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](#_Toc97061803)

[2 Duomenų aibė 4](#_Toc97061804)

[3 Atliktos analizės aprašymas 5](#_Toc97061805)

[3.1 Praleistos reikšmės 5](#_Toc97061806)

[3.2 Aprašomoji statistika 6](#_Toc97061807)

[3.3 Išskirčių analizė 8](#_Toc97061808)

[3.4 Duomenų normavimas 10](#_Toc97061809)

[3.5 Vizuali analizė 12](#_Toc97061810)

[3.6 Požymių koreliacijos 22](#_Toc97061811)

[4 Išvados 23](#_Toc97061812)

[Priedas 24](#_Toc97061813)

# 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

# Atliktos analizės aprašymas

## Aprašomoji statistika

Skaitiniams rodikliams apskaičiuotos pagrindinės aprašomosios statistikos charakteristikos (standartinis nuokrypis, vidurkis, mediana, mažiausia reikšmė (min), didžiausia reikšmė (max)). Rezultatai pateikti lentelėje (žr. 1 lentelė ).

1 lentelė Aprašomosios statistikos charakteristikos duomenų aibei

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | stand. nuokrypis | vidurkis | mediana | min | max |
| Inception | 3.23 | 2010.17 | 2011 | 1999 | 2014 |
| Employees | 393.11 | 145.59 | 56 | 1 | 7125 |
| Revenue | 3200082.76 | 10843584.61 | 10647231 | 1614585 | 21810051 |
| Expenses | 2119535.66 | 4313296.99 | 4366959.5 | 71219 | 9860686 |
| Profit | 3879083.89 | 6534258.87 | 6512379 | 12434 | 19624534 |
| Growth | 6.9 | 14.37 | 15 | -3 | 30 |

Tos pačios charakteristikos apskaičiuotos kiekvienai pramonės šakai atskirai (žr. 2 lentelė ). Lentelėje galime pamatyti, kad lyginimo charakteristika pasirinkus medianą, IT Services išsiskiria iš kitų pramonės šakų aukščiausiomis pajamomis ir pelnu (požymiai „Revenue“ ir „Profit“), Construction - žemiausiu darbuotojų skaičiumi (požymio „Employees“ medianinė reikšmė - 38), Health - žemiausiu pelnu (požymio „Profit“ medianinė reikšmė - 2514787) ir didžiausiomis išlaidomis (požymio „Expenses“ medianinė reikšmė – 6152151). Lyginant pagal standartinį nuokrypį stipriai išsiskiria Retail pramonės šaka dideliu standartiniu nuokrypiu darbuotojų skaičiui (požymio „Emplyees“ standartinis nuokrypis - 1045).

1 pav. Darbuotojų skaičiaus histograma

Dėl didelio išskirčių skaičiaus pagal darbuotojų skaičių, taikant statistinius metodus, naudojančius šio požymio reikšmes, būtina atsižvelgti į didelį išskirčių kiekį darbuotojų skaičiaus požymyje.

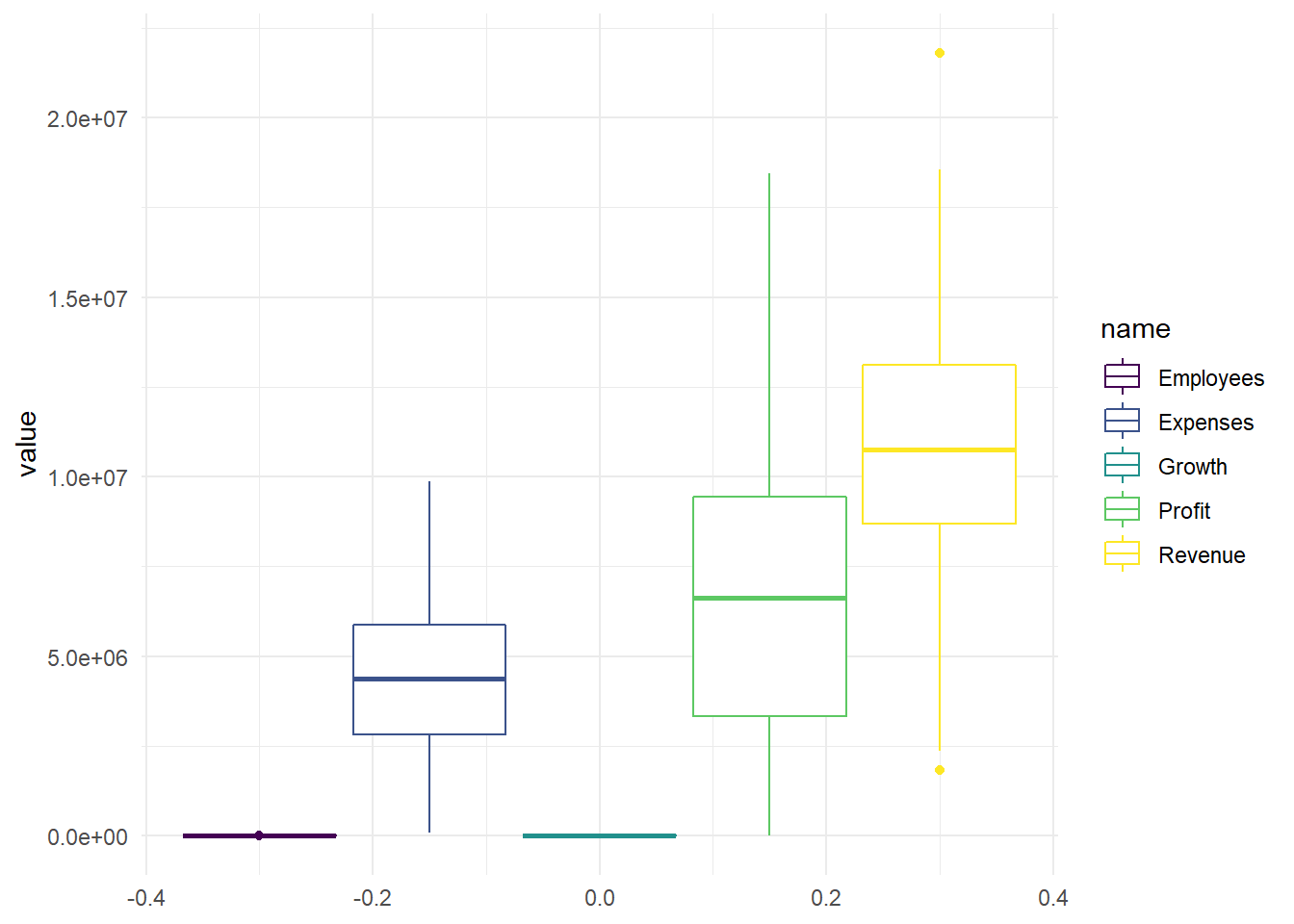
## Duomenų normavimas

Tarp skirtingų skaitinių požymių pastebėtas didelis reikšmių mastelio skirtumas (žr. Aprašomoji statistika). Dėl šios priežasties pasirinktiems taikyti statistiniams metodams gali būti reikalingas duomenų normavimas.

Duomenys sunormuoti naudojant min-max normavimą ir 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. 3 pav. ir 4 pav.).

Dėl didelio kiekio išskirčių (žr. Išskirčių analizė) nerekomenduojama taikyti standartizavimo metodą darbuotojų skaičiaus įmonėje požymiui.



2 pav. Kiekybinių požymių stačiakampė diagrama prieš atliekant normavimą

## PCA

Stačiakampėmis diagramomis pavaizduotas įmonių pelno pasiskirstymas pagal pramonės šaką (žr. 5 pav.). Pastebima, kad pelno vidurkiai IT Services, Financial Services ir Retail pramonės šakose didesni už likusių pramonės šakų.

Paveikslėlis, kuriame yra žinutė, dangus, žemėlapis

Automatiškai sugeneruotas aprašymas

5 pav. Įmonių pelnas pagal pramonės šaką

## MDS

## t-SNE

## Metodų palyginimas

# Išvados

Priklausomai nuo požymio specifikos, praleistos reikšmės užpildytos naudojant faktinį, išvestinį užpildymus, užpildymą tos pačios pramonės šakos medianine reikšme.

Rasti aprašomosios statistikos charakteristikų skirtumai tarp skirtingų industrijų: IT Services išsiskiria iš kitų pramonės šakų aukščiausiomis pajamomis ir pelnu tarp visų pramonės šakų (požymių „Revenue“ ir „Profit“ medianinės reikšmės atitinkamai 28% ir 21% didesnės už antroje vietoje pagal šiuos požymius esančias pramonės šakas), Construction – žemiausiu darbuotojų skaičiumi („Employees“ medianinė reikšmė 25% mažesnė už bet kokią kitą pramonės šaką), Health - žemiausiu pelnu („Profit“ mediana 36% mažesnė už antrą šiuo požymiu mažiausią) ir aukščiausiomis išlaidomis („Expenses“ mediana 13% didesnė už antrą šiuo požymiu didžiausią).

Duomenyse rasta 36 įmonių, išsiskiriančių pagal darbuotojų kiekį. Daroma išvada, kad taikant statistinius metodus, naudojančius šio požymio reikšmes, būtina atsižvelgti į didelį išskirčių kiekį duomenų aibėje. Pašalinus šias reikšmes iš duomenų aibės darbuotojų skaičiaus įmonėje standartinis nuokrypis sumažėtų 79%, vidurkis - 44%, mediana – 10%.

# Priedas

Žemiau programinis kodas: