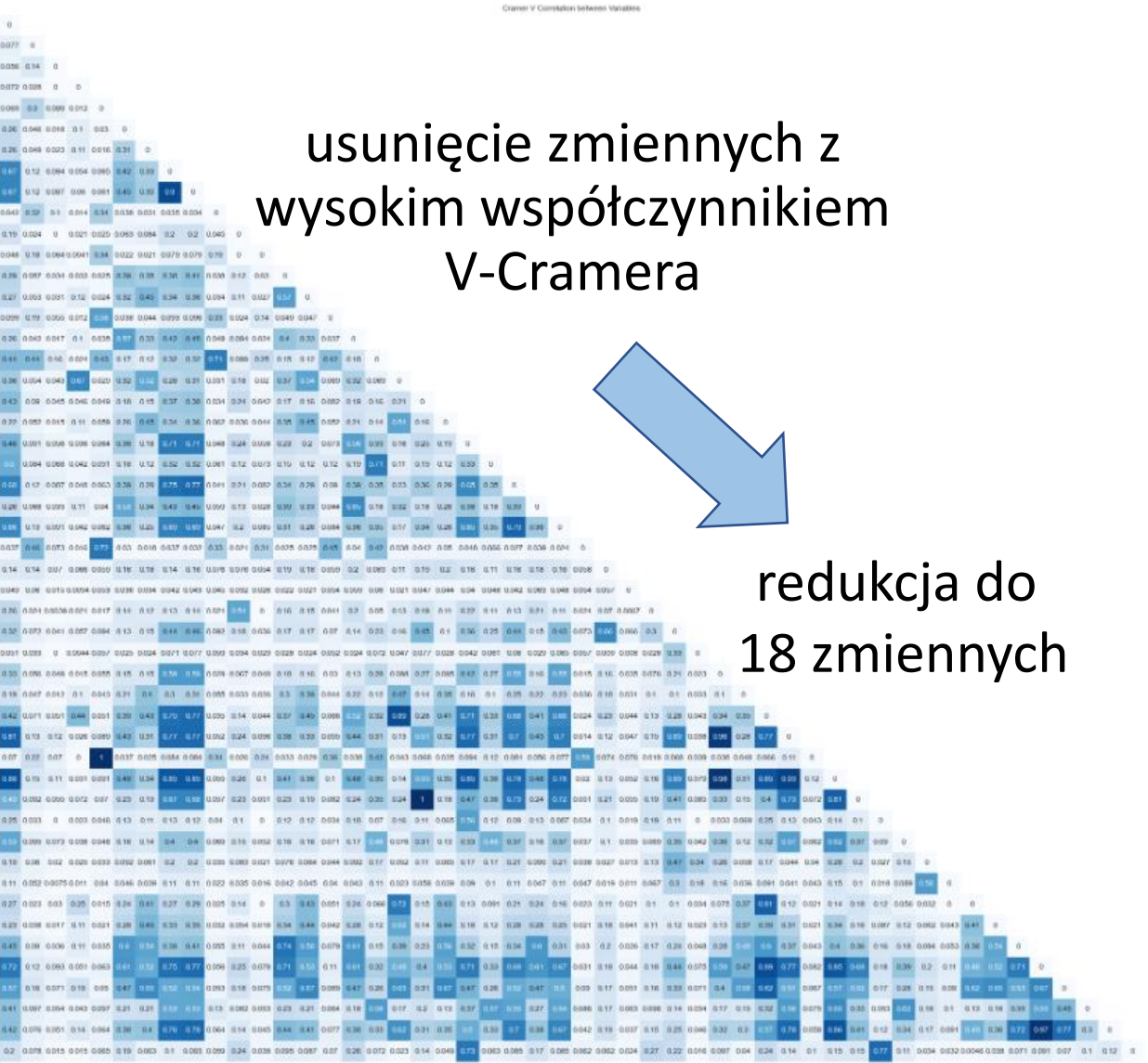
Served World War II September 1940 July
Served Vietnam Era August 1964 April 197
Served September 1980 or Later
Served February 1955 July 1964
Served Korean Conflict June 1950 January
Served May 1975 to August 1980
Served Any Other Time

ręczna redukcja zmiennych przekazujących te same informacje

iRvetserv

00	N/a Less Than 16 Yrs. Old, No Active Dut
01	September 1980 or Later Only
02	May 1975 to August 1980 Only
03	May 1975 to August 1980 and September 19
04	Vietnam Era, No Korean Conflict, No Wwii
05	Vietnam Era and Korean Conflict, No Wwii
06	Vietnam Era and Korean Conflict and Wwii
07	February 1955 to July 1964 Only
08	Korean Conflict, No Vietnam Era, No Wwii
09	Korean Conflict and Wwii, No Vietnam Era
10	Wwii, No Korean Conflict, No Vietnam Era
11	Other Srvc.

Inne testowane metody redukcji zmiennych

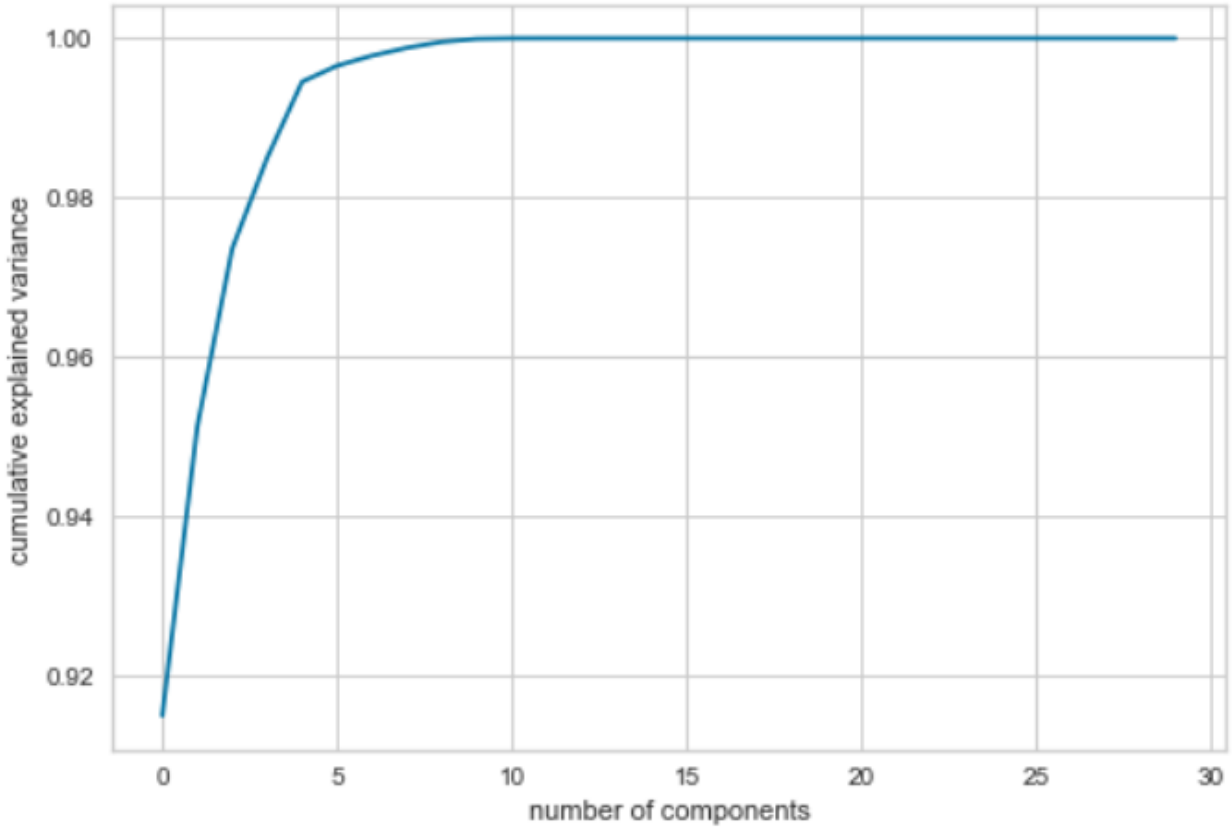


usunięcie zmiennych z wysokim współczynnikiem V-Cramera

redukcja do 18 zmiennych

MCA - Multiple Correspondence Analysis

odpowiednik PCA dla zmiennych kategorycznych



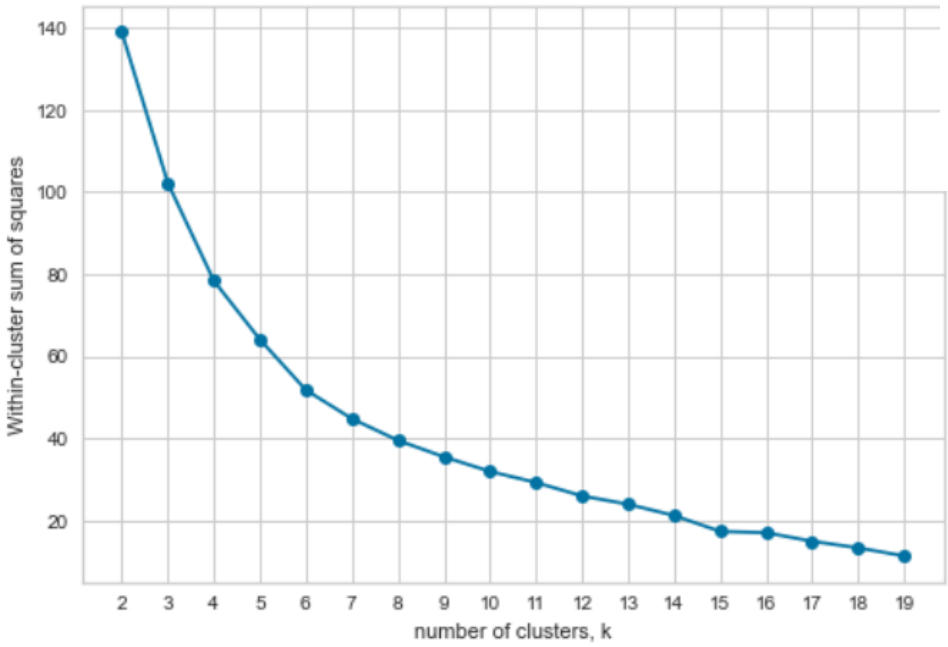
Skalowanie

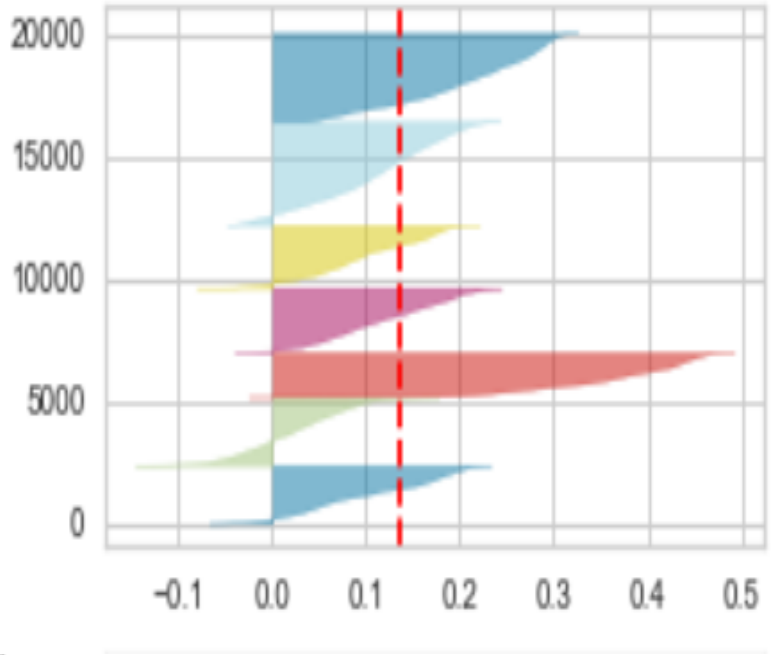
- Wybranie zmiennych, które nie reprezentują naturalnego porządku jako częstości wystąpień
- Skalowanie do przedziału [0, 1]

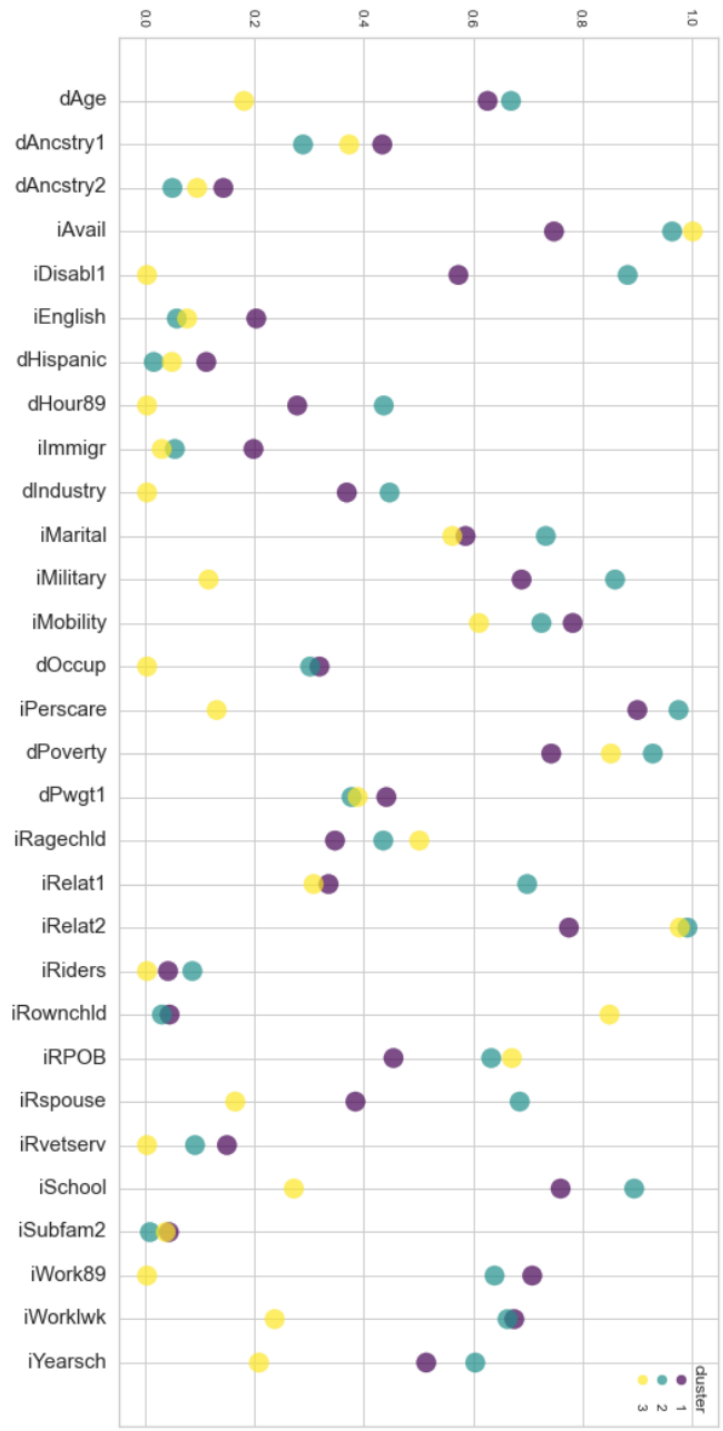
dAge	dAncstry1	dAncstry2	iAvail	iDisabl1	iEnglish	dHispanic	dHour89	iImmigr	dIndustry	...	iRiders	iRownchld	iRPOB	iRspous
0.571429	0.272727	1.0	0.034018	1.000000	0.0	1.0	0.0	0.0	0.363841	...	0.000	1.0	1.000000	0.42088
1.000000	1.000000	0.0	1.000000	0.041356	0.0	1.0	0.0	0.0	1.000000	...	0.000	1.0	0.188474	1.00000
0.428571	0.181818	0.0	1.000000	1.000000	0.0	1.0	0.2	0.0	0.646079	...	0.125	1.0	0.136714	1.00000
1.000000	0.090909	0.0	1.000000	1.000000	0.0	1.0	0.4	0.0	0.646079	...	0.000	1.0	0.188474	1.00000
0.571429	0.090909	0.0	1.000000	1.000000	0.0	1.0	0.0	0.0	1.000000	...	0.000	1.0	1.000000	1.00000

KMeans

wybór liczby klastrów
przy pomocy metody łokcia
oraz metodą Silhouette

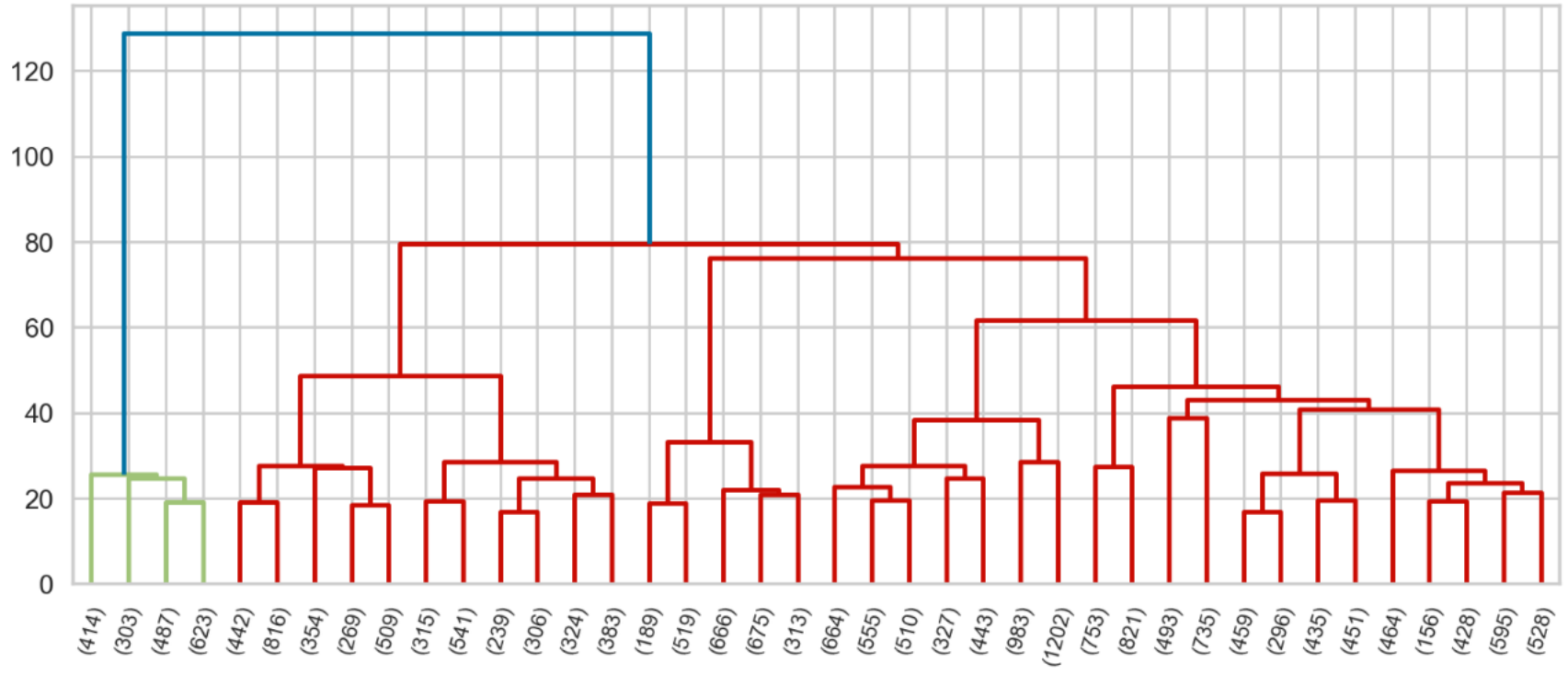


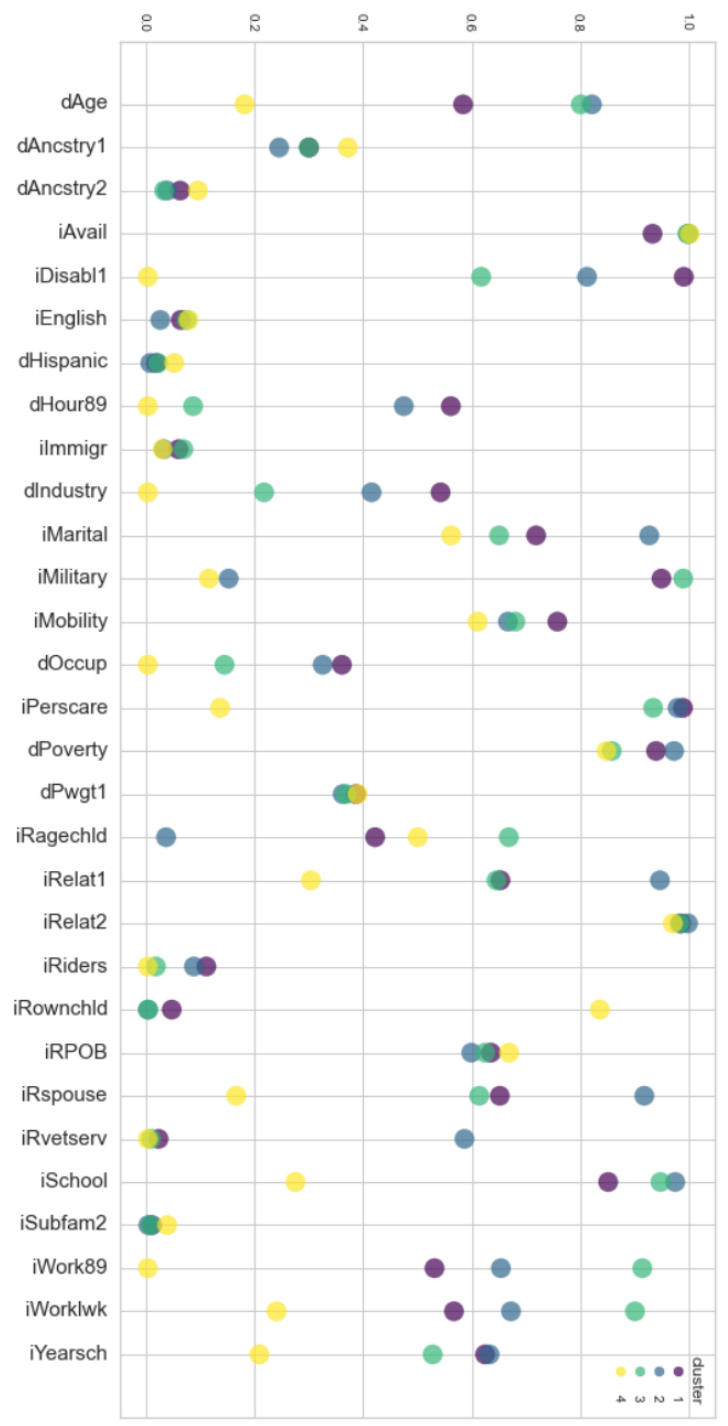




Klasteryzacja hierarchiczna

denodrogram

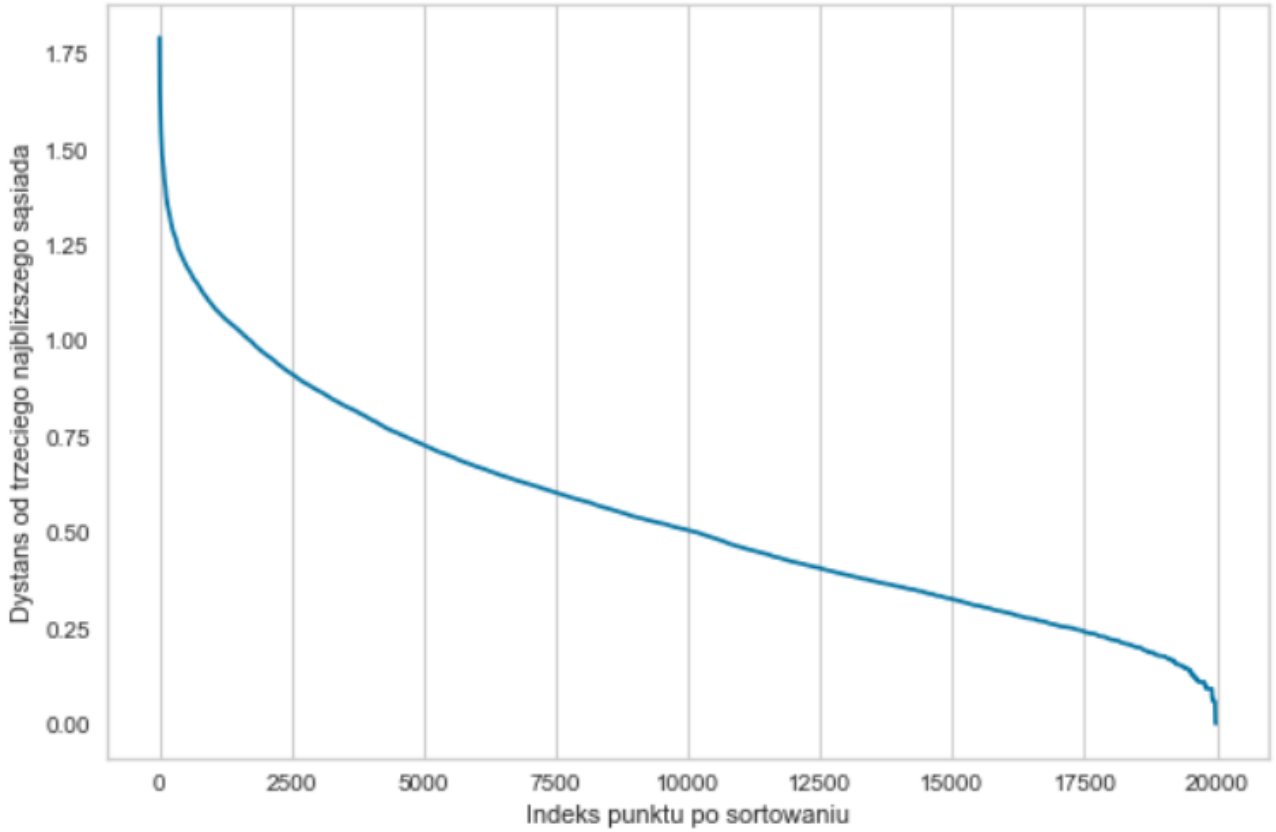


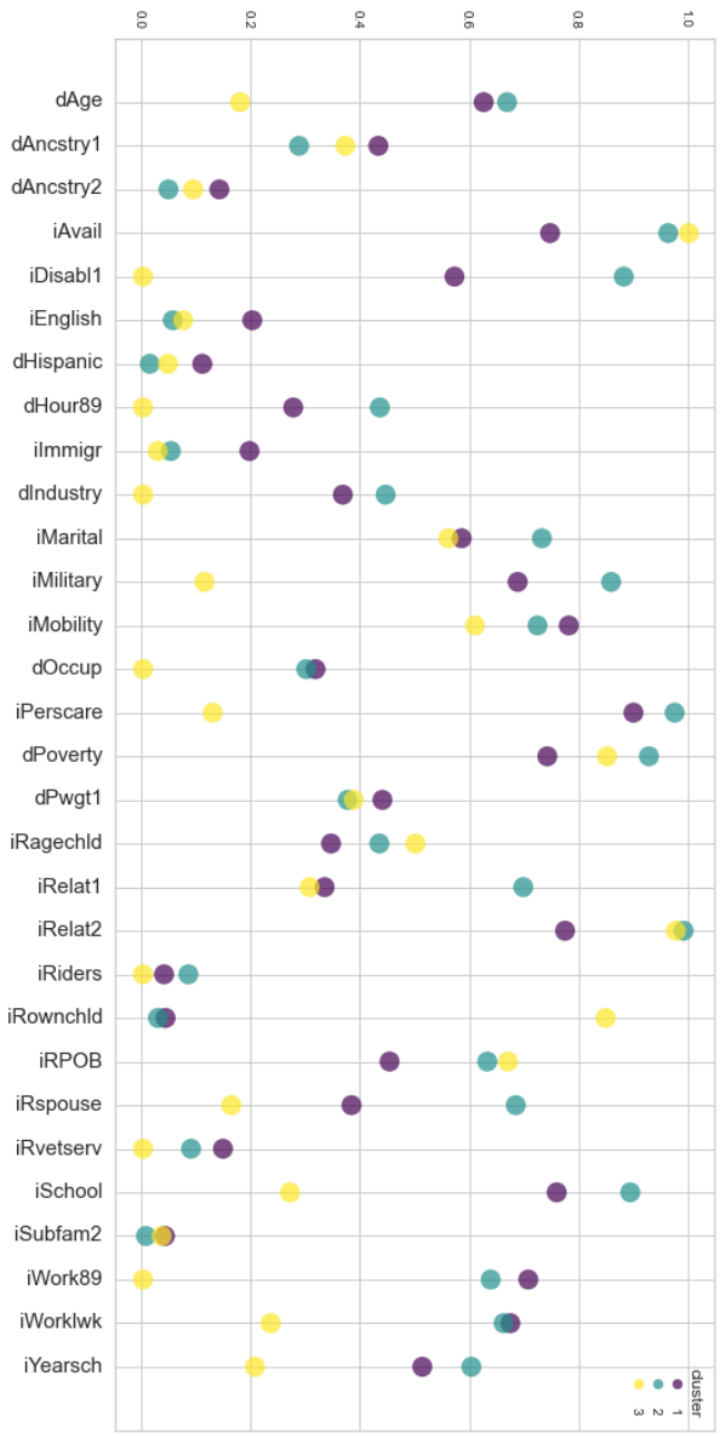


DBSCAN

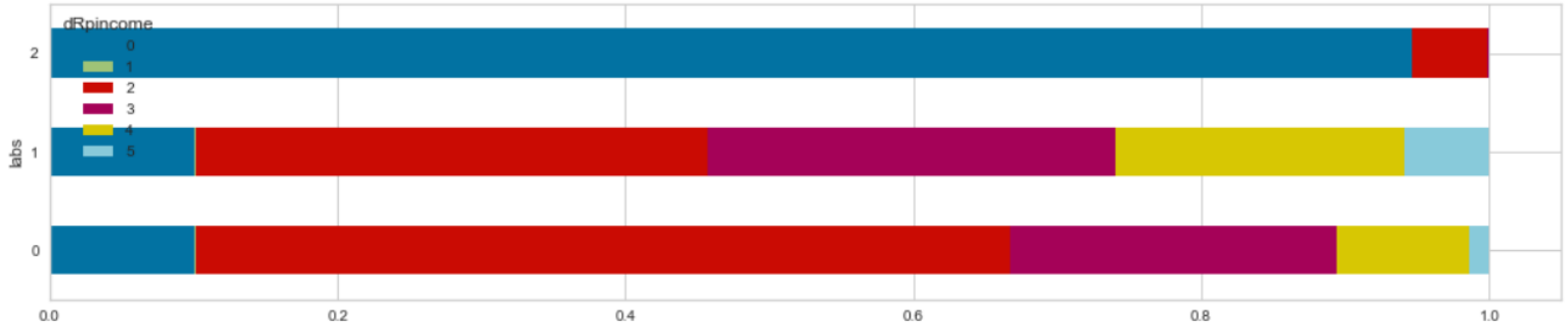
wybór epsilona

maksymalnej odległości, na jaką mogą być oddalone od siebie dwa punkty z tego samego klastra





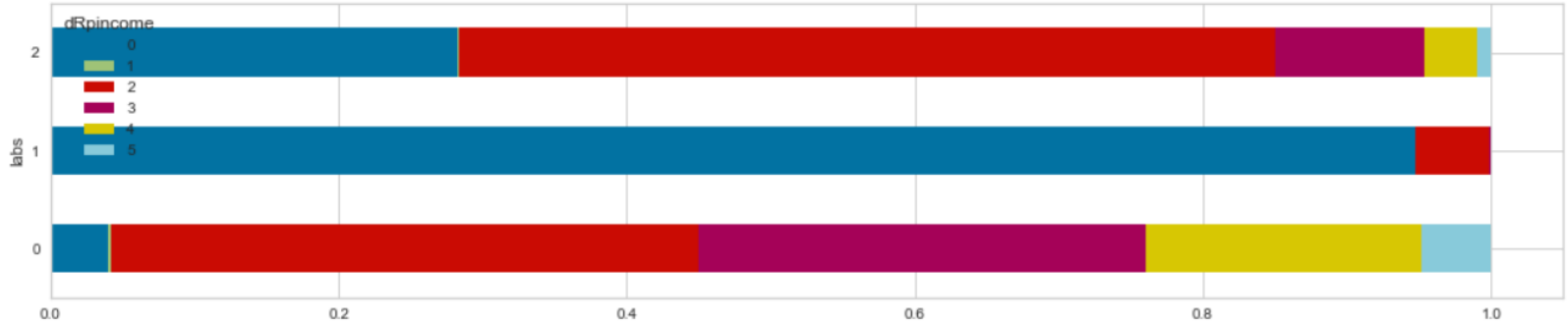
porównanie metod klasteryzacji

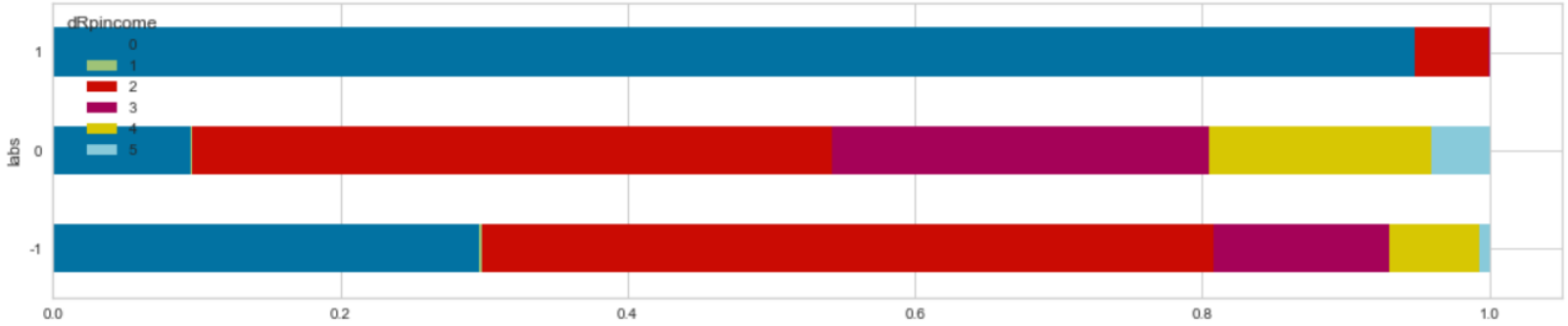


KMeans

klasteryzacja aglomeracyjna

wybrana przez nas jako najlepsza





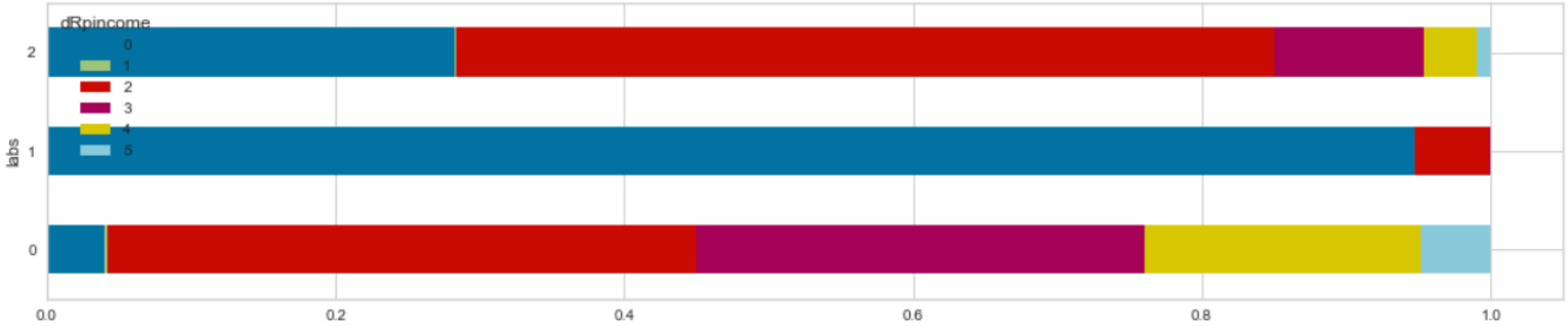
DBSCAN

Klasteryzacja aglomeracyjna – 3 klastry

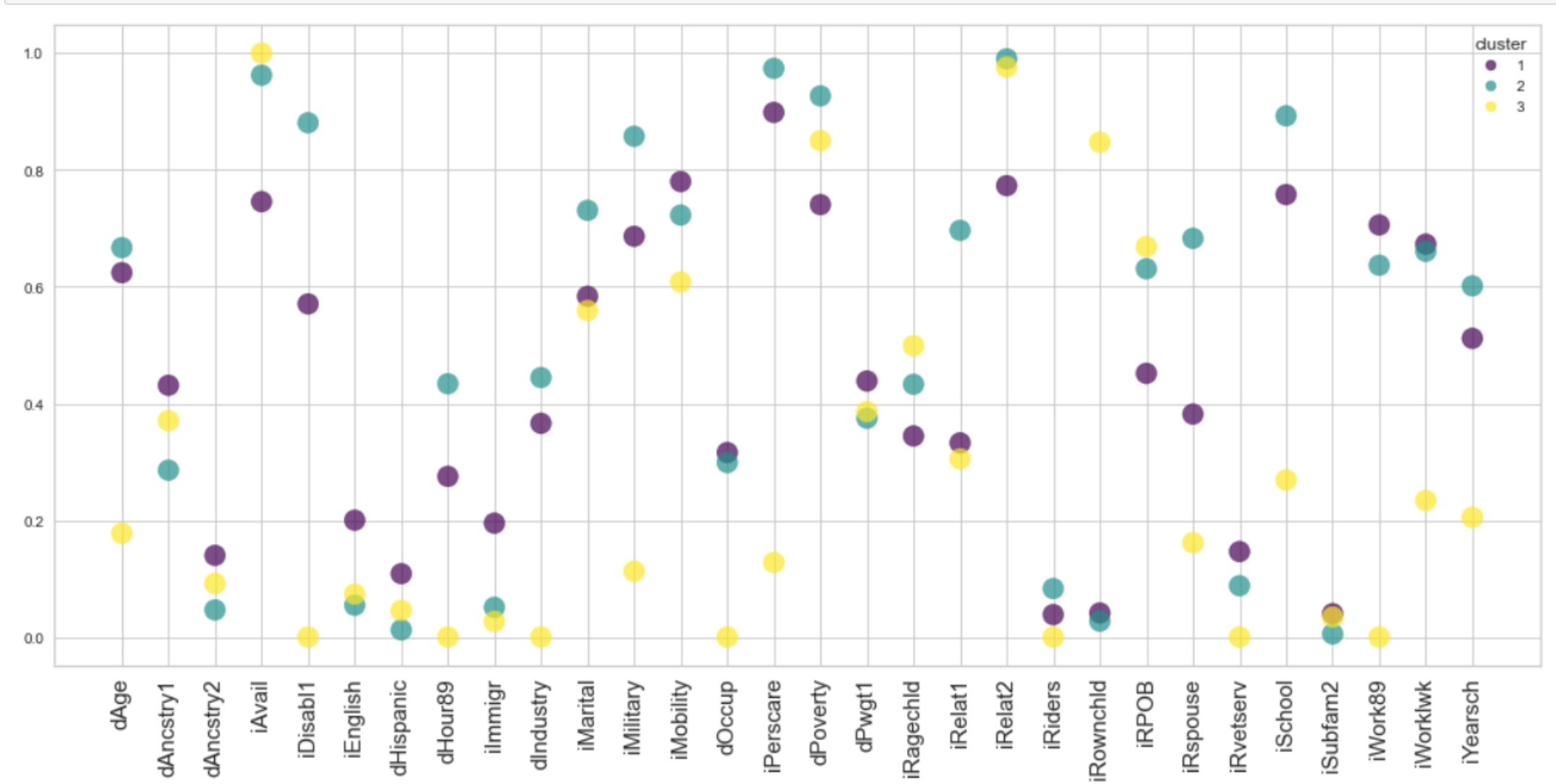
Klaster 2 – osoby o małym przychodzie (4498 obserwacji)

Klaster 1 – osoby bez przychodu (1827 obserwacji)

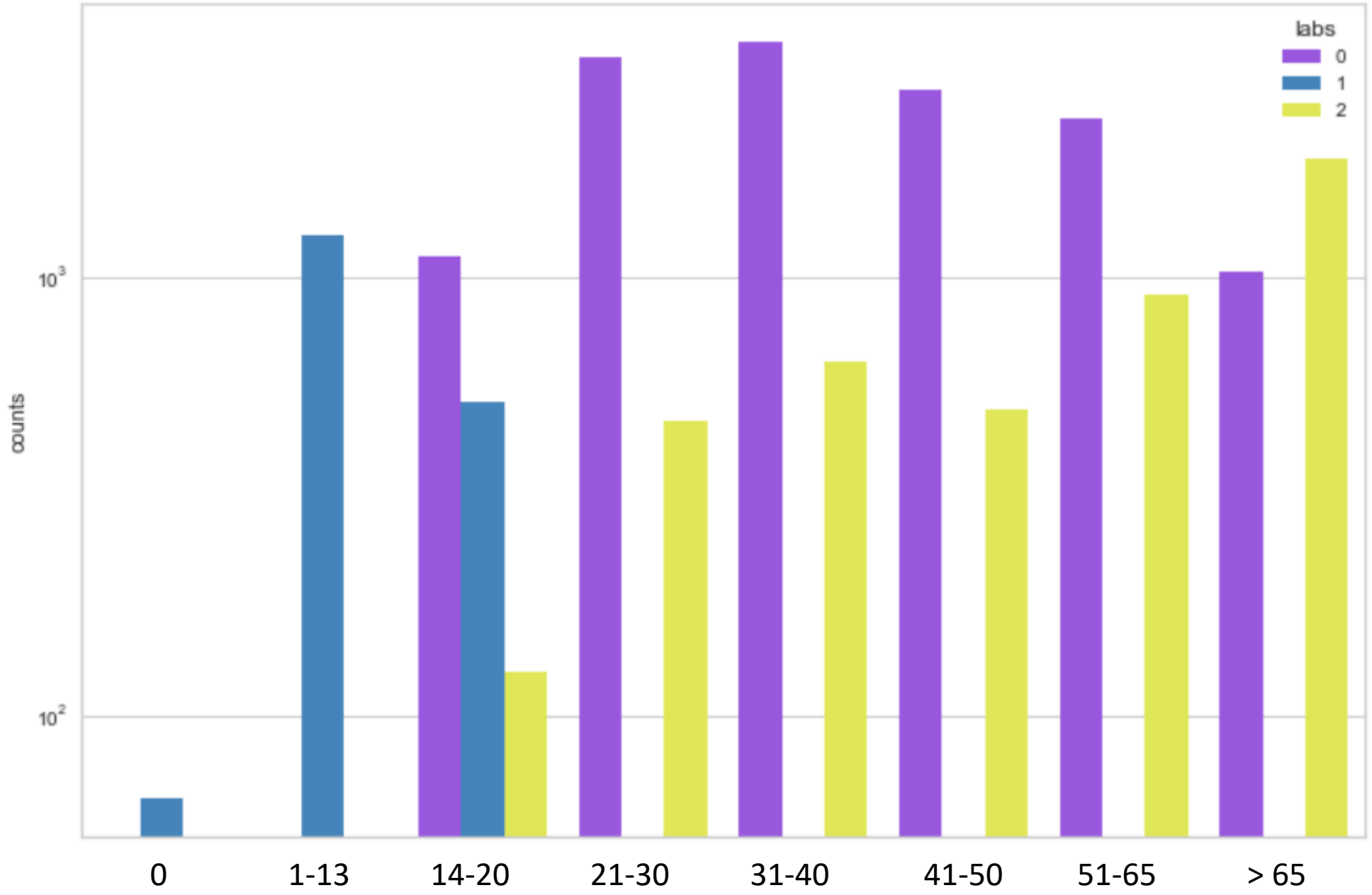
Klaster 0 – osoby o dużym przychodzie (13660 obserwacji)



Klasteryzacja aglomeracyjna – 3 klastry



wiek



Charakteryzacja klastrów

Klaster 1
osoby bez przychodu

- głównie dzieci i młodzież
- raczej nie są w żadnych relacjach ani związkach małżeńskich

Klaster 2
osoby z małym przychodem

- osoby dorosłe, których dochód jest przeważająco mały/średni lub nie mają przychody
- znajduje się tu sporo osób owdowiałych, a także wyraźnie starszych, być może na emeryturze

Klaster 0
osoby z dużym przychodem

- osoby dorosłe, których dochód jest przeważająco średni/duży/bardzo duży
- osoby w średnim wieku (20-40 lat)
- czynnie pracujące, często też służące w wojsku
- raczej zamężne

Dziękujemy za uwagę!