Online Shoppers' Purchase Intention

...

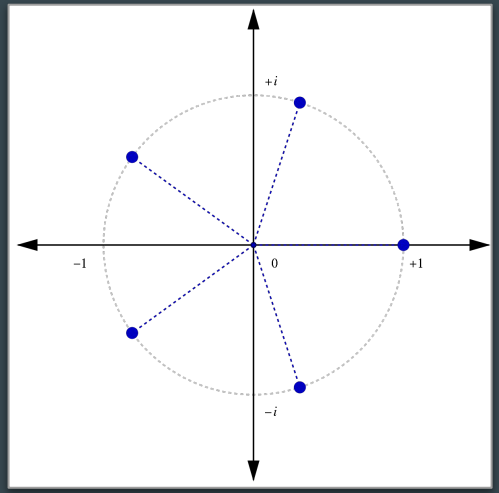
Patryk Tomaszewski
Mateusz Ziemła

Zbiór danych

`Administrative` - This is the number of pages of this type (administrative) that the user visited.
`Administrative_Duration` - This is the amount of time spent in this category of pages.
`Informational` - This is the number of pages of this type (informational) that the user visited.
`Informational_Duration` - This is the amount of time spent in this category of pages.
`ProductRelated` - This is the number of pages of this type (product related) that the user visited.
`ProductRelated_Duration` - This is the amount of time spent in this category of pages.
`BounceRates` - The percentage of visitors who enter the website through that page and exit without triggering any additional tasks.
`ExitRates` - The percentage of pageviews on the website that end at that specific page.
`PageValues` - The average value of the page averaged over the value of the target page and/or the completion of an eCommerce transaction.
`SpecialDay` - This value represents the closeness of the browsing date to special days or holidays (eg Mother's Day or Valentine's day) in which the transaction is more likely to be finalized.
`Month` - Contains the month the pageview occurred, in string form.
`OperatingSystems` - An integer value representing the operating system that the user was on when viewing the page.
`Browser` - An integer value representing the browser that the user was using to view the page.
`Region` - An integer value representing which region the user is located in.
`TrafficType` - An integer value representing what type of traffic the user is categorized into.
`VisitorType` - A string representing whether a visitor is New Visitor, Returning Visitor, or Other.
`Weekend` - A boolean representing whether the session is on a weekend.

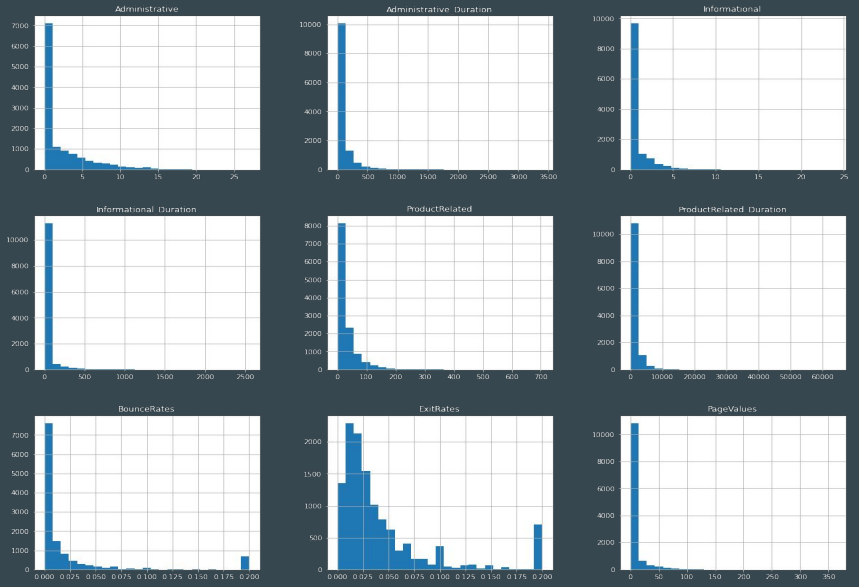
Przygotowanie danych

- tekstowe zmienne kategoryczne - one-hot encoding — Weekend, VisitorType
- numeryczne zmienne kategoryczne - one-hot encoding — OperatingSystems, Browser, Region, TrafficType
- zmienna cykliczna - rozbicie na sin i cos — Month

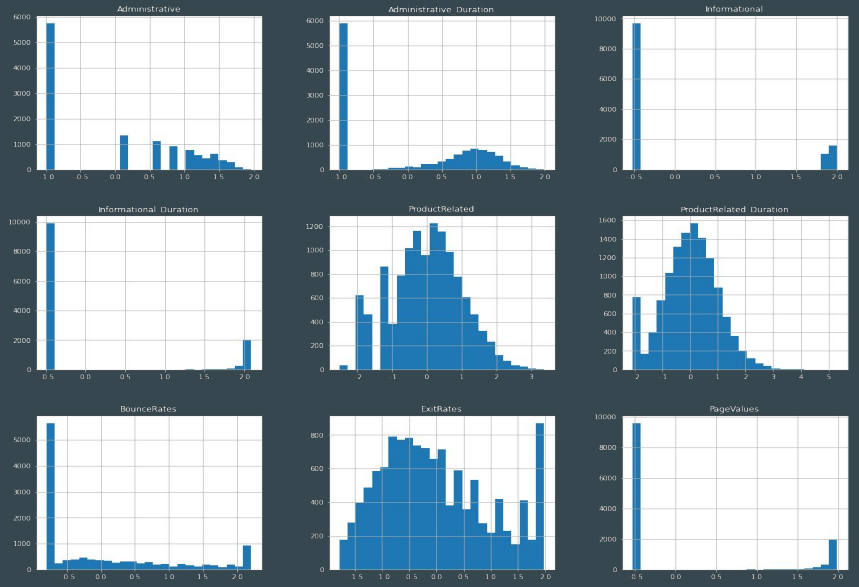


Przygotowanie danych

- zmienne mocno prawoskośne - przekształcenie Yeo-Johnson
- normalizacja do [-1,1]

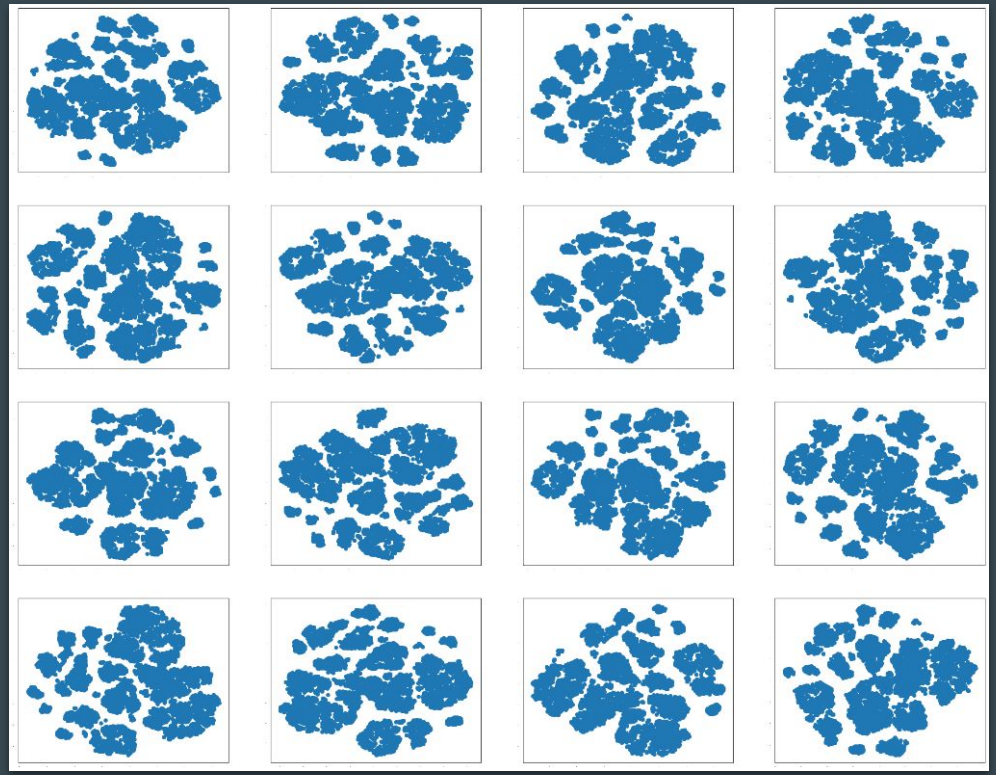


Przed przekształceniem



Po przekształceniu

Wizualizacja zbioru

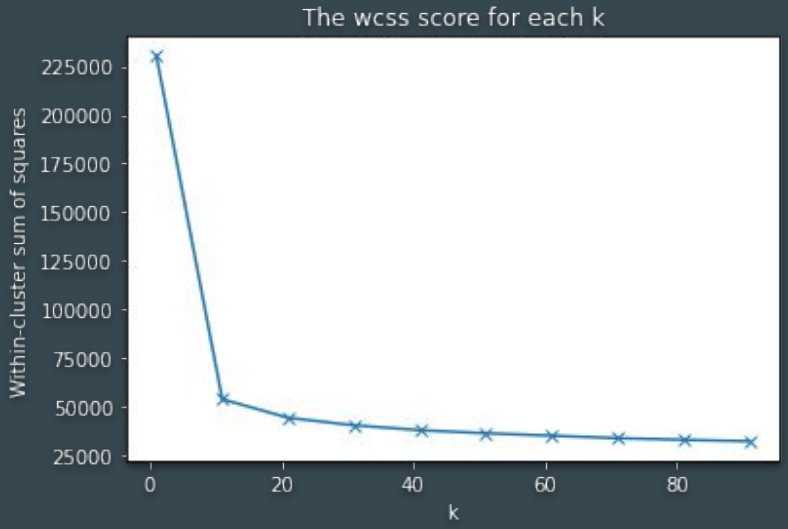


Badane modele

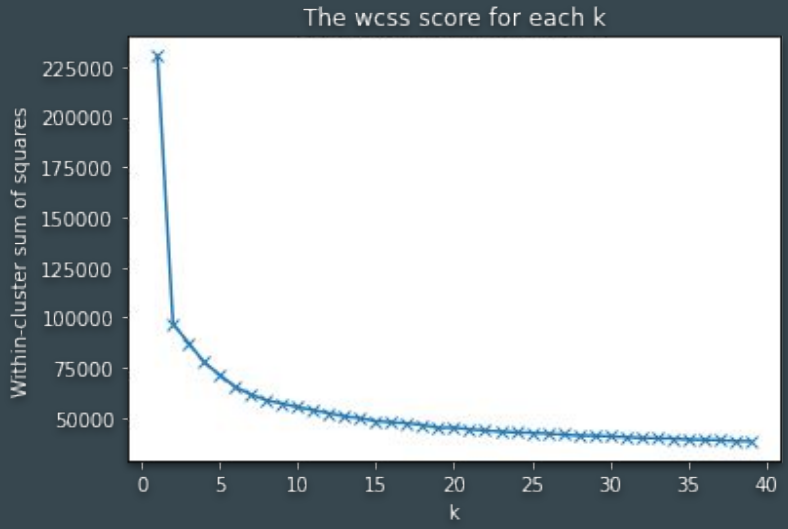
- KMeans
- Agglomerative clustering
- Optics

Optymalna ilość klastrów

Metoda łokcia



Od 1 do 100, skok o 10

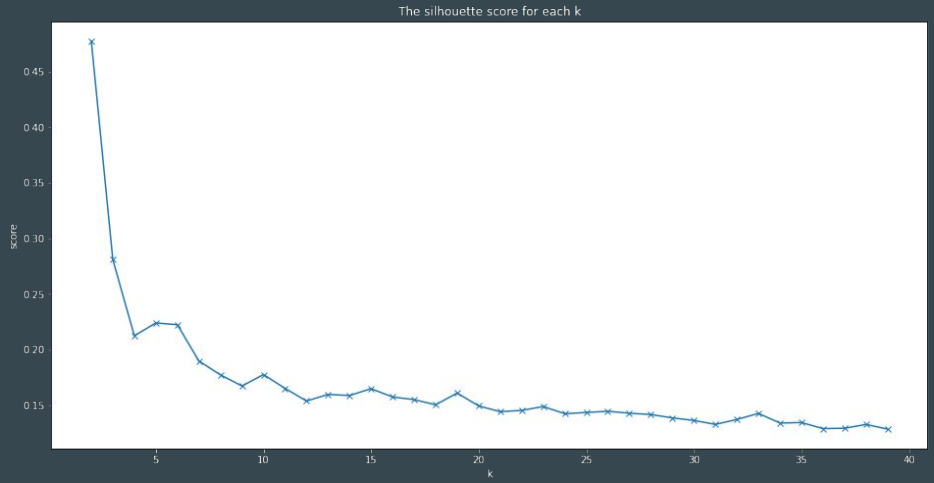


Od 1 do 40, skok o 1

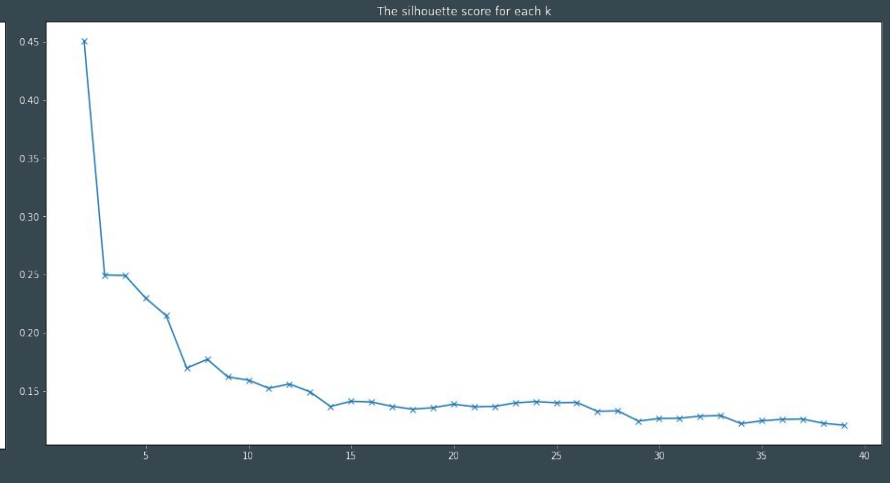
Optymalna ilość klastrów

Silhouette score

KMeans



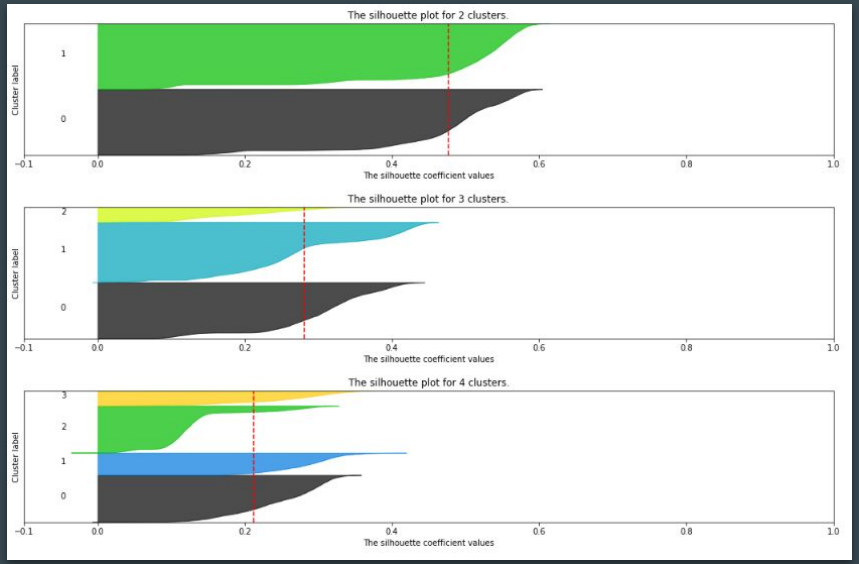
Agglomerative Clustering



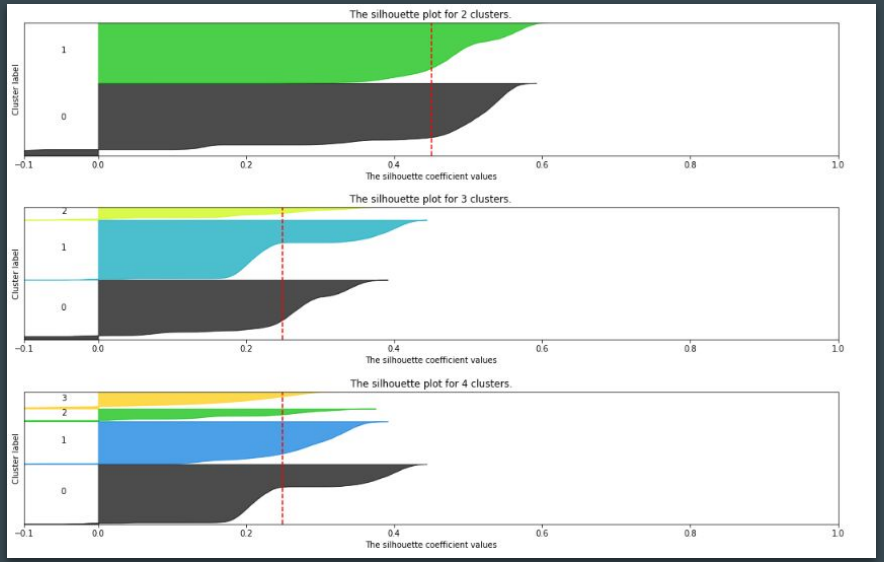
Optymalna ilość klastrów

Silhouette score

KMeans



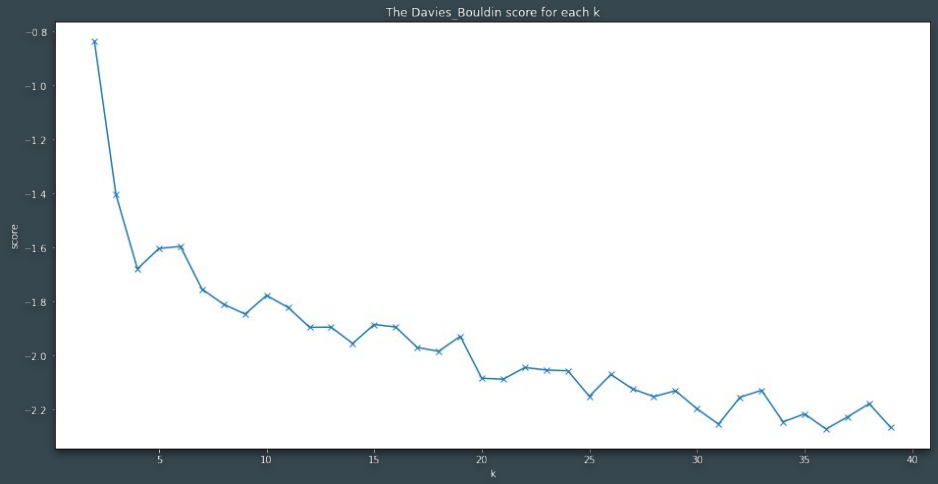
Agglomerative Clustering



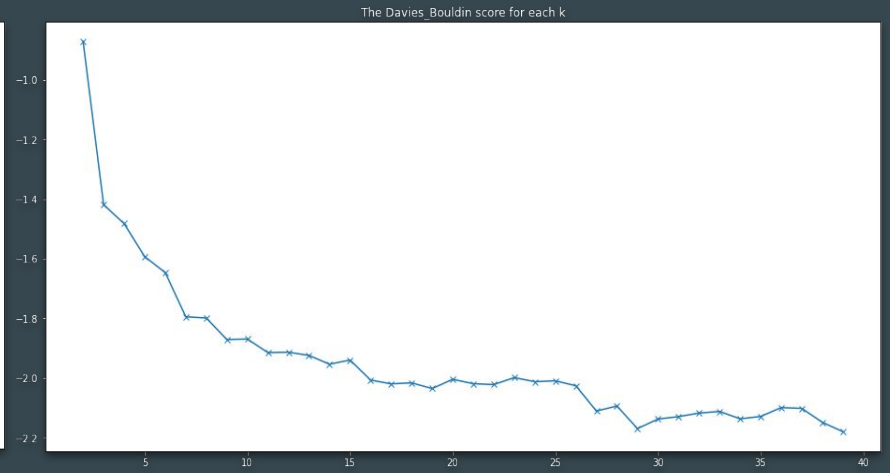
Optymalna ilość klastrów

Davies-Bouldin score

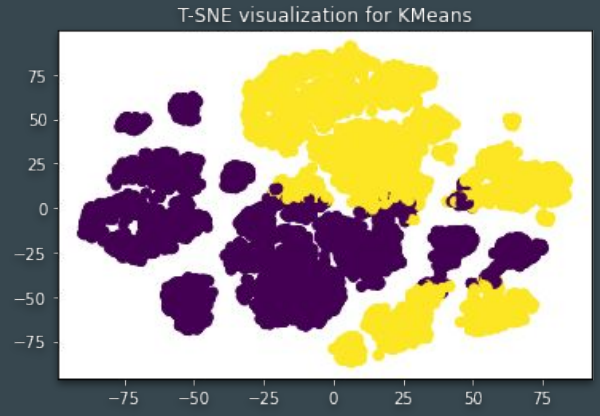
KMeans

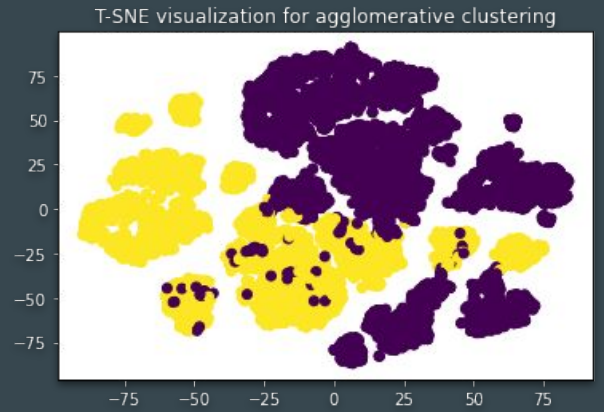


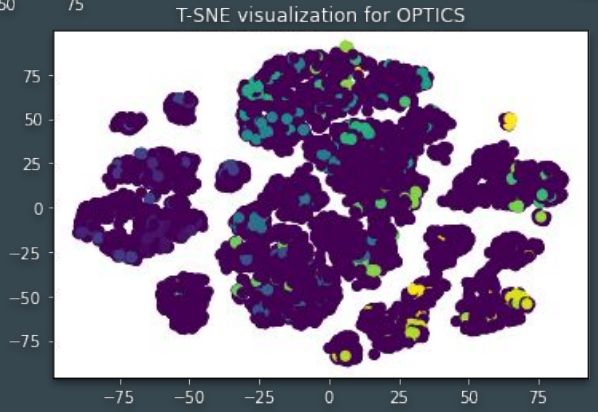
Agglomerative Clustering



Wizualizacja klastrów

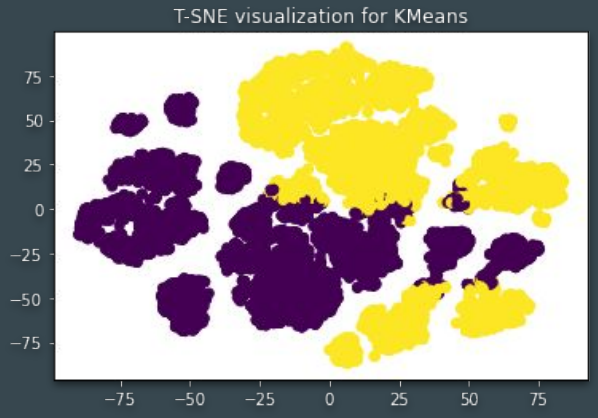




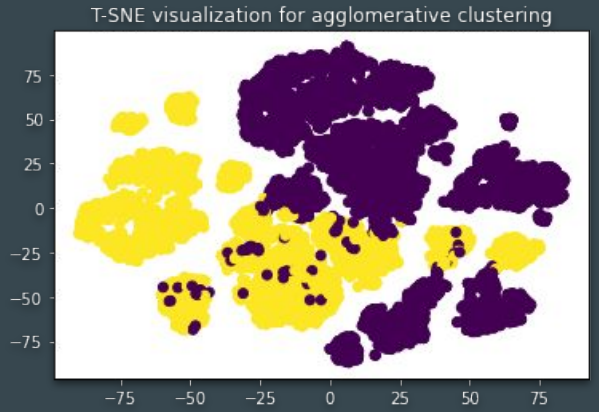


82% rekordów niezgrupowanych

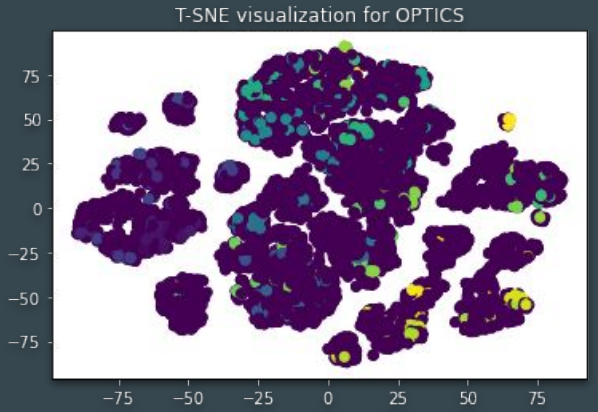
Wybór modelu



Silhouette score: 0.240
Davis-Bouldin score: 1.631

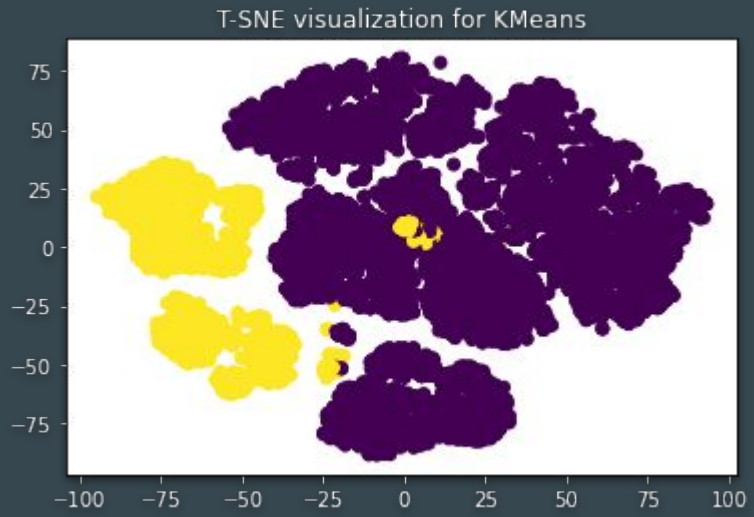


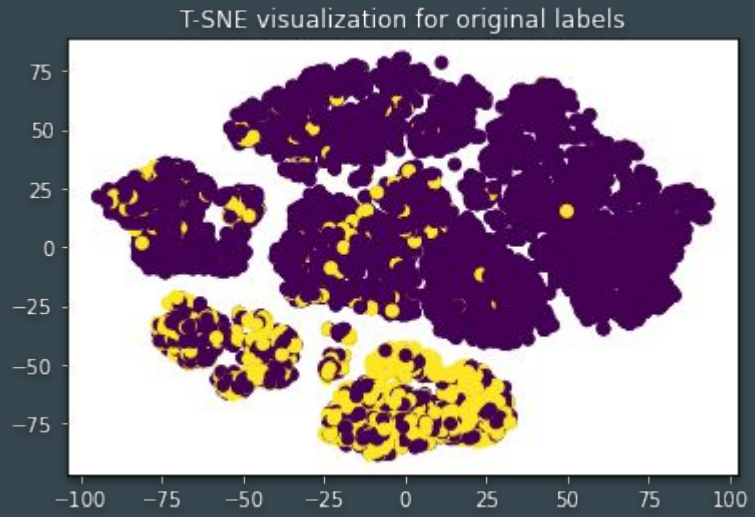
Silhouette score: 0.235
Davis-Bouldin score: 1.672



Silhouette score: 0.355
Davis-Bouldin score: 1.186
82% rekordów niezgrupowanych

Porównywanie z labelami ze zbioru

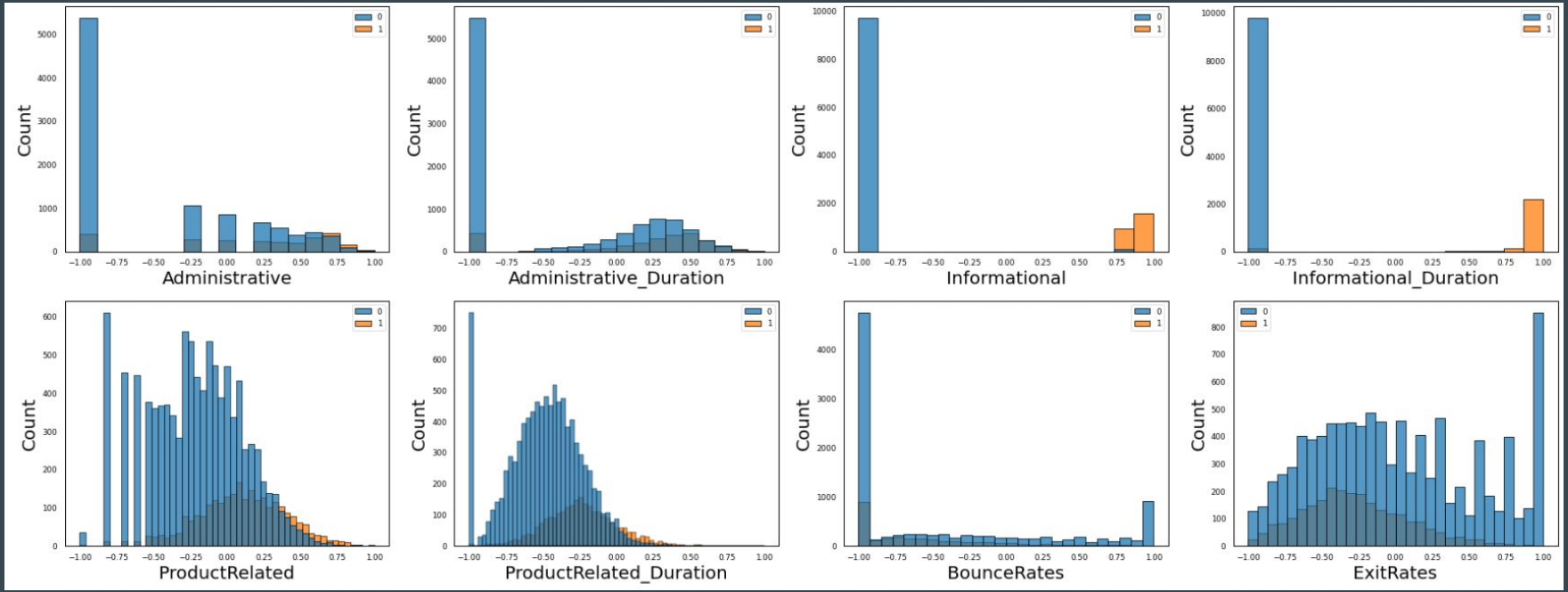




	Brak zakupu	**Zakup**
Klaster 1	68.9%	10.6%
Klaster 2	15.6%	4.9%

Accuracy: 73.8%

Analiza uzyskanych klastrów



Analiza klastrów ze zbioru danych