Pirminė duomenų analizė

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus
Duomenų Mokslas 3 kursas 2 gr.

Vilnius, 2022

Tikslas - nuskaityti duomenų aibę, atlikti duomenų apdorojimą ir ją išanalizuoti

Uždaviniai:

- Užpildyti praleistas reikšmes
- Ištirti išskirtis, įvertinti kaip pasikeičia aprašomoji statistika išėmus jas
- Pritaikyti duomenų normavimo metodus
- Atlikti aibės vizualią analizę
- Ištirti koreliacijas tarp požymių

Apie duomenis: 500 eilučių ir 11 stulpelių

Kintamasis	Tipas
Name	nominalus
Industry	nominalus
Inception	diskretus, intervalinis
Employees	diskretus, santykinis
State	nominalus
City	nominalus
Revenue	tolydus, santykinis
Expenses	tolydus, santykinis
Profit	tolydus, santykinis
Growth	tolydus, santykinis

Praleistos reikšmės

- Valstijos - randama pagal miesto reikšmę
- Profit, Revenue, Expenses - pagal sąryšį Profit = Revenue - Expenses arba industrijos medianą
- Employees, Growth, Inception - pagal įmonės industrijos medianą
- Industry - pašalinama iš aibės tik tada kai naudojama grafikuose

Aprašomoji statistika

1 lentelė 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

Išskirčių analizė

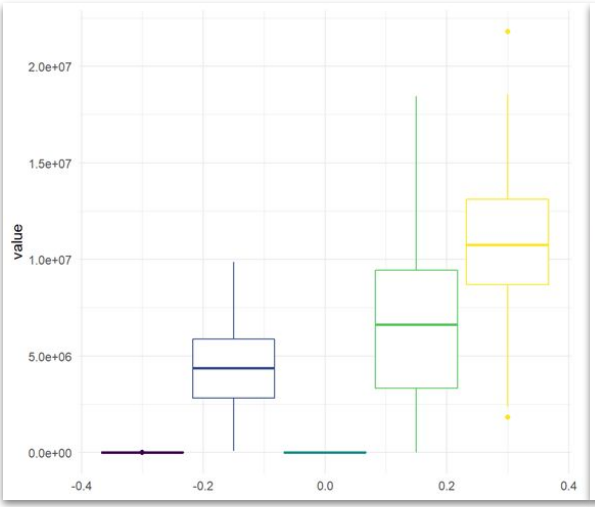
Vidinis barjeras [Q1 - 1.5H ; Q3 + 1.5H]:

- pagal “Revenue” : 4
- pagal “Profit” : 2
- Pagal “Employees” : 60

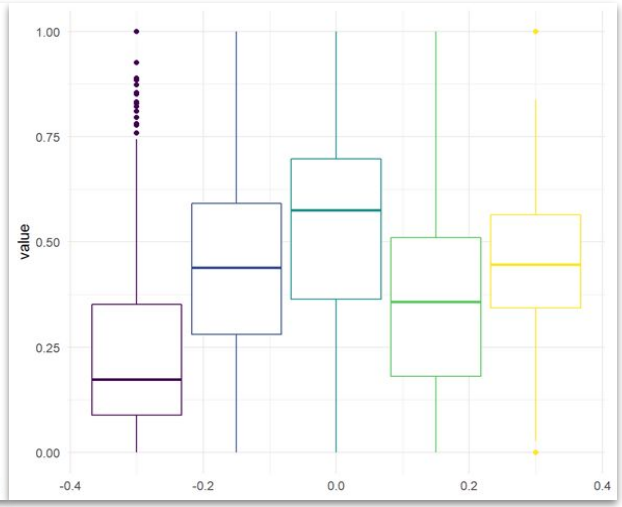
Išorinis barjeras [Q1 - 3H ; Q3 + 3H]:

- pagal “Employees” : 36

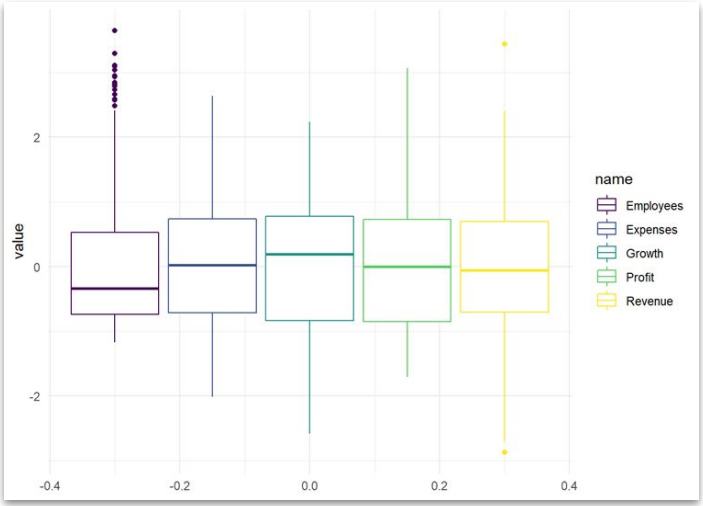
Duomenų normavimas



Nepakeista duomenų aibė

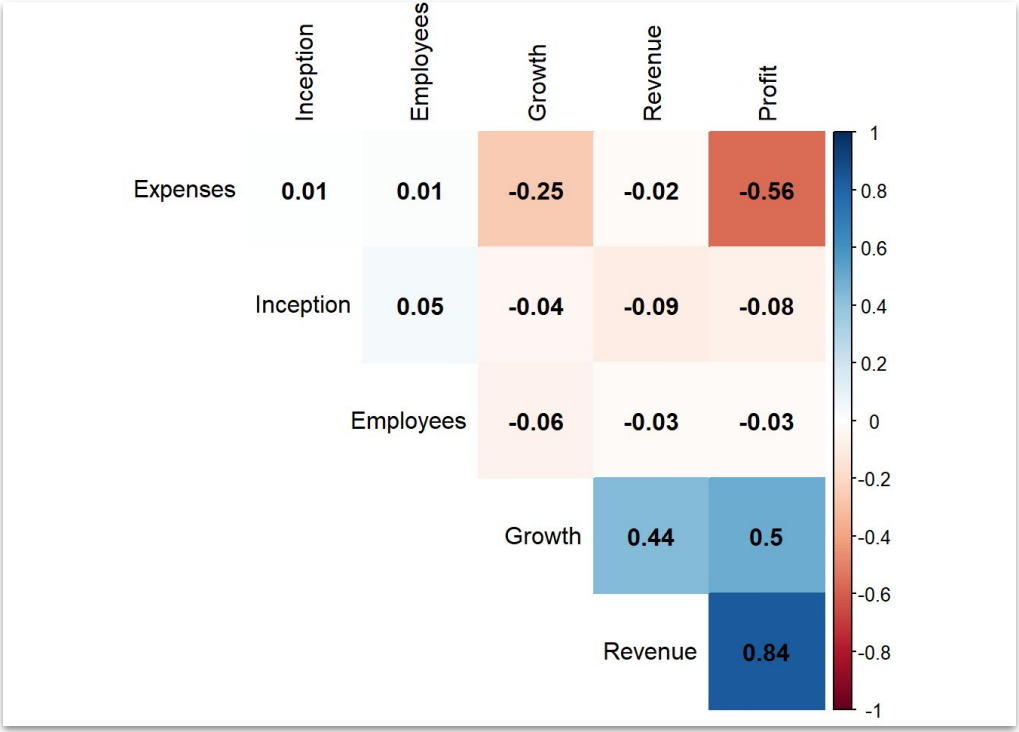


Min-max normuota duomenų aibė



Standartizuota duomenų aibė

Požymių koreliacijos

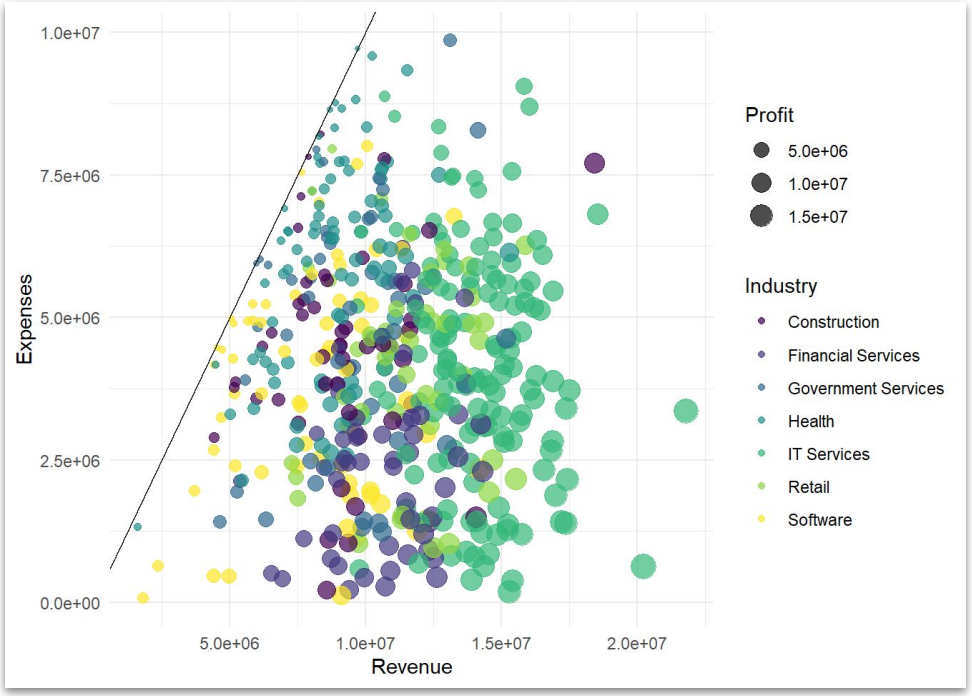


Pelnas. Pajamos. Išlaidos.

Sklaidos diagrama pavaizduotas įmonių pajamų ir išlaidų sklaidos diagrama kartu su palyginamąja tiese.

Įmonės duomenų aibėje nepatyrė nuostolių (nėra pavaizduotos kairėje palyginamosios tiesės pusėje).

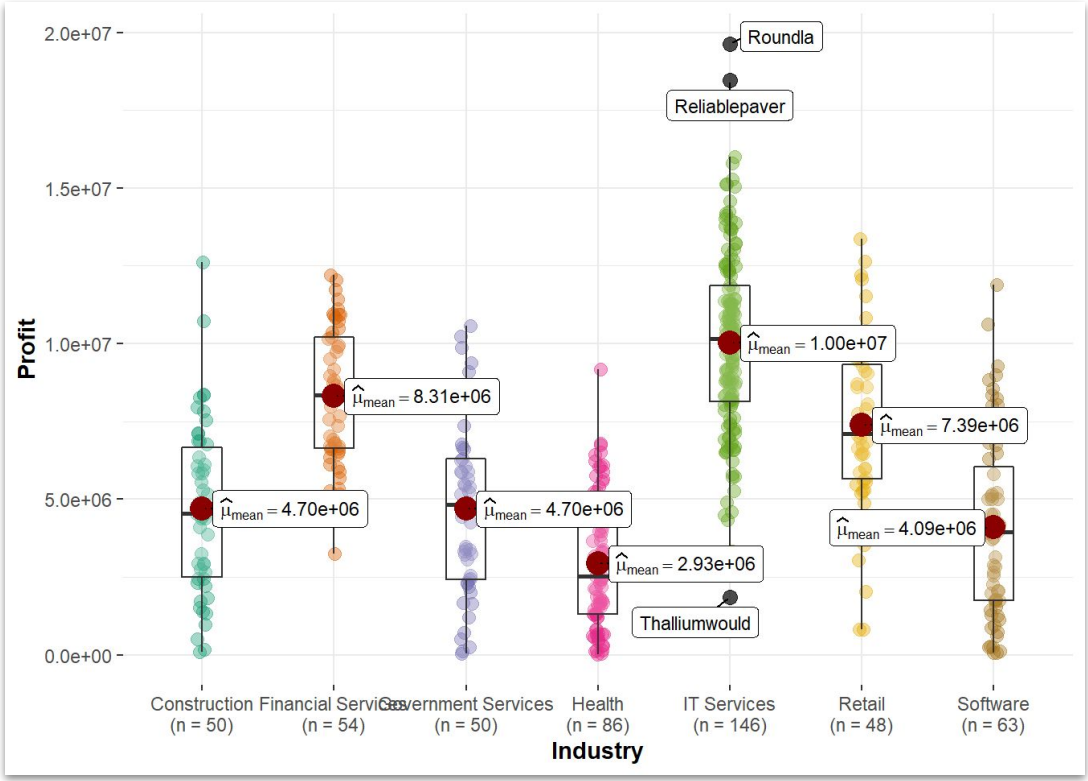
Pelningiausios yra IT Services įmonės.



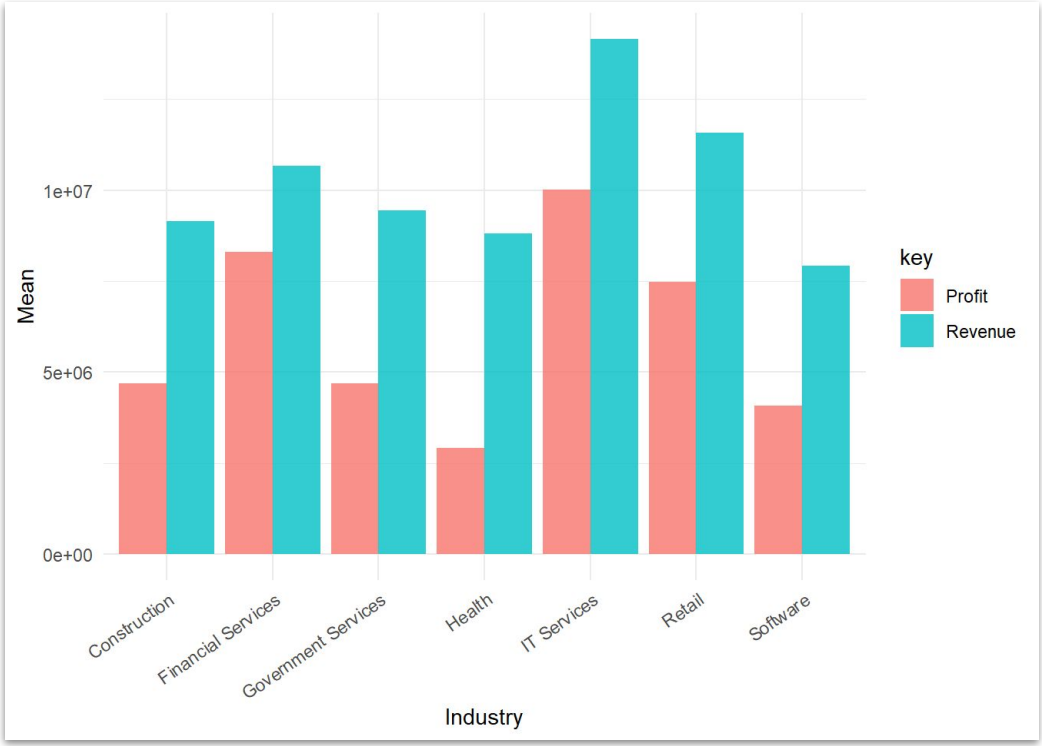
Pagal pramonės sritį

IT Services išsiskiria iš kitų pramonės šakų aukščiausiomis pajamomis ir pelnu (medianinės reikšmės atitinkamai 28% ir 21% didesnės už antroje vietoje pagal šiuos požymius esančias pramonės šakas)

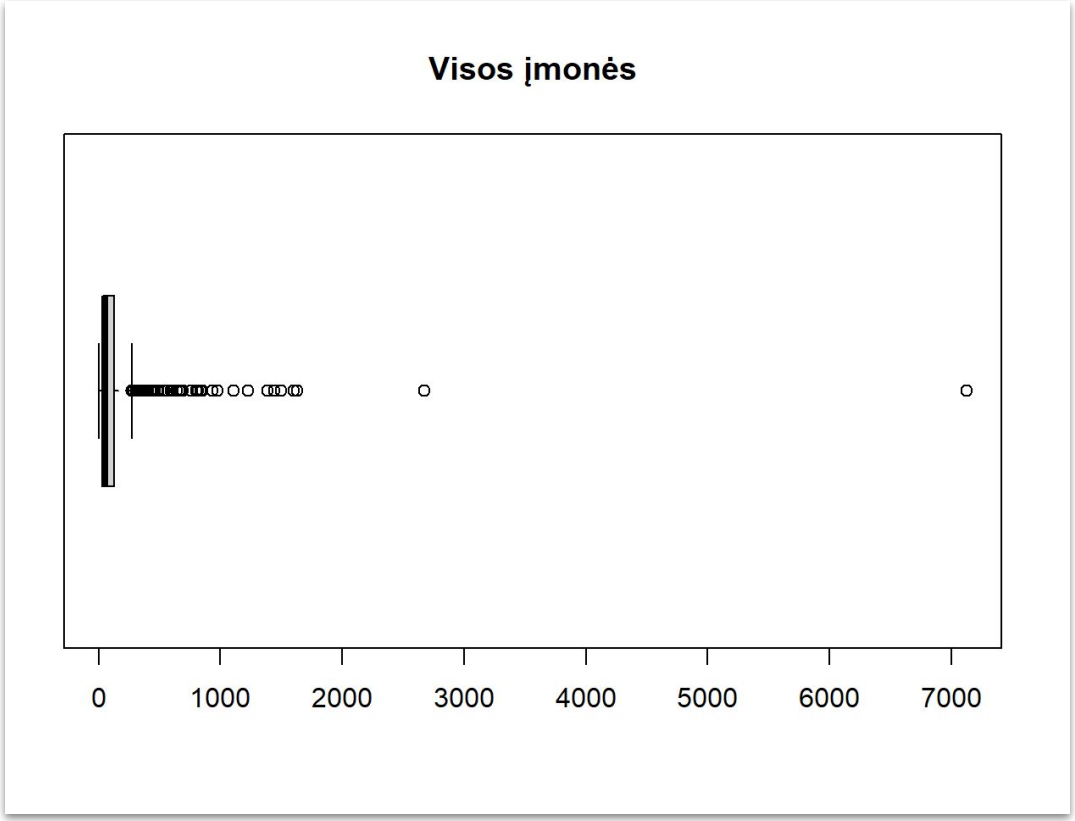
Health - žemiausiu pelnu (mediana 36% mažesnė už antrą šiuo požymiu mažiausią) ir aukščiausiomis išlaidomis (mediana 13% didesnė už antrą šiuo požymiu didžiausią).



Pagal pramonės sritį



Darbuotojų skaičiaus problematika



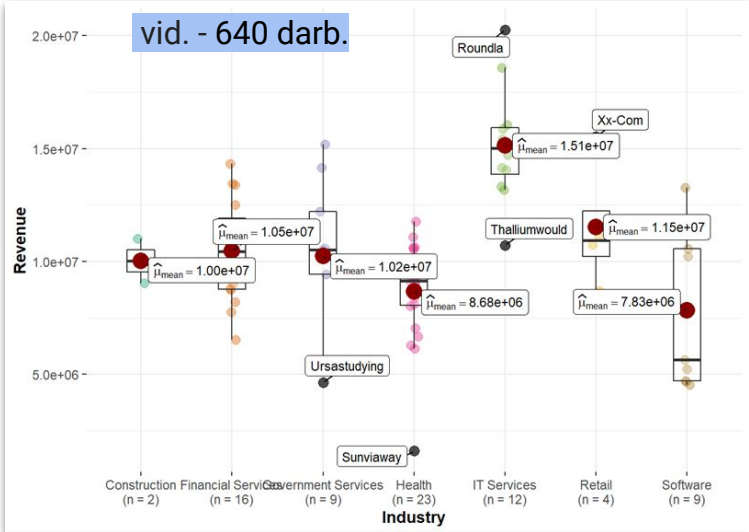
Darbuotojų skaičius

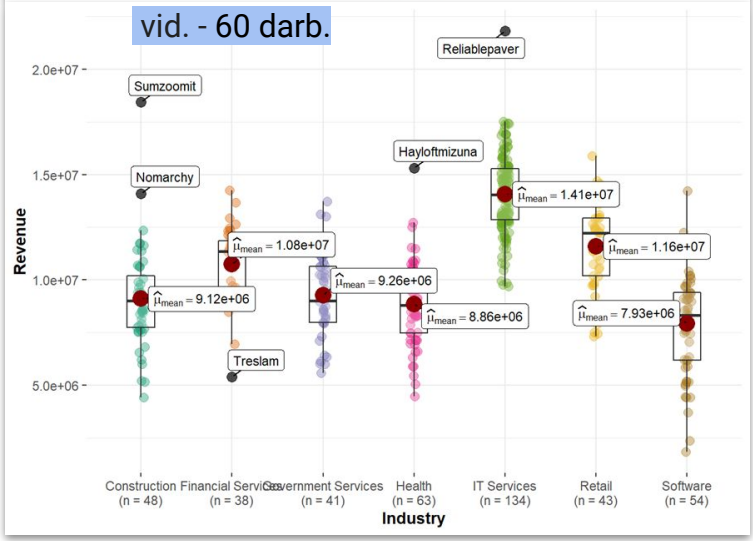
Darbuotojų skaičiaus problematika

Įmonės suskirstytos į dvi grupes:
15% didžiausių pagal darbuotojų skaičių ir likusios 85% įmonių.

Atskyrus įmones nebuvo rasta tendencijų pagal turimus požymius.

Tose pačiose pramonės šakose pelno vidurkių skirtumas tarp didelių ir mažų įmonių yra tarp -9% iki +9%.

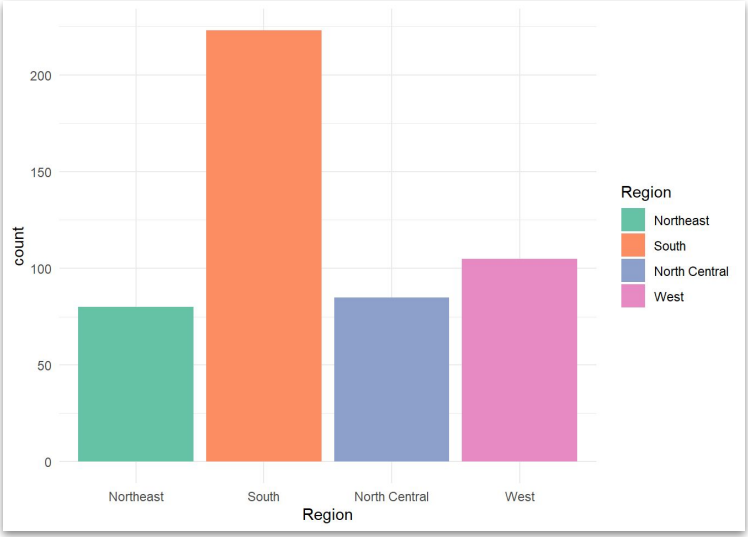


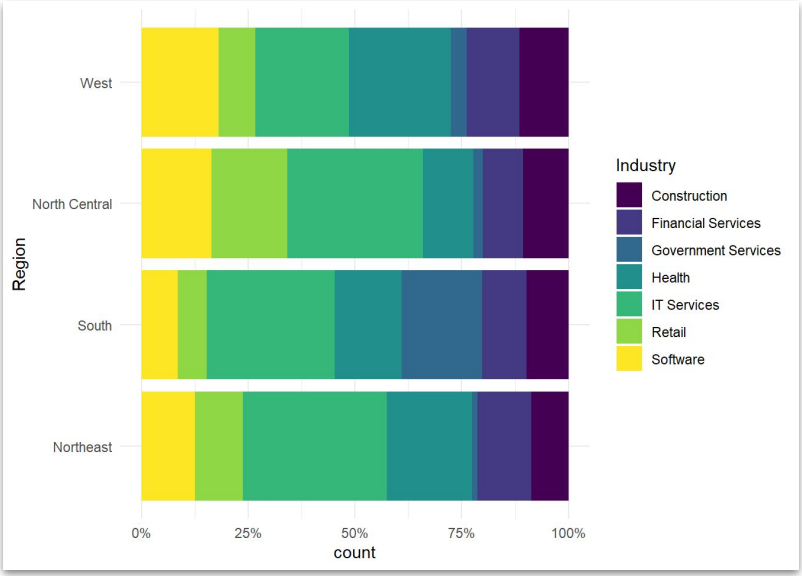


Pagal regioną

Stulpeline diagrama pavaizduotas įmonių kiekviename regione skaičius.

45% įmonių duomenų aibėje yra iš pietinio JAV regiono.

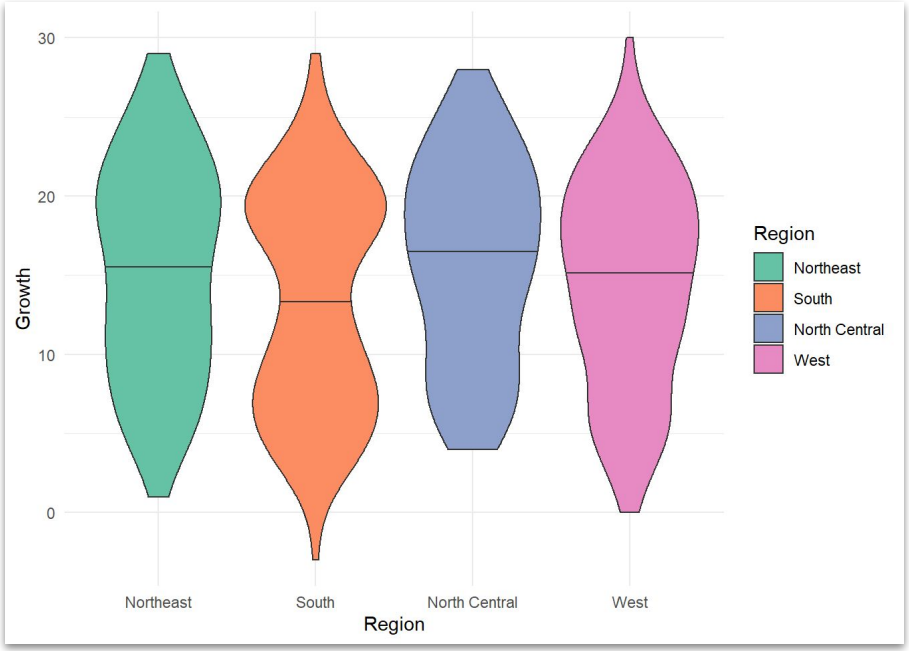




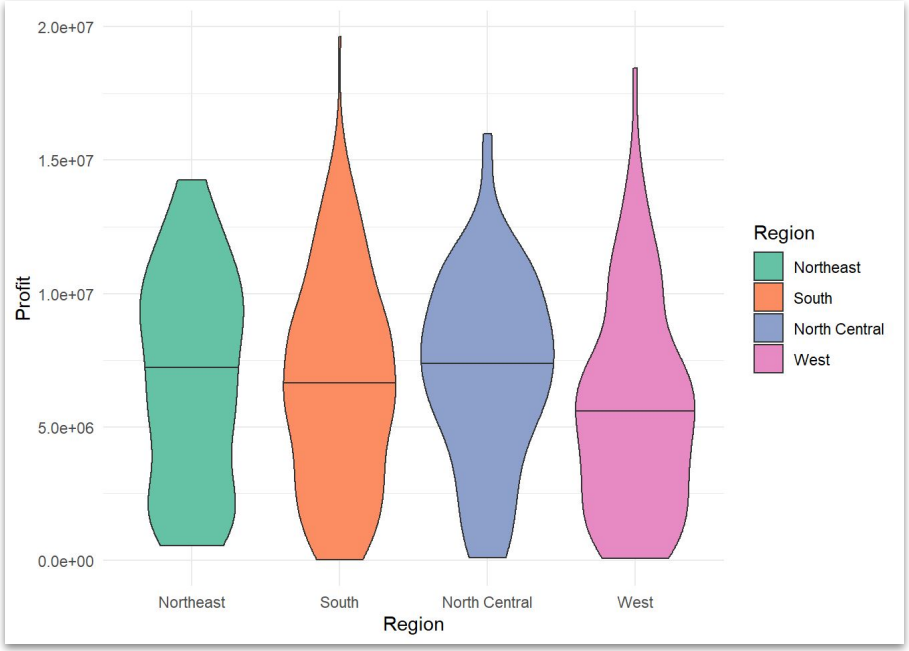
Kiekvienam regionui stulpeline diagrama pavaizduota kokią dalį įmonių sudaro tam tikrai pramonės šakai priklausančios įmonės.

Grafike galima matyti, kad įmonių pasiskirstymas labai panašus visuose 4 regionuose. Daugiausiai yra IT Services įmonių.

Pagal regioną

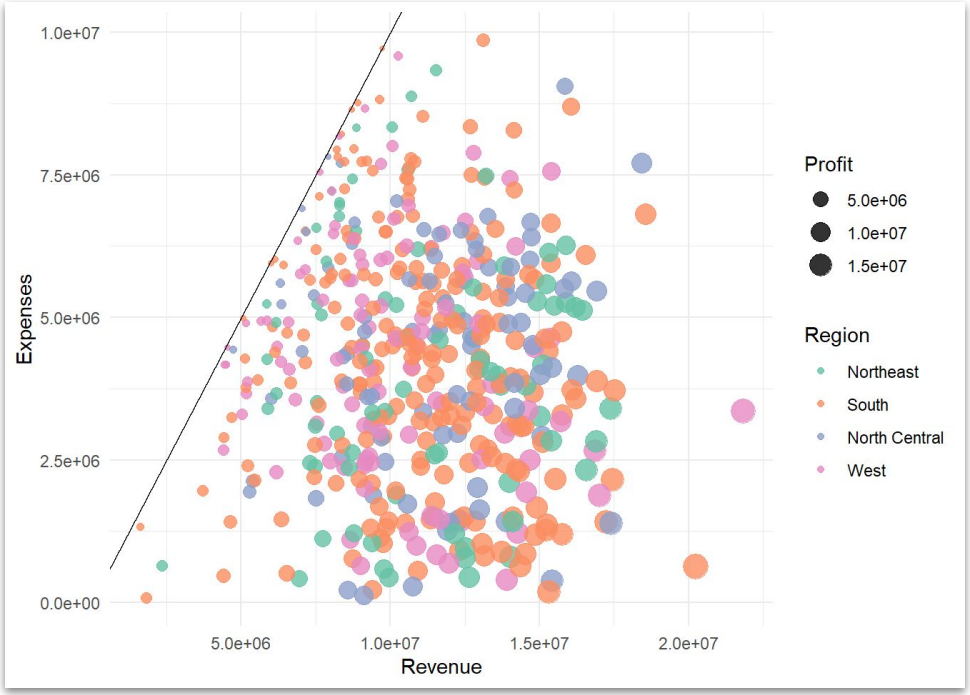


Įmonės augimas (%)

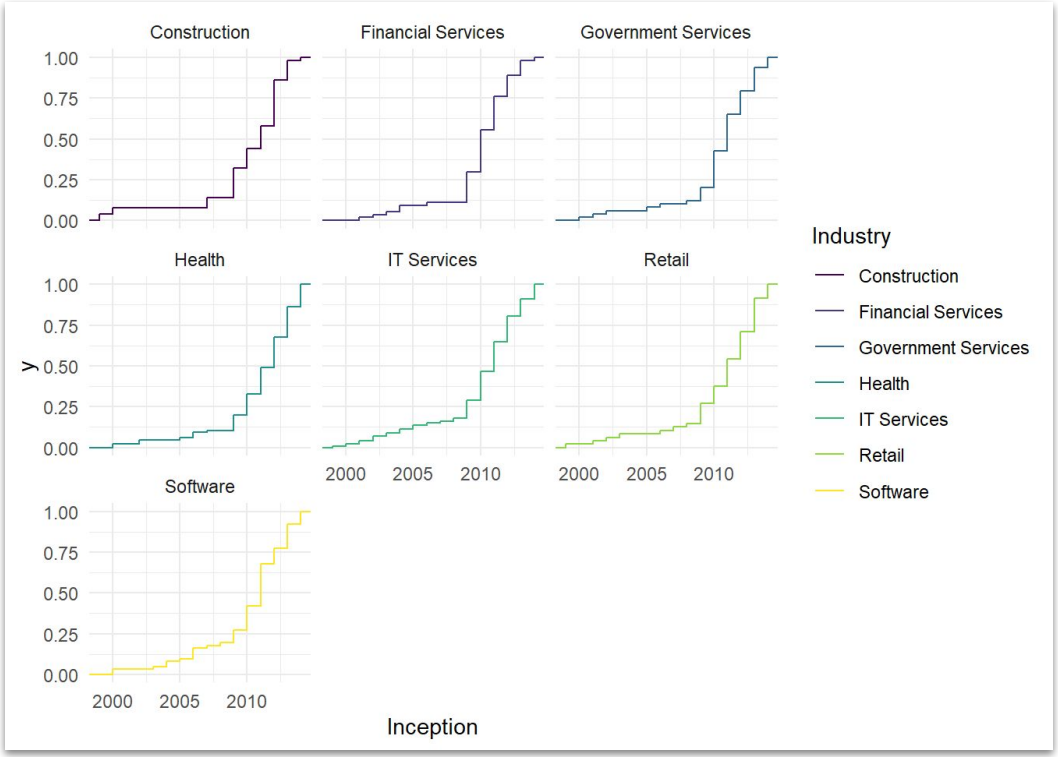


Pelnas ($)

Pagal regioną

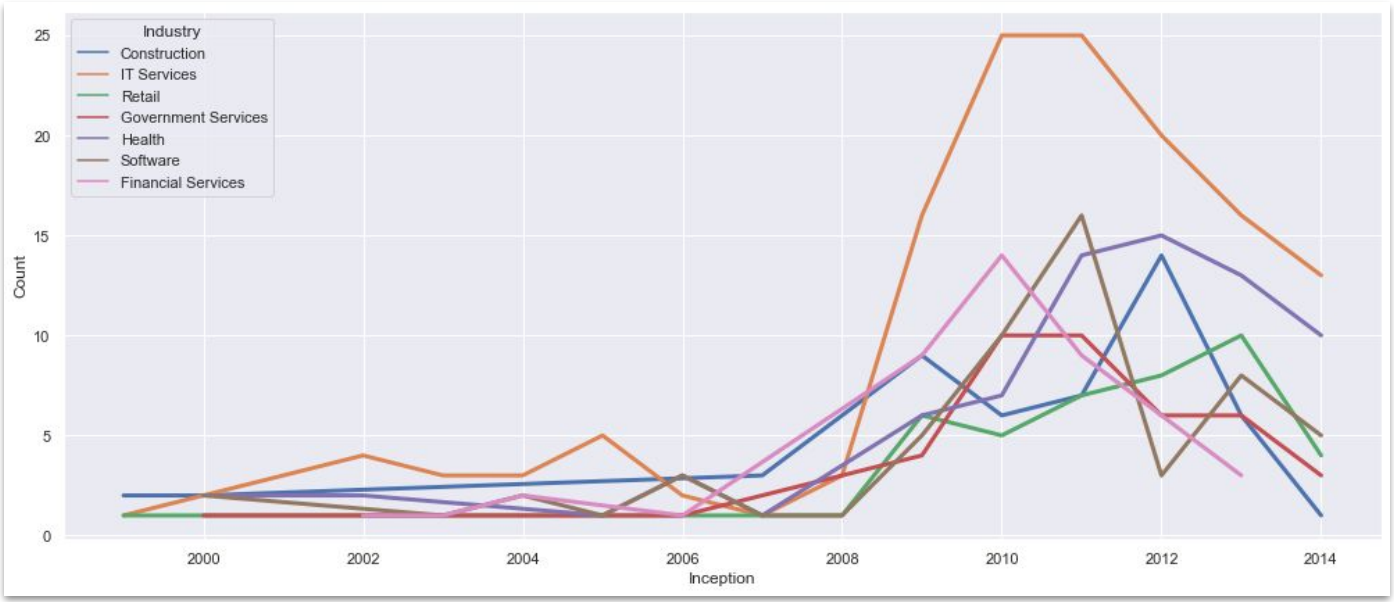


Pagal įkūrimo metus

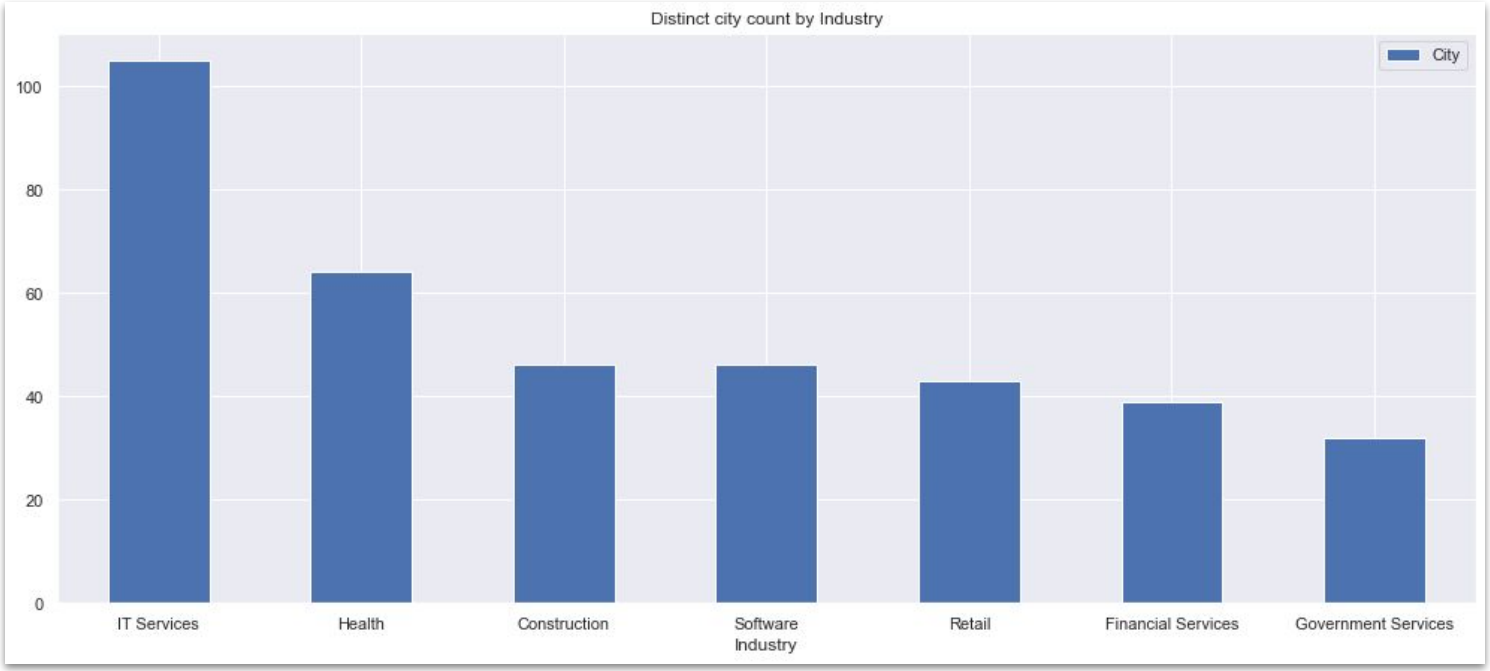


(Vaizduojamas empirinis pasiskirstymas)

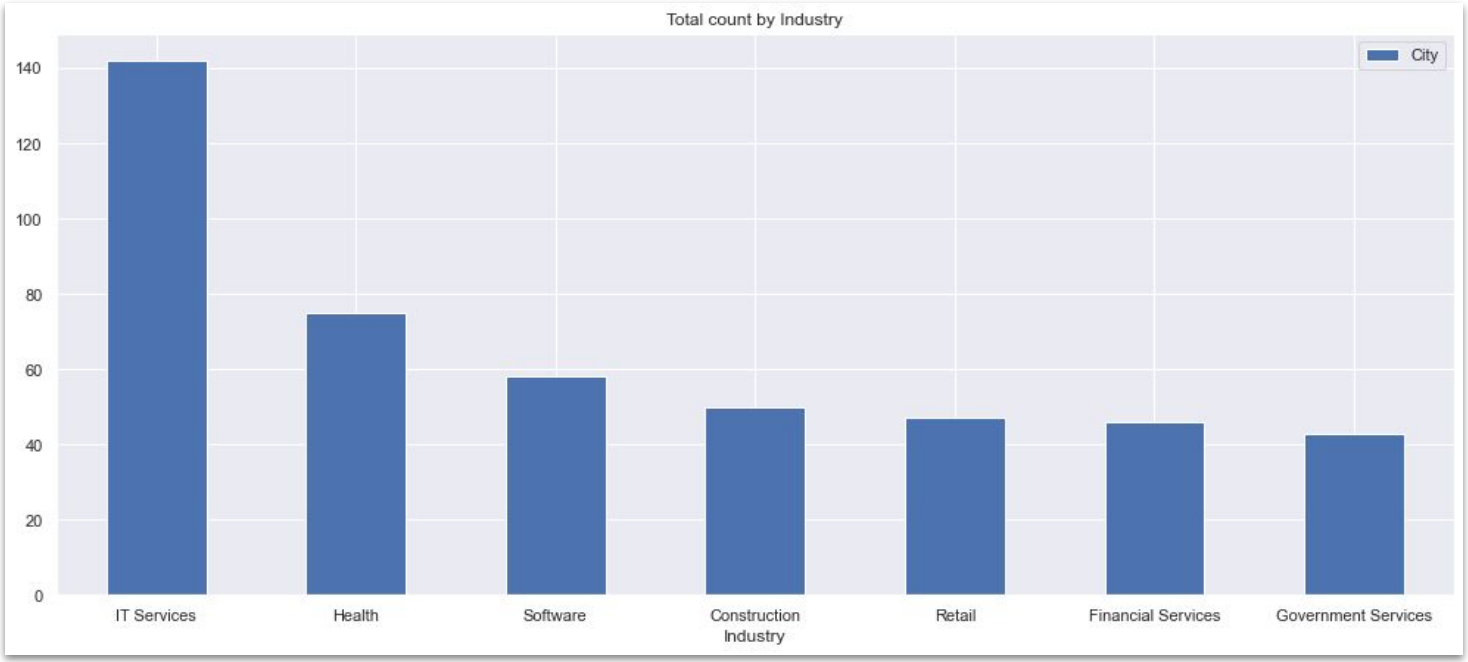
Pagal įkūrimo metus



Pagal miestą



Pagal miestą



Ačiū už dėmesį