Vilniaus Universitetas

Dimensijos mažinimas

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

Vilnius, 2022

**Turinys**

[1 Tikslas ir uždaviniai 3](#_Toc97915342)

[2 Duomenų aibė 4](#_Toc97915343)

[3 Atliktos analizės aprašymas 5](#_Toc97915344)

[3.1 Aprašomoji statistika 5](#_Toc97915345)

[3.2 Duomenų normavimas 9](#_Toc97915346)

[3.3 PCA 10](#_Toc97915347)

[3.4 MDS 11](#_Toc97915348)

[3.5 t-SNE 12](#_Toc97915349)

[3.6 Metodų palyginimas 14](#_Toc97915350)

[4 Išvados 15](#_Toc97915351)

[Priedas 16](#_Toc97915352)

# Tikslas ir uždaviniai

Tikslas:

Panaudoti dimensijos mažinimo metodus daugiamačių duomenų vizualizavimui, ištirti metodų galimybes bei pateikti pasirinktos aibės vizualizavimo rezultatus ir gautų rezultatų interpretaciją

Uždaviniai:

Pateikti pasirinktos aibės aprašomąją statistiką, aprašyti duomenų aibės specifiką.

Sunormuoti duomenų aibę pagal vidurkį ir dispersiją.

Naudojant tris dimensijos mažinimo metodus sumažinti duomenų aibės dimensiją iki dim=2.

Vizualizuoti dimensijos mažinimo rezultatus ir ištirti, kaip keičiasi vizualizavimo rezultatai, keičiant algoritmų parametrus.

Įvertinki gautus rezultatus ir padaryti išvadas, kuris metodas geriau atvaizduoja rezultatą.

Įvardinti tirtų dimensijos mažinimo metodų privalumus ir trūkumus.

# Duomenų aibė

Spotify Past Decades Songs duomenų aibė

Duomenų aibės šaltinis: Kaggle

Nuoroda per internetą: <https://www.kaggle.com/cnic92/spotify-past-decades-songs-50s10s?select=1990.csv>

Duomenų aibę sudaro tokie požymiai:

„Number“ – (kategorinis, nominalusis) dainą identifikuojantis kodas

„Title“ – (kategorinis, nominalusis) dainos pavadinimas

„Artist“ – (kategorinis, nominalusis) atlikėjas arba grupė

„Top Genre“ – (kategorinis, nominalusis) dainos žanras

„Year“ – (kiekybinis, diskretusis, intervalinė skalė) išleidimo metai

„Decade“ – (kiekybinis, diskretusis, intervalinė skalė) išleidimo dešimtmetis

„BPM“ – (kiekybinis, tolydus, santykių skalė) dainos tempas

“Loudness (dB)“ - (kiekybinis, tolydus, intervalų skalė) dainos garsumas

„Duration“ – (kiekybinis, tolydus, santykių skalė) dainos trukmė

„Energy“ – (kiekybinis, tolydus, santykių skalė) dainos energija

„Danceability“ – (kiekybinis, tolydus, santykių skalė) lengvumas šokti pagal dainą

„Liveness“ – (kiekybinis, tolydus, santykių skalė) kaip tikėtina, kad daina yra gyvas įrašas

„Valence“ – (kiekybinis, tolydus, santykių skalė) dainos pozityvumas

„Acousticness“ – (kiekybinis, tolydus, santykių skalė) dainos akustiškumas

„Speechiness“ – (kiekybinis, tolydus, santykių skalė) kiek dainoje yra kalbama

„Popularity“ - (kiekybinis, tolydus, santykių skalė) dainos populiarumas pagal perklausų skaičių

# Atliktos analizės aprašymas

## Aprašomoji statistika

Duomenų aibę sudaro trys klasės pagal dainos išleidimo dešimtmetį (požymis „Decade“). Kiekvienai klasei priklausančių objektų kiekis pateiktas lentelėje (žr. 1 lentelė).

1 lentelė Objektų kiekis duomenų aibėje pagal dešimtmetį

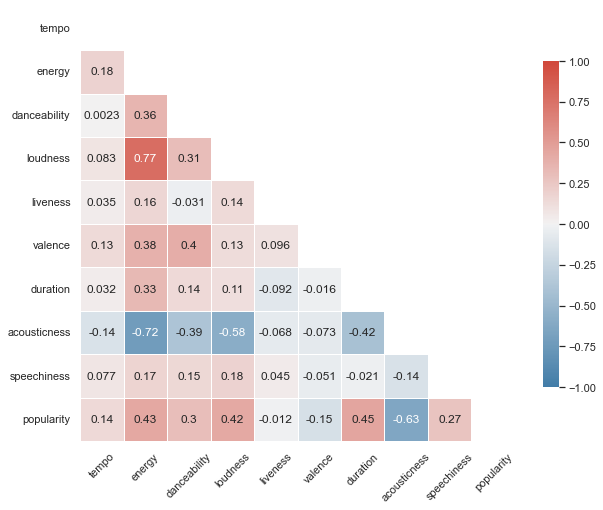
|  |  |
| --- | --- |
| Dešimtmetis | Skaičius |
| 2010-ieji | 100 |
| 1980-ieji | 705 |
| 1950-ieji | 73 |

Duomenų aibės skaitiniams požymiams apskaičiuotos pagrindinės aprašomosios statistikos charakteristikos (standartinis nuokrypis, vidurkis, mediana, mažiausia reikšmė (min), didžiausia reikšmė (max), pirmas ir trečias kvartiliai). Rezultatai pateikti lentelėje (žr. 2 lentelė). Papildomai pateiktos aprašomosios statistikos pagal dešimtmetį (žr. 1 priedas) Visi skaitiniai požymiai išskyrus „Tempo“, „Loudness“ ir „Duration“ matuojami skalėje nuo 0 iki 100.

2 lentelė Aprašomosios statistikos charakteristikos duomenų aibei

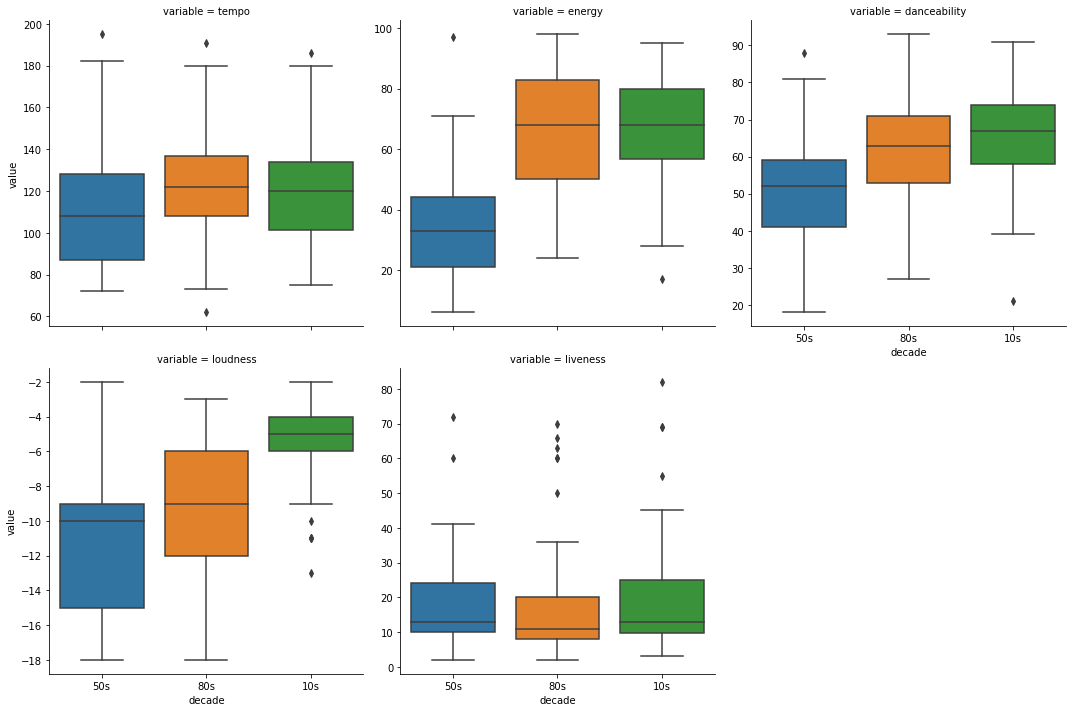
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | std | min | 25% | 50% | 75% | max |
| tempo | 118.2 | 25.3 | 62.0 | 100.0 | 117.0 | 135.0 | 195.0 |
| energy | 58.1 | 22.8 | 6.0 | 41.2 | 60.5 | 78.8 | 98.0 |
| danceability | 60.4 | 14.3 | 18.0 | 51.0 | 62.0 | 71.0 | 93.0 |
| loudness | -8.5 | 4.0 | -18.0 | -11.0 | -8.0 | -5.0 | -2.0 |
| liveness | 17.4 | 13.4 | 2.0 | 9.0 | 13.0 | 22.0 | 82.0 |
| valence | 55.6 | 25.0 | 9.0 | 34.0 | 55.0 | 77.8 | 99.0 |
| duration | 212.3 | 56.5 | 98.0 | 174.2 | 210.0 | 245.0 | 433.0 |
| acousticness | 33.7 | 31.1 | 0.0 | 7.0 | 20.5 | 61.0 | 100.0 |
| speechiness | 5.8 | 5.4 | 2.0 | 3.0 | 4.0 | 6.0 | 46.0 |
| popularity | 63.6 | 16.6 | 26.0 | 54.0 | 68.0 | 76.0 | 94.0 |

Tarp skaitinių rodiklių apskaičiuotos Pirsono koreliacijos koeficientų reikšmės (angl. Pearson correlation coefficient). Rasta stipri teigiama koreliacija tarp dainos garsumo (požymis „Loudness“) ir energijos („Energy“) (r = 0.77). Dainos akustiškumas (požymis „Acousticness“) neigiamai koreliuoja su beveik visais kitais požymiais. Iš jų didžiausia neigiama koreliacija su požymiais „Energy“ (r=-0.72) „Popularity“ (r=-0.63) ir „Loudness“ (r=-0.58). Rezultatai pateikti koreliacijų diagrama (žr. 2 pav.)

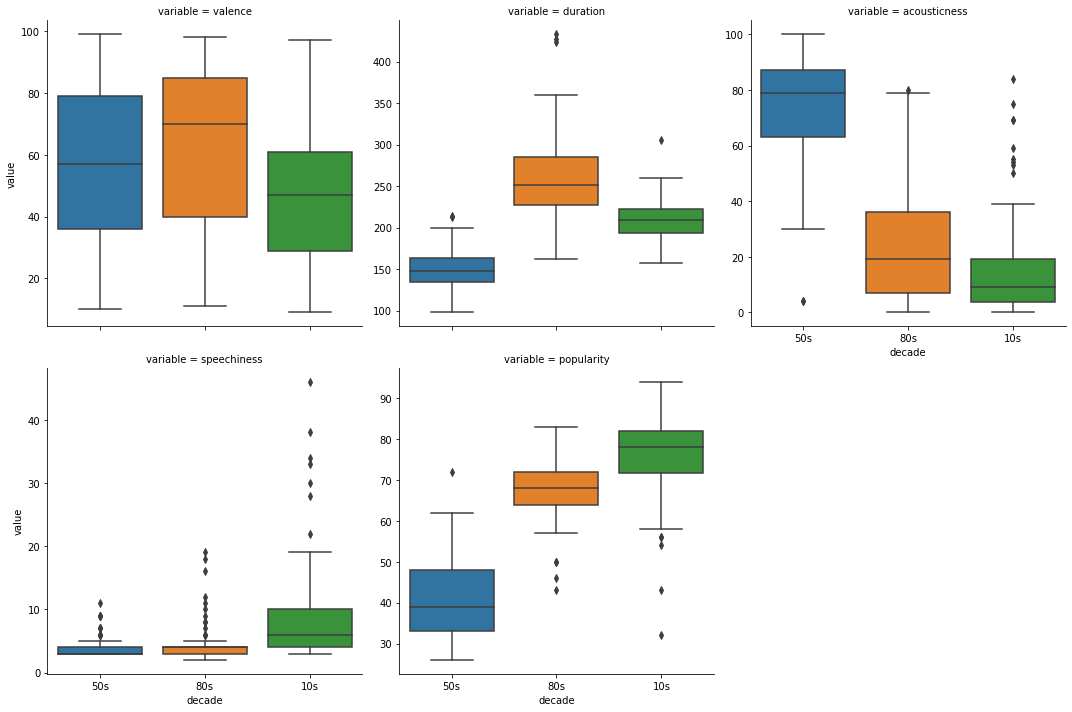


1 pav. Pirsono koreliacijos tarp požymių koeficientai

Kiekvieno skaitinio duomenų aibės požymio pasiskirstymas pagal dešimtmetį pavaizduotas stačiakampėmis diagramomis (žr. 3 pav. ir 4 pav.). Pastebėtas kad vėlesni dešimtmečiai pasižymi didėjančiomis garsumo, tinkamumo šokti, perklausų skaičiaus reikšmėmis ir mažėjančiomis dainų akustiškumo reikšmėmis.



2 pav. Skaitinių požymių stačiakampės diagramos pagal dešimtmetį



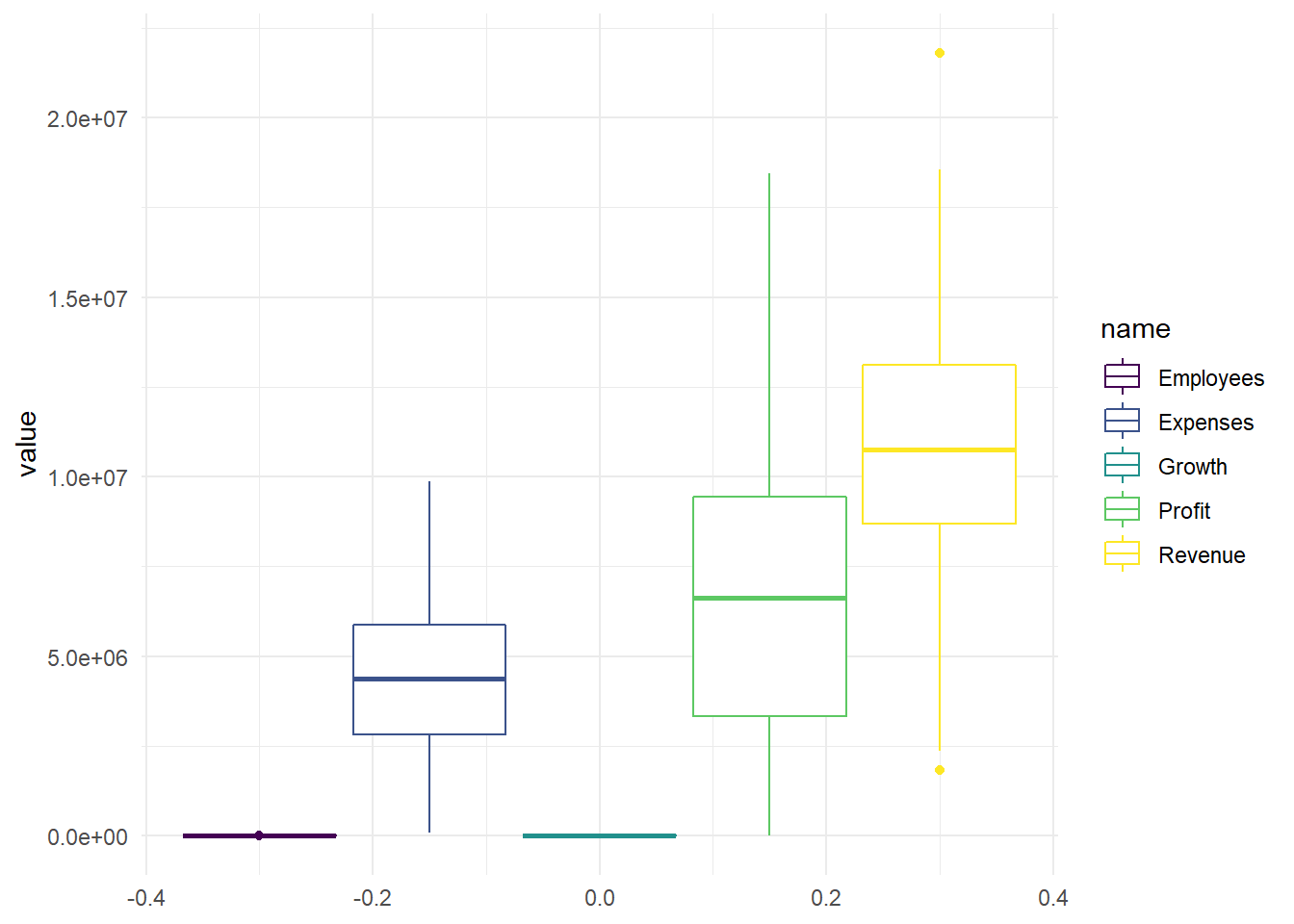
3 pav. Skaitinių požymių stačiakampės diagramos pagal dešimtmetį (2 dalis)

## Duomenų normavimas

Kadangi požymiai „Duration“, „Loudness“ ir „Tempo“ matuojami kitokio dydžio skalėse negu likusieji skaitiniai duomenų aibės požymiai (žr. Aprašomoji statistika), duomenys sunormuoti naudojant normavimą pagal vidurkį ir dispersiją (standartizavimas) , kur - požymio vidurkis požymio dispersija,   – požymio vidurkis.

Pradinis kiekybinių duomenų aibės požymių pasiskirstymas pavaizduotas stačiakampe diagrama (žr. 2 pav.). Pakartotinai pavaizduotas pasiskirstymas atlikus abu anksčiau minėtus normavimo metodus (žr. **Klaida! Nerastas nuorodos šaltinis.** ir **Klaida! Nerastas nuorodos šaltinis.**).

Dėl didelio kiekio išskirčių (žr. **Klaida! Nerastas nuorodos šaltinis.**) nerekomenduojama taikyti standartizavimo metodą darbuotojų skaičiaus įmonėje požymiui.



4 pav. Kiekybinių požymių stačiakampė diagrama prieš atliekant normavimą

## PCA

Pagrindinių komponenčių analizė (angl. Principal Component Analysis, toliau - PCA)

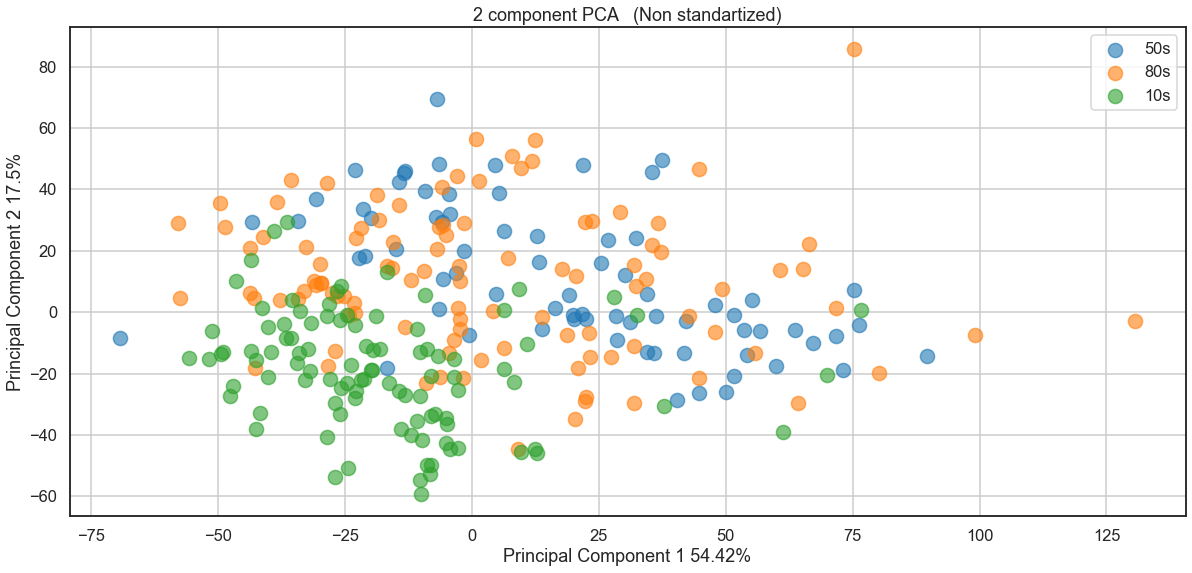
Įvertinti sąryšius tarp taškų plokštumoje (ar persidengia klasės).

Preserves local structures

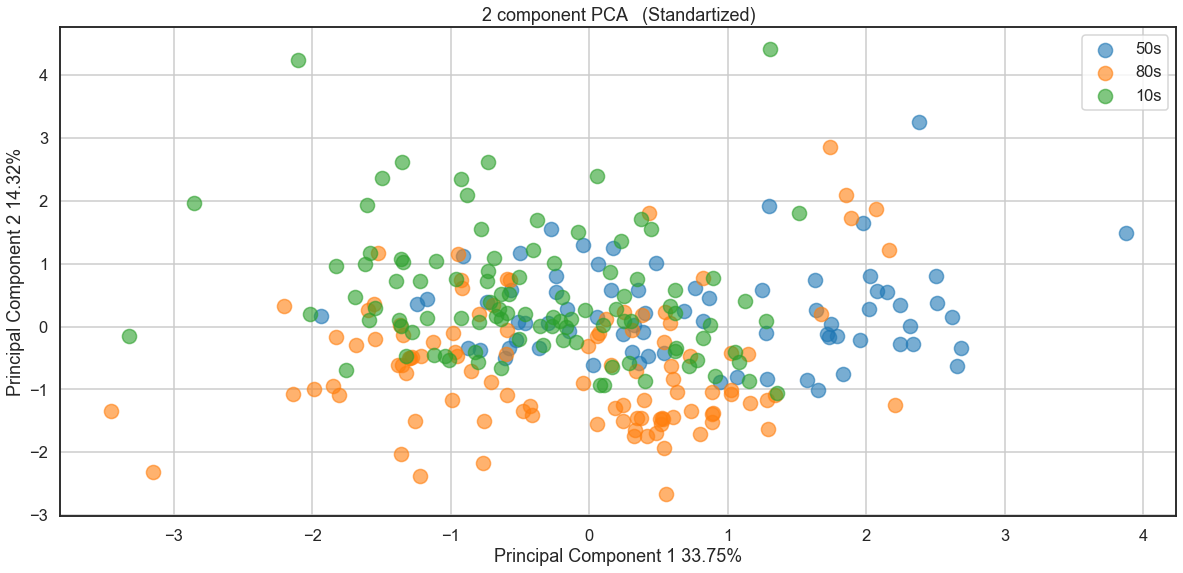
Only for numerical data

Stipriai veikiama išskirčių

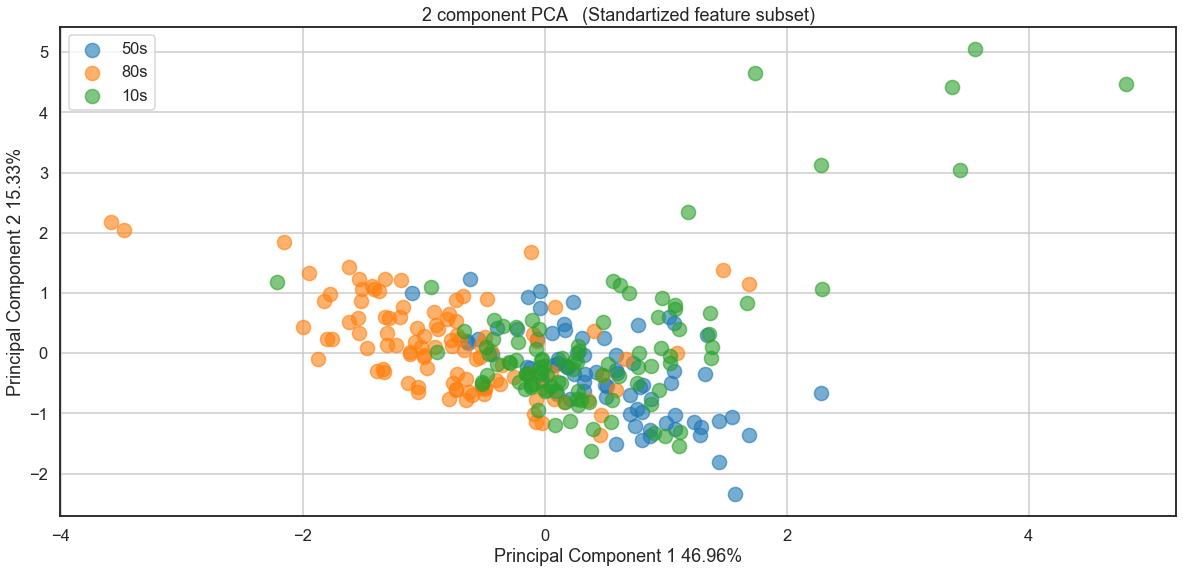
Palyginimui atlikta paėmus požymių rinkinį be požymių kurie pagal stačiakampes diagramas dešimtmečius atskiria mažiausiai (žr. )



Pav. 5 PCA su nestandartizuotais duomenimis



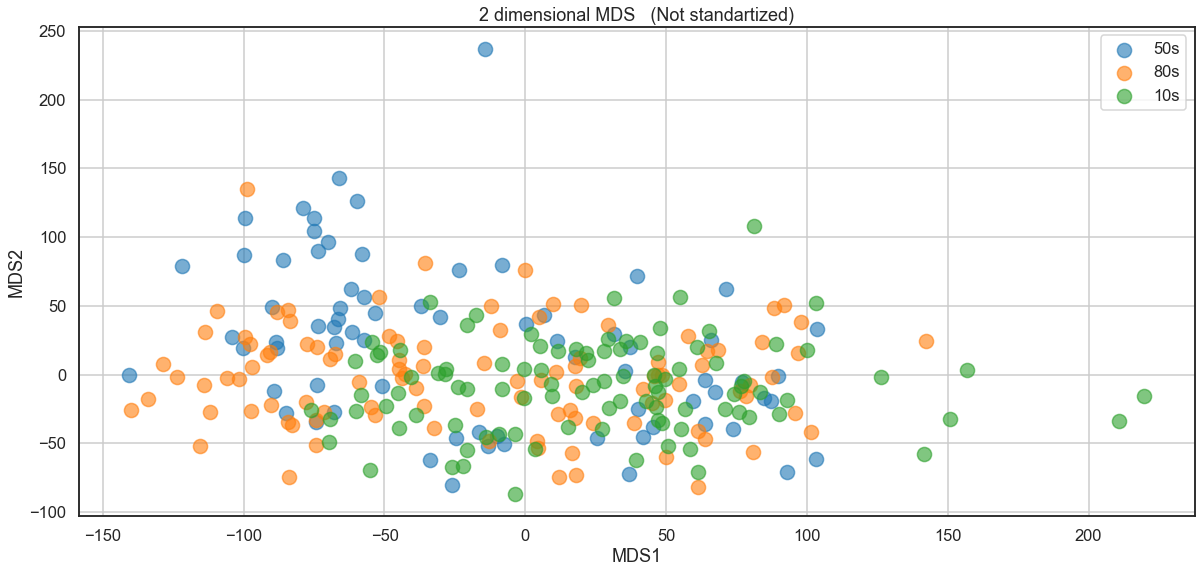
Pav. 6 PCA su standartizuotais duomenimis



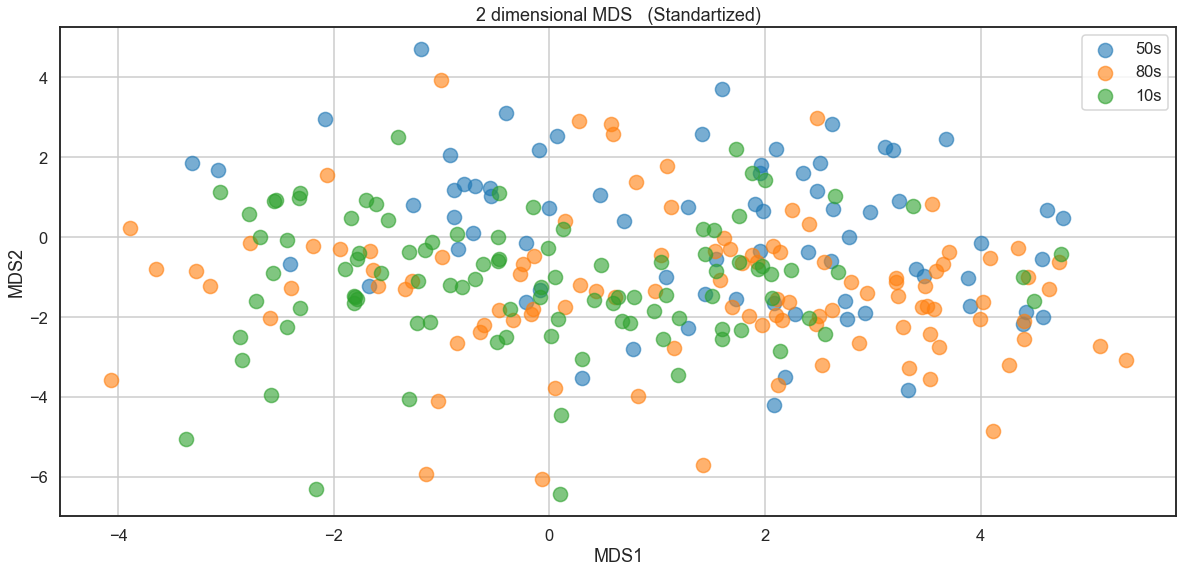
Pav. 7 PCA su standartizuotais duomenimis

## MDS

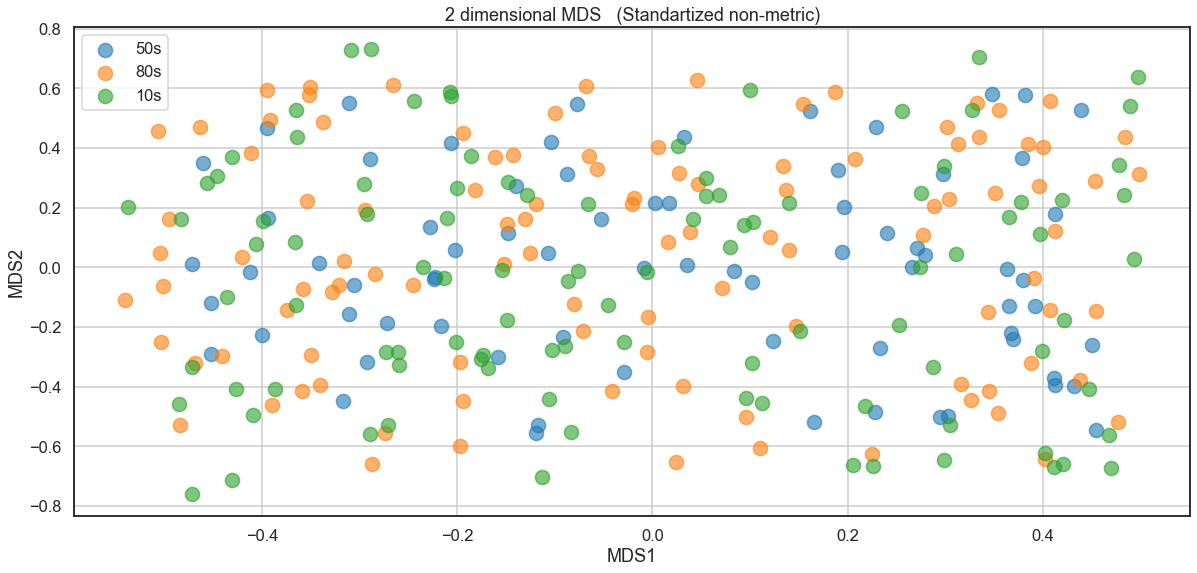
Daugiamatės skalės (angl. Multidimensional Scaling, toliau – MDS) yra netiesinis dimensijos mažinimo metodas. MDS ieškoma daugiamačių duomenų projekcijų mažesnės dimensijos erdvėje, siekiant išlaikyti atstumus tarp objektų.



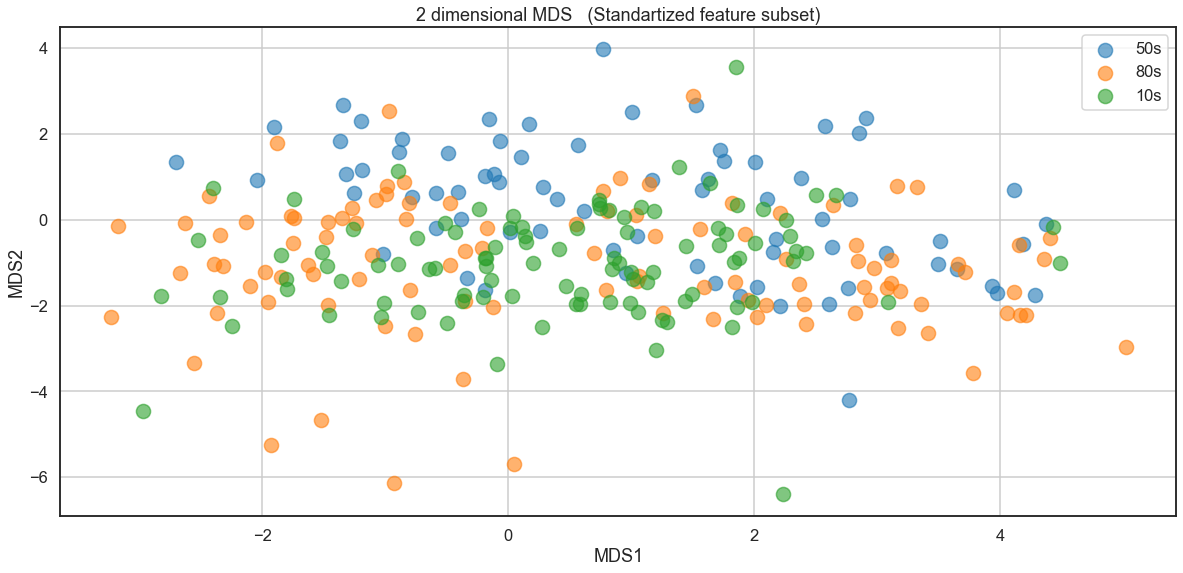
Pav. 8 MDS su nestandartizuotais duomenimis



Pav. 9 MDS su standartizuotais duomenimis



Pav. 10 Nemetrikinės MDS su standartizuotais duomenimis



Pav. 11 MDS su standartizuotais duomenis naudojant požymių poaibį

## t-SNE

Siekia išsaugoti kiekvieno objekto kaimynus. Orentuotas į vidinės struktūros išsaugojimą.

t-SNE works very well on large datasets but it also has it’s limitations, such as loss of large-scale information, slow computation time, and inability to meaningfully represent very large datasets

No crowding??s

Preserves local neighbors

Potential loss of global structures

Low preplexity -> artitrary small clusters start forming

Noise may look like a pattern

Recovers well separated clusters but may reate non existant clusters

The bottom line, however, is that you cannot see relative sizes of clusters in a t-SNE plot.

There may not be one perplexity value that will capture distances across all clusters—and sadly perplexity is a global parameter.

The basic message is that distances between well-separated clusters in a t-SNE plot may mean nothing.

There’s a reason that t-SNE has become so popular: it’s incredibly flexible, and can often find structure where other dimensionality-reduction algorithms cannot. Unfortunately, that very flexibility makes it tricky to interpret.

Only for visualization

## Metodų palyginimas

# Išvados

# Priedas

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | decade | mean | std | min | 25% | 50% | 75% | max |
| tempo | 10s | 118.7 | 22.4 | 75.0 | 101.5 | 120.0 | 134.0 | 186.0 |
| tempo | 50s | 111.2 | 28.0 | 72.0 | 87.0 | 108.0 | 128.0 | 195.0 |
| tempo | 80s | 122.6 | 25.1 | 62.0 | 108.0 | 122.0 | 137.0 | 191.0 |
| energy | 10s | 68.0 | 16.3 | 17.0 | 56.8 | 68.0 | 80.0 | 95.0 |
| energy | 50s | 34.9 | 17.4 | 6.0 | 21.0 | 33.0 | 44.0 | 97.0 |
| energy | 80s | 64.9 | 20.2 | 24.0 | 50.0 | 68.0 | 83.0 | 98.0 |
| danceability | 10s | 65.4 | 11.9 | 21.0 | 58.0 | 67.0 | 74.0 | 91.0 |
| danceability | 50s | 51.0 | 14.5 | 18.0 | 41.0 | 52.0 | 59.0 | 88.0 |
| danceability | 80s | 62.3 | 13.4 | 27.0 | 53.0 | 63.0 | 71.0 | 93.0 |
| loudness | 10s | -5.5 | 2.0 | -13.0 | -6.0 | -5.0 | -4.0 | -2.0 |
| loudness | 50s | -11.6 | 3.4 | -18.0 | -15.0 | -10.0 | -9.0 | -2.0 |
| loudness | 80s | -9.1 | 3.7 | -18.0 | -12.0 | -9.0 | -6.0 | -3.0 |
| liveness | 10s | 17.9 | 13.8 | 3.0 | 9.8 | 13.0 | 25.0 | 82.0 |
| liveness | 50s | 18.2 | 12.0 | 2.0 | 10.0 | 13.0 | 24.0 | 72.0 |
| liveness | 80s | 16.4 | 13.9 | 2.0 | 8.0 | 11.0 | 20.0 | 70.0 |
| valence | 10s | 46.4 | 20.9 | 9.0 | 29.0 | 47.0 | 61.0 | 97.0 |
| valence | 50s | 57.6 | 25.1 | 10.0 | 36.0 | 57.0 | 79.0 | 99.0 |
| valence | 80s | 63.1 | 25.9 | 11.0 | 40.0 | 70.0 | 85.0 | 98.0 |
| duration | 10s | 209.2 | 24.1 | 157.0 | 194.0 | 209.0 | 222.0 | 306.0 |
| duration | 50s | 149.5 | 22.5 | 98.0 | 135.0 | 148.0 | 163.0 | 214.0 |
| duration | 80s | 258.9 | 51.2 | 162.0 | 227.0 | 251.0 | 285.0 | 433.0 |
| acousticness | 10s | 14.7 | 17.9 | 0.0 | 3.8 | 9.0 | 19.0 | 84.0 |
| acousticness | 50s | 73.0 | 20.1 | 4.0 | 63.0 | 79.0 | 87.0 | 100.0 |
| acousticness | 80s | 24.6 | 21.7 | 0.0 | 7.0 | 19.0 | 36.0 | 80.0 |
| speechiness | 10s | 8.4 | 7.8 | 3.0 | 4.0 | 6.0 | 10.0 | 46.0 |
| speechiness | 50s | 4.2 | 1.8 | 3.0 | 3.0 | 3.0 | 4.0 | 11.0 |
| speechiness | 80s | 4.4 | 2.8 | 2.0 | 3.0 | 4.0 | 4.0 | 19.0 |
| popularity | 10s | 76.0 | 9.2 | 32.0 | 71.8 | 78.0 | 82.0 | 94.0 |
| popularity | 50s | 40.9 | 10.4 | 26.0 | 33.0 | 39.0 | 48.0 | 72.0 |
| popularity | 80s | 67.7 | 7.3 | 43.0 | 64.0 | 68.0 | 72.0 | 83.0 |

1 priedas Aprašomosios statistikos charakteristikos duomenų aibei pagal dešimtmetį

Žemiau programinis kodas: