

KLASTERYZACJA - LAPTOPY

Katsiaryna Bokhan, Martyna Leśniak,
Aleksandra Wójcik

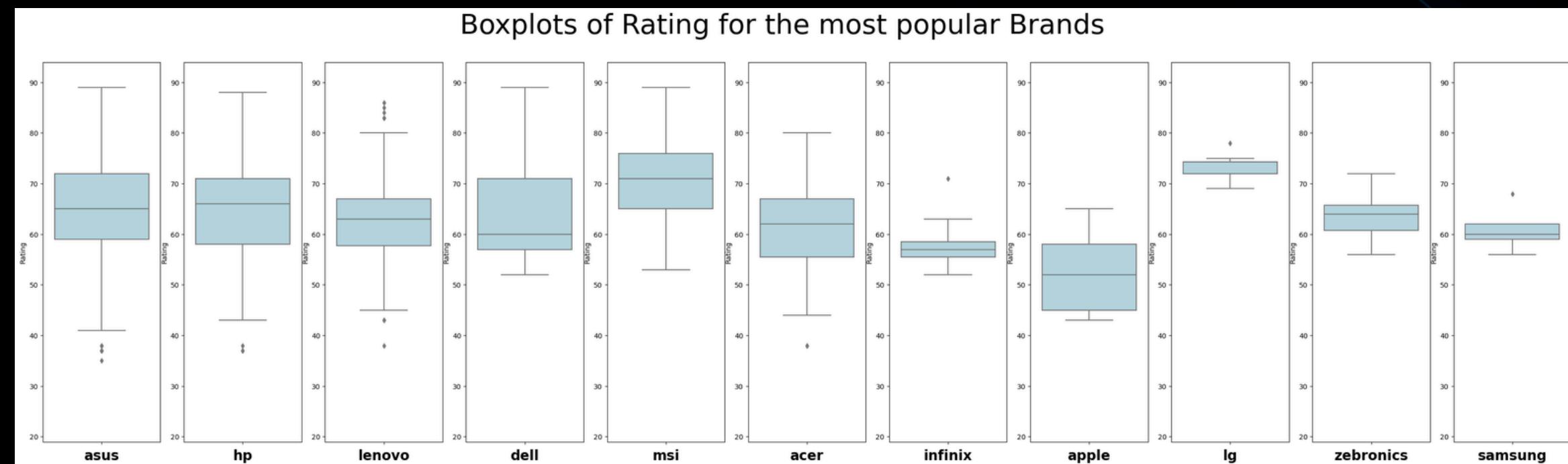
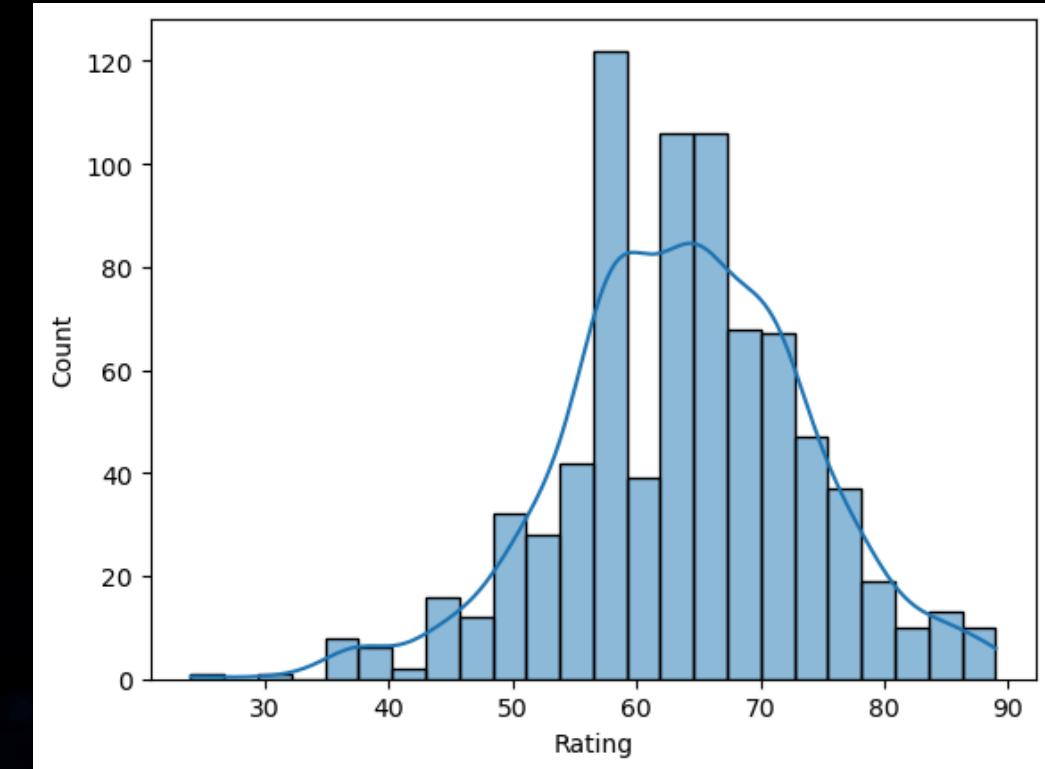
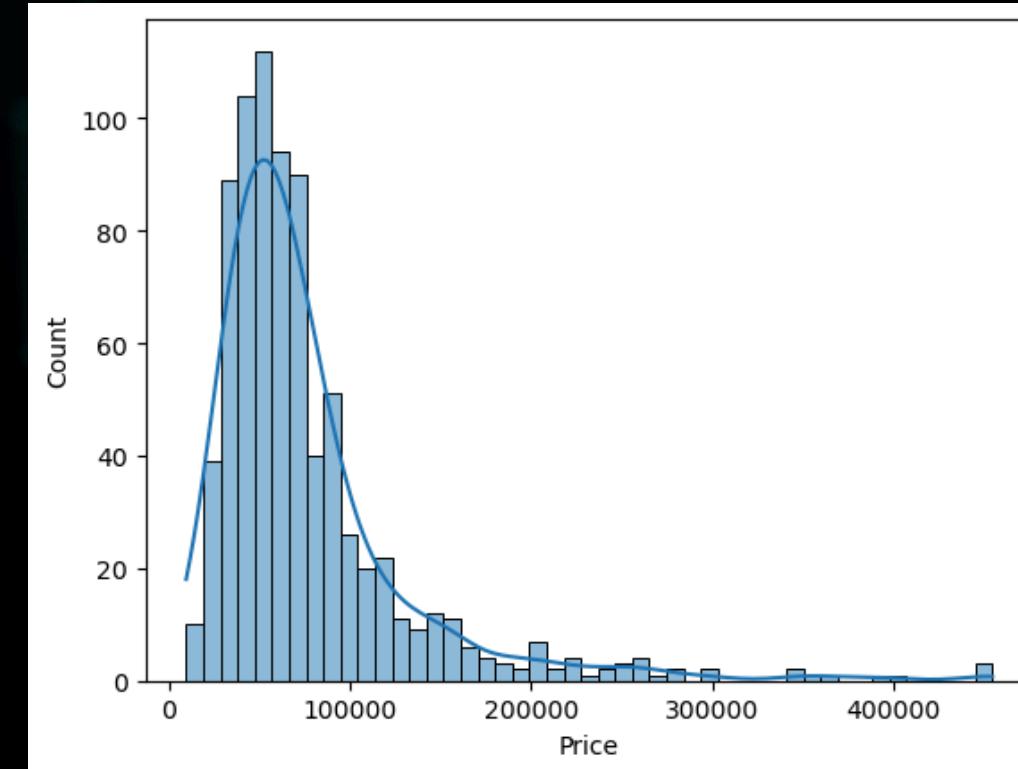
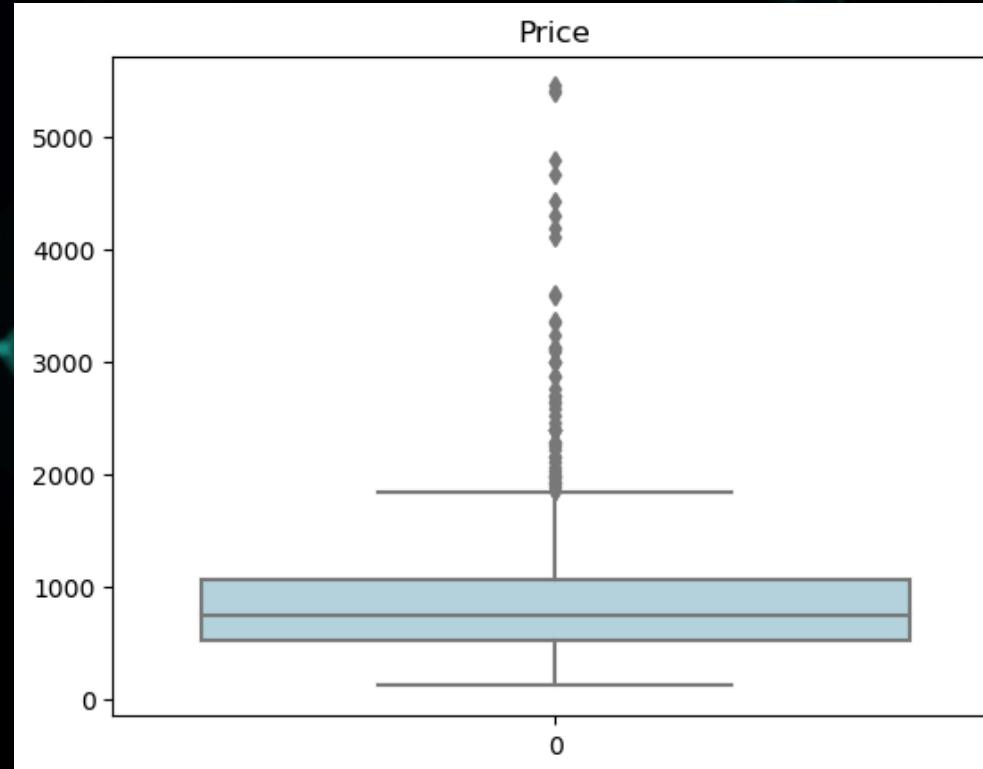
DANE

- Dane pochodzą z platformy kaggle:
<https://www.kaggle.com/datasets/bhavikjikadara/brand-laptops-dataset/>
- 22 kolumny, różne typy danych
- 20% danych dla zespołu walidacyjnego

Data columns (total 22 columns):			
#	Column	Non-Null Count	Dtype
0	index	991 non-null	int64
1	brand	991 non-null	object
2	Model	991 non-null	object
3	Price	991 non-null	int64
4	Rating	991 non-null	int64
5	processor_brand	991 non-null	object
6	processor_tier	991 non-null	object
7	num_cores	991 non-null	int64
8	num_threads	991 non-null	int64
9	ram_memory	991 non-null	int64
10	primary_storage_type	991 non-null	object
11	primary_storage_capacity	991 non-null	int64
12	secondary_storage_type	991 non-null	object
13	secondary_storage_capacity	991 non-null	int64
14	gpu_brand	991 non-null	object
15	gpu_type	991 non-null	object
16	is_touch_screen	991 non-null	bool
17	display_size	991 non-null	float64
18	resolution_width	991 non-null	int64
19	resolution_height	991 non-null	int64
20	OS	991 non-null	object
21	year_of_warranty	991 non-null	object

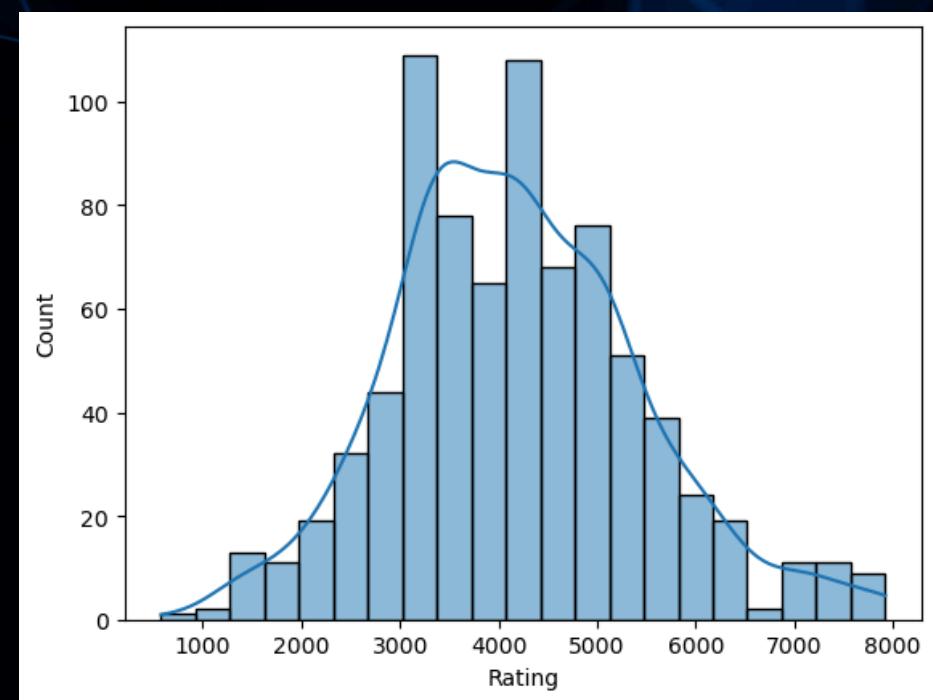
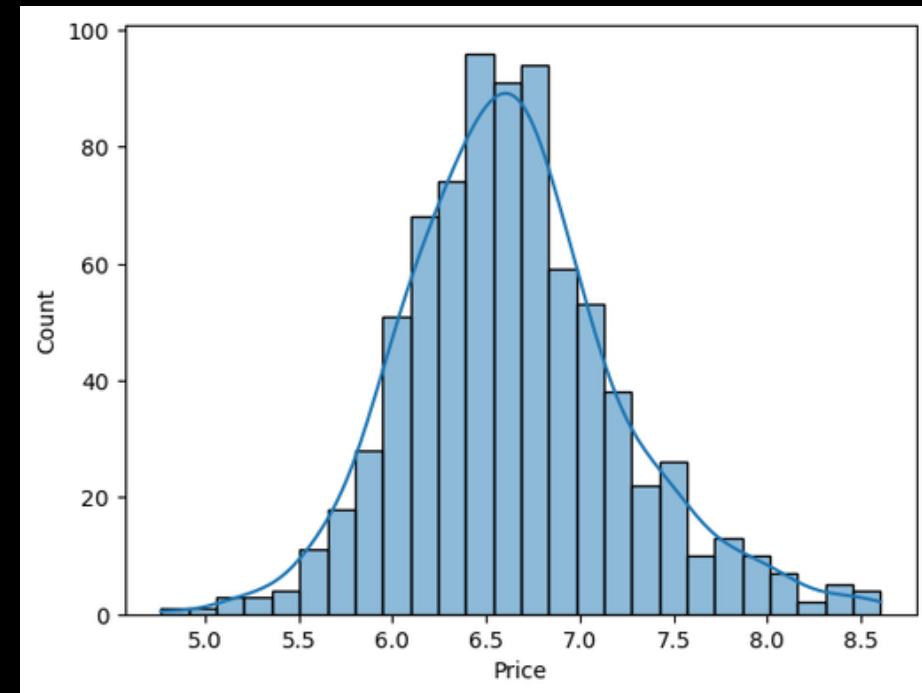
dtypes: bool(1), float64(1), int64(10), object(10)

EDA

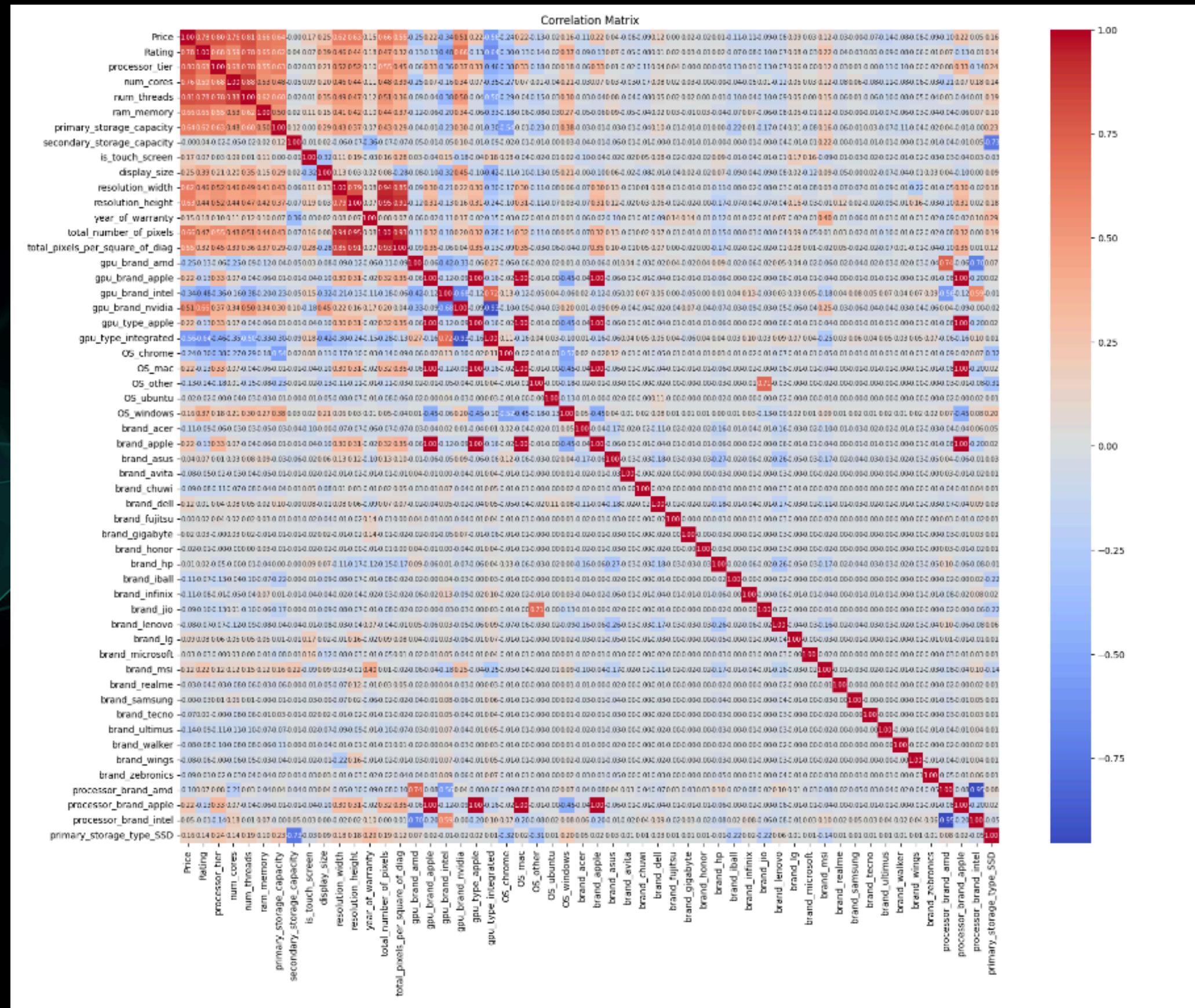


TRANSFORMACJA ZMIENNYCH

- Zmienne numeryczne: do rozkładu normalnego, testy Shapiro-Wilka, Q-Q, przeskalowanie do przedziału [0,1] za pomocą MinMaxScaler.
- Zmienne kategoryczne: OneHotEncoding, OrdinalEncoding.



Macierz korelacji po transformacji zmiennych

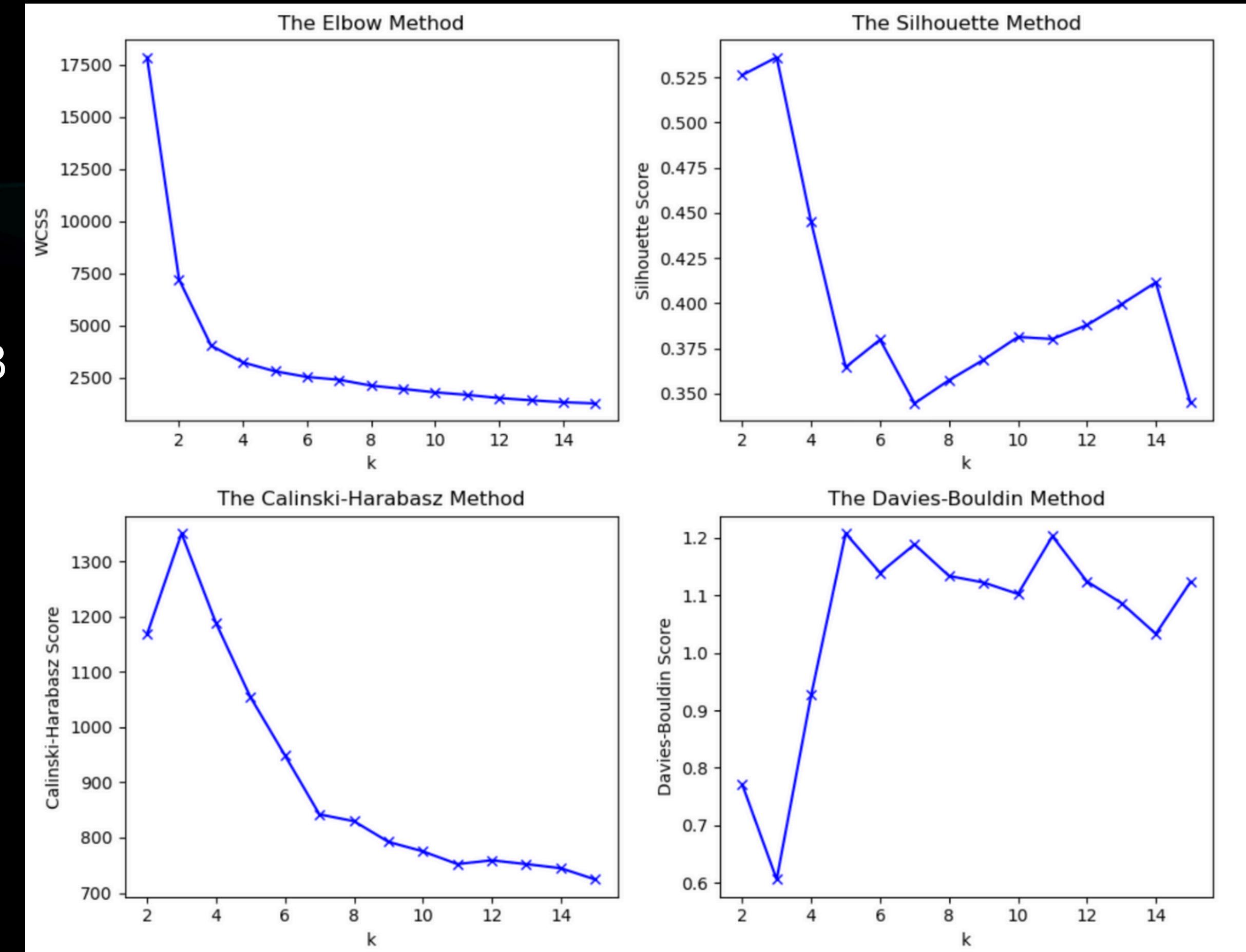


UMAP OGÓLNY POMYSŁ

1. Zmniejszyć wymiar danych korzystając z UMAP (Optuna) -> 10 wymiarów
2. Dla nowych danych znaleźć optymalną liczbę klastrów.
3. Zastosować różne metody klasteryzacji (KMeans, Agglomerative Clustering, DBSCAN, Gaussian Mixtures, Spectral Clustering)
4. Zmniejszyć wymiar do 2/3 z użyciem PCA, t-SNE, Multidimensional scaling w celu wizualizacji

METRYKI (KMeans)

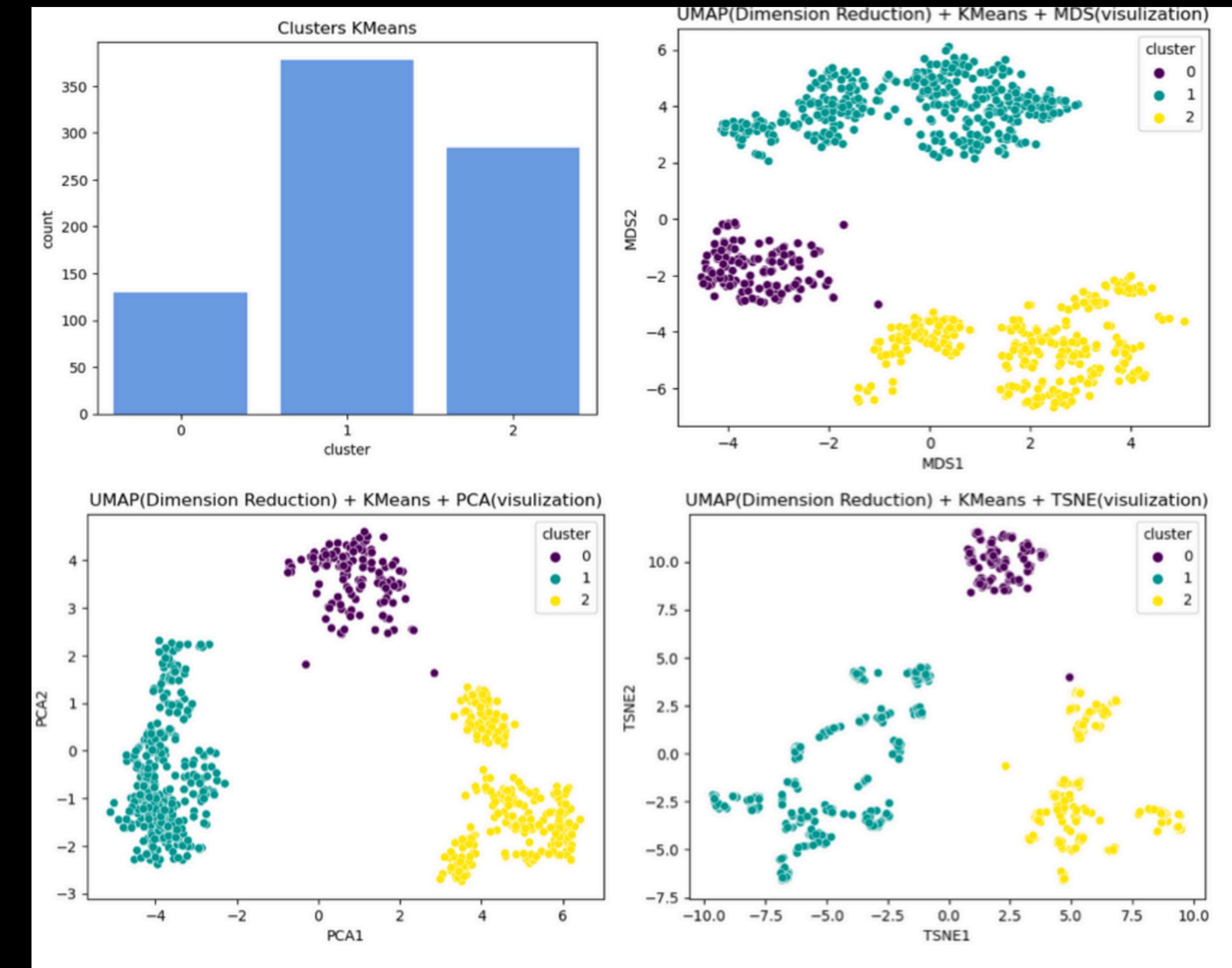
najlepsza liczba klastrów - 3



UMAP AND KMEANS (GMM, Spectral Clustering, Agglomerative Clustering)

Table 3: Metryki Kmeans

Metryka	Wartość
Silhouette Score	0.5360302
Calinski-Harabasz Score	1350.5186805091505
Davies-Bouldin Score	0.6061432453306356

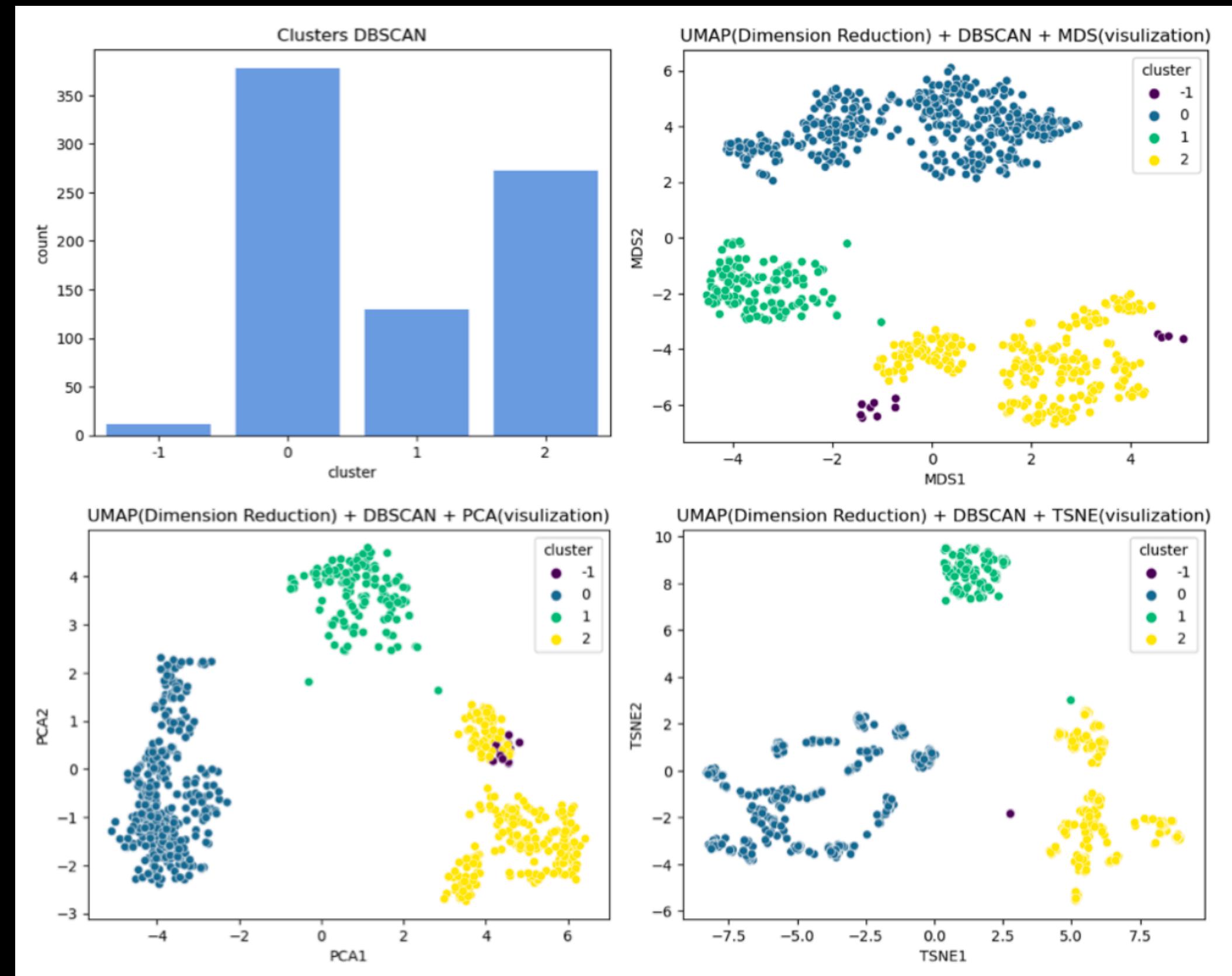


UMAP AND DBSCAN

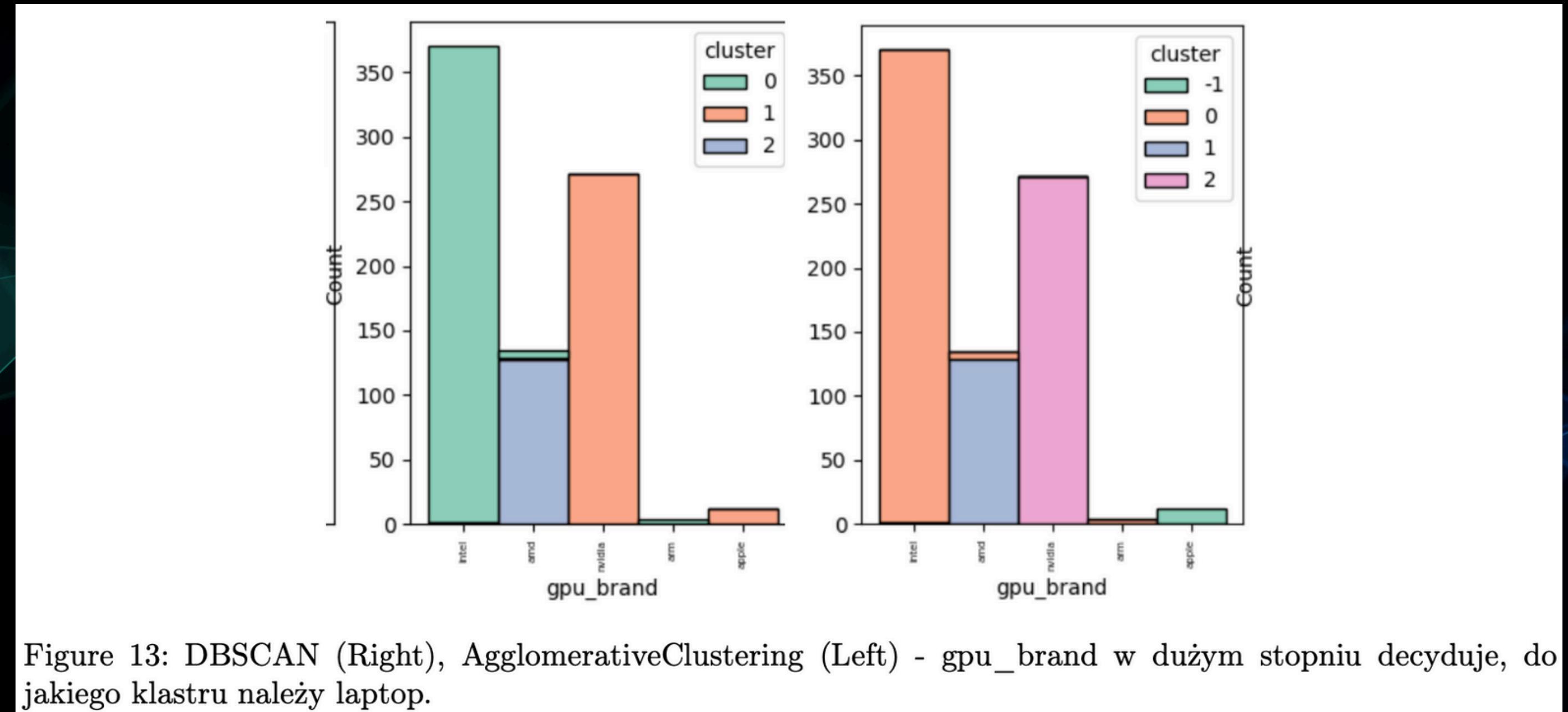
Table 5: DBSCAN

Metryka	Wartość
Silhouette Score	0.49322468
Calinski-Harabasz Score	960.0203006726656
Davies-Bouldin Score	0.6150387611468902

Wygląda dość fajnie?
Jednak jest problem...



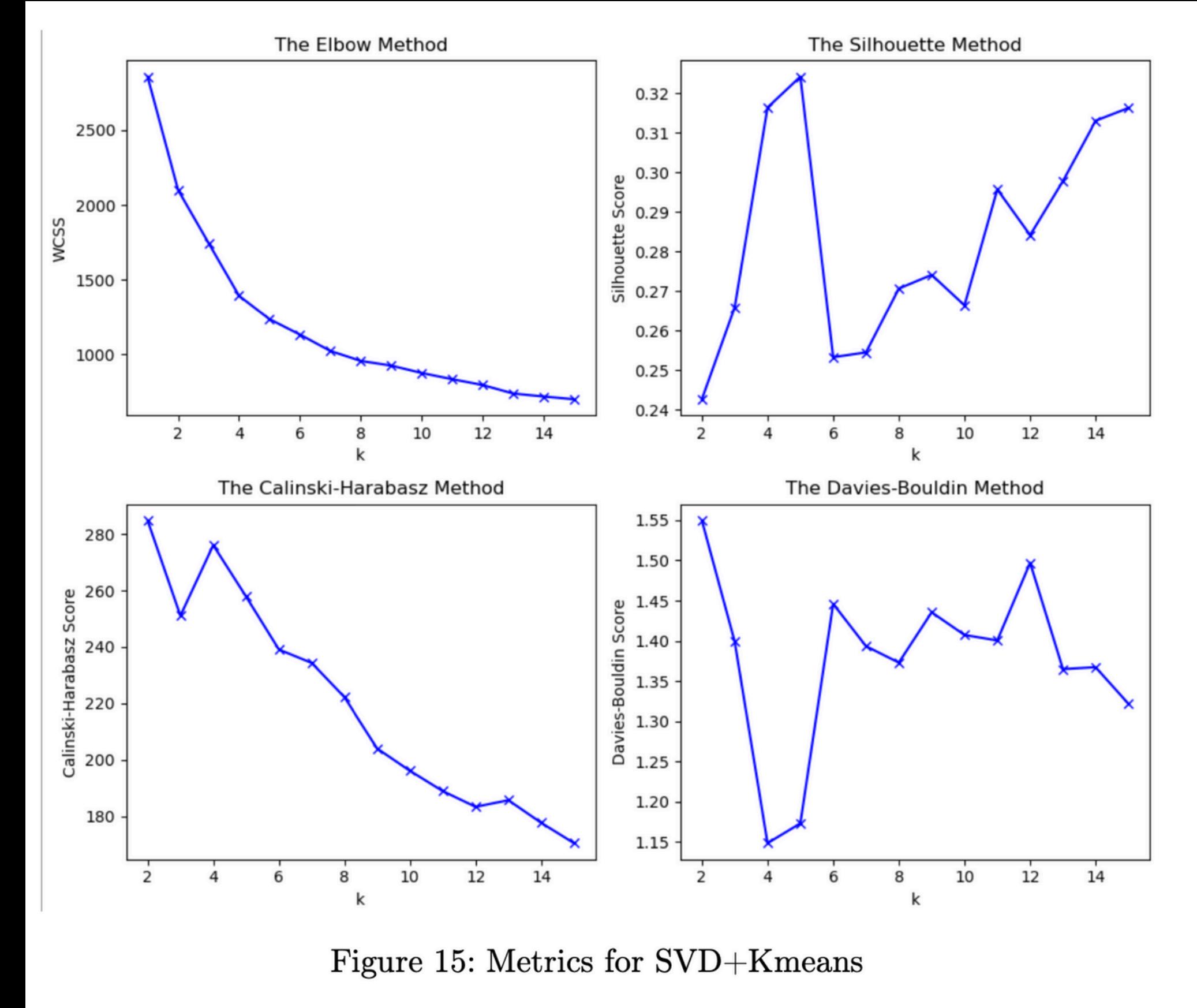
GŁÓWNY PROBLEM



Usunąć gpu_brand?

SVD AND KMEANS

- zdecydowałyśmy zachować co najmniej 90 % informacji - 11 wymiarów
- użycie innych metod klasteryzacji oraz 4 klastrów - znowu ten sam problem



SVD AND KMEANS

- duży brak równowagi w klastrach
- ale problem jak wcześniej już nie występuje

Table 8: SVD+KMeans (k=5) metrics

Metryka	Wartość
Silhouette Score	0.3242104496864801
Calinski-Harabasz Score	257.84248448945567
Davies-Bouldin Score	1.1731547842773122

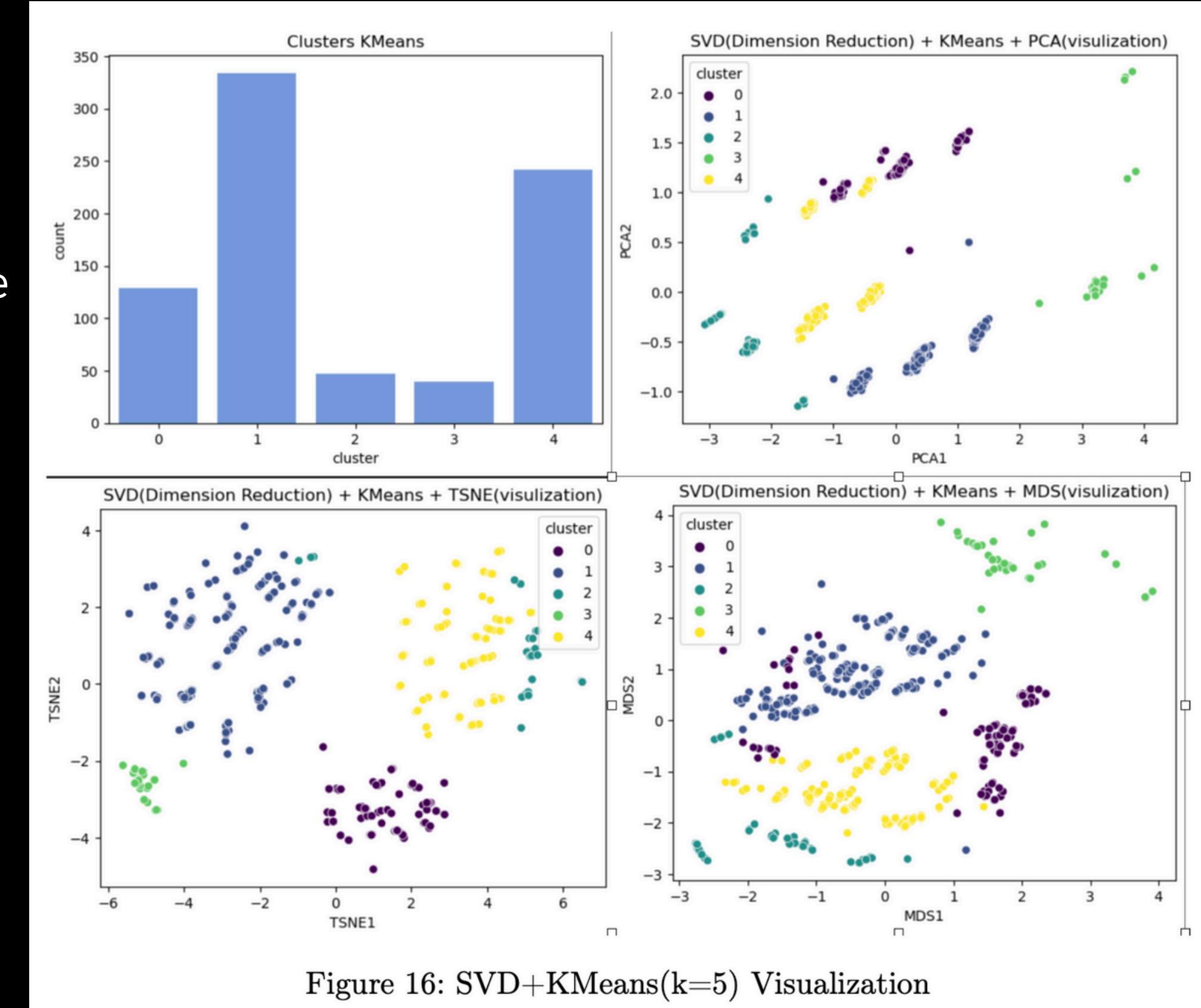


Figure 16: SVD+KMeans(k=5) Visualization

PCA AND KMEANS

- PCA do redukcji danych - najlepsze metryki dla 8 wymiarów
- Najlepsza liczba klastrów - 4

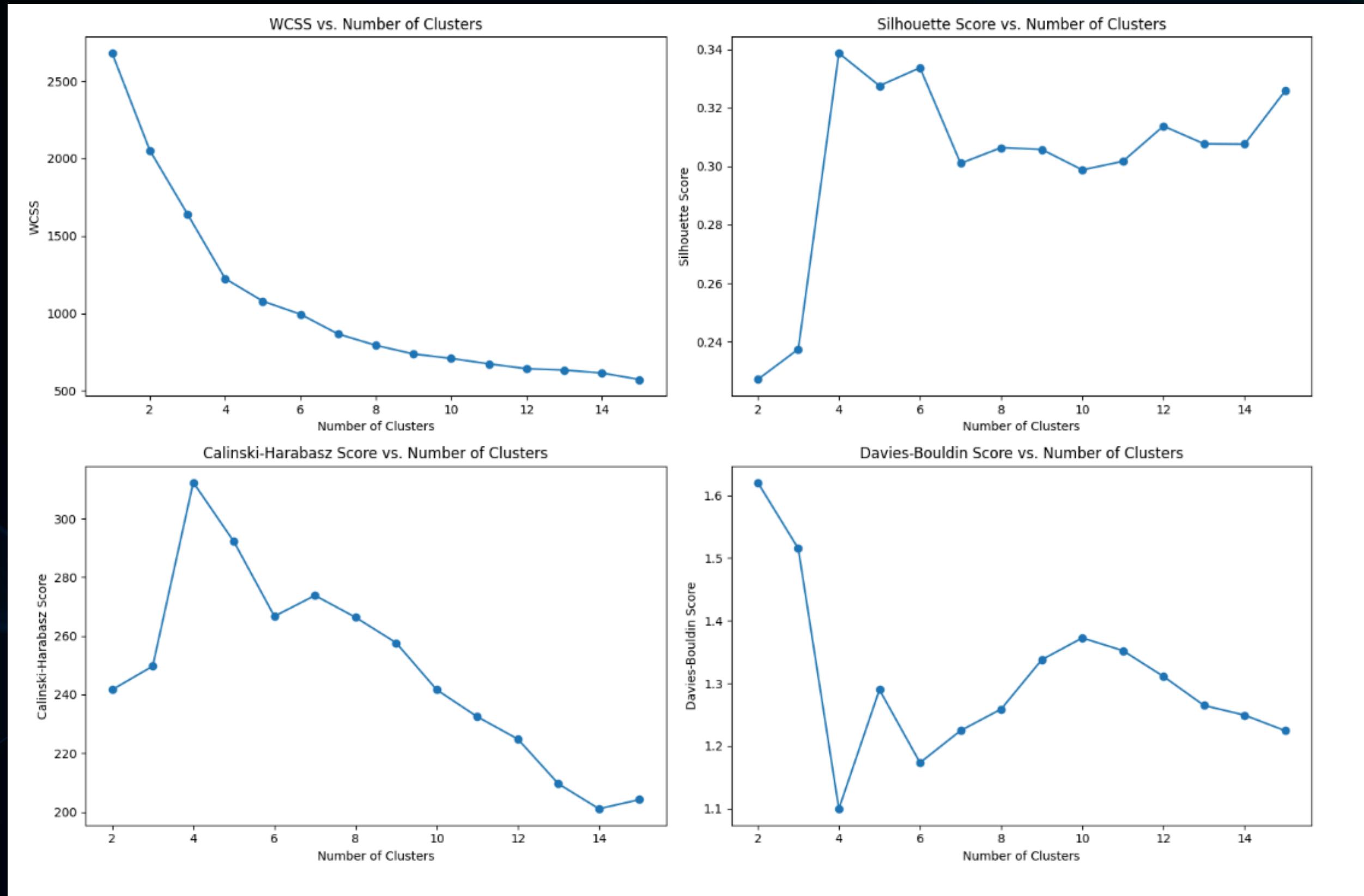
Cluster_k4

Silhouette Score: 0.3362499888715211

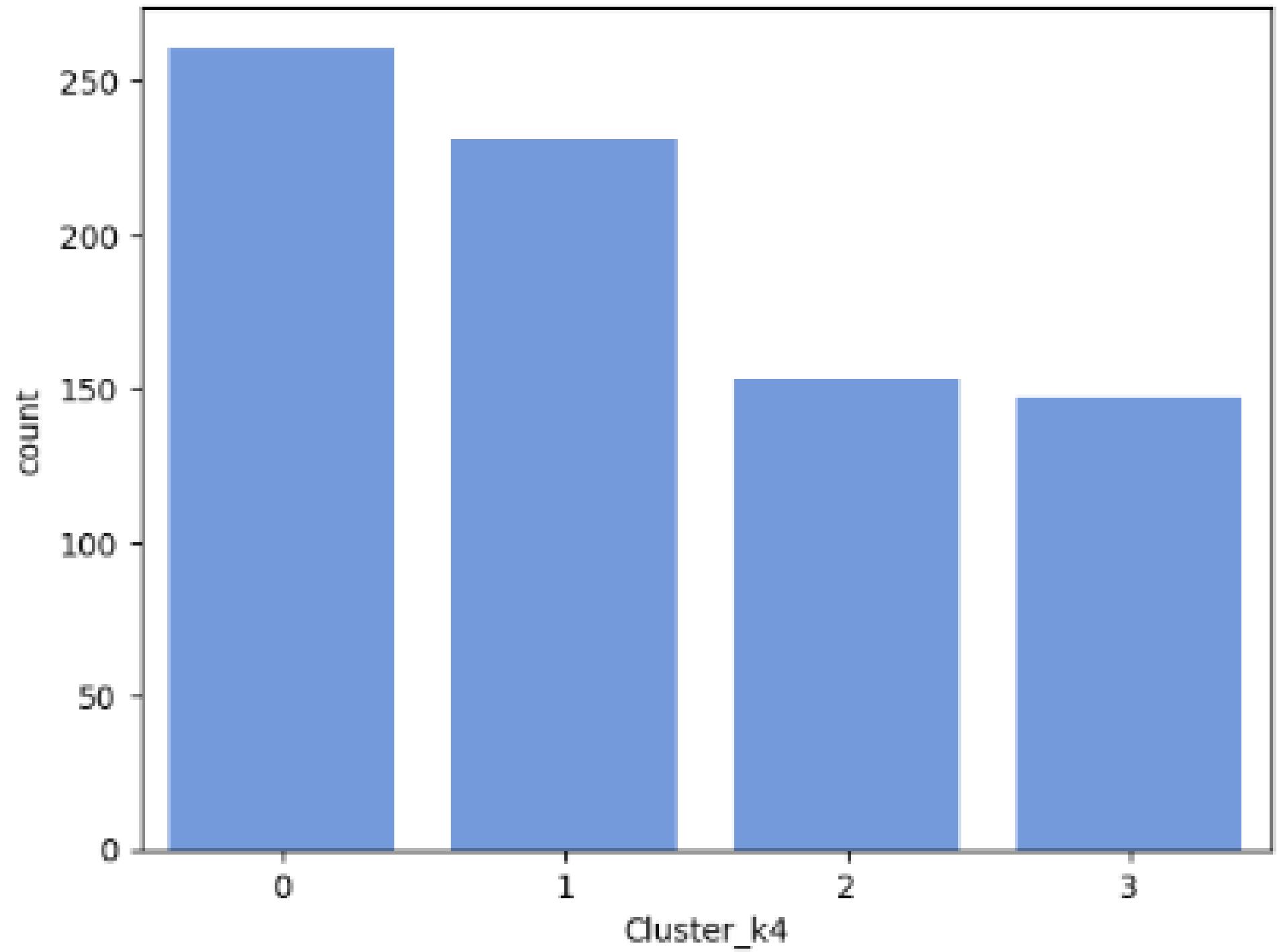
Calinski-Harabasz Score: 372.5242116666582

Davies-Bouldin Score: 1.2051595975573441

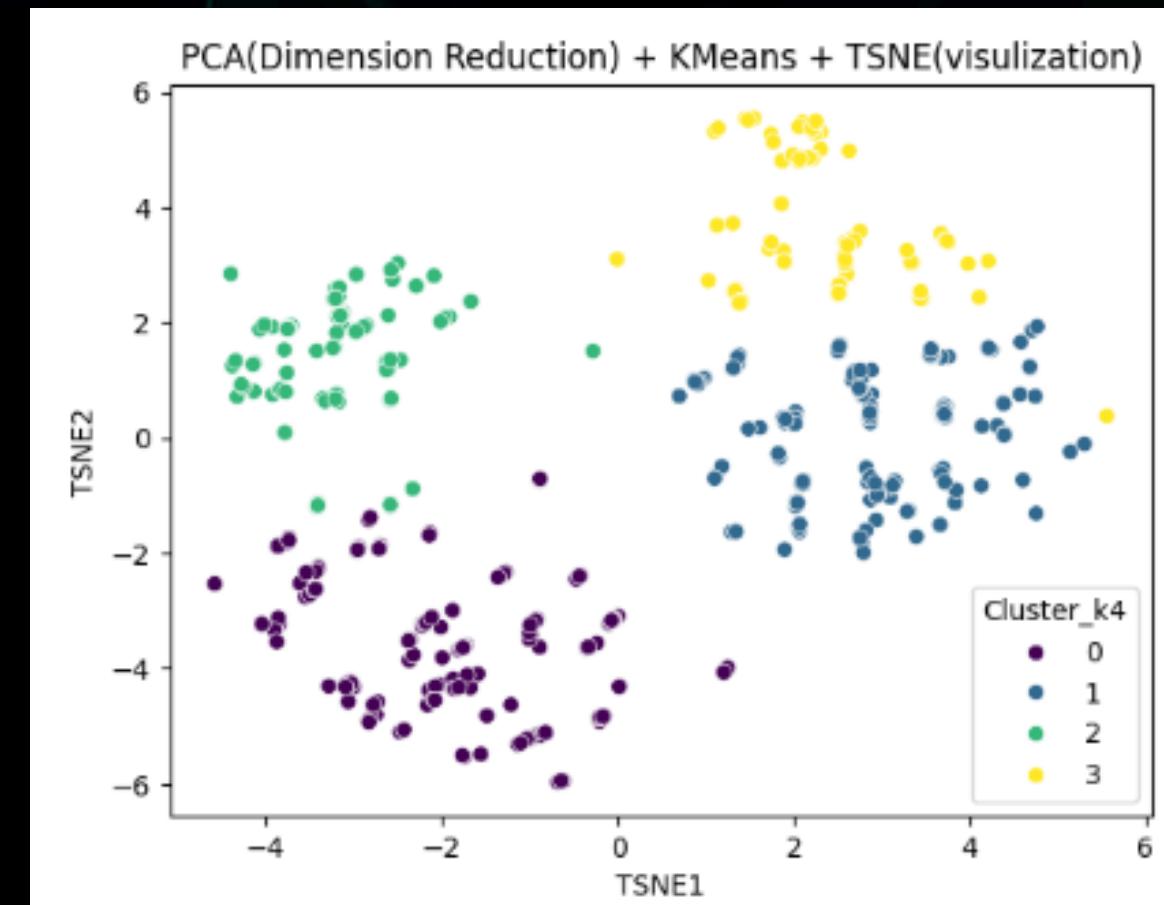
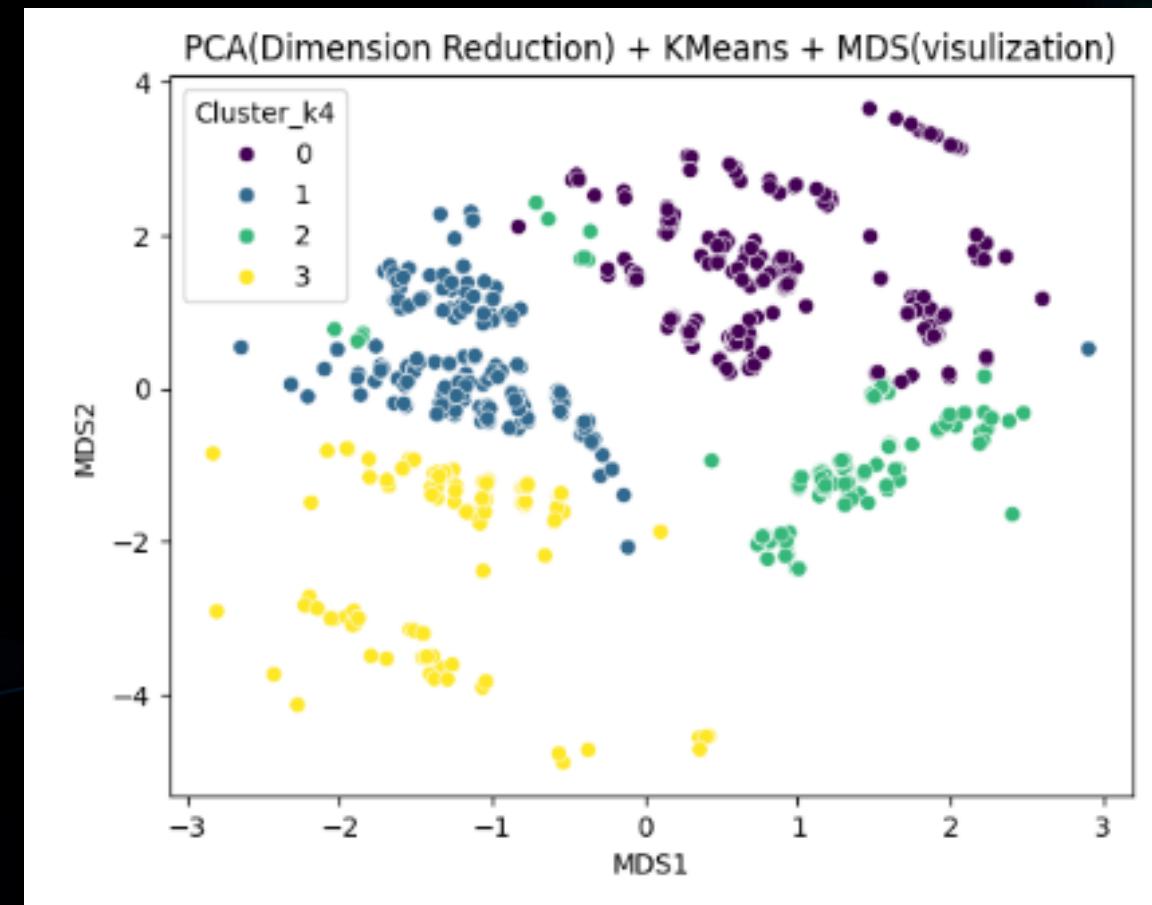
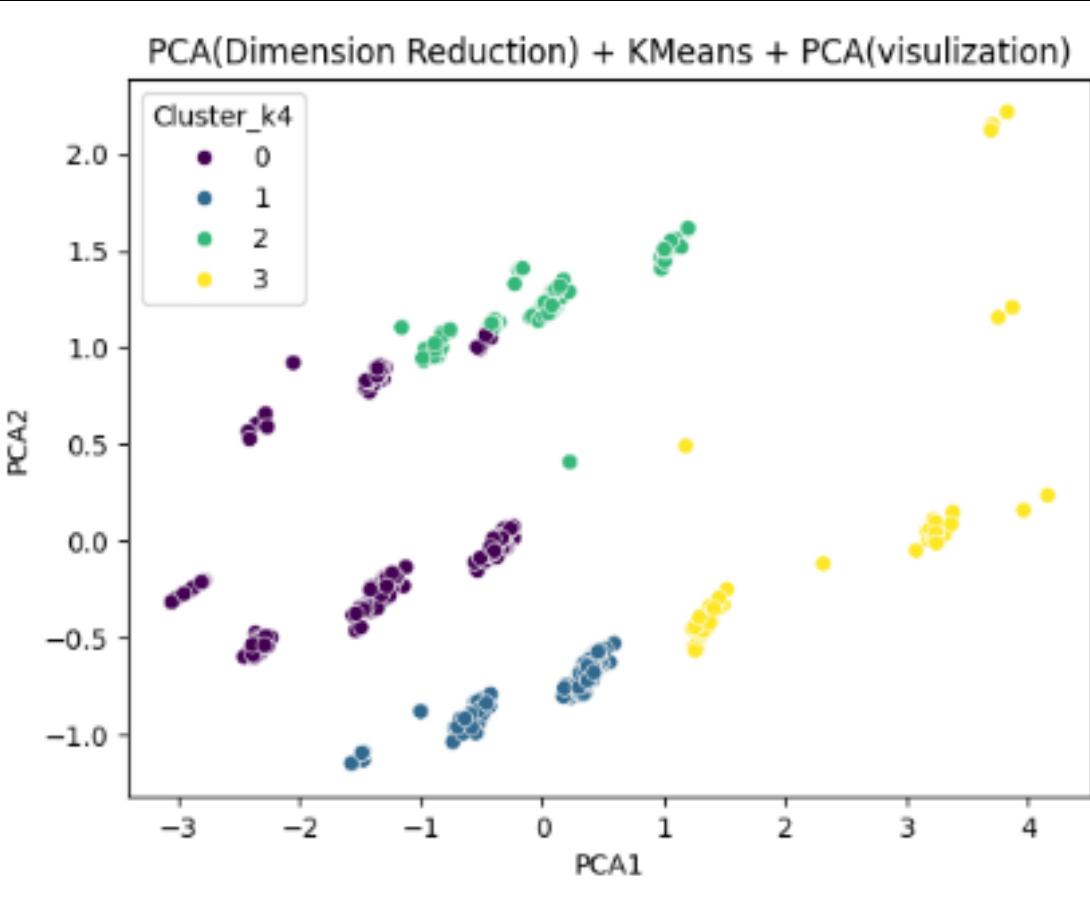
METRYKI



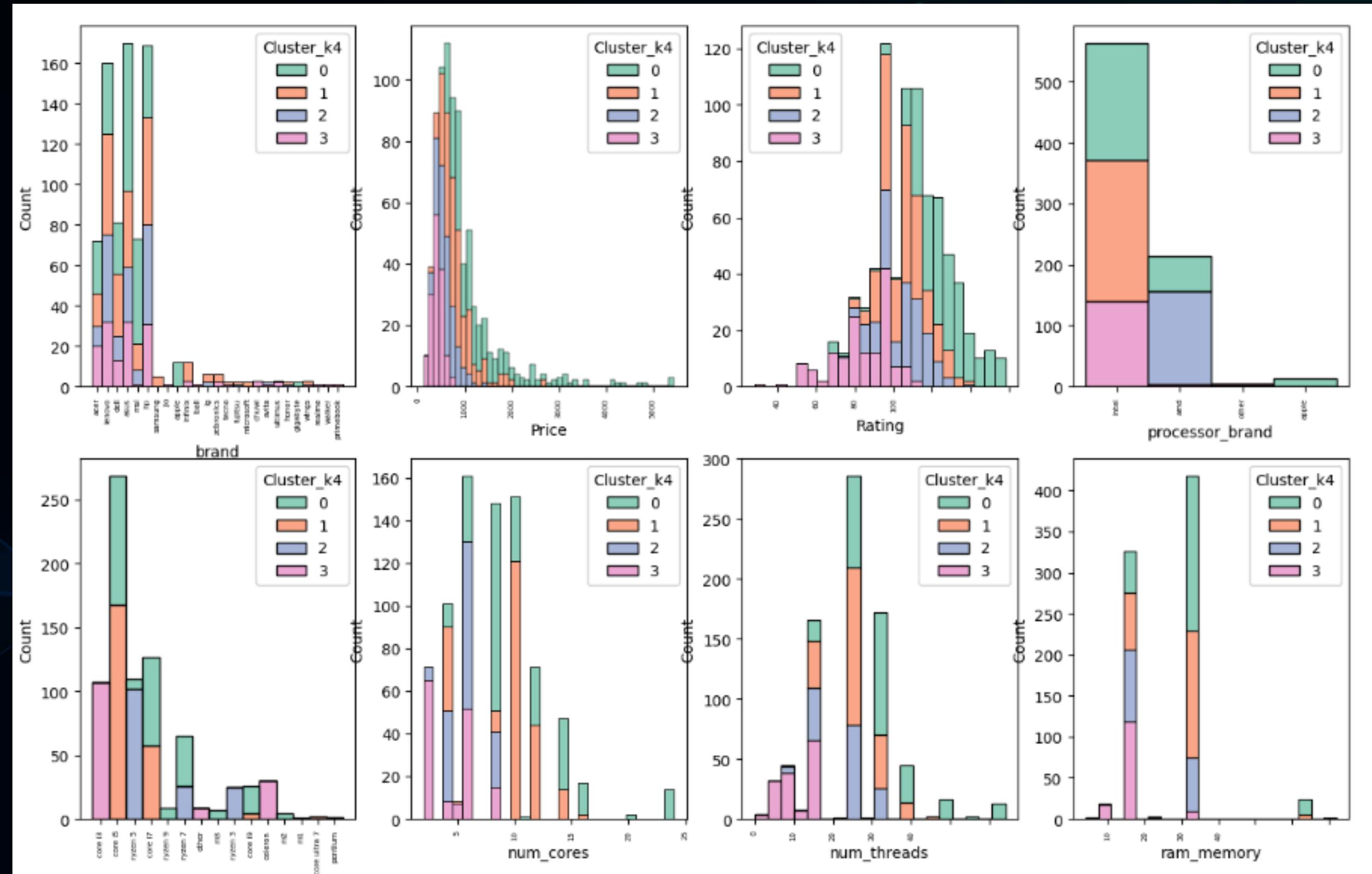
Clusters KMeans



WIZUALIZACJE



ANALIZA KLASTRÓW



ANALIZA KLASTRÓW

- **Klaster 0:** Laptopy od topowych marek, takich jak Apple, Lenovo, Asus. Laptopy z wysoką ceną i oceną. Laptopy z kartą graficzną Nvidia i najwyższą rozdzielczością ekranu.
- **Klaster 1:** Nieco tańsze i niżej oceniane laptopy. Głównie z procesorem Intel. Większość laptopów z ekranem dotykowym jest w tej grupie.
- **Klaster 2:** Wiele laptopów HP. Procesor i karta graficzna to AMD. Cena jest niższa niż 1000 USD.
- **Klaster 3:** Najtańsze i najniżej oceniane laptopy. Procesory to głównie Intel.

ANALIZA BIZNESOWA

- Segmentacja rynku
- Targetowanie kampanii marketingowych
- Dopasowanie oferty cenowej
- Planowanie asortymentu w sklepach