



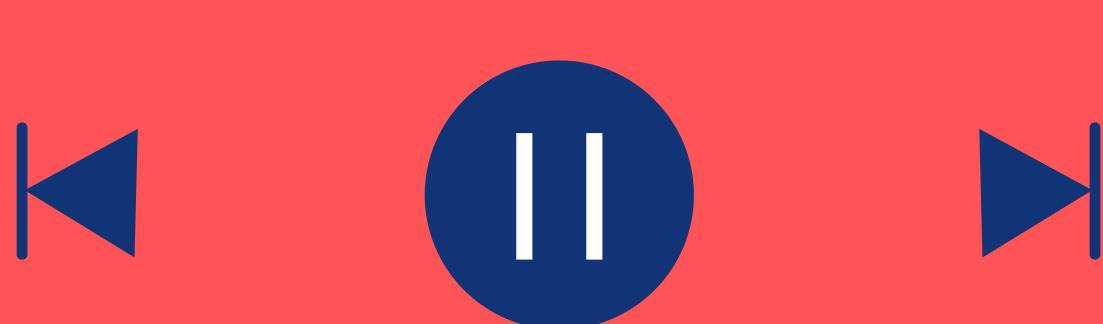
Guess The Genre

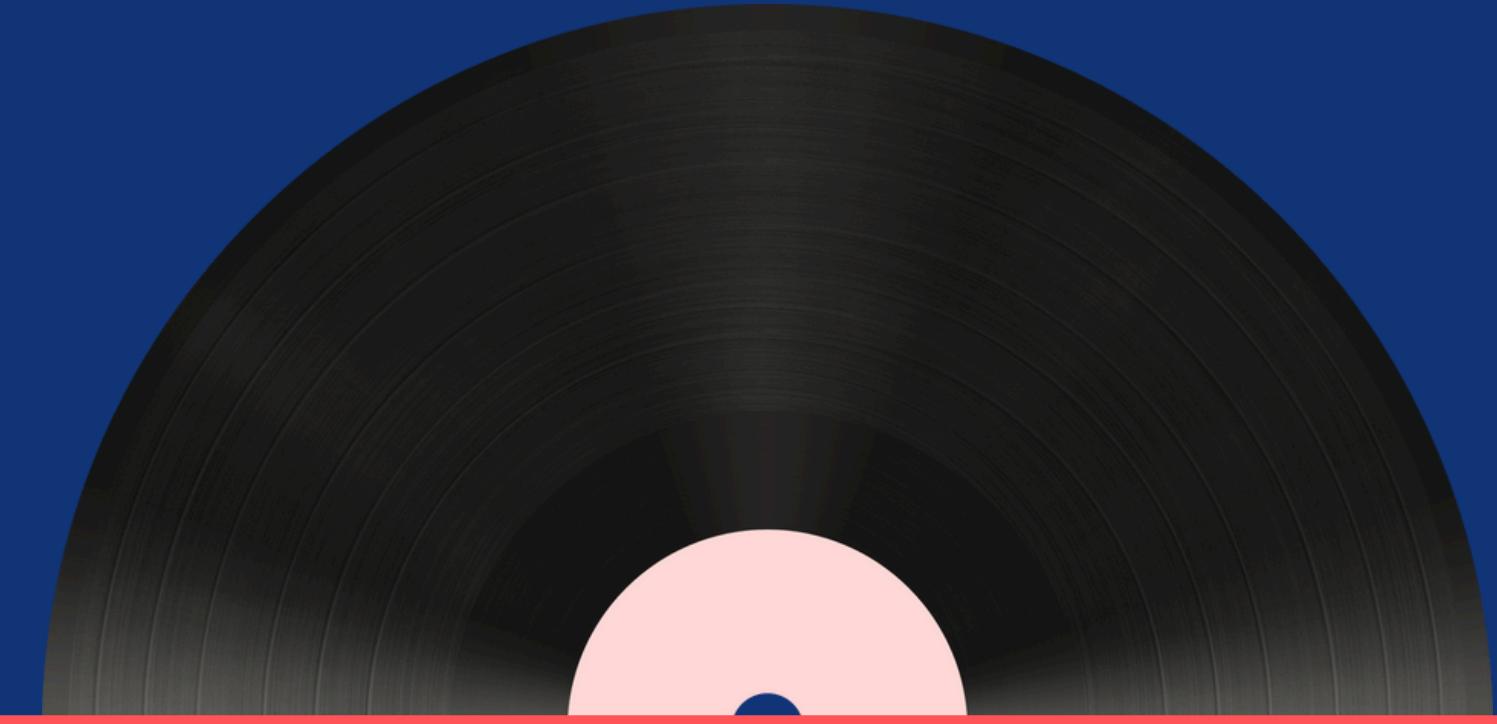
SPOTIFY DATA
CLUSTERING

ZADANIE: klasteryzacja utworów muzycznych na podstawie ich cech charakterystycznych, które opisują różne aspekty każdego utworu.

CEL: zidentyfikowanie grup utworów, które mają podobne właściwości muzyczne, a następnie przeprowadzenie na nich analizy tw celu znalezienia podobieństw i różnic między nimi.

Zamysł biznesowy





Dane

Dane dotyczą najpopularniejszych piosenek streamowanych na platformie Spotify w 2023 roku.

Zmienne można podzielić na 3 kategorie:

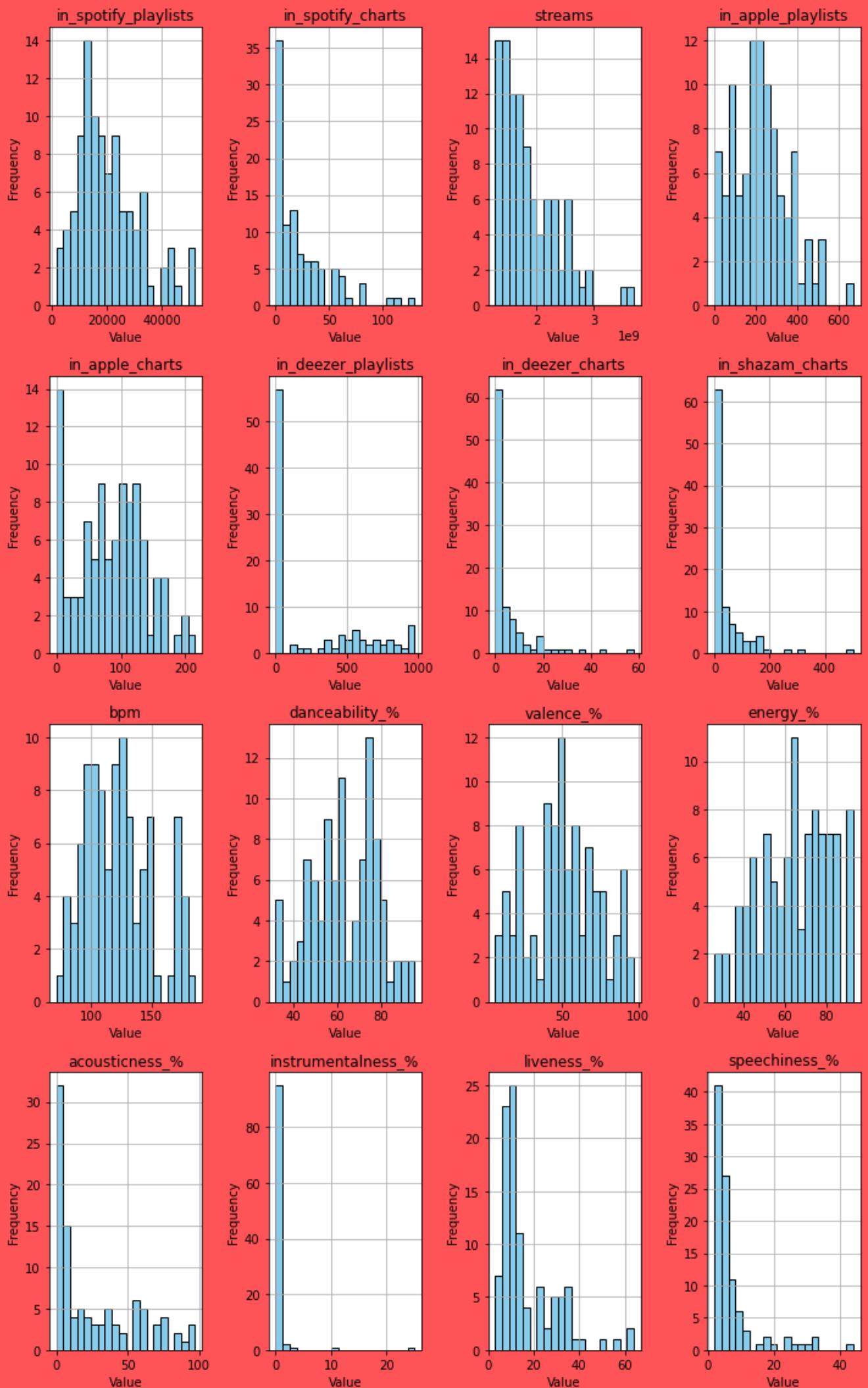
- twardé dane dotyczące twórców i daty wydania
- dotyczące popularności piosenki i obecności w różnego rodzaju rankingach
- opis muzycznych parametrów piosenki (gama, ton, muzyczność, energetyczność itp.)

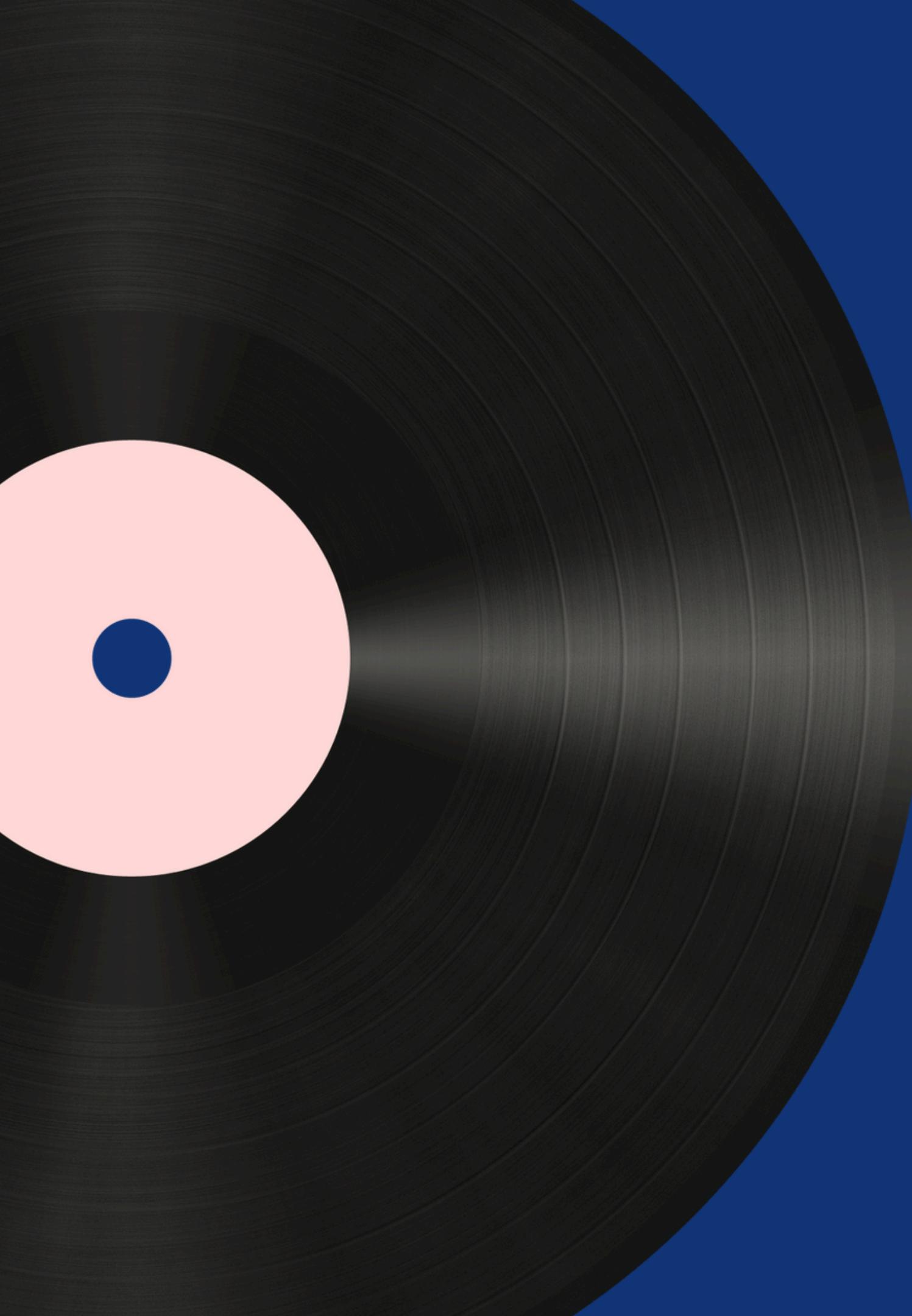
- przekonwertowanie typów danych na stringi i typy numeryczne
- zastąpienie braków danych
- usunięcie duplikatów
- transformacja zmiennych
- skalowanie kolumn numerycznych



podział danych na testowe i walidacyjne

Preprocessing danych





Klasteryzacja



Wybór danych

Celem naszej klasteryzacji była identyfikacja piosenek ze względu na ich cechy. Dlatego zmienne na których skupiłyśmy się to:

- bpm(beats per minute)
- danceability (taneczność)
- valence (pozytywność)
- acousticness
- instrumentalness
- liveness (czy piosenka ma elementy “na żywo”)
- speechiness (obecność elementów mówionych)



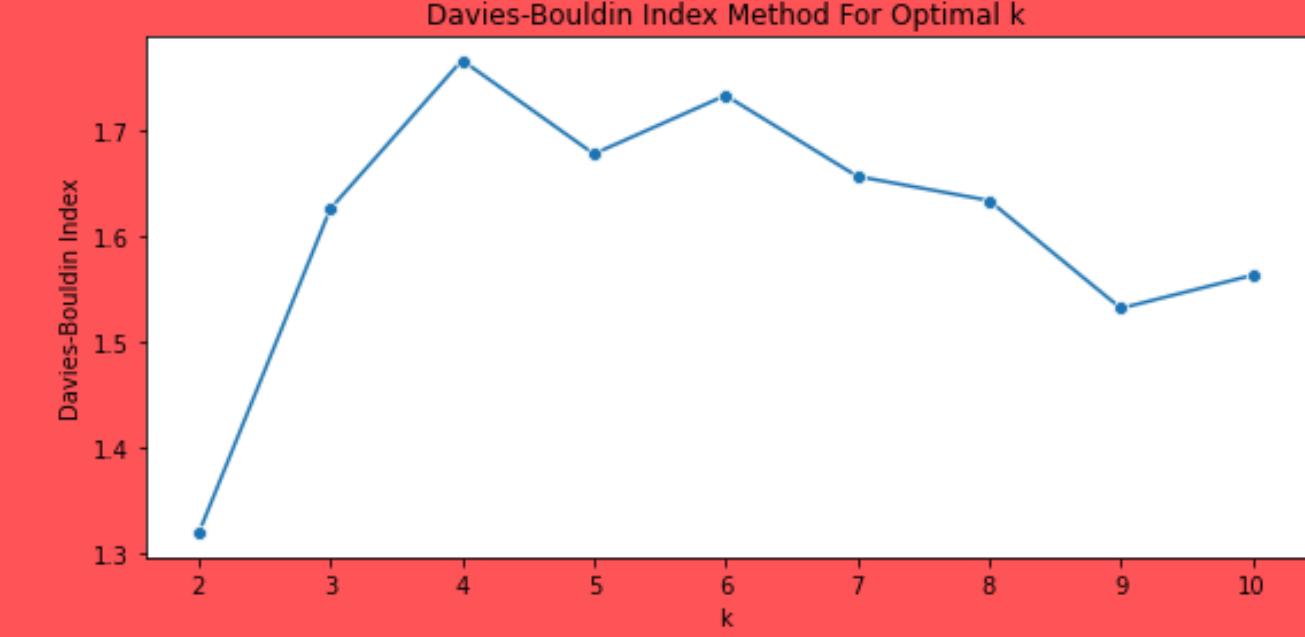
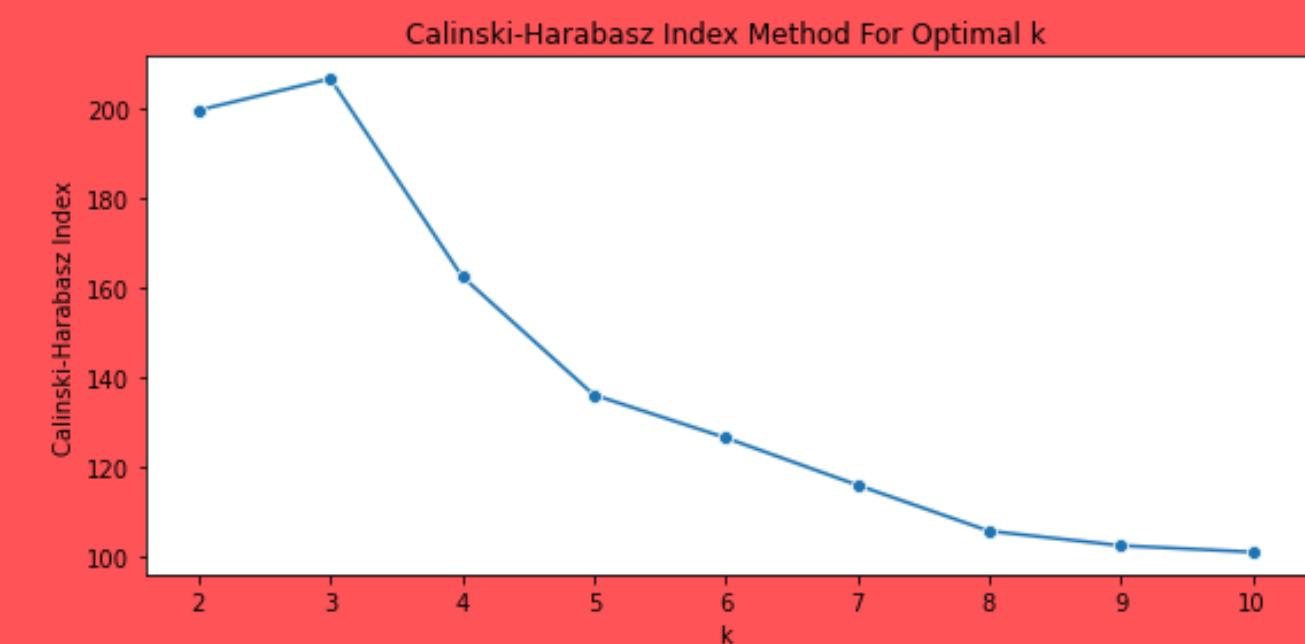
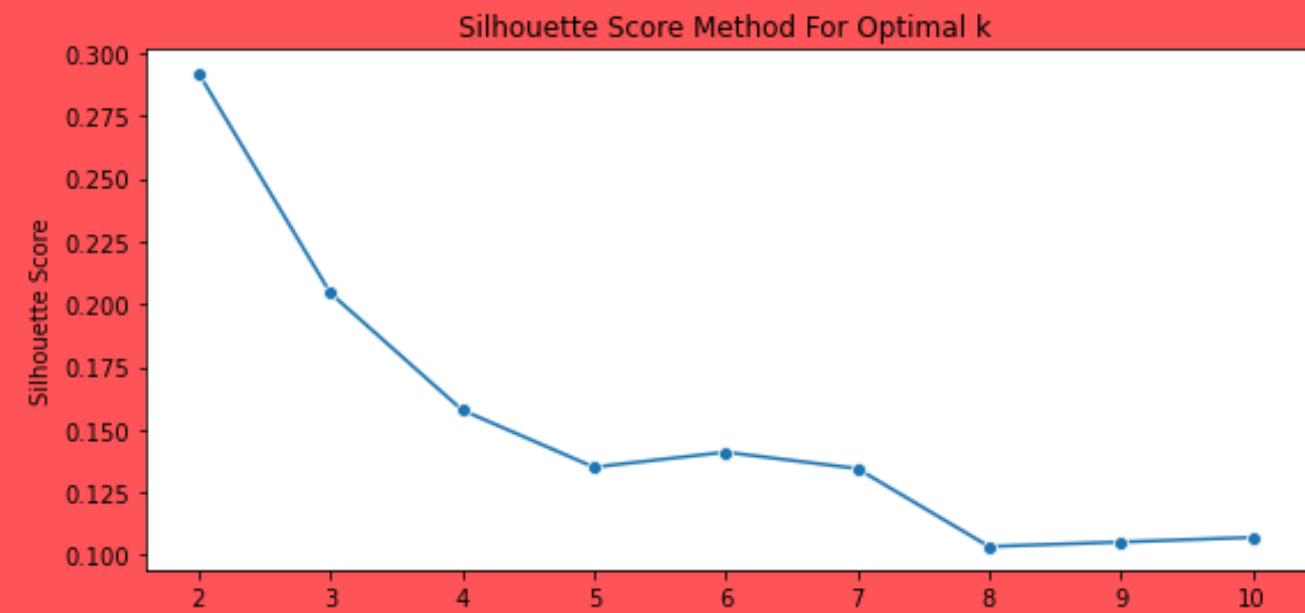
Rozwazane modele

- KMeans - przeanalizowałyśmy różne liczby klastrów (od 2 do 10)
- DBSCAN - wyniki wykazały nieskuteczność tej metody, ponieważ prawie wszystkie utwory zostały przypisane do jednego klastra.
- Spectral Clustering - pierwsze wyniki dość dobre, więc postanowiłyśmy za pomocą metryk znaleźć optymalną liczbę klastrów

Finalny model

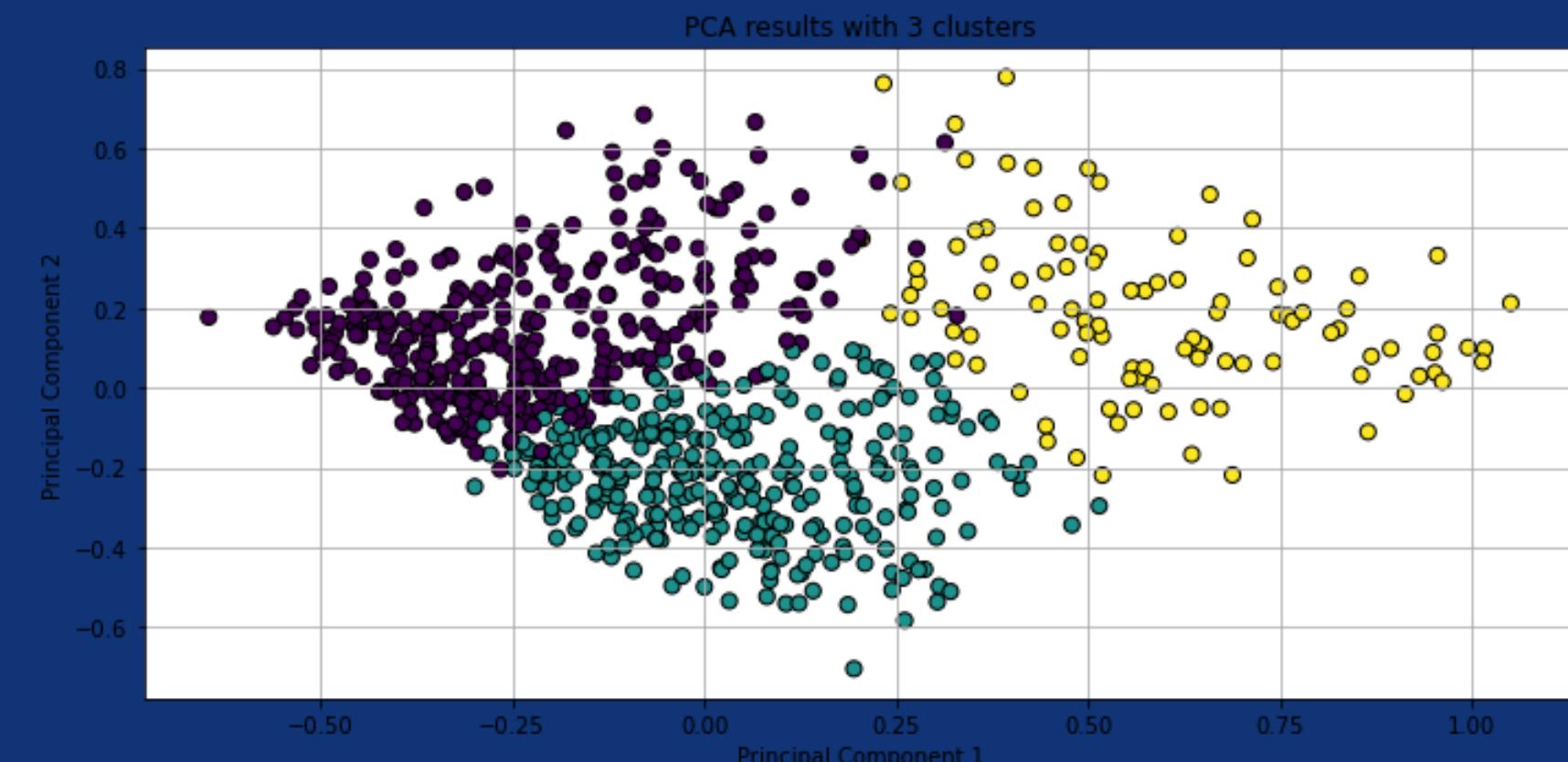
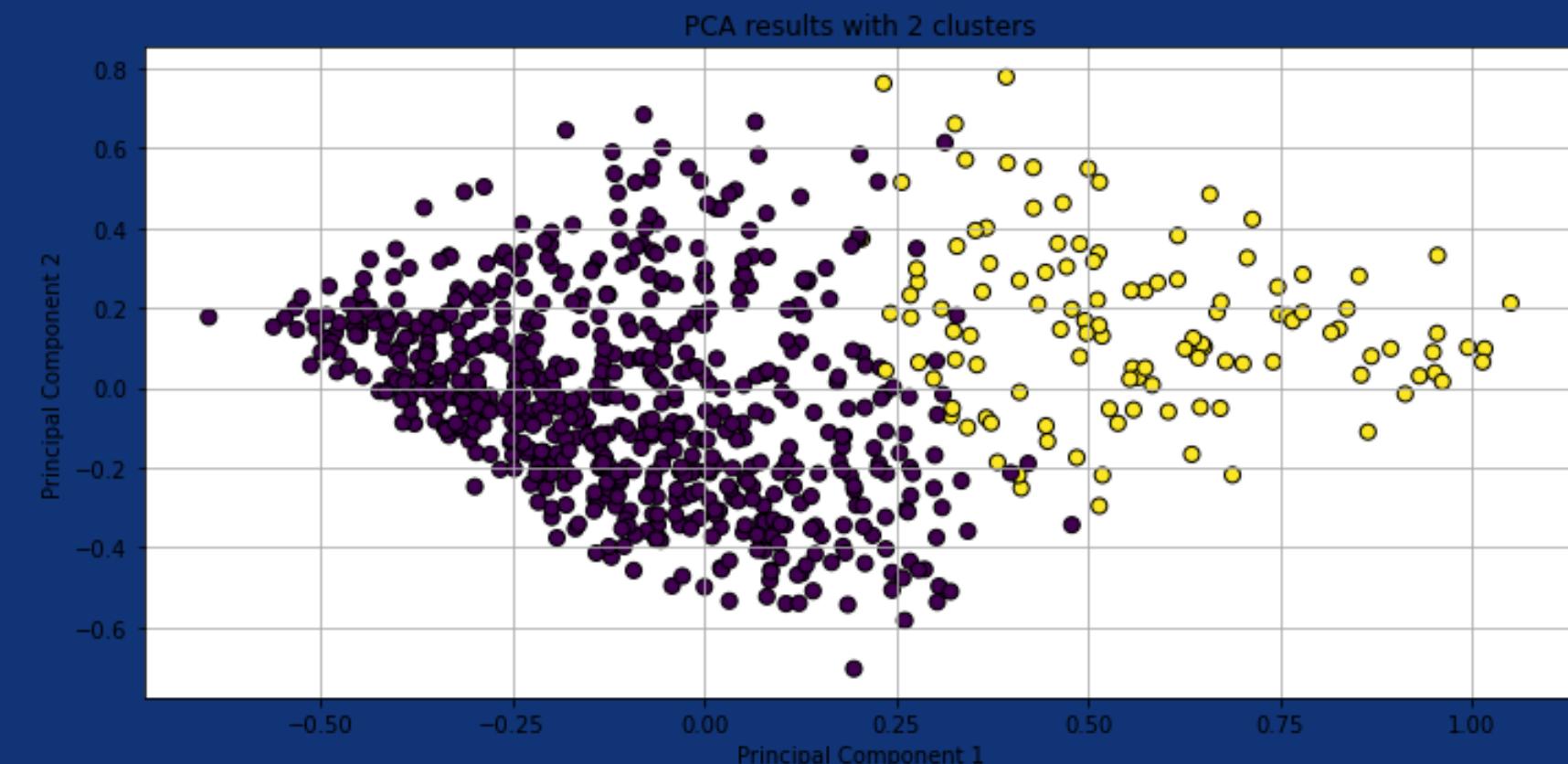
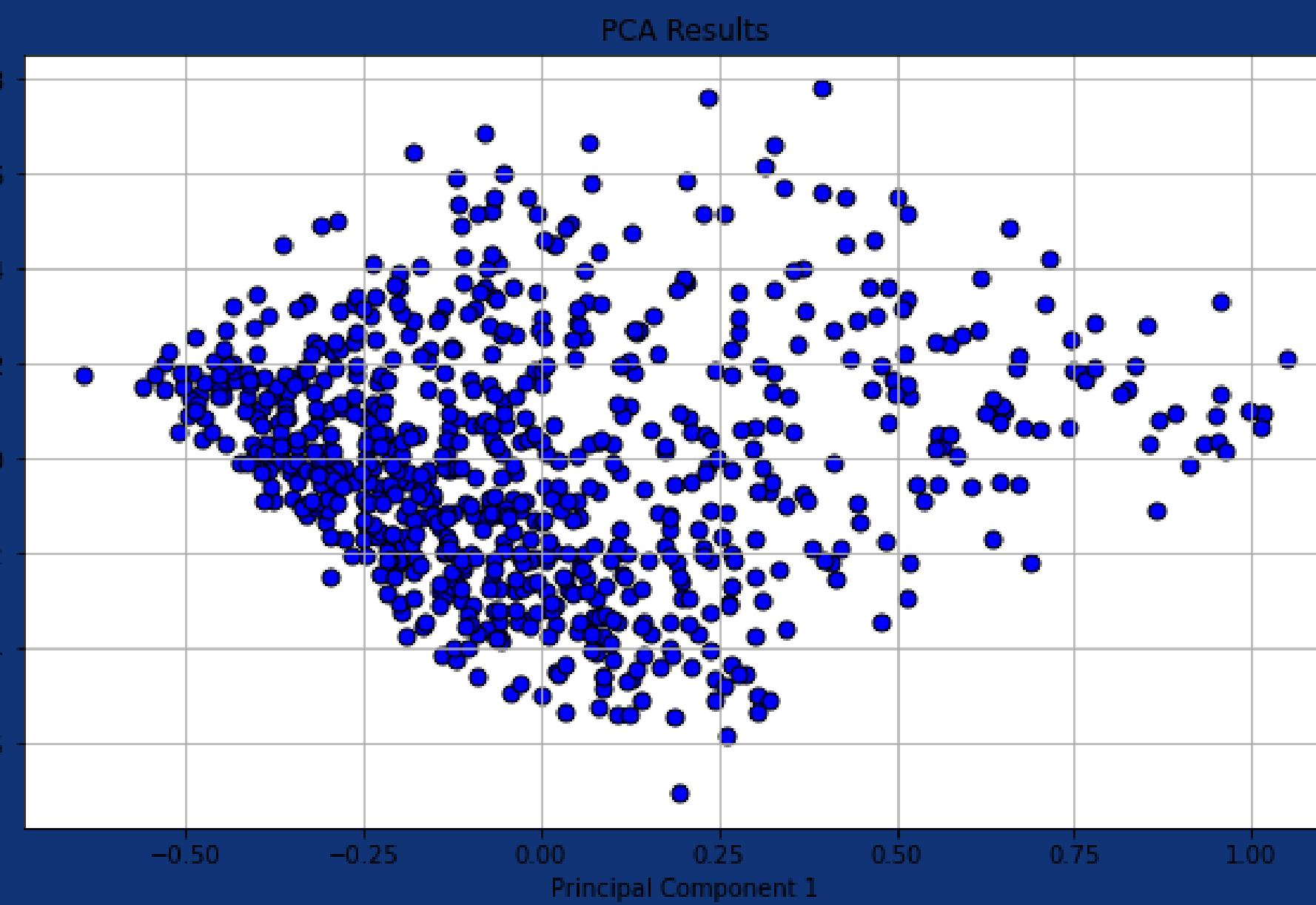
- Metryka Silhouette Score osiągnęła najwyższą wartość dla 2 klastrów, wskazując na dobrą spójność wewnętrz klastrów i rozdzielność między nimi.
- indeks Calinski-Harabasz wskazywał na najlepszą jakość klasteryzacji przy 3 klastrach, co sugeruje optymalną relację między rozproszeniem wewnętrz klastrów a ich rozdzielnością.
- Metryka Davies-Bouldin Score preferowała rozwiązanie z 2 klastrami, sugerując, że klastry w tym przypadku są bardziej zwarte i lepiej oddzielone.

Na podstawie tych wyników postanowiliśmy przetestować model z 3 klastrami.





Wizualizacja klastrów



Czy istnieje istotna różnica między klastrami?

{ Hipoteza zerowa H_0 : klastry się nie różnią
Hipoteza alternatywna H_1 : klastry się różnią



Test statystyczny ANOVA dla każdej z cech
poziom istotności = 0.05
 $p < 0.05 \Rightarrow$ odrzucenie H_0

Wyniki testów

- Dla niemal każdej cechy, wartość p jest mniejsza od poziomu istotności.
- Jedynie dla cechy liveness przyjmujemy hipotezę zerową, ponieważ $p = 0.113 > 0.05$, co oznacza, że w tej kategorii nie ma istotnych różnic między klastrami.
- Dla BPM odrzucamy hipotezę zerową, co wskazuje na różnice między klastrami, choć nie są one bardzo znaczące.
- W przypadku cech speechiness i instrumentalness różnice są istotne, ale nie tak znaczące jak dla pozostałych cech.
- Dla pozostałych cech, wartość p jest zdecydomnie mniejsza od poziomu istotności, co wskazuje na bardzo istotne różnice między klastrami.

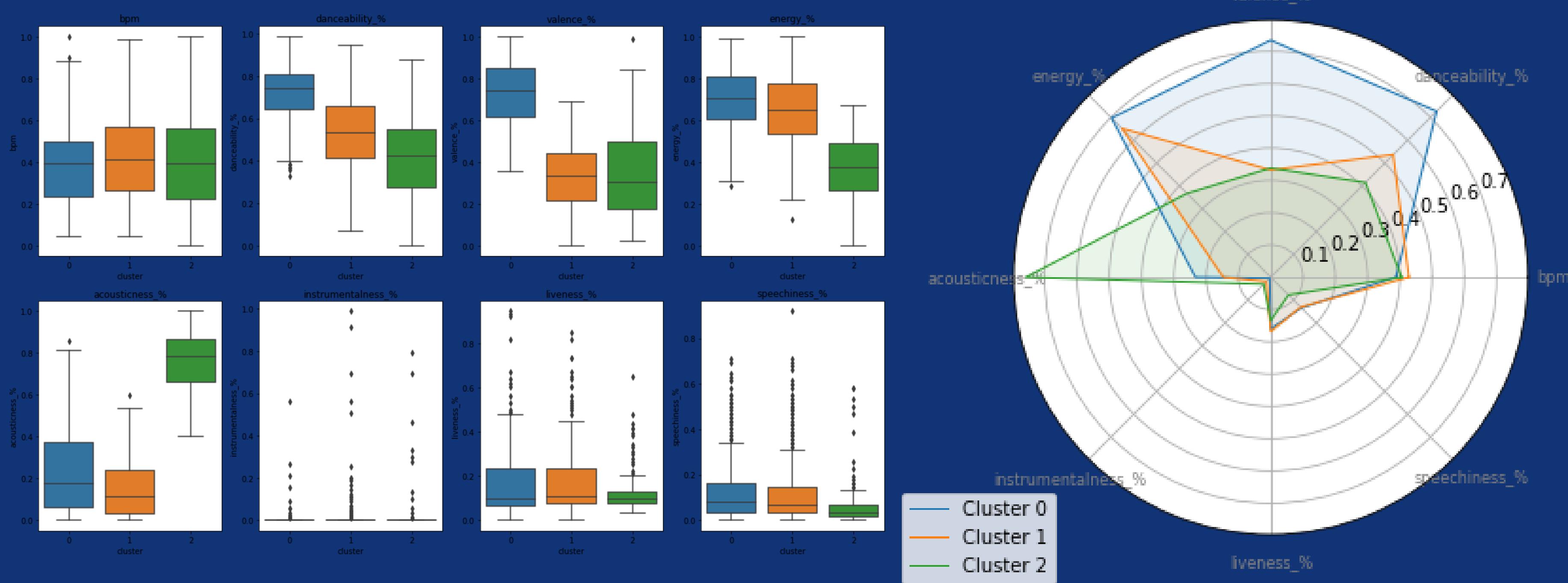
Cechy	Statystyka F	p-wartość
bpm	3.3681	0.0350
danceability%	188.4479	4.9155e-67
valence%	582.4478	2.2058e-153
energy%	197.1043	1.6163e-69
acousticness%	525.0887	2.5394e-143
instrumentalness%	5.0061	0.0069
liveness%	2.1888	0.1128
speechiness%	6.1512	0.0022



Analiza klastrów

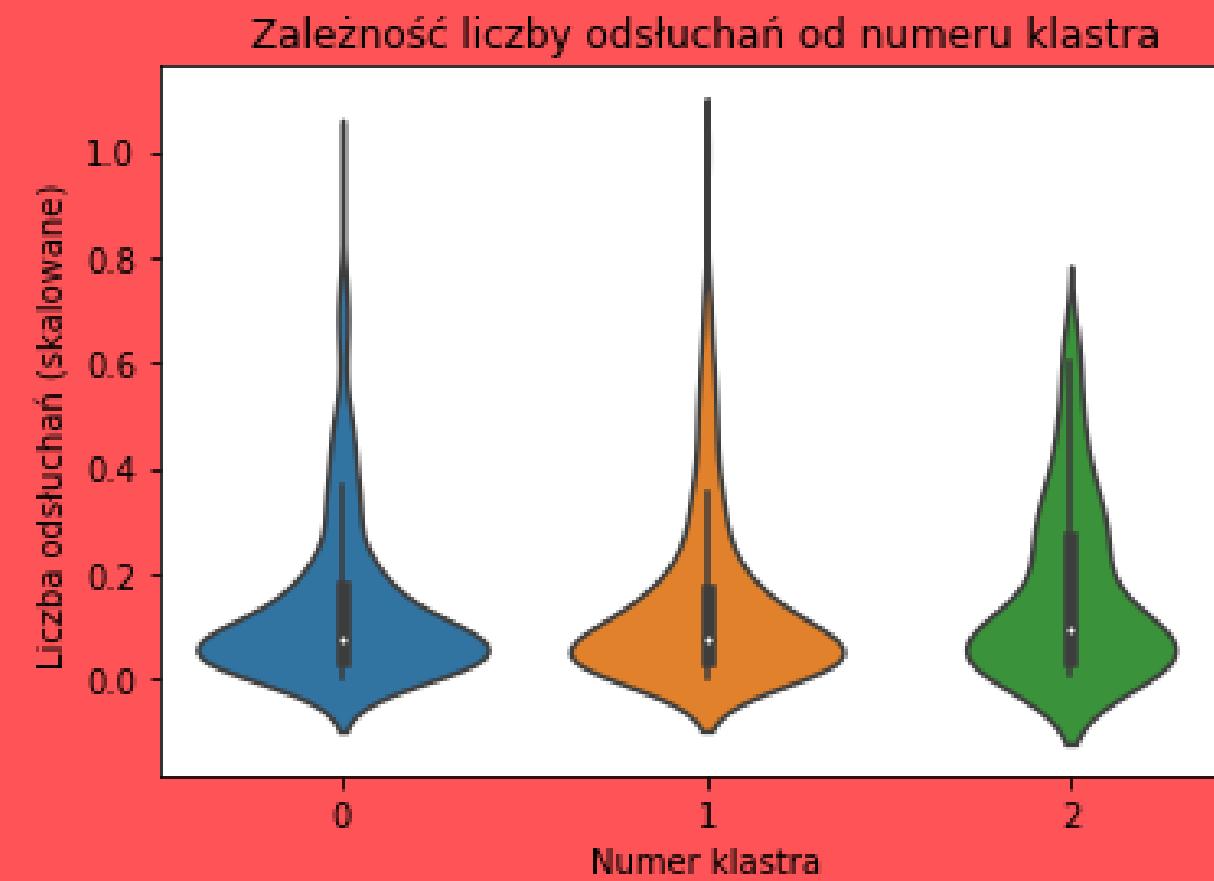


Cechy muzyczne w klastrach

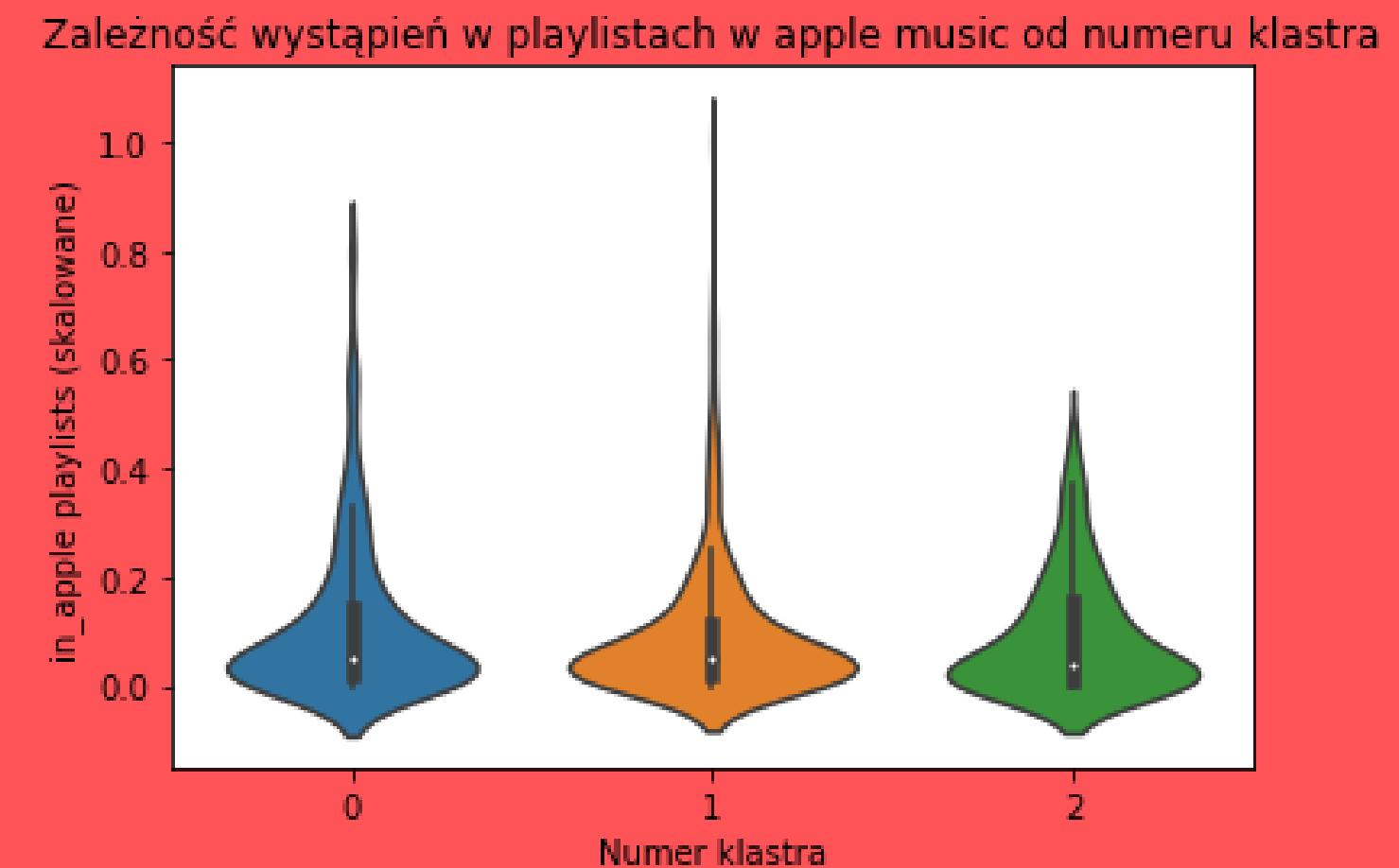


NR KLASTRA	0	1	2
TYP PIOSENEK	bardzo energetyczne, taneczne, pozytywne, nieakustyczne, nieinstrumentalne	najszybsze, taneczne, najmniej pozytywne, nieakustyczne, nieinstrumentalne	średnio szybkie, mało taneczne, mało pozytywne, bardzo akustyczne i najbardziej instrumentalne
MOŻLIWE GATUNKI	pop, dance pop, hip hop, funk, R&B	pop, trap, rap, techno	acoustic rock, classical, indie rock, indie folk, jazz
PRZYKŁADOWE PIOSENKI	"Shape of You" Ed Sheeran, "One Dance" Drake, WizKid, Kyla; "Sweater Weather" The Neighbourhood, "Do I Wanna Know?" Arctic Monkeys	"STAY" (with Justin Bieber), "Believer" Imagine Dragons, "Starboy" The Weeknd, Daft Punk, "Without Me" Eminem	"Shallow" Lady Gaga, Bradley Cooper, "Radio" Lana Del Rey, "As It Was" Harry Styles

Zwiazek klastrów z innymi zmiennymi

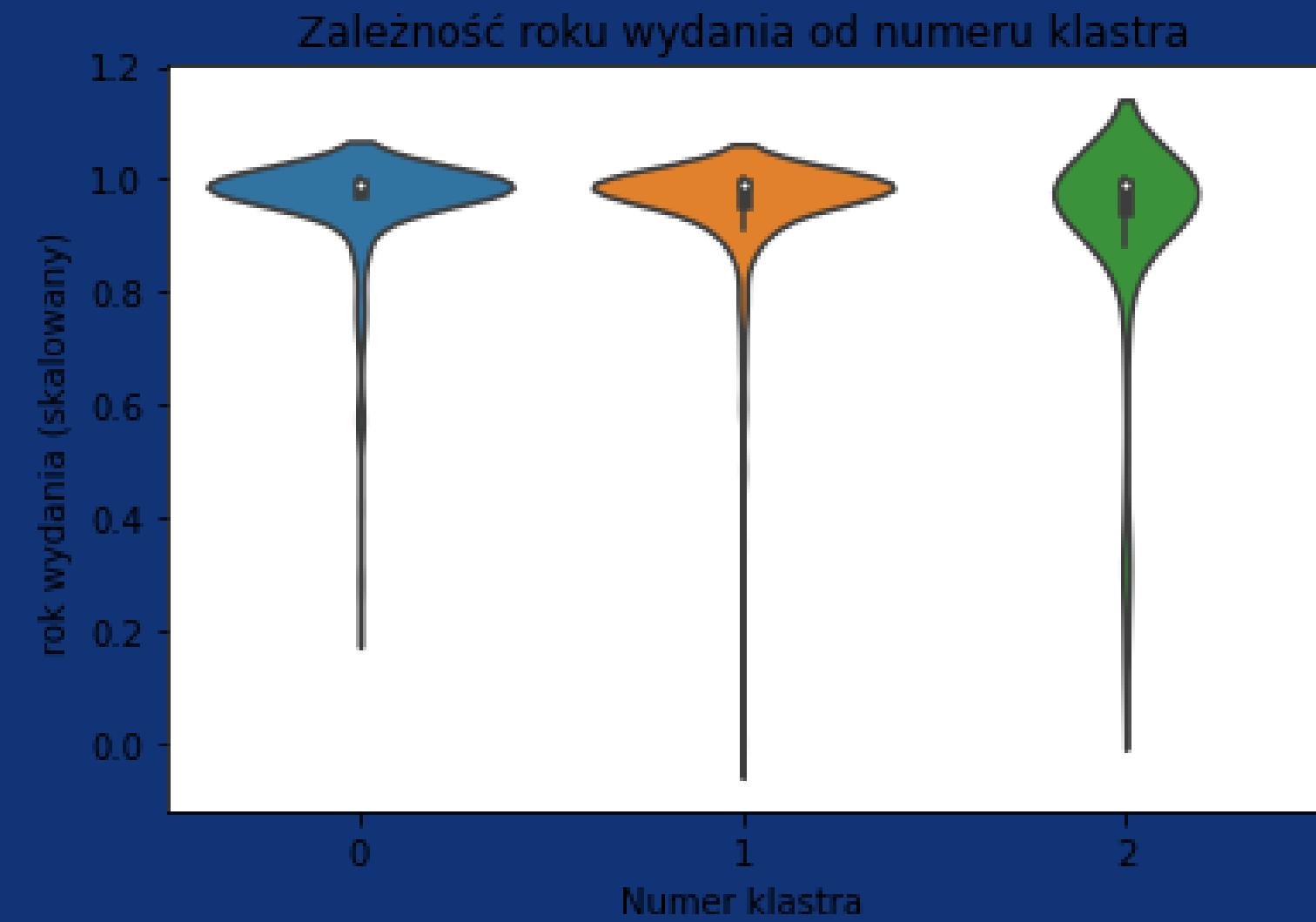


Brak istotnej różnicy w liczbie odtworzeń. Uważamy, że każdy lubi inny typ muzyki, więc nie przewidywałyśmy tutaj dużych różnic. Można zauważyć, że piosenkom z klastra 2 "ciężej jest się wybić", gdyż na wykresie wiele z nich znajduje się w średnio popularnych piosenkach i raczej mało w tych najbardziej popularnych.

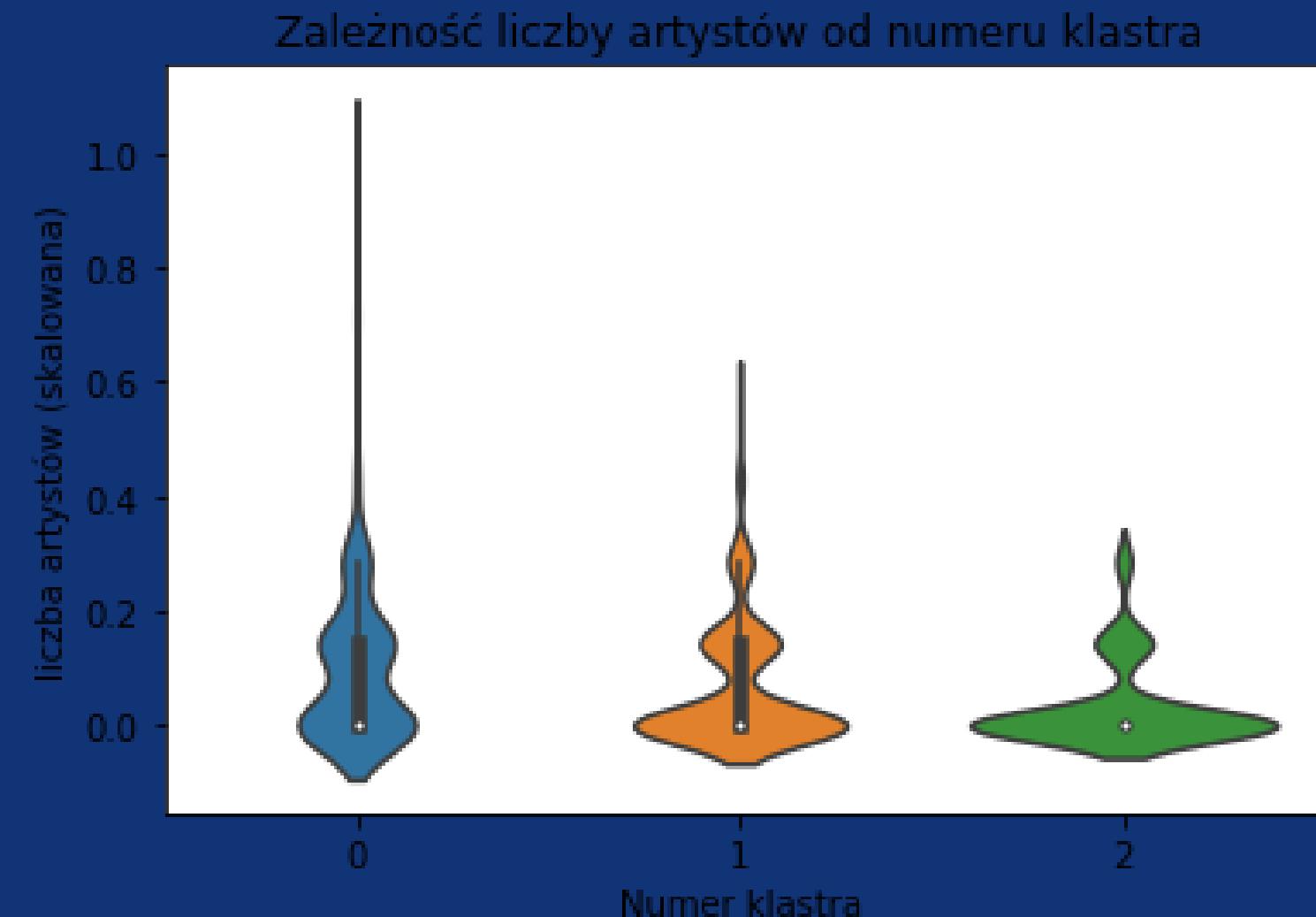


Najbardziej powtarzające się piosenki w playlistach w apple music znajdują się w klastrze 1 i 0.

Zwiazek klastrów z innymi zmiennymi



Wśród popularnych, obecnie produkowanych piosenek jest najmniej piosenek jest z klastra 2. Również największa część piosenek produkowanych wcześniej niż ostatnie kilka lat należy właśnie do klastra 2.



- Klaster 0 - często wykonywane piosenki przez kilku artystów, niemal tak samo często przez 1 i 2 artystów, albo nawet do 8 => prawdopodobnie są to piosenki z gatunku pop, rap, hip-hop - to właśnie tam często występuje kilku artysow
- Klaster 1 - najczęściej wykonywane solo, ale zdarzaja się piosenki kilku artystów, nawet do 5 artystow
- Klaster 2 - zdecydomana większość piosenek solo, zdarzaja się piosenki 2 lub 3 artystów, ale nie więcej => wskazuje na piosenki typu acoustic, indie, classical etc.