Projekt Klasteryzacja: Spotify 2023

Raport projektu

31 maja 2024

Spis treści

Authors

Barbara Seweryn, Urszula Szczęsna

1 Zamysł biznesowy

Naszym celem było przeprowadzenie klasteryzacji utworów muzycznych na podstawie ich cech charakterystycznych, które opisują różne aspekty każdego utworu. Skoncentrowałyśmy się szczególnie na takich cechach, jak taneczność, energetyczność, akustyczność oraz wiele innych. Proces klasteryzacji miał na celu zidentyfikowanie grup utworów, które mają podobne właściwości muzyczne, a następnie przeprowadzenie analizy tych grup w celu znalezienia podobieństw i różnic między nimi.

2 Dane

2.1 Źródło

Nasze dane pochodzą z `https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023`. Zawierają 943 recordów z 24 cechami.

```
Data columns (total 24 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   track_name            953 non-null    object
 1   artist(s)_name        953 non-null    object
 2   artist_count          953 non-null    int64
 3   released_year         953 non-null    int64
 4   released_month        953 non-null    int64
 5   released_day          953 non-null    int64
 6   in_spotify_playlists  953 non-null    int64
 7   in_spotify_charts     953 non-null    int64
 8   streams               953 non-null    object
 9   in_apple_playlists    953 non-null    int64
 10  in_apple_charts       953 non-null    int64
 11  in_deezer_playlists   953 non-null    object
 12  in_deezer_charts      953 non-null    int64
 13  in_shazam_charts      903 non-null    object
 14  bpm                   953 non-null    int64
 15  key                   858 non-null    object
 16  mode                  953 non-null    object
 17  danceability_%        953 non-null    int64
 18  valence_%             953 non-null    int64
 19  energy_%              953 non-null    int64
 20  acousticness_%        953 non-null    int64
 21  instrumentalness_%    953 non-null    int64
 22  liveness_%            953 non-null    int64
 23  speechiness_%         953 non-null    int64
```

Rysunek 1: Cechy danych

Cechy możemy podzielić na 3 grupy:

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

2.2 Podział danych

Nasze dane podziliśmy na dane treningowe na podstawie, których budowałyśmy nasz model oraz dane walidacyjne. Proporcje podziału

- 80% - train data,
- 20% - validation data,

2.3 Preprocessing danych

Podczas wstępnej analizy danych, wykonałyśmy następujące kroki, aby przygotować dane do dalszej analizy:

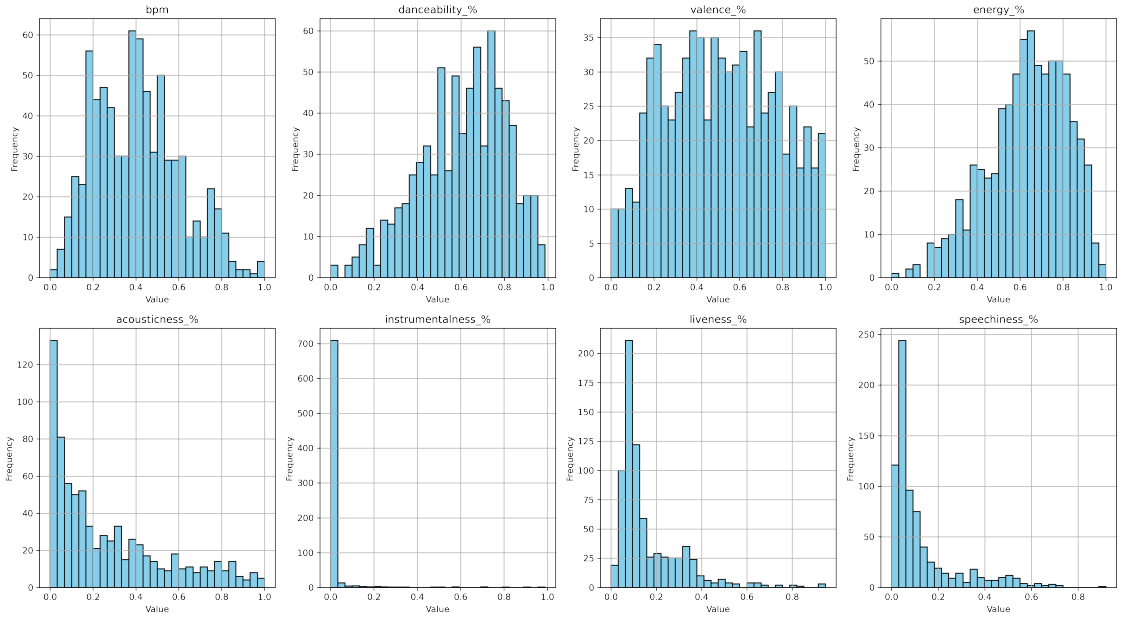
1. **Konwersja Typów Danych:**
   - Zmieniłyśmy wartości typu `object` na `string` lub `int64`, aby zapewnić spójność danych i ułatwić ich dalszą obróbkę.
2. **Zastępowanie Braków Danych:**
   - W kolumnach `streams`, `in_deezer_playlist` oraz `in_deezer_charts` brakujące wartości zastąpiłyśmy wartością 0. Uznałyśmy, że brak danych w tych kolumnach oznacza brak występowania w odpowiednich notowaniach lub playlistach.
3. **Uzupełnianie Braków w Kolumnie `key`:**
   - Braki danych w kolumnie `key` uzupełniłyśmy losowymi wartościami, zgodnie z rozkładem wartości tej kolumny w całym zbiorze danych.
4. **Usuwanie Duplikatów:**
   - Usunęłyśmy duplikaty w zbiorze danych, aby uniknąć powielania informacji i zapewnić dokładność analizy.
5. **Transformacja**
   - Niektóre zmienne miały bardzo skośny rozkład, dlatego podjęłyśmy próbę transformacji logarytmicznej. Niestety, wyniki klasteryzacji po zastosowaniu tej transformacji były gorsze, więc zdecydowałyśmy się z niej zrezygnować.
6. **Skalowanie**
   - Na sam koniec, przeskalowałyśmy całą ramkę danych za pomocą metody MinMaxScaler.

Po tym wszystkim nasze dane trenigowe miały 753 wiersze a dane walidacyjne 189 wierszy.

3 Klasteryzacja

Podczas klasteryzacji postanowiliśmy skupić się na cechach, które opisują parametry piosenki. Wybrane kolumny to:

- `bpm`
- `danceability_%`
- `valence_%`
- `energy_%`
- `acousticness_%`
- `instrumentalness_%`
- `liveness_%`
- `speechiness_%`



Rysunek 2: Rozkład wybranych zmiennych

Rozkład części cech jest zbliżony normalnego (pierwszy rząd wykresów), zatem nie było potrzeby transformacji ich transformacji. Dla pozostałych zmiennych gołym okiem zauważyć można, że występuje wiele wartości bliskich zera. Jest to szczególnie widoczne dla zmiennej `instrumentalness_%`, której histogram pokazuje, że prawie wszystkie wartości znajdują się w zakresie [0,0.05]. Logicznym postępowaniem wydawało się tutaj za ten zlogarytmizowanie tej kolumny, aby w dalszym modelowaniu miała ona zbliżone "znaczenie"do pozostałych zmiennych.
Okazało się jednak, że praktyka ta wcale nie ułatwiła procesu klasteryzacji, a nawet spowodowała, że klastry były mniej uporządkowane. Dlatego finalnie zdecydowałyśmy się zrezygnować z tego pomysłu.

3.1 Rozważane modele

W procesie kalsteryzacji rozważałyśmy kilka modeli:

- KMeans - przeanalizowałyśmy różne liczby klastrów (od 2 do 10) a następnie używając różnych metryk: Metoda Łokcia, Silhouette Score, Calinkski-Harabasz Index i Davies-Bouldin Index.

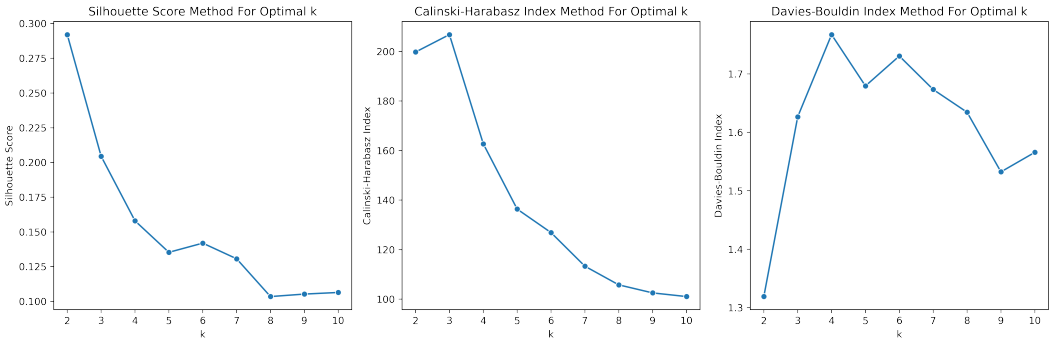
Metoda	**Optymalna liczba klastrów**
Metoda łokcia	3 klastry
Silhouette Score	2 klastry
Calinski-Harabasz	3 klastry
Davies-Bouldin	2 lub 9 klastrów

Tabela 1: Wyniki metryk klasteryzacji

- DBSCAN - wyniki wykazały nieskuteczność tej metody, ponieważ prawie wszystkie utwory zostały przypisane do jednego klastra.
- Spectral Clustering - pierwsze wyniki były dość dobre, więc postanowiłyśmy za pomocą metryk znaleźć optymalną liczbę klastrów

3.2 Wybrany model

Analiza Spectral Clustering na podstawie metryk: Silhouette Score, Calinkski-Harabasz Index i Davies-Bouldin Index.

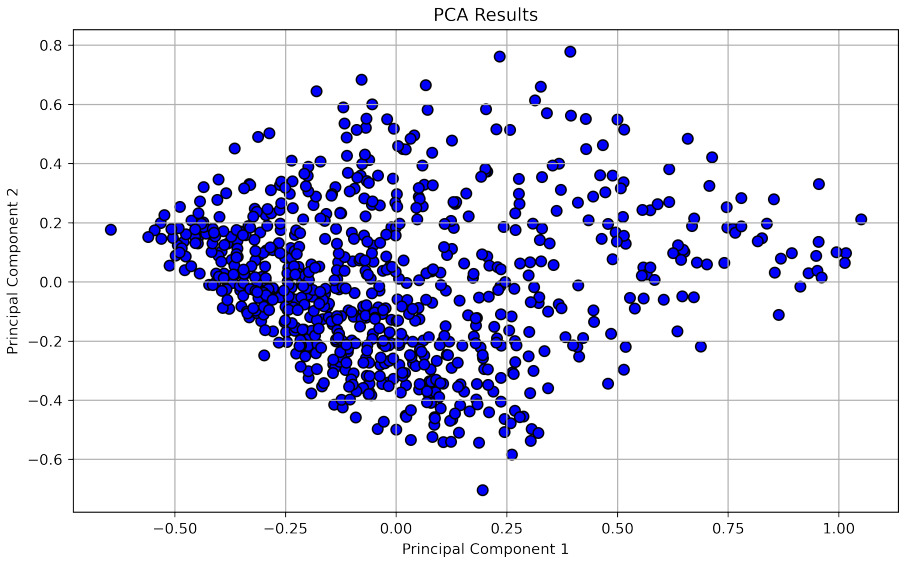


Rysunek 3: Metryki

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

3.3 Wizualizacja klastrów

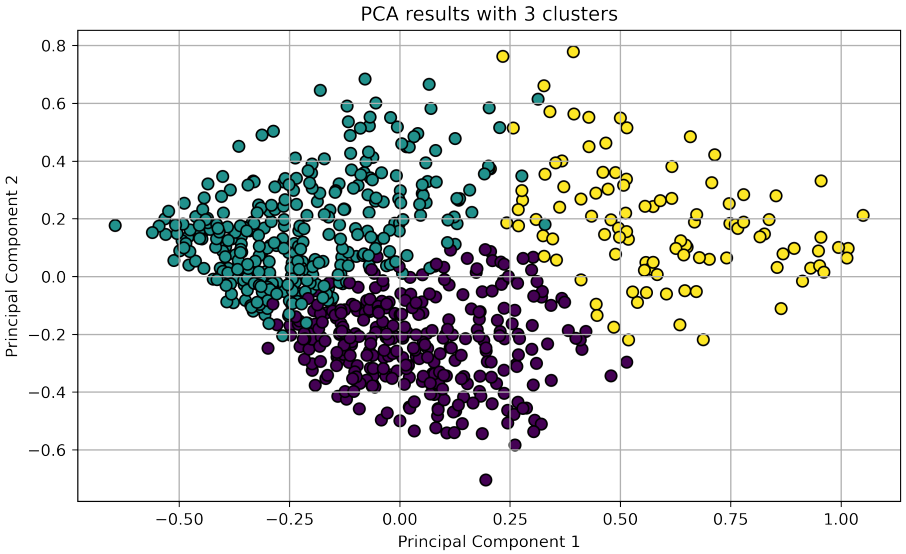
Kolejnym krokiem była wizualizacja danych w dwóch wymiarach przy użyciu analizy głównych składowych (PCA). Dzięki temu mogłyśmy zredukować złożoność danych i lepiej zobrazować struktury w nich zawarte. Po dokonaniu redukcji wymiarów, naniosłyśmy klastry na te dwa wymiary, co umożliwiło nam graficzne przedstawienie wyników klasteryzacji i lepsze zrozumienie relacji między utworami w różnych klastrach. Główne składowe (osie wykresu) reprezentują kombinacje oryginalnych cech danych, które maksymalizują wariancję. Pierwsza główna składowa (oś X) wyjaśnia 34% wariancji a druga główna składowa (oś Y) wyjaśnia 22 % wariancji.



Rysunek 4: PCA - 2 components

Choć klastrów nie widać wyraźnie, można zaobserwować pewne zagęszczenia punktów, szczególnie po lewej stronie wykresu. W lewej części wykresu punkty są bardziej skupione, co sugeruje, że utwory te mogą mieć bardziej zbliżone cechy. W prawej części wykresu punkty są bardziej rozproszone, co może wskazywać na większą różnorodność cech w tej grupie utworów.

Następnie nasiosłyśmy wyniki klasteryzacji.



Rysunek 5: Klastry

3.4 Analiza statystyczna klastrów

Przeprowadziłyśmy analizę statystyczną klastrów, aby sprawdzić, czy istnieją istotne różnice między nimi pod względem różnych cech. Do analizy wykorzystałyśmy test jednoczynnikowej analizy wariancji (ANOVA), który pozwala na porównanie średnich wartości cech w różnych klastrach.

Hipotezy:

- Hipoteza zerowa (H0): Nie ma różnic między klastrami.
- Hipoteza alternatywna (H1): Istnieją różnice między klastrami.

Ustalony poziom istotności wynosi 0,05. Jeśli wartość p jest mniejsza od 0,05, odrzucamy hipotezę zerową, co oznacza, że różnice między klastrami są istotne statystycznie.

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

Tabela 2: Wyniki testów ANOVA dla różnych cech

Na podstawie wyników można stwierdzić że:

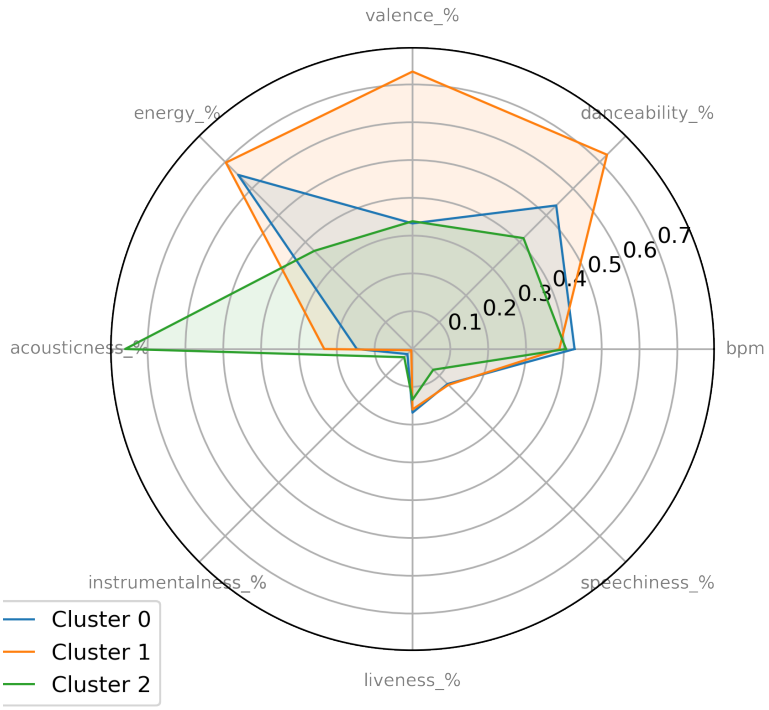
- Dla niemal każdej cechy, wartość p jest mniejsza od poziomu istotności.
- Jedynie dla cechy liveness przyjmujemy hipotezę zerową, ponieważ $p = 0.113 > 0.05$, co oznacza, że w tej kategorii nie ma istotnych różnic między klastrami.

- 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 zdecydowanie mniejsza od poziomu istotności, co wskazuje na bardzo istotne różnice między klastrami.

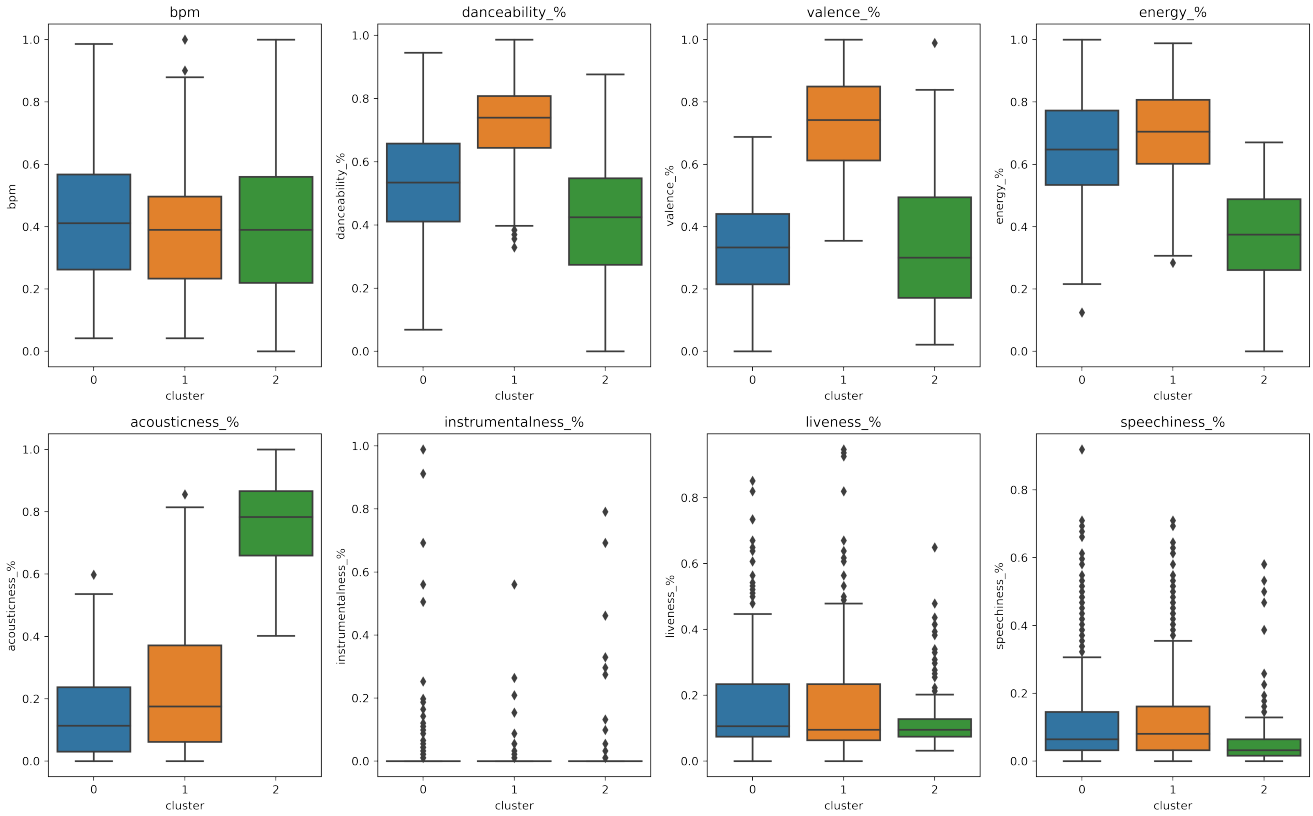
4 Analiza piosenek należących do klastrów

Na sam koniec zajęłyśmy się analizą piosenek z poszczególnych klastrów aby odkryć zależności pomiędzy piosenkami z tych samych klastrów.

4.1 Wykres radarowy dla poszczególnych klastrów oraz rozkład zmiennych w zależności od klastrów



Rysunek 6: wykres radarowy



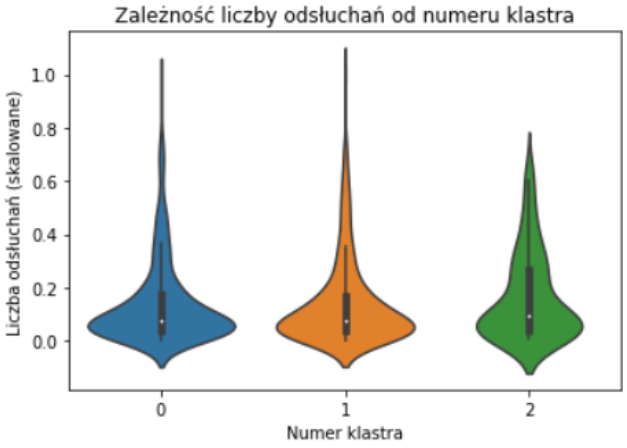
Rysunek 7: Rozkład klastrów względem zmiennych

Z wykresów można wywnioskować że:

1. Klaster 0:
   - piosenki bardzo energetyczne, taneczne, pozytywne, nieakustyczne, nieinstrumentalne
   - możliwe gatunki: pop, dance pop, hip hop, funk, r&b
   - przykładowe utwory z tego kalstra to m.in. "Shape of You"Ed Sheeran, "One Dance"Drake, WizKid, Kyla; "Sweater Weather"The Neighbourhood, "Do I Wanna Know?"Arctic Monkeys
2. Klaster 1:
   - piosenki najszybsze, taneczne, najmniej pozytywne, nieakustyczne, nieinstrumentalne
   - możliwe gatunki: trap, rap, techno
   - przykładowe utwory z tego klastra to m.in. "STAY"(with Justin Bieber), "Believer"Imagine Dragons, "Starboy"The Weeknd, Daft Punk, "Without Me"Eminem.
3. Klaser 2:
   - piosenki średnio szybkie, mało taneczne, mało pozytywne, bardzo akustyczne i najbardziej instrumentalne
   - możliwe gatunki: accoustic rock, classical, indie rock, indie folk, jazz
   - przykładowe utwory z tego klastra to m.in. "Shallow"Lady Gaga, Bradley Cooper, "Radio"Lana Del Rey, "As It Was"Harry Styles.

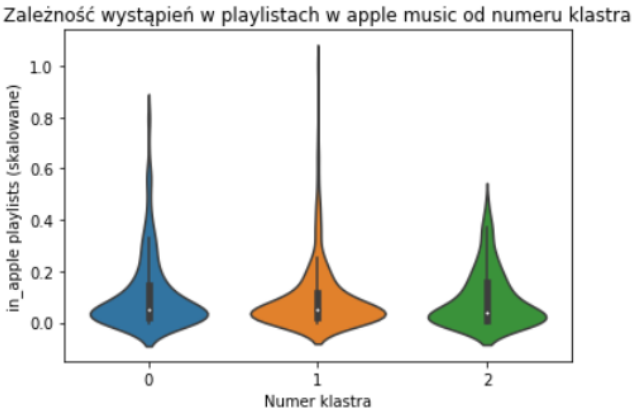
4.2 Związek klastrów z innymi zmiennymi

Warto też przyjrzeć się jak wygląda wpływ klastra na przykładowo zmienne takie jak `streams`, `in_apple_playlists`, `released_year` czy `artist_count`. Poniżej przedstawione zostały wykresy skrzypcowe zależności tych zmiennych od nr klastru.



Rysunek 8: Wykres zależnosci zmiennej `streams` od klastra

Łatwo zauważyć 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.

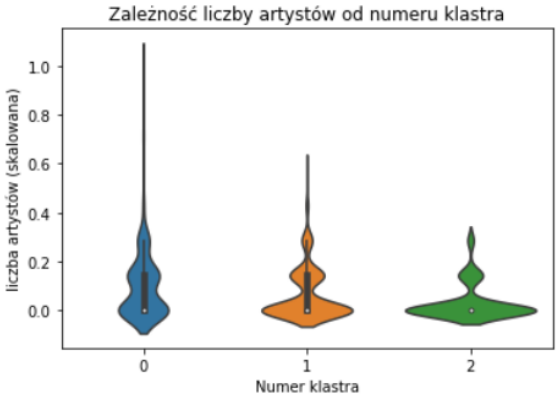


Rysunek 9: Wykres zależności zmiennej `in_apple_playlists` od klastra

Z powyższego wykresu wynika, że najbardziej powtarzajace sie piosenki w playlistach w apple music znajduja sie w klastrze 1 i 0.

Rysunek 10: Wykres zależności zmiennej `released_year` od klastra

Z wykresu wynika, że 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.



Rysunek 11: Wykres zależności zmiennej `artist_count` od klastra

Z tego wykresu wynika, że

- Klaster 0 - czesto wykonywane piosenki przez kilku artystow, niemal tak samo czesto przez 1 i 2 artystow, albo nawet do 8 => prawdopodobnie są to piosenki z gatunku pop, rap, hip-hop - to wlasnie tam czesto wystepuje kilku artysow
- Klaster 1 - najczesciej wykonywane solo, ale zdarzaja się piosenki kilku artystów, nawet do 5 artystow
- Klaster 2 - zdecydowana wiekszosc piosenek wykonywana jest solo, zdarzaja się piosenki 2 lub 3 artystów, ale nie więcej => znowu to wskazuje na piosenki typu acoustic, indie, classical etc.

4.3 Subiektywna ocena podziału

Niestety w ramce nie ma informacji na temat gatunku poszczególnych piosenek, dlatego nie możemy dokładnie zweryfikować naszych wniosków co do korelacji klastrów z gatunkami muzycznymi. Zamiast tego przyjrzałyśmy się kilku piosenkom i artystom, których muzykę znamy, aby do jakiego klastra zostały zakwalifikowane i czy zgadza się to z naszymi odczuciami.
W większości przypadków dopasowania były trafne, a gatunki i charaktery piosenek często zgadzały się z tymi które dopasowałyśmy do klastrów. Poniżej załączyłyśmy informację, do jakich klastrów zakwalifikowane zostało 15 najczęściej streamowanych piosenek.

cluster	track_name	artist(s)_name	artist_count	released_year	released_month	released_day	in_spotify_playlists	in_spotify_charts	streams
2	Blinding Lights	The Weeknd	1	2019	11	29	43899	69	3.703895e+09
1	Shape of You	Ed Sheeran	1	2017	1	6	32181	10	3.562544e+09
1	Someone You Loved	Lewis Capaldi	1	2018	11	8	17836	53	2.887242e+09
1	Dance Monkey	Tones and I	1	2019	5	10	24529	0	2.864792e+09
0	Sunflower - Spider-Man: Into the Spider-Verse	Post Malone, Swae Lee	2	2018	10	9	24094	78	2.808097e+09
1	One Dance	Drake, WizKid, Kyla	3	2016	4	4	43257	24	2.713922e+09
0	STAY (with Justin Bieber)	Justin Bieber, The Kid Laroi	2	2021	7	9	17050	36	2.665344e+09
0	Believer	Imagine Dragons	1	2017	1	31	18986	23	2.594040e+09
2	Closer	The Chainsmokers, Halsey	2	2016	5	31	28032	0	2.591224e+09
0	Starboy	The Weeknd, Daft Punk	2	2016	9	21	29536	79	2.565530e+09
1	Perfect	Ed Sheeran	1	2017	1	1	16596	13	2.559529e+09
0	Heat Waves	Glass Animals	1	2020	6	28	22543	63	2.557976e+09
2	As It Was	Harry Styles	1	2022	3	31	23575	130	2.513188e+09
0	SeÃ¯Â¿Â½Ã¯Â¿Â½o	Shawn Mendes, Camila Cabello	2	2019	6	19	15010	2	2.484813e+09
1	Say You Won't Let Go	James Arthur	1	2016	9	9	15722	16	2.420461e+09

Rysunek 12: Klasteryzacja dla 15 piosenek z największą wartością w kolumnie `streams`