Vilniaus Universitetas

Pirminė duomenų aibės analizė

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

Vilnius, 2022

Turinys

1 Tikslas ir uždaviniai

Tikslas:

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

Uždaviniai:

Surasti praleistas duomenų reikšmes ir pasirinkus tinkamus metodus jas užpildyti.

Apskaičiuoti aprašomosios statistikos charakteristikas, palyginti jas tarp skirtingų pramonės šalių.

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

Pritaikyti duomenų normavimo metodus.

Atlikti aibės vizualią analizę.

Ištirti koreliacijas tarp duomenų aibės požymių.

2 Duomenų aibė

Duomenų aibę sudaro duomenys apie 500 įmonių su tokiais požymiais:

„ID“ - (kategorinis, nominalusis) įmonę duomenyse identifikuojantis kodas

„Name“ – (kategorinis, nominalus) įmonės pavadinimas

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

„Inception“ – (kiekybinis, diskretusis) įmonės įkūrimo metai

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

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

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

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

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

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

3 Atliktos analizės aprašymas

3.1 Praleistos reikšmės

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

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

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

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

Kategoriniuose kintamamuosiuose esančios praleistos reikšmės paliktos nekeistos.

3.2 Aprašomoji statistika

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

Lentelė 1 Aprašomosios statistikos charakteristikos duomenų aibei

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

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

Lentelė 2 Aprašomosios statistikos charakteristikos atskirai kiekvienai pramonės šakai

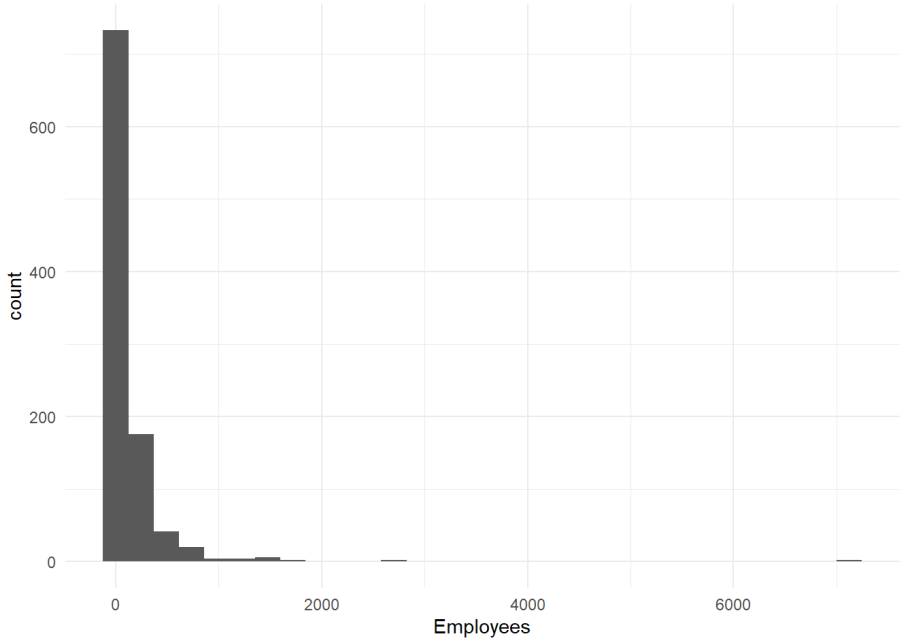
	Pramonės šaka	stand. nuokrypis	vidurkis	mediana	min	max
Inception	Construction	3.53	2009.94	2011	1999	2014
Employees	Construction	59.43	61.26	37.5	5	272
Revenue	Construction	2404913.29	9158737.12	9055058.5	4419277	18429577
Expenses	Construction	1793321.66	4453204.5	4506975.5	214470	8213905
Profit	Construction	2805089.4	4705532.62	4573280.5	96073	12616182
Growth	Construction	3.07	10.06	10	5	19
Inception	Financial Services	2.71	2009.83	2010	2001	2014
Employees	Financial Services	261.95	183.88	79	3	1387
Revenue	Financial Services	1935037.65	10711858.77	11175012.5	5387469	14330107
Expenses	Financial Services	1521249.39	2351572.02	2379097	223602	6212849
Profit	Financial Services	2166383.23	8363033.13	8348842.5	3259485	12205097
Growth	Financial Services	2.69	16.68	17	10	23
Inception	Government Services	3	2010.3	2011	2000	2014
Employees	Government Services	233.63	172.72	99	13	1224
Revenue	Government Services	2342556.62	9436792.34	9707475	4637647	15188113

Expenses	Government Services	2055429.6	4741746.34	4790732.5	1243956	9860686
Profit	Government Services	2776630.41	4605150.06	4776526	46851	10565044
Growth	Government Services	2.87	5	5	-3	11
Inception	Health	3.01	2010.89	2012	2000	2014
Employees	Health	308.32	205.51	86.5	6	1600
Revenue	Health	1978819.76	8811121.94	8855709.5	1614585	15312302
Expenses	Health	1892100.07	5881840.64	6162150.5	1323005	9712296
Profit	Health	2075213.51	2929281.3	2514786.5	12434	9174395
Growth	Health	2.6	6.59	6	0	14
Inception	IT Services	3.46	2009.9	2011	1999	2014
Employees	IT Services	257	107.81	52	2	2670
Revenue	IT Services	1950075.52	14175582.57	14121713	9691133	21810051
Expenses	IT Services	2043621.79	4149153.46	4068630	187655	9046498
Profit	IT Services	3003002.6	10019629.86	10160479	1841685	19624534
Growth	IT Services	3.09	21.4	21	15	30
Inception	Retail	3.38	2010.42	2011	1999	2014
Employees	Retail	1044.76	213.48	28	1	7125
Revenue	Retail	2183839.08	11581242.32	11654196	7307243	15880376
Expenses	Retail	1801630.91	4156855.09	4545730.5	968518	7957743
Profit	Retail	2897292.12	7482727.9	7326357	815381	13369247
Growth	Retail	2.59	12.5	12	8	19
Inception	Software	3.18	2010.08	2011	2000	2014
Employees	Software	179.75	121.06	58	3	850
Revenue	Software	2646904.18	7914512.71	8304480	1835717	14229411
Expenses	Software	1940555.97	3822601.62	4175332	71219	8007771
Profit	Software	2951684.76	4091911.1	3952602	68862	11902072
Growth	Software	2.89	18.89	19	13	26

3.3 Išskirčių analizė

Išskirtys vertintos naudojant vidinį $[Q_1 - 1.5H; Q_3 + 1.5H]$ ir išorinius $[Q_1 - 3H; Q_3 + 3H]$ barjerus, kur Q_1, Q_3 – atitinkamai pirmas ir trečias kvartiliai, $H = Q_3 - Q_1$ – interkvartilinis plotis.

Naudojant vidinį barjerą rastos 4 įmonės išsiskiriančios pagal pajamas (stulp. „Revenue“), 2 įmonės išsiskiriančios pagal pelną (stulpelis „Profit“) ir 60 įmonių išsiskiriančiu pagal didelį darbuotojų skaičių. Naudojant išorinį barjerą rastos 36 įmonės išsiskiria pagal darbuotojų skaičių. Darbuotojų skaičiaus įmonėje histogramoje galima pastebėti išsiskiriančias įmones (žr. 1 Pav.)



1 Pav. Darbuotojų skaičiaus histograma

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

Lentelėje žemiau (žr. Lentelė 3) pateikta kokiai pramonės šakai priklauso įmonės pagal išorinį barjerą išsiskiriančios bent vienu požymiu (šiuo atveju visos įmonės išsiskiria darbuotoju skaičiumi).

Lentelė 3 Pagal bet kurio požymio išorinį barjerą išsiskiriančių įmonių kiekis pagal pramonės šaką

Pramonės šaka	Pašalintų reikšmių skaičius
Financial Services	8
Government Services	7
Health	11
IT Services	4
Retail	1
Software	5

Pakartotinai apskaičiuotos duomenų aibės aprašomosios statistikos charakteristikos jeigu iš duomenų aibės būtų pašalintos prieš tai minėtos pagal išorinį barjerą išsiskiriančios įmonės (žr. Lentelė 4). Didžiausias pokytis pastebėtas darbuotojų kiekyje – pašalimus išskirtis darbuotojų kiekio vidurkis sumažėjo 45%, standartinis nuokrypis 79%, mediana – 11%.

Lentelė 4 Procentinis aprašomosios statistikos charakteristikų pokytis pašalinus išskirtis

	stand. nuokrypis	vidurkis	mediana	min	max
Inception	0.5	0	0	0	0
Employees	-79.52	-45.51	-12.28	0	-94.34
Revenue	-1.28	0.65	0.76	13.7	0
Expenses	-0.24	0.36	0	0	0
Profit	-0.41	0.85	1.45	0	-5.95
Growth	-0.58	1.5	6.67	0	0

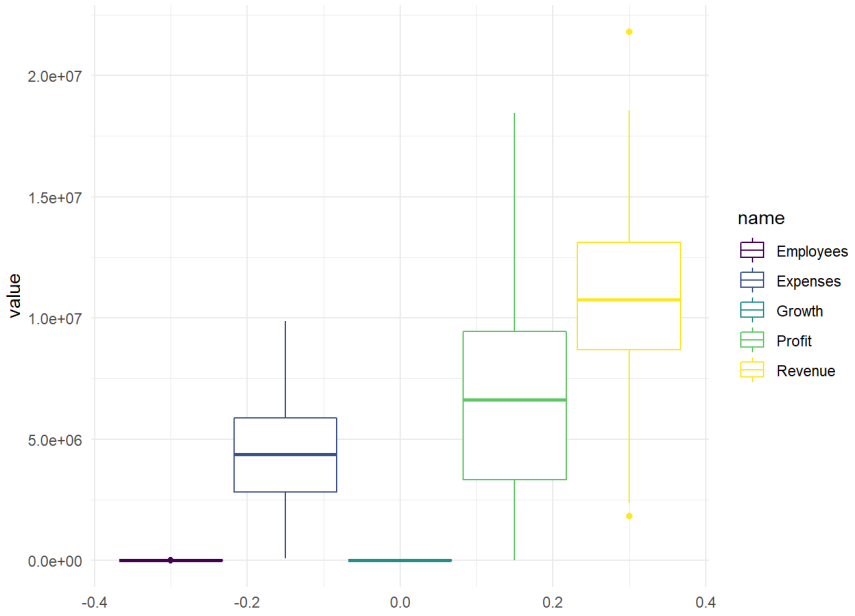
3.4 Duomenų normavimas

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

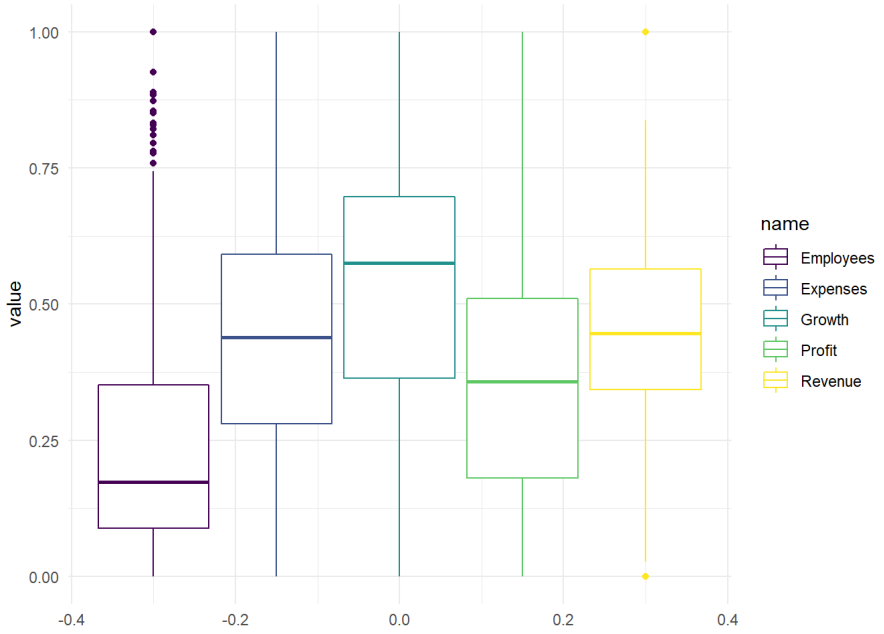
Duomenys sunormuoti naudojant min-max normavimą $x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$ ir normavimą pagal vidurkį ir dispersiją (standartizavimas) $x_{norm} = \frac{x - \bar{x}}{\sqrt{\sigma^2}}$, kur σ^2 - požymio vidurkis požymio dispersija, $\bar{x}$ – požymio vidurkis.

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

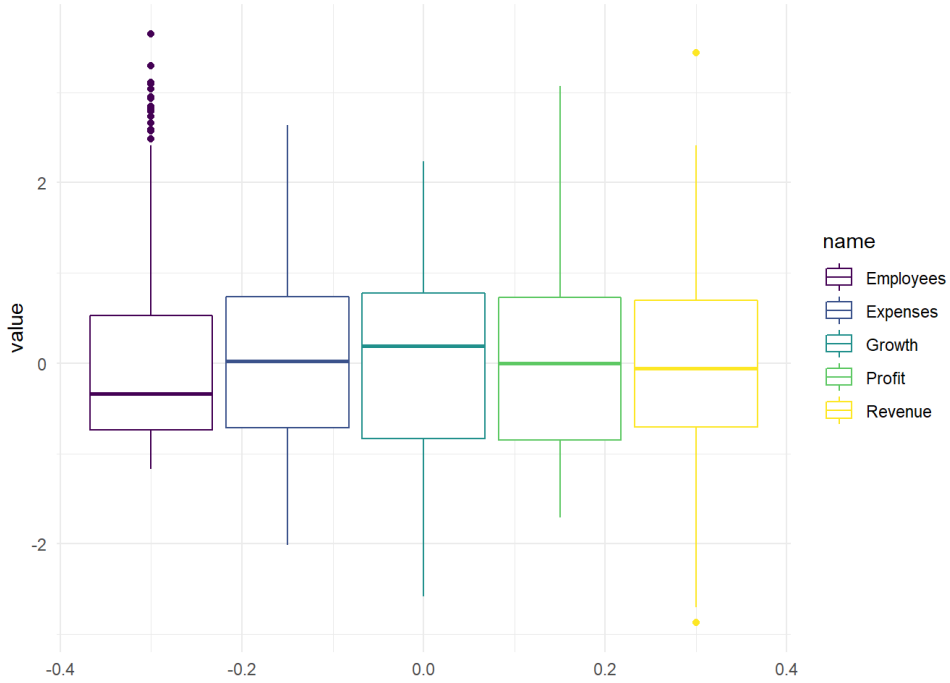
Dėl didelio kiekio išskirčių (žr. Išskirčių analizė) nerekomenduojama taikyti standartizavimo metodą darbuotojų skaičiaus įmonėje požymiui.



2 Pav. Kiekybinių požymių stačiakampė diagrama prieš atliekant normavimą



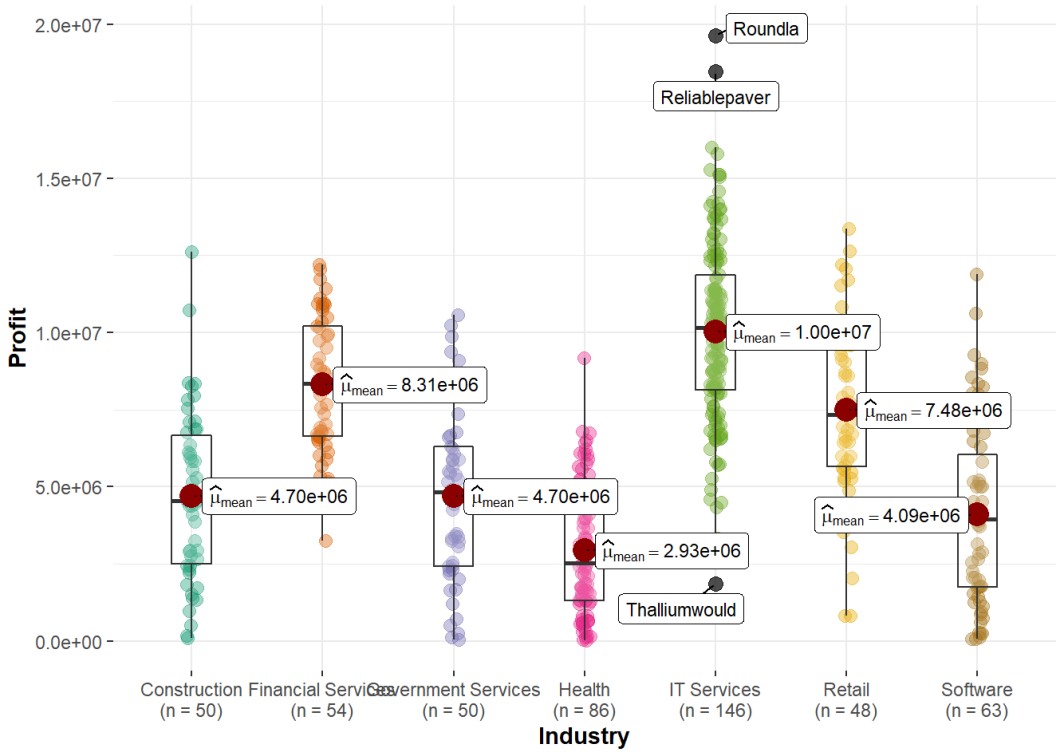
3 Pav. Stačiakampė diagrama atlikus min-max normavimą



4 Pav. Stačiakampė diagrama atlikus standartizaciją

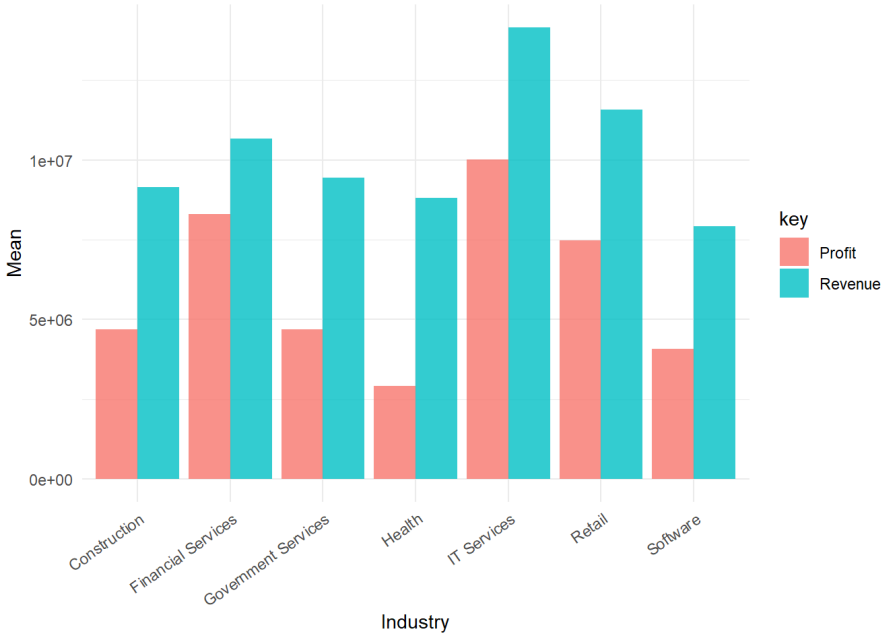
3.5 Vizuali analizė

Stačiakampėmis diagramomis pavaizduotas įmonių pelno pasiskirstymas pagal pramonės šaką (žr. 5 Pav.). Pastebimas didesnis pelnas IT Services, Financial Services ir Retail pramonės šakose.



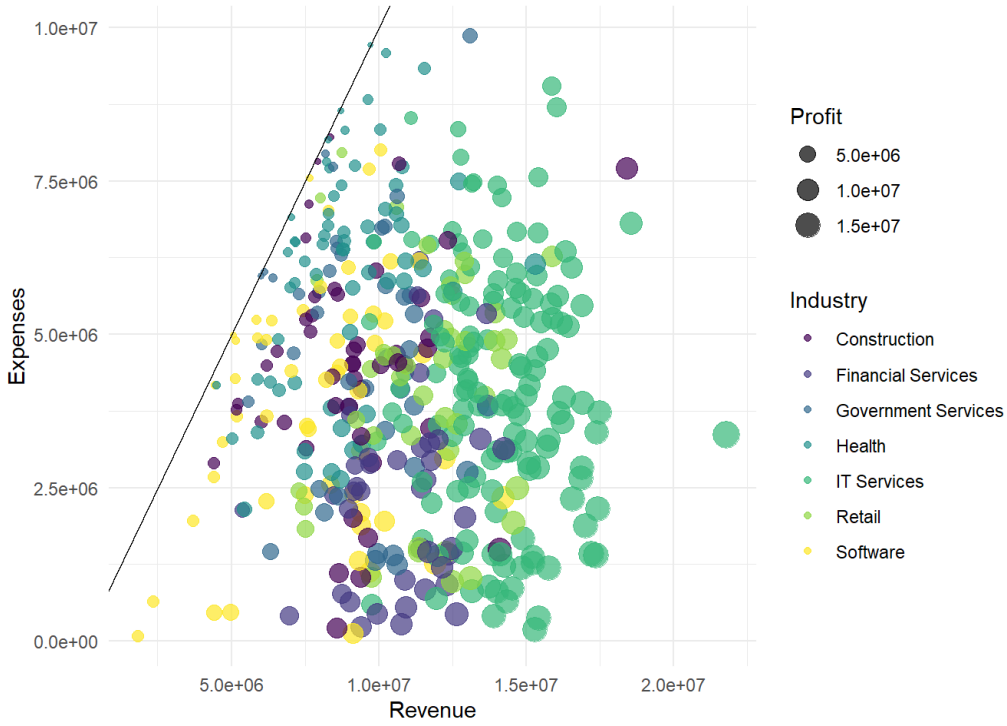
5 Pav. Įmonių pelnas pagal pramonės šaką

Stulpeline diagrama kiekvienai pramonės šakai pavaizduotas vidutinės gautos pajamos kartu su vidutiniu pelnu (žr. 6 Pav.). Pastebima, kad vidutiniškai nė viena pramonės šaka nepatyrė nuostolių. Taip pat rasta, kad Health srityje pelnas sudaro mažesnę dalį pajamų negu kitose pramonės šakose



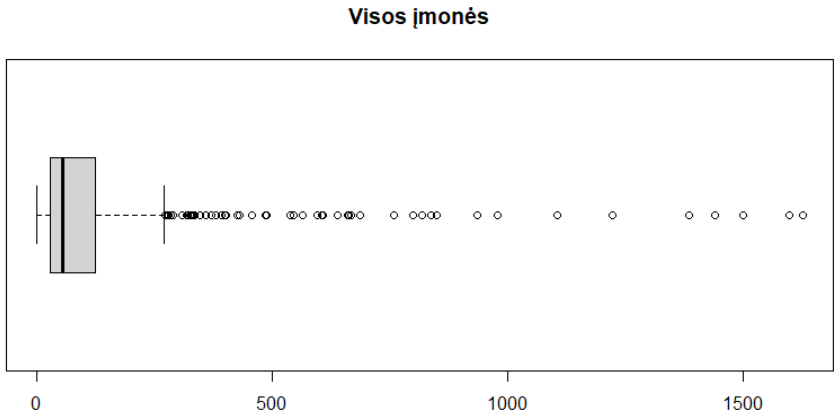
6 Pav. Įmonių vidutinių pajamų ir išlaidų stulpelinė diagrama pagal pramonės šaką

Sklaidos diagrama pavaizduotas įmonių pajamų ir išlaidų sklaidos diagrama kartu su palyginamąja tiese (žr. 7 Pav.) Iš grafiko matome, kad jokios įmonės duomenų aibėje nepatyrė nuostolių, pelningiausios yra IT Services įmonės.



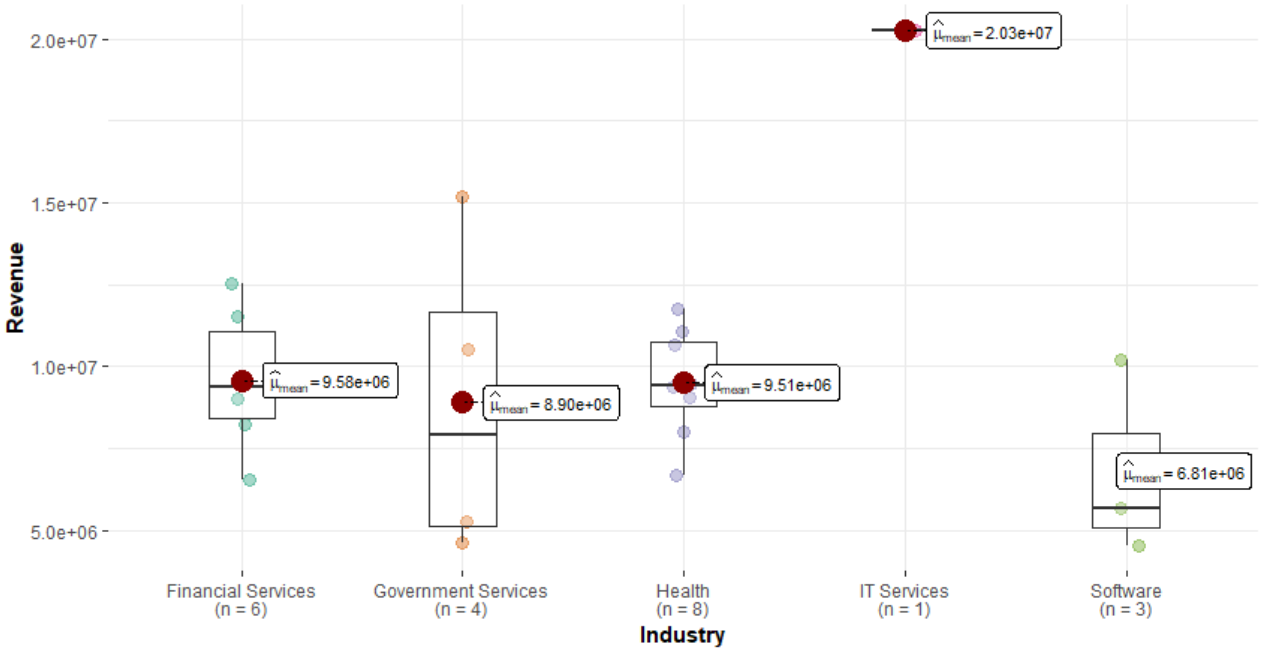
7 Pav. Įmonių pajamų ir išlaidų sklaidos diagrama

Stačiakampėmis diagramomis pavaizduotas darbuotojų skaičiaus įmonėse pasiskirstymas įmonėse (žr. 8 Pav.). Dėl didelio įmonių tankio arti nulio darbuotojų, galima išskaidyti įmones į dvi grupes. Mažos ir didelės įmones atskiriamos išrikiavus verslus mažėjimo tvarka pagal darbuotojų skaičių, didelių įmonių grupei priskiriant 15% daugiausiai įdarbinusių firmų.

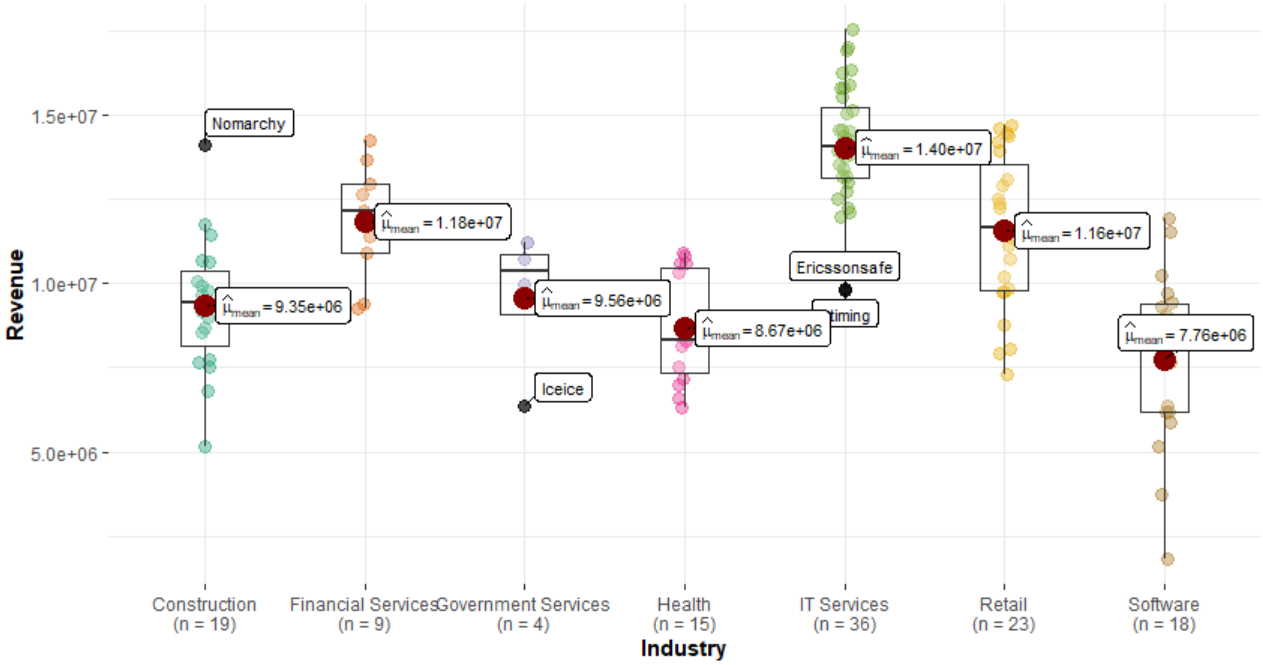


8 Pav. Darbuotojų skaičiaus pasiskirstymas.

Didelių (žr. 11 Pav.) ir mažų (žr. 11 Pav.) įmonių skaitinės charakteristikos stipriai nesiskiria. Tarp didelių įmonių nėra Retail pramonės šakos. Norint įtraukti Retail pramonės šakos firmas reikia didinti darbuotojų skaičiaus ribą, tačiau didinant šią ribą didėja šių įmonių įvairių skaitinių požymių dispersija.

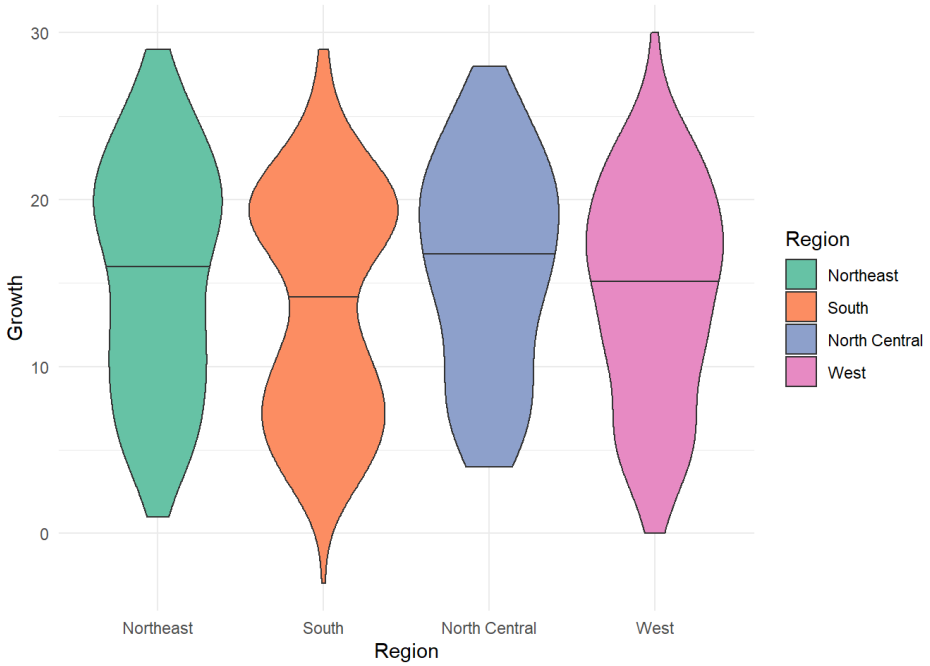


9 Pav. Didelių įmonių pajamų pasiskirstymas pagal pramonės šaką

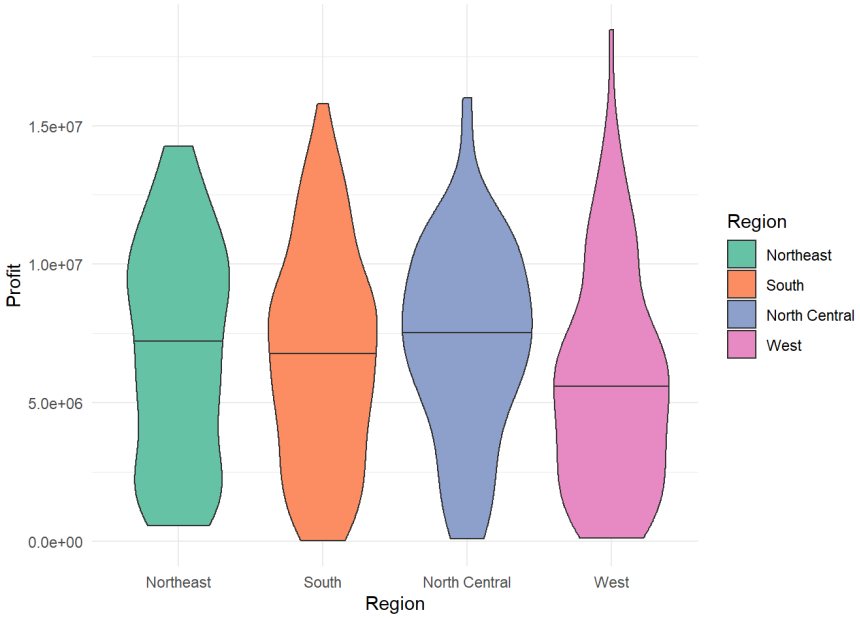


10 Pav. Mažų įmonių pajamų pasiskirstymas pagal pramonės šaką

JAV valstijos padalintos į keturis regionus ir smuiko formos grafikais kiekvienam regionui pavaizduotas įmonių augimo (žr. 11 Pav.) ir pelno (žr. 12 Pav.) pasiskirstymas (horizontalia linija papildomai pažymint medianinę reikšmę). Ryškių įmonių augimo ir pelno skirtumų tarp JAV regionų nepastebėta.

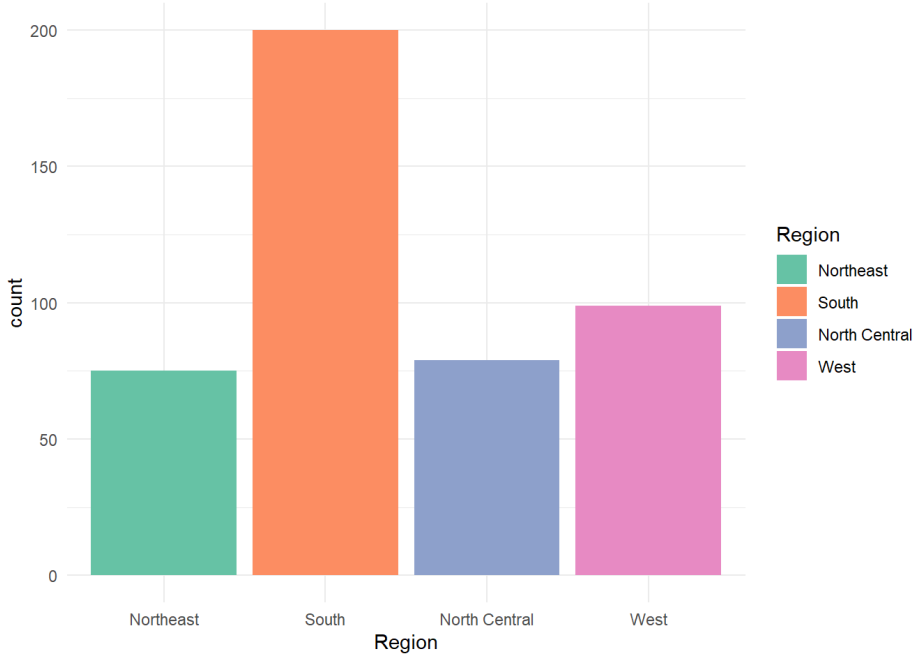


11 Pav. Įmonių augimo pasiskirstymas pagal JAV regionus

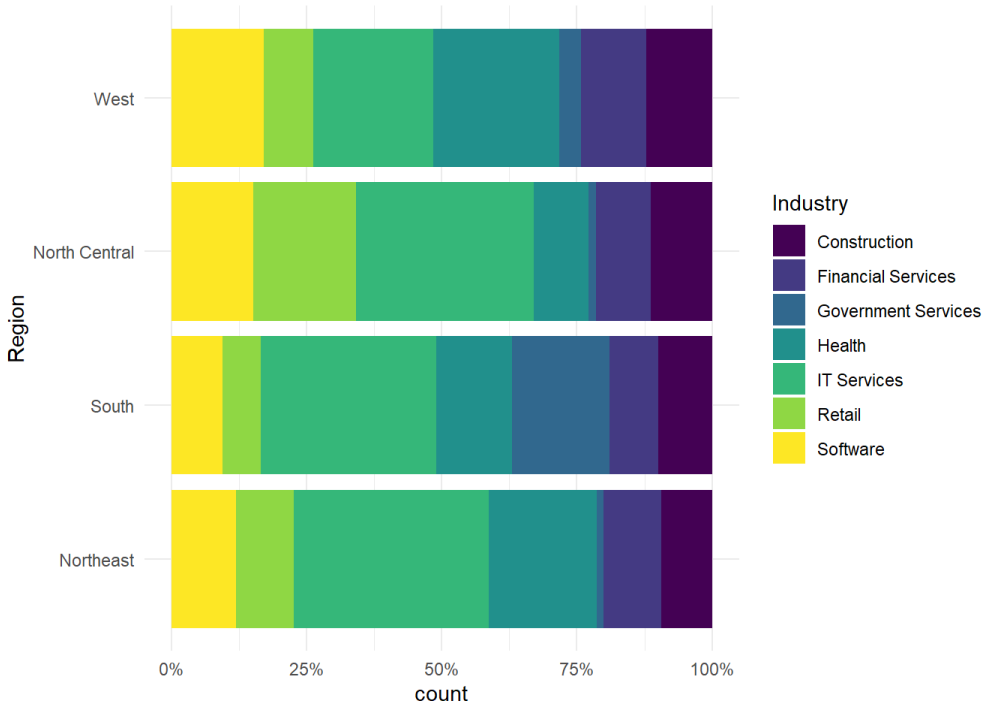


12 Pav. Įmonių pelno pasiskirstymas pagal JAV regioną

Stulpeline diagrama pavaizduotas įmonių kiekviename regione skaičius (žr. 13 Pav.). Didžiausia dalis įmonių duomenų aibėje yra iš pietinio JAV regiono. Papildomai kiekvienam regionui stulpeline diagrama pavaizduota kokią dalį įmonių sudaro tam tikrai pramonės šakai priklausančios įmonės (žr. 14 Pav.). Grafike galima matyti, kad įmonių pasiskirstymas labai panašus visuose 4 regionuose.

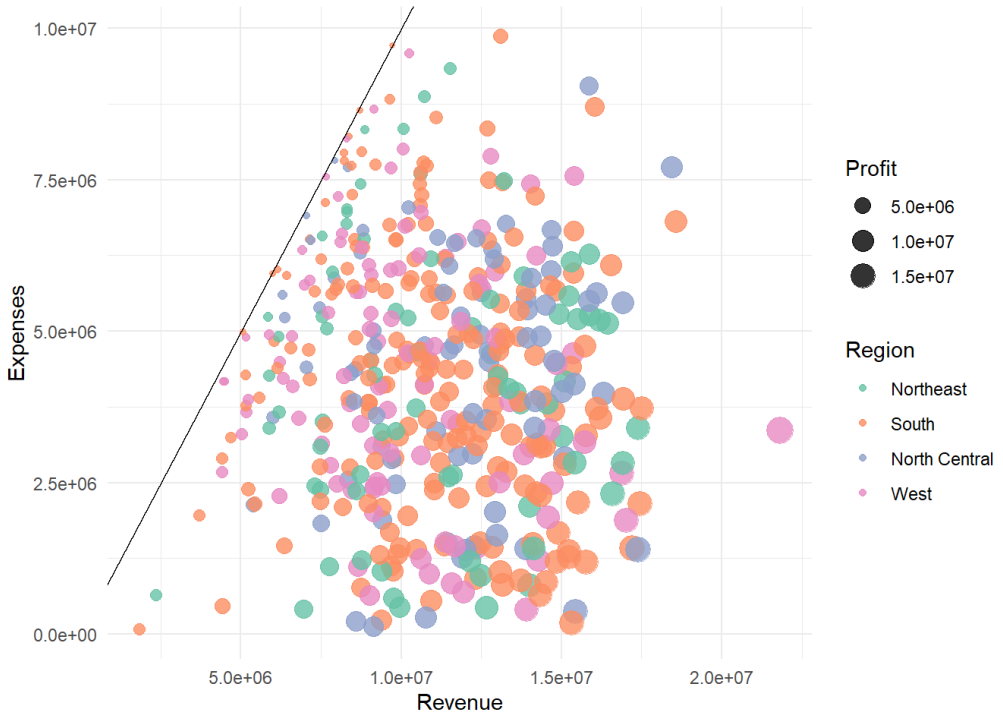


13 Pav. Įmonių skaičius pagal JAV regioną



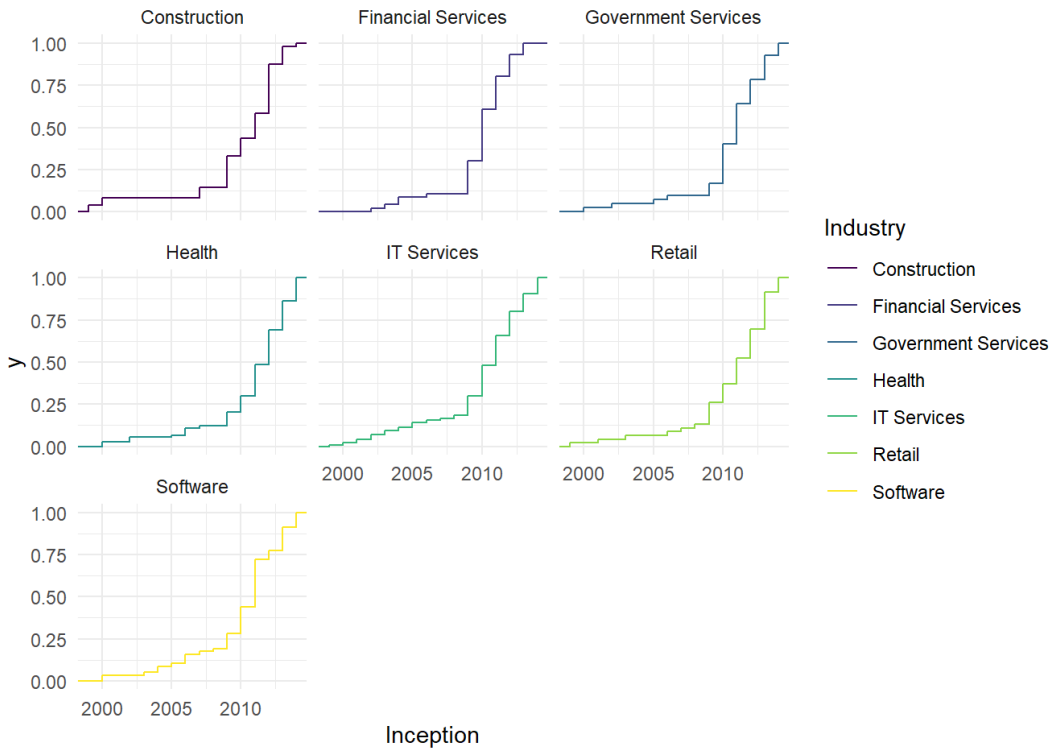
14 Pav. Įmonių pasiskirstymas pagal pramonės šaką kiekviename JAV regione

Dar kartą nubrėžta įmonių pajamų ir išlaidų sklaidos diagrama, tačiau šį kart nuspalvinant taškus pagal regioną (žr. 15 Pav.). Duomenų atsiskyrimas daug mažesnis negu taškus nuspalvinant pagal pramonės šaką (7 Pav.)



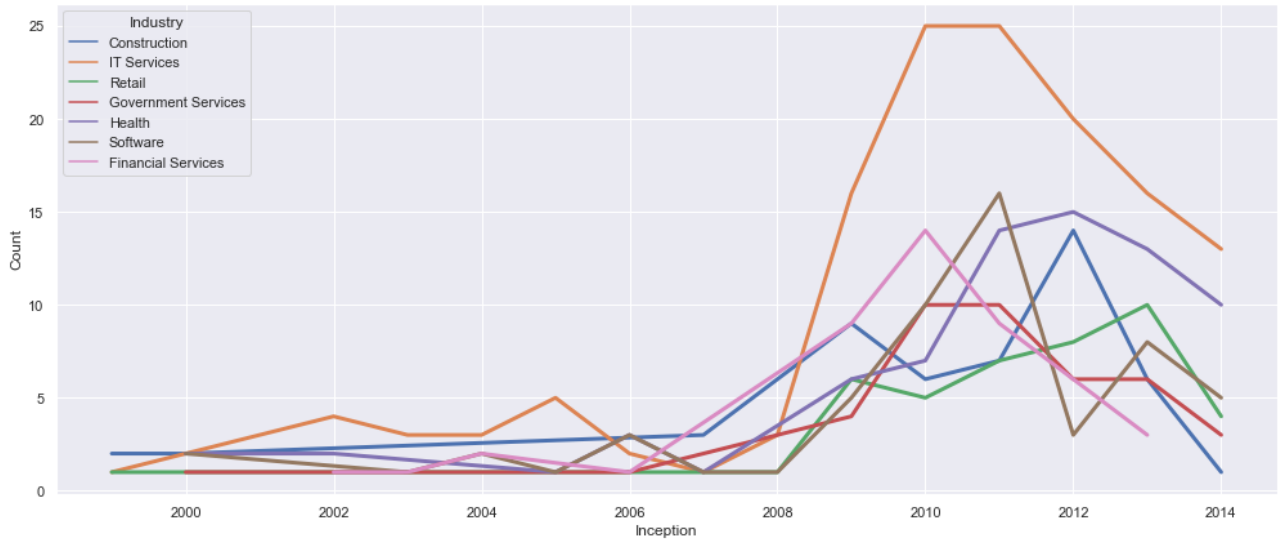
15 Pav. Įmonių pajamų ir išlaidų sklaidos diagrama pagal JAV regionus

Pavaizduota įmonių įsikūrimo metų empirinė pasiskirstymo funkcija (žr. 16 Pav.). Matoma, kad didžioji dalis įmonių, esančių duomenų aibėje, įkurta nuo maždaug 2009-2010 metų. Ši tendencija galioja visoms pramonėms šakomis.



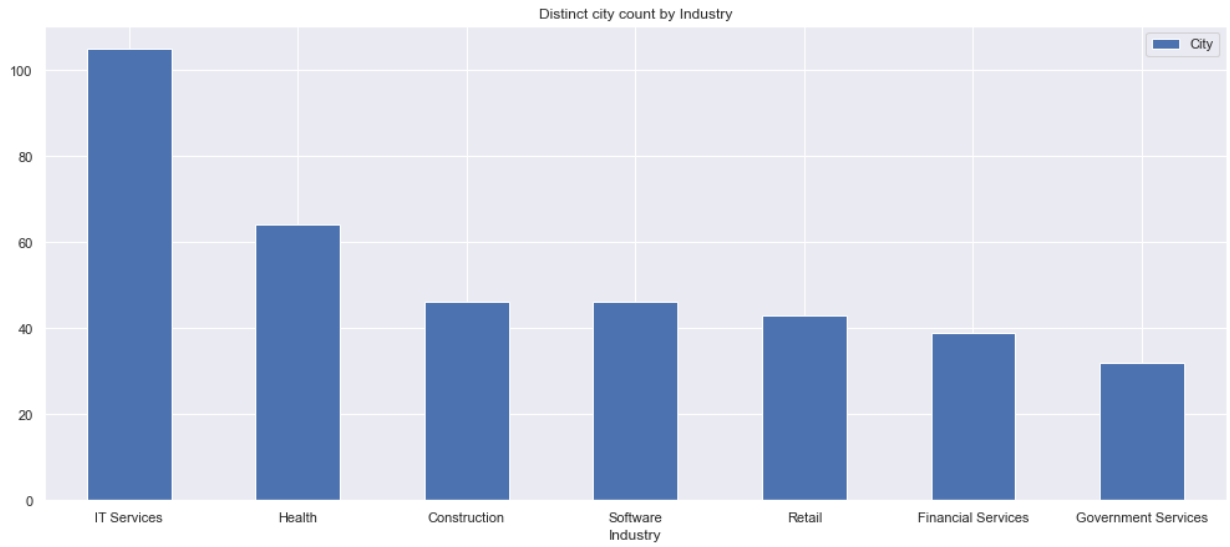
16 Pav. Įmonių įkūrimo metų empirinė pasiskirstymo funkcija pagal pramonės šaką

Linijine diagrama pateiktas kiekvienais metais įsikūrusių įmonių kiekis pagal pramonės šaką (žr. 17 Pav.). Matoma, kad daugiausia įmonių buvo įkurta 2010 ir 2011 metais IT Services srityje. Visoms pramonės šakoms matoma, kad 2013 ir 2014 metais įmonių duomenų aibėje įkurta mažiau negu kelis metus prieš tai.

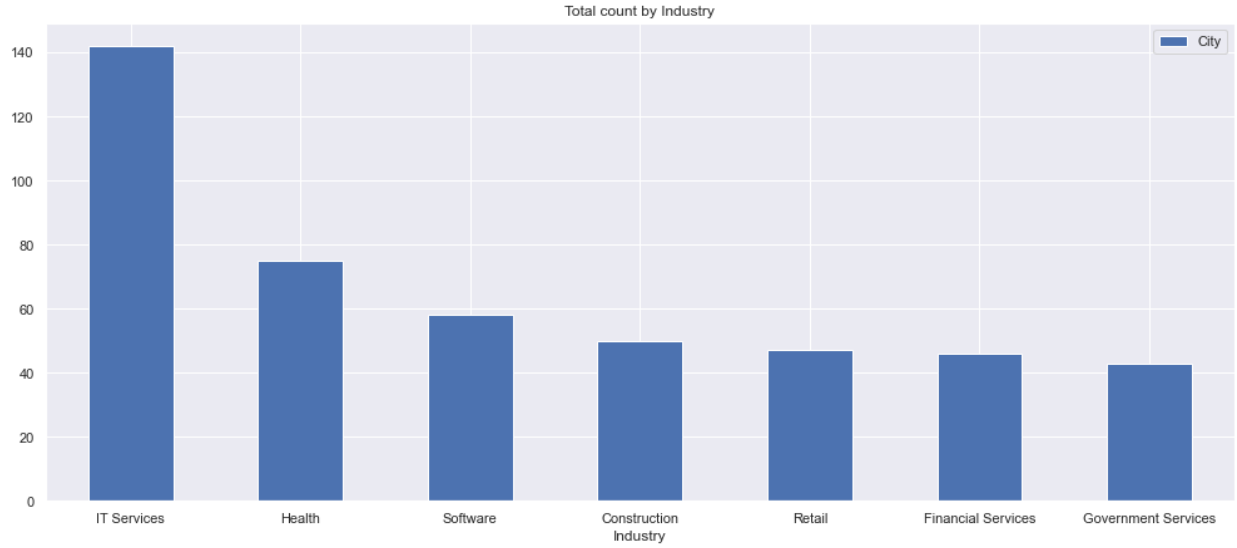


17 Pav. Kiekvienais metais įsikūrusių įmonių skaičius pagal pramonės šaką

Stulpeline diagrama pavaizduotas skirtingų miestų, kuriose yra įsikūrusios įmonės, skaičius pagal pramonės šakas (žr. 18 Pav.). Pastebima, kad gautas pasiskirstymas tik minimaliai skiriasi nuo bendro įmonių skaičiaus skirtingose pramonės šakose pasiskirstymo (žr. 19 Pav.). Taip yra todėl, nes didžiajai daliai miestų duomenų aibėje turimi duomenys tik apie vieną ten įsikūrusią įmonę.



18 Pav. Skirtingų miestų skaičius pagal pramonės šaką



19 Pav. Įmonių skaičius pagal pramonės šaką

3.6 Požymių koreliacijos

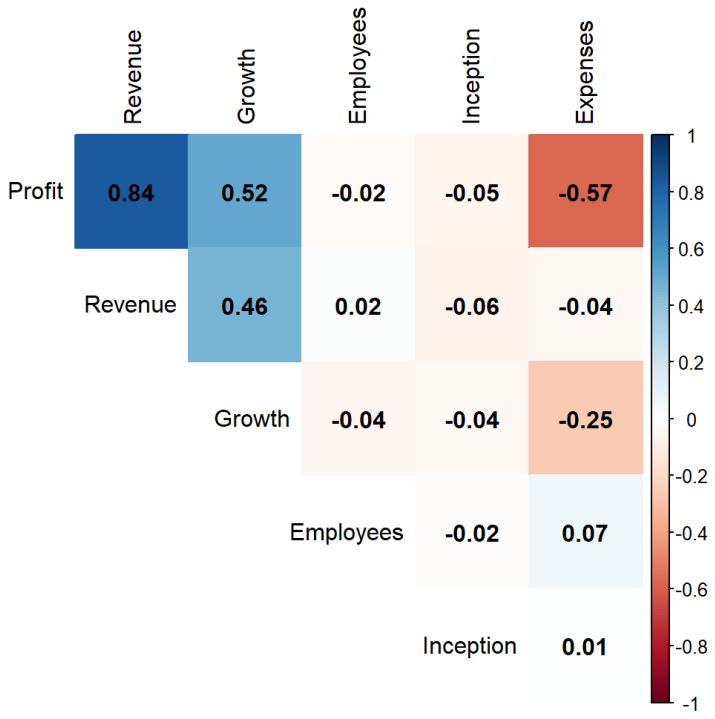
Tarp skaitinių rodiklių apskaičiuotos Pirsono koreliacijos koeficientų reikšmės (angl. Pearson correlation coefficient).

Gauti rezultatai pateikti lentelėje (žr. Lentelė 5). Rasta stipri teigiama tarp pajamų ir pelno (r = 0.84). Taip pat rastos vidutinio stiprumo teigiamos koreliacijos tarp pelno ir augimo (r = 0.52), pajamų ir augimo (r = 0.46) ir neigiama koreliacija tarp pelno ir išlaidų (r = -0.57)

Šios reikšmės papildomai vizualizuotos koreliacijų diagrama (žr. 20 Pav.).

Lentelė 5 Pirsono koreliacijos koeficientai tarp skaitinių įmonių požymių

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



20 Pav. Pirsono koreliacijos tarp skaitinių požymių koeficientai

4 Išvados

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

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

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

Galimas šios problemos sprendimas išskirti įmones į mažas ir dideles pagal darbuotojų skaičių. Tam reikia parinkti mažos ir didelės įmonės darbuotojų skaičiaus ribą arba naudojant procentinę dalį įmonių pagal dydį. Atskyrus įmones mažų ir didelių įmonių skaitinės charakteristikos stipriai nepasikeičia, tačiau reikia atkreipti dėmesį į požymių dispersiją, kuri priklauso nuo pasirinktos ribos.

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

Atlikus vizualią duomenų aibę rasta, kad duomenų aibėje didžiausią pelną gauna IT Services įmonės, tačiau pelningos ir visos kitos pramonės šakos. Lyginant JAV regionus rasta, kad didžioji dalis duomenų yra iš pietinio JAV regiono, tačiau pagal kitus požymius ryškių skirtumų tarp JAV regionų nerasta. Rasta, kad tik labai maža dalis įmonių duomenų aibėje įkurtos anksčiau negu 2009 metai, didžiausias įmonių įsikūrimo pikas buvo 2010-2012 metais. Didžiajai daliai miestų turimi duomenys tik apie vieną ten įsikūrusią įmonę.

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

Priedas

Žemiau pateiktas visas naudotas programinis kodas:

Naudojant R:

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

library(tidyverse)

# Duomenų įvesties klaidos (sutvarkysiu pačiame duomenų faile)
#lines <-readLines("Future-500-7.csv")
#lines[69]<- str_replace(lines[69],'\"',"")
#lines[79]<- str_replace(lines[79],'\"',"")

#writeLines(lines,"modified_csv.csv")
x <- read_csv("modified_csv.csv")


## ------------------------------------------------------------------------------------------------
-------------------------

# Pasivertimas į skaitinius kintamuosius
x_1 <- x %>%
  mutate(Revenue = as.numeric(str_replace_all(Revenue,"\\$|\\,","")),
         Expenses = as.numeric(str_replace_all(Expenses," Dollars|\\,","")),
         Growth = as.numeric(str_replace_all(Growth,"%","")),
         Profit = as.numeric(str_match(Profit,"\\d+")))


## ------------------------------------------------------------------------------------------------
-------------------------

library(psych)
x_1 %>% select(where(is.numeric),-"ID") %>% describe()



x_grouped <- x_1 %>% group_by(Industry)
names <- x_grouped %>% group_keys() %>% pull(Industry)

summary_list <- x_grouped %>% select(where(is.numeric)) %>% select(-"ID") %>%
  group_split() %>%
  purrr::map(~select(.x,-"Industry")) %>%
  purrr::map(describe) %>%
  purrr::map(~rownames_to_column(as.data.frame(.x)))

names(summary_list) <- names
summary_list


## ------------------------------------------------------------------------------------------------
-------------------------

# išveda į failus, siekiant nukopijuoti į word lentelę
x %>% describe %>% select(c("sd","mean","median","min","max")) %>% round(2) %>%
write.csv("out.csv",quote=FALSE)
```

```
temp <- summary_list %>% enframe() %>% unnest_longer("value")

cbind(temp$name,temp$value) %>% select(c("rowname","temp$name","sd","mean","median","min","max")) %>%
mutate(across(where(is.numeric),round,2)) %>% write.csv("out_2.csv",quote=FALSE,row.names = FALSE)


## ------------------------------------------------------------------------------------------------------------
-------------------------

# turimi vienų metų duomenys. esant praeitų metų duomenims NA reikšmes būtų galima pakeisti praeitomis
x_1 %>% group_by(Name) %>% count() %>% arrange(desc(n))

x_1 %>% ungroup() %>% summarize(across(everything(),~sum(is.na(.)))) # pradiniai kiekiai praleistų
reikšmių


replace_with_group_median<- function(x,y) {
  group_median <- median(x,na.rm = TRUE)
  if_else(is.na(x),group_median,x)
}


library(maps)
cities <-  us.cities$country.etc
names(cities) <- str_replace(us.cities$name,paste("",us.cities$country.etc),"")


x_2 <- x_1 %>%
  # faktinis užpildymas
  mutate(State = if_else(is.na(State),cities[City],State)) %>%
  # išvestinės reikšmės
  mutate(Expenses = if_else(is.na(Expenses) & !is.na(Profit),Revenue - Profit,Expenses),
         Revenue = if_else(is.na(Revenue) & !is.na(Profit),Expenses + Profit,Revenue)) %>%
  group_by(Industry) %>%
  filter(!(is.na(Revenue) & is.na(Expenses) & !is.na(Profit))) %>%
  mutate(Expenses = replace_with_group_median(Expenses),
          Revenue = replace_with_group_median(Revenue),
          Profit = Revenue - Expenses) %>%
  rbind(x_1  %>% filter((is.na(Revenue) & is.na(Expenses) & !is.na(Profit)))) %>%
  mutate(Employees = replace_with_group_median(Employees),
          Growth = replace_with_group_median(Growth)) %>%
  ungroup()




## ------------------------------------------------------------------------------------------------------------
-------------------------

# likusios praleistos reikšmės paliekamos duomenyse (daugiausia reikšmės nominaliuose kintamuosiouse)
x_2 %>% summarize(across(everything(),~sum(is.na(.))))


## ------------------------------------------------------------------------------------------------------------
-------------------------

names <- c("Employees","Revenue","Expenses","Profit","Growth")
```

```
x_2 %>%  select(names) %>% purrr::map(~boxplot.stats(.x,coef = 1.5)$out) # sąlyginės išskirtys ("mild" outliers)

(outliers <- x_2 %>% select(names) %>% purrr::map(~boxplot.stats(.x,3)$out)) # išskirtys ("extreme" outliers)


## ------------------------------------------------------------------------------------------------------------------------

ggplot(x_2,aes(Employees)) + geom_histogram() + theme_minimal() # įmonės darbuotojų skaičiaus pasiskirstymas yra stiprios dešininės asimetrijos (right skewed)
# toliau pašalinsiu šias išskirtis


x_2 %>% filter(Employees %in% outliers$Employees)
# išsiskiriančios įmonės t.y. tyrimo objektai. kai kurios iš šių įmonių turi ne tik didelius darbuotojų kiekis, bet ir didelius Expenses/Revenue

x_2 %>% filter(Employees %in% outliers$Employees) %>% count(Industry)

x_3 <- x_2 %>% filter(!Employees %in% outliers$Employees)


## ------------------------------------------------------------------------------------------------------------------------

x_3 %>% select(names) %>% purrr::map(~boxplot.stats(.x,3)$out) # daugiau išskirčių pagal dominančius stulpelius nerasta


## ------------------------------------------------------------------------------------------------------------------------

# Kaip skiriasi imties statistiniai duomenys pašalinus išskirtis
summary_1 <- x_2 %>% select(where(is.numeric),-"ID") %>% describe()

summary_2 <- x_3 %>% select(where(is.numeric),-"ID") %>% describe()

(summary_2 - summary_1) / summary_1 * 100 # procentinis imties statistinių duomenų pokytis pašalinus išskirtis


## ------------------------------------------------------------------------------------------------------------------------

# normalizavimas
normalized <- x_2 %>% select(where(is.numeric),-c("ID","Inception")) %>% drop_na() %>% map_df(~((.x-min(.x))/(max(.x)-min(.x))))
# standartizavimas
standartized <- x_2 %>% select(where(is.numeric),-c("ID","Inception")) %>% drop_na() %>% map_df(~(.x-mean(.x))/sd(.x))


## ------------------------------------------------------------------------------------------------------------------------

normalized  %>% pivot_longer(1:5) %>% ggplot(aes(value,color=name)) + geom_boxplot() + coord_flip() + theme_minimal() + scale_color_viridis_d()
```

```
standartized  %>% pivot_longer(1:5) %>% ggplot(aes(value,color=name)) + geom_boxplot() + coord_flip() +
theme_minimal() + scale_color_viridis_d()

x_2 %>% select(where(is.numeric),-c("ID","Inception")) %>% pivot_longer(1:5) %>%
ggplot(aes(value,color=name)) + geom_boxplot() + coord_flip() + theme_minimal() +
scale_color_viridis_d()


## ------------------------------------------------------------------------------------------------------------
-------------------------

library(corrplot)
x_corr <- x_2[,-1] %>% drop_na()

numerical <- unlist(lapply(x_corr, is.numeric))
correlation_matrix <- cor(as.matrix(x_corr[,numerical]))
correlation_matrix


corrplot(correlation_matrix, order = "FPC", method = "color",type="upper",diag=FALSE,tl.col = "black",
addCoef.col = "black")


## ------------------------------------------------------------------------------------------------------------
-------------------------

length(unique(x_2$Industry)) # 7 industrijos

x_industry <- x_2 %>% drop_na()


x_industry %>% ggplot(aes(Revenue,Expenses,color=Industry)) + geom_point(aes(size=Profit),alpha=0.7) +
  scale_color_viridis_d() + geom_abline(slope=1,intercept=0) + theme_minimal()

min(x_2$Profit)


## ------------------------------------------------------------------------------------------------------------
-------------------------
library(datasets)
states <-state.region
names(states) <- state.abb

x_regions <- x_2 %>% mutate(Region = states[State]) %>% drop_na()


## ------------------------------------------------------------------------------------------------------------
-------------------------

x_regions %>% ggplot(aes(Region,Growth,fill=Region)) + geom_violin(draw_quantiles = 0.5) +
theme_minimal() + scale_fill_brewer(palette = "Set2")

x_regions %>% ggplot(aes(Region,Profit,fill=Region)) + geom_violin(draw_quantiles = 0.5) +
theme_minimal() + scale_fill_brewer(palette = "Set2")



x_regions %>% ggplot(aes(Region,fill=Region))+ geom_bar() + scale_fill_brewer(palette = "Set2") +
theme_minimal()

x_regions %>% ggplot(aes(Region,fill=Industry))+ geom_bar(position="fill")+ coord_flip() +
```

```
  scale_y_continuous(labels=scales::label_percent()) + scale_fill_viridis_d() + theme_minimal()

## ------------------------------------------------------------------------------------------------
-------------------------

library(ggstatsplot)

emp.data <- x_2 %>% arrange(desc(Employees))
emp.data_1 <- emp.data[1:as.integer(nrow(emp.data)*0.15),]
emp.data <- emp.data %>% arrange(Employees)
emp.data_2 <- emp.data[1:as.integer(nrow(emp.data)*0.85),]
ggbetweenstats(data = emp.data_1,
               x = Industry,
               y = Revenue            ,
               plot.type = "box", mean.plotting=FALSE,
               results.subtitle=FALSE,
               outlier.tagging = TRUE, outlier.label = "Name" ,pairwise.comparisons = FALSE)

ggbetweenstats(data = emp.data_2,
               x = Industry,
               y = Revenue            ,
               plot.type = "box", mean.plotting=FALSE,
               results.subtitle=FALSE,
               outlier.tagging = TRUE, outlier.label = "Name" ,pairwise.comparisons = FALSE)

boxplot(x_2$Employees, horizontal = TRUE, main="Visos įmonės")
boxplot(emp.data_1$Employees, horizontal = TRUE, main="Didelių įmonių grupė")
boxplot(emp.data_2$Employees, horizontal = TRUE, main="Mažų įmonių grupė")

## ------------------------------------------------------------------------------------------------
-------------------------

x_regions %>% ggplot(aes(Revenue,Expenses,color=Region)) + geom_point(aes(size=Profit),alpha=0.8) +
  scale_color_brewer(palette="Set2") + geom_abline(slope=1,intercept=0) + theme_minimal()


## ------------------------------------------------------------------------------------------------
-------------------------

x_regions %>% drop_na() %>% ggplot(aes(Inception,color=Industry)) + stat_ecdf() +
facet_wrap(vars(Industry)) + theme_minimal() + scale_color_viridis_d()


ggbetweenstats(data = x_2,
               x = Industry,
               y = Employees                  ,
               plot.type = "box", mean.plotting=FALSE,
               results.subtitle=FALSE,
               outlier.tagging = TRUE, outlier.label = "Name",pairwise.comparisons = FALSE,point.args =
list(position="jitter"))

Revenue.data <- x_2 %>% group_by(Industry) %>% dplyr::summarize(Mean = mean(Revenue, na.rm=TRUE))
Revenue.data$key <- "Revenue"
Profit.data <- x_2 %>% group_by(Industry) %>% dplyr::summarize(Mean = mean(Profit, na.rm=TRUE))
Profit.data$key <- "Profit"

mean.data <- rbind(Revenue.data,Profit.data)
mean.data <- mean.data[complete.cases(mean.data),]

ggplot(mean.data, aes(fill=key, y=Mean, x=Industry)) +
```

```
  geom_bar(position='dodge', stat="identity", alpha = 0.8) + theme_minimal() +
guides(x=guide_axis(angle=35))
ggbetweenstats(data = x_2,
               x = Industry,
               y = Profit,                  ,
               plot.type = "box", mean.plotting=FALSE,
               results.subtitle=FALSE,
               outlier.tagging = TRUE, outlier.label = "Name",pairwise.comparisons = FALSE)
```

Naudojant Python (laikant, kad kintamasis „x_2“ yra duomenų aibė, sutvarkyta pagal šį aprašą):

```
import pandas as pd
pd.set_option('display.float_format', lambda x: '%.3f' % x)
import numpy as np
from scipy import stats
import seaborn as sns
sns.set(rc = {'figure.figsize':(17,7)})
import matplotlib.pyplot as plt


## ------------------------------------------------------------------------------------------------------------
-------------------------

tempdf = pd.DataFrame(x_2.groupby('Industry').City.nunique()).reset_index().sort_values(by="City",
ascending = False)
tempdf.plot.bar(x = 'Industry', y = 'City', figsize = (17, 7), rot = 0, title = "Distinct city count by
Industry")


## ------------------------------------------------------------------------------------------------------------
-------------------------

tempdf = pd.DataFrame(x_2.groupby('Industry').City.count()).reset_index().sort_values(by="City",
ascending = False)
tempdf.plot.bar(x = 'Industry', y = 'City', figsize = (17, 7), rot = 0, title = "Total count by
Industry")


## ------------------------------------------------------------------------------------------------------------
-------------------------

values = x_2.groupby(['Inception', 'Industry']).Industry.count().values
cols = x_2.groupby(['Inception', 'Industry']).Industry.count().index.values
tempdf = pd.DataFrame([(cols[i][0], cols[i][1], values[i]) for i in range(len(values))], columns =
['Inception', 'Industry', 'Count'])
sns.lineplot(data=tempdf,x='Inception', y = 'Count', hue='Industry', linewidth = 3)
```