Vilniaus Universitetas

Pirminė duomenų aibės analizė

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

Vilnius, 2022

# Tikslas ir uždaviniai

## Tikslas:

Nusiskaityti duomenų aibę, atlikti pirminį duomenų apdorojimą ir ją išanalizuoti (žr. Duomenų aibė).

## Uždaviniai:

Apskaičiuoti aprašomąsias duomenų statistikas.

Pasirinkus tinkamus metodus užpildyti praleistas duomenų reikšmes.

Ištirti duomenų aibės taškus atsiskyrėlius, įvertinti kaip pasikeičia duomenų aibės aprašomosios charakteristikos pašalinus šiuos taškus.

Pritaikyti duomenų normavimo metodus.

Atlikti aibės vizualią analizę.

Ištirti koreliacijas tarp duomenų aibės požymių.

# Duomenų aibė

Duomenų aibę sudaro duomenys apie 499 įmones su tokiais požymiais:

„ID“ - (nominalusis) įmonę duomenyse identifikuojantis kodas

„Name“ – (nominalus) įmonės pavadinimas

„Industry“ – (nominalus) pramonės šaka, kurioje veikia įmonė

„Inception“ – (kiekybinis) įmonės įkūrimo metai

„State“ - (nominalus) JAV valstija, kurioje įsikūrusi įmonė

„City“ – (nominalus) miestas, kuriame įsikūrusi įmonė

„Revenue“ – (kiekybinis) įmonės pajamos (JAV doleriais)

„Expenses“ – (kiekybinis) įmonės išlaidos (JAV doleriais)

„Profit“ – (kiekybinis) Įmonės pelnas (JAV doleriais)

„Growth“ – (kiekybinis) įmonės augimas (%)

# 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. Lentelė 1).

Lentelė 1 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. Lentelė 2). Lentelėje galime pamatyti, kad lyginimo charakteristika pasirinkus medianą, IT Services išsiskiria iš kitų pramonės šakų aukštomis pajamomis ir pelnu (požymiai „Revenue“ ir „Profit“), Construction - žemu darbuotojų skaičiumi („Employees“), Health - žemu pelnu („Profit“). Lyginant pagal standartinį nuokrypį stipriai išsiskiria Services pramonės šaka dideliu standartiniu nuokrypi darbuotojų skaičiui („Retail“).

Lentelė 2 Aprašomosios statistikos charakteristikos atskirai kiekvienai pramonės šakai

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Pramonės šaka | stand. nuokrypis | vidurkis | mediana | min | max |
| Inception | Construction | 3.53 | 2009.94 | 2011 | 1999 | 2014 |
| Employees | Construction | 59.43 | 61.26 | 37.5 | 5 | 272 |
| Revenue | Construction | 2404913.29 | 9158737.12 | 9055058.5 | 4419277 | 18429577 |
| Expenses | Construction | 1793321.66 | 4453204.5 | 4506975.5 | 214470 | 8213905 |
| Profit | Construction | 2805089.4 | 4705532.62 | 4573280.5 | 96073 | 12616182 |
| Growth | Construction | 3.07 | 10.06 | 10 | 5 | 19 |
| Inception | Financial Services | 2.71 | 2009.83 | 2010 | 2001 | 2014 |
| Employees | Financial Services | 261.95 | 183.88 | 79 | 3 | 1387 |
| Revenue | Financial Services | 1935037.65 | 10711858.77 | 11175012.5 | 5387469 | 14330107 |
| Expenses | Financial Services | 1521249.39 | 2351572.02 | 2379097 | 223602 | 6212849 |
| Profit | Financial Services | 2166383.23 | 8363033.13 | 8348842.5 | 3259485 | 12205097 |
| Growth | Financial Services | 2.69 | 16.68 | 17 | 10 | 23 |
| Inception | Government Services | 3 | 2010.3 | 2011 | 2000 | 2014 |
| Employees | Government Services | 233.63 | 172.72 | 99 | 13 | 1224 |
| Revenue | Government Services | 2342556.62 | 9436792.34 | 9707475 | 4637647 | 15188113 |
| Expenses | Government Services | 2055429.6 | 4741746.34 | 4790732.5 | 1243956 | 9860686 |
| Profit | Government Services | 2776630.41 | 4605150.06 | 4776526 | 46851 | 10565044 |
| Growth | Government Services | 2.87 | 5 | 5 | -3 | 11 |
| Inception | Health | 3.01 | 2010.89 | 2012 | 2000 | 2014 |
| Employees | Health | 308.32 | 205.51 | 86.5 | 6 | 1600 |
| Revenue | Health | 1978819.76 | 8811121.94 | 8855709.5 | 1614585 | 15312302 |
| Expenses | Health | 1892100.07 | 5881840.64 | 6162150.5 | 1323005 | 9712296 |
| Profit | Health | 2075213.51 | 2929281.3 | 2514786.5 | 12434 | 9174395 |
| Growth | Health | 2.6 | 6.59 | 6 | 0 | 14 |
| Inception | IT Services | 3.46 | 2009.9 | 2011 | 1999 | 2014 |
| Employees | IT Services | 257 | 107.81 | 52 | 2 | 2670 |
| Revenue | IT Services | 1950075.52 | 14175582.57 | 14121713 | 9691133 | 21810051 |
| Expenses | IT Services | 2043621.79 | 4149153.46 | 4068630 | 187655 | 9046498 |
| Profit | IT Services | 3003002.6 | 10019629.86 | 10160479 | 1841685 | 19624534 |
| Growth | IT Services | 3.09 | 21.4 | 21 | 15 | 30 |
| Inception | Retail | 3.38 | 2010.42 | 2011 | 1999 | 2014 |
| Employees | Retail | 1044.76 | 213.48 | 28 | 1 | 7125 |
| Revenue | Retail | 2183839.08 | 11581242.32 | 11654196 | 7307243 | 15880376 |
| Expenses | Retail | 1801630.91 | 4156855.09 | 4545730.5 | 968518 | 7957743 |
| Profit | Retail | 2897292.12 | 7482727.9 | 7326357 | 815381 | 13369247 |
| Growth | Retail | 2.59 | 12.5 | 12 | 8 | 19 |
| Inception | Software | 3.18 | 2010.08 | 2011 | 2000 | 2014 |
| Employees | Software | 179.75 | 121.06 | 58 | 3 | 850 |
| Revenue | Software | 2646904.18 | 7914512.71 | 8304480 | 1835717 | 14229411 |
| Expenses | Software | 1940555.97 | 3822601.62 | 4175332 | 71219 | 8007771 |
| Profit | Software | 2951684.76 | 4091911.1 | 3952602 | 68862 | 11902072 |
| Growth | Software | 2.89 | 18.89 | 19 | 13 | 26 |

## Praleistos reikšmės

Praleistos valstijų reikšmėms užpildytos naudojant faktinį užpildymą naudojant esamas miestų, kuriuose įsikūrusi įmonė pavadinimus.

Laikant, kad stulpelius „Revenue“, „Expenses“ ir „Profit“ sieja ryšys , esant praleistai vienai reikšmei iš šių trijų likusi apskaičiuota išvestiniu būdu.

To negalint padaryti, praleistos reikšmės stulpeliuose „Revenue“ ir „Expenses“ užpildytos pramonės šakos, kurioje veikia įmonė medianinėmis reikšmėmis.

Toks pat praleistų reikšmių užpildymo metodas taikytas ir stulpeliams „Employees“ ir „Growth“.

Kategoriniuose kintamamuosiuose esančios praleistos reikšmės paliktos nekeistos.

## Išskirčių analizė

Išskirtys vertintos naudojant vidinį barjerą , kur – atitinkamai pirmas ir trečias kvartiliai, – interkvartilinis plotis.

Naudota šį metodą rastos 4 įmonės išsiskiriančios pagal pajamas (stulp. „Revenue“), 2 įmonės išsiskiriančios pagal pelną (stulpelis „Profit“) ir 60 įmonių išsiskiriančiu pagal didelį darbuotojų skaičių. Darbuotojų skaičiaus įmonėje histogramoje galima pastebėti išsiskiriančias įmones (žr. 1 Pav.)



1 Pav. Darbuotojų skaičiaus histograma

Pasirinkta toliau tirti tik mažo-vidutinio dydžio (pagal darbuotojų skaičių) įmones ir šios išsiskiriančios reikšmės pašalintos. Pašalinus šias išskirtis kartu pašalintos ir įmonės išsiskyrusios pagal kitus požymius.

Pašalinus išskirtis pakartotinai apskaičiuotos duomenų aibės aprašomosios statistikos charakteristikos (žr. Lentelė 3). Didžiausias pokytis pastebimas darbuotojų kiekyje – pašalimus išskirtis darbuotojų kiekio vidurkis sumažėjo 44%, standartinis nuokrypis 79%, mediana – 11%.

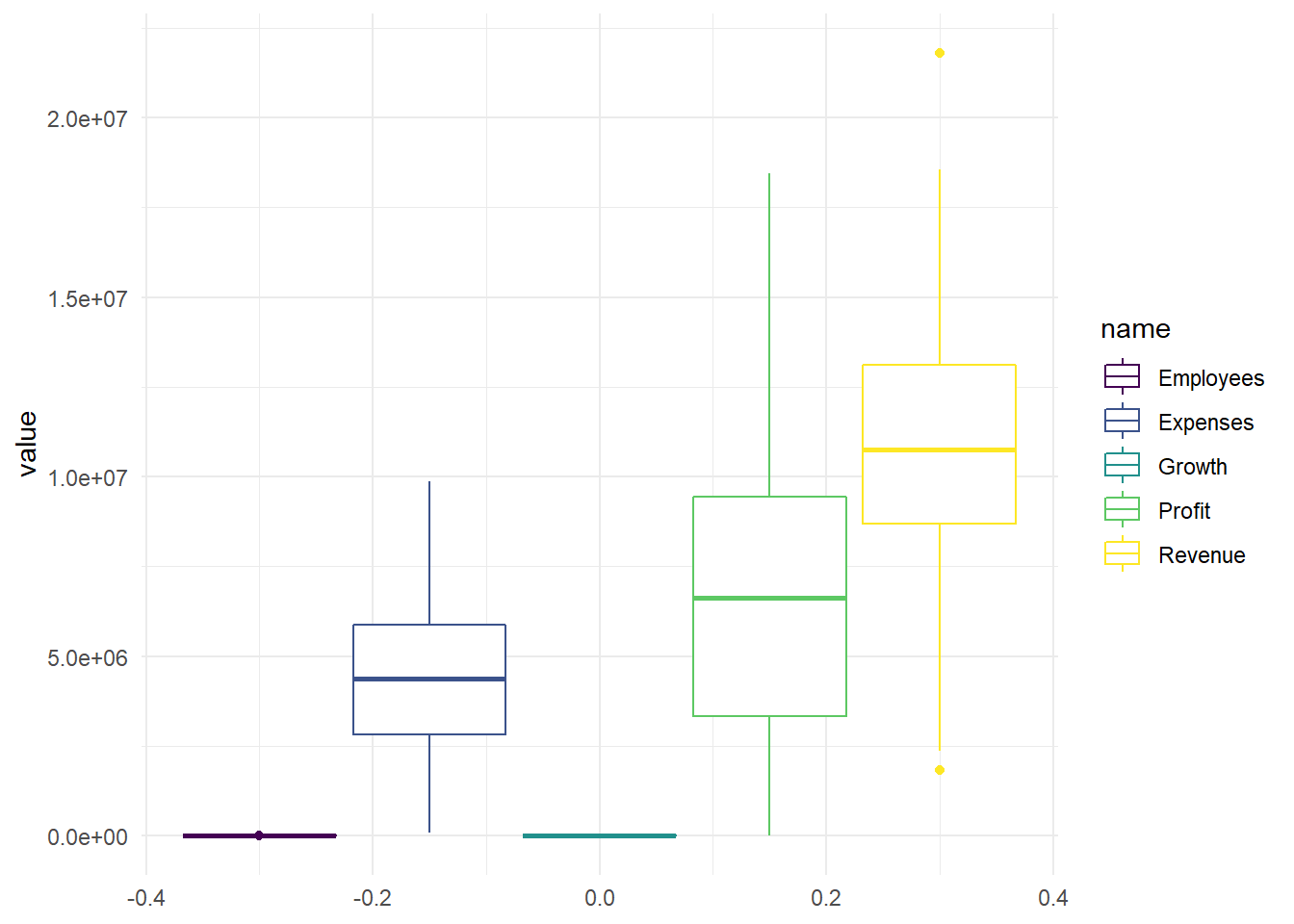
Lentelė 3 Procentinis aprašomosios statistikos charakteristikų pokytis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | stand. nuokrypis | vidurkis | mediana | min | max |
| Inception | 0.48 | 0 | 0 | 0 | 0 |
| Employees | -79.25 | -44.39 | -10.71 | 0 | -94.34 |
| Revenue | -1.32 | 0.6 | 0.71 | 13.7 | 0 |
| Expenses | -0.31 | 0.3 | -0.07 | 0 | 0 |
| Profit | -0.51 | 0.8 | 1.31 | 0 | -5.95 |
| Growth | -0.68 | 1.53 | 6.67 | 0 | 0 |

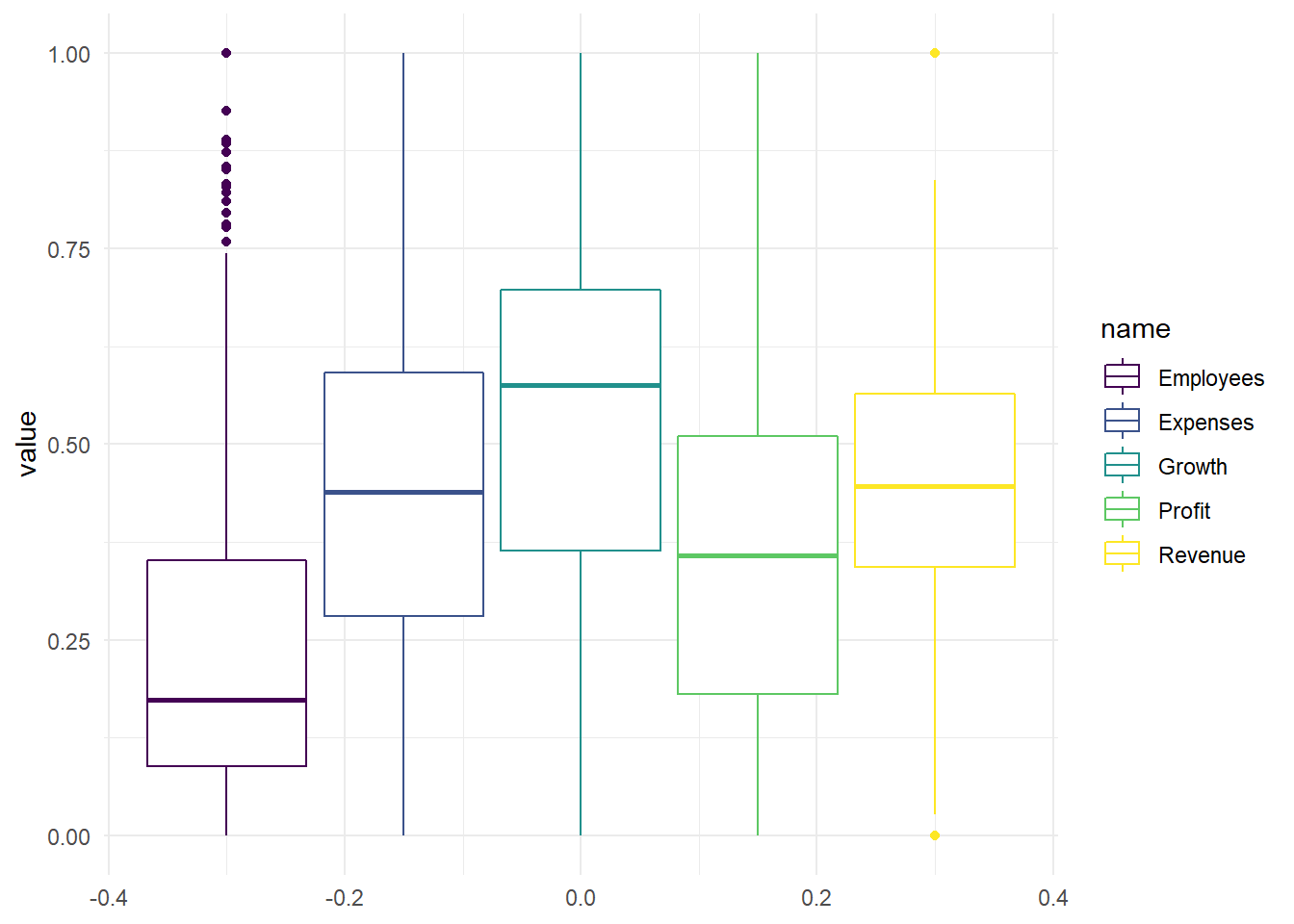
## Duomenų normavimas

Duomenys sunormuoti naudojant min-max normavimą ir normavimą pagal vidurkį ir dispersiją (standartizavimas) , kur - požymio vidurkis požymio dispersija,   – požymio vidurkis.

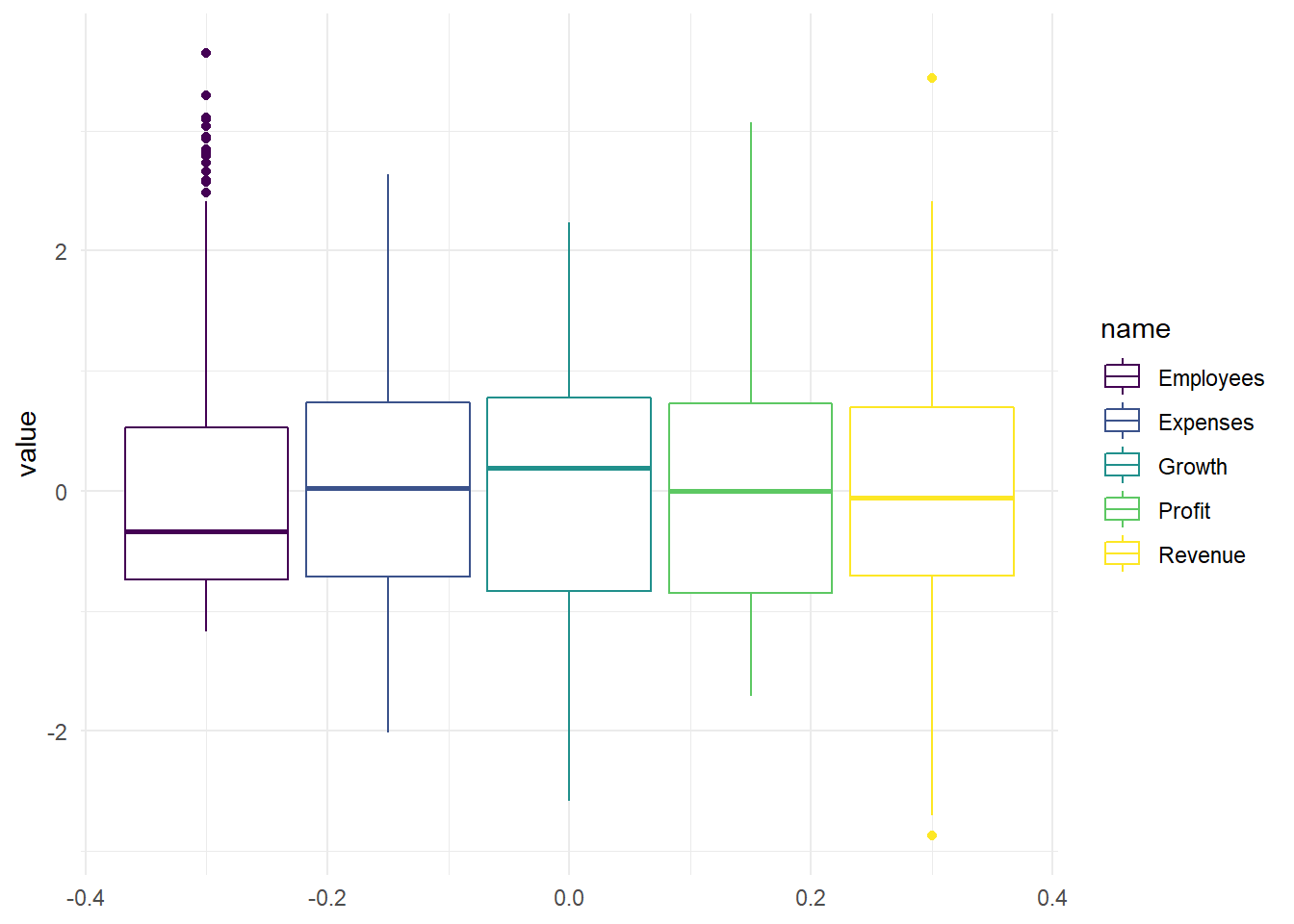
Pradinis kiekybinių duomenų aibės požymių pasiskirstymas pavaizduotas stačiakampe diagrama (žr. 2 Pav.). Pakartotinai pavaizduotas pasiskirstymas atlikus abu anksčiau minėtus normavimo metodus (žr. 3 Pav. ir 4 Pav.)



2 Pav. Kiekybinių požymių stačiakampė diagrama prieš atliekant normavimą



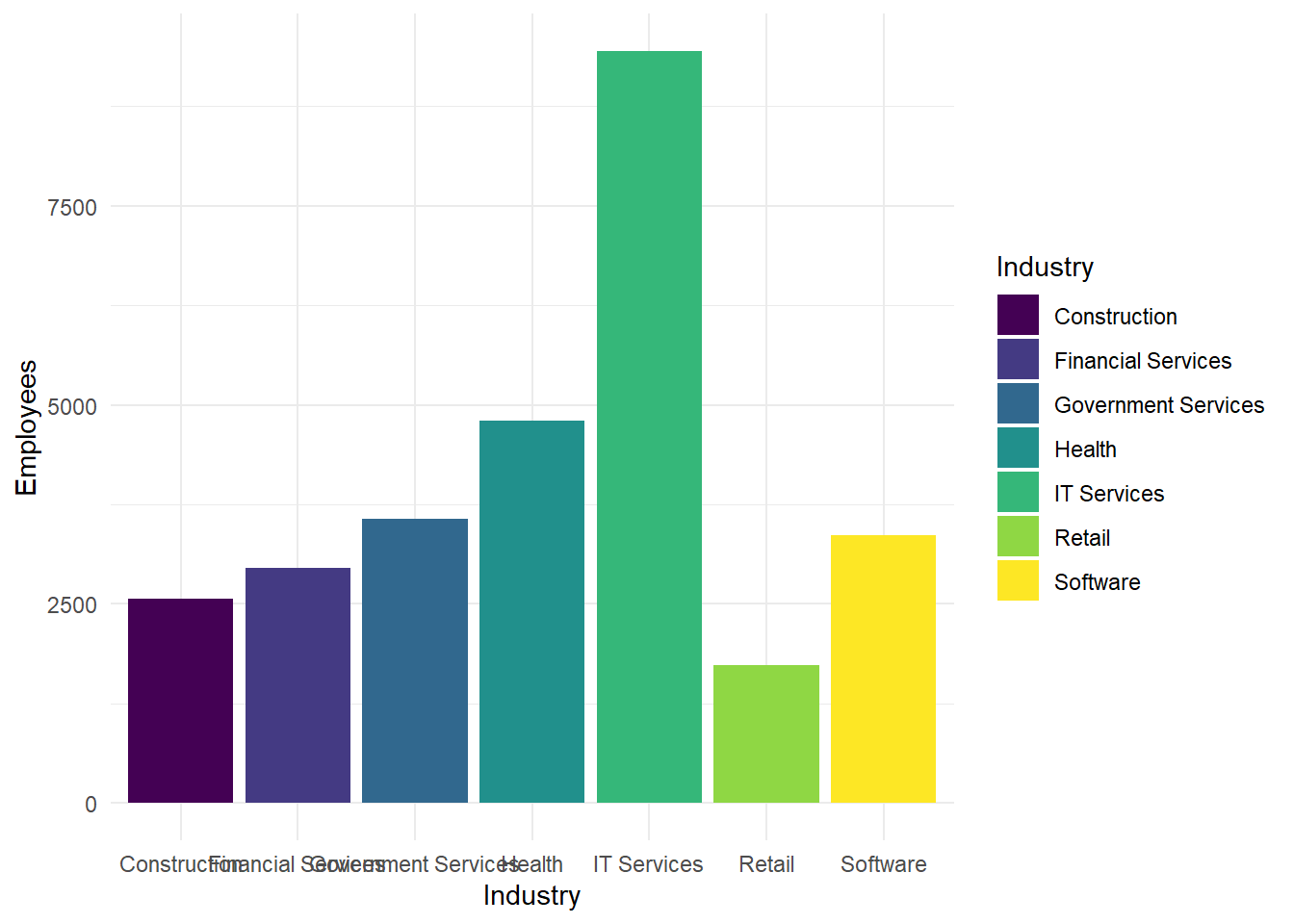
3 Pav. Stačiakampė diagrama atlikus min-max normavimą



4 Pav. Stačiakampė diagrama atlikus standartizaciją

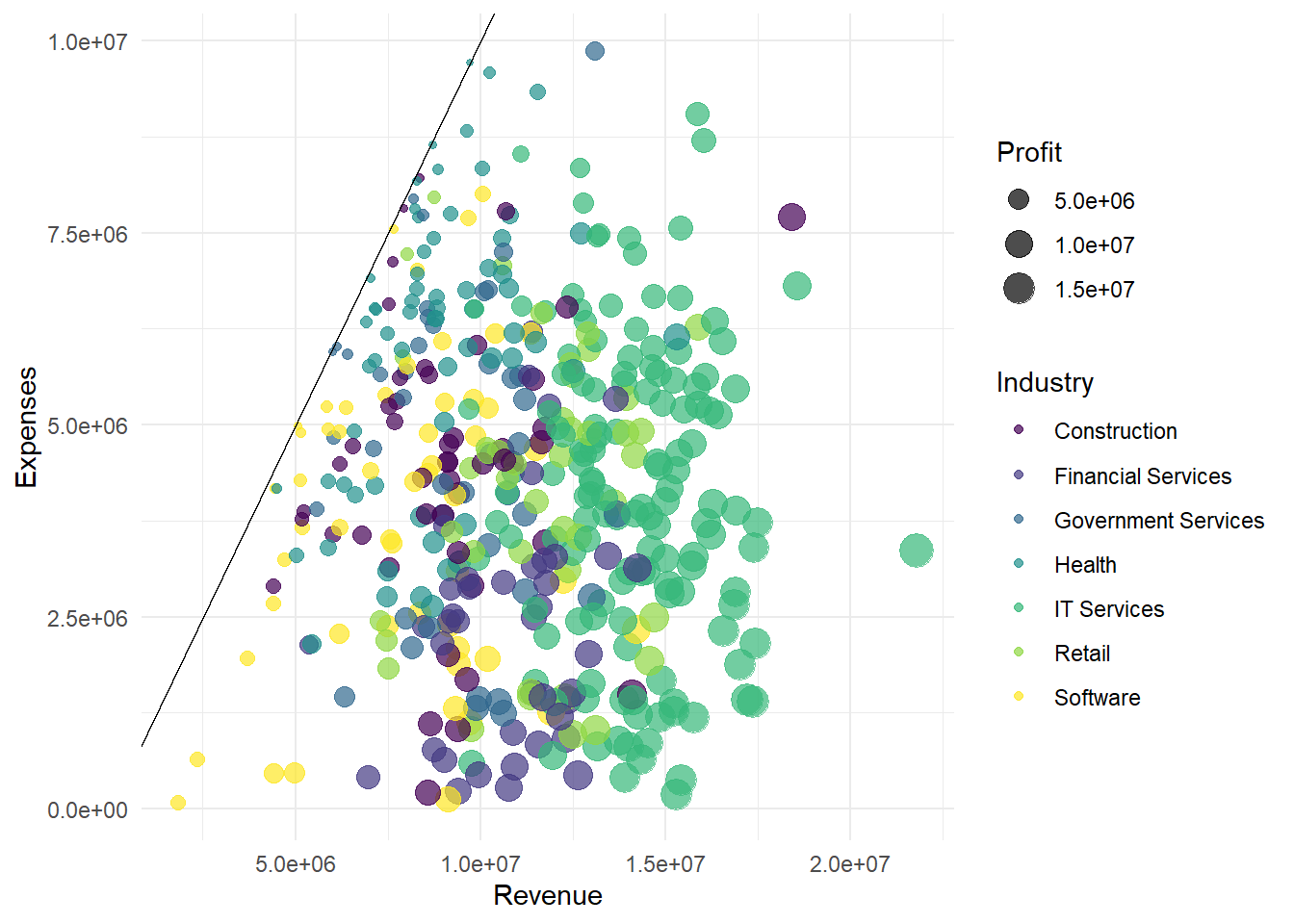
## Vizuali analizė

Darbuotojų kiekiai įmonėse susumuoti pagal pramonės šaką ir pavaizduoti stulpeline diagrama (žr. 5 Pav.). Matoma, kad didžiausias darbuotojų skaičius duomenų aibėje dirba IT Services, mažiausias – Retail.



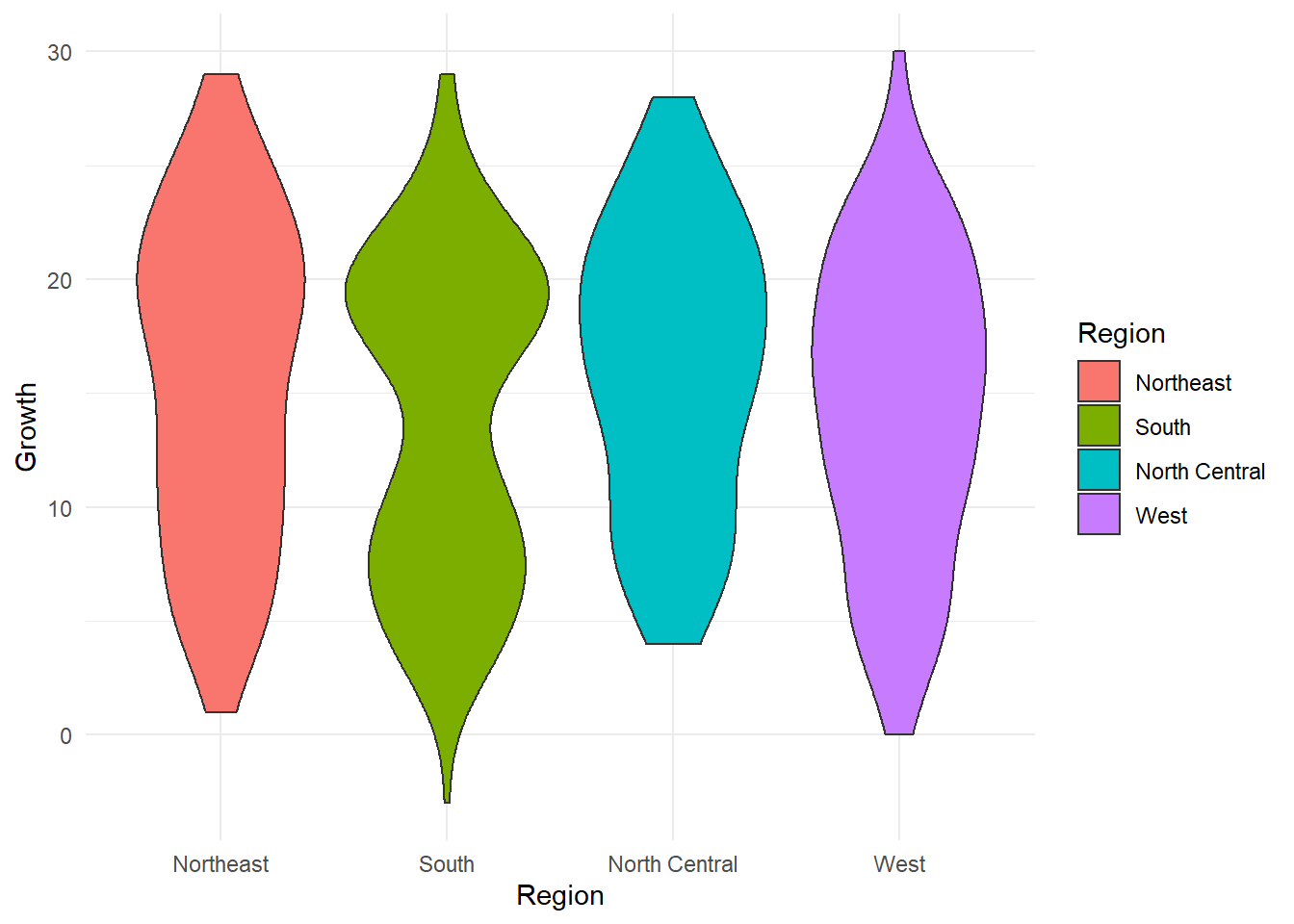
5 Pav. Bendras darbuotojų skaičius pagal pramonės šaką

Sklaidos diagrama pavaizduotas įmonių pajamų ir išlaidų sklaidos diagrama kartu su palyginamąja tiese (žr. 6 Pav.) Iš grafiko matome, kad jokios įmonės duomenų aibėje nepatyrė nuostolių, pelningiausios yra IT Services įmonės.



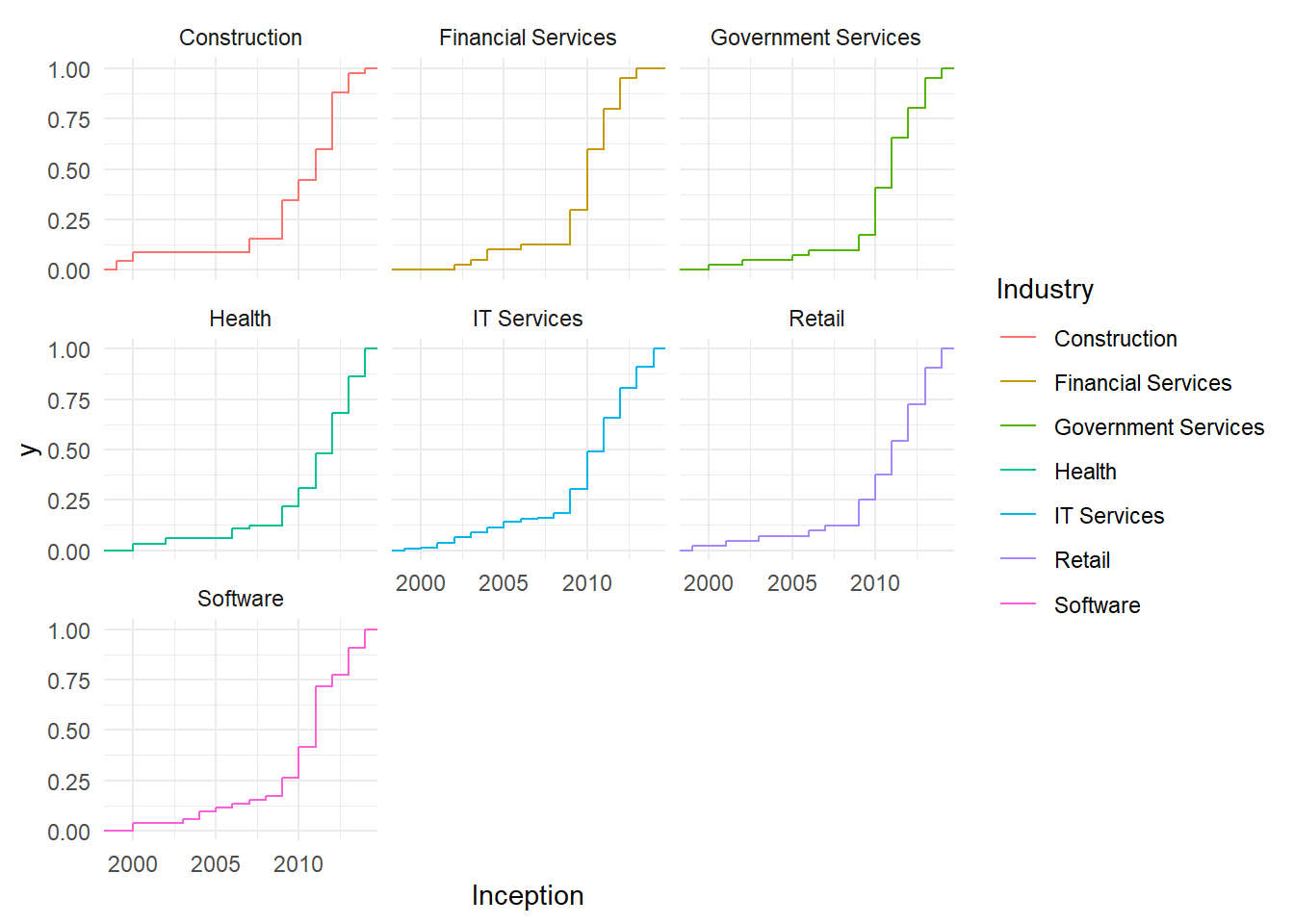
6 Pav. Įmonių pajamų ir išlaidų sklaidos diagrama

JAV valstijos padalintos į keturis regionus ir smuiko formos grafikais kiekvienam regionui pavaizduotas įmonių augimo pasiskirstymas (žr. 7 Pav.) Ryškių įmonių augimo skirtumų tarp JAV regionų nepastebima.



7 Pav. Įmonių augimo pasiskirstymas pagal JAV regionus

Pavaizduota įmonių įsikūrimo metų empirinė pasiskirstymo funkcija (žr. 8 Pav.). Matome, kad didžioji dalis įmonių, esančių duomenų aibėje, įkurta nuo maždaug 2009-2010 metų. Ši tendencija galioja visoms pramonėms šakomis.



8 Pav. Įmonių įkūrimo metų empirinė pasiskirstymo funkcija pagal pramonės šaką

## Požymių koreliacijos

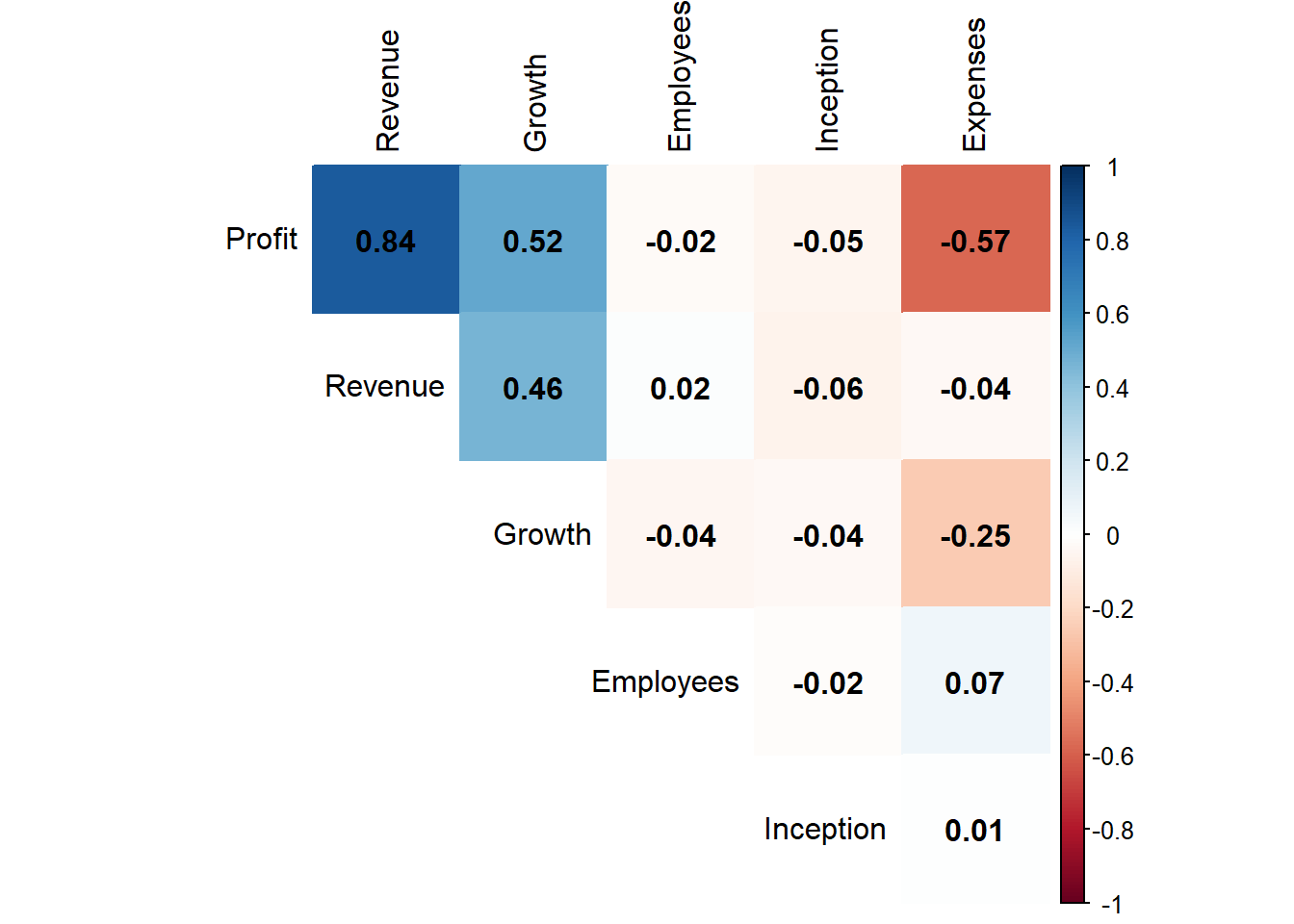
Tarp skaitinių rodiklių apskaičiuotos Pirsono koreliacijos koeficientų reikšmės (angl. Pearson correlation coefficient).

Gauti rezultatai pateikti lentelėje (žr. Lentelė 4). Rasta stipri teigiama tarp pajamų ir pelno (r = 0.84). Taip pat rastos vidutinio stiprumo teigiamos koreliacijos tarp pelno ir augimo (r = 0.52), pajamų ir augimo (r = 0.46) ir neigiama koreliacija tarp pelno ir išlaidų (r = -0.57)

Šios reikšmės papildomai vizualizuotos koreliacijų diagrama (žr. 14 Pav.).

Lentelė 4 Pirsono koreliacijos koeficientai tarp skaitinių įmonių požymių

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Inception | Employees | Revenue | Expenses | Profit | Growth |
| Inception | 1 | 0 | -0.08 | 0.01 | -0.07 | -0.05 |
| Employees | 0 | 1 | -0.01 | 0.07 | -0.05 | -0.08 |
| Revenue | -0.08 | -0.01 | 1 | -0.03 | 0.84 | 0.46 |
| Expenses | 0.01 | 0.07 | -0.03 | 1 | -0.57 | -0.25 |
| Profit | -0.07 | -0.05 | 0.84 | -0.57 | 1 | 0.52 |
| Growth | -0.05 | -0.08 | 0.46 | -0.25 | 0.52 | 1 |



14 Pav. Pirsono koreliacijos tarp skaitinių požymių koeficientai

# Išvados

Rasti aprašomosios statistikos charakteristikų skirtumai tarp skirtingų industrijų: IT Services išsiskiria iš kitų pramonės šakų aukštomis pajamomis ir pelnu (požymiai „Revenue“ ir „Profit“), Construction - žemu darbuotojų skaičiumi („Employees“), Health - žemu pelnu („Profit“).

22 praleistos reikšmės užpildytos naudojant faktinį, išvestinį užpildymus, užpildymą tos pačios pramonės šakos mediana. Įmonių su likusiomis praleistomis reikšmėmis (daugiausia šių reikšmių yra nominaliuose požymiuose) pasirinkta nešalinti iš duomenų aibės.

Duomenyse rasta 60 įmonių, išsiskiriančių pagal darbuotojų kiekį. Pasirinkus susiaurinti duomenų aibės analizę iki mažo-vidutinio dydžio įmonių, šios išskirtys pašalintos iš duomenų aibės. Pašalinus šias reikšmes iš duomenų aibės darbuotojų skaičiaus įmonėje standartinis nuokrypis sumažėjo 79%, vidurkis - 44%, mediana – 10%

Kiekybiniams požymiams atliktos min-max normavimas ir normavimas pagal vidurkį ir dispersiją (standartizacija) ir gauti rezultatai tarpusavyje palyginti.

Atlikus vizualią duomenų aibę rasta, kad duomenų aibėje didžiausią pelną gauna IT Services įmonės. Lyginant JAV regionus nerasta ryškių skirtumų įmonių augime.

Apskaičiavus koreliacijas tarp požymių rasta stipri teigiama koreliacija tarp pajamų ir pelno (r = 0.84), vidutinio stiprumo teigiamos koreliacijos tarp pelno ir augimo (r = 0.52), pajamų ir augimo (r = 0.46) ir neigiama koreliacija tarp pelno ir išlaidų (r = -0.57).

# Priedas

Žemiau pateiktas visas naudotas programinis kodas: