
VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFOMATIKOS FAKULTETAS

DUOMENŲ MOKSLO BAKALAURAS

# PIRMINĖ DUOMENŲ AIBĖS ANALIZĖ

Ataskaita

Atliko: Simona Gelžinytė,

Ugnė Kniukštaitė, Rugilė Bagdonaitė

duomenų mokslas 3 k.

Vilnius, 2023

# TURINYS

# ĮVADAS

## Tikslas

Ištirti duomenų aibę.

## Uždaviniai

- Nusiskaityti duotų duomenų aibę.
- Pateikti aprašomąsias duomenų statistikas.
- Užpildyti arba pašalinti praleistus duomenis.
- Ištirti taškus atsiskyrėlius.
- Pritaikyti porą normavimo metodų.
- Ištirti požymių koreliacijas.
- Išanalizuoti gautus rezultatus.

## Duomenys

Duomenyse yra pateikta informacija apie įvairias kompanijas iš JAV: įmonės pavadinimas ir id numeris, industrijos šaka, kuriai priklauso konkreti įmonė, įkūrimo metai, darbuotojų skaičius, valstija ir miestas, kuriame yra įmonė, pajamos, išlaidos, pelnas ir įmonės prieaugio procentas. Duomenų rinkinį sudaro 500 stebėjimų, kurie susideda iš 12 anksčiau išvardintų atributų. Rinkinyje yra keletas nominaliųjų kategorinių kintamųjų: įmonės pavadinimas, industrijos šaka, apimanti programinės įrangos, informacinių technologijų paslaugų, mažmeninės prekybos, finansinių paslaugų, statybos, sveikatos ir vyriausybinių paslaugų pramonės sektorius, valstijų sutrumpinimai, apimantys 42 valstijas, taip pat pateiktos 297 skirtingų miestų kategorijos. Įmonės įkūrimo metai, darbuotojų skaičius, pajamų, išlaidų, pelno ir įmonės padidėjimo procentas yra skaitiniai kintamieji. Įmonių įkūrimo metai kinta nuo 1999 m. iki 2014 m.. Bendrovėse dirba nuo 1 iki 7125 darbuotojų. Įmonių pajamos, išlaidos ir pelnas kinta atitinkamai nuo 1,6 mln. iki 21,8 mln. dolerių, nuo 71 tūkst. iki 9,8 mln. dolerių, nuo 12 tūkst. iki 675 mln. dolerių. Bendrovių prieaugis kito nuo -3 iki 30 proc.. Duomenų rinkinyje turime praleistų reikšmių.

*1 lentelė. Duomenų tipai ir jų skalės.*

| | Name | Industry | Inception | Employees | State |
|---|---|---|---|---|---|
| Duomenų tipas | Nominalus | Nominalus | Diskretieji | Diskretieji | Nominalus |
| Skalė | Nominali | Nominali | Intervalinė | Santykių | Nominali |
| | City | Revenue | Expenses | Profit | Growth |
| Duomenų tipas | Nominalus | Tolydieji | Tolydieji | Tolydieji | Tolydieji |
| Skalė | Nominali | Santykių | Santykių | Santykių | Santykių |

# PIRMINIS DUOMENŲ APDOROJIMAS

## Pradinė aprašomoji statistika

Skaitiniams rodikliams apskaičiuotos pagrindinės aprašomosios statistikos charakteristikos (standartinis nuokrypis, vidurkis, mediana, mažiausia reikšmė (min), didžiausia reikšmė (max), 1 ir 3 kvartilis).

*2 lentelė. Pagrindinės aprašomosios statistikos charakteristikos.*

| | stand. nuokr. | vidurkis | mediana | min | max | $Q_1$ | $Q_3$ |
|---|---|---|---|---|---|---|---|
| Inception | 3,23 | 2010 | 2011 | 1999 | 2014 | 2009 | 2012 |
| Employees | 398,08 | 149,09 | 57 | 1 | 7125 | 27,75 | 126 |
| Revenue | 3,191 mln. | 10,848 mln. | 10,647 mln. | 1,615 mln. | 21,810 mln. | 8,697 mln. | 13,101 mln. |
| Expenses | 2,118 mln. | 4,318 mln. | 4,367 mln. | 0,071 mln. | 9,861 mln. | 2,762 mln. | 5,835 mln. |
| Profit | 30,207 mln. | 7,881 mln. | 6,513 mln. | 0,012 mln. | 675,081 mln. | 3,272 mln. | 9,365 mln. |
| Growth | 6,90 | 14,36 | 15 | -3 | 30 | 8 | 20 |

Tos pačios charakteristikos apskaičiuotos ir skirtingoms pramonėms šakoms (3 lentelė - 9 lentelė). Pasirinkus lyginimo charakteristiką – medianą, matome, jog informacinių technologijų pramonės šaka išsiskiria aukščiausiomis pajamomis, pelnu bei įmonės augimu. Mažiausiu darbuotojų skaičiumi pasižymi – statybų sektorius (medianinė reikšmė - 38), o didžiausiu – vyriausybės paslaugų sektorius (medianinė reikšmė - 99), šioje pramonės šakoje taip pat pastebimas mažiausias įmonės augimas (medianinė reikšmė - 5). Sveikatos sektorius – išsiskiria didžiausiomis išlaidomis (medianinė reikšmė - 6186394), bet turi mažiausią pelną (medianinė reikšmė - 2514786), mažiausiai išlaidų turi finansų sektorius. Mažmeninės prekybos šaka stipriai išsiskiria standartiniu nuokrypiu darbuotojų skaičiui (medianinė reikšmė – 1045). Lentelėse mėlyna spalva nuspalvoti langeliai žymi – mažiausią reikšmę įgyjusį atributą, oranžine – didžiausią reikšmę.

*3 lentelė. Construction industrijos charakteristikos.*

| | vidurkis | stand. nuokrypis | mediana | min | max |
|---|---|---|---|---|---|
| Inception | 2009,94 | 3,53 | 2011 | 1999 | 2014 |
| Employees | 62,1 | 59,74 | 38 | 5 | 272 |
| Revenue | 9158737 | 2404913 | 9055058 | 4419277 | 18429577 |
| Expenses | 4452433 | 1812701 | 4515112 | 214470 | 8213905 |
| Profit | 18628449 | 96806902 | 4573280 | 96073 | 675080995 |
| Growth | 10,06 | 3,07 | 10 | 5 | 19 |

*4 lentelė. Financial Services industrijos charakteristikos.*

| | vidurkis | stand. nuokrypis | mediana | min | max |
|---|---|---|---|---|---|
| Inception | 2009,89 | 2,72 | 2010 | 2001 | 2014 |
| Employees | 211,13 | 326,57 | 80 | 3 | 1628 |
| Revenue | 10658702 | 1928785 | 10978676 | 5387469 | 14330107 |
| Expenses | 2362818 | 1509094 | 2412491 | 223602 | 6212849 |
| Profit | 8295883 | 2167587 | 8301998 | 3259485 | 12205097 |
| Growth | 16,67 | 2,67 | 17 | 10 | 23 |

*5lentelė. Government Services industrijos charakteristikos.*

| | vidurkis | stand. nuokrypis | mediana | min | max |
|---|---|---|---|---|---|
| Inception | 2010,3 | 3 | 2011 | 2000 | 2014 |
| Employees | 172,72 | 233,63 | 99 | 13 | 1224 |
| Revenue | 9436792 | 2342557 | 9707475 | 4637647 | 15188113 |
| Expenses | 4741746 | 2055430 | 4790732 | 1243956 | 9860686 |
| Profit | 4695046 | 2820709 | 4836706 | 46851 | 10565044 |
| Growth | 5 | 2,87 | 5 | -3 | 11 |

*6 lentelė. Health industrijos charakteristikos.*

| | vidurkis | stand. nuokrypis | mediana | min | max |
|---|---|---|---|---|---|
| Inception | 2010,89 | 3,01 | 2012 | 2000 | 2014 |
| Employees | 205,87 | 305,92 | 86,5 | 6 | 1600 |
| Revenue | 8837185 | 1975660 | 8873078 | 1614585 | 15312302 |
| Expenses | 5935474 | 1836392 | 6186394 | 2140030 | 9712296 |
| Profit | 2929281 | 2075214 | 2514786 | 12434 | 9174395 |
| Growth | 6,59 | 2,6 | 6 | 0 | 14 |

*7 lentelė. IT Services industrijos charakteristikos.*

| | vidurkis | stand. nuokrypis | mediana | min | max |
|---|---|---|---|---|---|
| Inception | 2009,9 | 3,46 | 2011 | 1999 | 2014 |
| Employees | 107,81 | 257 | 52 | 2 | 2670 |
| Revenue | 14156466 | 1966260 | 14121713 | 9691133 | 21810051 |
| Expenses | 4155306 | 2049407 | 4083060 | 187655 | 9046498 |
| Profit | 10019630 | 3003003 | 10160479 | 1841685 | 19624534 |
| Growth | 21,4 | 3,09 | 21 | 15 | 30 |

*8 lentelė. Retail industrijos charakteristikos.*

| | vidurkis | stand. nuokrypis | mediana | min | max |
|---|---|---|---|---|---|
| Inception | 2010,36 | 3,39 | 2011 | 1999 | 2014 |
| Employees | 207,13 | 1045,06 | 52 | 2 | 2670 |
| Revenue | 11558774 | 2147426 | 11654196 | 7307243 | 15880376 |
| Expenses | 4201291 | 1782205 | 4600156 | 968518 | 7957743 |
| Profit | 7357483 | 2794186 | 7136631 | 815381 | 12651172 |
| Growth | 12,51 | 2,62 | 12 | 8 | 19 |

*9 lentelė. Software industrijos charakteristikos.*

| | vidurkis | stand. nuokrypis | mediana | min | max |
|---|---|---|---|---|---|
| Inception | 2010,06 | 3,16 | 2011 | 2000 | 2014 |
| Employees | 122,32 | 179,46 | 62 | 3 | 850 |
| Revenue | 7930751 | 2629024 | 8333446 | 1835717 | 14229411 |
| Expenses | 3826462 | 1925341 | 4129542 | 71219 | 8007771 |
| Profit | 4104289 | 2929839 | 3957674 | 68862 | 11902072 |
| Growth | 18,95 | 2,91 | 19 | 13 | 26 |

## Praleistų reikšmių sutvarkymas

Pradinę duomenų aibę sudaro 500 įrašų. Iš pradžių buvo rasti 24 įrašai su praleista bent viena reikšme. Atlikus faktinį valstijų užpildymą pagal duotus miestus bei, kur buvo galima, paskaičiavus išlaidas arba pajamas pagal formulę:

$$pajamos = išlaidos + pelnas,$$

praleistų reikšmių liko 12, t. y. iš 500 įrašų tik 12 stebėjimų negalėjome užpildyti pagal turimus duomenis, tai sudaro 2,4 proc. visų duomenų. Galime teigti, jog duomenys yra patikimi. Pirmiausia buvo pašalinti įrašai, kur buvo nenurodyta industrija arba metai, kadangi šių duomenų gauti negalėjome. Tuomet darbuotojų skaičius buvo užpildytas pagal industrijos medianą. Įrašai,

kuriuose buvo praleistos 3 - 4 reikšmės finansiniuose rodikliuose, buvo pašalinti. Likusias praleistas reikšmes pajamų ir išlaidų stulpeliuose buvo galima apskaičiuoti pasinaudojus anksčiau pateikta formule, o prieaugio stulpelyje 2 likusios praleistos reikšmės buvo užpildytos mediana, gauta atsižvelgus į industrijos šaką. Šį metodą pasirinkome, nes nėra jautrus išskirtims.

Praleistų reikšmių tvarkymą kartojome dar porą kartų - medianą pakeisdamos į vidurkį ir paskui į modą bei palyginome gautus rezultatus.

Employees

*10 lentelė. Darbuotojų skaičiaus charakteristikos*

| | min | 1 kvart. | mediana | vidurkis | 3 kvart. | max |
|---|---|---|---|---|---|---|
| Mediana | 1 | 28 | 55,5 | 148,5 | 125,8 | 7125 |
| Vidurkis | 1 | 28 | 57 | 149,3 | 146 | 7125 |
| Moda | 1 | 27 | 55 | 148,2 | 125,8 | 7125 |

Revenue

*11 lentelė. Pajamų charakteristikos.*

| | min | 1 kvart. | mediana | vidurkis | 3 kvart. | max |
|---|---|---|---|---|---|---|
| Mediana | 1614585 | 8662529 | 10647231 | 10835609 | 13097316 | 21810051 |
| Vidurkis | 1614585 | 8662529 | 10647231 | 10835609 | 13097316 | 21810051 |
| Moda | 1614585 | 8662529 | 10647231 | 10835609 | 13097316 | 21810051 |

Expenses

*12 lentelė. Išlaidų charakteristikos.*

| | min | 1 kvart. | mediana | vidurkis | 3 kvart. | max |
|---|---|---|---|---|---|---|
| Mediana | 71219 | 2758418 | 4307867 | 4297138 | 5794227 | 9860686 |
| Vidurkis | 71219 | 2758418 | 4307867 | 4297138 | 5794227 | 9860686 |
| Moda | 71219 | 5758418 | 4307867 | 4297138 | 2794227 | 9860686 |

Profit

*13 lentelė. Pelno charakteristikos.*

| | min | 1 kvart. | mediana | vidurkis | 3 kvart. | max |
|---|---|---|---|---|---|---|
| Mediana | 12434 | 3314984 | 6513366 | 7891305 | 9365441 | 675080995 |
| Vidurkis | 12434 | 3314984 | 6513366 | 7891305 | 9365441 | 675080995 |
| Moda | 12434 | 3314984 | 6513366 | 7891305 | 9365441 | 675080995 |

Growth

*24 lentelė. Prieaugio charakteristikos.*

| | min | 1 kvart. | mediana | vidurkis | 3 kvart. | max |
|---|---|---|---|---|---|---|
| Mediana | -3 | 8 | 16 | 14.42 | 20 | 30 |
| Vidurkis | -3 | 8 | 16 | 14.42 | 20 | 30 |
| Moda | -3 | 8 | 16 | 14.42 | 20 | 30 |

Pagal (11 lentelė - 24 lentelė) pateiktus rezultatus galima matyti, kad užpildžius praleistas reikšmes vidurkiu, moda arba mediana gautos statistikos visiškai nesiskiria pajamoms, išlaidoms ir pelnui, o darbuotojų skaičiui reikšmes užpildžius mediana ir moda gautos charakteristikos yra labai panašios, tik su vidurkiu rodikliai truputį didesni.

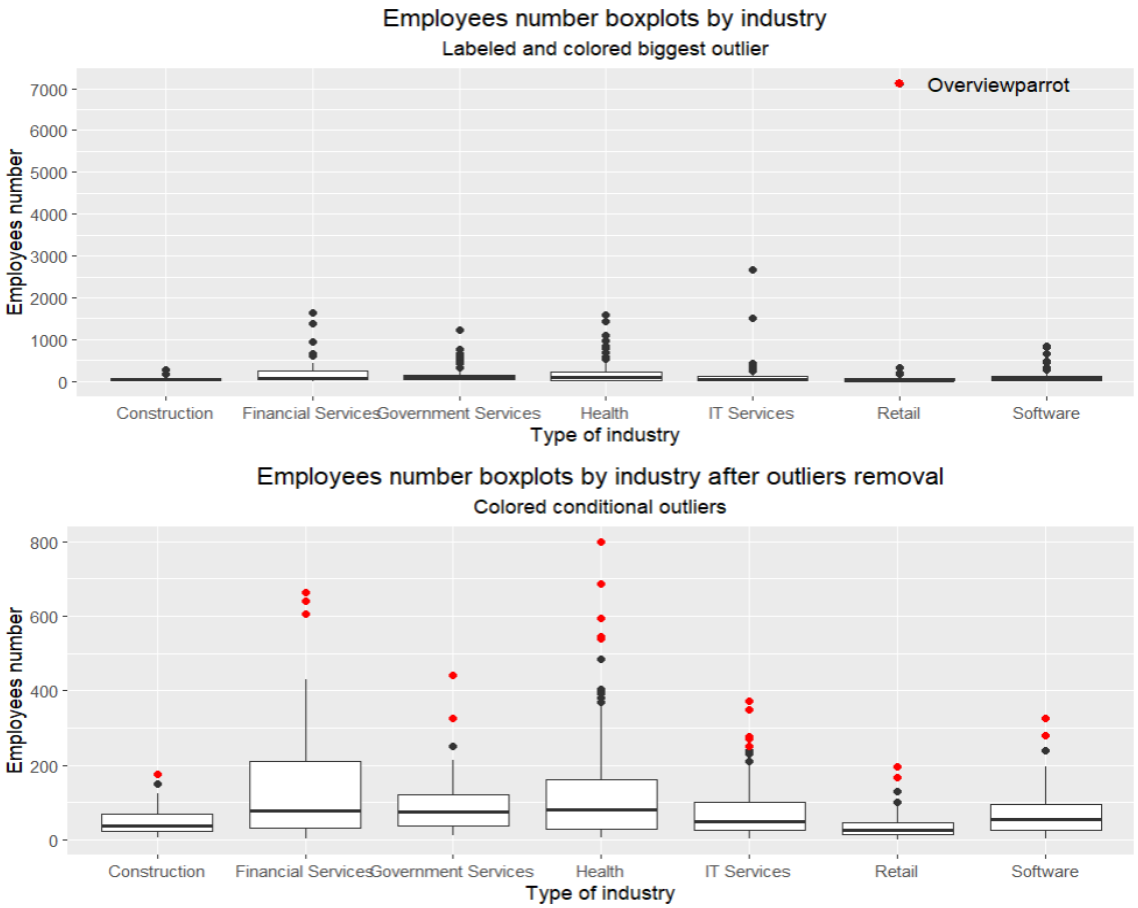
### Išskirčių identifikavimas bei sutvarkymas

Norint atpažinti išskirtis duomenų aibėje, pasinaudojome statistiniais išskirčių apibrėžimais:

- Sąlyginė išskirtis identifikuojama, kai stebėjimo reikšmė yra tarp vidinio ir išorinio barjero, t. y. kai stebėjimas pakliūva į intervalą $(Q_1 - 3 \times (Q_3 - Q_1); Q_1 - 1{,}5 \times (Q_3 - Q_1)]$ arba $[Q_1 + 1{,}5 \times (Q_3 - Q_1); Q_1 + 3 \times (Q_3 - Q_1))$.
- Išskirtis nustatoma, kai stebėjimas yra už išorinio barjero ribos, t. y. reikšmė $< Q_1 - 3 \times (Q_3 - Q_1)$ arba $> Q_1 + 3 \times (Q_3 - Q_1)$.

Išskirčių ieškojome darbuotojų skaičiui, pajamoms, išlaidoms, pelnui ir prieaugiui atsižvelgus į industrijos rūšį.

Tarp darbuotojų skaičiaus rodiklio buvo 27 išskirtys ir 23 sąlyginės išskirtys. Buvo rasta viena stipriai nuo kitų stebėjimų išsiskirianti reikšmė – 99 procentais didesnė nei mediana toje industrijos šakoje. Tai įmonė „Overviewparrot“ iš mažmeninės prekybos sektoriaus. Šioje bendrovėje dirba 7125 darbuotojai. Be šios išskirties dar buvo rasta 26 ir visos pašalintos iš duomenų imties.




*1 pav. Darbuotojų skaičiaus stačiakampės diagramos pagal industrijas prieš ir po išskirčių pašalinimo.*

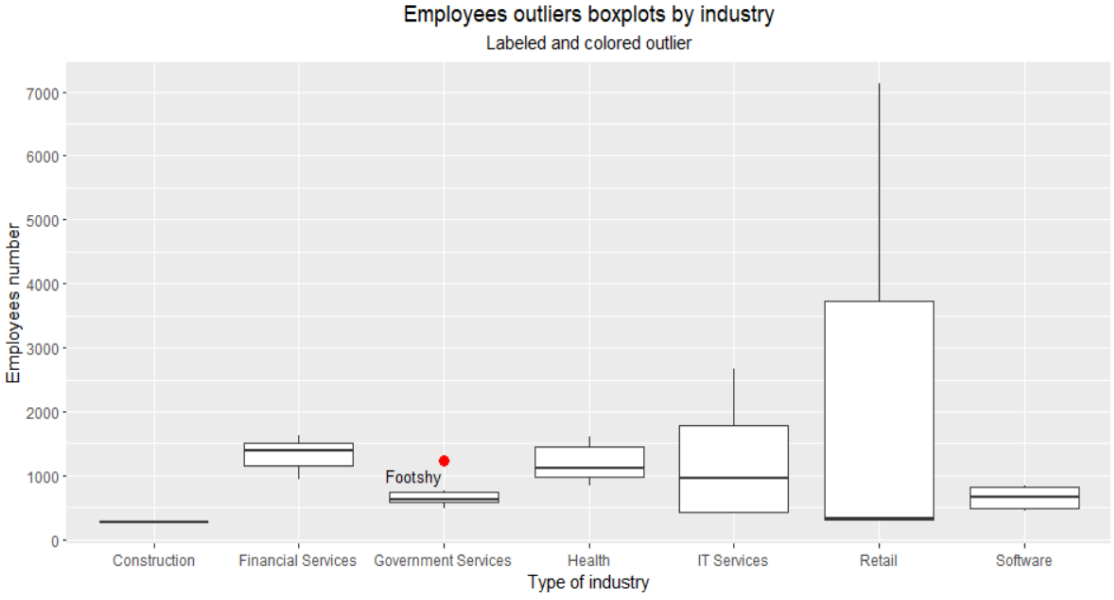
Iš (1 pav.) viršutiniame grafike matome, kaip „Overviewparrot“ įmonė stipriai išsiskiria iš visų likusių įmonių darbuotojų skaičiumi, o apatiniame grafike matome, kad, pašalinus išskirtis, darbuotojų skaičiaus reikšmių plotis ženkliai susiaurėjo, tačiau reikšmių išsibarstymas vis tiek nemažas. Pastarajame grafike matome pažymėtas sąlygines išskirtis, kurių randame kiekvienoje industrijoje. Be to, galime pasakyti, jog mažiausiai darbuotojų dirba statybų bei mažmeninės prekybos sektoriuose. Taip pat iš šių grafikų matome, jog sveikatos sektoriuje labiausiai varijuoja darbuotojų skaičius.

Toliau buvo atsižvelgta į sąlygines išskirtis. Pabandžius sąlygines išskirtis pakeisti 5 ir 95 kvantiliu atitinkamai, naujai atsirado 67 sąlyginės išskirtys, o, išbandžius sąlygines išskirtis pakeisti darbuotojų skaičiaus mediana, liko 10 sąlyginių išskirčių.


*2 pav. Darbuotojų skaičiaus sąlyginių išskirčių keitimo metodų palyginimai.*

(2 pav.) matome, kad panaudojus medianą keisti sąlyginėms išskirtims, tarpkvartilinis plotas yra siauresnis, o tai reiškia, kad duomenų išsibarstymas yra mažesnis, reikšmės koncentruojasi į siauresnį intervalą negu, kai sąlygines išskirtis keičiame į 5 ir 95 kvantilį. Nors reikšmių variacija išlieka panaši, naudojantis abu sąlyginių tikimybių keitimo metodus, pasirinkome jas modifikuoti mediana.

Išsamiau ištyrėme 36 taškus atsiskyrėlius, kuriuos radome, tirdamos darbuotojų skaičių. Šioje naujoje imtyje taip pat radome 1 išskirtį: „Footshy“ įmonė iš valstybinių paslaugų sektoriaus.




*3 pav. Išskirčių stačiakampių diagramos pagal industrijas*

Iš (3 pav.) matome, jog anksčiau minėta įmonė iš mažmeninės prekybos sektoriaus nebėra traktuojama kaip išskirtis.

*3 lentelė. Išskirčių industrijų dažniai*

| Industrija | Construction | Financial Services | Government Services | Health | IT Services | Retail | Software |
|---|---|---|---|---|---|---|---|
| Dažnis | 1 | 3 | 6 | 5 | 4 | 3 | 5 |

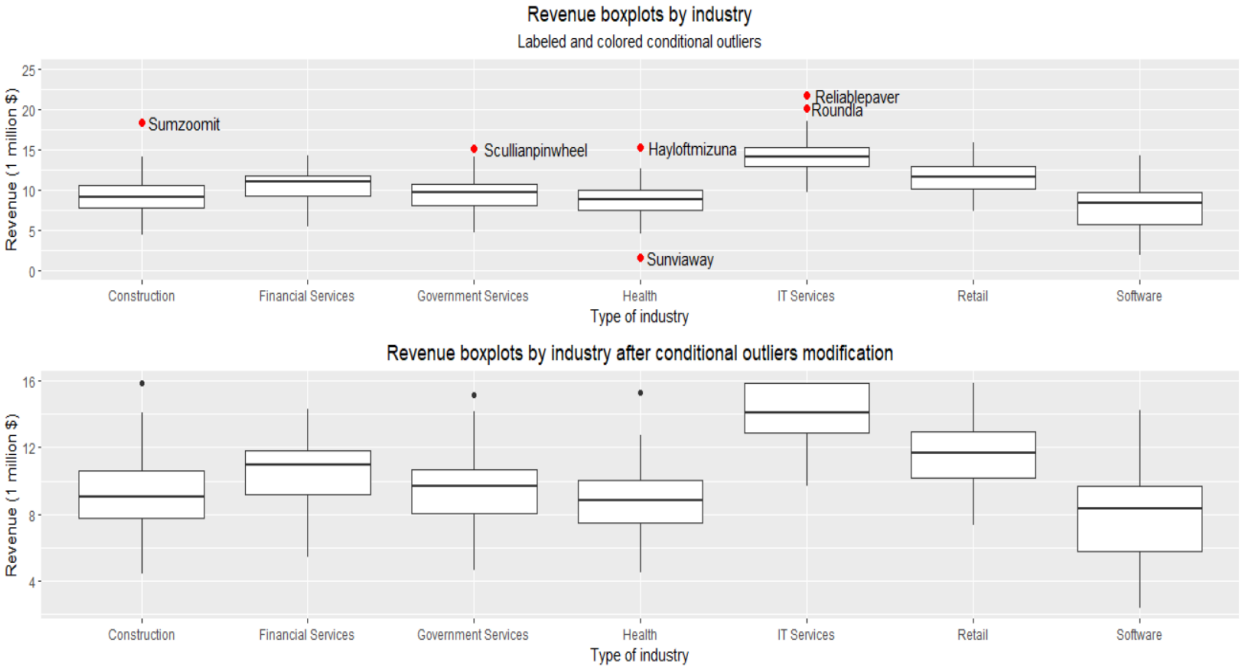
Taip pat iš (3 lentelė) matome, jog daugiausiai darbuotojų skaičiaus išskirčių radome valstybinių paslaugų sektoriuje, o mažiausiai statybų pramonėje.

*4 lentelė. Išskirčių imties darbuotojų skaičiaus charakteristikos*

| Min | 1 kvartilis | Mediana | Vidurkis | 3 kvartilis | Max |
|---|---|---|---|---|---|
| 272 | 487 | 818 | 1132 | 1306 | 7125 |

Matome (4 lentelė), kad išskirčių imtyje darbuotojų skaičiaus charakteristikos ženkliai padidėjo.

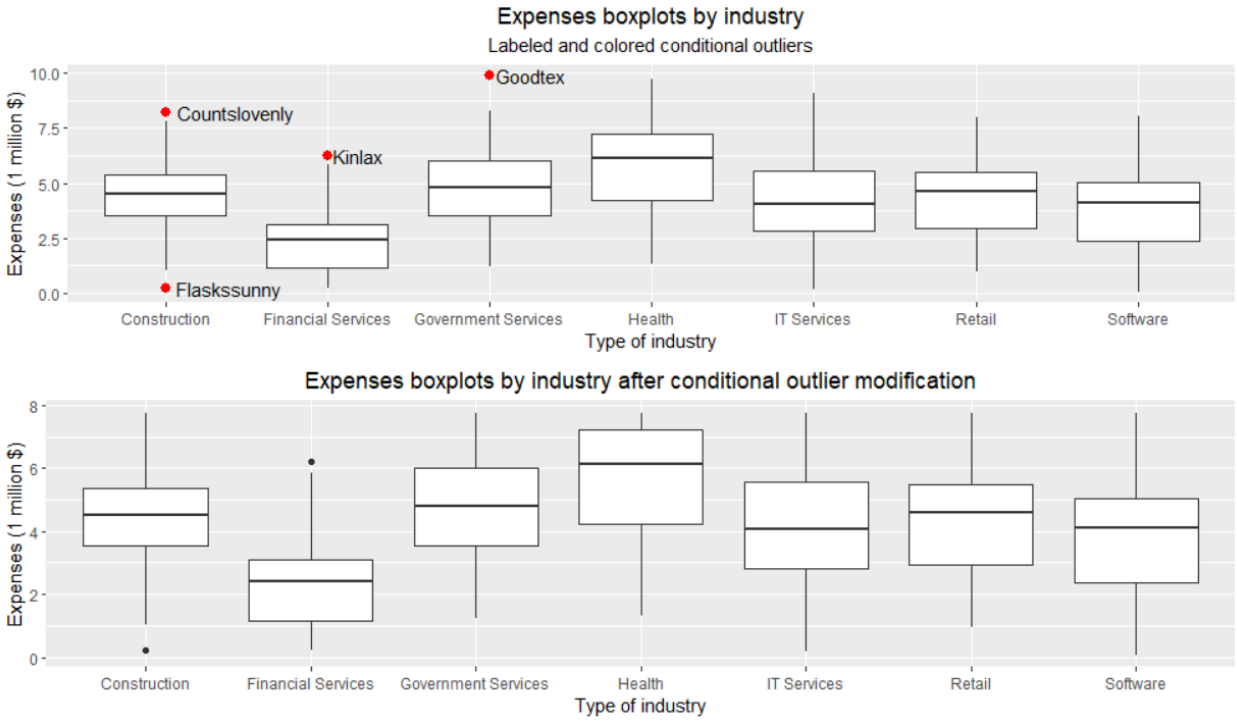
Pajamose buvo tik 6 sąlyginės tikimybės, jas keitėme į 5 ir 95 kvantilių reikšmes atitinkamai.




*4 pav. Pajamų stačiakampės diagramos pagal industrijas prieš ir po sąlyginių išskirčių modifikavimo.*

(4 pav.) viršutiniame grafike matome situaciją prieš sąlyginių išskirčių modifikavimą, o apatinis grafikas atspindi situaciją po pakeitimų ir galime pastebėti, jog variacija reikšmių sumažėjo nuo apytiksliai 25 mln. iki 16 mln. dolerių per visas industrijas. Taip pat galime iškart

pastebėti, jog IT sektoriaus pajamos yra didžiausios palyginus su kitais likusiais tiriamais sektoriais. Programinės įrangos sektoriuje pajamos yra vienos iš labiausiai varijuojančių.



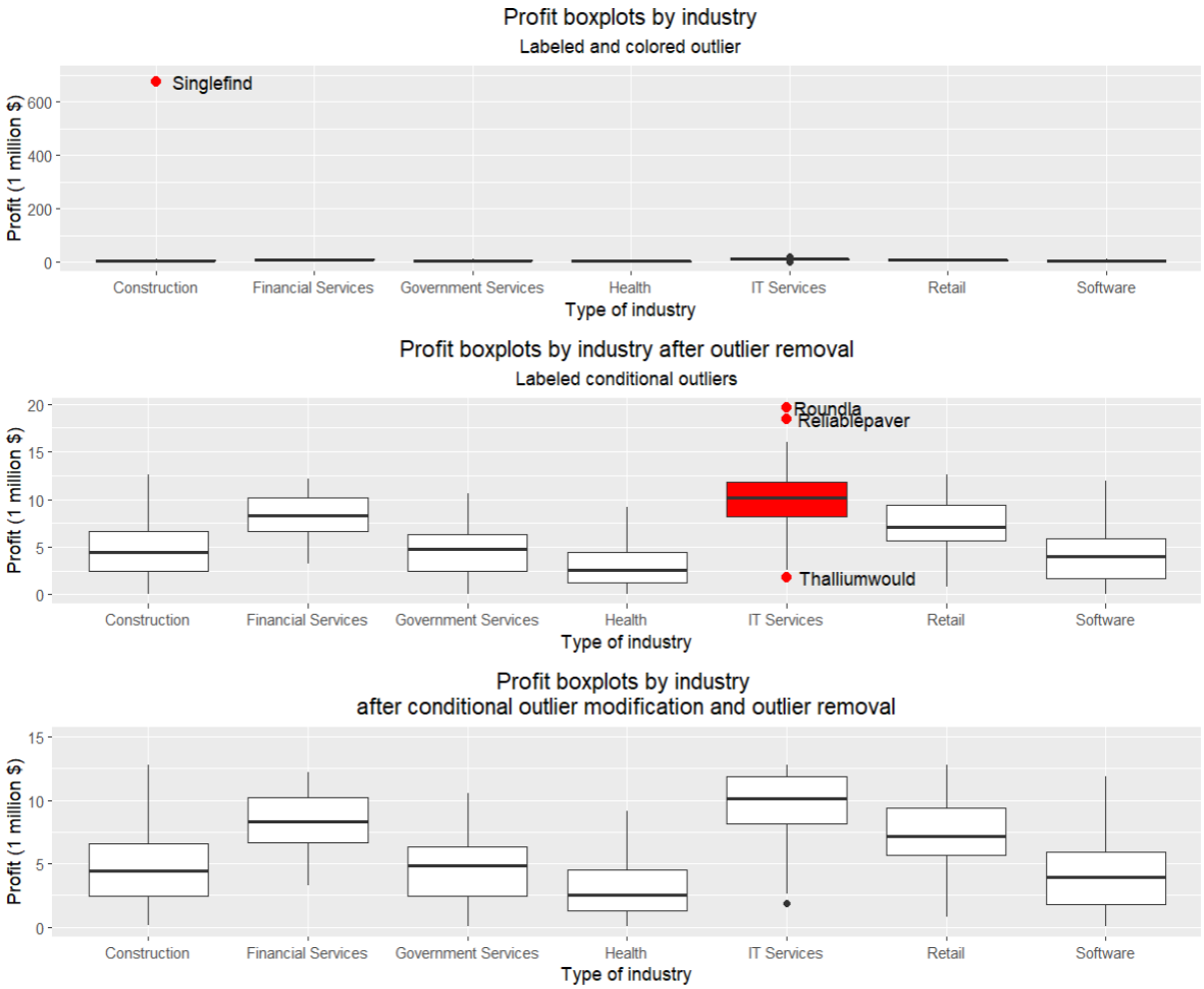




*5 pav. Išlaidų stačiakampės diagramos prieš ir po sąlyginių išskirčių modifikavimą*

Iš (5 pav.) matome, kad yra pažymėti 4 įtartini taškai, kurie galėtų būti laikomi sąlyginėmis išskirtimis išlaidose. Daugiausiai sąlyginių išskirčių pastebime statybų sektoriuje. Be to, apatiniame sutvarkytame grafike galime įžvelgti, jog finansiniame sektoriuje yra mažiausios išlaidos, o sveikatos sektoriuje - didžiausios.

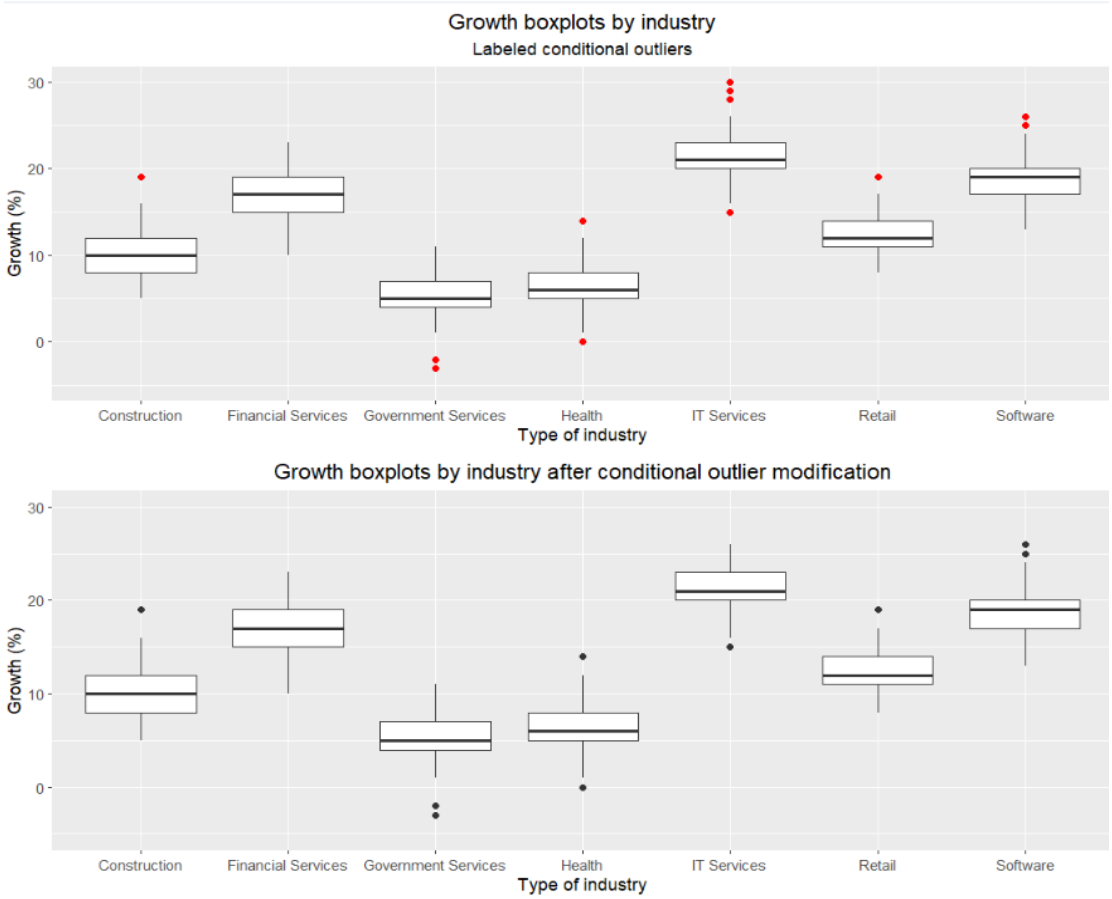
Nagrinėjant pelno reikšmes buvo rasta 1 išskirtis ir 3 sąlyginės. Įmonė „Singlefind“, atstovaujanti statybų sektorių, 9 procentiniais vienetais viršijo pelną skaičiuojant nuo medianos arba kitaip virš 1 mln. turėjo didesni pelną nei vidurinė reikšmė.




*6 pav. Pelno stačiakampės diagramos pagal industrijas prieš, po išskirčių pašalinimo ir po sąlyginių išskirčių modifikavimo.*

Iš (6 pav.) taip pat vizualiai pastebime aptartą išskirtį, ją pašalinus, matome, jog pelno variaciją sumažėjo nuo kažkur 700 mln. iki 20 mln. dolerių. Antrajame grafike iš panelės yra pažymėtos sąlyginės išskirtis. Matome, kad jas teturi IT įmonės, kurių yra 3. Jos buvo pakeistos į 5 ir 95 kvantilių reikšmes atitinkamai ir apatiniame grafike jau matome sutvarkytus pelno duomenis. Pastebime, kad modifikavus sąlygines išskirtis pelno variacija sumažėjo iki maždaug 13 mln. dolerių. Taip pat galime įžvelgti, jog IT paslaugų sektorius daugiausiai sugeneruoja pelno, o sveikatos - mažiausiai. Be to, statybų sektorius gali sugeneruoti įvairiausią, t. y. pelno išsibarstymas yra plačiausias iš visų sektorių.

Tiriant įmonių prieaugį radome 19 sąlyginių išskirčių ir jas keitėme į 5 ir 95 kvantilių reikšmes atitinkamai.




*7 pav. Prieaugio stačiakampės diagramos pagal industrijas.*

(7 pav.) viršutiniame grafike matome, jog visi sektoriai išskyrus finansinių paslaugų turėjo sąlyginių išskirčių. Pakeitus jas, pastebime, jog sėkmingiausiai vystytis sekėsi IT sektoriaus bei programinės įrangos sektoriams, lėčiausias augimas vyko valstybiniuose bei sveikatos sektoriuose.

Taigi, iš viso buvo rastos 28 išskirtys bei 55 sąlyginių išskirčių. Bendras likusių stebėjimų skaičius yra 457.

*15 lentelė. Pagrindinės charakteristikos prieš ir po išskirčių sutvarkymo.*

| Prieš išskirčių šalinimą | | | | | | |
|---|---|---|---|---|---|---|
| | vidurkis | mediana | min | max | 1 kvartilis | 3 kvartilis |
| Employees | 149 | 56 | 1 | 7125 | 28 | 126 |
| Revenue | 10,835 mln. | 10,647 mln. | 1,615 mln. | 21,810 mln. | 8,663 mln. | 13,097 mln. |
| Expenses | 4,297 mln. | 4,307 mln. | 0,071 mln. | 9,861 mln. | 2,758 mln. | 5,794 mln. |
| Profit | 7,891 mln. | 6,513 mln. | 0,012 mln. | 675,081 mln. | 3,315 mln. | 9,365 mln. |
| Growth | 14,42 | 16,00 | -3,00 | 30,00 | 8,00 | 20,00 |
| Po išskirčių šalinimo | | | | | | |
| | vidurkis | mediana | min | max | 1 kvartilis | 3 kvartilis |
| Employees | 52 | 50 | 1 | 140 | 26 | 67 |
| Revenue | 10,829 mln. | 10,645 mln. | 2,368 mln. | 15,882 mln. | 8,662 mln. | 13,107 mln. |
| Expenses | 4,263 mln. | 4,277 mln. | 0,071 mln. | 7,756 mln. | 2,755 mln. | 5,857 mln. |
| Profit | 6,490 mln. | 6,560 mln. | 0,012 mln. | 12,757 mln. | 3,336 mln. | 9,383 mln. |
| Growth | 14,51 | 16,00 | -3,00 | 26,00 | 8,00 | 20,00 |

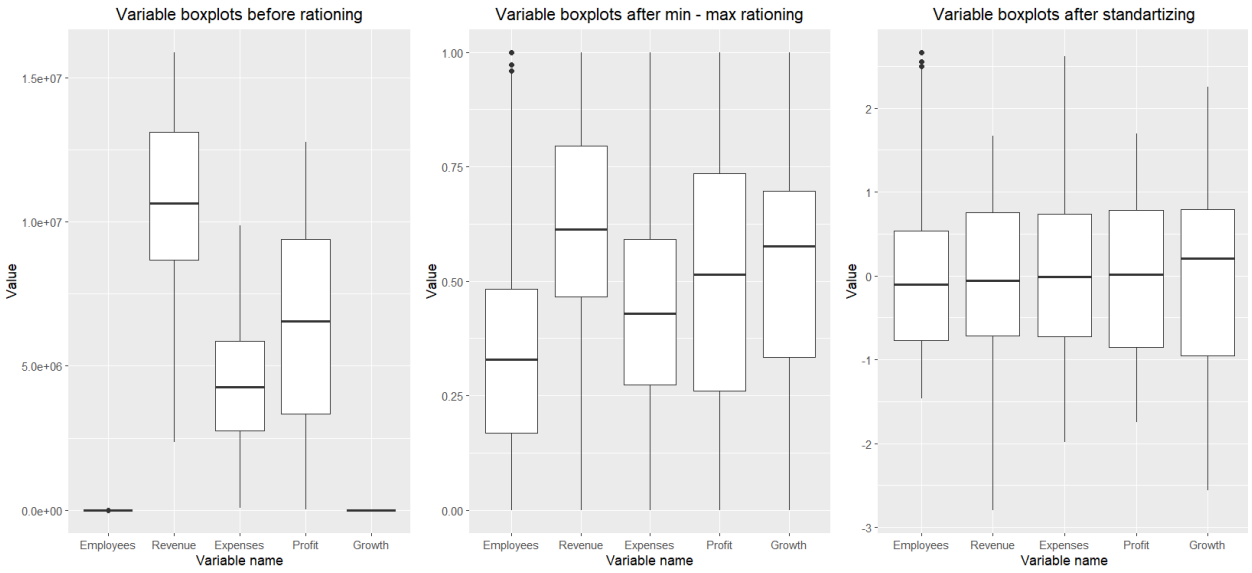
Iš (15 lentelė) matome, jog labiausiai pasikeitė maksimalios bei vidurkių reikšmės - didžiausią skirtumą matome pelne.

## Duomenų normavimas

Analizuojamos duomenų aibės reikšmės kinta skirtinguose intervaluose, todėl taikysime duomenų normavimą, kuris leidžia suvienodinti reikšmių mąstelius. Taikysime 2 normavimo būdus:

1. Normavimas pagal min – max metodą:
$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}.$$
2. Normavimas pagal vidurkį ir dispersiją:
$$x_{norm} = \frac{x - \bar{x}}{\sqrt{\delta^2}},$$

kur $\bar{x}$ – požymio vidurkis, $\delta^2$ - požymio dispersija.




*8 pav. Duomenys prieš normavimą, po min - max normavimo ir po standartizavimo*

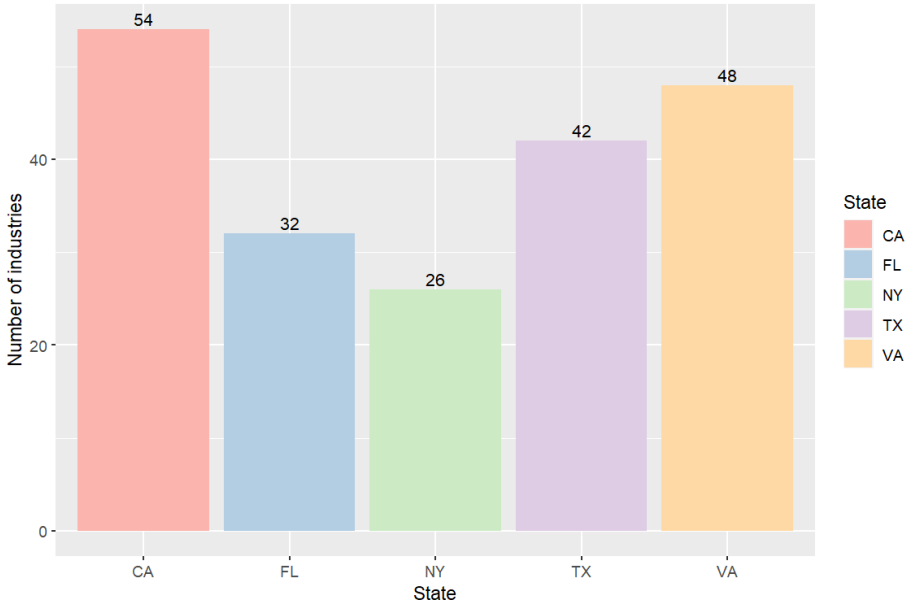
Pradinis kiekybinių duomenų aibės požymių pasiskirstymas pavaizduotas stačiakampe diagrama. Pakartotinai pavaizduotas pasiskirstymas atlikus abu anksčiau minėtus normavimo metodus. (8 pav.) Matome, kad po standartizavimo labiausiai suvienodėjo kintamųjų vidurkiai, o po min – max transformavimo supanašėjo reikšmių išsibarstymo plotis.

## Duomenų statistikos

Siekiant sužinoti, kaip pelnas, išlaidos ir darbuotojų skaičius keičiasi pagal pramonės šakas, paėmėme jau sutvarkytą duomenų aibę, kurioje užpildytos reikšmės, pašalintos išskirtys bei atliktas normavimas.


*9 pav. Pelno, darbuotojų skaičiaus ir išlaidų pasiskirstymas industrijose*

(9 pav.) matome, jog daugiausia išlaidų yra sveikatos sektoriuje, išlaidos stipriai išsiskiria. Mažiausios išlaidos - Finansų sektoriuje. Galime pastebėti, jog didžiausią darbuotojų skaičių turime Vyriausybės sektoriuje. Panašus darbuotojų skaičius yra sveikatos, IT bei programinės įrangos sektoriuose. Mažiausias darbuotojų skaičius – mažmeninės prekybos sektoriuje. Didžiausias pelnas – IT, toliau eina finansų sektorius, kuri pasižymi dar ir mažomis išlaidomis. Mažiausiai pelno - sveikatos sektoriuje, kurioje taip pat yra ir daugiausia išlaidų.




*10 pav. Daugiausiai įmonių turinčios valstijos.*

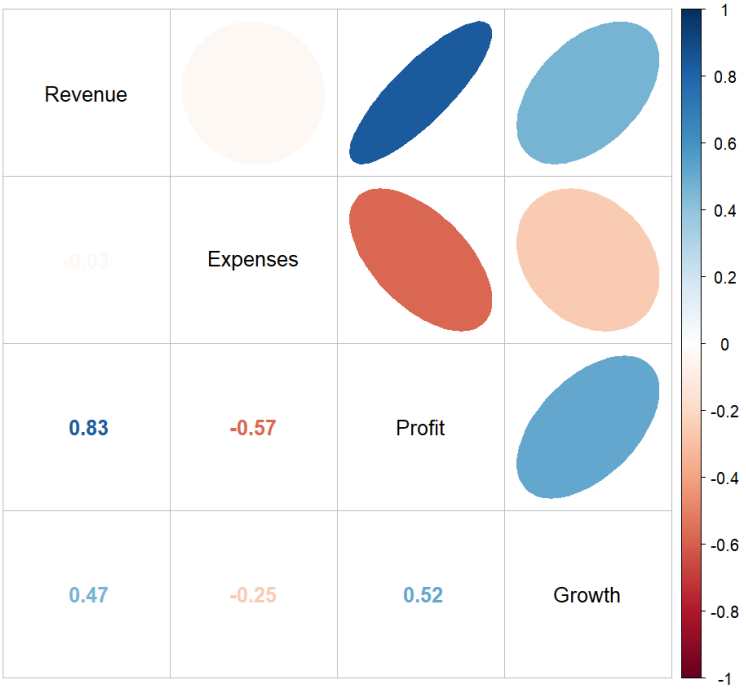
Iš (10 pav.) sužinojome, kad iš visų valstijų Kalifornijos, Vašingtono, Teksaso, Floridos ir Niujorko valtijos turi daugiausiai įmonių. Šias valstijas ištirsime detaliau – aiškinsimės labiausiai paplitusias pramonės šakas.

*16 lentelė. Įmonių skaičius pagal pramonės šakas valstijose.*

| | | |
|---|---|---|
| California | Software | 13 |
| | Health | 12 |
| | IT Services | 12 |
| Florida | Financial Services | 9 |
| | IT Services | 7 |
| | Construction | 4 |
| New York | IT Services | 8 |
| | Retail | 5 |
| | Construction | 4 |
| Texas | IT Services | 12 |
| | Health | 10 |
| | Financial Services | 7 |
| Virginia | IT Services | 21 |
| | Government Services | 18 |
| | Software | 5 |

Iš (16 lentelė) matome, jog beveik visose penkiose valstijose daugiausia įmonių yra iš tų pačių pramonės šakų. Tiek Niujorke, tiek Virdžinijoje bei Teksaso valstijose daugiausia yra IT įmonių. Kalifornijoje – programinės įrangos, o Floridoje – Finansų.

## Koreliacija




*11 pav. Finansinių rodiklių koreliacija.*

Patikrinus koreliaciją finansiniams rodikliams galima matyti iš (11 pav.), kad pajamų ir išlaidų sąryšis yra labai mažas ir artimas 0. Stipriausia koreliacija yra tarp pajamų ir pelno – 0,83, o išlaidos ir pelnas yra atvirkščiai proporcingi ir turi vidutinio stiprumo koreliaciją – 0,53. Taip pat galime pastebėti, jog išlaidos ir prieaugis turi atvirkštinę silpną koreliaciją.

# IŠVADOS

Iš viso buvo rasta 24 įrašai su bent viena praleista reikšme. Po faktinio užpildymo ir kur buvo galima finansinių rodiklių paskaičiavimo, įrašų su praleistomis reikšmėmis liko 12.

Iš viso buvo rastos 28 išskirtys bei 55 sąlyginių išskirčių. Bendras stebėjimų skaičius po duomenų išvalymo yra 457. Pašalinus išskirtis labiausiai pasikeitė maksimalios reikšmės kiekybinių kintamųjų reikšmės bei vidurkiai - didžiausia skirtumą matome pelne.

Mažiausiai darbuotojų dirba statybų bei mažmeninės prekybos sektoriuose, o sveikatos sektoriuje labiausiai varijuoja darbuotojų skaičius. IT sektoriaus pajamos yra didžiausios, daugiausiai sugeneruoja pelno bei jis sėkmingiausiai vystėsi kartu su programinės įrangos pramone, kurios pajamos yra vienos iš labiausiai varijuojančių. Finansiniame sektoriuje mažiausios išlaidos, o sveikatos pramonėje priešingai - didžiausios. Taip pat šis sektorius sugeneruoja mažiausiai pelno ir mažiausiai augo kartu su valstybinėmis paslaugomis teikiančiomis įmonėmis. Be to, statybų sektorius gali sugeneruoti įvairiausią pelną, t. y. pelno išsibarstymas yra plačiausias iš visų sektorių.

Daugiausiai įmonių turinčios valstijos - Kalifornija, Virdžinija bei Teksasas. Tiek Virdžinijoje, tiek Teksase IT sektorius pirmauja pagal įmonių skaičių valstijoje. Kalifornijoje daugiausia yra programinės įrangos sektoriaus įmonių ir per vieną įmonę, antroje vietoje yra sveikatos sektorius.

Stipriausia koreliacija pastebima tarp pajamų ir pelno, kurios koeficientas lygus 0,83, o silpniausias ryšys yra tarp pajamų ir išlaidų, kurio koreliacijos koeficientas beveik lygus 0. Išlaidos ir pelnas yra atvirkščiai proporcingi ir turi vidutinio stiprumo koreliaciją lygią 0,53.

# PRIEDAS

```
###################################
#duomenu nuskaitymas
###################################

duomenys <-
read.csv("C:/Users/ugneo/Downloads/Sample_Code_R/Duomenys/Future-500-
5.csv")

###############################
#susipazinimas su duomenimis, numeric ir factor priskyrimas
###############################

#-------------------------
#tarpus pakeiciame i NA
#-------------------------
duomenys <- replace(duomenys, duomenys =='', NA)


#-------------------------
# procentine praleistu reiksmiu dalis
#-------------------------------------
data_na <- duomenys[!complete.cases(duomenys),]
View(data_na)
# eilutciu kuriose nenurodyta valstija arba finansinis rodiklis, kuri
galima
# paskaiciuoti pasalinimas
data_na <- data_na[-c(3, 6, 8, 11, 12, 14, 17:20, 22, 24),]
View(data_na)

nrow(data_na)
nrow(data_na)/nrow(duomenys)*100
#suzinome, kiek turime stebejimu su trukstamomis reiksmemis


#--------------------------------
#kategoriju skaiciaus suradimas
#--------------------------------------
length(unique(duomenys$Name))
length(unique(duomenys$Industry))-1#del NA
length(unique(duomenys$State))-1#del NA
length(unique(duomenys$City))


#-----------------------------------------
#kategoriniams duomenims pridedame kategorijas
#ir skaitinius pasivertciame numeric
#-----------------------------------------------

str(duomenys)
```

```
duomenys$Industry<-as.factor(duomenys$Industry)
duomenys$State<-as.factor(duomenys$State)
duomenys$City<-as.factor(duomenys$City)

duomenys$Profit <- as.numeric(as.character(duomenys$Profit))
#salinam nereikalingus simbolius
duomenys$Expenses <- gsub(" Dollars","",duomenys$Expenses)
duomenys$Expenses <- gsub(",","",duomenys$Expenses)
#darom numeric
duomenys$Expenses <- as.numeric(as.character(duomenys$Expenses))

#salinam nereikalingus simbolius
duomenys$Revenue <- gsub("\\$","",duomenys$Revenue)
duomenys$Revenue <- gsub(",","",duomenys$Revenue)
#darom numeric
duomenys$Revenue <- as.numeric(as.character(duomenys$Revenue))

#salinam nereikalingus simbolius
duomenys$Growth <- gsub("\\%","",duomenys$Growth)
#darom numeric
duomenys$Growth <- as.numeric(as.character(duomenys$Growth))

str(duomenys)


################################
#pirmas vaizdas apie duomenis
################################
summary(duomenys)

#---------------------------
#Statistika pagal pramones sakas
#---------------------------
library(psych)
describeBy(duomenys$Inception,
           group = duomenys$Industry, digits = 4)

describeBy(duomenys$Employees,
           group = duomenys$Industry, digits = 4)

describeBy(duomenys$Revenue,
           group = duomenys$Industry, digits = 4)

describeBy(duomenys$Expenses,
           group = duomenys$Industry, digits = 4)

describeBy(duomenys$Profit,
           group = duomenys$Industry, digits = 4)

describeBy(duomenys$Growth,
           group = duomenys$Industry, digits = 4)
```

```
#####################################
#praleistos reiksmes
#####################################

#----------------------------
#pradinis reiksmiu uzpildymas
#----------------------------
library(dplyr)
# irasai su praleistomis reiksmemis
duomenys[!complete.cases(duomenys),]

# istriname irasus kuriuose nenurodyta industrija
duomenys <- duomenys[!is.na(duomenys$Industry),]
duomenys[!complete.cases(duomenys),]

# praleistu metu iraso tvarkymas - istriname
duomenys <- duomenys[!is.na(duomenys$Inception),]
duomenys[!complete.cases(duomenys),]

# valstiju prasleistu reiksmiu uzpildymas pagal miestus
duomenys[is.na(duomenys$State) &duomenys$City=="New York", "State"] <-
"NY"
duomenys[is.na(duomenys$State) &duomenys$City=="Newport Beach", "State"]
<- "CA"
duomenys[is.na(duomenys$State) &duomenys$City=="San Francisco", "State"]
<- "CA"
duomenys[is.na(duomenys$State) &duomenys$City=="Alpharetta", "State"] <-
"GA"
duomenys[is.na(duomenys$State) &duomenys$City=="Chicago", "State"] <-
"IL"
duomenys[!complete.cases(duomenys),]


#----------------------
#uzpildymas mediana
#----------------------

data_med <- duomenys

data_med[!complete.cases(data_med),]

# darbuotoju praleistu reiksmiu uzpildymas mediana pagal industrija

data_med$Employees <- replace(data_med$Employees, data_med$Industry ==
"Retail" & is.na(data_med$Employees) == T,
                              median(filter(data_med, data_med$Industry
== "Retail")$Employees, na.rm = T))

data_med$Employees <- replace(data_med$Employees, data_med$Industry ==
"Construction" & is.na(data_med$Employees) == T,
```

```
                              median(filter(data_med, data_med$Industry
== "Construction")$Employees, na.rm = T))

data_med$Employees <- replace(data_med$Employees, data_med$Industry ==
"Software" & is.na(data_med$Employees) == T,
                              median(filter(data_med, data_med$Industry
== "Software")$Employees, na.rm = T))

data_med$Employees <- replace(data_med$Employees, data_med$Industry ==
"Financial Services" & is.na(data_med$Employees) == T,
                              median(filter(data_med, data_med$Industry
== "Financial Services")$Employees, na.rm = T))

data_med[!complete.cases(data_med),]


# 8 ir 44 eilutes pasalinamos del bent 3 nenurodytu finansiniu rodikliu

data_med <- data_med[-c(8, 41),]
data_med[!complete.cases(data_med),]


# pajamu uzpildymas sudedant islaidas ir pelna

data_med[is.na(data_med$Revenue), "Revenue"] <-
data_med[is.na(data_med$Revenue), "Expenses"] +
  data_med[is.na(data_med$Revenue), "Profit"]

data_med[!complete.cases(data_med),]


# islaidos uzpildomos is pajamu atemus pelna

data_med[is.na(data_med$Expenses), "Expenses"] <-
data_med[is.na(data_med$Expenses), "Revenue"] -
  data_med[is.na(data_med$Expenses), "Profit"]

data_med[!complete.cases(data_med),]


# augimo praleistos reiksmes

data_med$Growth <- replace(data_med$Growth, data_med$Industry ==
"Software" & is.na(data_med$Growth) == T,
                           median(filter(data_med, data_med$Industry ==
"Software")$Growth, na.rm = T))

data_med[!complete.cases(data_med),]


#----------------------
#reiksmiu uzpildymas vidurkiu
```

```
# ---------------

data_vid <- duomenys

# darbuotoju praleistu reiksmiu uzpildymas vidurkiu pagal industrija

data_vid$Employees <- replace(data_vid$Employees, data_vid$Industry ==
"Retail" & is.na(data_vid$Employees) == T,
                              mean(filter(data_vid, data_vid$Industry ==
"Retail")$Employees, na.rm = T))

data_vid$Employees <- replace(data_vid$Employees, data_vid$Industry ==
"Construction" & is.na(data_vid$Employees) == T,
                              mean(filter(data_vid, data_vid$Industry ==
"Construction")$Employees, na.rm = T))

data_vid$Employees <- replace(data_vid$Employees, data_vid$Industry ==
"Software" & is.na(data_vid$Employees) == T,
                              mean(filter(data_vid, data_vid$Industry ==
"Software")$Employees, na.rm = T))

data_vid$Employees <- replace(data_vid$Employees, data_vid$Industry ==
"Financial Services" & is.na(data_vid$Employees) == T,
                              mean(filter(data_vid, data_vid$Industry ==
"Financial Services")$Employees, na.rm = T))

data_vid[!complete.cases(data_vid),]

# 8 ir 44 eilutes pasalinamos del bent 3 nenurodytu finansiniu rodikliu

data_vid <- data_vid[-c(8, 41),]
data_vid[!complete.cases(data_vid),]

# pajamu uzpildymas sudedant islaidas ir pelna

data_vid[is.na(data_vid$Revenue), "Revenue"] <-
data_vid[is.na(data_vid$Revenue), "Expenses"] +
  data_vid[is.na(data_vid$Revenue), "Profit"]

data_vid[!complete.cases(data_vid),]

# islaidos uzpildomos is pajamu atemus pelna

data_vid[is.na(data_vid$Expenses), "Expenses"] <-
data_vid[is.na(data_vid$Expenses), "Revenue"] -
  data_vid[is.na(data_vid$Expenses), "Profit"]

data_vid[!complete.cases(data_vid),]

# augimo praleistos reiksmes uzpildomos vidurkiu pagal industrija
```

```
data_vid$Growth <- replace(data_vid$Growth, data_vid$Industry ==
"Software" & is.na(data_vid$Growth) == T,
                           mean(filter(data_vid, data_vid$Industry ==
"Software")$Growth, na.rm = T))

data_vid[!complete.cases(data_vid),]


#---------------------
#reiksmiu uzpildymas moda
# ---------------

mode <- function(x) {
  return(as.numeric(names(which.max(table(x)))))
}

data_mod <- duomenys

# darbuotoju praleistu reiksmiu uzpildymas moda pagal industrija
library(dplyr)

data_mod$Employees <- replace(data_mod$Employees, data_mod$Industry ==
"Retail" & is.na(data_mod$Employees) == T,
                              mode(filter(data_mod, data_mod$Industry ==
"Retail")$Employees))

data_mod$Employees <- replace(data_mod$Employees, data_mod$Industry ==
"Construction" & is.na(data_mod$Employees) == T,
                              mode(filter(data_mod, data_mod$Industry ==
"Construction")$Employees))

data_mod$Employees <- replace(data_mod$Employees, data_mod$Industry ==
"Software" & is.na(data_mod$Employees) == T,
                              mode(filter(data_mod, data_mod$Industry ==
"Software")$Employees))

data_mod$Employees <- replace(data_mod$Employees, data_mod$Industry ==
"Financial Services" & is.na(data_mod$Employees) == T,
                              mode(filter(data_mod, data_mod$Industry ==
"Financial Services")$Employees))

data_mod[!complete.cases(data_mod),]

# 8 ir 44 eilutes pasalinamos del bent 3 nenurodytu finansiniu rodikliu

data_mod <- data_mod[-c(8, 41),]
data_mod[!complete.cases(data_mod),]

# pajamu uzpildymas sudedant islaidas ir pelna

data_mod[is.na(data_mod$Revenue), "Revenue"] <-
data_mod[is.na(data_mod$Revenue), "Expenses"] +
```

```
  data_mod[is.na(data_mod$Revenue), "Profit"]

data_mod[!complete.cases(data_mod),]

# islaidos uzpildomos is pajamu atemus pelna

data_mod[is.na(data_mod$Expenses), "Expenses"] <-
data_mod[is.na(data_mod$Expenses), "Revenue"] -
  data_mod[is.na(data_mod$Expenses), "Profit"]

data_mod[!complete.cases(data_mod),]

# augimo praleistos reiksmes

data_mod$Growth <- replace(data_mod$Growth, data_mod$Industry ==
"Software" & is.na(data_mod$Growth) == T,
                           mode(filter(data_mod, data_mod$Industry ==
"Software")$Growth))

data_mod[!complete.cases(data_mod),]



#-------------------------------
#statistika pagal uzpildymo metodus
# ------------------ ---------

med<-summary(data_med)
vid<-summary(data_vid)
mod<-summary(data_mod)
med

med[,c(5,8:11)]
vid[,c(5,8:11)]
mod[,c(5,8:11)]


##########################################
#tiriame isskirtis
##########################################
library(ggplot2)
library(dplyr)
summ<-data.frame(summary(data_med))
duomenys<-data_med
mean(is.na(duomenys))#=>nebera praleistu reiksmiu, dirbame su duomenimis,
kur praleistos reiksmes uzpildytos su mediana

#-----------------------------------------------------------------
#apsirasome isskirciu ir salyginiu isskirciu skaiciavimo taisykles
#-----------------------------------------------------------------
isskirtis <- function(x) {
```

```
  return(x <= quantile(x, .25) - 3*IQR(x) | x >= quantile(x, .75) +
3*IQR(x))
}

salygine_isskirtis<-function(x){
  return(((x <= quantile(x, .25) - 1.5*IQR(x))&(x > quantile(x, .25) -
3*IQR(x))) | (x >= quantile(x, .75) + 1.5*IQR(x))&
           (x < quantile(x, .75) + 3*IQR(x)))
}


#------------------------------
#susikuriame atskiras duomenu imtis ir susizymime tai isskirtis ar
salygine isskirtis
#----------------------------------

duomenys_darbuotojai <- duomenys %>%
  group_by(Industry) %>%
  mutate(outlier = ifelse(isskirtis(Employees), Name, NA)) %>%
  mutate(salygine_outlier = ifelse(salygine_isskirtis(Employees), Name,
NA))

duomenys_pajamos <- duomenys %>%
  group_by(Industry) %>%
  mutate(outlier = ifelse(isskirtis(Revenue), Name, NA))%>%
  mutate(salygine_outlier = ifelse(salygine_isskirtis(Revenue), Name,
NA))

duomenys_islaidos <- duomenys %>%
  group_by(Industry) %>%
  mutate(outlier = ifelse(isskirtis(Expenses), Name, NA)) %>%
  mutate(salygine_outlier = ifelse(salygine_isskirtis(Expenses), Name,
NA))

duomenys_pelnas <- duomenys %>%
  group_by(Industry) %>%
  mutate(outlier = ifelse(isskirtis(Profit), Name, NA))%>%
  mutate(salygine_outlier = ifelse(salygine_isskirtis(Profit), Name, NA))

duomenys_prieaugis <- duomenys %>%
  group_by(Industry) %>%
  mutate(outlier = ifelse(isskirtis(Growth), Name, NA))%>%
  mutate(salygine_outlier = ifelse(salygine_isskirtis(Growth), Name, NA))

#------------------------------
#Nagrinesime darbuotoju skaiciu
#------------------------------

#pasiziurime, ar turime isskirciu
mean(is.na(duomenys_darbuotojai$outlier)) #yra isskirciu

#suskaiciuojame, kiek ju turime
```

```
sum(!is.na(duomenys_darbuotojai$outlier)) #27 isskirtys
sum(!is.na(duomenys_darbuotojai$salygine_outlier)) #23 salygines
isskirtys

#nusibraizome staciakampes diagramas paziureti vaizdiskai isskirtis
g1<-ggplot(duomenys_darbuotojai, aes(x=Industry, y=Employees)) +
  geom_boxplot() + ggtitle("Employees number boxplots by industry") +
  xlab("Type of industry") + ylab("Employees number") +
ggrepel::geom_text_repel(aes(label = outlier), size = 3.5)  +
  scale_y_continuous(breaks = c(0, seq(0, 8000, 2000))) +
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
element_text(hjust = 0.5)) + labs(subtitle = "Labeled outliers")
g1


#Matome, kad yra viena didele isskirtis retail sektoriuje, suzinome apie
ja
retail<-duomenys_darbuotojai[duomenys_darbuotojai$Industry == "Retail",]
retail_isskirtis<-retail[which.max(retail$Employees),]
retail_isskirtis

#kiek virsijo mediana
retail_isskirtis$Employees-median(duomenys_darbuotojai$Employees)#7069
#kiek procentais
100-
((100*median(duomenys_darbuotojai$Employees))/retail_isskirtis$Employees)
#99 proc

duomenys_darbuotojai$max_outlier<-NA
duomenys_darbuotojai$max_outlier[duomenys_darbuotojai$Name ==
retail_isskirtis$Name]<-retail_isskirtis$Name


#nusibraizome staciakampes diagramas paziureti vaizdiskai isskirtis, cia
bus pazymeta didziausioji isskirtis
g2<-ggplot(duomenys_darbuotojai, aes(x=Industry, y=Employees)) +
  geom_boxplot()+ geom_point(data = subset(duomenys_darbuotojai,
max_outlier != "NA"),
                             aes(x = Industry, y  = Employees), size = 2,
color = "red")+
  geom_text(aes(label=max_outlier), hjust=-.2) + ggtitle("Employees
number boxplots by industry") +
  xlab("Type of industry") + ylab("Employees number") +
  scale_y_continuous(breaks = c(0, seq(1000, 7000, 1000))) +
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
element_text(hjust = 0.5)) +
  labs(subtitle = "Labeled and colored biggest outlier")

g2

#sudedame isskirtis i atskira duomenu masyva
```

```
duomenys_isskirtys<-subset(duomenys_darbuotojai,
duomenys_darbuotojai$outlier != "NA")


#issamiau panagrinesime isskirtis
#--------------------------------
duomenys_darbuotoju_isskirtys <- duomenys_isskirtys %>%
  group_by(Industry) %>%
  mutate(outlier = ifelse(isskirtis(Employees), Name, NA)) %>%
  mutate(salygine_outlier = ifelse(salygine_isskirtis(Employees), Name,
NA))

sum(!is.na(duomenys_darbuotoju_isskirtys$outlier)) #2 isskirtys
sum(!is.na(duomenys_darbuotoju_isskirtys$salygine_outlier))#salyginiu
nera

duomenys_darbuotoju_isskirtys$outlier[duomenys_darbuotoju_isskirtys$Indus
try == "Construction"]<-NA

ggplot(duomenys_darbuotoju_isskirtys, aes(x=Industry, y=Employees)) +
  geom_boxplot() + geom_point(data =
subset(duomenys_darbuotoju_isskirtys, outlier != "NA"),
                              aes(x = Industry, y  = Employees), size =
3, color = "red") +
  ggrepel::geom_text_repel(aes(label = outlier), size = 3.5)  +
ggtitle("Employees outliers boxplots by industry") +
  xlab("Type of industry") + ylab("Employees number") +
  scale_y_continuous(breaks = c(0, seq(1000, 7000, 1000))) +
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
element_text(hjust = 0.5)) +
  labs(subtitle = "Labeled and colored outlier")
table(duomenys_darbuotoju_isskirtys$Industry)

summary(duomenys_darbuotoju_isskirtys)

#---------------------------
#pasaliname visas isskirtis
duomenys_darbuotojai<-filter(duomenys_darbuotojai, is.na(outlier)==T)

#pasiziurime staciakampes diagramas
g3<-ggplot(duomenys_darbuotojai, aes(x=Industry, y=Employees)) +
  geom_boxplot() + ggtitle("Employees number boxplots by industry after
outliers removal") +
  xlab("Type of industry") + ylab("Employees number") +
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
element_text(hjust = 0.5)) +
  labs(subtitle = "Colored conditional outliers")  +
  geom_point(data = subset(duomenys_darbuotojai, salygine_outlier !=
"NA"),
             aes(x = Industry, y  = Employees), size = 1.5, color =
"red")
```

```
g3
#isskirciu nebera, taciau yra salyginiu isskirciu

#panele pries ir po isskirciu pasalinimo
gridExtra::grid.arrange(g2, g3, nrow = 2)

duomenys_darb_2<-duomenys_darbuotojai

#pakeiciame salygines isskirtis i 5 kvantili arba 95
#-------------------------------------------------
caps <- quantile(duomenys_darbuotojai$Employees, probs=c(.05, .95))

duomenys_darbuotojai$Employees[duomenys_darbuotojai$Employees <
quantile(duomenys_darbuotojai$Employees, .25) -
1.5*IQR(duomenys_darbuotojai$Employees)] <- caps[1]
duomenys_darbuotojai$Employees[duomenys_darbuotojai$Employees >
quantile(duomenys_darbuotojai$Employees, .25) +
1.5*IQR(duomenys_darbuotojai$Employees)] <- caps[2]

sum(isskirtis(duomenys_darbuotojai$Employees))
sum(salygine_isskirtis(duomenys_darbuotojai$Employees))#salyginiu 67

#isskaido pagal industrijas
g4<-ggplot(duomenys_darbuotojai, aes(x=Industry, y=Employees)) +
  geom_boxplot() + ggtitle("Employees number boxplots by industry when 5
and 95 quantiles for conditional outliers") +
  xlab("Type of industry") + ylab("Employees number") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits = c(0,150)) + labs(subtitle = "Colored
conditional outliers")  +
  geom_point(data = subset(duomenys_darbuotojai, salygine_outlier !=
"NA"),
             aes(x = Industry, y  = Employees), size = 1.5, color =
"red")

g4

#bandom keisti su mediana
#----------------------------------------
duomenys_darb_2$Employees[duomenys_darb_2$Employees <
quantile(duomenys_darb_2$Employees, .25) -
1.5*IQR(duomenys_darb_2$Employees)] <- median(duomenys_darb_2$Employees)
duomenys_darb_2$Employees[duomenys_darb_2$Employees >
quantile(duomenys_darb_2$Employees, .25) +
1.5*IQR(duomenys_darb_2$Employees)] <- median(duomenys_darb_2$Employees)



sum(isskirtis(duomenys_darb_2$Employees))
sum(salygine_isskirtis(duomenys_darb_2$Employees))#salyginiu 10
#isskaido pagal industrijas
g5<-ggplot(duomenys_darb_2, aes(x=Industry, y=Employees)) +
```

```
  geom_boxplot()+ ggtitle("Employees number boxplots when using median
for conditional outliers") +
  xlab("Type of industry") + ylab("Employees number") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits = c(0,150)) + labs(subtitle = "Colored
conditional outliers")  +
  geom_point(data = subset(duomenys_darb_2, salygine_outlier != "NA"),
             aes(x = Industry, y  = Employees), size = 1.5, color =
"red")

g5

gridExtra::grid.arrange(g4, g5, nrow = 1)

#paziurime vaizdiskai, kuris metodas geresnis
#-------------------------------------------
boxplot(duomenys_darb_2$Employees, duomenys_darbuotojai$Employees,
ylim=c(0, 150), ylab = "Employees number", main = "Employees number
boxplot across all industries",
        xlab = "Methods of modifying conditional outliers",
        names = c("Using median", "Using 5 and 95 quantiles"))


#pasirenkame medianos metoda, nes sklaida mazesne bei maziau paciu
salyginiu isskirciu


#-------------------------------
#pajamos
#-----------------------------
mean(is.na(duomenys_pajamos$outlier)) #=>1=> nera tikruju isskirciu
sum(!is.na(duomenys_pajamos$outlier)) #0
sum(!is.na(duomenys_pajamos$salygine_outlier))#6 salygines isskirtys

#sukuriame nauja stulpeli, kuriame pajamos milijonais
duomenys_pajamos$Revenue1mln<-duomenys_pajamos$Revenue/1000000

g6<-ggplot(duomenys_pajamos, aes(x=Industry, y=Revenue1mln)) +
  geom_boxplot() + ggtitle("Revenue boxplots by industry") +
  xlab("Type of industry") + ylab("Revenue (1 million $)") +
  scale_y_continuous(limits=c(0, 25))  +
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
element_text(hjust = 0.5)) +
  geom_text(aes(label=salygine_outlier), hjust= -0.1)  + labs(subtitle =
"Labeled and colored conditional outliers")+
  geom_point(data = subset(duomenys_pajamos, salygine_outlier != "NA"),
             aes(x = Industry, y  = Revenue1mln), size = 2, color =
"red")

g6

#tvarkysime salygines isskirtis, jas pakeisime i 5 arba 95 kvantili
```

```
caps <- quantile(duomenys_pajamos$Revenue, probs=c(.05, .95))

duomenys_pajamos$Revenue[duomenys_pajamos$Revenue <
quantile(duomenys_pajamos$Revenue, .25) -
1.5*IQR(duomenys_pajamos$Revenue)] <- caps[1]
duomenys_pajamos$Revenue[duomenys_pajamos$Revenue >
quantile(duomenys_pajamos$Revenue, .25) +
1.5*IQR(duomenys_pajamos$Revenue)] <- caps[2]

sum(isskirtis(duomenys_pajamos$Revenue))
sum(salygine_isskirtis(duomenys_pajamos$Revenue))
#nebeliko jokiu isskirciu nei salyginiu

#sukuriame nauja stulpeli, kuriame pajamos milijonais
duomenys_pajamos$Revenue1mln<-duomenys_pajamos$Revenue/1000000

#dar nusibraizome keleta boxplot diagramu pasiziurejimui vizualiam
#cia isskaido pagal industrijas
g7<-ggplot(duomenys_pajamos, aes(x=Industry, y=Revenue1mln)) +
  geom_boxplot()+ggtitle("Revenue boxplots by industry after conditional
outliers modification") +
  xlab("Type of industry") + ylab("Revenue (1 million $)")+
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(breaks=c(0, 4,8,12,16))

g7

#pries ir po salyginiu isskirciu pasalinimo
gridExtra::grid.arrange(g6, g7, nrow = 2)

#---------------------------------------
#islaidos
#---------------------------------------

mean(is.na(duomenys_islaidos$outlier)) #=>1=> nera tikruju isskirciu
sum(!is.na(duomenys_islaidos$salygine_outlier))#4 salyginiu isskirciu

#sukuriame nauja stulpeli, kuriame islaidos milijonais
duomenys_islaidos$Expenses1mln<-duomenys_islaidos$Expenses/1000000

#vizualiai pasiziurime boxplotus
#cia isskaido pagal industrijas
g8<-ggplot(duomenys_islaidos, aes(x=Industry, y=Expenses1mln)) +
  geom_boxplot()+ggtitle("Expenses boxplots by industry") +
  xlab("Type of industry") + ylab("Expenses (1 million $)")+
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
element_text(hjust = 0.5)) +
  geom_text(aes(label=salygine_outlier), hjust= -0.1)  + labs(subtitle =
"Labeled and colored conditional outliers")+
  geom_point(data = subset(duomenys_islaidos, salygine_outlier != "NA"),
             aes(x = Industry, y  = Expenses1mln), size = 2.5, color =
"red")
```

```
g8


#tvarkysime salygines isskirtis, jas pakeisime i 5 arba 95 kvantili
caps <- quantile(duomenys_islaidos$Expenses, probs=c(.05, .95))

duomenys_islaidos$Expenses[duomenys_islaidos$Expenses <
quantile(duomenys_islaidos$Expenses, .25) -
1.5*IQR(duomenys_islaidos$Expenses)] <- caps[1]
duomenys_islaidos$Expenses[duomenys_islaidos$Expenses >
quantile(duomenys_islaidos$Expenses, .25) +
1.5*IQR(duomenys_islaidos$Expenses)] <- caps[2]

sum(isskirtis(duomenys_islaidos$Expenses))
sum(salygine_isskirtis(duomenys_islaidos$Expenses))
#nebeliko jokiu isskirciu nei salyginiu

#sukuriame nauja stulpeli, kuriame islaidos milijonais
duomenys_islaidos$Expenses1mln<-duomenys_islaidos$Expenses/1000000

#pasalinus salygines isskirtis
#vizualiai pasiziurime boxplotus
#cia isskaido pagal industrijas
g8_2<-ggplot(duomenys_islaidos, aes(x=Industry, y=Expenses1mln)) +
  geom_boxplot()+ggtitle("Expenses boxplots by industry after conditional
outlier modification") +
  xlab("Type of industry") + ylab("Expenses (1 million $)")+
  theme(plot.title = element_text(hjust = 0.5))

g8_2


gridExtra::grid.arrange(g8, g8_2, nrow = 2)
#-----------------------------------
#pelnas
#-----------------------------------

mean(is.na(duomenys_pelnas$outlier)) #turime isskirciu
sum(!is.na(duomenys_pelnas$outlier)) #1 isskirtis
sum(!is.na(duomenys_pelnas$salygine_outlier)) #3 salygines isskirtys

#sukuriame nauja stulpeli, kuriame pelnas milijonais
duomenys_pelnas$Profit1mln<-duomenys_pelnas$Profit/1000000

#pasibraizome boxplot
g9<-ggplot(duomenys_pelnas, aes(x=Industry, y=Profit1mln)) +
  geom_boxplot() +
  geom_text(aes(label=outlier), hjust = -0.2) + ggtitle("Profit boxplots
by industry") +
  xlab("Type of industry") + ylab("Profit (1 million $)")+
```

```
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
element_text(hjust = 0.5)) +
  labs(subtitle = "Labeled and colored outlier") +
scale_y_continuous(limits=c(0, 700)) +
  geom_point(data = subset(duomenys_pelnas, outlier != "NA"),
             aes(x = Industry, y  = Profit1mln), size = 2.5, color =
"red")

g9
#matome, viena labai didele isskirti statybu sektoriuje

#pasitikrinsime, ar sis stebejimas jau yra isskirciu lenteleje
pelno_isskirtis<-duomenys_pelnas[duomenys_pelnas$Industry ==
"Construction",]
pelno_isskirtis<-pelno_isskirtis[which.max(pelno_isskirtis$Profit),]
pelno_isskirtis<-pelno_isskirtis[,1:13]
pelno_isskirtis$maxoutlier<-NA
pelno_isskirtis
#doleriais
pelno_isskirtis$Revenue-median(duomenys_pelnas$Revenue) #1078448
#procentais kiek virsijo
100-((100*median(duomenys_pelnas$Revenue))/pelno_isskirtis$Revenue)#9
proc

library(plyr)

match_df(duomenys_isskirtys, pelno_isskirtis, on="Name")
#nera tokio dar stebejimo isskirtyse

#sudedame isskirtis i atskira duomenu masyva
duomenys_isskirtys[nrow(duomenys_isskirtys) + 1,] <- pelno_isskirtis

#Matome, kad yra viena didele isskirtis construction imoneje, ja
pasaliname
duomenys_pelnas<-subset(duomenys_pelnas, Name != pelno_isskirtis$Name)

duomenys_pelnas<-mutate( duomenys_pelnas, type=ifelse(Industry=="IT
Services","Highlighted","Normal"))

g10<-ggplot(duomenys_pelnas, aes(x=Industry, y=Profit1mln, fill = type))
+
  geom_boxplot() +
  geom_text(aes(label=salygine_outlier), hjust = -0.1)+ ggtitle("Profit
boxplots by industry after outlier removal") +
  xlab("Type of industry") + ylab("Profit (1 million $)")+
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
element_text(hjust=0.5))+
  labs(subtitle = "Labeled conditional outliers") +
  geom_point(data = subset(duomenys_pelnas, salygine_outlier != "NA"),
             aes(x = Industry, y  = Profit1mln), size = 2.5, color =
"red")+
  scale_fill_manual(values=c("red", "white")) +
```

```
  theme(legend.position = "none")

g10
#pazymetos salygines isskirtys

gridExtra::grid.arrange(g9, g10, nrow=2)

#dabar tvarkysime salygines isskirtis, jas pakeisi i 5 ir 95 kvantili

caps <- quantile(duomenys_pelnas$Profit, probs=c(.05, .95))

duomenys_pelnas$Profit[duomenys_pelnas$Profit <
quantile(duomenys_pelnas$Profit, .25) - 1.5*IQR(duomenys_pelnas$Profit)]
<- caps[1]
duomenys_pelnas$Profit[duomenys_pelnas$Profit >
quantile(duomenys_pelnas$Profit, .25) + 1.5*IQR(duomenys_pelnas$Profit)]
<- caps[2]

#patikriname, ar liko salyginiu ar tikru isskirciu
sum(isskirtis(duomenys_pelnas$Profit))
sum(salygine_isskirtis(duomenys_pelnas$Profit))
#neliko

#sukuriame nauja stulpeli, kuriame pelnas butu milijonais
duomenys_pelnas$Profit1mln<-duomenys_pelnas$Profit/1000000

#nubreziame boxplot
g11<-ggplot(duomenys_pelnas, aes(x=Industry, y=Profit1mln)) +
  geom_boxplot() + ggtitle("Profit boxplots by industry \nafter
conditional outlier modification and outlier removal") +
  xlab("Type of industry") + ylab("Profit (1 million $)")+
  theme(plot.title = element_text(hjust = 0.5)) +
scale_y_continuous(limits=c(0, 15))

g11
gridExtra::grid.arrange(g9, g10, g11, nrow=3)

#-------------------------------------
#prieaugis
#-------------------------------------
mean(is.na(duomenys_prieaugis$outlier)) #=>1=>nera isskirciu
sum(!is.na(duomenys_prieaugis$salygine_outlier))#19 salyginiu

#pasibraizome boxplot
#cia isskaido pagal industrijas
g12<-ggplot(duomenys_prieaugis, aes(x=Industry, y=Growth)) +
  geom_boxplot() + ggtitle("Growth boxplots by industry") +
  xlab("Type of industry") + ylab("Growth (%)")+
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
element_text(hjust=0.5))+
  labs(subtitle = "Labeled conditional outliers") +
  geom_point(data = subset(duomenys_prieaugis, salygine_outlier != "NA"),
```

```
            aes(x = Industry, y  = Growth), size = 1.5, color = "red") +
  scale_y_continuous(limits = c(-5, 30))

g12
#dabar tvarkysime salygines isskirtis, jas pakeisi i 5 ir 95 kvantili

caps <- quantile(duomenys_prieaugis$Growth, probs=c(.05, .95))

duomenys_prieaugis$Growth[duomenys_prieaugis$Growth <
quantile(duomenys_prieaugis$Growth, .25) -
1.5*IQR(duomenys_prieaugis$Growth)] <- caps[1]
duomenys_prieaugis$Growth[duomenys_prieaugis$Growth >
quantile(duomenys_prieaugis$Growth, .25) +
1.5*IQR(duomenys_prieaugis$Growth)] <- caps[2]

#patikriname, ar liko salyginiu ar tikru isskirciu
sum(isskirtis(duomenys_prieaugis$Growth))
sum(salygine_isskirtis(duomenys_prieaugis$Growth))
#neliko


g13<-ggplot(duomenys_prieaugis, aes(x=Industry, y=Growth)) +
  geom_boxplot() + ggtitle("Growth boxplots by industry after conditional
outlier modification") +
  xlab("Type of industry") + ylab("Growth (%)")+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_y_continuous(limits = c(-5, 30))

g13
gridExtra::grid.arrange(g12, g13, nrow=2)

#------------------------------
#Isskirtys
#------------------------------
#pasaliname is visu duomenu rinkiniu isskirtis
duomenys_darb_2<-anti_join(duomenys_darb_2, duomenys_isskirtys, by =
"Name")
duomenys_islaidos<-anti_join(duomenys_islaidos, duomenys_isskirtys, by =
"Name")
duomenys_pajamos<-anti_join(duomenys_pajamos, duomenys_isskirtys, by =
"Name")
duomenys_pelnas<-anti_join(duomenys_pelnas, duomenys_isskirtys, by =
"Name")
duomenys_prieaugis<-anti_join(duomenys_prieaugis, duomenys_isskirtys, by
= "Name")

#pasaliname isskirtis, kuriu yra 37
duomenys<-anti_join(duomenys, duomenys_isskirtys, by = "Name")

#pakeiciame salyginiu isskirciu reiksmes
duomenys$Employees<-duomenys_darb_2$Employees
duomenys$Revenue<-duomenys_pajamos$Revenue
```

```
duomenys$Expenses<-duomenys_islaidos$Expenses
duomenys$Profit<-duomenys_pelnas$Profit
duomenys$Growth<-duomenys_prieaugis$Growth

#toliau dirbame su duomenys rinkiniu

#---------------------------------
#statistika pagal industrijas
#---------------------------------
summary(duomenys)

#---------------------------------
#duomenu normavimas
#---------------------------------
# boxplot -pries normavima
library(reshape)

duomenys$Revenue1mln<-duomenys$Revenue/1000000
duomenys$Expenses1mln<-duomenys$Expenses/1000000
duomenys$Profit1mln<-duomenys$Profit/1000000

data_mod1 <- melt(duomenys, id.vars = "Inception",
                  measure.vars = c('Employees','Revenue', 'Expenses',
'Profit','Growth' ))

p1 <- ggplot(data_mod1) +
  geom_boxplot(aes( y=value, x = as.factor(variable))) +
ggtitle("Variable boxplots before rationing") +
  xlab("Variable name") + ylab("Value") +
  theme(plot.title = element_text(hjust = 0.5))

p1

#normavimas pagal min-max
min_max_norm <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}

min_max_norm.data <- duomenys #xsukuriame duomenu aibes kopija
#normavimas min-max
min_max_norm.data$Growth <-  min_max_norm(min_max_norm.data$Growth)
min_max_norm.data$Employees <-  min_max_norm(min_max_norm.data$Employees)
min_max_norm.data$Revenue <-  min_max_norm(min_max_norm.data$Revenue)
min_max_norm.data$Expenses <-  min_max_norm(min_max_norm.data$Expenses)
min_max_norm.data$Profit <-  min_max_norm(min_max_norm.data$Profit)

#boxplot sunormuotiems duomenims pagal min-max
data_mod2 <- melt(min_max_norm.data, id.vars = "Inception",
                  measure.vars = c('Employees','Revenue', 'Expenses',
'Profit','Growth' ))

p2 <- ggplot(data_mod2) +
```

```
  geom_boxplot(aes( y=value, x = as.factor(variable))) +
ggtitle("Variable boxplots after min - max rationing") +
  xlab("Variable name") + ylab("Value") +
  theme(plot.title = element_text(hjust = 0.5))
p2

#normavimas pagal vidurki ir dispersija
mean_sd_norm <- function(x) {
  (x - mean(x))/ sd(x)
}

mean_sd_norm.data <- duomenys #xsukuriame duomenu aibes kopija
#normavimas pagal vidurki ir dispersija
mean_sd_norm.data$Growth <- mean_sd_norm(mean_sd_norm.data$Growth)
mean_sd_norm.data$Profit <- mean_sd_norm(mean_sd_norm.data$Profit)
mean_sd_norm.data$Employees <- mean_sd_norm(mean_sd_norm.data$Employees)
mean_sd_norm.data$Revenue <- mean_sd_norm(mean_sd_norm.data$Revenue)
mean_sd_norm.data$Expenses <- mean_sd_norm(mean_sd_norm.data$Expenses)

#boxplot sunormuotiems duomenims pagal vidurki ir dispersija
data_mod3 <- melt(mean_sd_norm.data, id.vars = "Inception",
                  measure.vars = c('Employees','Revenue', 'Expenses',
'Profit','Growth' ))

p3 <- ggplot(data_mod3) +
  geom_boxplot(aes( y=value, x = as.factor(variable))) +
ggtitle("Variable boxplots after standartizing") +
  xlab("Variable name") + ylab("Value") +
  theme(plot.title = element_text(hjust = 0.5))
p3

gridExtra::grid.arrange(p1, p2, p3,  nrow=1)

# lyginimas pagal pramones sakas ----------------------------------------
--

pelnas <- min_max_norm.data %>% select('Industry','Profit')

# pelno pasiskirstymas pagal pramones sakas
cons <- with(pelnas, sum(Profit[Industry == 'Construction']))
fs <- with(pelnas, sum(Profit[Industry == 'Financial Services']))
gs <- with(pelnas, sum(Profit[Industry == 'Government Services']))
he <- with(pelnas, sum(Profit[Industry == 'Health']))
it <- with(pelnas, sum(Profit[Industry == 'IT Services']))
re <- with(pelnas, sum(Profit[Industry == 'Retail']))
sof <- with(pelnas, sum(Profit[Industry == 'Software']))
count(pelnas, "Industry") #suzinome kiek eiluciu kiekviena pramones saka
turi

df <- data.frame(Industry = c("Construction","Financial
Services","Government Services","Health","IT Services", "Retail",
"Software"),
```

```
                  Profit =
c(cons/46,fs/51,gs/44,he/80,it/142,re/44,sof/59))

sj1 <- ggplot(df, aes(x = Industry, y = Profit, fill = Industry)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle("Profit distribution among industries") +
  scale_fill_manual(values=c("lightgrey",  "lightgrey", "lightgrey",
"lightgrey","darkgrey", "lightgrey", "lightgrey" )) +
  theme(legend.position="none")

sj1

# islaidu pasiskirstymas pagal pramones sakas

islaidos <- min_max_norm.data %>% select('Industry','Expenses')

cons <- with(islaidos, sum(Expenses[Industry == 'Construction']))
fs <- with(islaidos, sum(Expenses[Industry == 'Financial Services']))
gs <- with(islaidos, sum(Expenses[Industry == 'Government Services']))
he <- with(islaidos, sum(Expenses[Industry == 'Health']))
it <- with(islaidos, sum(Expenses[Industry == 'IT Services']))
re <- with(islaidos, sum(Expenses[Industry == 'Retail']))
sof <- with(islaidos, sum(Expenses[Industry == 'Software']))

df <- data.frame(Industry = c("Construction","Financial
Services","Government Services","Health","IT Services", "Retail",
"Software"),
                  Expenses =
c(cons/46,fs/51,gs/44,he/80,it/142,re/44,sof/59))

sj2 <- ggplot(df,                                    # Grouped barplot
using ggplot2
             aes(x = Industry,
                 y = Expenses, fill = Industry)) +
  geom_bar(stat = "identity",
           position = "dodge") + ggtitle("Expenses distribution among
industries") +
  scale_fill_manual(values=c("lightgrey",  "lightgrey", "lightgrey",
"darkgrey","lightgrey", "lightgrey", "lightgrey" )) +
  theme(legend.position="none")

sj2

# darbuotoju pasiskirstymas pagal pramones sakas

darbuotojai <- min_max_norm.data %>% select('Industry','Employees')

cons <- with(darbuotojai, sum(Employees[Industry == 'Construction']))
fs <- with(darbuotojai, sum(Employees[Industry == 'Financial Services']))
gs <- with(darbuotojai, sum(Employees[Industry == 'Government
Services']))
he <- with(darbuotojai, sum(Employees[Industry == 'Health']))
```

```
it <- with(darbuotojai, sum(Employees[Industry == 'IT Services']))
re <- with(darbuotojai, sum(Employees[Industry == 'Retail']))
sof <- with(darbuotojai, sum(Employees[Industry == 'Software']))

df <- data.frame(Industry = c("Construction","Financial
Services","Government Services","Health","IT Services", "Retail",
"Software"),
                 Employees =
c(cons/46,fs/51,gs/44,he/80,it/142,re/44,sof/59))

sj3 <- ggplot(df,                                    # Grouped barplot
using ggplot2
             aes(x = Industry,
                 y = Employees, fill = Industry)) +
  geom_bar(stat = "identity",
           position = "dodge") + ggtitle("Employees distribution among
industries") +
  scale_fill_manual(values=c("lightgrey",  "lightgrey",
"darkgrey","lightgrey", "lightgrey", "lightgrey", "lightgrey" )) +
  theme(legend.position="none")


sj3

library(ggpubr)
ggarrange(sj1, sj2, sj3, ncol=1)

# Kur daugiausia imoniu ----------------------------------------------
---

df <- count(duomenys, "State") #suzinome kiek eiluciu kiekviena pramones
saka turi
df

df <- head(df[order(df$freq, decreasing = TRUE),c(1,2)], 5) #randame
kiek, kiekviena valstija turi imoniu

sg <- ggplot(df, aes(x = State, y = freq, fill = State)) +
  geom_bar(stat = "identity",position = "dodge") +
geom_text(aes(label=freq), vjust=-0.3, size=3.5) +
  ylab("Number of industries") +
  scale_fill_brewer(palette="Set1")
sg


# Kuriose pramones sakose --------------------------------------------
--
# pirma issirenki valstija ir tada randi max is stulpelio

#--------
#Kokiu pramones saku daugiausiai CA
#--------
```

```
CA <- duomenys %>% select('State','Industry')
CA <- with(CA, count(Industry[State == 'CA']))
CA[order(CA$freq, decreasing = TRUE),c(1,2)]

#--------
#Kokiu pramones saku daugiausiai FL
#--------
FL <- duomenys %>% select('State','Industry')
FL <- with(FL, count(Industry[State == 'FL']))
FL[order(FL$freq, decreasing = TRUE),c(1,2)]

#--------
#Kokiu pramones saku daugiausiai NY
#--------
NY <- duomenys %>% select('State','Industry')
NY <- with(NY, count(Industry[State == 'NY']))
NY[order(NY$freq, decreasing = TRUE),c(1,2)]

#--------
#Kokiu pramones saku daugiausiai TX
#--------
TX <- duomenys %>% select('State','Industry')
TX <- with(TX, count(Industry[State == 'TX']))
TX[order(TX$freq, decreasing = TRUE),c(1,2)]

#--------
#Kokiu pramones saku daugiausiai VA
#--------
VA <- duomenys %>% select('State','Industry')
VA <- with(VA, count(Industry[State == 'VA']))
VA[order(VA$freq, decreasing = TRUE),c(1,2)]

# -------------------------------
# koreliacija
# -------------------------------
library(corrplot)

corrplot.mixed(cor(duomenys[,8:11]),
               lower = "number",
               upper = "ellipse",
               tl.col = "black")
```