Vilniaus Universitetas

Pirminė duomenų aibės analizė

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

Vilnius, 2022

# Tikslas ir uždaviniai

## Tikslas:

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

## Uždaviniai:

Apskaičiuoti aprašomąsias duomenų statistikas.

Pasirinkus tinkamus metodus užpildyti praleistas duomenų reikšmes.

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

Pritaikyti duomenų normavimo metodus.

Atlikti aibės vizualią analizę.

Ištirti koreliacijas tarp duomenų aibės požymių.

# Duomenų aibė

Duomenų aibę sudaro duomenys apie 500 įmonių su tokiais požymiais:

„ID“ - (nominalusis) įmonę duomenyse identifikuojantis kodas

„Name“ – (nominalus) įmonės pavadinimas

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

„Inception“ – (kiekybinis) įmonės įkūrimo metai

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

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

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

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

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

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

Aprašomosios statistikos dalyje (žr. Aprašomoji statistika) palygintos trijų duomenų aibės variantų pagrindinės aprašomosios statistikos charakteristikos. Siekiant nuoseklumo, toliau aprašyti tik vieno duomenų aibės varianto rezultatai.

# Atliktos analizės aprašymas

## Aprašomoji statistika

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

Lentelė 1 Aprašomosios statistikos charakteristikos duomenų aibei

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

Papildomai palyginimui pateiktos dviejų kitų duomenų aibės variantų pagrindinės charakteristikos (žr. Lentelė 3 ir Lentelė 4). Tarp aprašomosios statistikos charakteristikų pastebimi tik minimalūs skirtumai.

Lentelė 2 Aprašomosios statistikos charakteristikos pirmai alternatyviai duomenų aibei

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | stand. nuokrypis | vidurkis | mediana | min | max |
| Inception | 3.23 | 2010.17 | 2011 | 1999 | 2014 |
| Employees | 398.66 | 148.66 | 56 | 1 | 7125 |
| Revenue | 3199673.19 | 10846282.41 | 10671779.5 | 1614585 | 21810051 |
| Expenses | 2122629.34 | 4307620.18 | 4316632 | 71219 | 9860686 |
| Profit | 3873644.87 | 6537765.72 | 6512379 | 12434 | 19624534 |
| Growth | 6.9 | 14.34 | 15 | -3 | 30 |

Lentelė 3 Aprašomosios statistikos charakteristikos antrai alternatyviai duomenų aibei

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | stand. nuokrypis | vidurkis | mediana | min | max |
| Inception | 3.23 | 2010.17 | 2011 | 1999 | 2014 |
| Employees | 398.08 | 149.08 | 57 | 1 | 7125 |
| Revenue | 3191399.82 | 10847567.24 | 10647231 | 1614585 | 21810051 |
| Expenses | 2115947.99 | 4315806.73 | 4365512 | 71219 | 9860686 |
| Profit | 3869933.65 | 6539474.01 | 6513366 | 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 4 Aprašomosios statistikos charakteristikos atskirai kiekvienai pramonės šakai

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

## Praleistos reikšmės

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

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

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

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

Kategoriniuose kintamamuosiuose esančios praleistos reikšmės paliktos nekeistos.

Iš viso užpildytos 22 praleistos reikšmės.

## Išskirčių analizė

Išskirtys vertintos naudojant vidinį ir išorinius barjerus, kur – atitinkamai pirmas ir trečias kvartiliai, – 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

Iš duomenų aibės pašalintos visos įmonės, išsiskiriančios pagal išorinį barjerą bent vienu požymiu (šiuo atveju visos tokios įmonės išsiskiria pagal darbuotojų skaičių).Lentelėje žemiau pateikta kokiai pramonės šakai priklauso pašalintos reikšmės (žr. Lentelė 5).

Lentelė 5 Iš duomenų aibės pašalintų reikšmių kiekis pagal pramonės šaką

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

Pašalinus išskirtis pakartotinai apskaičiuotos duomenų aibės aprašomosios statistikos charakteristikos (žr. Lentelė 3). Didžiausias pokytis pastebėtas darbuotojų kiekyje – pašalimus išskirtis darbuotojų kiekio vidurkis sumažėjo 45%, standartinis nuokrypis 79%, mediana – 11%.

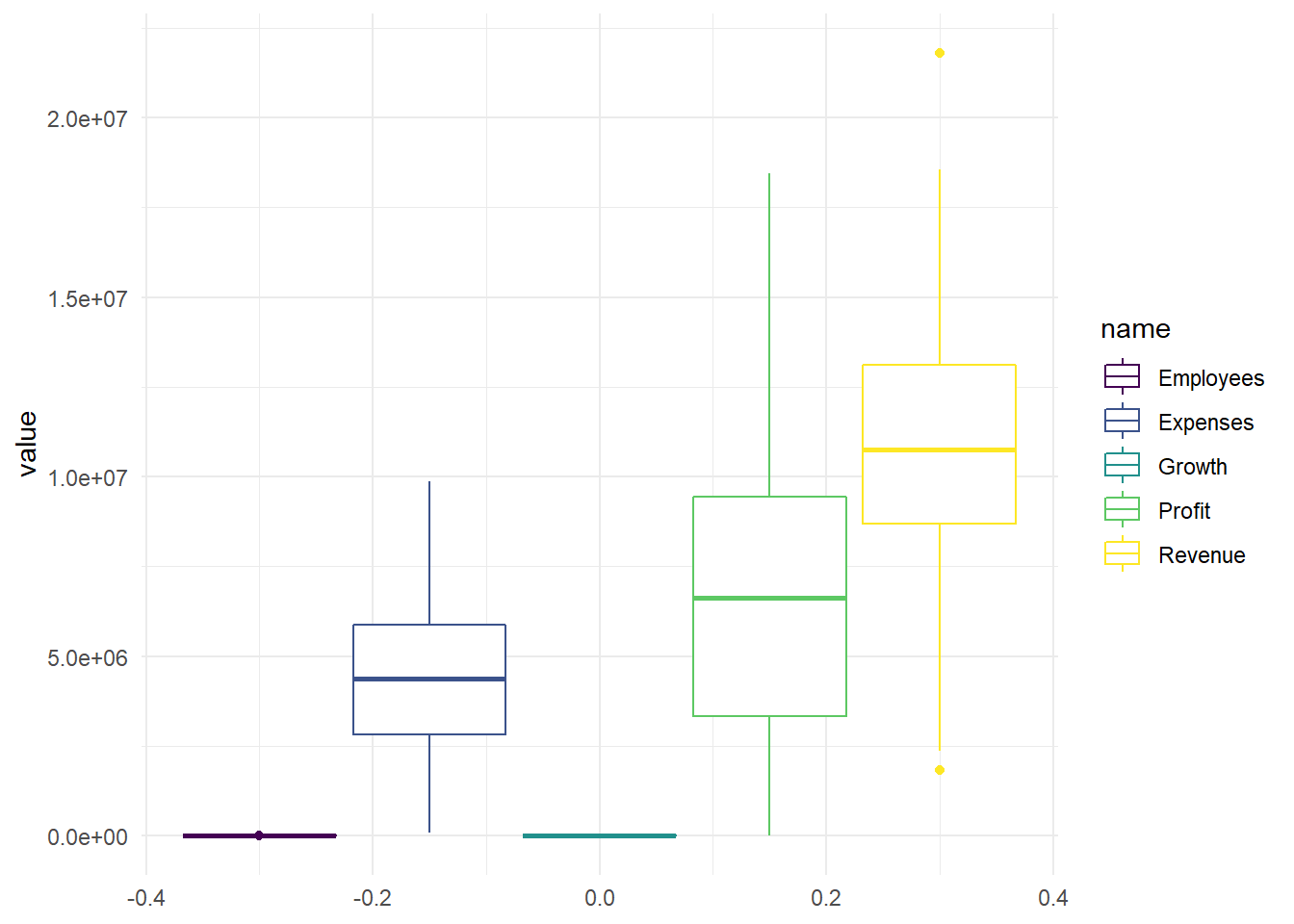
Lentelė 6 Procentinis aprašomosios statistikos charakteristikų pokytis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 |

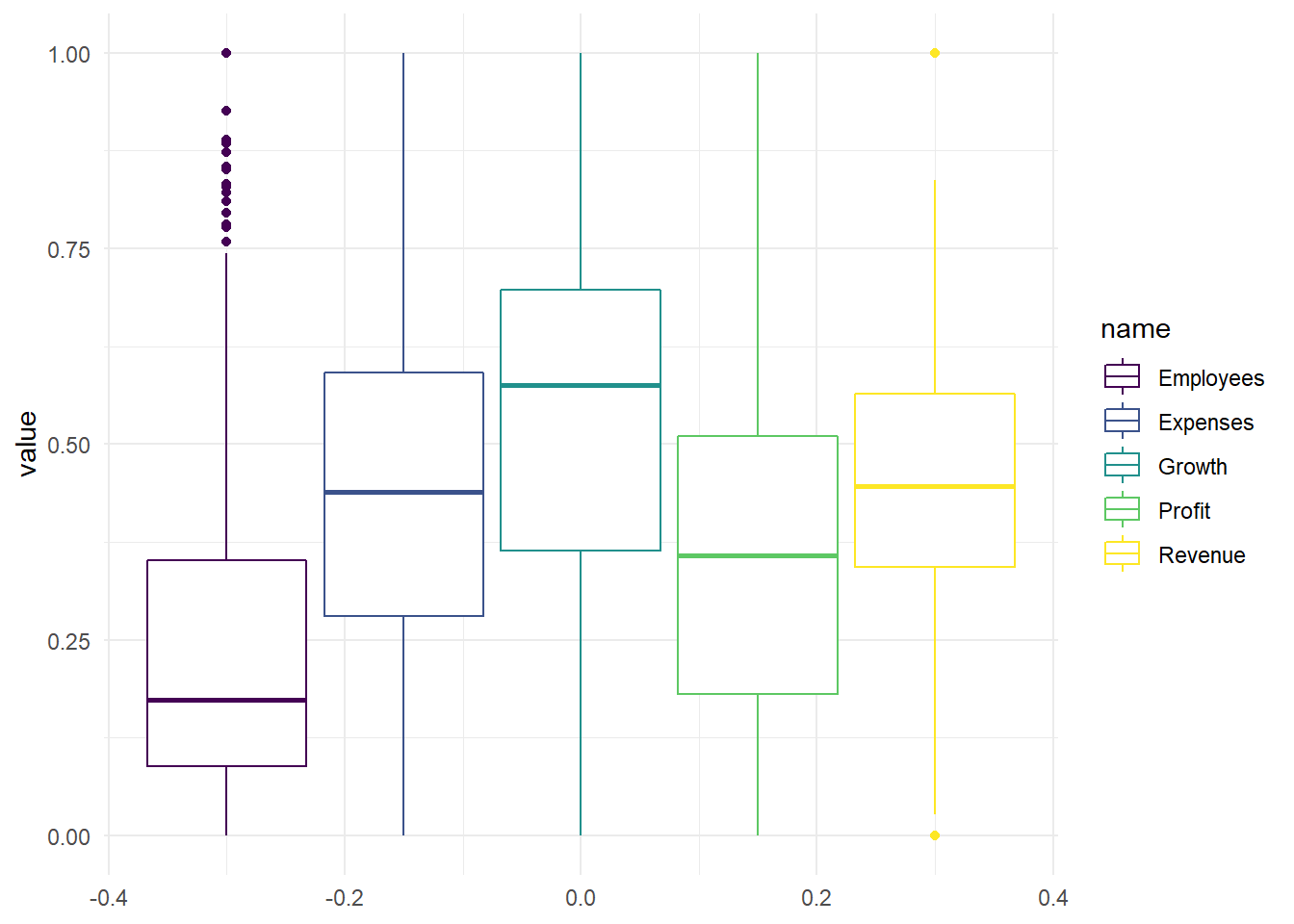
## Duomenų normavimas

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

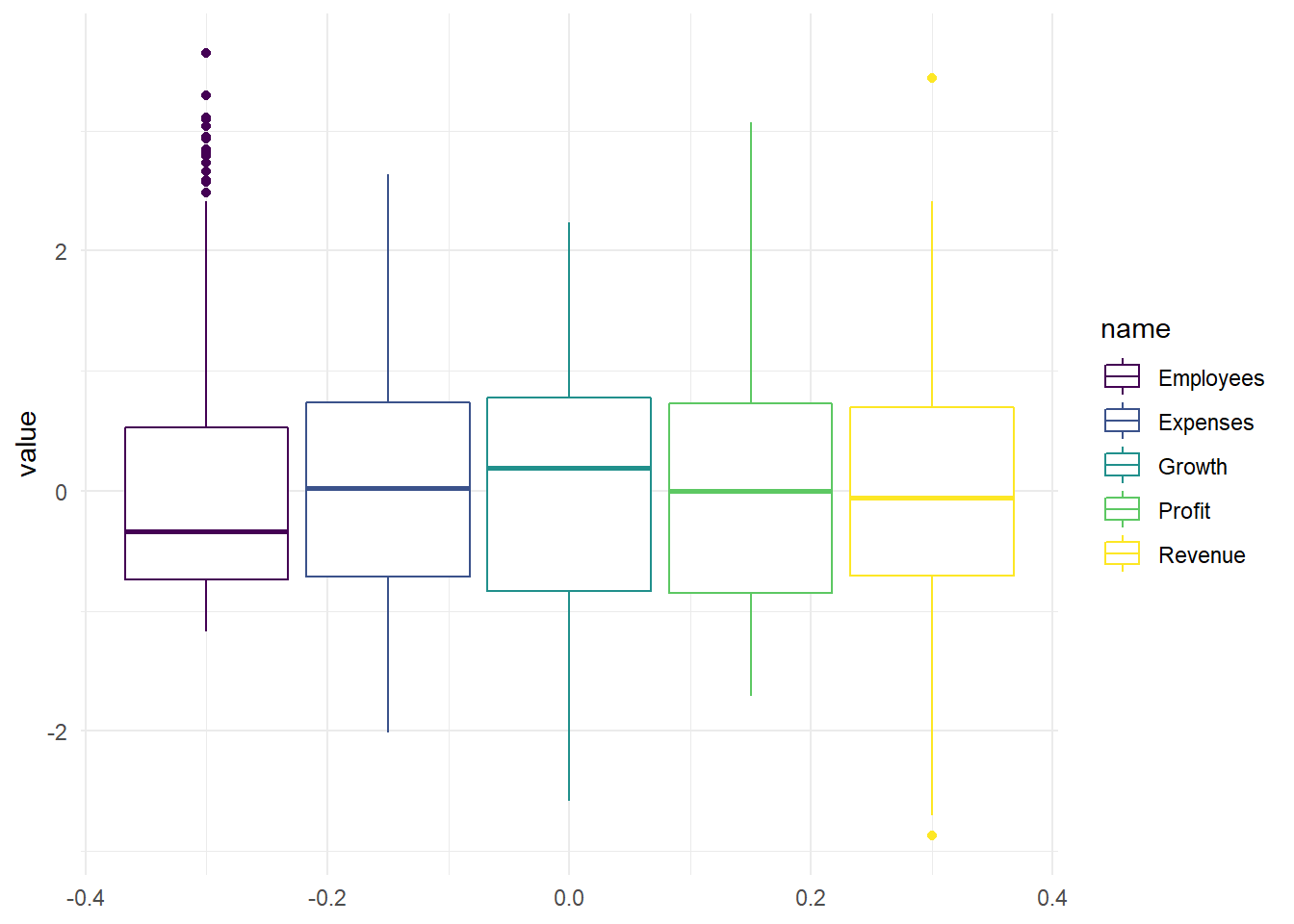
Pradinis kiekybinių duomenų aibės požymių pasiskirstymas pavaizduotas stačiakampe diagrama (žr. 2 Pav.). Pakartotinai pavaizduotas pasiskirstymas atlikus abu anksčiau minėtus normavimo metodus (žr. 3 Pav. ir 4 Pav.)



2 Pav. Kiekybinių požymių stačiakampė diagrama prieš atliekant normavimą



3 Pav. Stačiakampė diagrama atlikus min-max normavimą

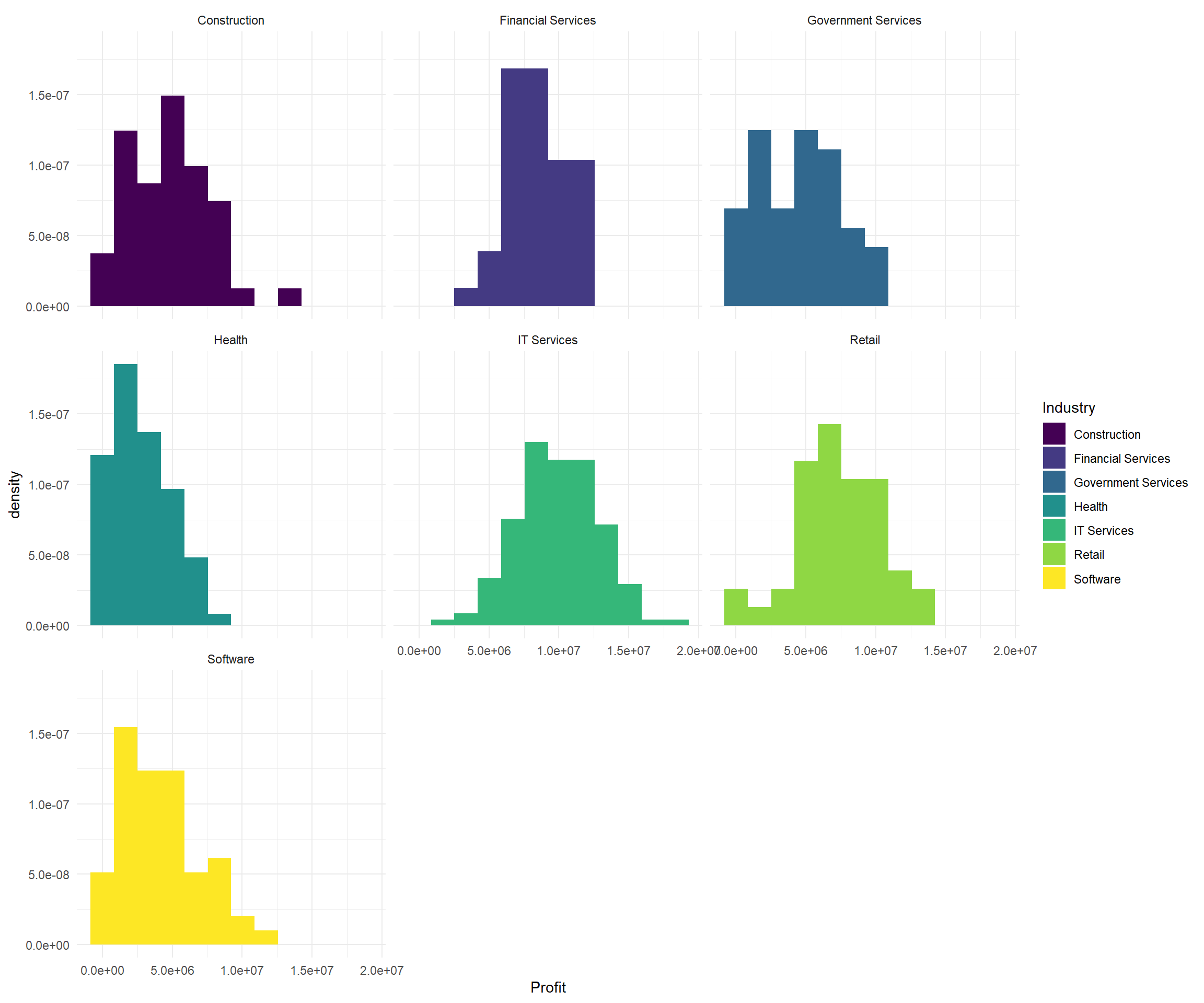


4 Pav. Stačiakampė diagrama atlikus standartizaciją

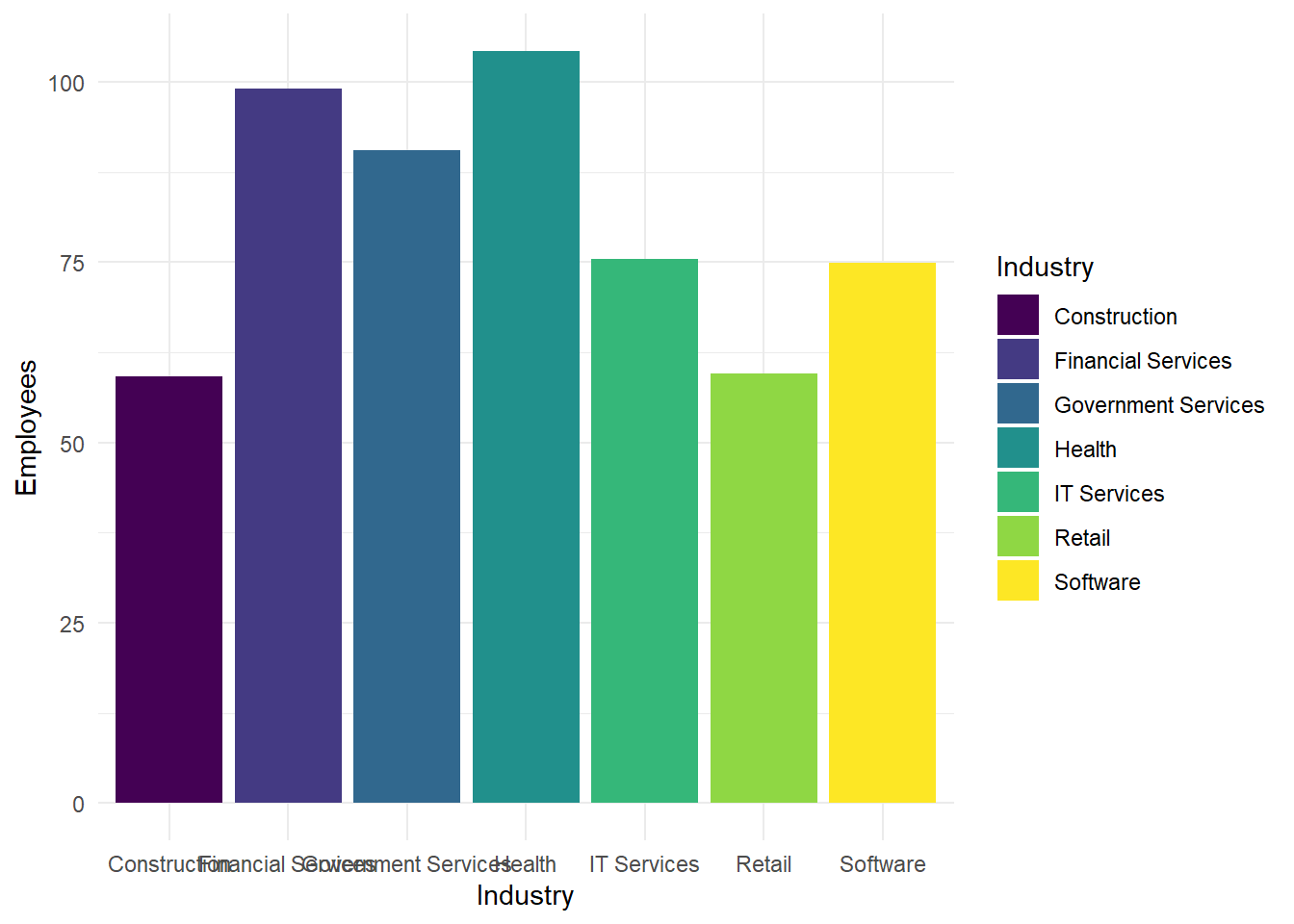
## Vizuali analizė

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

Kiekvienai pramonės šakai suskaičiuotas ir stulpeline diagrama pavaizduotas vienoje įmonėje dirbančių darbuotojų vidurkis (žr. 6 Pav.). Matoma, kad didžiausias vidutinis darbuotojų skaičius įmonėje duomenų aibėje yra IT Health ir Financial Services pramonės šakose, mažiausias – Retail ir Construction.

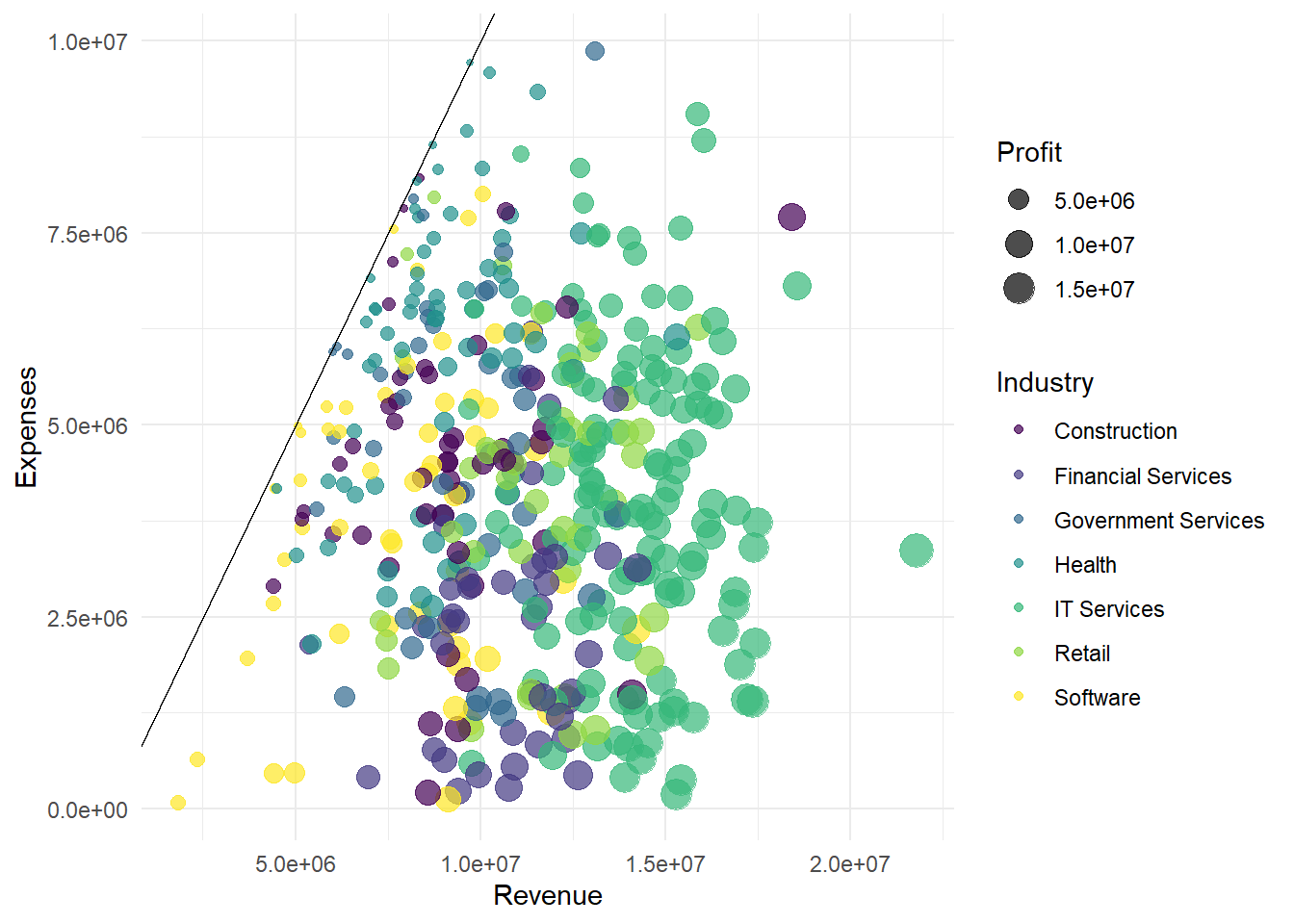


5 Pav. Įmonių pelnas pagal pramonės šaką



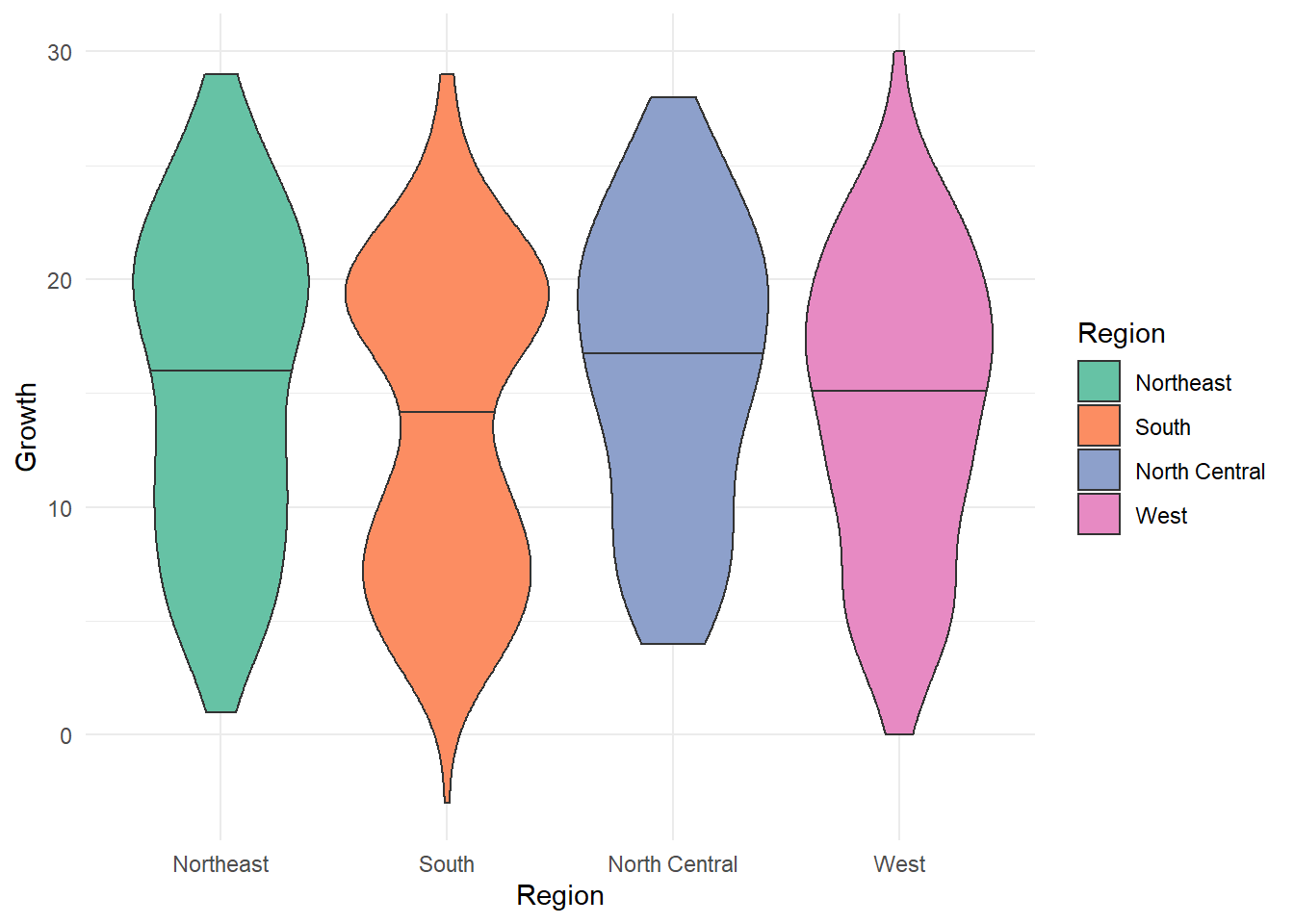
6 Pav. Vidutinis darbuotojų skaičius pagal pramonės šaką

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.

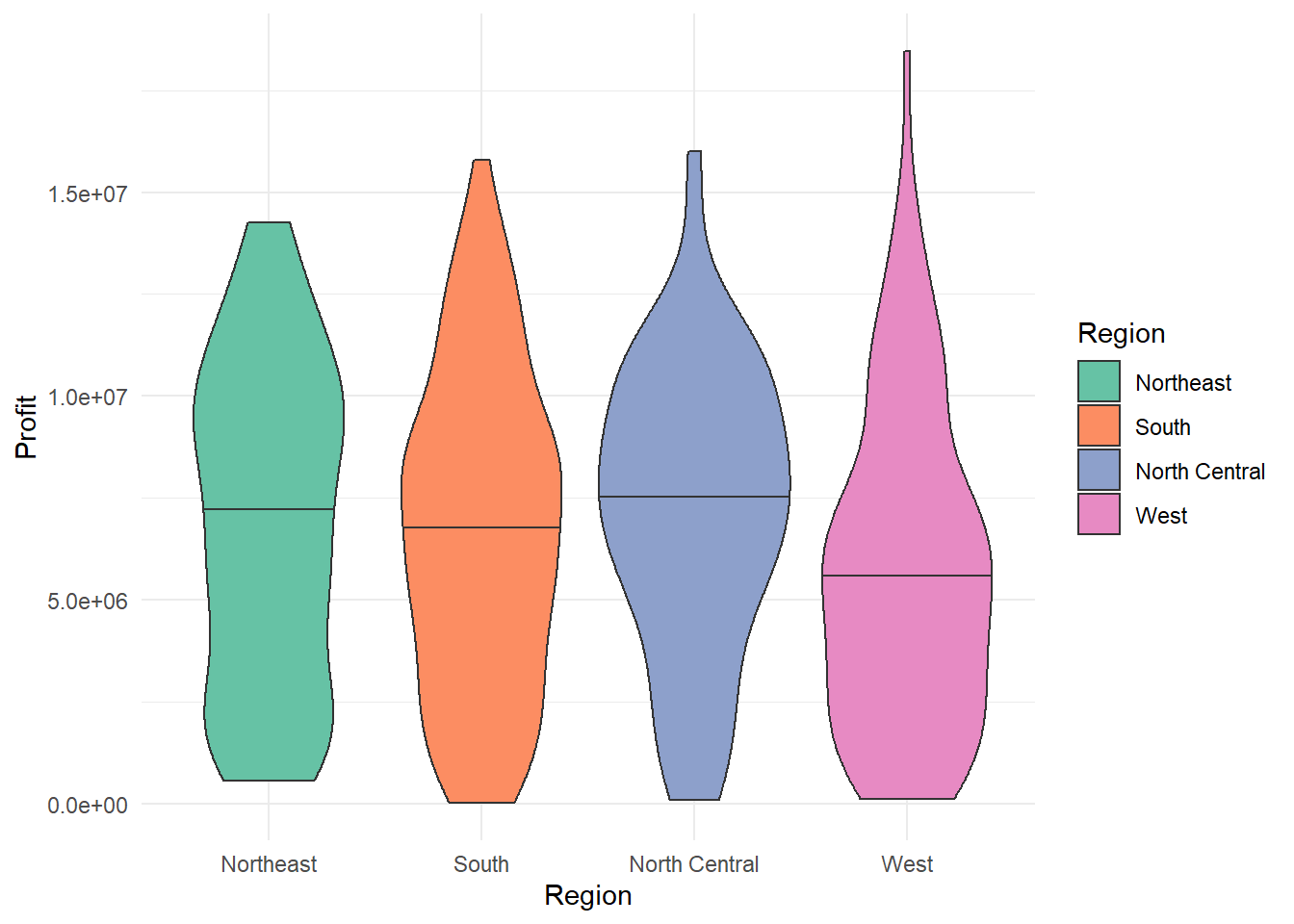


7 Pav. Įmonių pajamų ir išlaidų sklaidos diagrama

JAV valstijos padalintos į keturis regionus ir smuiko formos grafikais kiekvienam regionui pavaizduotas įmonių augimo (žr. 8 Pav.) ir pelno (žr. 9 Pav.) pasiskirstymas (horizontalia linija papildomai pažymint medianinę reikšmę). Ryškių įmonių augimo ir pelno skirtumų tarp JAV regionų nepastebėta.

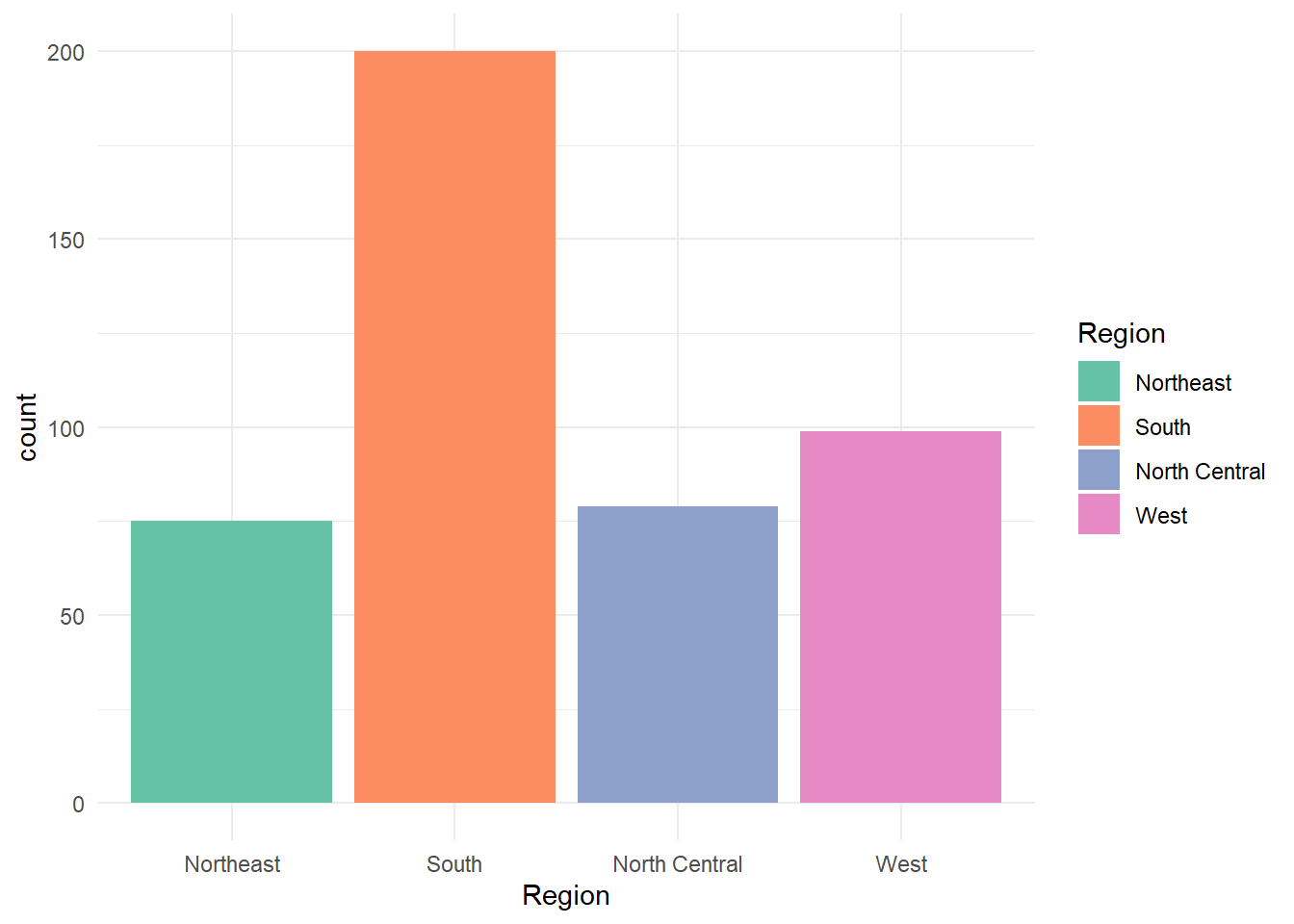


8 Pav. Įmonių augimo pasiskirstymas pagal JAV regionus

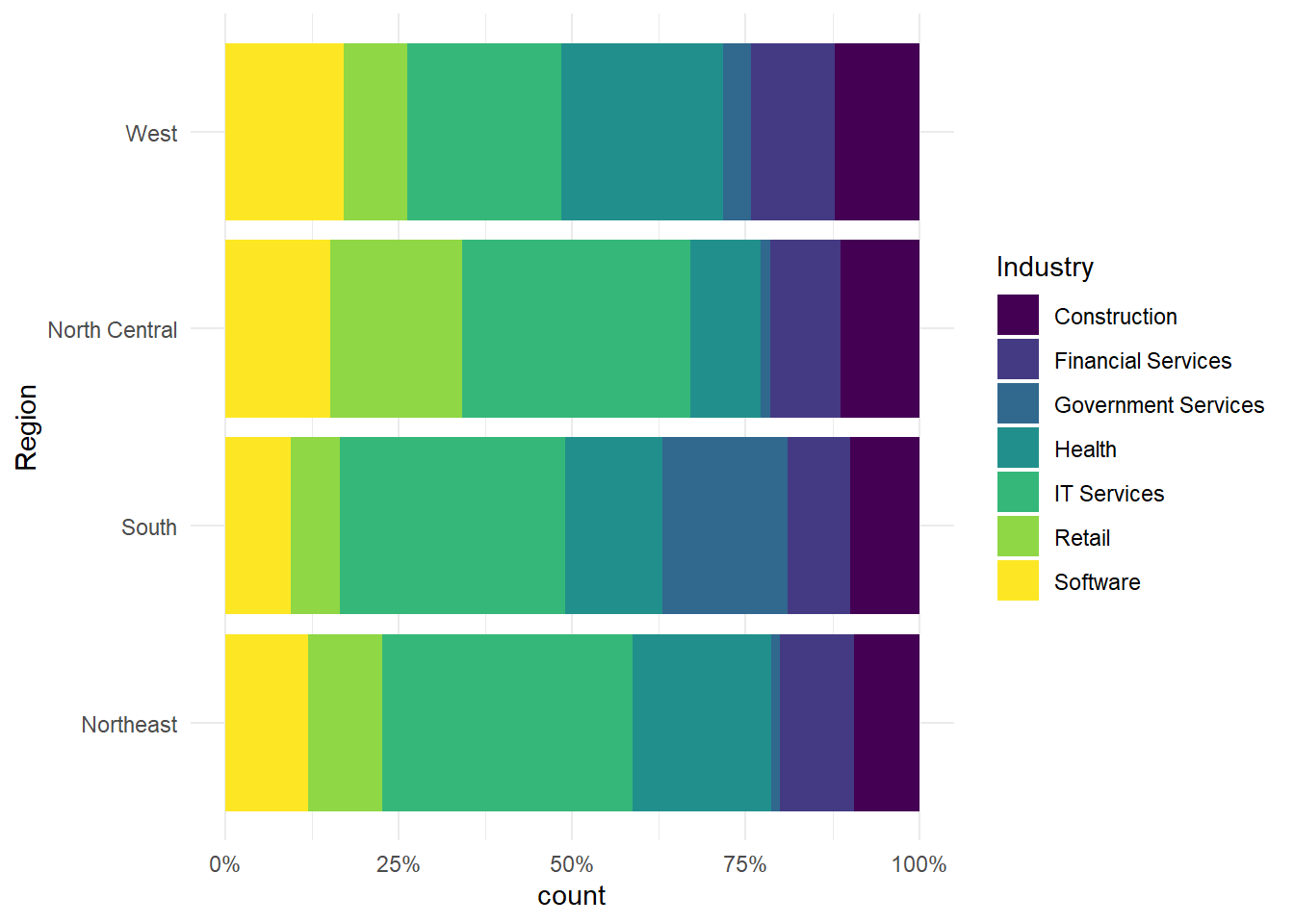


9 Pav. Įmonių pelno pasiskirstymas pagal JAV regioną

Stulpeline diagrama pavaizduotas įmonių kiekviename regione skaičius (žr. 10 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. 11 Pav.). Grafike galima matyti, kad įmonių pasiskirstymas labai panašus visuose 4 regionuose.

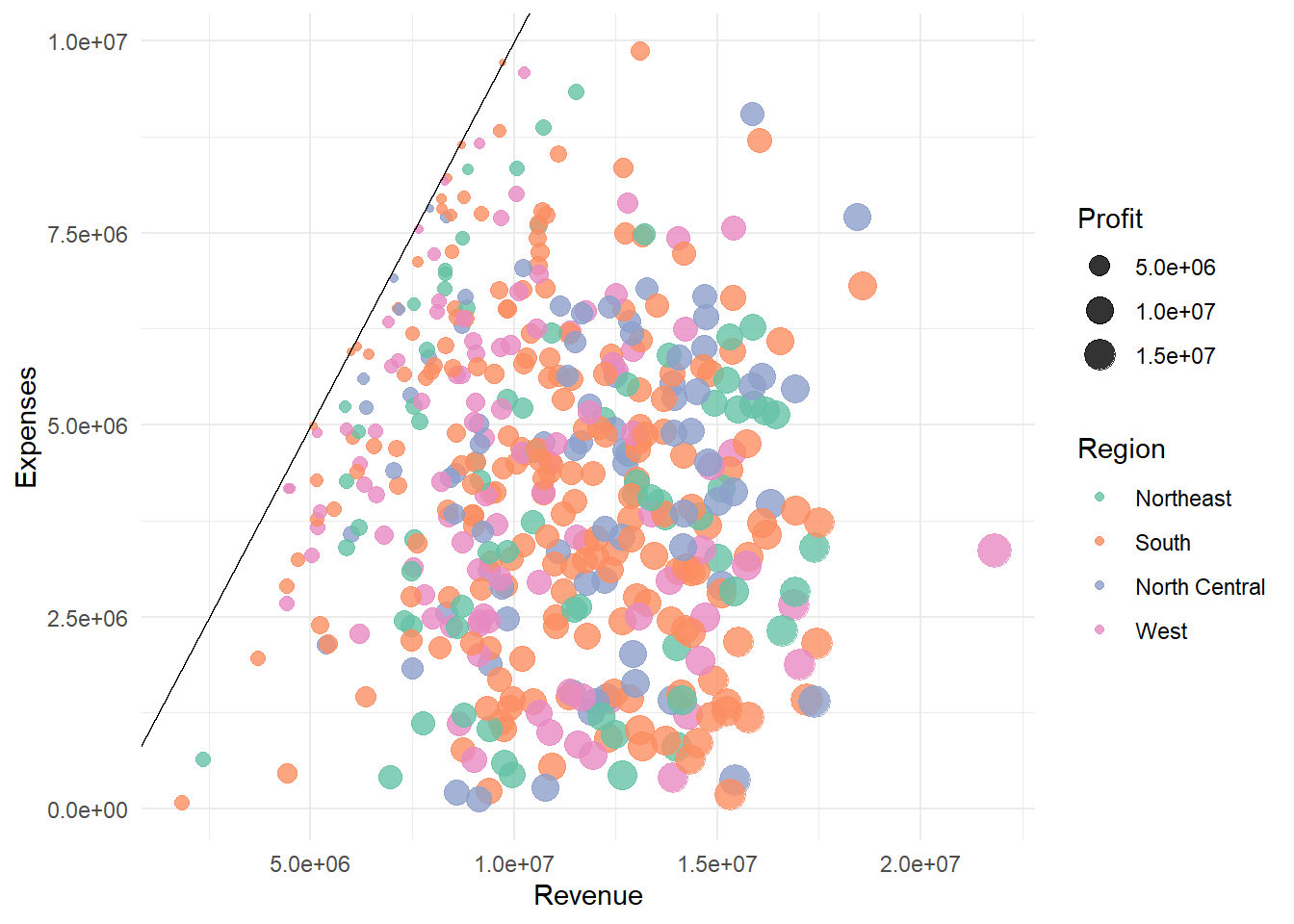


10 Pav. Įmonių skaičius pagal JAV regioną



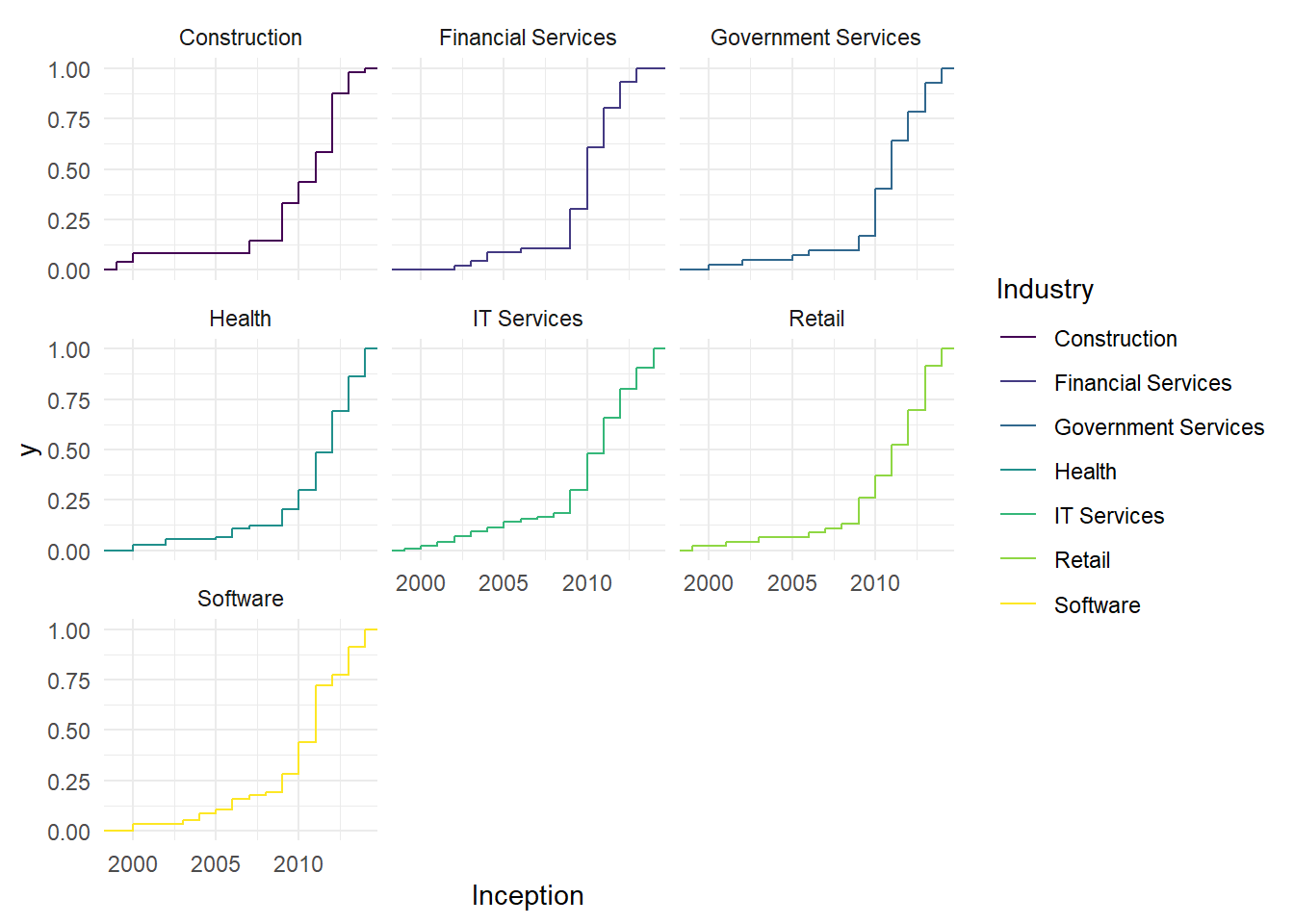
11 Pav. Įmonių pasiskirstymas pagal pramonės šaką kiekviename JAV regione

Dar kartą nubrėžta įmonių pajamų ir išlaidų sklaidos diagrama, tačiau šį kart nuspalvinant taškus pagal regioną (žr. 12 Pav.). Duomenų atsiskyrimas daug mažesnis negu taškus nuspalvinant pagal pramonės šaką (7 Pav.)



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

Pavaizduota įmonių įsikūrimo metų empirinė pasiskirstymo funkcija (žr. 13 Pav.). Matoma, kad didžioji dalis įmonių, esančių duomenų aibėje, įkurta nuo maždaug 2009-2010 metų. Ši tendencija galioja visoms pramonėms šakomis.



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

## Požymių koreliacijos

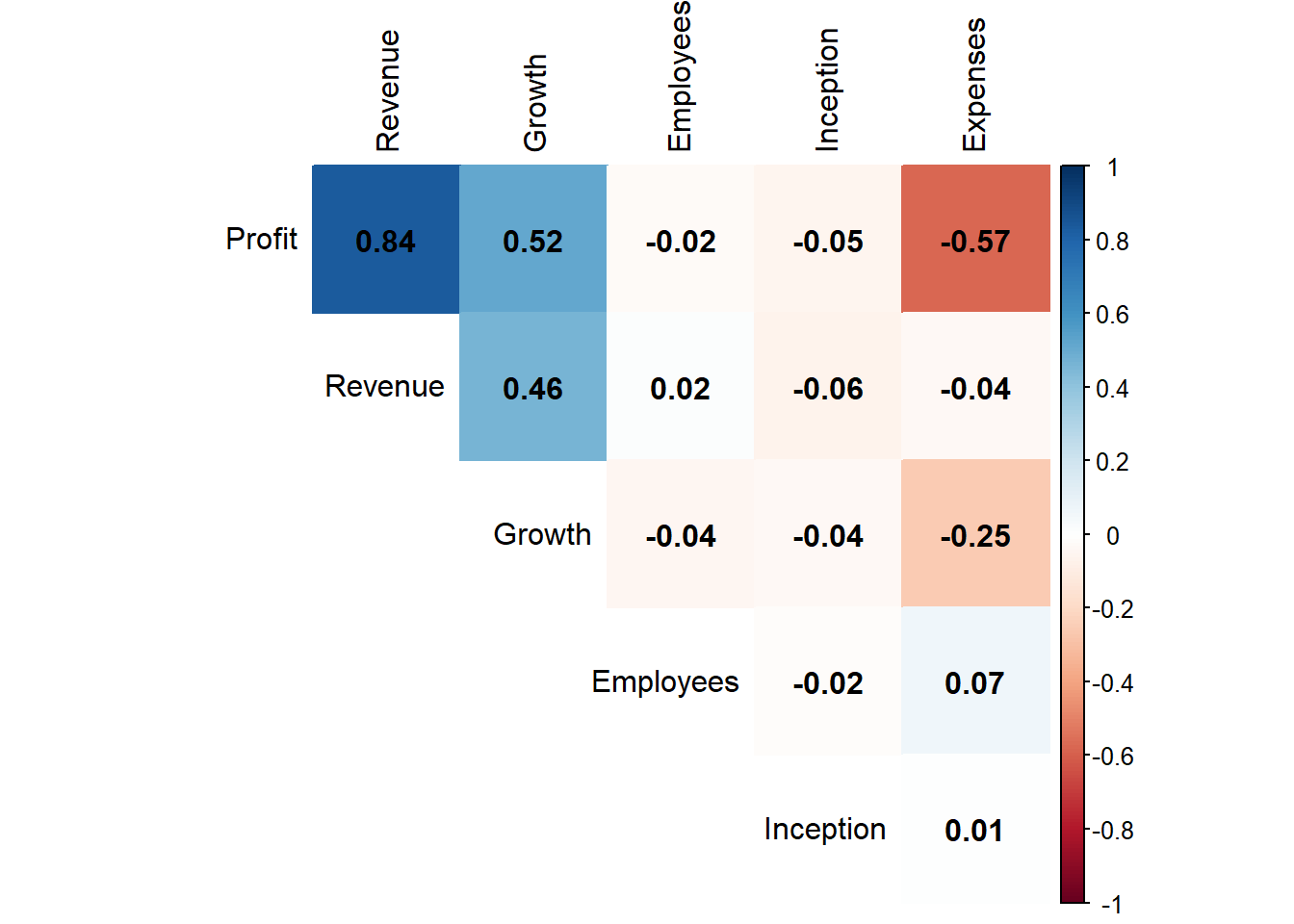
Tarp skaitinių rodiklių apskaičiuotos Pirsono koreliacijos koeficientų reikšmės (angl. Pearson correlation coefficient).

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

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

Lentelė 7 Pirsono koreliacijos koeficientai tarp skaitinių įmonių požymių

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



14 Pav. Pirsono koreliacijos tarp skaitinių požymių koeficientai

# Išvados

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

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

Duomenyse rasta 36 įmonių, išsiskiriančių pagal darbuotojų kiekį. Šios išskirtys pašalintos iš duomenų aibės. Pašalinus šias reikšmes iš duomenų aibės darbuotojų skaičiaus įmonėje standartinis nuokrypis sumažėjo 79%, vidurkis - 44%, mediana – 10%

Kiekybiniams požymiams 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. Lyginant JAV regionus rasta, kad didžioji dalis duomenų yra iš pietinio JAV regiono, tačiau pagal kitus požymius ryškių skirtumų tarp JAV regionų nerasta.

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

# Priedas

Žemiau pateiktas visas naudotas programinis kodas:

## ---- include=FALSE-----------------------------------------------------------------------------------------------------------

knitr::opts\_chunk$set(warning=FALSE,message=FALSE)

options("digits" = 5)

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

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(.x)))) # 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(.x))))

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

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\_3 %>% select(where(is.numeric),-c("ID","Inception")) %>% drop\_na() %>% map\_df(~((.x-min(.x))/(max(.x)-min(.x))))

# standartizavimas

standartized <- x\_3 %>% 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\_3 %>% 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\_3[,-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\_3$Industry)) # 7 industrijos

x\_industry <- x\_3 %>% drop\_na()

x\_industry %>% ggplot(aes(x=Profit,fill=Industry)) + geom\_histogram(aes(y=after\_stat(density)),bins = 12) + facet\_wrap(vars(Industry)) + scale\_fill\_viridis\_d() + theme\_minimal() + scale\_y\_continuous(n.breaks = 5)

x\_industry %>% ggplot(aes(x=Industry,y=Employees,fill=Industry)) + stat\_summary(fun=mean,geom="bar") + scale\_fill\_viridis\_d() + theme\_minimal()

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

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\_3$Profit)

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

library(datasets)

states <-state.region

names(states) <- state.abb

x\_regions <- x\_3 %>% 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()

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

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()