Wstęp do uczenia maszynowego projekt 2

Wydział Matematyki i Nauk Informacyjnych Politechniki Warszawskiej

Jakub Fołtyn
Paulina Jaszczuk

Problem:

- Klasteryzacja zbioru klientów strony typu e-commerce (oryginalnie dwie klasy: klient, który dokonał transakcji podczas sesji lub nie)

Dane

- Zbiór danych dotyczących pojedynczych sesji klientów strony typu e-commerce
- link do zbioru danych

Plan pracy

- EDA
- Inżynieria cech
- Modelowanie i ocena jakości klastrowania

EDA - informacje ogólne

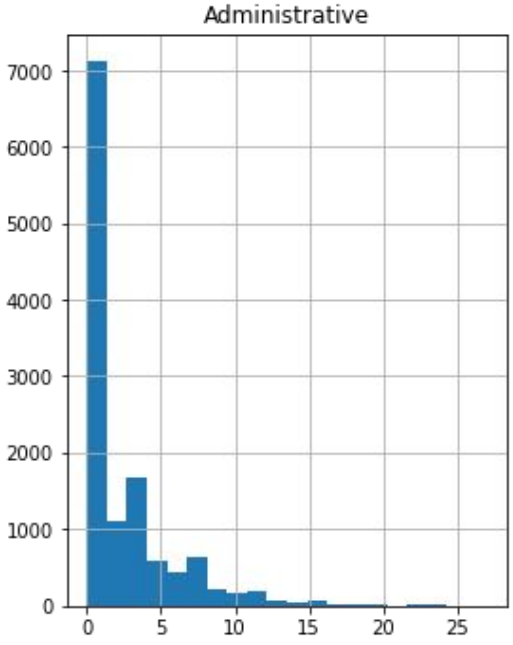
- Dane dotyczą aktywności użytkowników na podstronach strony typu e-commerce.
- Każdy wiersz to osobna sesja osobnego użytkownika.
- Obserwacje zawierają dane dotyczące typu odwiedzonych podstron, czasu spędzonego na nich, pewnych współczynników charakteryzujących podstrony, a także informacje na temat samego użytkownika.
- Targetem jest zmienna Revenue, która mówi nam o tym, czy podczas sesji została sfinalizowana transakcja (głównie czy klient coś kupił).

- 12330 obserwacji
- 3 cechy kategoryczne
- 13 cech numerycznych
  (w tym 4 określające kategorie)
- brak wartości brakujących

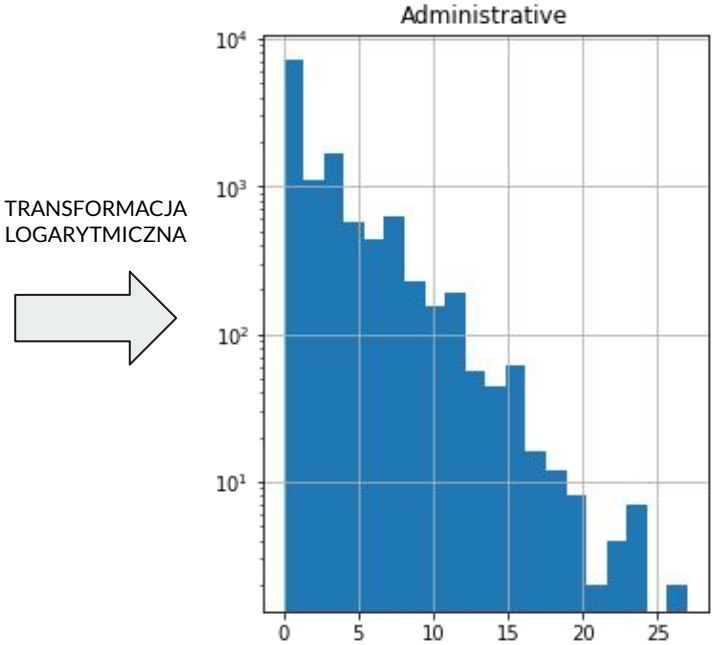
EDA - rozkłady danych numerycznych

- W znacznej większości przypadków wgląd w dane zaburzało to, że wartość 0 występuje kilkakrotnie razy więcej niż inne wartości. Aby sobie z tym poradzić zlogarytmowaliśmy dane.

Dane oryginalne



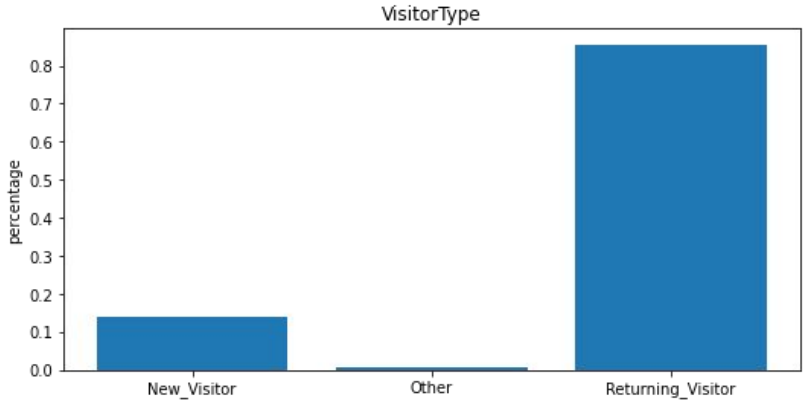
TRANSFORMACJA LOGARYTMICZNA

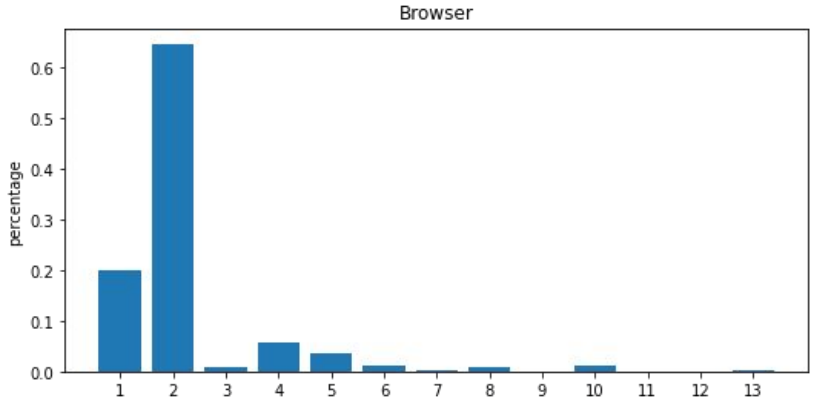


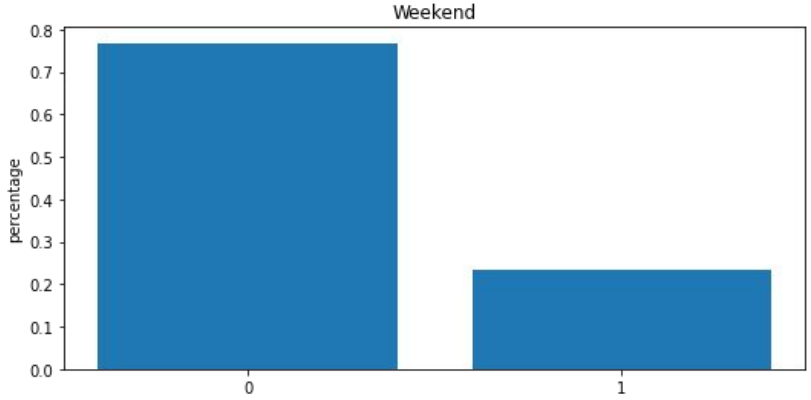
Dane zlogarytmowane

EDA - rozkłady danych kategorycznych

- W przypadku większości zmiennych widać dominację jednej z wartości.
- Brak bardziej szczegółowego opisu danych uniemożliwia głębszą interpretację niektórych cech.
- Większość sesji miała miejsce w maju i listopadzie, w dni nie weekendowe oraz były one dokonane przez powracających użytkowników.

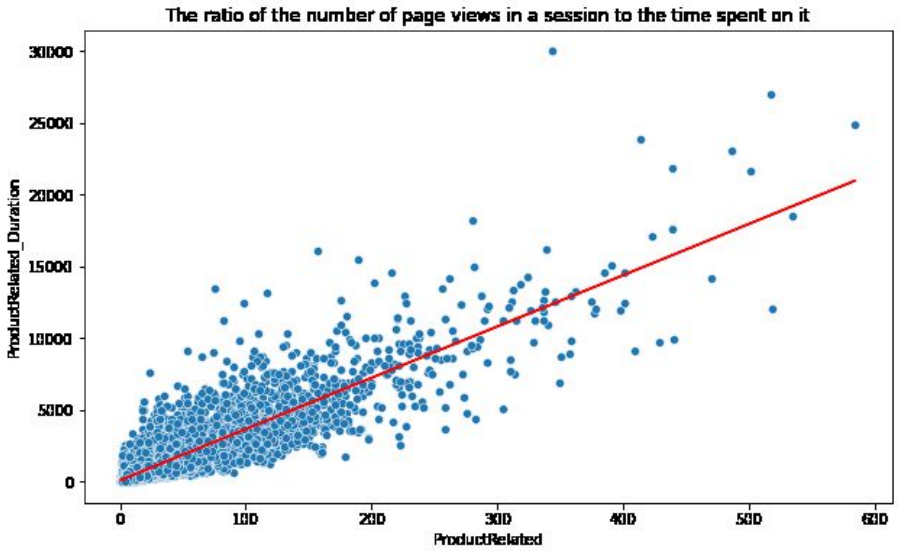


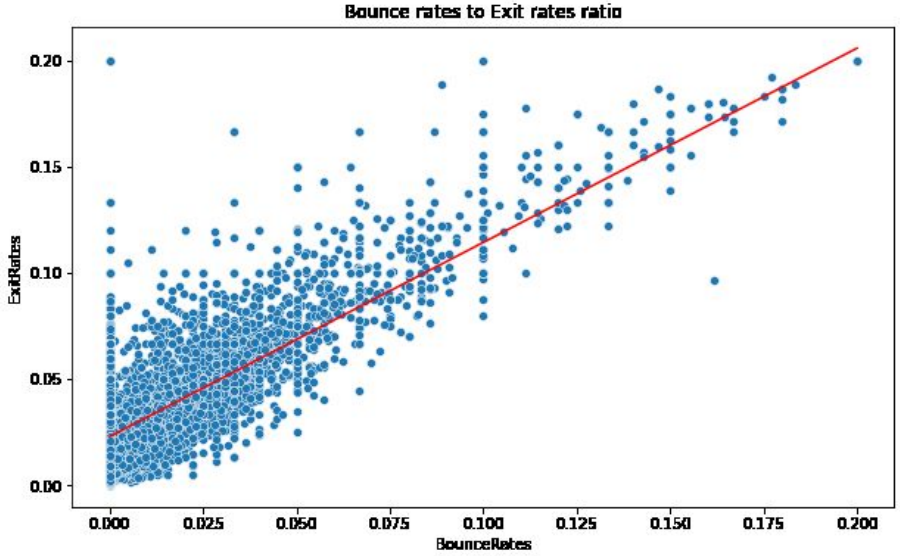




EDA - korelacja zmiennych

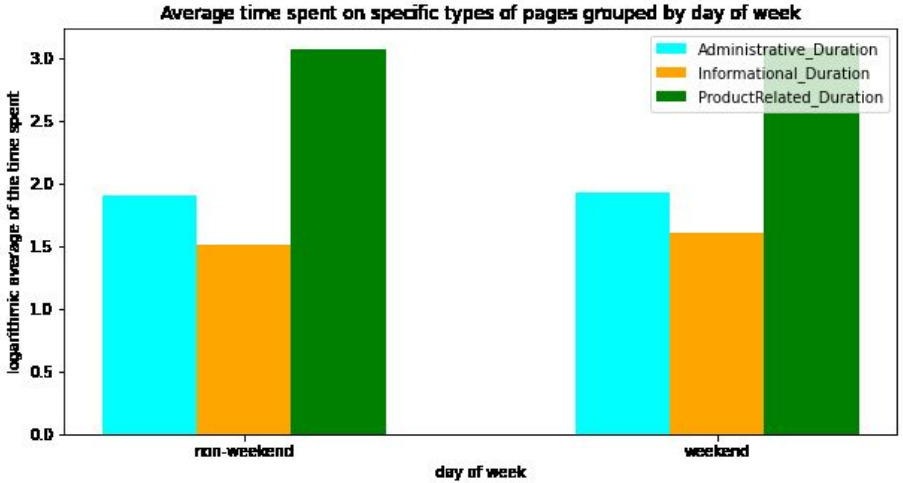
- Korelacja pomiędzy zmiennymi typu ‘page’ a ‘page_duration’, co jest dosyć intuicjne - im więcej stron danego typu odwiedzamy, tym więcej czasu na nich spędzamy.
- Korelacja między ‘ExitRates’ a ‘BounceRates’.

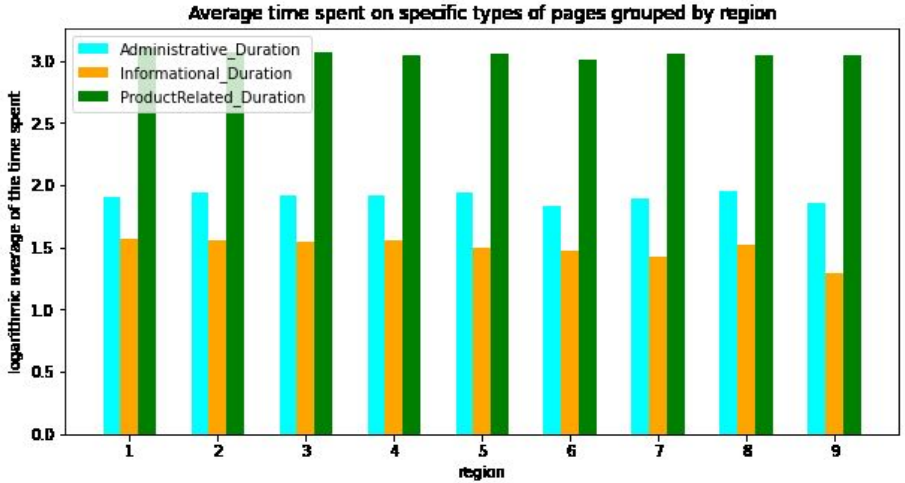




EDA - szczegółowa analiza

- Wśród danych pogrupowanych po cechach użytkownika lub sesji zdecydowanie przoduje czas spędzony na podstronach z produktami.
- Rozkłady są we wszystkich przypadkach bardzo podobne i nie występują w nich żadne ciekawy zjawiska.

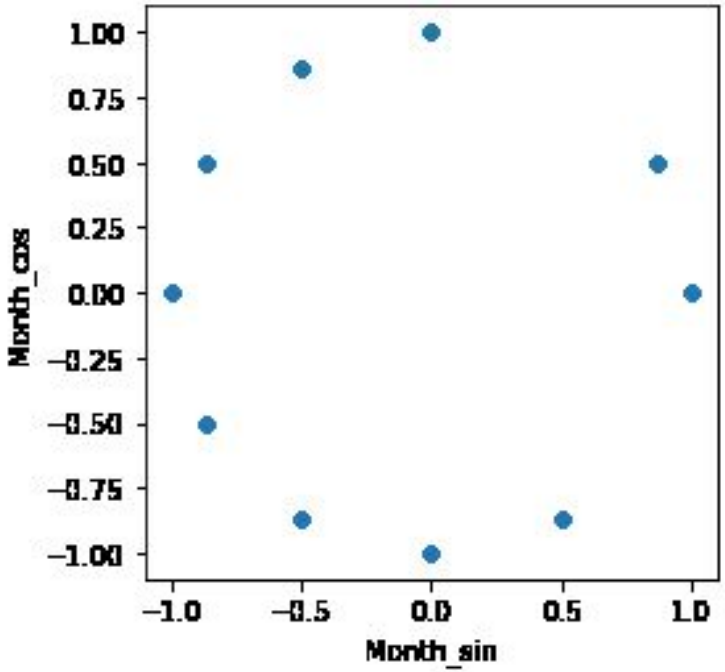




Inżynieria cech - kodowanie cech kategorycznych

Zakodowaliśmy 3 cechy kategoryczne z naszego zbioru:

- ‘Weekend’ zmapowaliśmy na wartości 1/0,
- ‘VisitorType’ zakodowaliśmy za pomocą One Hot Encoding,
- ‘Month’ zakodowaliśmy za pomocą enkodingu cyklicznego.

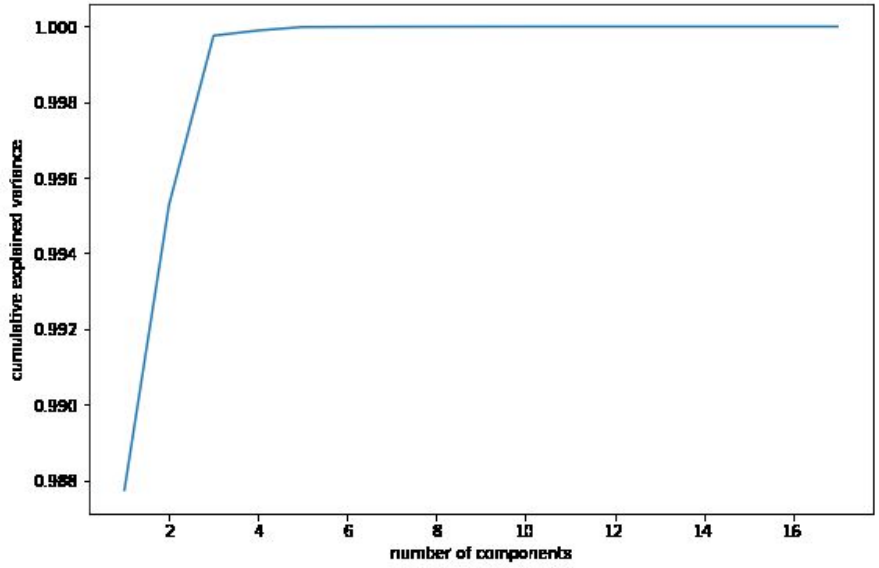


Inżynieria cech - transformacja logarytmiczna i standaryzacja

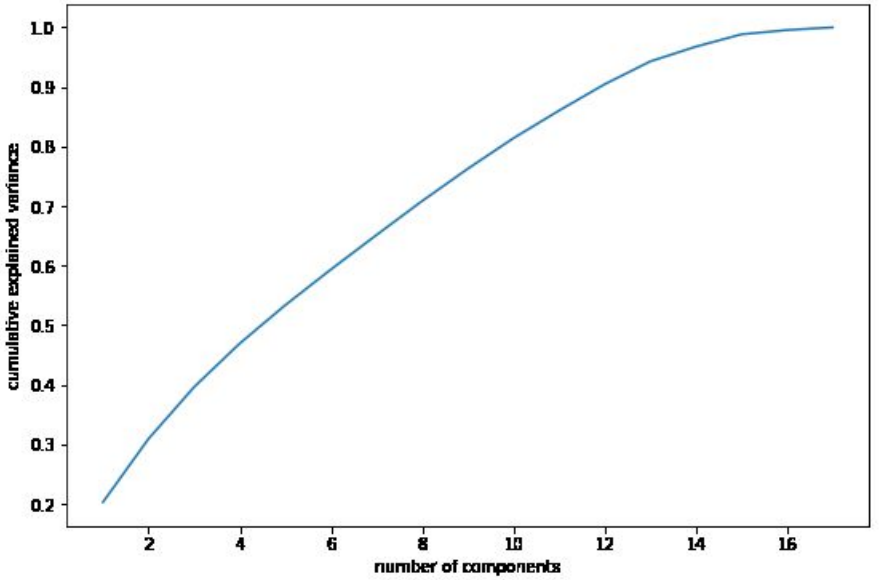
- Z powodu nierównomiernego rozkładu danych utworzyliśmy w celach eksperymentalnych dodatkowe ramki danych: ze zlogarytmowanymi kolumnami dot. stron oraz ze zlogarytmowanymi kolumnami dot. stron i wskaźników stron.
- Wystandaryzowaliśmy dane z powodu dużego rozrzutu w niektórych kolumnach, by działanie PCA było wiarygodne.

Inżynieria cech - redukcja wymiarowości

- Na danych przeprowadziliśmy PCA.
- Wyniki na oryginalnej ramce były mocno niewiarygodne.
- Przyczyną okazał się brak standaryzacji.



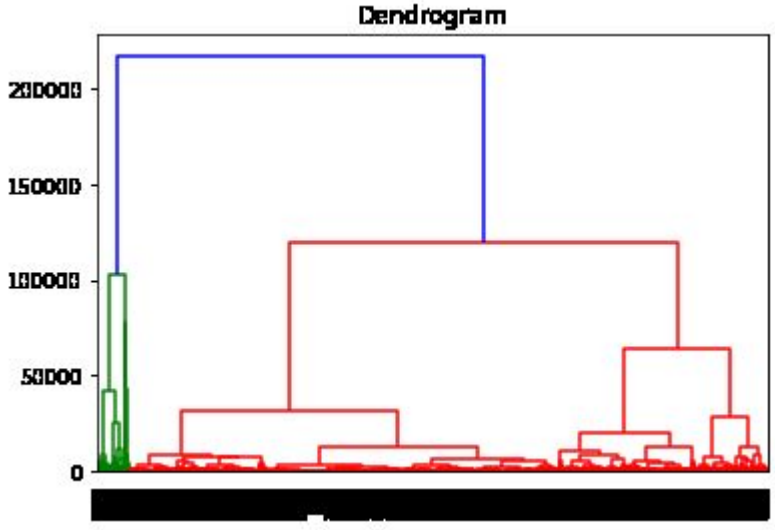
STANDARYZACJA

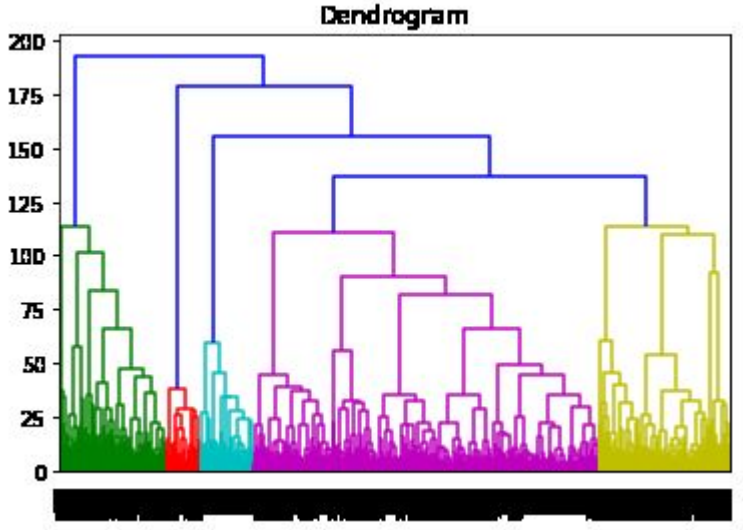


Modelowanie - wyznaczenie optymalnej liczby klastrów

- Dla danych oryginalnych i zlogarytmowanych dendrogramy i wykresy łokciowe jasno wskazywały 2 jako optymalną liczbę klastrów.
- Dla danych ustandaryzowanych trudno było dobrać odpowiednią liczbę.
- Finalnie zdecydowaliśmy się na 2 klastry.

Dane oryginalne



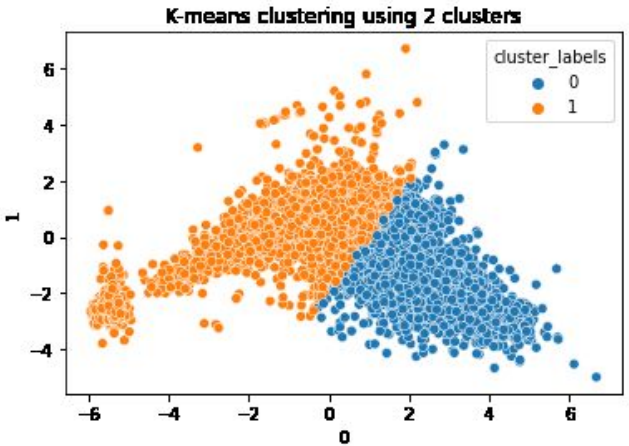


Dane ustandaryzowane

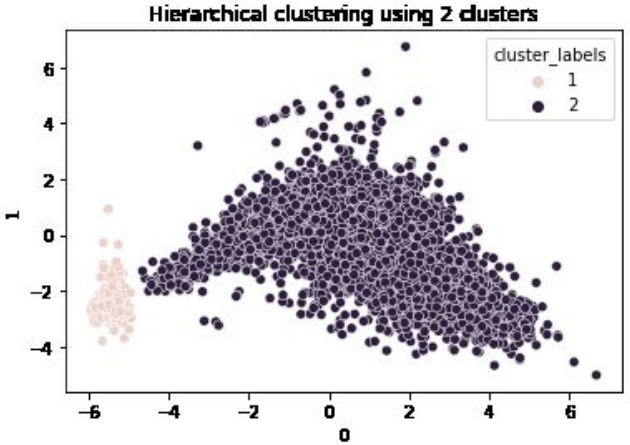
Modelowanie - wizualizacja dla 2 wymiarów

- Wizualizacja klasteringu metodą hierarchiczną i k-means danych sprowadzonych do 2 wymiarów przy pomocy PCA.

Dane zlogarytmowane ustandaryzowane



Dane zlogarytmowane ustandaryzowane



Dane oryginalne ustandaryzowane