|  | Politechnika Bydgoska im. J.J. Śniadeckich Wydział Telekomunikacji, Informatyki i Elektrotechniki | | |  |
| --- | --- | --- | --- | --- |
| **Przedmiot** | Algorytmy i eksploracja danych | | | |
| **Prowadzący** | dr inż. Michał Kruczkowski | | | |
| **Temat** | Uczenie bez nadzoru | | | |
| **Studenci** | Adam Szreiber, Cezary Naskręt | | | |
| **Nr ćw.** | 3 | **Data wykonania** | 20.10.2023 | |

## Cel ćwiczenia

Uczenie maszynowe bez nadzoru to obszar uczenia maszynowego, w którym algorytmy są trenowane na danych, które nie mają przypisanych etykiet lub wyników, w przeciwieństwie do uczenia maszynowego nadzorowanego, gdzie algorytmy uczą się na danych z etykietami (wynikami). W uczeniu maszynowym bez nadzoru celem jest odnalezienie ukrytych wzorców, struktury lub zależności w danych, a także grupowanie danych w klastry lub redukcja wymiarów.

## Przebieg laboratorium

### Zadanie 1

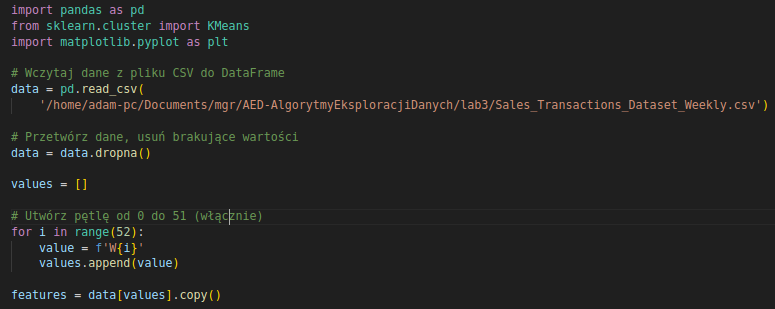
Zbiór danych: http://archive.ics.uci.edu/ml/machine-learning-databases/00396/ Dane dotyczące transakcji różnych produktów na przestrzeni 51-tygodni zawierają zarówno wartości oryginalne jak i znormalizowane.

### Zadanie 2

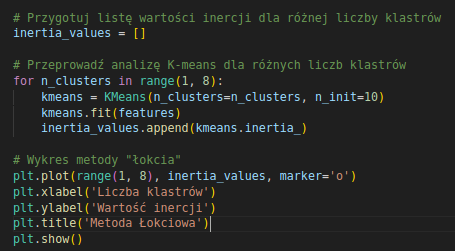
Wykonaj analizę skupień danych dla następujących algorytmów:

* K-MEANS
* K-MEANS++
* \*K-MEDIOD

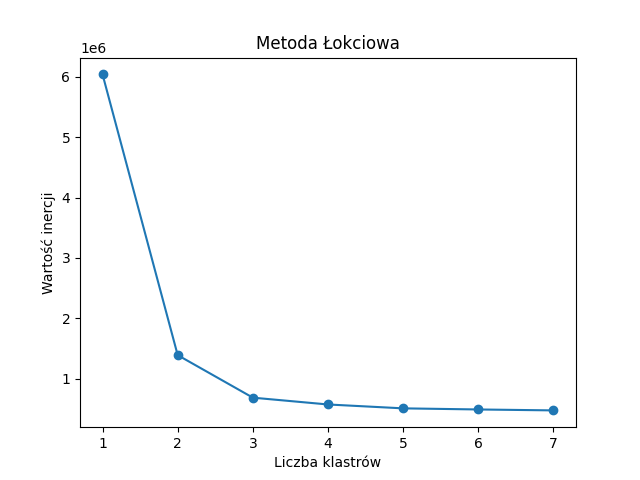
Na początku importuje wszystkie biblioteki.  
Następnie przygotowuję dane do obróbki.



Kod przedstawiający metodę łokciową - służącą w wyznaczaniu ilości klastrów.

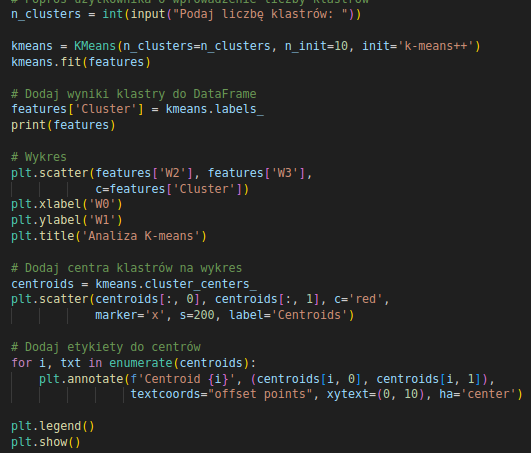


Na wykresie poszukaj "łokcia" (punkt zwany "punkt łokcia"), który jest miejscem, w którym spadek wartości funkcji kosztu zaczyna się wyhamowywać. Z wykresu wynika że przy wartości 3 jest wyrównanie na wykresie - co oznacza że to jest nasza ilość klastrów.

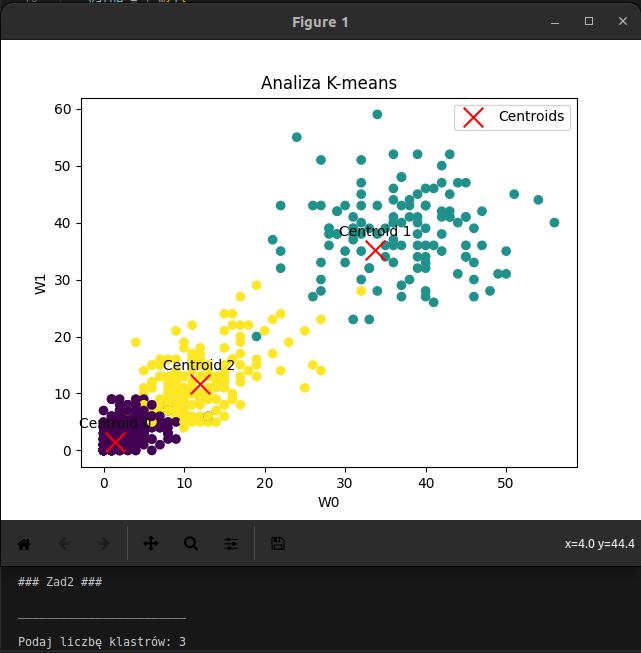


Kod służący wczytaniu ilości klastrów od użytkownika, a następnie wykonaniu uczenia modelu sieci KMeans.

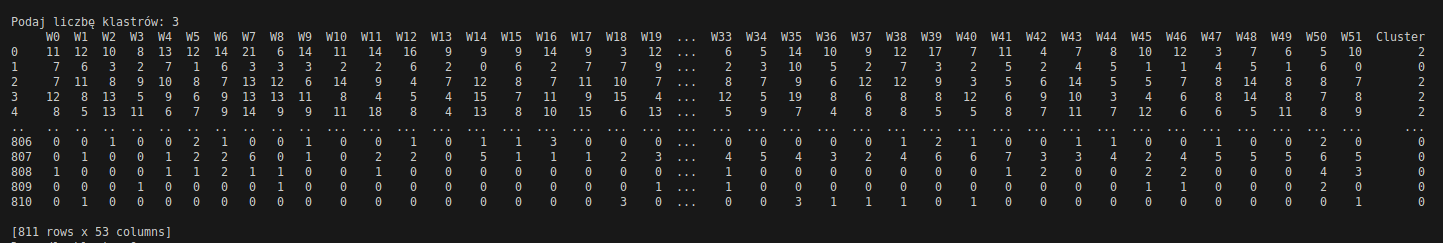
Do danych w `features` zostanie dodana kolumna ‘Cluster’ i w niej będą przechowywane etykiety klastra do którego przynależą.  
W ostatniej fazie zostaje tworzony wykres w raz z centroidami.



Przedstawienie wykresu z pogrupowanymi punktami.



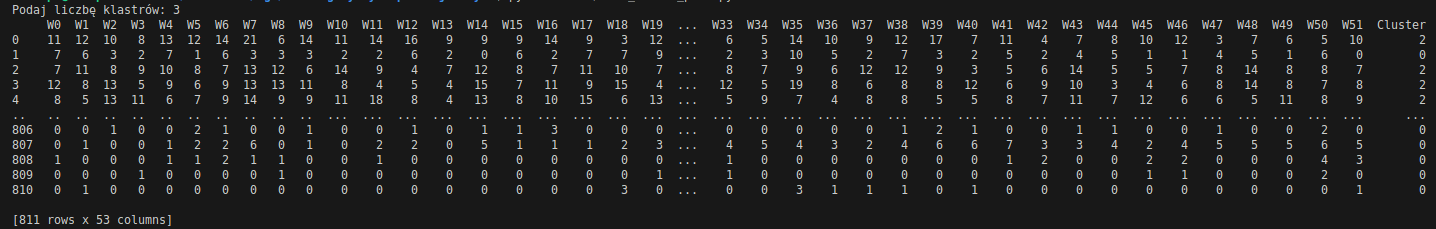
Widok danych wraz z etykietą klastra w ostatniej kolumnie.



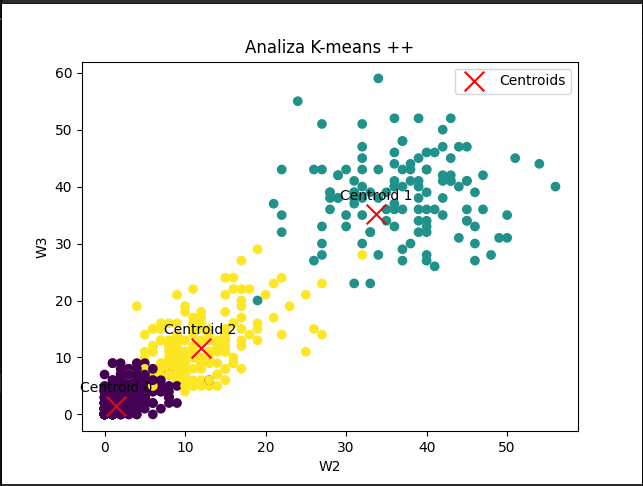
W kolejnym podpunkcie należało użyć metody `k-means++`.  
Nie nastąpiła żadna większa zmiana w kodzie.   
Poniżej przedstawiono fragment kodu zawierającego inicjalizację modelu `KMeans` lecz parametr `init` świadczy o tym iż będzie użyty algorytm `k-means++`

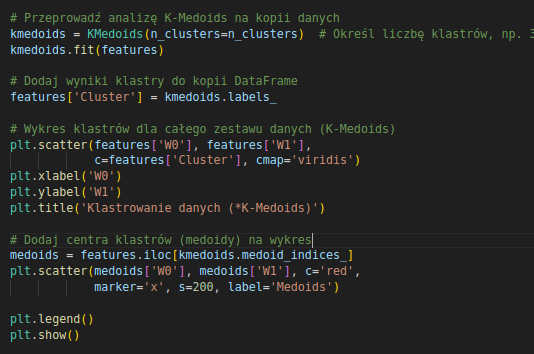


Wyniki z konsoli.

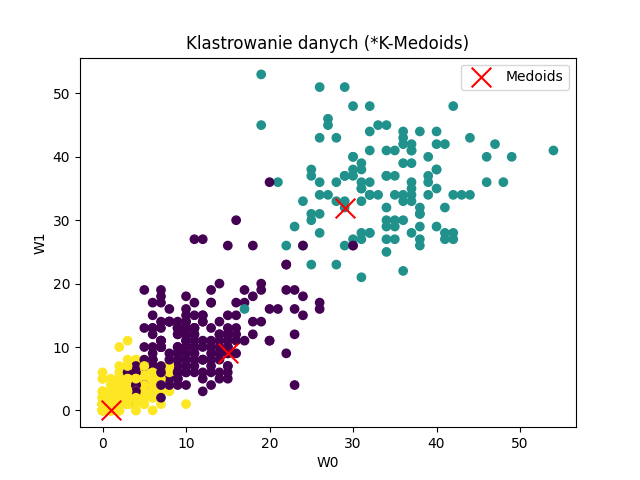


Wykres przedstawiający dane pogrupowane za pomocą metody `k-means++`



W kolejnym podpunkcie należało wykorzystać metodę KMedoids.   
Poniżej przedstawiono kod który realizuje uczenie oraz tworzenie wykres

Wykres grupowania metodą KMedoids z centroidami.

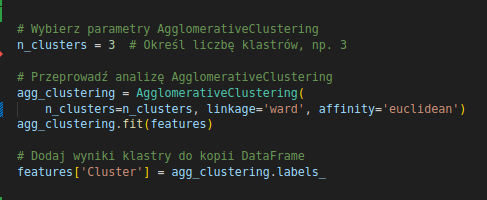


### Zadanie 3

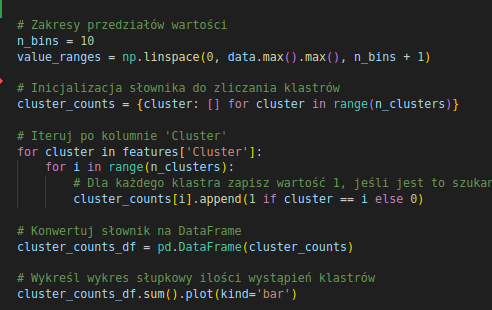
Wykonaj klasteryzację dla tego samego zbioru danych na podstawie algorytmów:

* Agglomerative Clustering z dowolnym typem klasteryzacji, typem wiązania oraz metryką odległości
* DBSCAN z domyślnymi wartościami argumentów

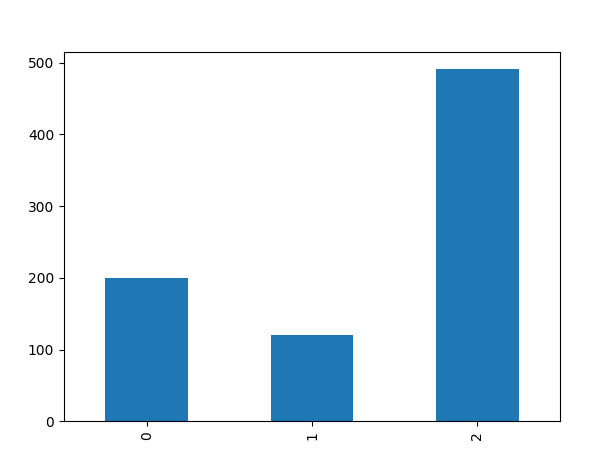
Inicjalizacja modelu AgglomerativeClustering,



Zliczanie ilości wystąpień elementów danej klasy.

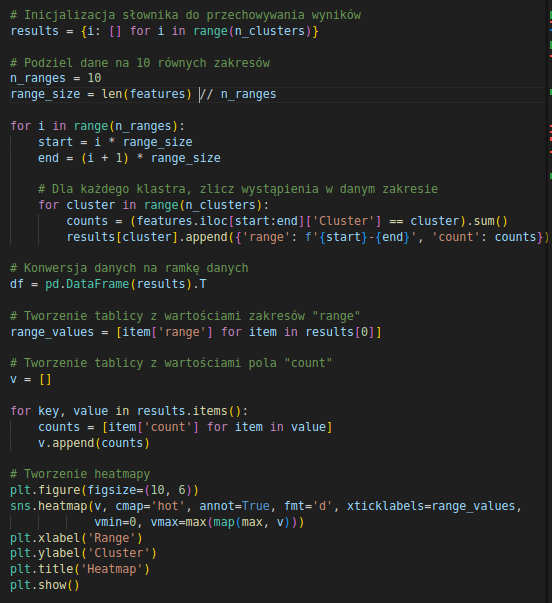


Wyświetlenie obfitości wystąpień klas.



Następnie przeszedłem do obróbki danych pod `heatmap`  
Poniżej tworze 10 zakresów liczbowych, od wartości 0 do maksymalnej wartości występującej w oryginalnej tablicy “data”.

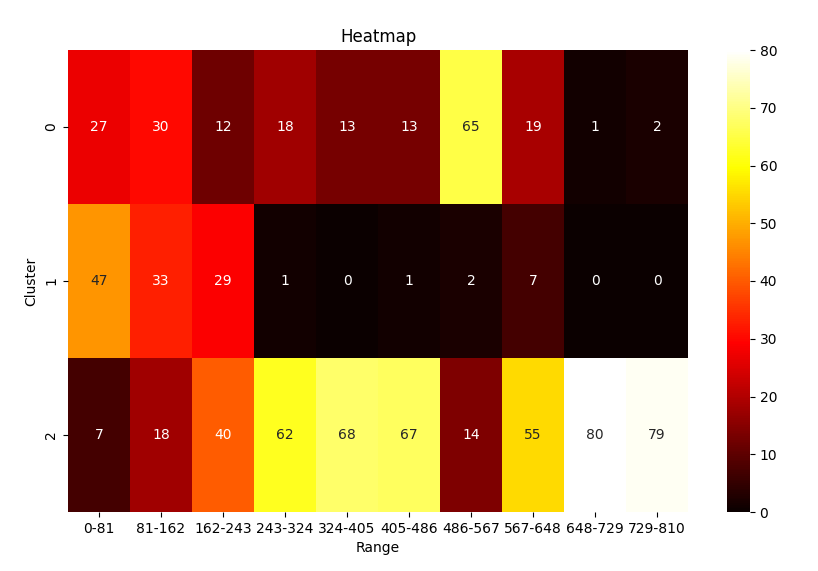
Liczymy ilość wystąpień danej wystąpień danych z danego przedziały w każdej z klas



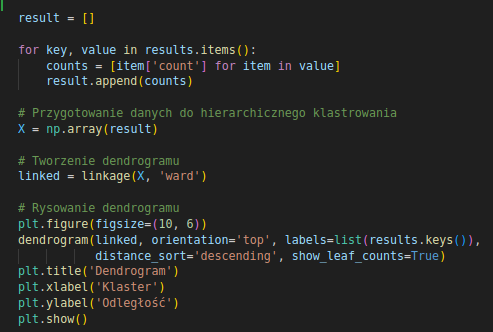
Po obróbce danych możemy wyświetlić heatmap

Interpretacja danych:

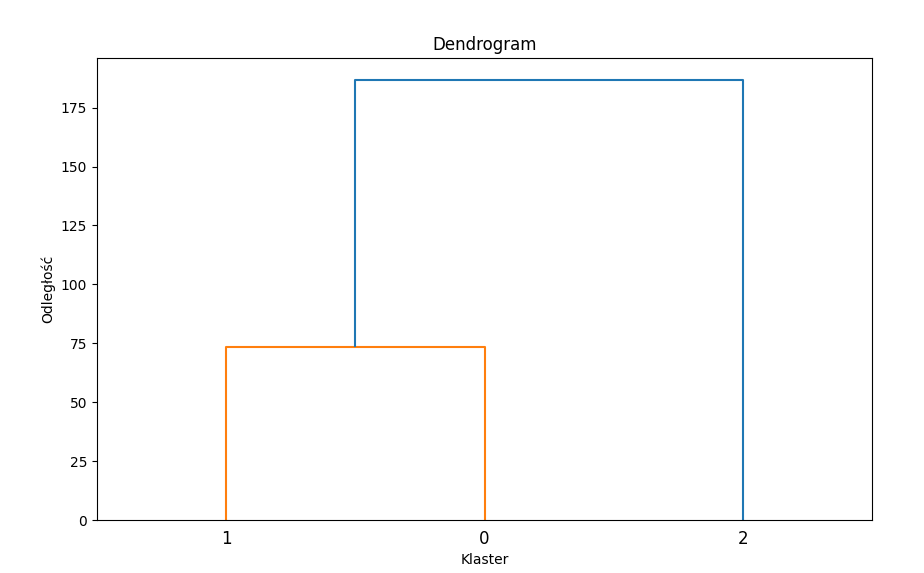
Heatmapa przedstawia 3 klastry ( oś y ).  
Oś x przedstawia zakresy indeksy danych w tablicy zwróconej poprzez predykcję modelu.  
Natomiast nałożone kolory tj. im jaśniejszy kolor tym większa ilość wystąpień.  
  
Z poniższego wykresu możemy zauważyć, że klaster 2 najwięcej występował w pod koniec datasetu, natomiast najmniej na początku.  
Zupełnie inaczej wygląda klaster 1, który to można powiedzieć, występował tylko na początkowych 250 pozycjach.

Elementy przynależące do klaster 0, były niemal równomiernie rozłożone po całym dataset, za wyjątkiem zakresów 485-565 gdzie występował pik częstotliwości występowania (65 razy - co stanowi 80% predykcji), 650-end gdzie klaster 0 praktycznie nie występował.  


W kolejnym kroku należało przygotować dane do denogramu.

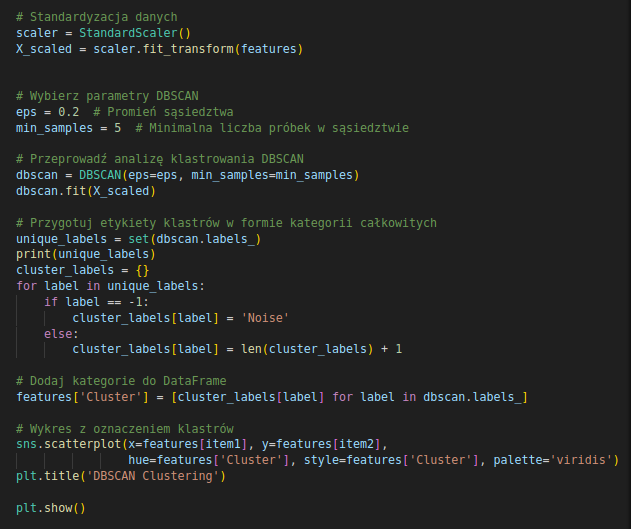


Wyświetlenie denogramu.



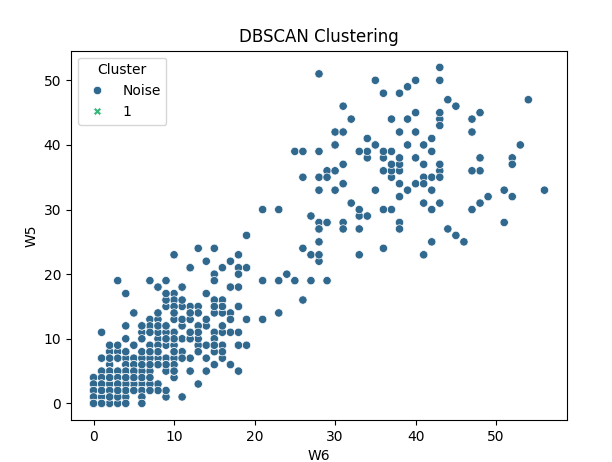
W pierwszej kolejności, po wczytaniu danych, należy ustandaryzować dane.  
W przypadku, gdy niektóre cechy mają większy wpływ na wyniki analizy niż inne, to może prowadzić to do błędnych wyników lub zniekształcenia struktury klastrów.

Standaryzacja pomaga uniknąć takich sytuacji.

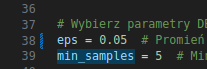


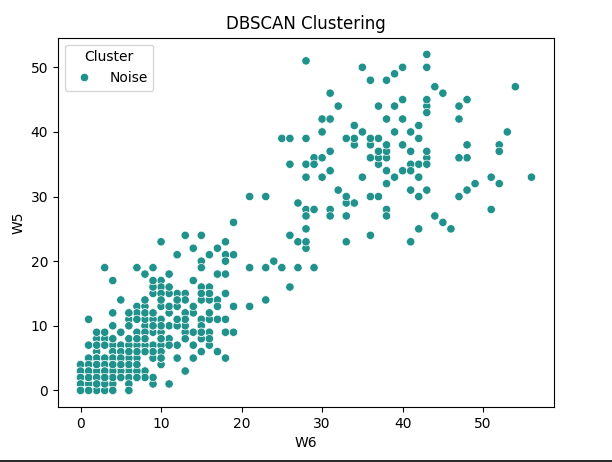
Wykres standaryzowanych danych.

Z wykresu wynika, że wszystkie punkty to “szum”.  
Zakwalifikowanych danych było



Natomiast zmiana parametrów przyniosła tylko negatywne skutki.  
Np dla wartości min\_samples = 5

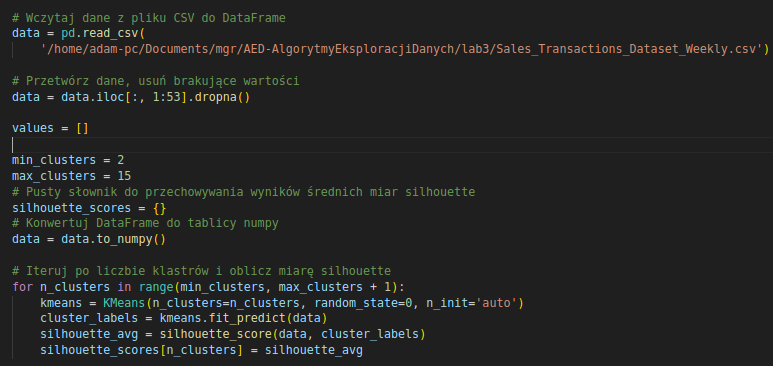


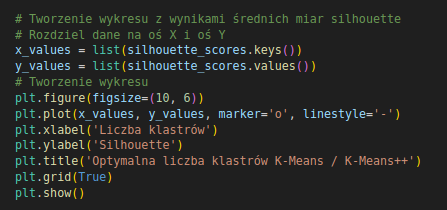


### Zadanie 4

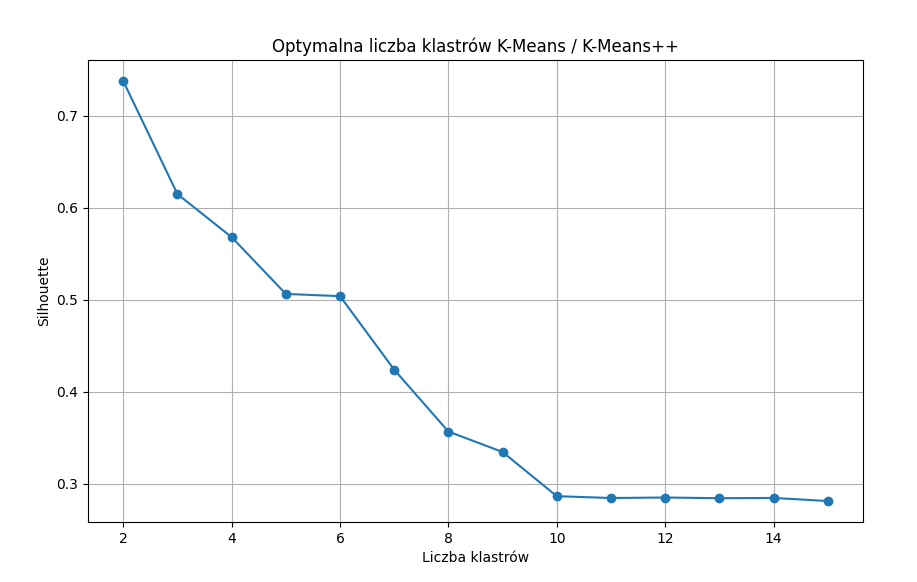
Korzystając z miary silhouette, a dokładnie średniej miary silhouette dla wszystkich klastrów oblicz i zaprezentuj graficznie optymalne liczby klastrów dla każdego z analizowanych algorytmów w zakresie od 2 do 15 klastrów.

Poniższy kod realizuje wczytanie danych z pliku, inicjalizację podstawowych zmiennych, oraz generowanie danych potrzebnych do utworzenia wykresu silhouette.

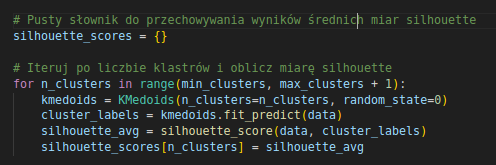


Następnie mając dane mogłem utworzyć wykres  


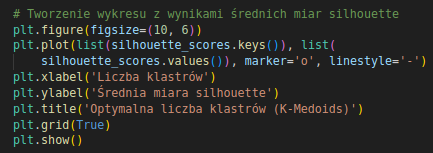
Przedstawienie wykresu który otrzymałem



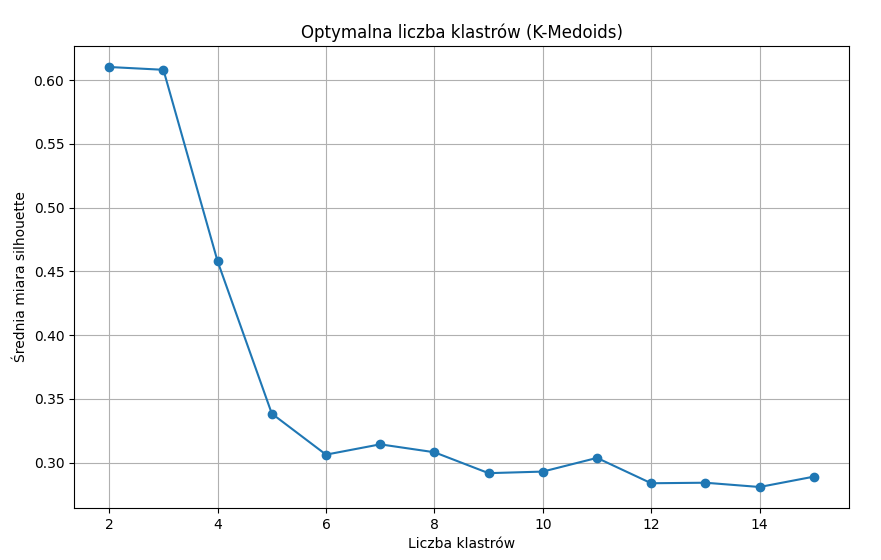
W kolejnym przeszedłem do realizacji wykresu silhouette dla metody KMedoids, co ukazuje poniższy kod:



Dane zostały zgromadzone, mogę je następnie zaprezentować na wykresie.



Wykres optymalnej ilości klas dla metody K-Medoids



## Wnioski

K-means (K-średnich) to popularny algorytm w uczeniu maszynowym, wykorzystywany do klasteryzacji danych. Jest to technika uczenia maszynowego bez nadzoru, która grupuje punkty danych w klastry na podstawie ich podobieństwa. Istnieją techniki, takie jak metoda łokcia (elbow method) do wyboru optymalnej liczby klastrów. Algorytm K-means jest prosty w implementacji, ale jego skuteczność może zależeć od charakterystyki danych i wyboru parametrów.

K-means jest stosowany w wielu dziedzinach, takich jak analiza danych, segmentacja klientów, analiza obrazów, analiza tekstu, przetwarzanie sygnałów i wiele innych.

K-means++ to ulepszona wersja algorytmu K-means, która poprawia sposób inicjalizacji początkowych pozycji centroidów klastrów. Standardowa inicjalizacja w algorytmie K-means polega na losowym wyborze początkowych centroidów, co może prowadzić do wyników zależnych od wylosowanego początkowego stanu. K-means++ wprowadza bardziej inteligentny sposób inicjalizacji, który pomaga zbiegnąć do lepszych wyników klastrowania.

K-medoid, który jest modyfikacją algorytmu K-średnich (K-means) i bazuje na koncepcji medoidów.

K-medoid to algorytm klasteryzacji danych, który jest podobny do K-średnich, ale zamiast używać centroidów klastrów, używa medoidów.

Medoid to punkt danych w klastrze, który minimalizuje sumę odległości między nim a pozostałymi punktami w klastrze. W przeciwieństwie do centroidów, medoid nie jest średnią arytmetyczną punktów w klastrze.

Algorytm K-medoid ma na celu znalezienie k medoidów, reprezentujących klastry w danych.

Agglomerative Clustering to algorytm klasteryzacji hierarchicznej wykorzystywany w uczeniu maszynowym do grupowania danych. Jest to algorytm aglomeracyjny, co oznacza, że zaczyna od pojedynczych punktów i stopniowo łączy je w większe klastry.

Algorytm DBSCAN to metoda klasteryzacji danych, która grupuje punkty na podstawie gęstości punktów w przestrzeni.

Wykorzystuje dwa parametry: promień otoczenia (eps) i minimalną liczbę punktów w otoczeniu (min\_samples), aby określić klastry i punkty odstające.

Dendrogram ukazuje hierarchiczną strukturę danych lub obserwacji w kontekście skupień. Skupienia są reprezentowane jako gałęzie drzewa, a odległość między nimi jest interpretowana jako miara podobieństwa lub odmienności

Wykres Silhouette pomaga on w określeniu optymalnej liczby klastrów poprzez mierzenie odległości między punktami wewnątrz klastrów i między klastrami. Wyższe wartości indeksu Silhouette wskazują na lepszą jakość klastrów.