VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Papildomi duomenų vizualizavimo skyriai

1 užduotis

Atliko: 3 kurso 1 grupės studentai:

Matas Amšiejus

Roland Gulbinovič

Darbo vadovė: dr. Jolita Bernatavičienė

Vilnius, 2022

Turinys

1. Įvadas

1.1 Tikslas

Atlikti pradinį imties iš JAV įmonių duomenų apdarojimą bei vizualią ir koreliacinę analizę.

1.2 Uždaviniai

1. Įsigilinti į duomenis, išsiaiškinti, ką reiškia kiekvienas argumentas;
2. Išvalyti nekorektiškai suvestus duomenis;
3. Atlikti duomenų priešanalizę;
4. Užpildyti praleistas reiškmes atitinkamais metodais;
5. Identifikuoti ir pašalinti išskirtis;
6. Sunormuoti skaitinius duomenis;
7. Atlikti duomenų vizualią analizę;
8. Patikrinti koreliacijas tarp duomenų stulpelių.

2. Duomenys

Duomenų faile buvo 500 įrašų JAV įmonių imtis. Duomenų atributai buvo:

1. ID –įmonės identifikacinis numeris duomenyse, kategorinis kintamasis (matavimų skalė (toliau m.s.) skalė nominalioji);
2. Name – įmonės pavadinimas, kategorinis kint., (m. s. nominalioji);
3. Industry – pramonės šaka, kategorinis kint., (m. s. nominalioji);
4. Inception – įsteigimo metai, kiekybinis diskretus kint., (m. s. intervalų);
5. Employees – darbuotojų skaičius, kiekybinis diskretus kint., (m. s. santykių);
6. State – valstija, kategorinis kint., (m. s. nominalioji);
7. City – miestas, kategorinis kint., (m. s. nominalioji);
8. Revenue – pajamos (doleriais), kiekybinis tolydus kint., (m. s. santykių);
9. Profit – pelnas (doleriais), kiekybinis tolydus kint. (m. s. intervalų);
10. Expenses – išlaidos (doleriais), kiekybinis tolydus kint. (m. s. santykių);
11. Growth – įmonės augimas, procentais, kiekybinis tolydus kint. (m. s. intervalų);

Kai kurie duomenys buvo suvesti nekorektiškai (pridėti nereikalingi simboliai), todėl prieš pereinant prie tolimesnės analizės ištaisome klaidas (kodas: 2. Duomenų nuskaitymas / tvarkymas).

3. Aprašomoji statistika

3.1 Bendra statistika

Pirma ištirsime duomenų aprašomąją statistiką.

1 lentelė. Bendra aprašomoji statistika

Atributas	n	Vidurkis	Stand. nuok	Mediana	Q1	Q3	Min	Max	Variacijos žingsnis
Įsteigimas	499	2010,17	3,23	2011	2009	2012	1999	2014	15
Darbuotojai	496	149,04	398,1	56	27,75	126	1	7125	7124
Pajamos	496	10850256,23	3199034,56	10671779,5	8684289	13112127	1614585	21810051	20195466
Išlaidos	494	4309096,01	2120729,17	4341072	2758418	5834723	71219	9860686	9789467
Pelnas	498	6539474,01	3869933,65	6513366	3272074	9303951	12434	19624534	19612100
Augimas	497	14,35	6,9	15	8	20	-3	30	33

Iš lentelės svarbiausia atkreipti dėmesį reikia į darbuotojų skaičių. Mediana ir vidurkis smarkiai skiriasi, o skirtumas tarp didžiausios ir mažiausios reikšmių yra žymiai didesnis už standartinį nuokrypį. Tai gali indikuoti vieną ar kelias smarkias išskirtis duomenyse. Taip pat reiškia, kad vidurkiu kol kas pasikliauti nederėtų.

3.2 Statistika pagal pramonės tipą

Šią lentelę vizualizuoti ir interpretuoti yra žymiai sunkiau, todėl šioje dalyje įkelsiu tik dalį lentelės (2 lentelė (pilna). Aprašomoji statistika pagal pramonę).

2 lentelė. Aprašomoji statistika pagal pramonę (tik darbuotojų)

Column1	Pramonė	n	Vidurkis	Stand. nuok.	Mediana	Min	Max	Intervalo ilgis	Q1	Q3
Darbuotojai1	Construction	50	61,26	59,43	37,5	5	272	267	23,25	75
Darbuotojai2	Financial Services	51	217,75	331,24	85	3	1628	1625	33,5	267,5
Darbuotojai3	Government Services	50	172,72	233,63	99	13	1224	1211	49	150
Darbuotojai4	Health	85	207,99	307,1	88	6	1600	1594	31	230
Darbuotojai5	IT Services	145	107,63	257,88	51	2	2670	2668	28	110
Darbuotojai6	Retail	47	209,28	1033,74	28	1	7125	7124	15,5	70
Darbuotojai7	Software	64	121,06	178,31	60	3	850	847	26	122,25

Matome, kad darbuotojų skaičius smarkiai skiriasi tarp skirtingų pramonių. Tą verta įsiminti ateičiai. Taip pat konkrečiai matome, kad mūsų tikslumą labai mažina išskirtis / išskirtys iš *pardavimų* (*retail*) pramonės (kodas: 3. Aprašomoji statistika).

4. Praleistų reikšmių tvarkymas

Iš pirmos lentelės nepastovaus n (įrašų skaičiaus) buvo galima susidaryti išvadą, kad duomenyse yra praleistų reikšmių. Mūsų tikslas yra užpildyti jas kuo tikslesnėmis reikšmėmis. Pirma išmetame eilutes, kur tuščių reikšmių užpildyti nepavyks (*pramonės*, *metų*). Aiškiausias užpildymo metodas yra pasinaudojimas faktiniais duomenimis (užpildome *valstijų* praleistas reikšmes pagal *miestą*) . Tam, kad užpildytume *darbuotojų skaičių*, pasitelksime papildomą informaciją, t.y. *pramonę*. Iš 2 lentelės matėme, kad duomonys tarp jų gana smarkiai skiriasi, todėl taip galime padidinti įstatytų reikšmių tikslumą. Naudosime medianas, nes duomenys kol kas turi išskirčių. Toliau pildysime *pajamas* pasinaudodami matematiniu sąryšiu *Pelnas = Pajamos – Išlaidos*. Gauname identiškas reikšmes, kurios turėjo būti vietoje praleistų langelių. Tačiau jei

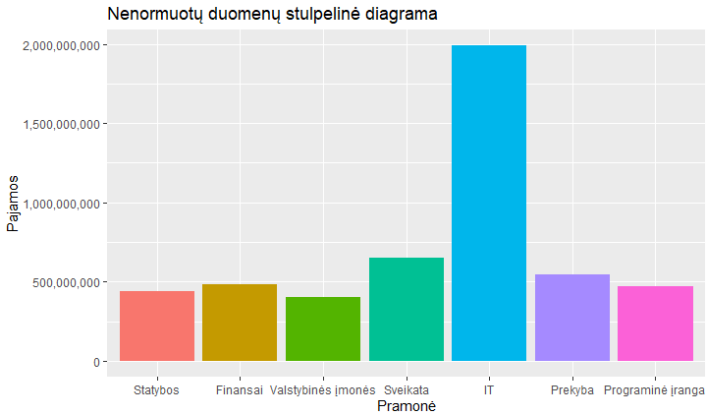
kažkuri iš likusių formulės dalių irgi buvo tuščia, užpildome pasinaudodami mediana (vėl įtraukdami pramones). Tą pačią formulę naudojame užpildyti *pelno* ir *išlaidų* reikšmes, tačiau jei šios abi tuščios, eilutes šaliname (per daug išvestinių duomenų, pradeda neatitikti realybės). Užpildome *augimo* reikšmes vėl panaudodami medianą pagal pramonę (kodas: 4. Praleistų reikšmių tvarkymas).

5. Išskirčių šalinimas

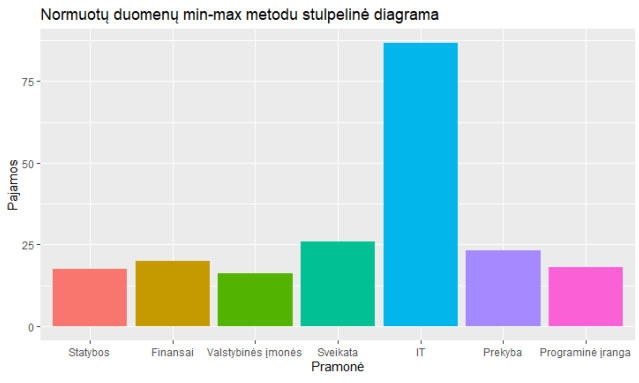
Išskirtis išmesime, jei jos nepateks tarp išorinių barjerų ([$Q1$ – 3H; $Q3$ + 3H], kur H = $Q3$ – $Q1$). Patikrinus *pajamų*, *išlaidų* ir *pelno* reikšmes rastos tik sąlyginės išskirtys, kurių nusprendžiame nešalinti dėl buvimo niekuo neypatingomis tarp kitų kintamųjų. Taip pat paliekame visas *augimo* reikšmes. Tačiau su *darbuotojų skaičiumi* gauname labai daug išskirčių, kurias šaliname. Nors iš naujo patikrinus išskirtis jos buvo, tačiau jų nebetriname, tik pažymime papildomai, kad būtų lengviau jas atskirti. Prieš pašalinant išskirtis *darbuotojų skaičiaus* vidurkis buvo *149,04*, standartinis nuokrypis – *398,1*, o mediana – *56*. Išmetus išskirtis, atitinkamai gavome *81,07*, *82,74*, *50*. Taigi, vidurkis sumažėjo beveik dvigubai, standartinis nuokrypis – beveik 5 kartus, o mediana beveik nepakito (pagrindimas, kodėl naudota anksčiau įstatant į tuščias reikšmes) (kodas: 5. Išskirčių šalinimas).

6. Duomenų normavimas

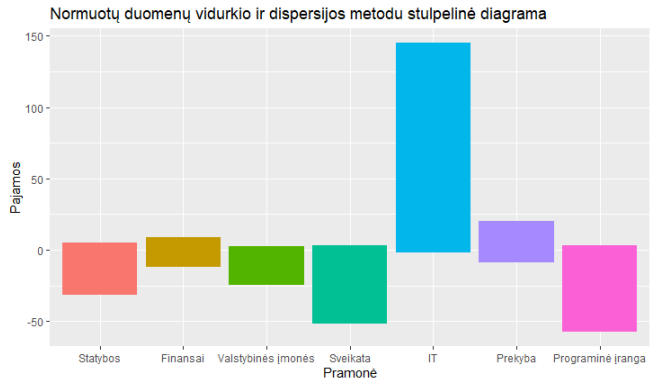
Šiame darbe duomenų normavimą gana sunku ištirti, nes šiam etapui jis iš esmės yra nereikalingas (nevykdomas algortitmų mokymas). Tačiau pateiksiu vizualizaciją, kaip duomenys pasiskirstę naudojant histogramą.



2 pav. Nenormuotų duomenų stuleplinė diagrama



1 pav. Normuotų duomenų min-max metodu stulpelinė diagrama

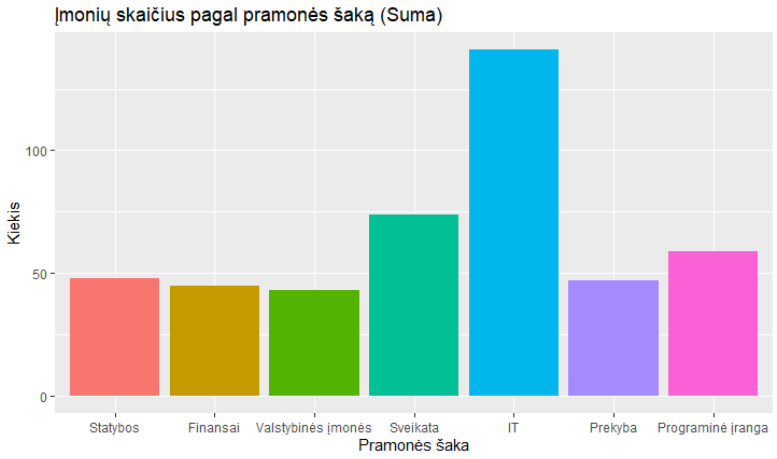


3 pav. Normuotų duomenų vidurkio ir dispersijos metodu stulpelinė diagrama

Matome, kad normuojant vidurkio ir dispersijos metodu, galimos neigiamos reikšmės. Tai reiškia, kad tos pramonės pajamos yra žemiau bendro vidurkio.

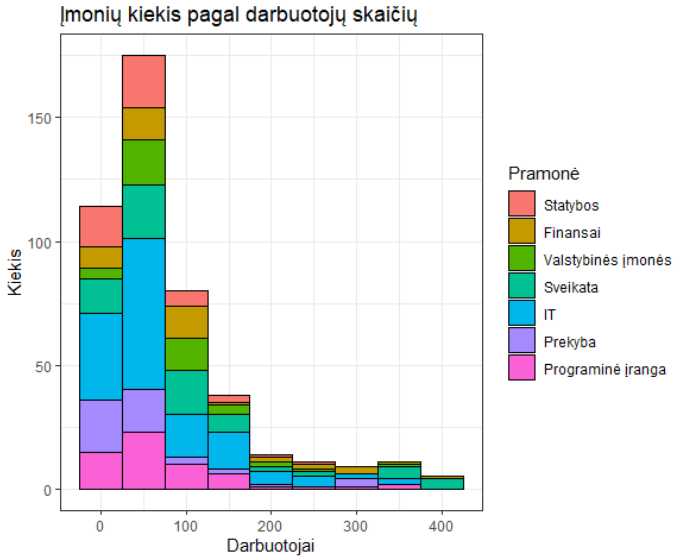
7. Vizuali analizė

Pirma norėjome patikrinti kaip pasiskirsčiusios pramonės pagal įvairius rodiklius (kodas: 7. Vizuali analizė).



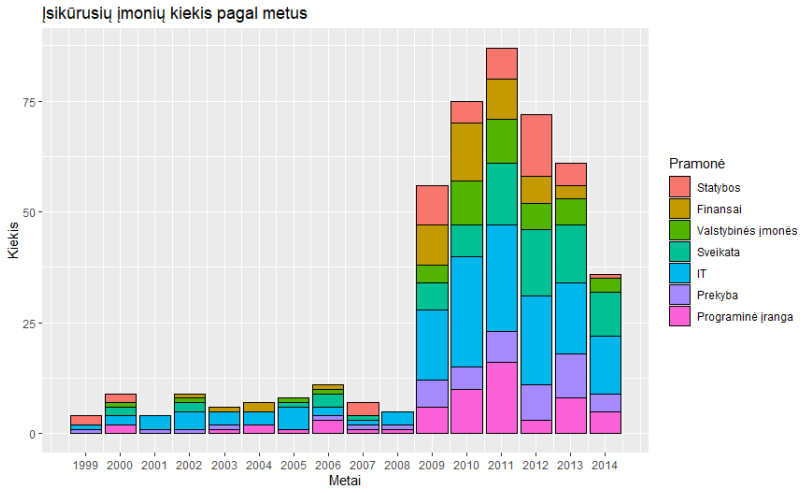
4 pav. Įmonių skaičius pagal pramonės šaką (Suma)

Matome, kad *IT* ir *sveikatos* įmonės yra pačios dažniausios (141 ir 74 įmonės), likusios įmonės yra panašaus populiarumo (vidutiniškai po 50).



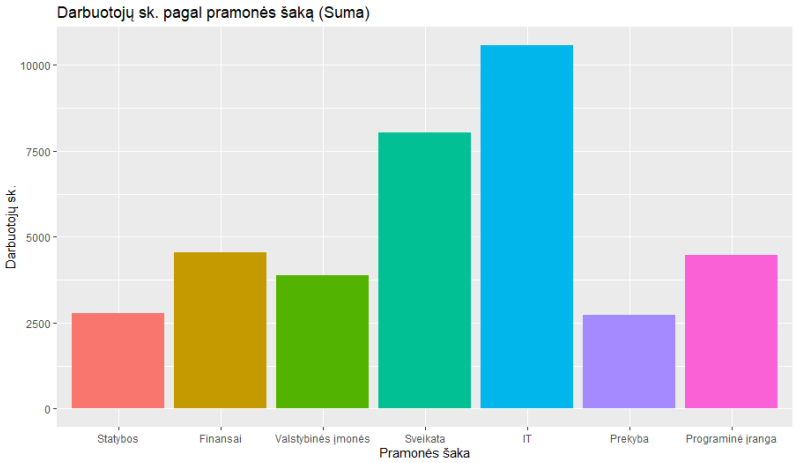
5 pav. Įmonių kiekis pagal darbuotojų skaičių

Ši histograma parodo, kad didžiausia dalis įmonių turi mažiau negu 150 darbuotojų. Didžioji dalis darbuotojų skaičiumi išsiskyrusių įmonių yra *sveikatos* pramonės.

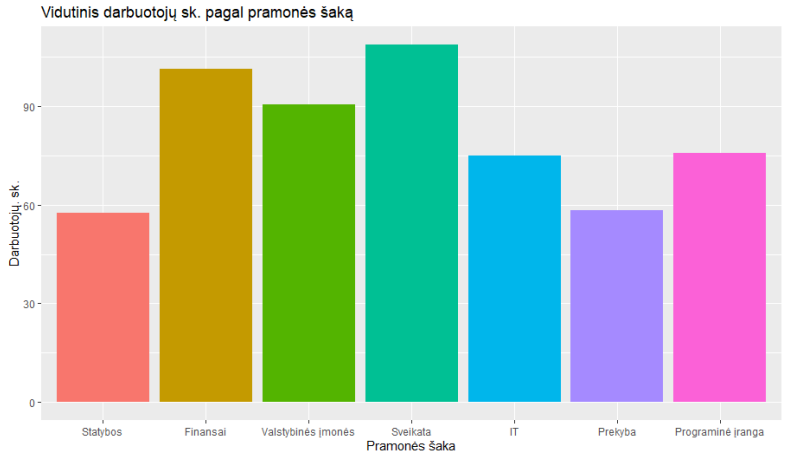


6 pav. Įsikūrusių įmonių kiekis pagal metus

Iš stulpelinės diagramos matome, kad didžioji dalis į imtį patekusių įmonių įsikūrusios nuo 2009.



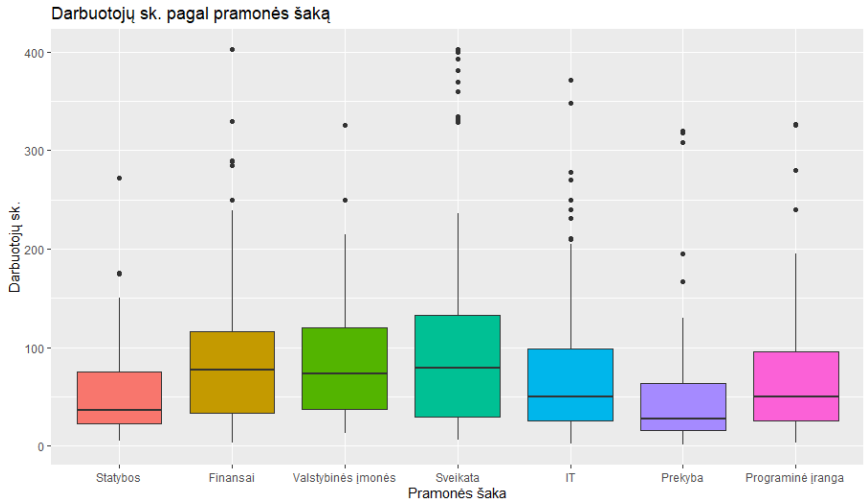
7 pav. Darbuotojų sk. pagal pramonės šaką (Suma)



8 pav. Vidutinis darbuotojų sk. pagal pramonės šaką

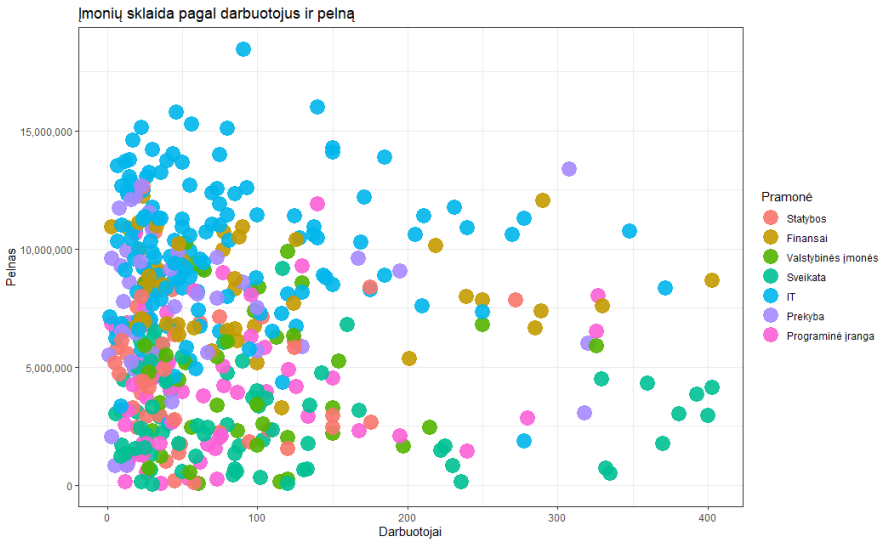
IT ir *sveikatos* pramonės šakos turi didžiausius darbuotojų skaičius (10582 ir 8045 darbuotojai). Tačiau, nors *finansų* įmonių yra nedaug (45), jų darbuotojų skaičius yra gana didelis

(4560 darbuotojai), nes jos išsiskiria dideliu vidutiniu darbuotojų skaičiumi (vid. 101 darbuotojas, o visų pramonių vidurkis - 81) (*7, 8 pav.*).

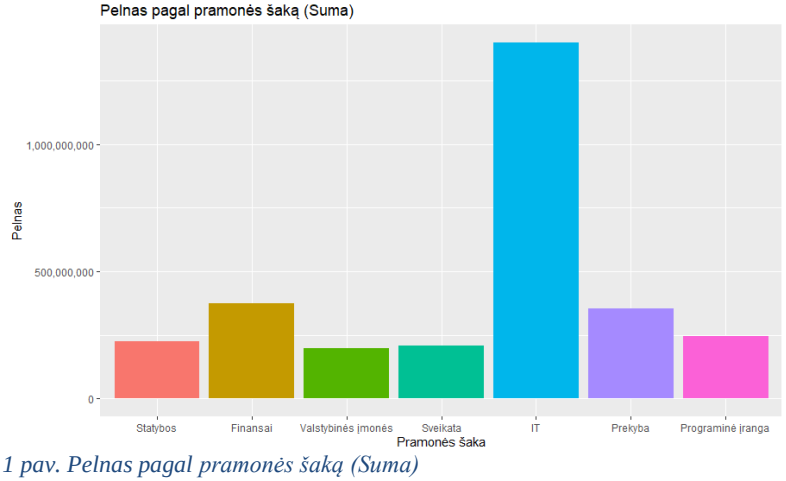


9 pav. Stačiakampė darbuotojų sk. pagal pramonės šaką diagrama

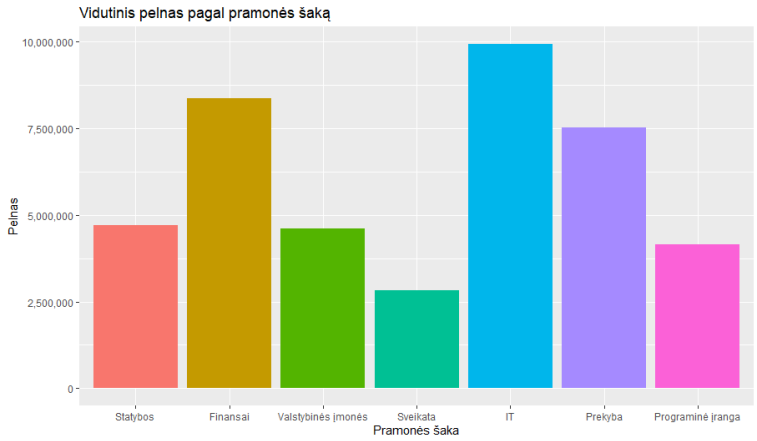
Iš stačiakampės diagramos matome, kad *sveikatos* ir *IT* pramonės turi didžiausią kiekį išskirčių. Galime pastebėti tai, kad visos išskirtys yra tik dėl didesnio nei įprasto darbuotojų skaičiaus.



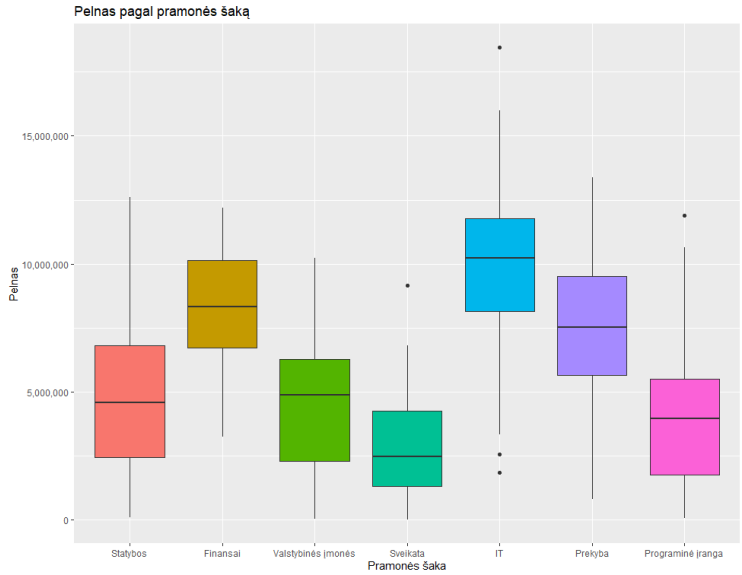
10 pav. Įmonių sklaida pagal darbuotojus ir pelną



11 pav. Pelnas pagal pramonės šaką (Suma)

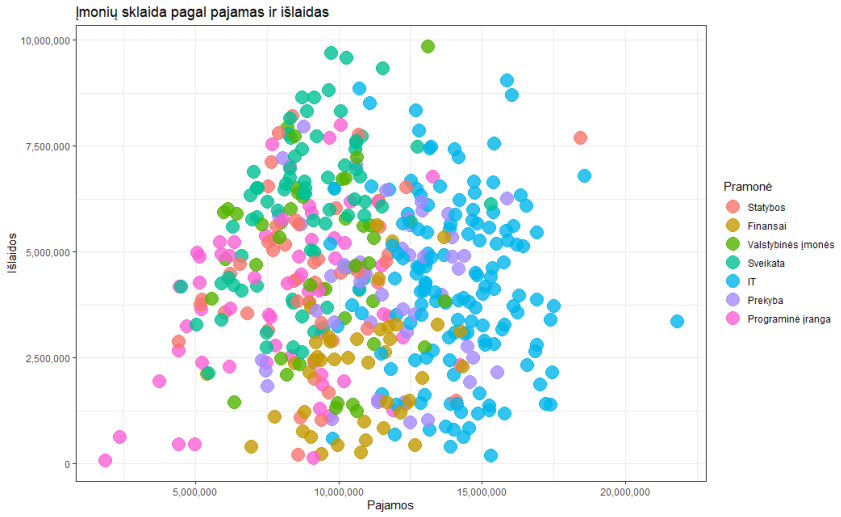


12 pav. Vidutinis pelnas pagal pramonės šaką

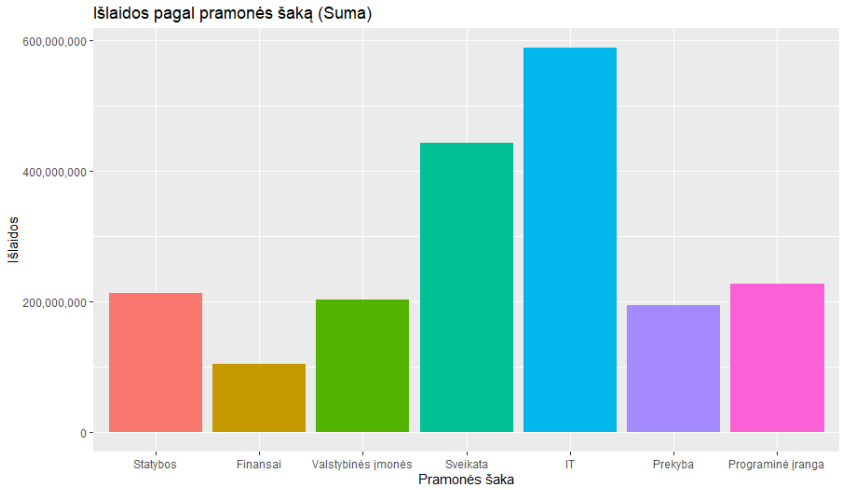


13 pav. Pelno pagal pramonės šaką stačiakampė diagrama

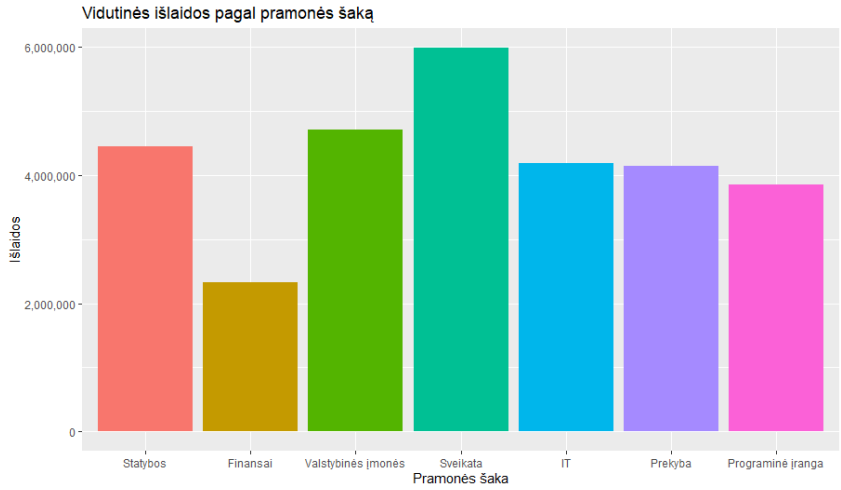
Matome, kad *IT* įmonės turi didžiausią pelną (vidutinis pelnas lygus 9 935 720,46 $ palyginus su visų įmonių vid. pelnu 6 580 702,26 $), nors dauguma jų turi mažesnį darbuotojų skaičių (vid. darbuotojų sk. IT = 75, visų įmonių = 81). Mažiausiai pelningos atrodo *sveikatos* įmonės (2 826 353,18 $), tą atspindi ir žemas vidutinis pelnas. Nors *finansų* įmonių yra nedaug (45), jie yra antri pagal sumį ir vidutinį pelningumą (8 356 685,40 $). Taip pat dideliu vidutiniu pelnu pasižymi prekybos įmonės (7 514 645,13 $). Yra tik 5 išskirtys, visos – sąlyginės.



14 pav. Įmonių sklaida pagal pajamas ir išlaidas



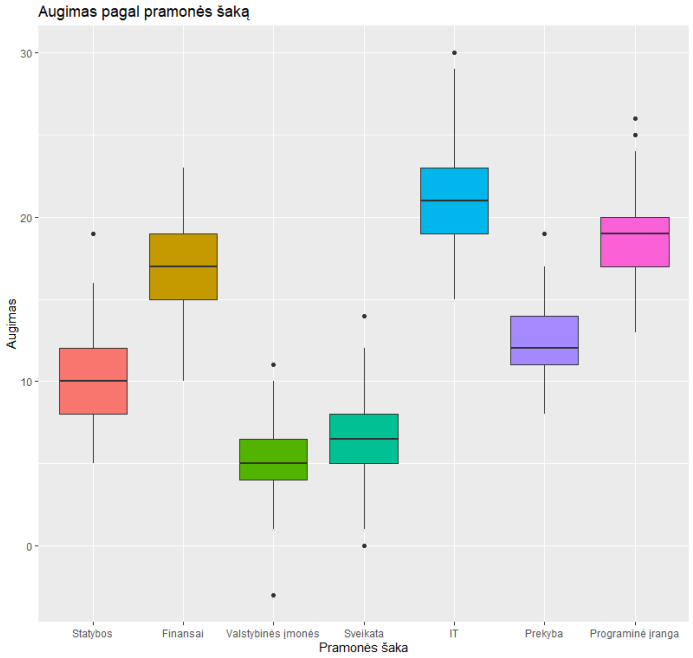
15 pav. Išlaidos pagal pramonės šaką (suma)



16 pav. Vidutines išlaidos pagal pramonės šaką

Iš 14 paveikslėlio matosi, kad didžiausias pajamų sektorius yra *IT* (IT vidurkis = 14 116 291,11 $, bendras vidurkis = 10 903 867,87 $). *Programinė įranga* ir *sveikata* gauna mažiausias pajamas (vid. 7 997 605,83 $ ir 8 816 952,23 $ atitinkamai), tačiau pastarosios išlaidos

yra didesnės (vid. 5 990 599,05 $, kai bendras vidurkis yra 4 323 165,61 $). *Finansai* turi mažiausias išlaidas (vid. 2 324 115,44 $). Nors bendros *IT* sektoriaus išlaidos yra didžiausios, taip yra todėl, kad jis yra populiariausias.



17 pav. Augimo pagal pramonės šaką stačiakampė diagrama

Iš stačiakampės diagramos matome, kad labiausiai augančios (besivystančios) įmonės yra iš IT sektoriaus (vid. 21,34 %). Taip pat sparčiai auga programinės įrangos (19,08 %) ir finansų (16,73 %) sektoriai. Mažiausias pokytis pasireiškia valstybinėse (5,16 %) ir sveikatos (6,61 %) įmonėse.

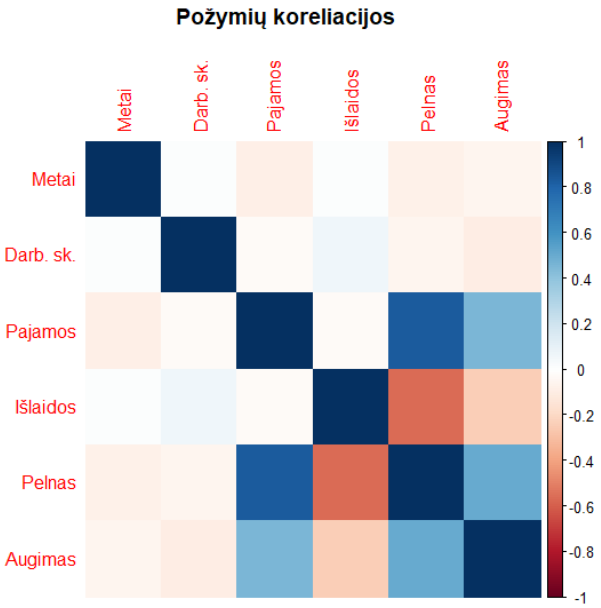
3 lentelė Įmonių valstijų kiekiai

	State	Kiekis_vals
1	CA	54
2	VA	47
3	TX	40
4	FL	31
5	NY	26
6	IL	22
7	MD	22
8	GA	21
9	NJ	16
10	MN	14

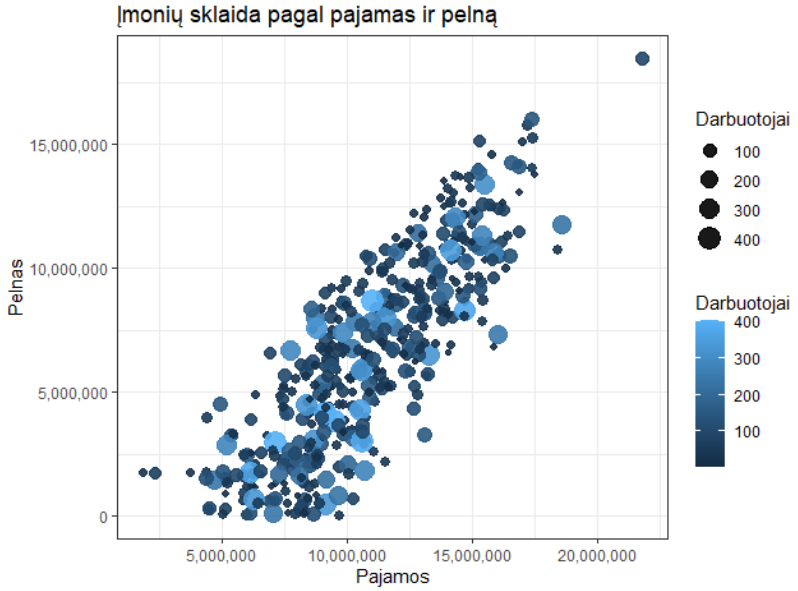
Iš lentelės matome, kad daugiausiai įmonių yra Kalifornijos, Virdžinijos ir Teksaso valstijose, toliau įmonių skaičiai pradeda sparčiai kristi.

4 lentelė Koreliacijos koeficientu matrica

	Metai	Darb. sk.	Pajamos	Išlaidos	Pelnas	Augimas
Metai	1.00	0.01	-0.08	0.02	-0.08	-0.05
Darb. sk.	0.01	1.00	-0.03	0.06	-0.06	-0.09
Pajamos	-0.08	-0.03	1.00	-0.03	0.84	0.45
Išlaidos	0.02	0.06	-0.03	1.00	-0.57	-0.24
Pelnas	-0.08	-0.06	0.84	-0.57	1.00	0.50
Augimas	-0.05	-0.09	0.45	-0.24	0.50	1.00



18 pav. Požymių koreliacijos



19 pav. Įmonių sklaida pagal pajamas ir pelną.

Sudarius koreliacijos matricą matome, kad egzistuoja stipri koreliacija tarp *pelno* ir *pajamų*, tai atspindi ir *18* paveikslas, tačiau ši informacija buvo žinoma iš anksčiau, nes šie požymiai yra susiję matematinių ryšiu. Tą pati galime pasakyti apie *išlaidų* ir *pelno* sąryšį, tik, kad jis atvirkštinis. Dar yra vidutinio stiprumo koreliacija tarp *augimo* ir *pelno*, *pajamų* (kodas: 8. Koreliacijos).

8. Išvados

Įsigilinus ir išvalius duomenis, atlikus duomenų priešanalizę pastebėjome, kad darbuotojų skaičius gali sukelti problemų tolimesnėje analizėje dėl nepastovių duomenų pasirinkus medianą kaip praleistų duomenų užpildymo reikšmę gavome daug išskirčių su darbuotojų skaičiumi. Pašalinus išskirtis atsirado naujos, tačiau jas užfiksavus mes tęsėme darbą. Iš vizualios analizės pastebėjome, kad nors *IT* skyrius nepasižymi dideliu darbuotojų skaičiumi, jos vidutinės pajamos yra didžiausios. Priešingai, sveikatos pramonė pasižymėjo dideliu darbuotojų skaičiumi bei mažomis pajamomis. Nors *finansų* skyriaus įmonių yra nedaug, bet darbuotojų skaičius ir pelningumas yra gana didelis. Pastebėjome, kad didžioji dalis įmonių yra įsikūrusios nuo 2009 metų. Duomenys patvirtino koreliaciją tarp pelno, pajamų ir išlaidų. Taip pat pastebėjome vidutinio stiprumo koreliaciją tarp *augimo* ir *pelno, pajamų,* tačiau, kadangi nežinome daug apie augimo požymį, negalime susidaryti gilesnių įžvalgų.

9. Priedai

2 lentelė (pilna). Aprašomoji statistika pagal pramonę

Stulpelis	Pramonė	n	Vidurkis	Stand. nuok.	Mediana	Min	Max	Intervalo ilgis	Q1	Q3
Darbuotojai 1	Statyba	50	61,26	59,43	37,5	5	272	267	23,25	75
Darbuotojai 2	Finansai	51	217,75	331,24	85	3	1628	1625	33,5	267,5
Darbuotojai 3	Valstybinės įmonės	50	172,72	233,63	99	13	1224	1211	49	150
Darbuotojai 4	Sveikata	85	207,99	307,1	88	6	1600	1594	31	230
Darbuotojai 5	IT	145	107,63	257,88	51	2	2670	2668	28	110
Darbuotojai 6	Prekyba	47	209,28	1033,74	28	1	7125	7124	15,5	70
Darbuotojai 7	Programinė įranga	64	121,06	178,31	60	3	850	847	26	122,25
Pajamos1	Statyba	47	9145391,49	2429115,8	8982358	4419277	18429577	14010300	7718803,5	10634699
Pajamos2	Finansai	53	10627179,6	1933148,8	10928801	5387469	14330107	8942638	9205547	11779555
Pajamos3	Valstybinės įmonės	50	9436792,34	2342556,6	9707475	4637647	15188113	10550466	8035933,8	10706253
Pajamos4	Sveikata	86	8811121,94	1978819,8	8855709,5	1614585	15312302	13697717	7588070,3	10013635
Pajamos5	IT	145	14146014	1963516,1	14053058	9691133	21810051	12118918	12882726	15359369
Pajamos6	Prekyba	48	11641572,4	2200542,7	11936371,5	7307243	15880376	8573133	10183493	12989172
Pajamos7	Programinė įranga	63	7907718,67	2643624,6	8304480	1835717	14229411	12393694	5755508	9684355
Išlaidos1	Statyba	48	4453204,5	1793321,7	4506975,5	214470	8213905	7999435	3539327,5	5376596,3
Išlaidos2	Finansai	53	2390108,36	1510023,8	2445885	223602	6212849	5989247	1207273	3133190
Išlaidos3	Valstybinės įmonės	50	4741746,34	2055429,6	4790732,5	1243956	9860686	8616730	3533079,3	5999725
Išlaidos4	Sveikata	86	5881840,64	1892100,1	6162150,5	1323005	9712296	8389291	4231219,3	7249760,8
Išlaidos5	IT	143	4164930,15	2029385,9	4068630	187655	9046498	8858843	2823765	5558355
Išlaidos6	Prekyba	48	4158844,48	1787500,8	4545730,5	968518	7957743	6989225	2703718,8	5421299,3
Išlaidos7	Programinė įranga	62	3824478,08	1951319,7	4129542	71219	8007771	7936552	2340480	5156943,3
Pelnas1	Statyba	48	4705532,62	2805089,4	4573280,5	96073	12616182	12520109	2442148,3	6801062
Pelnas2	Finansai	53	8237071,26	2144392,7	8282728	3259485	12205097	8945612	6636007	10151080
Pelnas3	Valstybinės įmonės	50	4695046	2820709,1	4836705,5	46851	10565044	10518193	2425506,5	6302041,3
Pelnas4	Sveikata	86	2929281,3	2075213,5	2514786,5	12434	9174395	9161961	1311362,5	4480562,8
Pelnas5	IT	145	9984962,73	2983951,6	10104104	1841685	19624534	17782849	8138717	11765611
Pelnas6	Prekyba	48	7482727,9	2897292,1	7326357	815381	13369247	12553866	5658476,5	9490338
Pelnas7	Programinė įranga	64	4104288,7	2929838,8	3957673,5	68862	11902072	11833210	1749145	5936768,5
Augimas1	Statyba	49	10,06	3,07	10	5	19	14	8	12
Augimas2	Finansai	53	16,6	2,66	17	10	23	13	15	19
Augimas3	Valstybinės įmonės	50	5	2,87	5	-3	11	14	4	7

Augimas4	Sveikata	86	6,59	2,6	6	0	14	14	5	8
Augimas5	IT	144	21,37	3,09	21	15	30	15	19,75	23
Augimas6	Prekyba	48	12,5	2,59	12	8	19	11	11	14
Augimas7	Programinė įranga	63	18,92	2,9	19	13	26	13	17	20

2. Duomenų nuskaitymas / tvarkymas

```
# Matas Amšiejus
# 1 užduotis. Pirminio duomenų apdorojimo metodų taikymas


library(readr)
library(psych) # naudojama describe fjai
library(tidyverse)
library(ggpubr)
library(scales)
library(corrplot)
library(gridExtra)

################################
# 1. DUOMENU LENTELES PARUOSIMAS
################################

duom <- read_csv("Future-500-1.csv")

str(duom)
# Matome, kad revenue, expenses ir growth turetu buti skaitiniai duomenys,
# taciau nera.

# Matome, kad revenue, expenses ir growth turetu buti skaitiniai duomenys,
# taciau nera. Taip pat id priskirkime kaip kategorini kintamaji:

duom$ID <- as.character(duom$ID)
# tvarkome revenue stulpeli
duom$Revenue <- gsub("\\$","",duom$Revenue)
duom$Revenue <- gsub(",","",duom$Revenue)
duom$Revenue <- as.numeric(duom$Revenue)

# tvarkome expenses stulpeli
duom$Expenses <- gsub(" Dollars","",duom$Expenses)
duom$Expenses <- gsub(",","",duom$Expenses)
duom$Expenses <- as.numeric(duom$Expenses)

# tvarkome expenses stulpeli
duom$Growth <- gsub("\\%","",duom$Growth)
duom$Growth <- as.numeric(duom$Growth)
```

3. Aprašomoji statistika

```
################################
# 2. APRASOMOJI STATISTIKA IR DUOMENU PRIESANALIZE
################################

# a)
```

```
# Bendra aprasomoji statistika + praleistos reiksmes:
as.table(summary(duom))

# Patogu tuom, kad graziai sudeda i lentele
apras_stat <- describe(duom[-c(1,2,3,6,7)], quant = c(0.25,0.75))
apras_stat <- round(apras_stat, 2)
apras_stat <- apras_stat[,-c(1,6,7,11,12,13)]

#write.csv(apras_stat, file = "pirma_lent.csv")

# b)
# Statistika pagal pramones sakas (nespausdina 1 ir 3 kvartiliu, kaip?)
apras_stat_pram <- describeBy(duom[-c(1,2,3,6,7)], group = duom$Industry, mat =
T,
                              digits = 2, quant = c(0.25, 0.75))

apras_stat_pram <- apras_stat_pram[,-c(1,3,8,9,13,14,15)]

#write.csv(apras_stat_pram, file = "antra_lent.csv")
```

4. Praleistų reikšmių tvarkymas

```
################################
# 3. PRALEISTU REIKSMIU TVARKYMAS
################################
# Atsargine kopija
duom_backup <- duom
#duom <- duom_backup

# Pirma isrinksime reiksmes, kuriu nera kaip uzpildyti.
# Industry
duom <- duom[!is.na(duom$Industry),]
# Inception
duom <- duom[!is.na(duom$Inception),]


# Toliau pildysime employees

#Patikrinkime, kur truksta reiksmiu:
duom[is.na(duom$Employees),]
#Truksta Retail, Health ir Financial Sector.

# a) pagal visos imties mediana
# is lenteles apras_stat matome, kad mediana 56, o vidurkis 149. Matome dideli
skirtuma,
# pabandykime paanalozuoti, kas geriau.
hist(duom$Employees)
# Turime akivaizdzia isskirti, tad arba tektu imti mediana, arba nupjautini
vidurki, kuris
# yra ~81

# b) pagal mediana pramones grupese (industry)
# is lenteles apras_stat_pram matome, kad tiek vidurkiai, tiek medianos tarp kai
kuriu grupiu
# gana smarkiai skiriasi. Tai sufleruoja, jog vertetu atsizvelgti i pramones
tipa uzpildant
# praleistas reiksmes.

# grupuota <- duom %>% group_by(Industry) %>% summarise(mean)
temp <- duom
temp <- temp[!is.na(temp$Employees),]
```

```
temp <- temp[!c(temp$Employees>6000),]

ggerrorplot(data = temp, x = 'Industry', y = 'Employees',
            desc_stat = "median_iqr",
            add = "mean") +
  xlab("Pramonė") + ylab("Darb. sk.") + theme_bw()

#Si karta pakeiskime praleistas reiksmes pagal medianas pramones grupese
duom$Employees = round(ifelse(is.na(duom$Employees),
                      ave(duom$Employees, duom$Industry,
                          FUN = function(x) median(x, na.rm = TRUE)),
                      duom$Employees), 0)

# Patikriname, ar nebeliko tusciu reiksmiu:
duom[is.na(duom$Employees),]


# Dabar uzpildysime praleistas State reiksmes
duom[is.na(duom$State),]

#Turime uzpildyti 2 eil NY ir 2 eil CA
duom[is.na(duom$State) & duom$City=="New York", "State"] <- "NY"
duom[is.na(duom$State) & duom$City=="San Francisco", "State"] <- "CA"


# Revenue
duom[is.na(duom$Revenue),]

#Jei imanoma, pagal formule
duom[is.na(duom$Revenue), "Revenue"] <- duom[is.na(duom$Revenue), "Profit"] +
  duom[is.na(duom$Revenue), "Expenses"]

# Uzpildome pagal mediana pramonese. Isiminkime eilutes 8 ir 44
duom$Revenue = ifelse(is.na(duom$Revenue),
                      ave(duom$Revenue, duom$Industry,
                          FUN = function(x) median(x, na.rm = TRUE)),
                          duom$Revenue)

# Profit
duom[is.na(duom$Profit),]

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

# Expenses
duom[is.na(duom$Expenses),]

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

# Istriname reiksmes, kurias pildyti nekorektiska
duom <- duom[!is.na(duom$Expenses),]

# Uzpildome tuscias growth reiksmes
duom[is.na(duom$Growth),]

duom$Growth = ifelse(is.na(duom$Growth),
                      ave(duom$Growth, duom$Industry,
                          FUN = function(x) median(x, na.rm = TRUE)),
                      duom$Growth)
```

```
# Patikriname, ar nebeliko tusciu reiksmiu
duom[!complete.cases(duom), ]
```

5. Išskirčių šalinimas

```
################################
# 4. ISSKIRCIU SALINIMAS
################################

isskirtys <- function(stulp, daugikl){
  iqr <- IQR(stulp)
  Q1<-as.numeric(summary(stulp)[2])
  Q3<-as.numeric(summary(stulp)[5])
  lower_bound <- Q1 - daugikl * iqr
  upper_bound <- Q3 + daugikl * iqr


  outliers <- which(stulp < lower_bound | stulp > upper_bound)
}

# Revenue
# Salygines isskirtys
eilut <- isskirtys(duom$Revenue, 1.5)
duom[eilut,]

# Isskirtys
eilut <- isskirtys(duom$Revenue, 3)
duom[eilut,]

#boxplot(duom$Revenue)


# Expenses
# Salygines isskirtys
eilut <- isskirtys(duom$Expenses, 1.5)
duom[eilut,]

# Isskirtys
eilut <- isskirtys(duom$Expenses, 3)
duom[eilut,]

#boxplot(duom$Expenses)


# Profit
eilut <- isskirtys(duom$Profit, 1.5)
duom[eilut,]

# Isskirtys
eilut <- isskirtys(duom$Profit, 3)
duom[eilut,]

#boxplot(duom$Profit)


# Employees
# Salygines isskirtys
eilut <- isskirtys(duom$Employees, 1.5)
duom[eilut,]

# Isskirtys
```

```
eilut <- isskirtys(duom$Employees, 3)
duom[eilut,]

#duom %>% arrange(desc(Employees)) %>% head(10)

#boxplot(duom$Employees)

#duom_backup <- duom
#duom <- duom_backup

# Istriname isskirtis
duom <- duom[-c(eilut),]

duom$Outlier <- 0
# Naujos isskirtys, kurias deretu uzfiskuoti
eilut <- isskirtys(duom$Employees, 3)
duom$Outlier[c(eilut)] <- 1


ggplot(duom, aes(x=Revenue, y=Expenses, color=Employees)) + geom_point(size =
5.5,
                                                                       alpha =
0.9) +
  scale_y_continuous(labels = comma) +
  scale_x_continuous(labels = comma) +
  labs(title = "Sklaidos diagrama", y = "Išlaidos", x = "Pajamos") +
  scale_colour_continuous("Darbuotjų sk.")

# Matome, kad imones, su daug darbuotoju yra susimaisiusios tarp kitu, t. y. jos
# neissiskiria kitais bruozais.

# Employees
apras_stat[2,2] #vidurkis
apras_stat[2,3] #stand nuok
apras_stat[2,4] #mediana

# -||- po iskirciu salinimo:
apras_stat_red <- round(describe(duom[-c(1,2,3,6,7)]), 2)
apras_stat_red[2,3] #vidurkis
apras_stat_red[2,4] #stand nuok
apras_stat_red[2,5] #mediana

#Revenue
apras_stat[3,2] #vidurkis
apras_stat[3,3] #dispersija
apras_stat[3,4] #mediana

# -||- po iskirciu salinimo:
apras_stat_red[3,3] #vidurkis
apras_stat_red[3,4] #dispersija
apras_stat_red[3,5] #mediana
```

6. Duomenų normavimas

```
################################
# 5. NORMAVIMAS
################################

duom2 <- duom
```

```r
# Normavimo funkcija pagal min max
min_max_func <- function(stulp) {
  mini <- min(stulp)
  maxi <- max(stulp)
  normStulp <- (stulp - mini) / (maxi - mini)
}

# Normuojame Employees
normuota <- min_max_func(duom2$Employees)
duom2$Employees <- normuota

# Normuojame Revenue
normuota <- min_max_func(duom2$Revenue)
duom2$Revenue <- normuota

# Normuojame Profit
normuota <- min_max_func(duom2$Profit)
duom2$Profit <- normuota

# Normuojame Expenses
normuota <- min_max_func(duom2$Expenses)
duom2$Expenses <- normuota


# Normavimo funkcija pagal vidurki ir dispersija
norm_func <- function(stulp) {
  vid <- mean(stulp)
  stNuok <- sd(stulp)
  stulp_norm <- (stulp - vid) / stNuok
}

duom3 <- duom

# Normuojame Employees
normuota <- norm_func(duom3$Employees)
duom3$Employees <- normuota

# Normuojame Revenue
normuota <- norm_func(duom3$Revenue)
duom3$Revenue <- normuota

# Normuojame Profit
normuota <- norm_func(duom3$Profit)
duom3$Profit <- normuota

# Normuojame Expenses
normuota <- norm_func(duom3$Expenses)
duom3$Expenses <- normuota

# Taskines diagramos
#ggplot(duom, aes(x=Revenue, y=Expenses, color=Employees)) + geom_point(size =
4)
#ggplot(duom2, aes(x=Revenue, y=Expenses, color=Employees)) + geom_point(size =
4)
#ggplot(duom3, aes(x=Revenue, y=Expenses, color=Employees)) + geom_point(size =
4)


# Stulpelines diagramos
ggplot(duom, aes(x=Industry, y=Revenue, fill = Industry)) + geom_col() +
  scale_y_continuous(labels = comma) +
```

```
  theme(legend.position="none") +
  labs(title = "Nenormuotų duomenų stulpelinė diagrama", y = "Pajamos", x =
"Pramonė") +
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
                              "Government Services" = "Valstybinės įmonės",
"Health" =
                                "Sveikata", "IT Services" = "IT", "Retail" =
                                "Prekyba", "Software" = "Programinė įranga"))

ggplot(duom2, aes(x=Industry, y=Revenue, fill = Industry)) + geom_col() +
  theme(legend.position="none") +
  labs(title = "Normuotų duomenų min-max metodu stulpelinė diagrama",
       y = "Pajamos", x = "Pramonė") +
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
                              "Government Services" = "Valstybinės įmonės",
"Health" =
                                "Sveikata", "IT Services" = "IT", "Retail" =
                                "Prekyba", "Software" = "Programinė įranga"))

ggplot(duom3, aes(x=Industry, y=Revenue, fill = Industry)) + geom_col()+
  theme(legend.position="none") +
  labs(title = "Normuotų duomenų vidurkio ir dispersijos metodu stulpelinė
diagrama",
       y = "Pajamos", x = "Pramonė") +
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
                              "Government Services" = "Valstybinės įmonės",
"Health" =
                                "Sveikata", "IT Services" = "IT", "Retail" =
                                "Prekyba", "Software" = "Programinė įranga"))
```

7. Vizualizavimas

```
################################
# 6. VIZUALIZAVIMAS
################################
# TASKINES DIAGRAMOS

ggplot(duom, aes(x=Employees, y=Profit, color=Industry)) +
  geom_point(size = 6, alpha = 0.9) +
  theme_bw() + scale_x_continuous(labels = comma) +
  scale_y_continuous(labels = comma) +
  labs(title = "Įmonių sklaida pagal darbuotojus ir pelną",
       y = "Pelnas", x = "Darbuotojai") +
  scale_color_discrete(name = "Pramonė", labels = c("Construction" = "Statybos",
                                "Financial Services" = "Finansai",
                              "Government Services" = "Valstybinės įmonės",
"Health" =
                                "Sveikata", "IT Services" = "IT", "Retail" =
                                "Prekyba", "Software" = "Programinė įranga"))

ggplot(duom, aes(x=Revenue, y=Expenses, color=Industry)) +
  geom_point(size = 5.5, alpha = 0.8) +
  theme_bw() + scale_x_continuous(labels = comma) +
  scale_y_continuous(labels = comma)+
  labs(title = "Įmonių sklaida pagal pajamas ir išlaidas",
       y = "Išlaidos", x = "Pajamos") +
  scale_color_discrete(name = "Pramonė", labels = c("Construction" = "Statybos",
```

```
                          "Financial Services" = "Finansai",
                          "Government Services" = "Valstybinės įmonės", "Health" =
                            "Sveikata", "IT Services" = "IT", "Retail" =
                            "Prekyba", "Software" = "Programinė įranga"))


test <- duom
test$Darbuotojai<- test$Employees

ggplot(test, aes(x=Revenue, y=Profit, size = Darbuotojai, color=Employees)) +
  geom_point(alpha = 0.9) +
  theme_bw() + scale_x_continuous(labels = comma) +
  scale_y_continuous(labels = comma) +
  labs(title = "Įmonių sklaida pagal pajamas ir pelną",
       y = "Pelnas", x = "Pajamos") +
  scale_colour_continuous("Darbuotojai")


# DAZNIU DIAGRAMOS
ggplot(duom, aes(x=Employees)) +
  geom_histogram(aes(fill = Industry), binwidth = 50, colour = "black", size =
0.5) +
  theme_bw() +
  scale_fill_discrete(name = "Pramonė", labels = c("Construction" = "Statybos",
                                      "Financial Services" = "Finansai",
                                      "Government Services" = "Valstybinės
įmonės", "Health" =
                                        "Sveikata", "IT Services" = "IT", "Retail"
=
                                        "Prekyba", "Software" = "Programinė
įranga")) +
  labs(title = "Įmonių kiekis pagal darbuotojų skaičių",
       y = "Kiekis", x = "Darbuotojai")


ggplot(duom, aes(x=Inception)) +
  geom_bar(aes(fill = Industry), colour = "black", size = 0.5) +
  scale_x_continuous(breaks = seq(1999, 2014, by = 1)) +
  scale_fill_discrete(name = "Pramonė", labels = c("Construction" = "Statybos",
                                          "Financial Services" = "Finansai",
                                "Government Services" = "Valstybinės įmonės",
"Health" =
                                  "Sveikata", "IT Services" = "IT", "Retail" =
                                  "Prekyba", "Software" = "Programinė įranga")) +
  labs(title = "Įsikūrusių įmonių kiekis pagal metus",
       y = "Kiekis", x = "Metai")


#Suma
ggplot(duom, aes(x=Industry, y=Employees, fill = Industry)) +
  geom_col() + theme(legend.position="none") +
  labs(title = "Darbuotojų sk. pagal pramonės šaką (Suma)", y = "Darbuotojų
sk.", x = "Pramonės šaka") +
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
                              "Government Services" = "Valstybinės įmonės",
"Health" =
                                "Sveikata", "IT Services" = "IT", "Retail" =
                                "Prekyba", "Software" = "Programinė įranga"))

ggplot(duom, aes(x=Industry, y=Profit, fill = Industry)) +
```

```
  geom_col() + theme(legend.position="none") + scale_y_continuous(labels =
comma) +
  labs(title = "Pelnas pagal pramonės šaką (Suma)", y = "Pelnas", x = "Pramonės
šaka") +
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
                              "Government Services" = "Valstybinės įmonės",
"Health" =
                                "Sveikata", "IT Services" = "IT", "Retail" =
                                "Prekyba", "Software" = "Programinė įranga"))


ggplot(duom, aes(x=Industry, y=Expenses, fill = Industry)) +
  geom_col() + theme(legend.position="none") + scale_y_continuous(labels =
comma) +
  labs(title = "Išlaidos pagal pramonės šaką (Suma)", y = "Išlaidos", x =
"Pramonės šaka") +
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
                              "Government Services" = "Valstybinės įmonės",
"Health" =
                                "Sveikata", "IT Services" = "IT", "Retail" =
                                "Prekyba", "Software" = "Programinė įranga"))


ggplot(duom, aes(x=Industry, fill = Industry)) +
  geom_bar() + theme(legend.position="none") +
  labs(title = "Darbuotojų sk. pagal pramonės šaką (Suma)", y = "Kiekis", x =
"Pramonės šaka") +
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
                              "Government Services" = "Valstybinės įmonės",
"Health" =
                                "Sveikata", "IT Services" = "IT", "Retail" =
                                "Prekyba", "Software" = "Programinė įranga"))


#Vidurkiai
test <- as.data.frame(aggregate(duom$Employees, list(duom$Industry), FUN=mean))
names(test) <- c("Industry", "Mean employees")
ggplot(test, aes(x=Industry, y=`Mean employees`, fill = Industry)) +
  geom_col() + theme(legend.position="none") +
  labs(title = "Vidutinis darbuotojų sk. pagal pramonės šaką", y = "Darbuotojų.
sk.", x = "Pramonės šaka")+
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
                              "Government Services" = "Valstybinės įmonės",
"Health" =
                                "Sveikata", "IT Services" = "IT", "Retail" =
                                "Prekyba", "Software" = "Programinė įranga"))


test <- as.data.frame(aggregate(duom$Profit, list(duom$Industry), FUN=mean))
names(test) <- c("Industry", "Mean profit")
ggplot(test, aes(x=Industry, y=`Mean profit`, fill = Industry)) +
  geom_col() + theme(legend.position="none") + scale_y_continuous(labels =
comma) +
  labs(title = "Vidutinis pelnas pagal pramonės šaką", y = "Pelnas", x =
"Pramonės šaka") +
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
```

```
                                "Government Services" = "Valstybinės įmonės",
"Health" =
                                  "Sveikata", "IT Services" = "IT", "Retail" =
                                  "Prekyba", "Software" = "Programinė įranga"))


test <- as.data.frame(aggregate(duom$Expenses, list(duom$Industry), FUN=mean))
names(test) <- c("Industry", "Mean expenses")
ggplot(test, aes(x=Industry, y=`Mean expenses`, fill = Industry)) +
  geom_col() + theme(legend.position="none") + scale_y_continuous(labels =
comma) +
  labs(title = "Vidutinės išlaidos pagal pramonės šaką", y = "Išlaidos", x =
"Pramonės šaka") +
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
                                "Government Services" = "Valstybinės įmonės",
"Health" =
                                  "Sveikata", "IT Services" = "IT", "Retail" =
                                  "Prekyba", "Software" = "Programinė įranga"))



pop_valst <- duom %>% group_by(State) %>%
  summarise(Kiekis_vals = n()) %>%
  arrange(desc(Kiekis_vals)) %>%
  head(10)


# Staciakampes diagramos
ggplot(duom, aes(x = Industry, y = Employees, fill = Industry)) + geom_boxplot()
+
  theme(legend.position="none") +
  labs(title = "Darbuotojų sk. pagal pramonės šaką", y = "Darbuotojų sk.", x =
"Pramonės šaka") +
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
                                "Government Services" = "Valstybinės įmonės",
"Health" =
                                  "Sveikata", "IT Services" = "IT", "Retail" =
                                  "Prekyba", "Software" = "Programinė įranga"))


ggplot(duom, aes(x = Industry, y = Revenue, fill = Industry)) + geom_boxplot() +
  scale_y_continuous(labels = comma) + theme(legend.position="none") +
  labs(title = "Pelnas pagal pramonės šaką", y = "Pelnas", x = "Pramonės šaka")
+
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
                                "Government Services" = "Valstybinės įmonės",
"Health" =
                                  "Sveikata", "IT Services" = "IT", "Retail" =
                                  "Prekyba", "Software" = "Programinė įranga"))

ggplot(duom, aes(x = Industry, y = Growth, fill = Industry)) + geom_boxplot() +
  scale_y_continuous(labels = comma) + theme(legend.position="none") +
  labs(title = "Augimas pagal pramonės šaką", y = "Augimas", x = "Pramonės
šaka") +
  scale_x_discrete(labels = c("Construction" = "Statybos", "Financial Services"
= "Finansai",
                                "Government Services" = "Valstybinės įmonės",
"Health" =
```

```
                                    "Sveikata", "IT Services" = "IT", "Retail" =
                                    "Prekyba", "Software" = "Programinė įranga"))
```

8. Koreliacijos

```
################################
# 7. KORELIACIJOS
################################
m <- cor(duom[,c(4,5,8,9,10,11)])
colnames(m) <- c("Metai", "Darb. sk.", "Pajamos", "Išlaidos", "Pelnas",
"Augimas")
rownames(m) <- c("Metai", "Darb. sk.", "Pajamos", "Išlaidos", "Pelnas",
"Augimas")
corrplot(m, method = "color", title = "Požymių koreliacijos", mar=c(0,0,1,0))
```