Dimensijos mažinimas klasterizavime

Matas Gaulia, Vainius Gataveckas, Dovydas Martinkus
Duomenų Mokslas 3 kursas 2 gr.

Vilnius, 2022

Naudoti duomenys

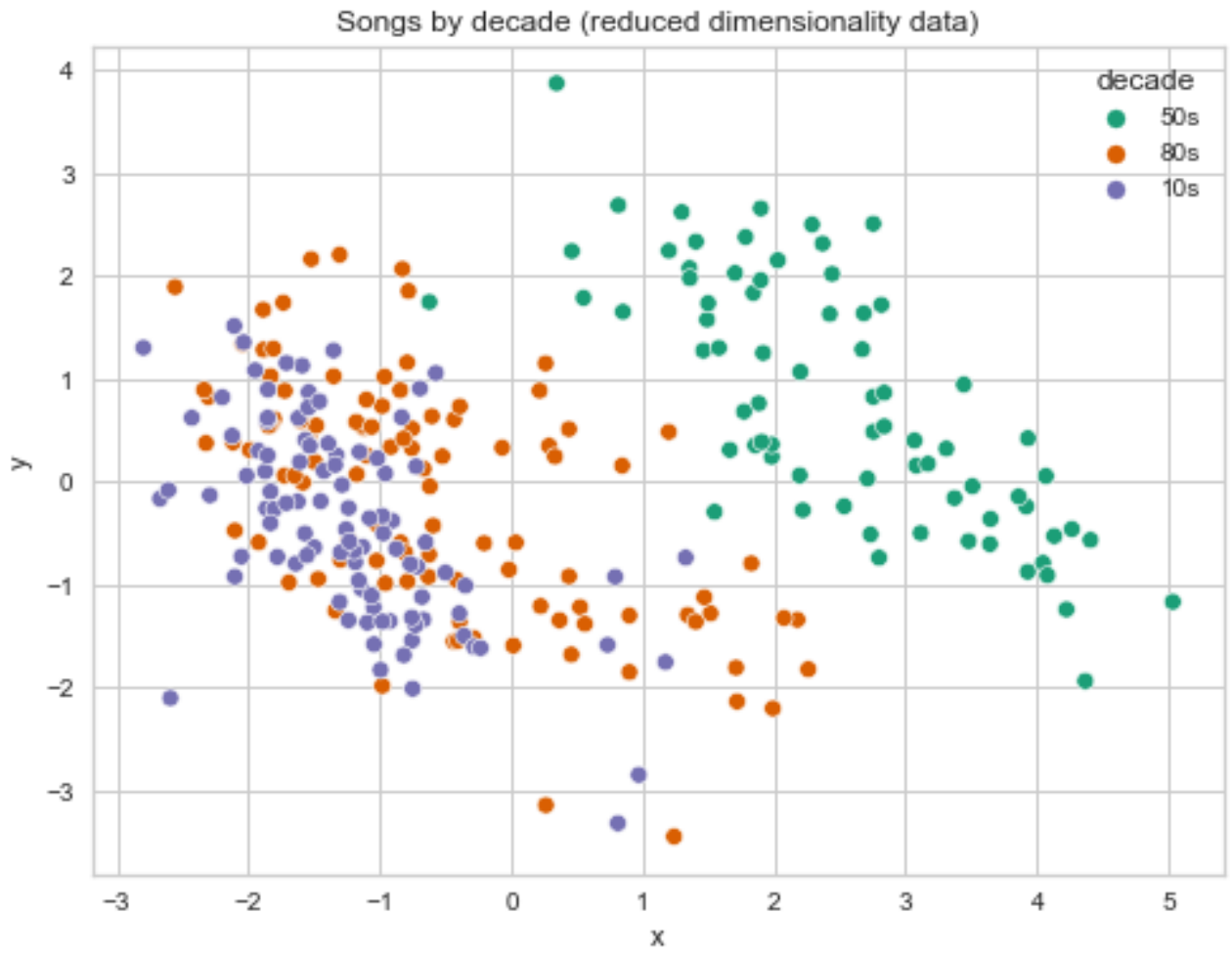
- decade - dainos sukūrimo metų dešimtmetis (50-ieji, 80-ieji ar 2010-ieji)
- tempo - greitis
- energy - energiškumas
- danceability - šokamumas
- loudness – garsumas
- liveness - gyvumas
- valence – pozityvumas
- duration - trukmė
- acousticness - akustiškumas
- speechiness - žodžių kiekis dainoje
- popularity - populiarumas

Prieš tai su dimensijos mažinimo metodais naudotas Spotify dainų duomenų rinkinys.

Požymių matavimo skalės suvienodintos standartizuojant.

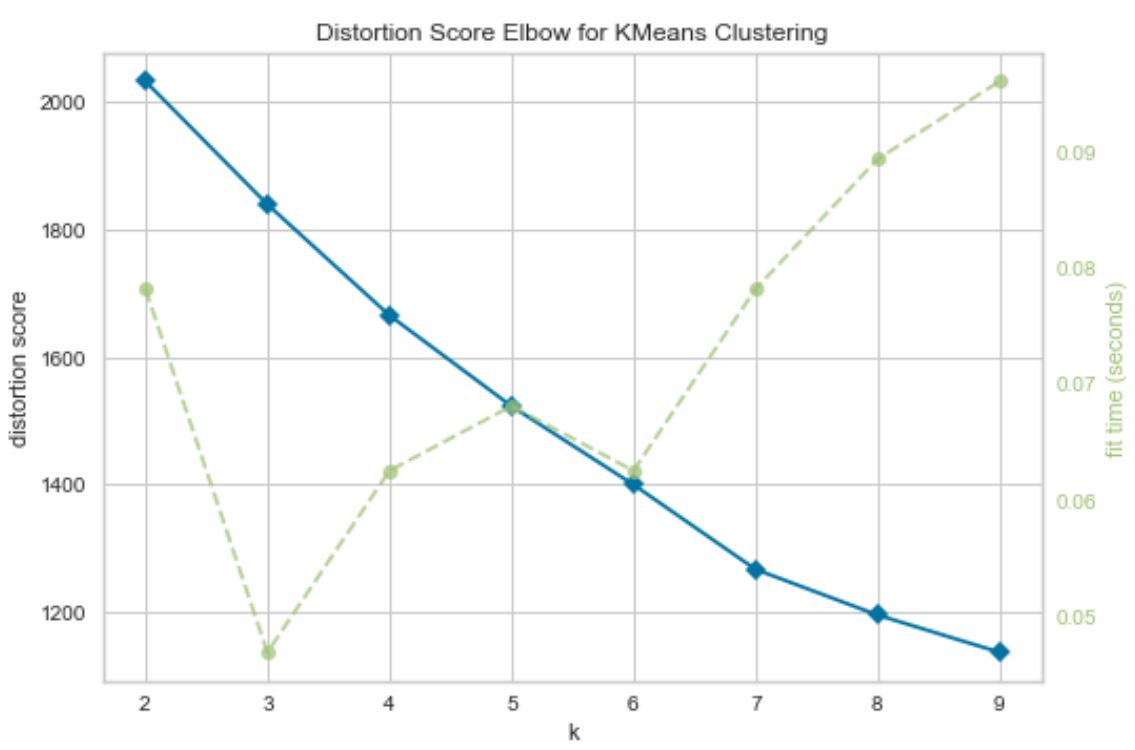
Vizualizavimui dimensija sumažinta iki n=2 naudojantis PCA metodu.

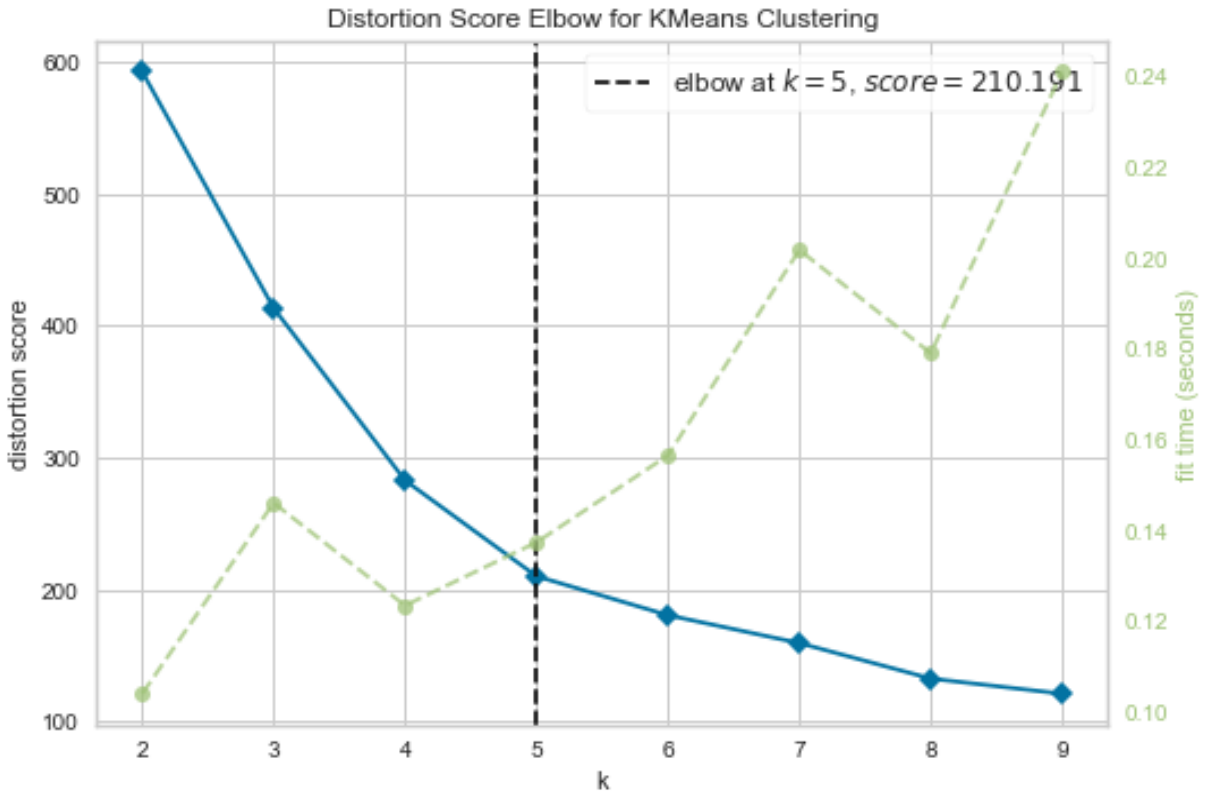
Klasterizuota k-means metodu naudojant originalios dimensijos (n=10) ir PCA metodu sumažintos dimensijos (n=2) duomenis.



k-vidurkių (k-means) metodas

Alkūnės metodas

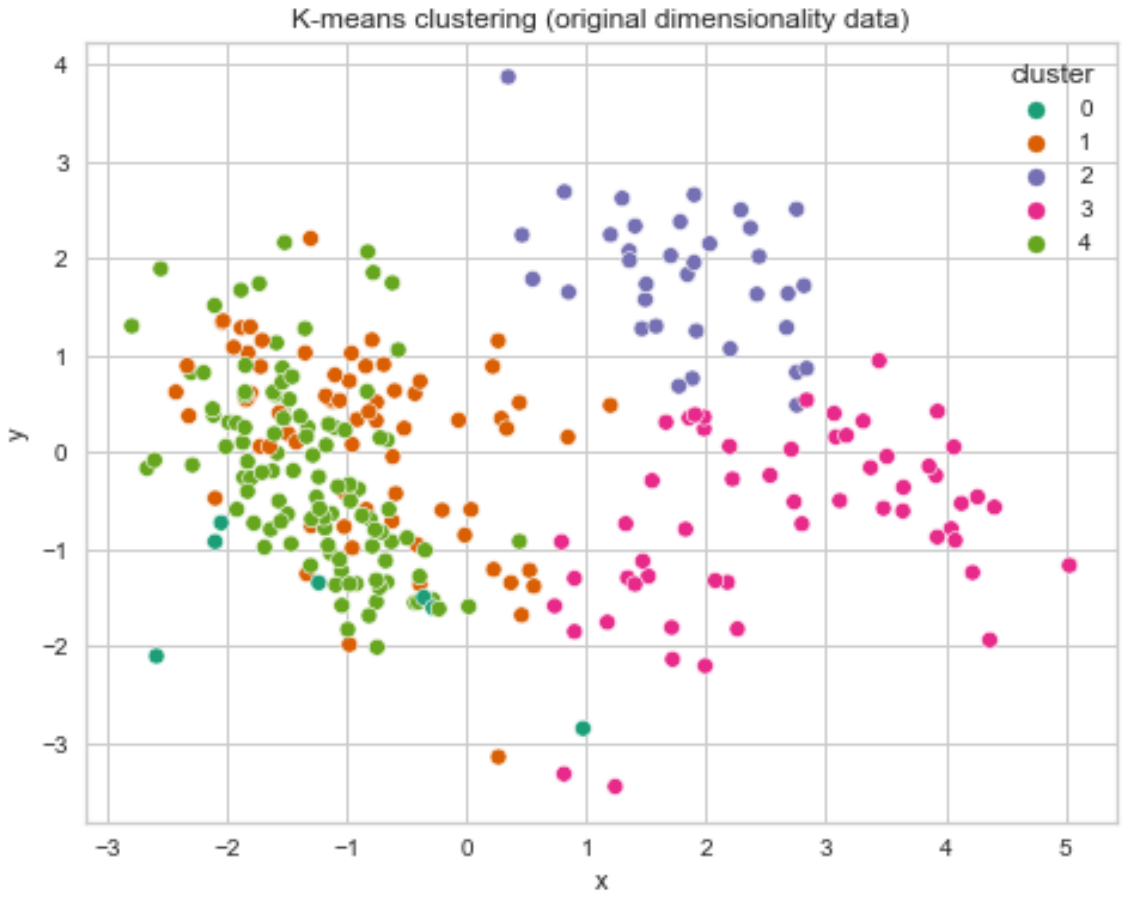


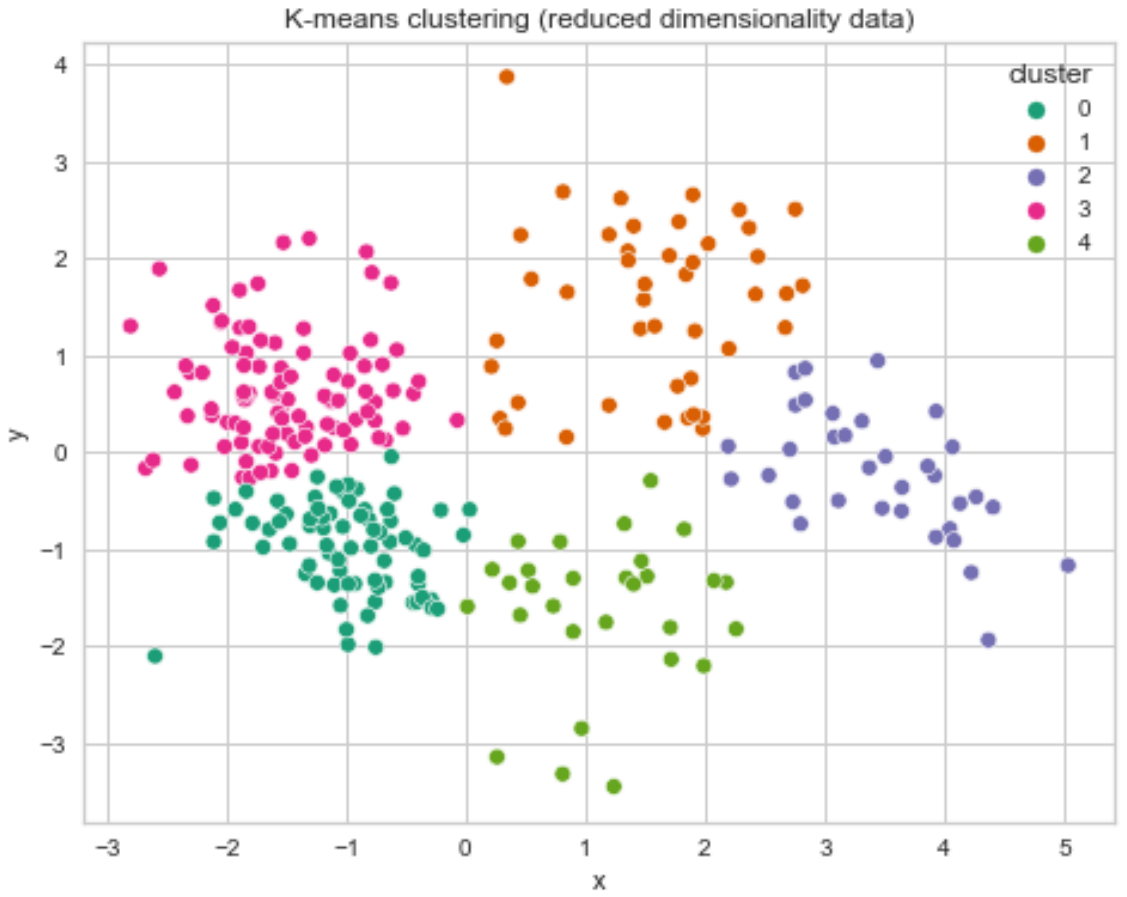


- Euklidinių atstumų nuo klasterio vidurkio taško kvadratų sumos (within cluster sum of squares, scikit-learn vadinama "distortion") alkūnės grafike nėra aiškkių linkio taškų.

- Vienas galimas variantas yra imti k=5.

Gauti klasteriai

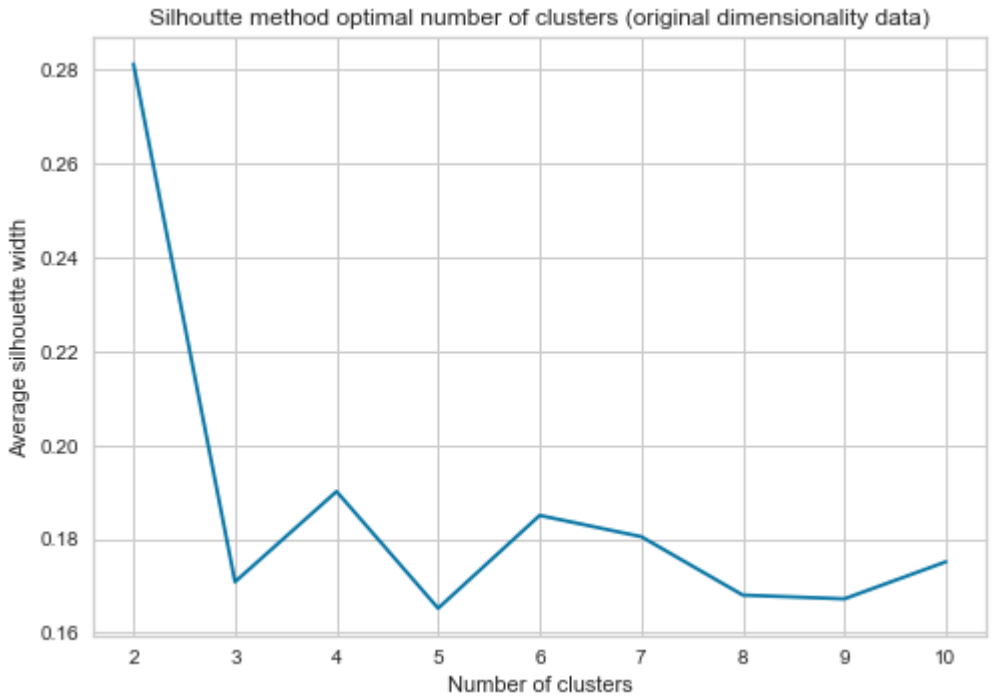


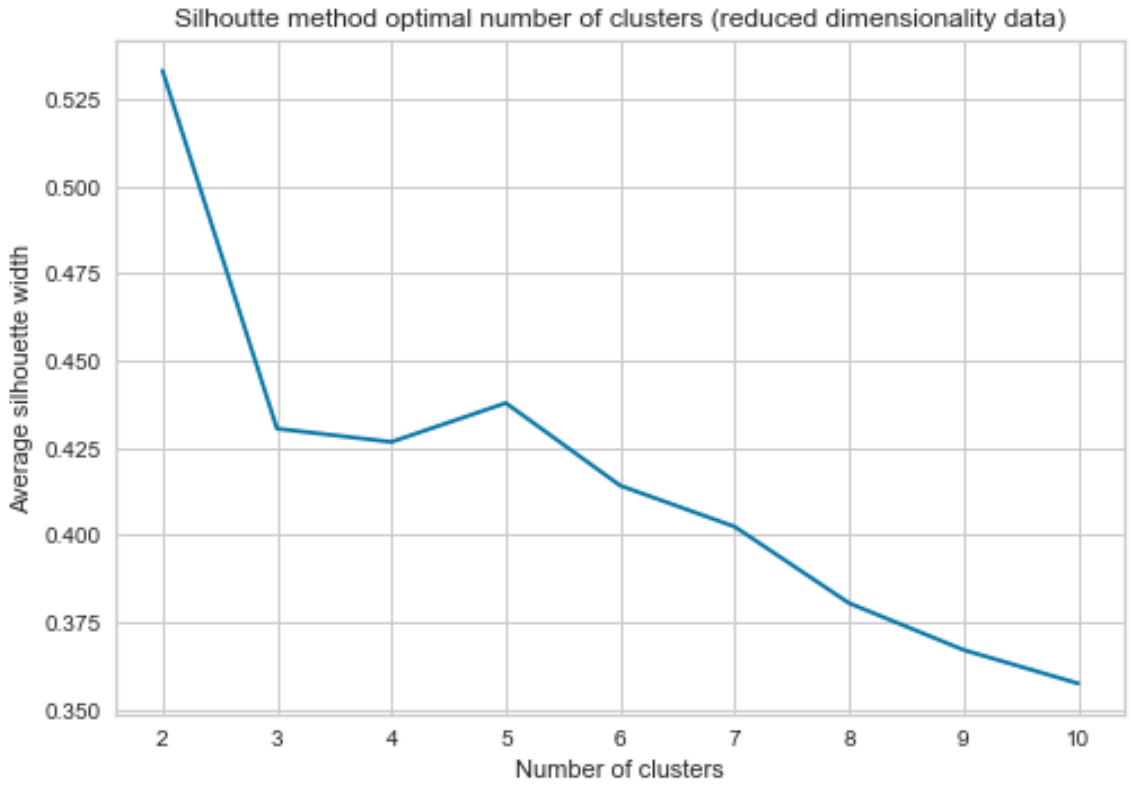


- Gauti klasteriai nestabilūs - rezultatai (dainai priskirtas klasteris) skiriasi prieš tai sumažinus duomenų dimensiją.

- Sprendimas yra parinkti kitą klasterių skaičių.

Silueto metodas

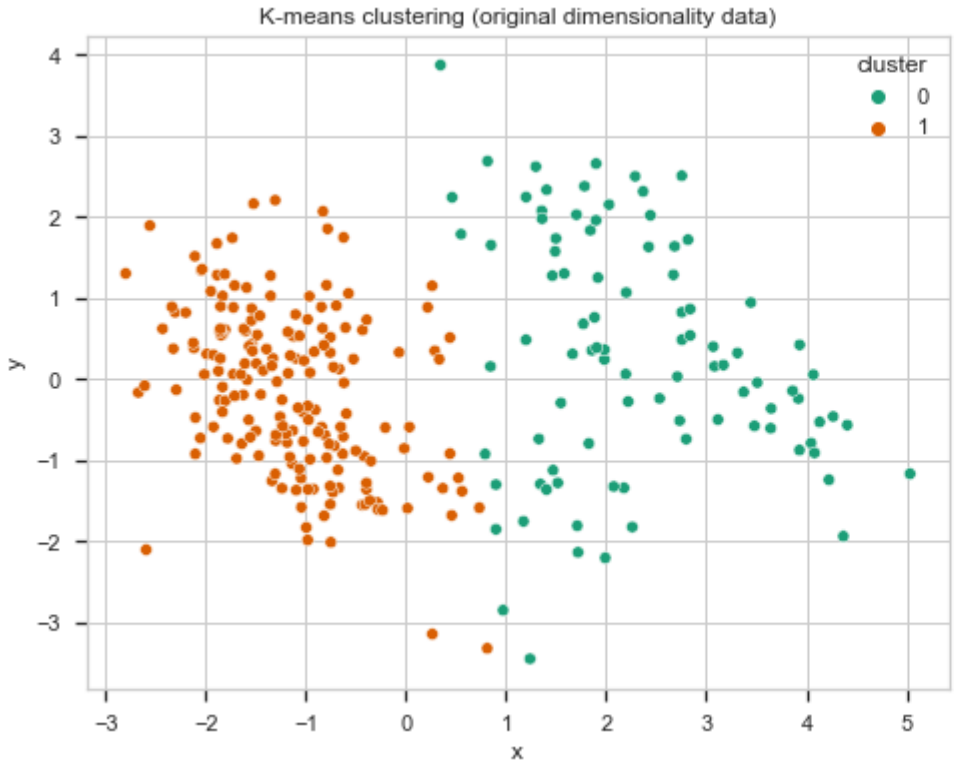


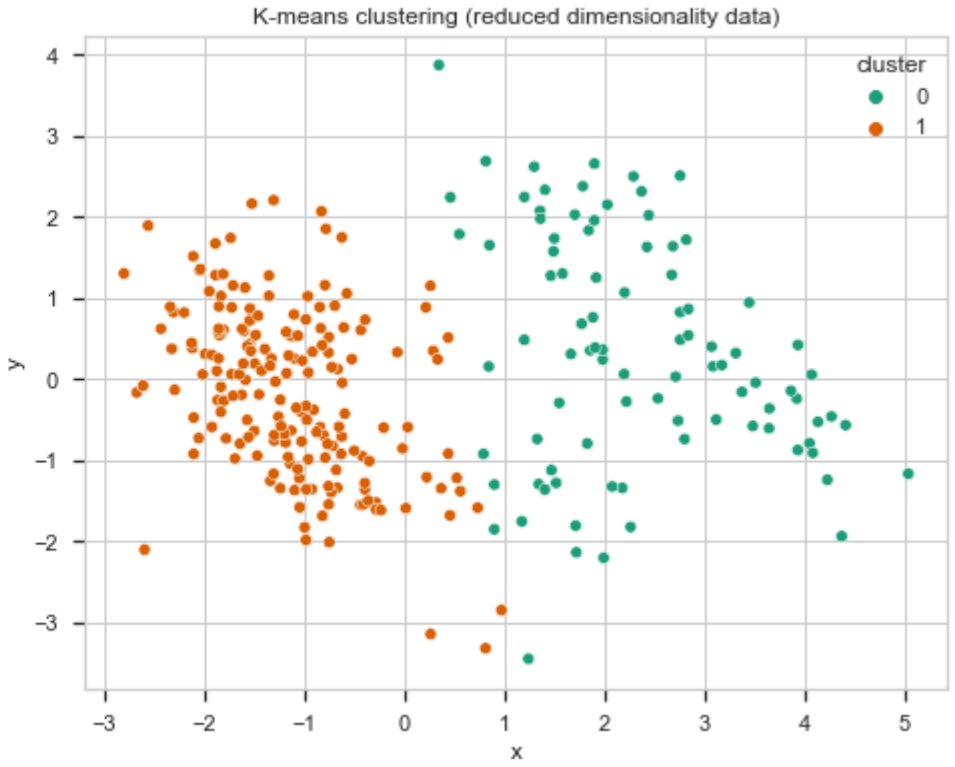


- Vidutinio silueto koeficiento metodu stipriai matomas optimalus klasterių skaičius – 2.

- Toks pat skaičius gaunamas ir empiriniu metodu k ≈ $\sqrt{\frac{n}{2}}$ ≈ 2.23, nes šiuo atveju turima n=10.

Gauti klasteriai

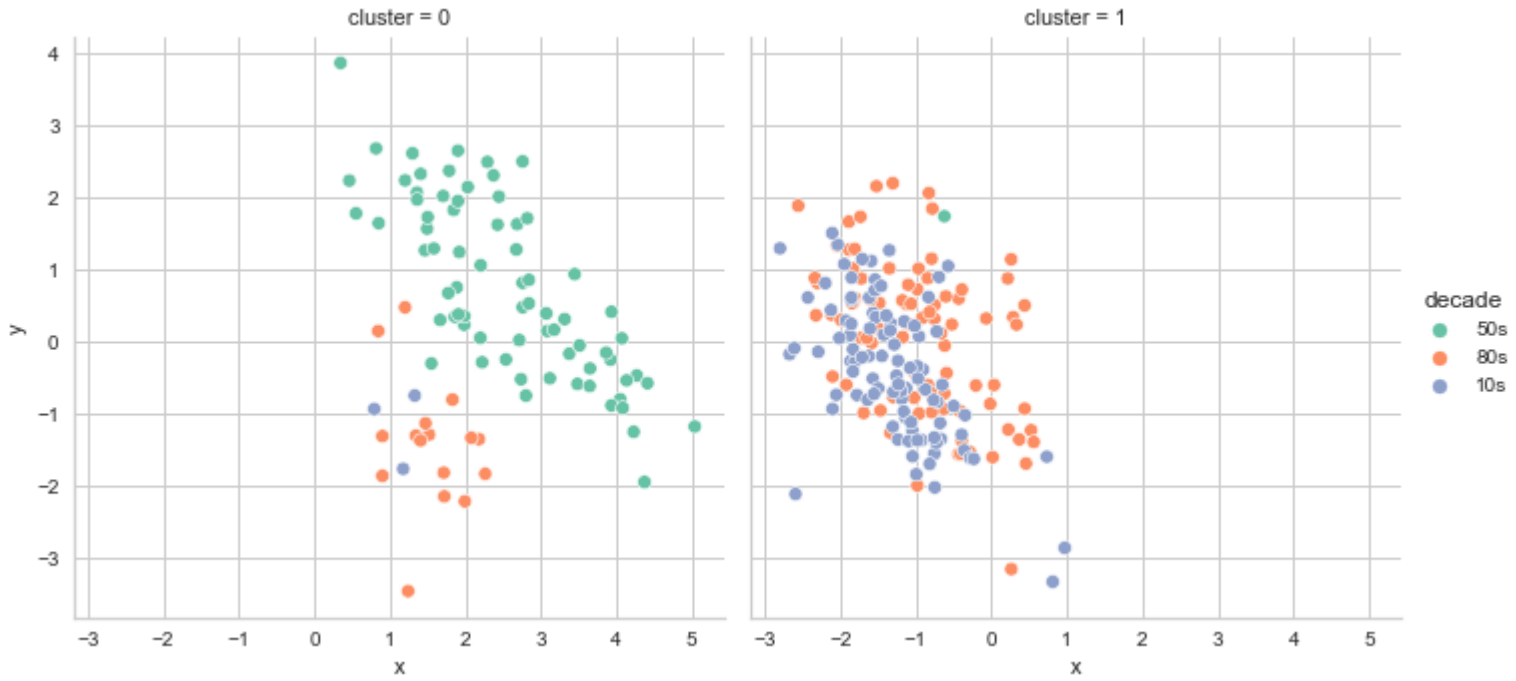




- Klasteriai stabilūs - gaunami tokie patys rezultatai prieš tai sumažinus duomenų dimensiją ir to nedarant.

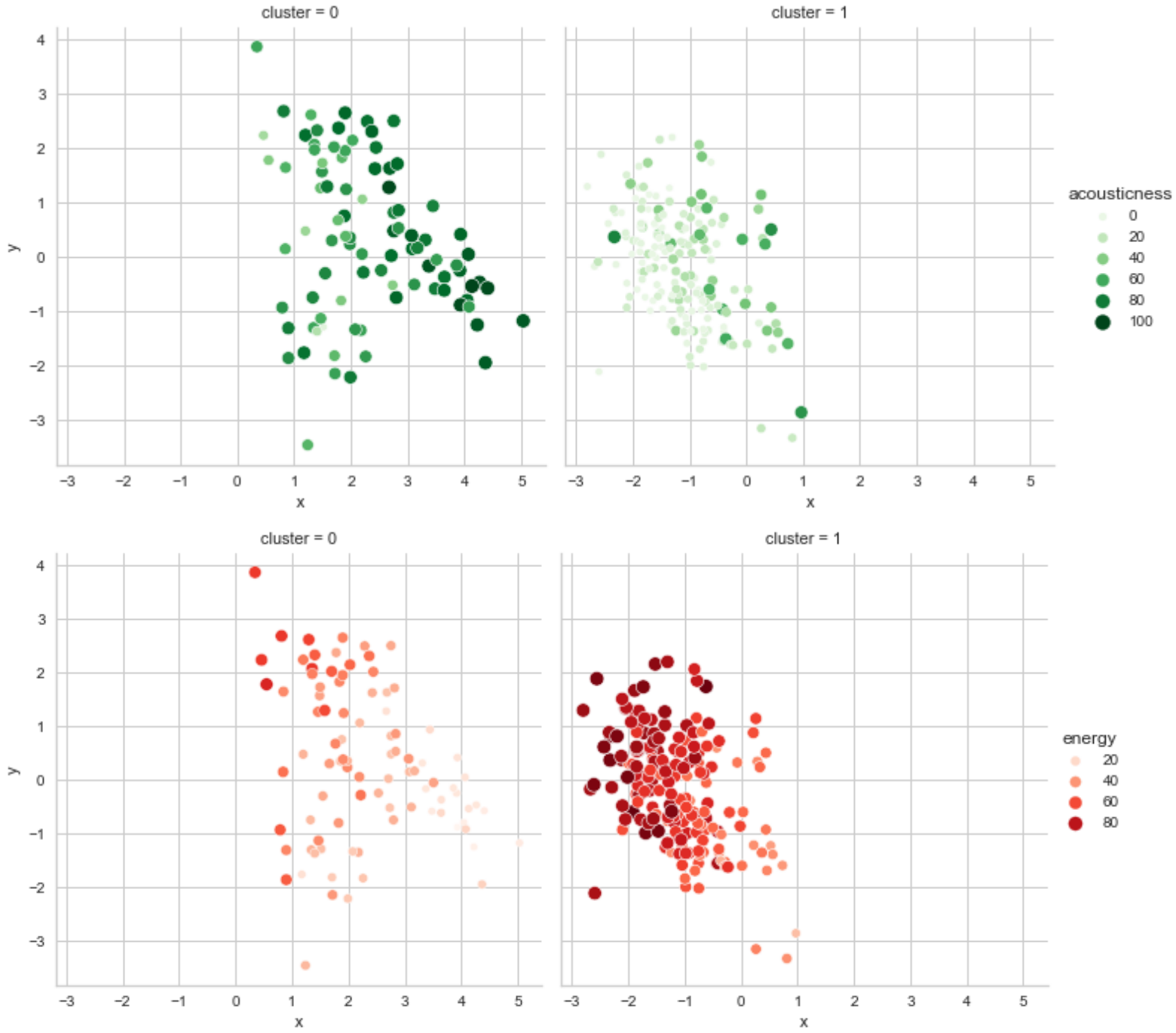
- Be to, vizualizavus sumažintos dimensijos erdvėje nėra daug duomenų taškų, kuriems "iš akies" priskirtas ne tas klasteris.

Klasteriuose matomos ryškios tendencijos pagal dešimtmetį: Vienam klasteriui priklauso daugiausia 50-ųjų dainos, tuo tarpu kitam – 80-ųjų ir 2010-ųjų.



Kita ryški tendencija klasteriuose:

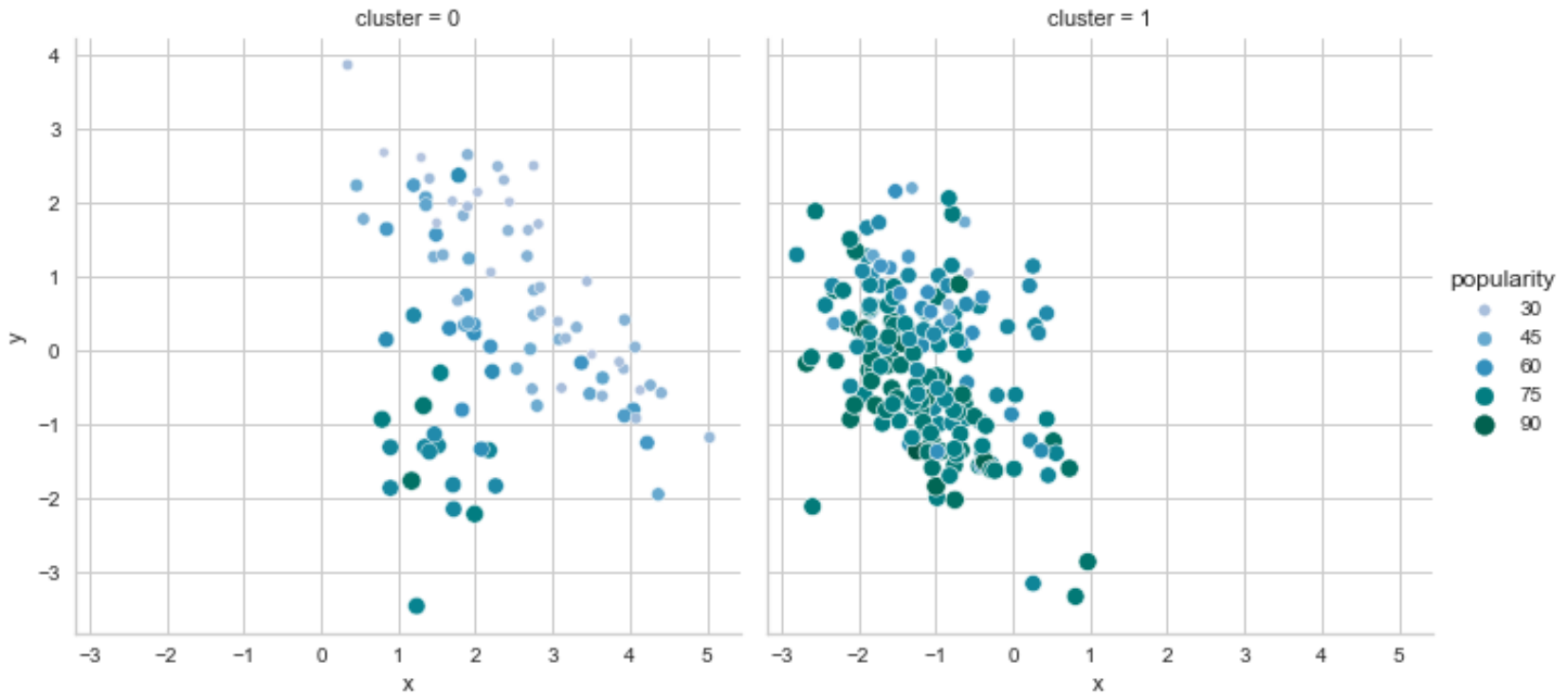
Akustiškumo ir energijos dichotomija: Vienam klasteryje aukštos vieno, kitam – kito po požymio reikšmės.



Antrajam klasteriui taip pat priklauso ir didesnio populiarumo dainos.

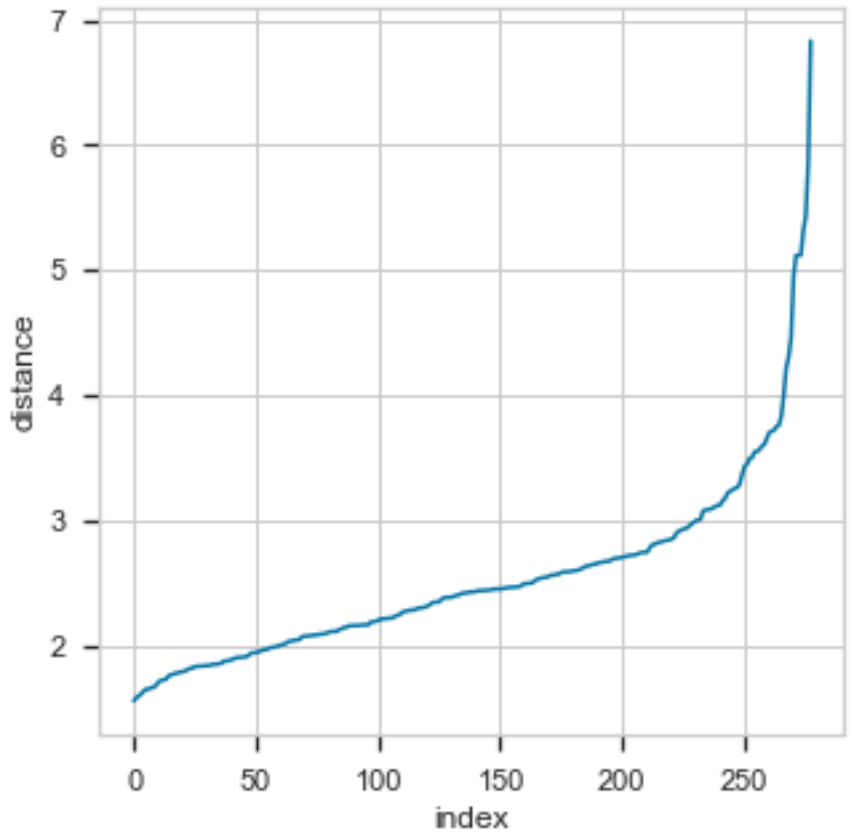
Kitos (ne tokios ryškios tendencijos):

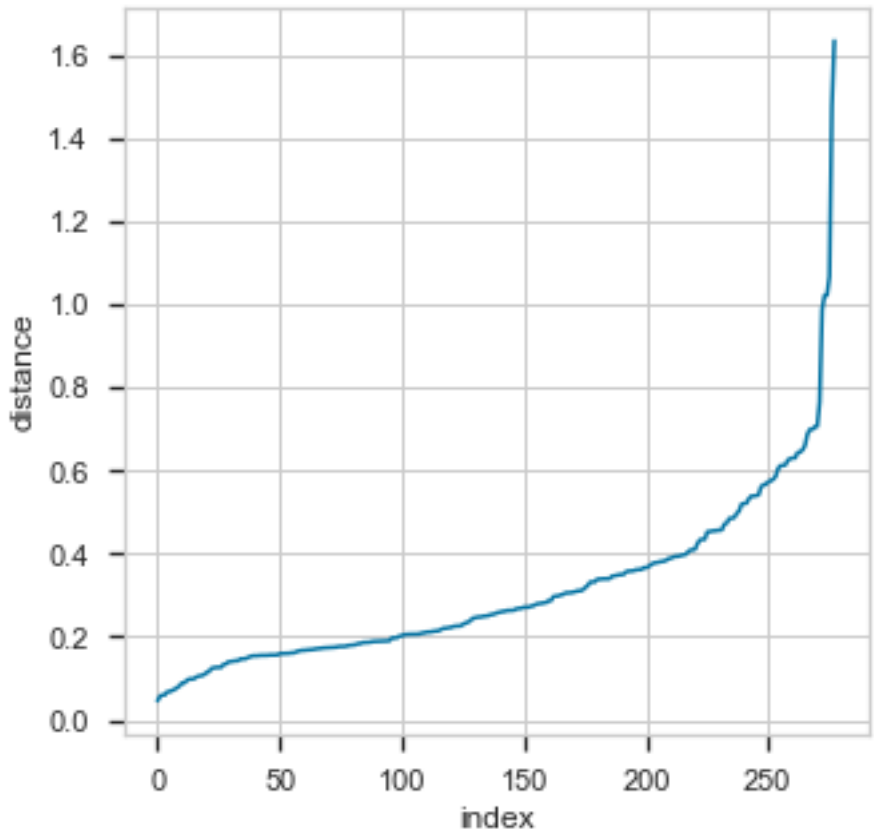
Antrajame klasteryje dainos vidutiniškai ilgesnės, garsesnės, labiau tinkamos šokti, greitesnio tempo.



DBSCAN metodas

Eps paieška pagal kelio (knee) tašką

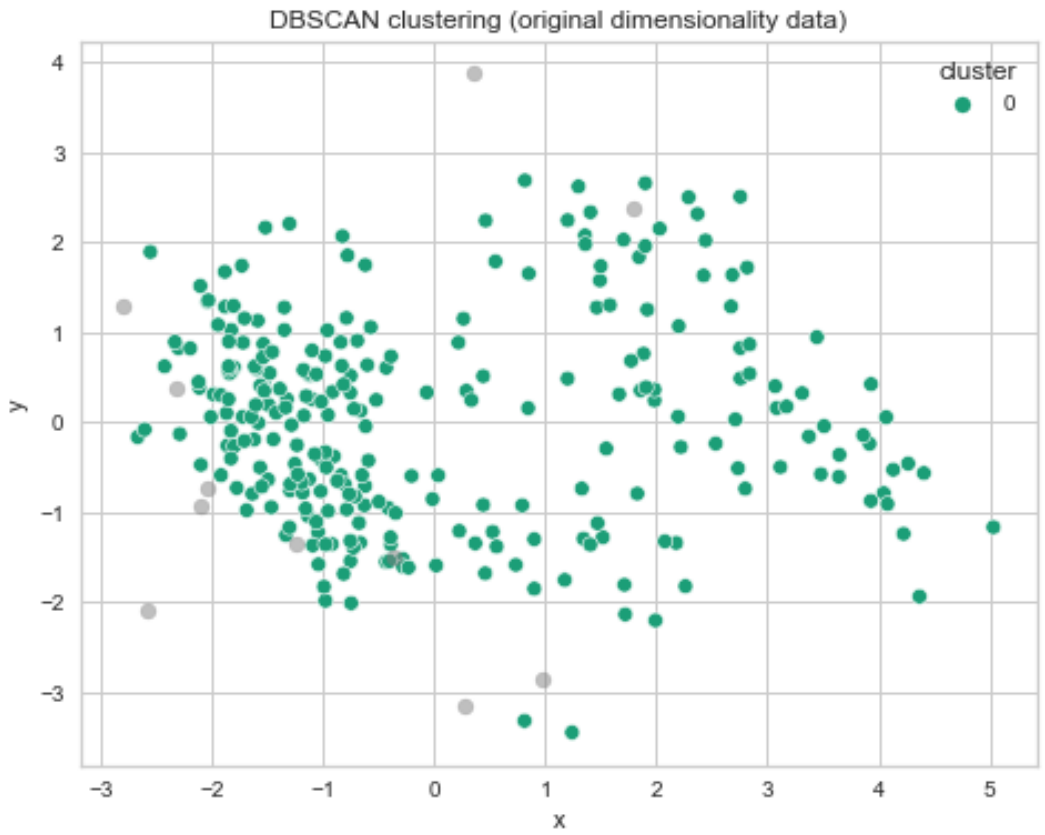


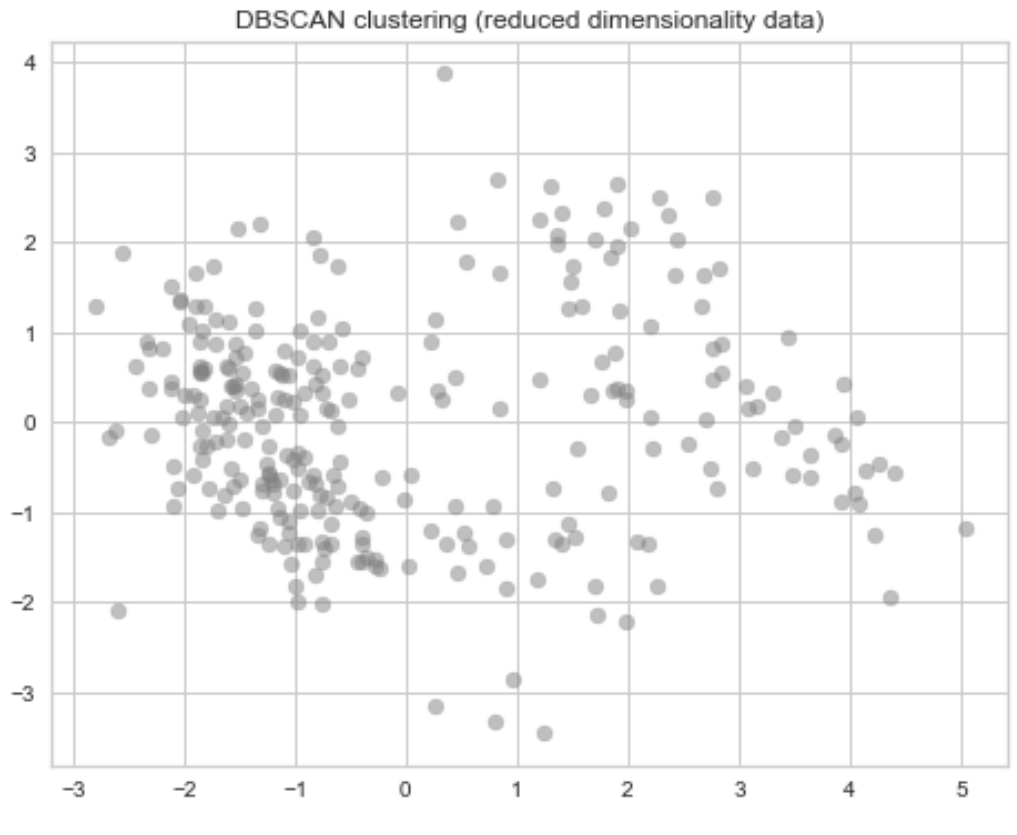


- $MinPts$ parinktas naudojantis nykščio taisykle imant $MinPts = 2n$, kur n – požymių skaičius duomenų aibėje.

- Naudojant prieš tai pavaizduotą parametro parinkimo metodą, eps reikšmės originalios ir sumažintos dimensijos duomenims gautos atitinkamai 3.5 ir 0.7.

Gauti klasteriai



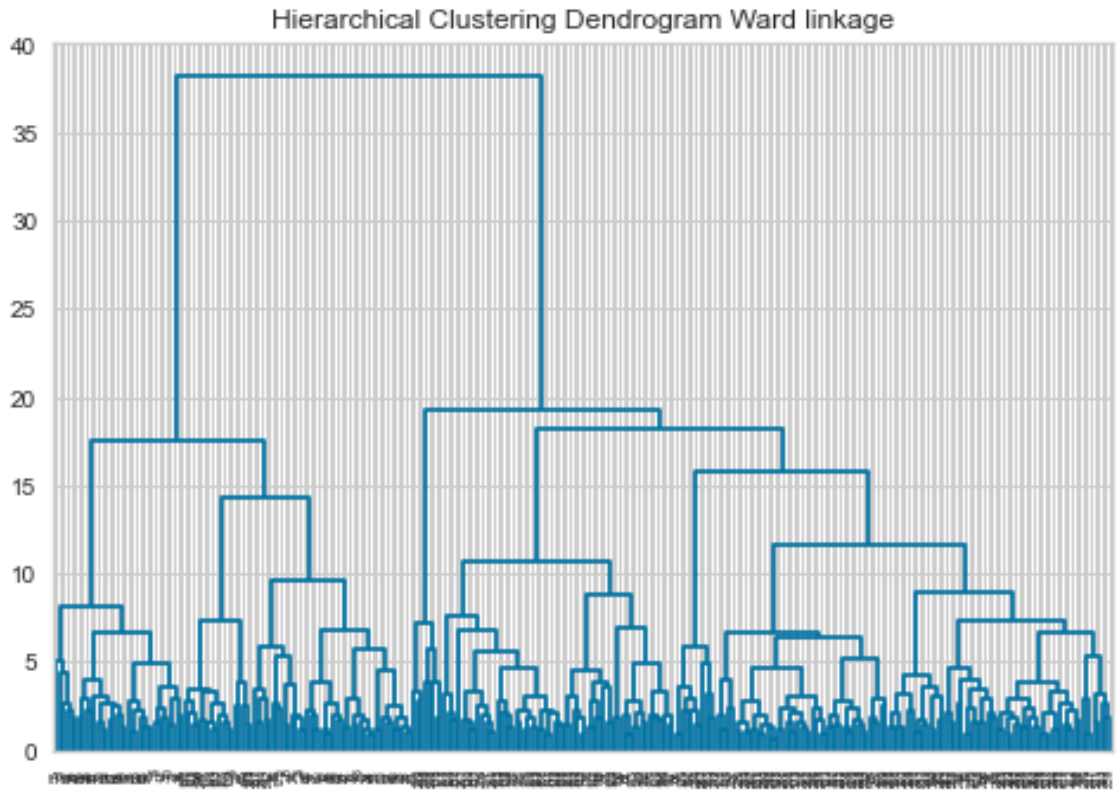


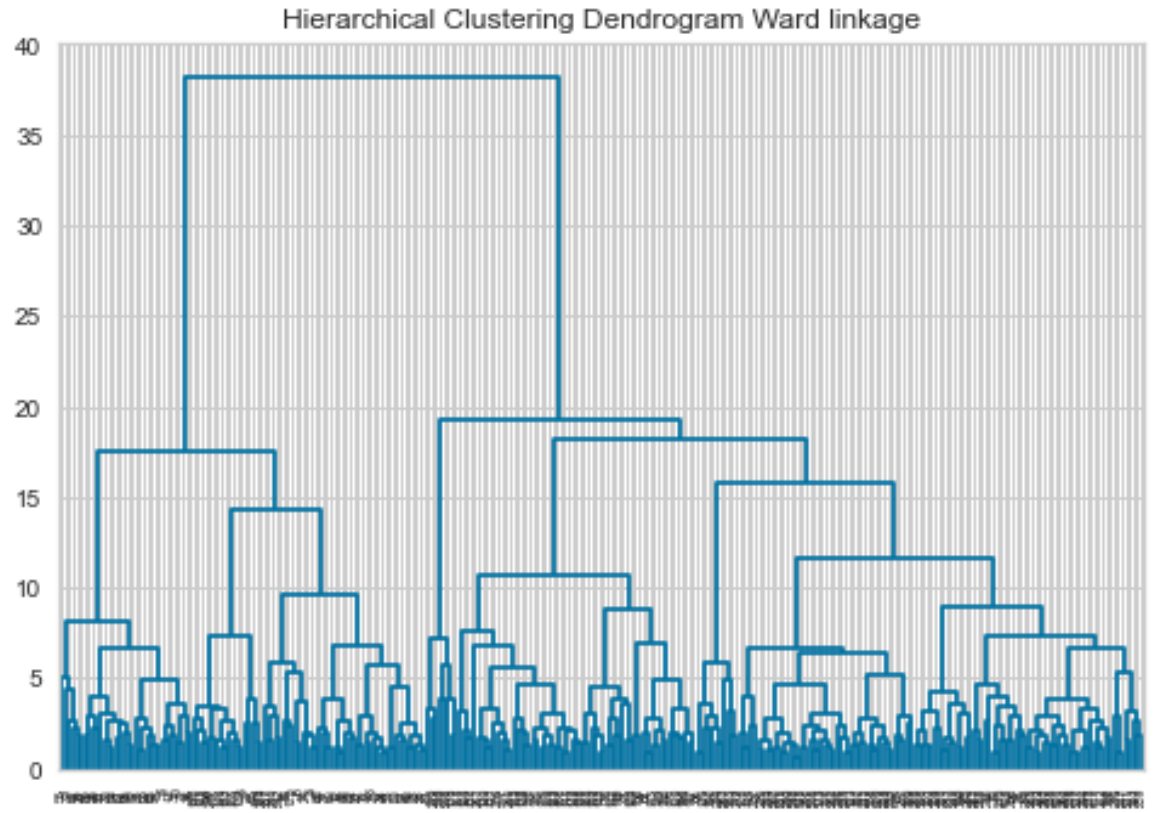
- Gauti nestabilūs klasteriai, negana to originalios dimensijos duomenyse beveik visi taškai priskirti vienam klasteriui, sumažintos dimensijos duomenyse visi taškai laikyti triukšmo taškais.

- Dėl šių priežasčių laikyta, kad DBSCAN metodu naudingų įžvalgų turimam duomenų rinkiniui negauta.

Hierarchinio klasterizavimo metodas

Klasterių skaičiaus parinkimas naudojantis dendrograma

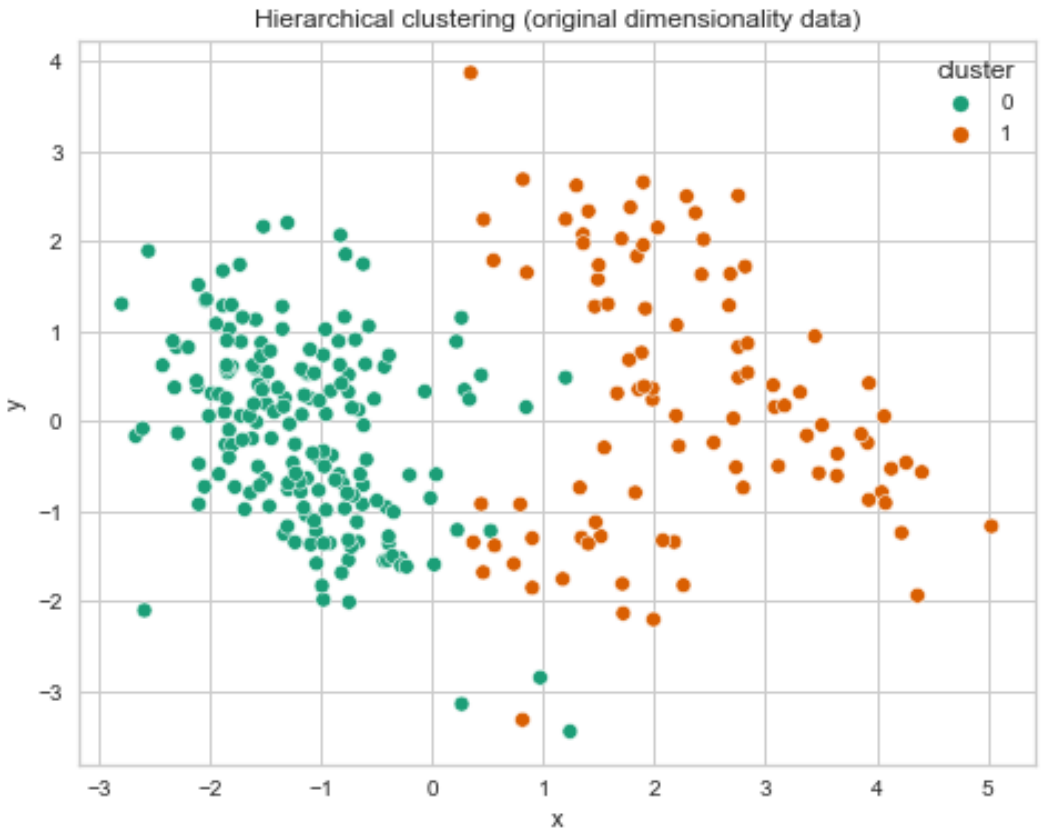


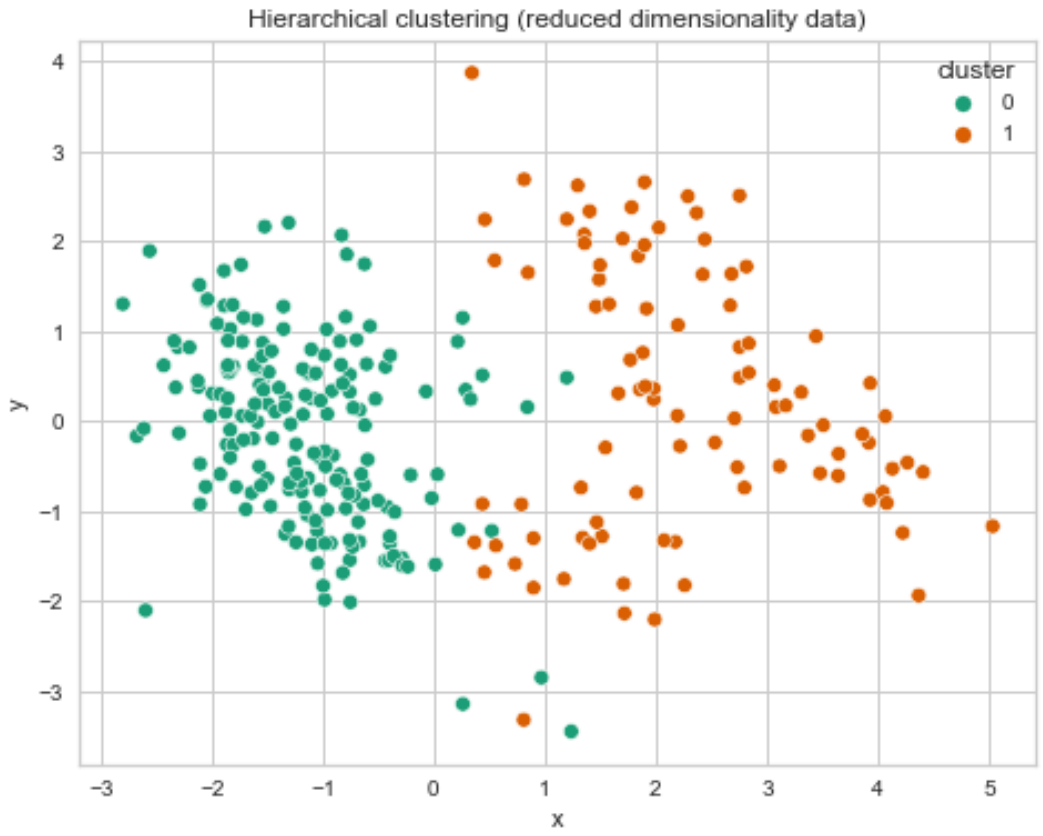


- Geriausi rezultatai (jungiamų klasterių dydžių subalansavimo prasme) gauti naudojant Ward jungimo matą.

- Tiek originalios, tiek sumažintos erdvės duomenyse naudojant Ward jungimo matą pagal dendrogramą pasirinkta duomenų aibę dalinti į 2 dalis (sudaryti 2 klasterius).

Gauti klasteriai





- Gauti klasteriai stabilūs.

- Klasteriai tik minimaliai skiriasi nuo klasterių, gautų naudojant k-means metodą pasirinkus naudoti 2 klasterius, todėl galioja tos pačios požymių tendencijos.

Išvados

Naudojant klasterizavimą k-means ir hierarchinio klasterizavimo metodais, gautas optimalus klasterių skaičius k=2. Be to, abu metodai beveik visus taškus priskiria tam pačiam klasteriui.

Galima gautų rezultatų interpretacija:

Duomenų aibėje egzistuoja 2 pagrindiniai dainų tipai.

- Pirmajam priklauso didesnio akustiškumo, ramesnės dainos (iš tirtų dešimtmečių šio tipo dainos buvo sukuriamos daugiausiai 50-aisiais),
- Antrajam priklauso energiškesnės, garsesnės, labiau tinkamos šokti dainos (šios dainos buvo sukuriamos 80-aisiais ir 2010-aisias).
- Spotify vartotojai dažniau klausosi antrojo dainų tipo dainų.