US Census Data

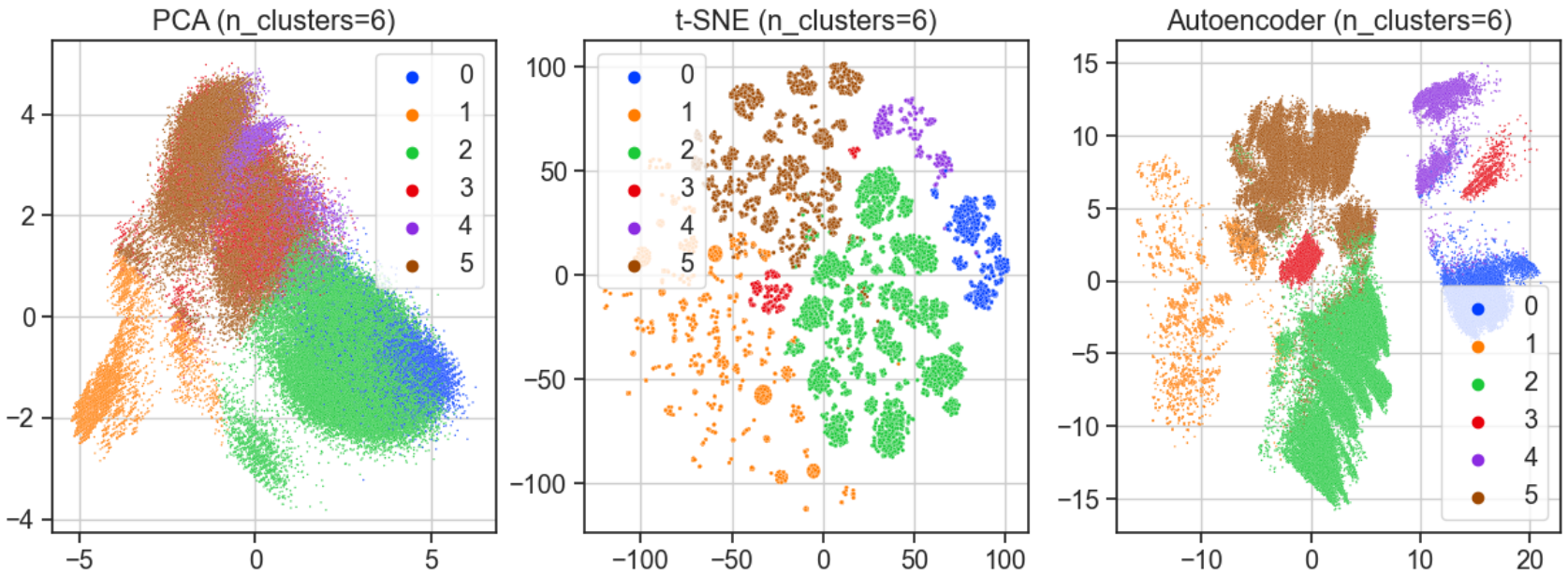
Klasteryzacja zbioru danych

Grzegorz Zakrzewski
Tomasz Modzelewski

MiNI

2022

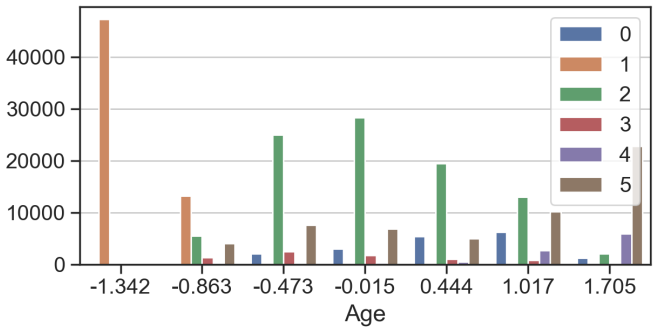
KMeans - n_clusters=6

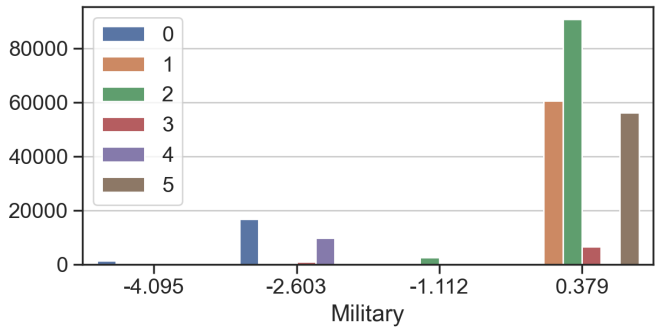


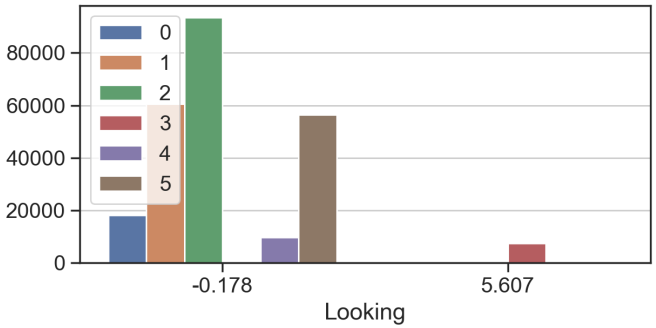
KMeans - interpretacja

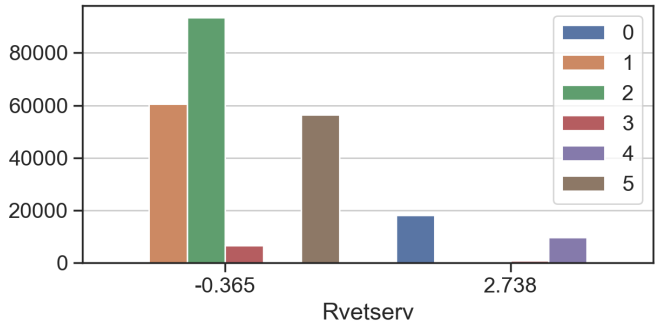
- Sprawdziliśmy od 2 do 16 klastrów.
- Jakość podziału mierzyliśmy *silhouette score* oraz *davies bouldin score*.
- Najlepsze wyniki zostały osiągnięte dla 3, 4, 5 i 6 klastrów.
- Osiągnięty podział możemy interpretować jako:
  - 1 - dzieci (25%),
  - 2 - normalni, pracujący dorośli (38%),
  - 5 - niepracujący, często z powodu wieku podeszłego (23%),
  - 0 - pracujący weterani (7%),
  - 3 - bezrobotni (3%),
  - 4 - weterani w podeszłym wieku, niepracujący (4%).

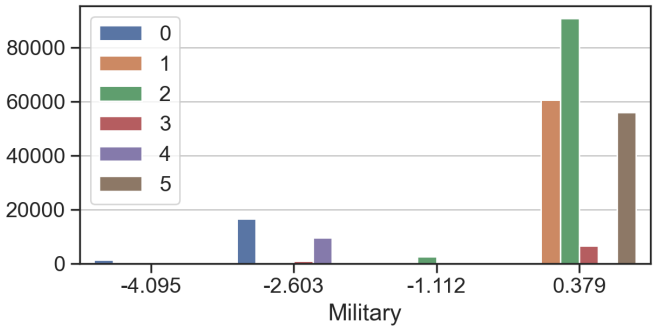
KMeans - interpretacja

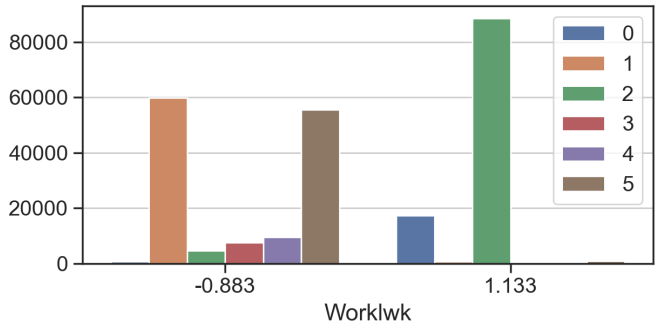










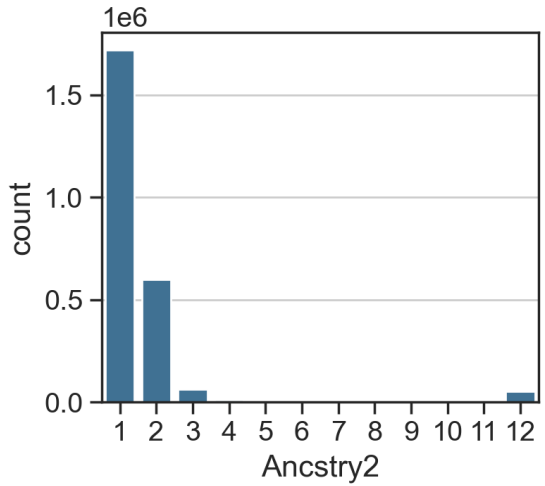


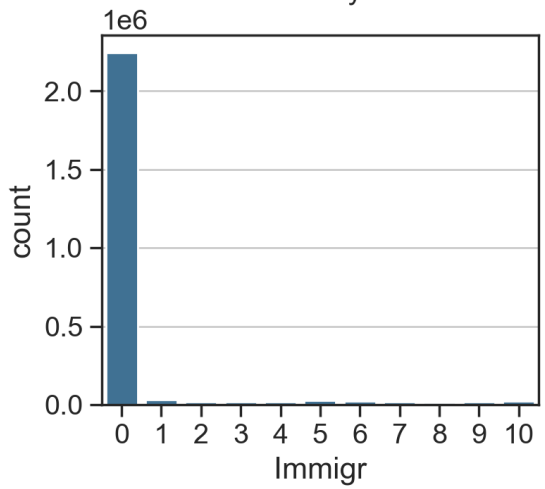
Eksploracja danych

- Zbiór danych składa się z:
  - 68 zmiennych,
  - 2 458 285 obserwacji.
- Wszystkie zmienne są kategoryczne:
  - zmienne `false`, `true` jak `Sex`,
  - zmienne `N/A`, `true`, `false` jak `Mobillim`,
  - zmienne nominalne jak `Marital`,
  - zmienne uporządkowane jak `Age`.
- Zbiór danych był wstępnie przetworzony.
- Nie ma braków danych ani duplikatów.

Problemy ze zmiennymi

- Dużo kategorii o znikomej liczności.
- Informacje niemożliwe do odtworzenia i interpretacji, np. `Ancstry`, `Industry`.
- Kategorie bez naturalnego porządku, np. `Class`, `Marital`.
- Nieinteresujące, zbyt szczegółowe informacje, np. `Hispanic`, `Feb55`
- Pokrywanie się z innymi zmiennymi, np. `POB`, `RPOB`.





Preprocessing

- Łączenie kategorii, np:
  - `Marital`: 0 - *married*, 1-4 - *not married*.
- Usuwanie kolumn:
  - powtarzających się,
  - silnie skorelowanych,
  - zbyt szczegółowych, niemożliwych do interpretacji,
  - o dużej liczbie nieuporządkowanych, mało licznych kategorii.
- Skalowanie zmiennych za pomocą `StandardScaler`.

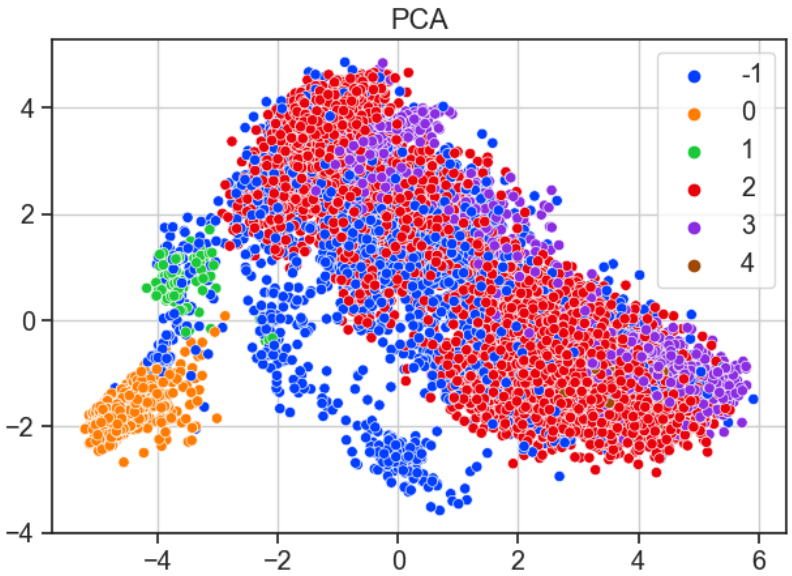
- W wyniku preprocessingu zostało 28 zmiennych.
- Zmienne nieuporządkowane zostały zredukowane do binarnych lub wyrzucone.

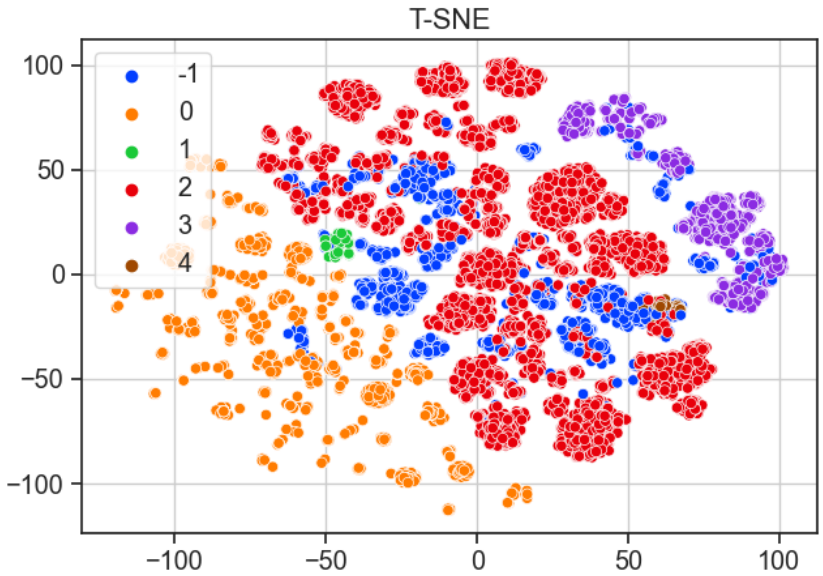
Ogólny zarys

- Cztery metody klasteryzacji:
  - KMeans,
  - DBSCAN,
  - Agglomerative,
  - Birch.
- Wizualizacja klastrów za pomocą:
  - PCA dla dwóch komponentów - ok. 50% wariancji,
  - T-SNE,
  - autokodera zaimplementowanego w Keras.
- Interpretacja klastrów za pomocą:
  - modelu regresji logistycznej wytrenowanego do przewidywania przypisanych klastrów - spojrzenie na przyznane współczynniki,
  - wykresów słupkowych.

DBSCAN

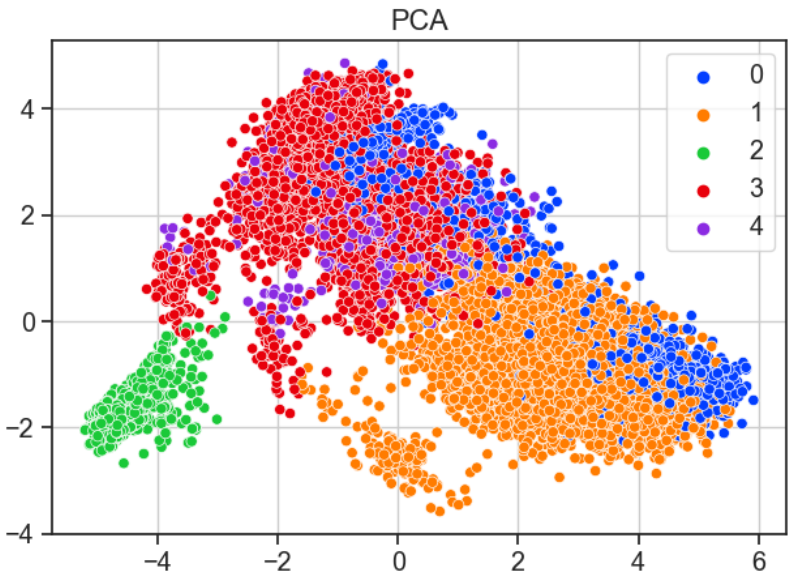
- Problem ze znalezieniem odpowiedniego eps.
- Dużo outlierów.
- Dwa nieliczne, bezużyteczne klastry.
- Niemożliwość w interpretacji.

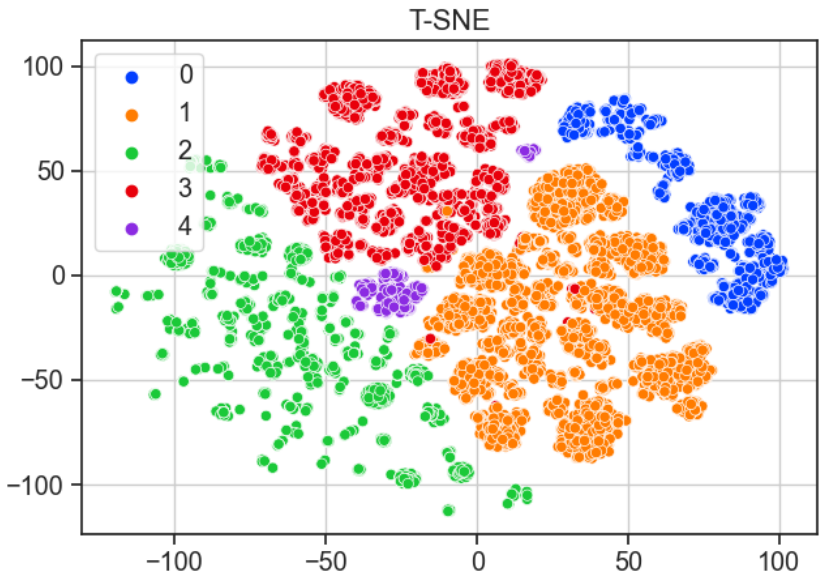




Agglomerative Clustering

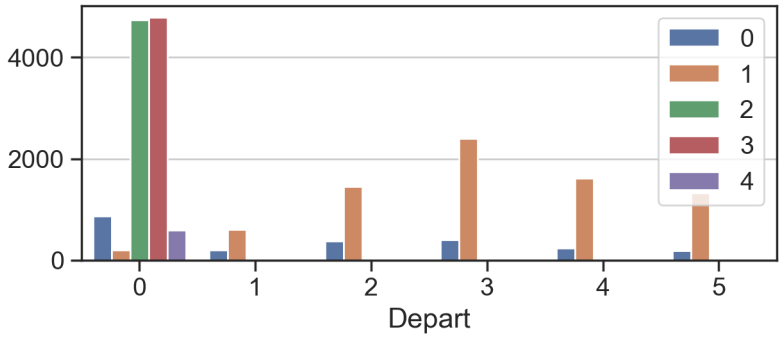
- Wynik wybrany na podstawie *silhouette score*.
- 5 klastrów, kryterium łączenia `ward`.
- Obiecująco również wyglądały 3 klastry.
- Kryteria łączenia `average` i `single` okazały się bardzo słabe.

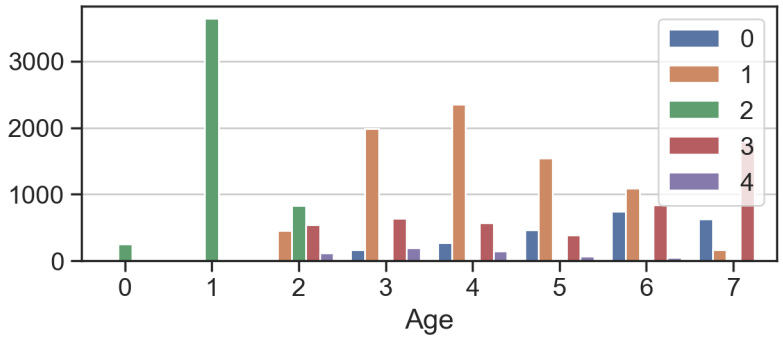


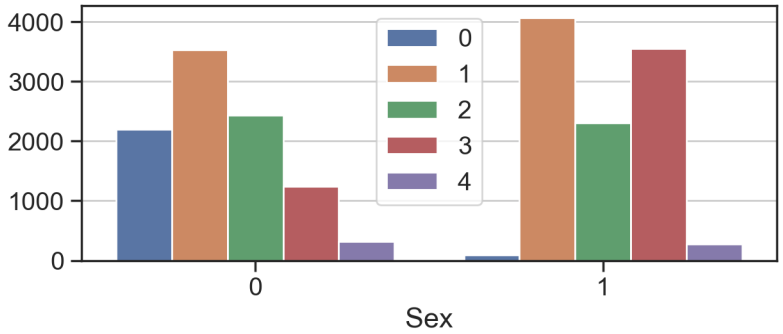


Agglomerative Clustering

- 0 - dorośli mężczyźni
- 1 - dorośli pracujący
- 2 - dzieci
- 3 - dorośli niepracujący, kobiety mające dzieci
- 4 - niepracujący, w średnim wieku

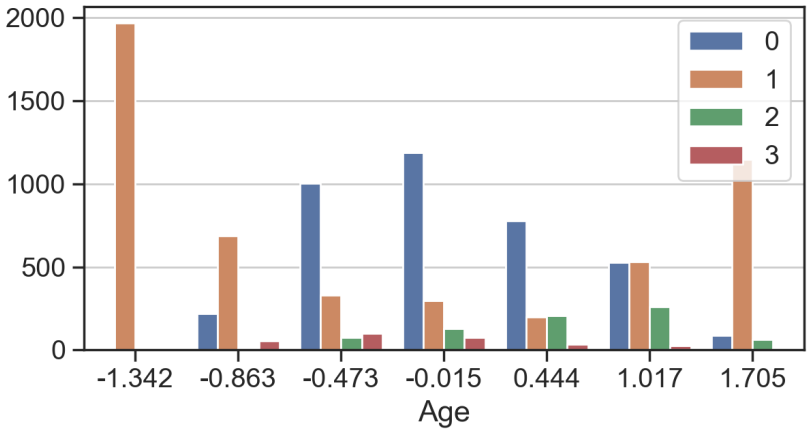


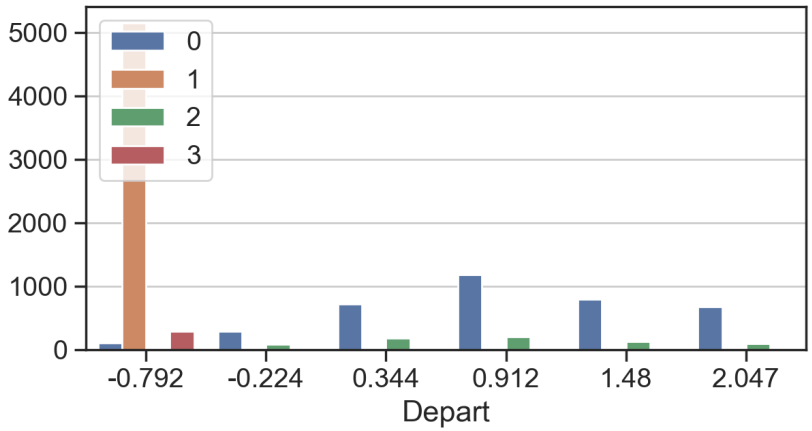


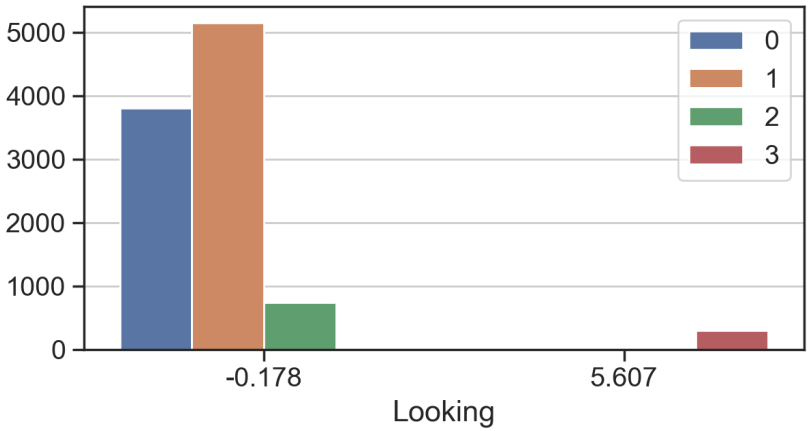


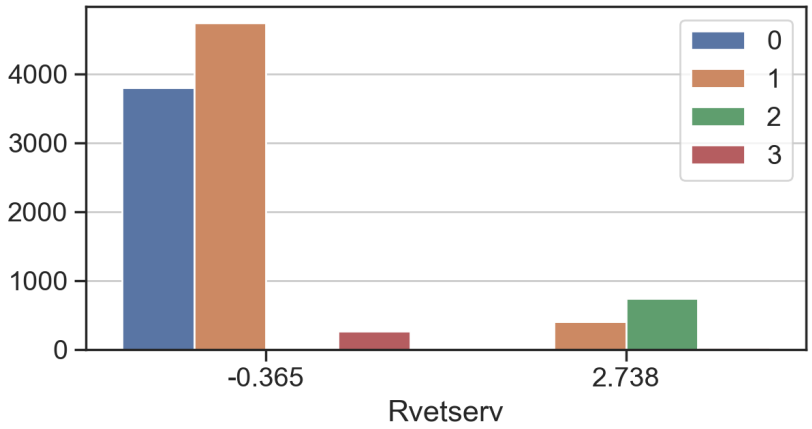
Birch

0 - pracujący dorośli
1 - osoby niepracujące (podeszły wiek)
2 - bezrobotni (szukający pracy)
3 - weterani wojenni









Podsumowanie

- Wszystkie algorytmy klasteryzujące zwracały uwagę na:
  - status zawodowy,
  - wiek,
  - stan cywilny,
  - przeszłość w wojsku,
  - płeć.
- Najlepszy podział uzyskaliśmy korzystając z KMeans dla sześciu klastrów.