KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FUKULTETAS

Intelektikos pagrindai

(P176B101)

Laboratorinis darbas Nr. 1
Duomenų apdorojimas ir analizė

Darbą atliko:
IFF-7/6 gr. studentė
Indrė Pabijonavičiūtė

Darbą priėmė:
lekt. BUDNIKAS Germanas
doc. PAULAUSKAITĖ-TARASEVIČIENĖ Agnė

KAUNAS 2020

Turinys

1 Įvadas

1.1 Tikslas

Duomenų apdorojimas ir analizė

1.2 Uždaviniai

1. Pasirinkti (susikurti) duomenų rinkinį
2. Atlikti duomenų rinkinio kokybės analizę
3. Nupaišyti atributų histogramas
4. Identifikuoti duomenų kokybės problemas
5. Nustatyti sąryšius tarp atributų panaudojant vizualizacijos būdus
6. Paskaičiuoti kovariacijos ir koreliacijos reikšmes
7. Atlikti duomenų normalizaciją
8. Kategorinio tipo kintamuosius paversti į tolydinio tipo kintamuosius

2 Duomenų kokybės analizė

2.1 Pasirinktas duomenų rinkinys

Pasirinktas „Google Play App Store“ mobiliesiems įrenginiams skirtų programų duomenų rinkinys, aprašantis daugiau nei 40000 programų charakteristikas:

1 lentelė. Duomenų rinkinio atributai

Atributas	Tipas	Prasmė	Pavyzdys
App	Kategorinis	Programėlės pavadinimas	Coloring book moana
Category	Kategorinis	Programėlės kategorija	ART_AND_DESIGN
Rating	Tolydinis	Reitingas X/5	3.9
Reviews	Tolydinis	Atsiliepimų skaičius	967
Size	Tolydinis	Programėlės dydis	14M
Installs	Tolydinis	Apibendrintas parsisiuntimų skaičius	500,000+
Price	Tolydinis	Kaina	0
Content Rating	Kategorinis	Amžiaus grupė, kam skirta programėlė	Everyone
Last Updated	Kategorinis	Paskutinio atnaujinimo data	January 15, 2018
Latest Version	Kategorinis	Dabartinė programos versija	2.0.0
Minimum Version	Kategorinis	Žemiausia Minimum Versionsija, palaikanti programą	4.0.3 and up

2.2 Duomenų rinkinio analizė

Analizuojamos pagrindinės atributų charakteristikos.

2.2.1 Tolydinio tipo atributai

2 lentelė. Tolydinio tipo atributų charakteristikos

Atributas	Reikšmių skaičius	Trūkstamos reikšmės, %	Kardinalumas	Minimali	Maksimali	1 ir 3 kvartiliai		Vidurkis	Mediana	Standartinis nuokrypis
Rating	42841	3.448	42	1.0	19.0	4.1	4.6	4.107	4.4	0.522
Reviews	42841	0.002	18730	0	86214292	108	19978	185945	1555	1794496.925
Size	42841	11.414	730	0.004	347.0	6.200	45.000	19.691	17.0	21.775
Price	42841	0.005	146	0.0	1458.0	0.00	0.00	0.511	0.0	11.058
Installs	42841	0.005	23	0	5000000000	10000	1000000	6994973	100000	66548527.962

2.2.2 Kategorinio tipo atributai

3 lentelė. Kategorinio tipo atributų charakteristikos

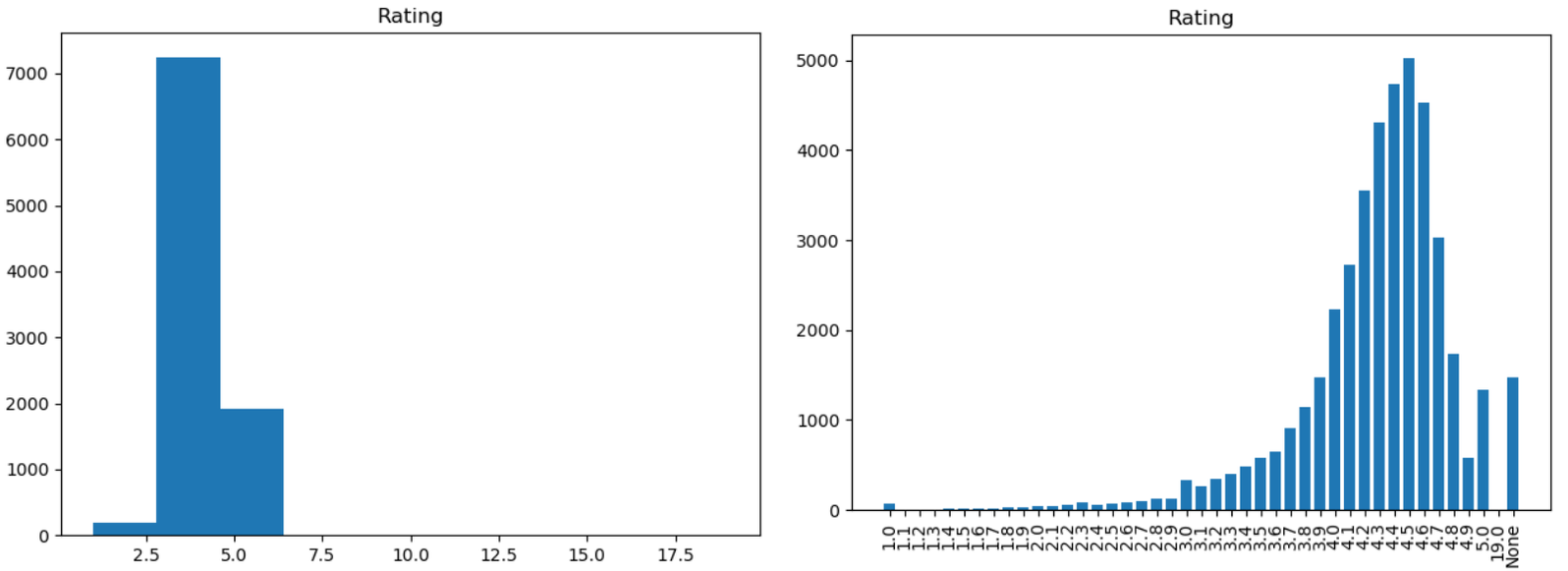
Atributas	Kiekis	Trūkst. reikšm.,%	Kardinalumas	Moda	Modos dažn.	Moda, %	2-oji Moda	2-osios Modos dažn.	2-oji Moda,%
App Name	42841	0.000	38425	????	52	0.121	?????	36	0.08
Category	42841	0.000	55	EDUCATION	3587	8.373	TOOLS	3242	7.57
Content Rating	42841	0.002	10	Everyone	36169	84.426	Teen	4068	9.50
Last Updated	42841	0.000	2042	April 2, 2019	1090	2.544	April 1, 2019	859	2.01
Latest Version	42841	0.021	8218	Varies with device	3926	9.164	1	1784	4.16
Minimum Version	42841	0.007	43	4.1 and up	10916	25.480	4.0.3 and up	6714	15.67

3 Duomenų grafinis atvaizdavimas ir analizė

3.1 Histogramos

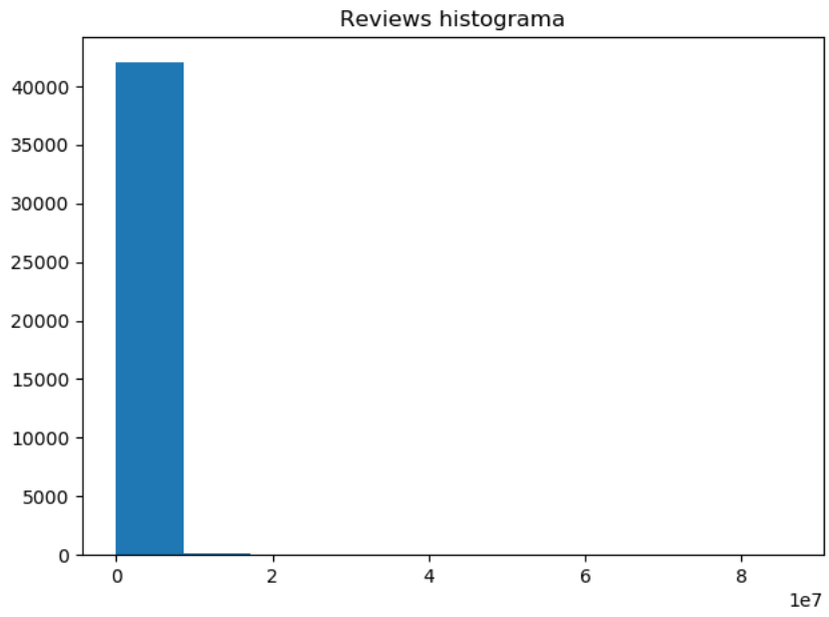
Brėžiamos atributų reikšmių dažnių diagramos – histogramos. Braižomos tik tos stulpelinės diagramos, kurių numatomas stulpelių skaičius neviršija slenksčio, lygaus 1 + 3.22 * ln(reikšmių skaičius) * 3.

3.1.1 Tolydiniai atributai



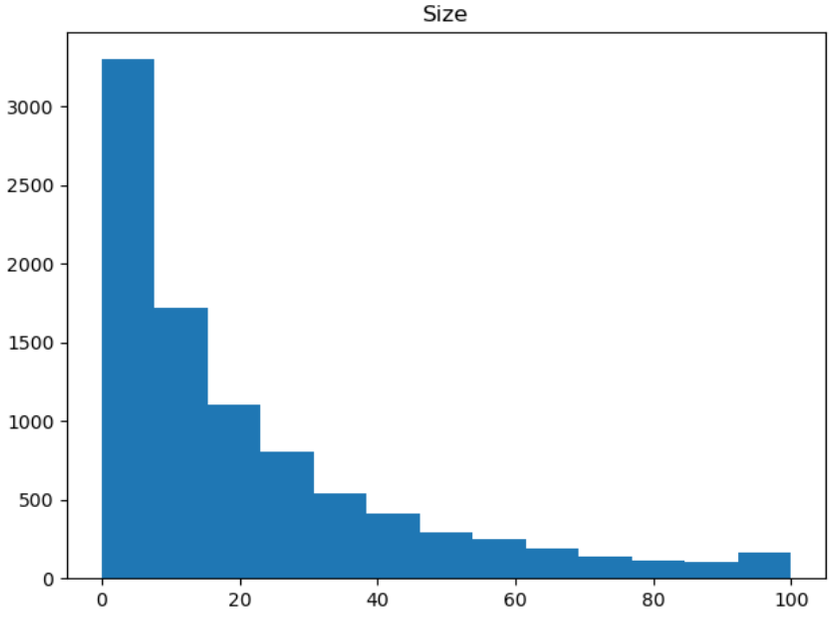
1 pav. Atributo "Rating" histograma ir stulpelinė diagrama

Atributo „Rating" reikšmės histogramoje pasiskirsčiusios netolygiai. Matomas duomenų defektas, kadangi maksimalus galimas vertinimas yra 5, o duomenyse pateiktas maksimalus vertinimas – 19. Brėžiant stulpelinę diagramą, matyti, kad reitingo reikšmės sudaro pasislinkusį į dešinę normalų pasiskirstymą, susitelkęs aplinkui 4.4 medianą. Matoma, jog didelės dalies duomenų trūksta.



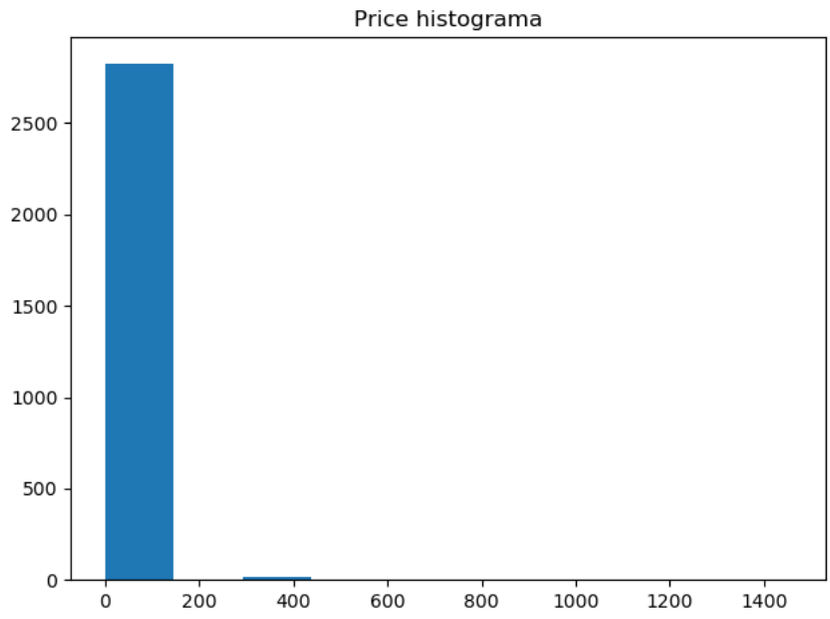
2 pav. Atributo "Reviews" histograma

Atributo „Reviews" reikšmės pasiskirsčiusios pagal atvirkštinį eksponentinį skirstinį. Daugiausia programėlių turi nedidelį kiekį atsiliepimų.



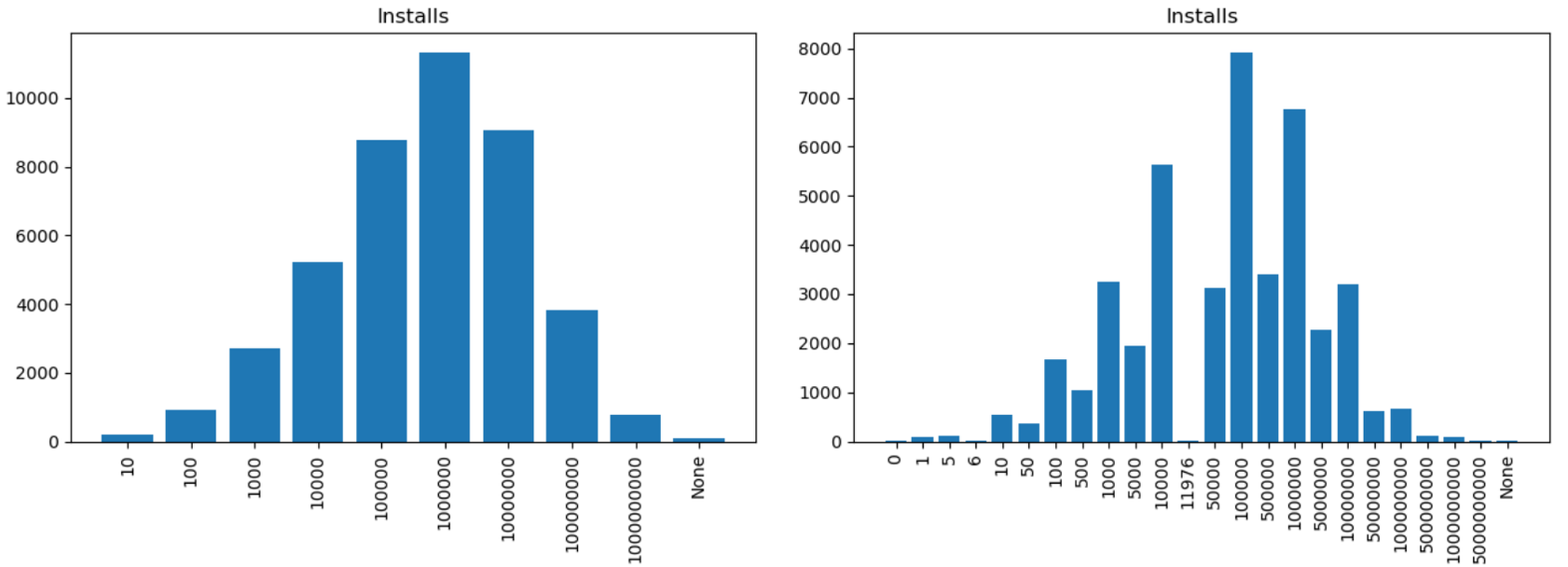
3 pav. Atributo "Size" historama

Atributo „Size" reikšmės pasiskirsčiusios pagal atvirkštinį eksponentinį skirstinį, su nuokrypiu tiriamo intervalo pabaigoje. Galima daryti išvadą, kad daugiausia programėlių yra mažiausio dydžio, ir atvirkščiai. Iš lentelės žinoma, kad trūksta dalies reikšmių.



4 pav. Atributo "Price" histograma

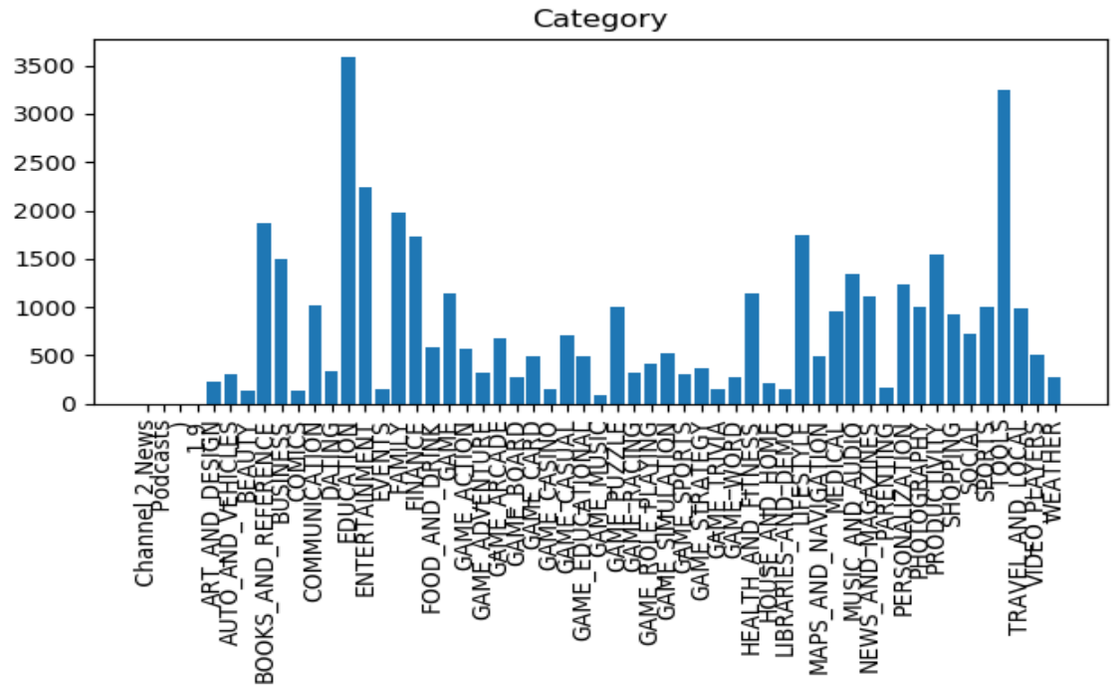
Atributo „Price" reikšmės pasiskirsčiusios netolygiai. Matoma, jog absoliuti dauguma programėlių yra nemokamos, išskyrus nedidelį kiekį nukrypusių programėlių nuo šios taisyklės.



5 pav. Atributo "Installs" stulpelinė diagrama

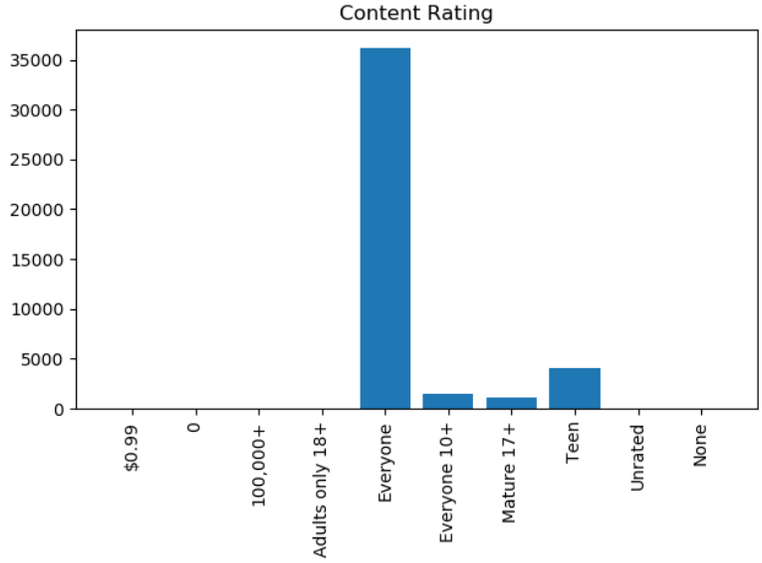
Pašalinus tarpines reikšmes (prasidedančias skaitmeniu 5), matomas atributo „Installs" reikšmių pasiskirstymas, panašus į normalųjį. Tai reiškia, kad reikšmės nenutolusios nuo medianos. Matoma, kad nedidelės dalies duomenų trūksta, įsimaišiusios neatitinkančios formato reikšmės.

3.1.2 Kategoriniai atributai



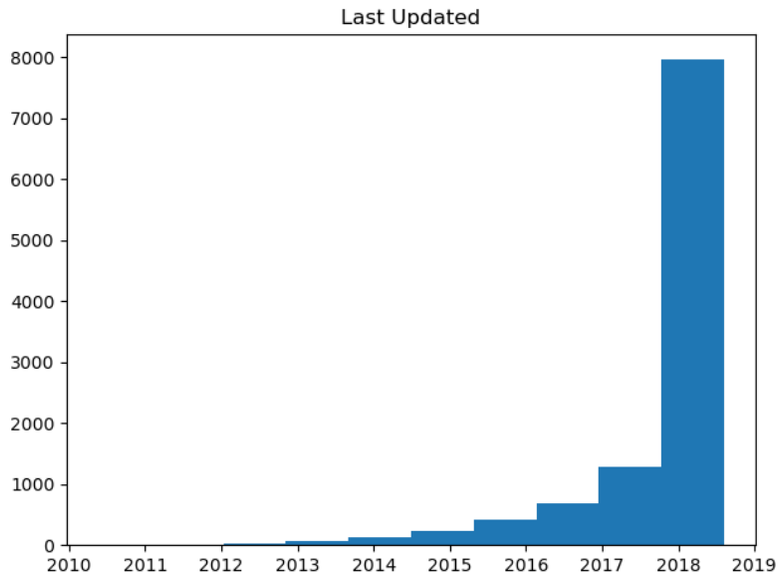
6 pav. Atributo "Category" stulpelinė diagrama

Atributo „Category" reikšmės pasiskirsčiusios netolygiai. Matomas išskirtinai didelis reikšmės „FAMILY" dažnis, taigi daug programėlių yra orientuotos į šeimą. Tikėtinos reikšmės – tekstinės, vienodo formato. Matoma duomenų klaida – skaitinės reikšmės, formato neatitinkančios tekstinės reikšmės.



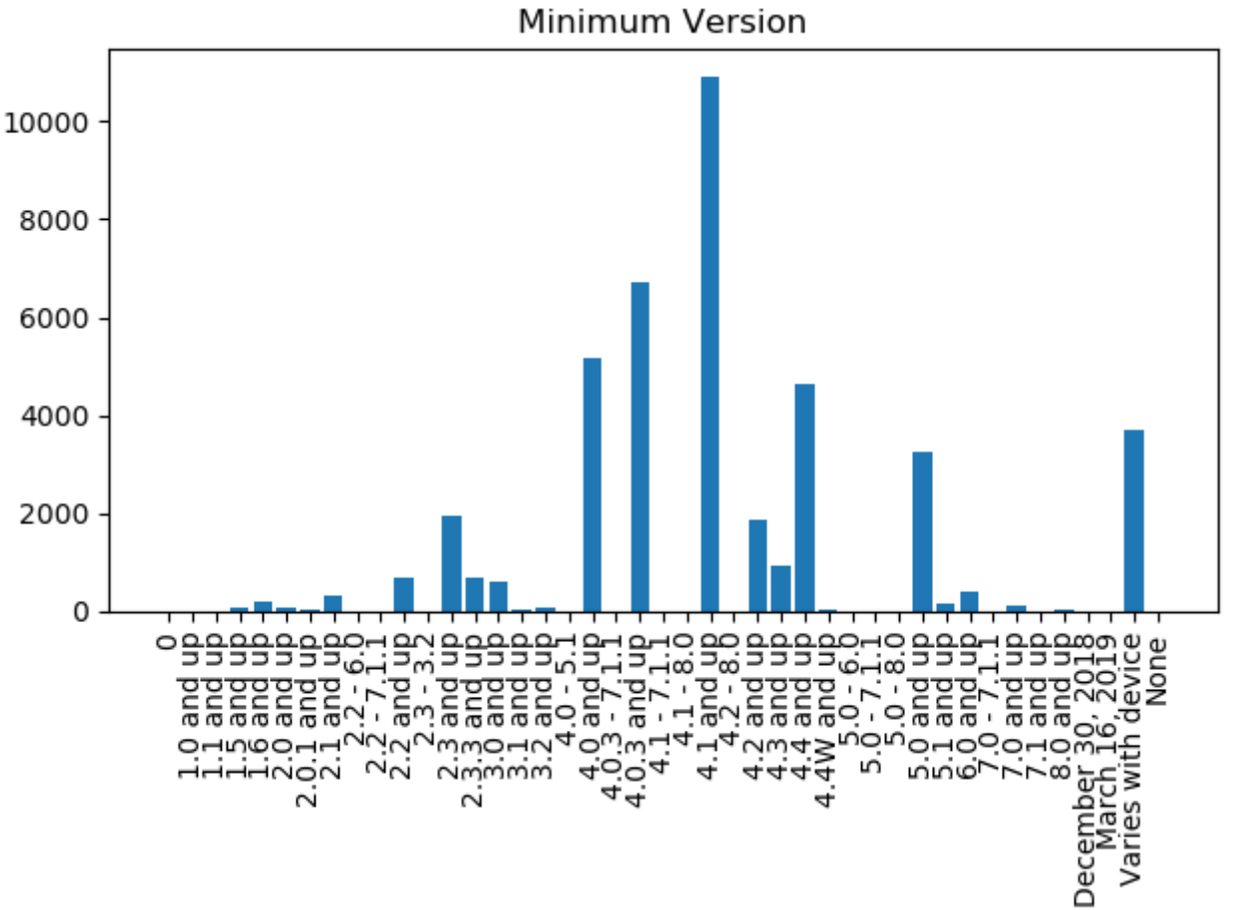
7 pav. Atributo "Content Rating" stulpelinė diagrama

Atributo „Content Rating" pasiskirstymas netolygus. Daugiausia programėlių neriboja amžiaus, kuriam tinkamas jų turinys. Pastebimos netikėtos reikšmės.



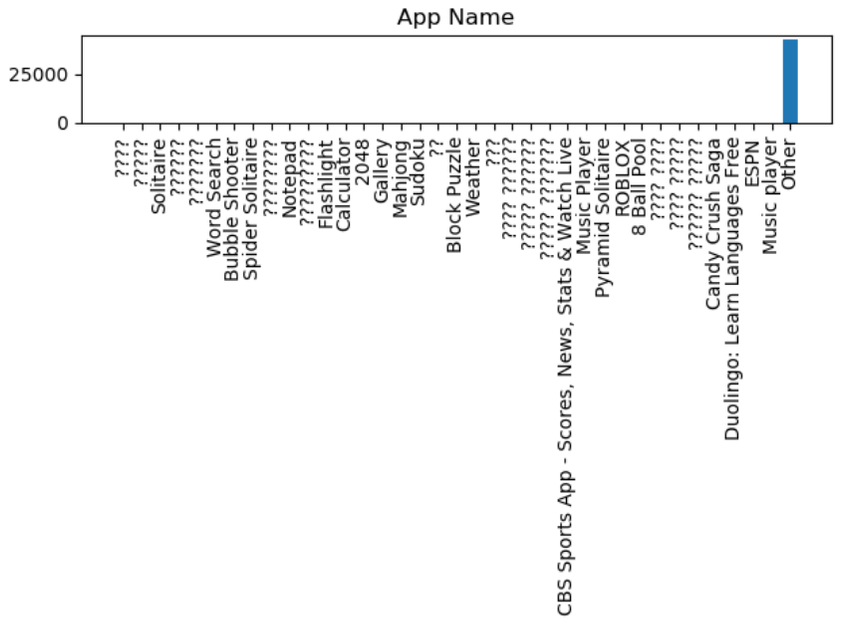
8 pav. Atributo „Last Updated“ histograma

Atributo „Last Updated“ pasiskirstymas eksponentinis, tai reiškia, kad labai maža dalis programų atnaujintos anksčiau, nei duomenų rinkinio sudarymo metais – 2018. Absoliuti dauguma programų buvo aktyviai naujinamos.

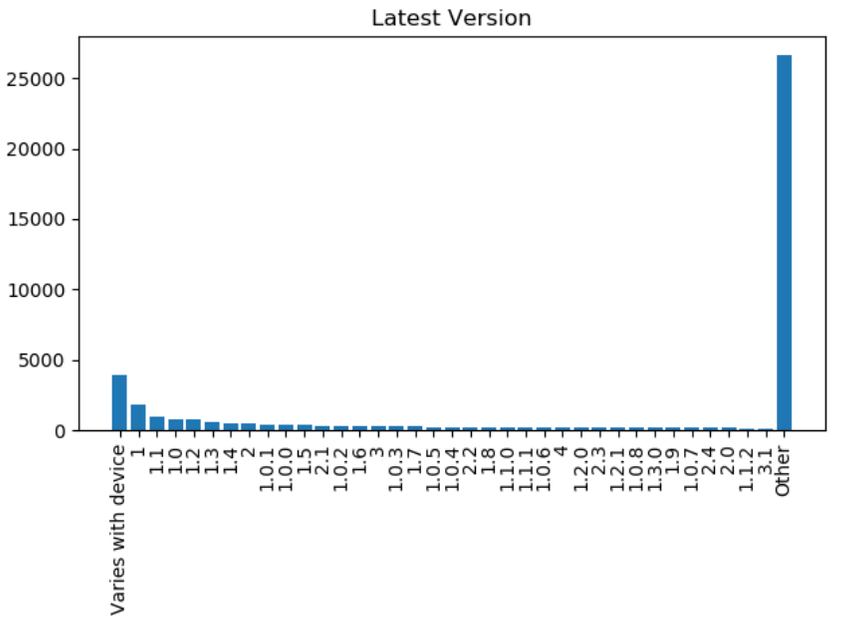


9 pav. Atributo „Minimum Version“ stulpelinė diagrama

Atributo „Minimum Version“ reikšmės pasisikirsčiusios netolygiai. Matoma, kad daugiausia programų kuriama su jų palaikymu Minimum Versionsijoms nuo 4.1. Matoma, jog yra trūkstamų reikšmių, bei neatitinkančių tekstinio formato „X.Y and up“.



10 pav. Atributo „App Name“ stulpelinė diagrama



11 pav. Atributo „Latest Version“ stulpelinė diagrama

Abiejų atributų reikšmių pasiskirstymas tolydinis. Reikšmės unikalios (sutapimai atsitiktiniai), kadangi programos pavadinimas unikalus, o versijų numeracija nestandartizuota.

3.2 Duomenų kokybės problemų identifikacija ir sprendimas

Duomenų problemų analizė pateikiama lentele:

4 lentelė. Duomenų kokybės problemų analizė

Atributas	Tipas	Problema	Sprendimas
App	Kategorinis	Reikšmės unikalios (kardinalumas apytiksliai lygus reikšmių skaičiui)	Atmesti atributą tolimesnėje analizėje
Category	Kategorinis	Įsimaišiusios netikėtos skaitinės ir tekstinės reikšmės	Pakeisti reikšmes moda
Rating	Tolydinis	Reikšmės, iškrentančios iš intervalo [0;5]	Pakeisti reikšmes mediana
		Tuščios reikšmės	

Reviews	Tolydinis	Ekstremalios reikšmės	Pritaikyti korekciją, pagrįstą Q1 ir Q3 skirtumu
Size	Tolydinis	Tuščios reikšmės	Įterpti atsitiktinę reikšmę iš imties su apskaičiuota tikimybe
Installs	Tolydinis	Tuščios reikšmės	Įterpti atsitiktinę reikšmę iš imties su apskaičiuota tikimybe
Price	Tolydinis	Ekstremalios reikšmės	Nekoreguojama, nes tikimasi rasti sąsajų su kitais atributais, kai kainos ekstremalios
Content Rating	Kategorinis	Netikėtos reikšmės, tuščios reikšmės	Įterpti atsitiktinę reikšmę iš imties su apskaičiuota tikimybe
Last Updated	Kategorinis	Trūkstamos reikšmės	Įterpti atsitiktinę reikšmę iš imties su apskaičiuota tikimybe
Latest Version	Kategorinis	Reikšmės unikalios (sutapimai atsitiktiniai), kadangi versijų numeracija nestandartizuota	Atmesti atributą tolimesnėje analizėje
Minimum Version	Kategorinis	Trūkstamos reikšmės	Įterpti atsitiktinę reikšmę iš imties su apskaičiuota tikimybe
		Neatitinkančios formato reikšmės	

Pasirinktų problemų sprendimo būdų apibendrinimas:

a) Jei yra trūkstamų reikšmių:

Pirmiausia bandyta vietoje trūkstamų reikšmių įrašyti medianą/modą, tačiau pastebėta, jog, dėl nemažo trūkstamo reikšmių kiekio, reikšmių pasiskirstymas išsikraipo, todėl išvystytas sprendimas, vietoje trūkstamų įterpiantis atsitiktinę reikšmę iš jau egzistuojančių reikšmių imties, su tų reikšmių pasikartojimo imtyje tikimybe.

b) Jei yra netikėto formato reikšmės:

Nustatytos regulairiosios išraiškos neatitinkančios reikšmės traktuojamos kaip trūkstamos

c) Jei yra triukšmas:

Taikomas ekstremalių reikšmių koregavimas remiantis Q1 ir Q3

d) Jei visos atributo reikšmės unikalios, o sutapimai atsitiktiniai:

Atributas šalinamas iš tolimesnės analizės

Duomenų analizė po korekcijos:

Tolydniai atributai:

5 lentelė. Koreguotų tolydinio tipo atributų charakteristikos

Atributas	Reikšmių skaičius	Trūkstamos reikšmės, %	Kardinalumas	Minimali	Maksimali	1 ir 3 kvartiliai		Vidurkis	Mediana	Standartinis nuokrypis
Rating	42841	0.0	41	1.0	5.0	4.1	4.6	4.220	4.4	0.484
Reviews	42841	0.0	11783	0	49783	108	19974	12941	1555	18994.734
Size	42841	0.0	720	0.004	103.2	6.2	27.0	21.619	17.0	21.596
Price	42841	0.0	146	0.0	1458	0.00	0.00	0.511	0.0	11.058
Installs	42841	0.0	14	1	6994972	10000	1000000	624727	100000	890541.919

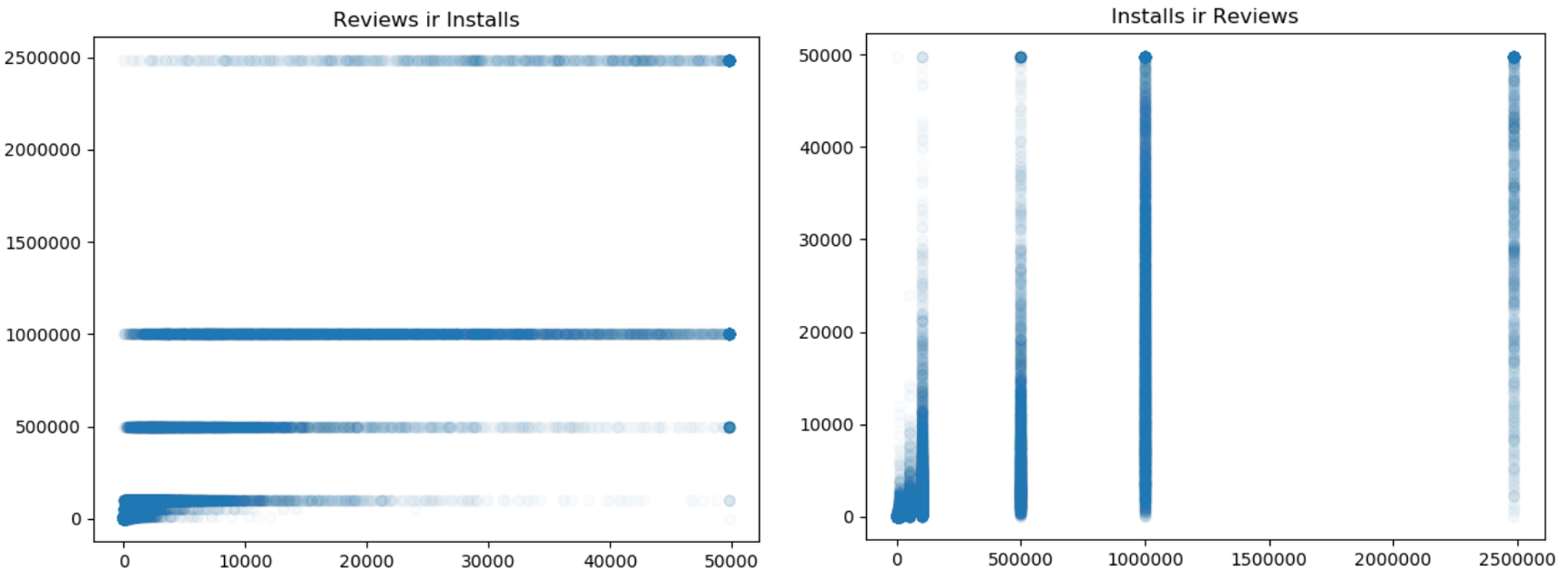
Kategoriniai atributai:

6 lentelė. Koreguotų kategorinio tipo atributų charakteristikos

Atributas	Reikšmių skaičius	Trūkstamos reikšmės, %	Kardinalumas	Moda	Modos dažnumas	Moda, %	2-oji Moda	2-osios Modos dažnumas	2-oji Moda, %
Category	42841	0.0	51	EDUCATION	3591	8.382	TOOLS	3242	7.57
Type	42841	0.0	2	Free	10041	92.62	Paid	800	7.38
Content Rating	42841	0.0	6	Everyone	36173	84.435	Teen	4068	9.50
Minimum Version	42841	0.0	25	4.1 and up	14678	34.262	4.0.3 and up	6714	15.67

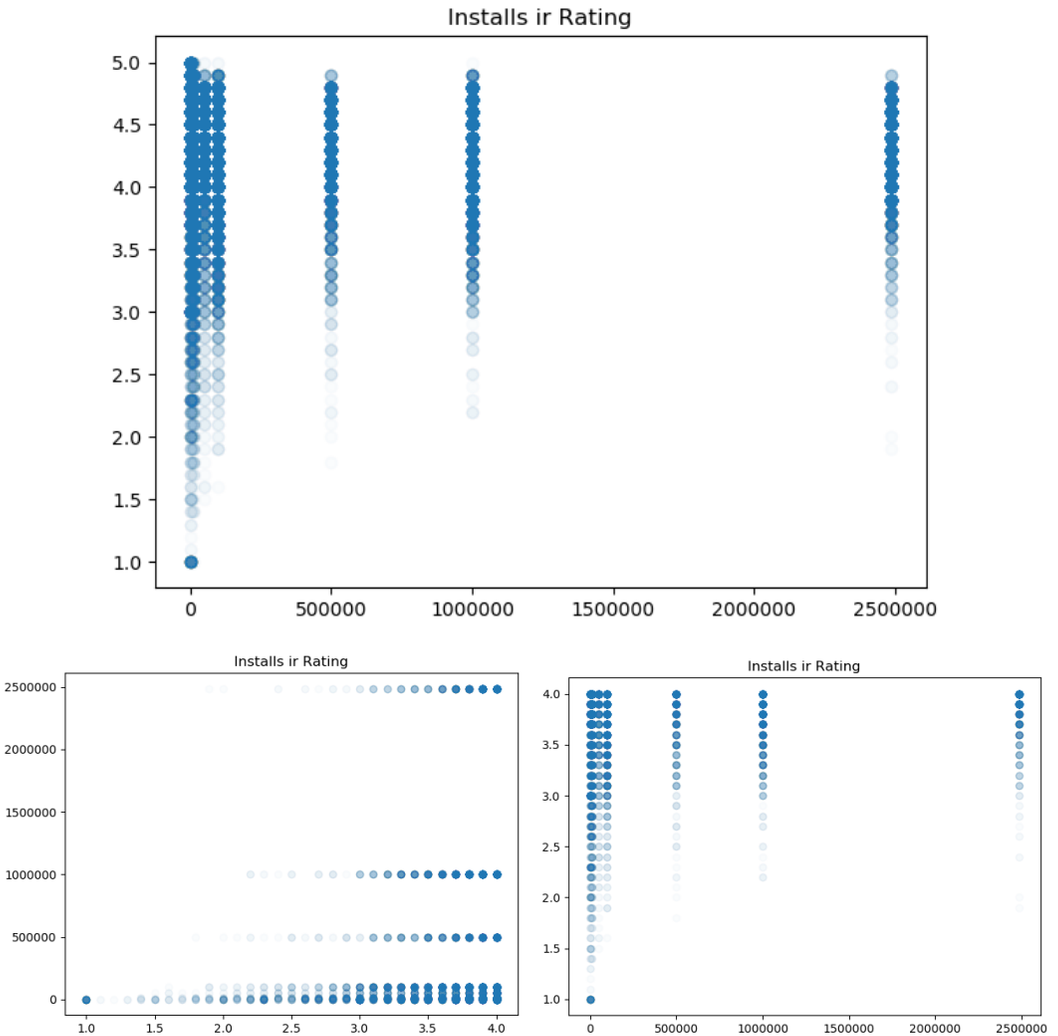
4 Sąryšių tarp atributų vizualizacija

4.1 Sąryšiai tarp tolydinio tipo atributų



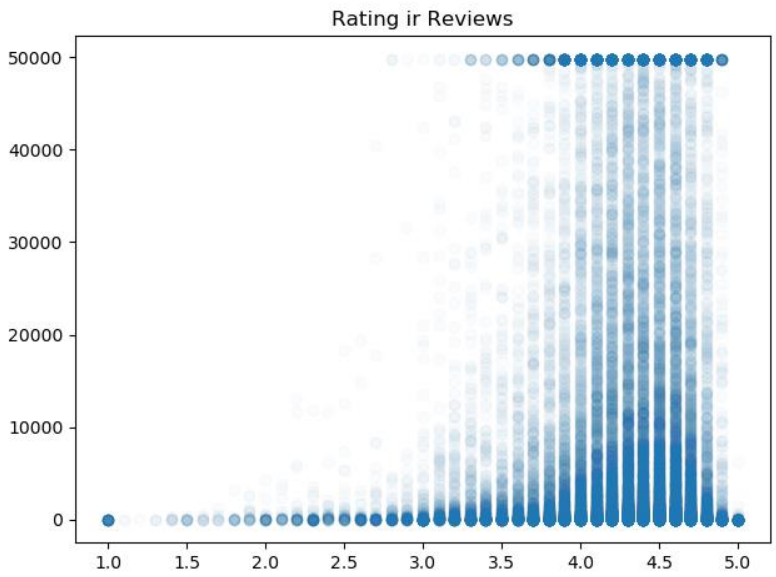
12 pav. Reviews ir Installs atributų sklaidos grafikas

Atsiliepimų skaičiaus ir parsisiuntimų skaičiaus sąryšis pastebimai stiprus, tačiau grafikas neišvaizdus, kadangi Installs – intervalinis atributas.



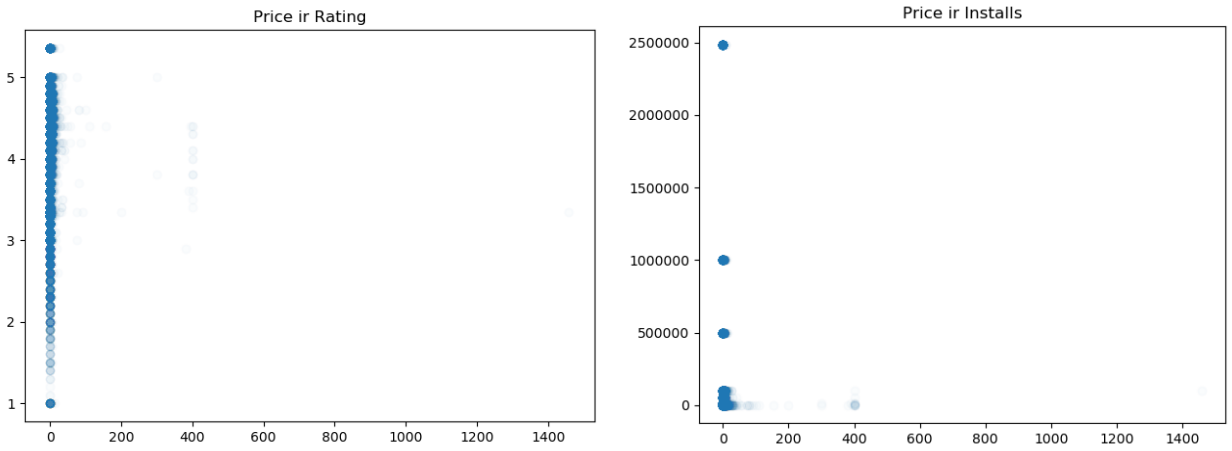
13 pav Rating ir Installs sklaidos grafikas.

Tarp parsisiuntimų ir reitingo priklausomybė neišreikšta (koreliacija = 0.07). Matoma, kad itin mažai parsisiuntimų turinčios programėlės dažniau gauna mažą įvertinimą, bet nuo tam tikro parsisiuntimų (reitingo) taško, reitingo (parsisiuntimų) reikšmės išlaiko tą patį intervalą, nepriklausomai nuo parsiuntimų (reitingo) reikšmės, tačiau ir šis sąryšis silpnas: koreliacija=0.23



14 pav. Reviews ir Rating sklaidos grafikas

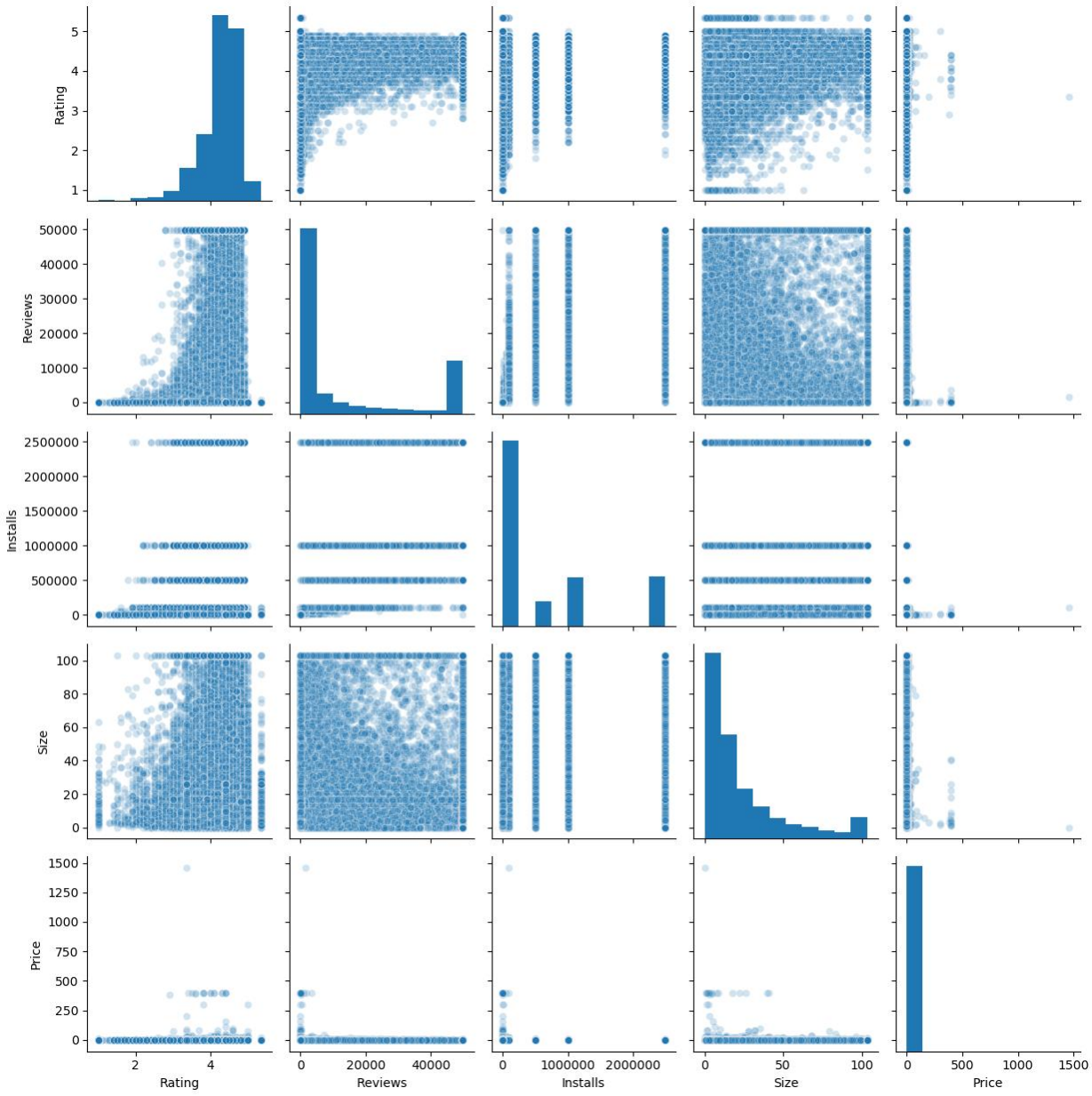
Matomas nežymus sąryšis tarp „Reviews“ ir „Rating“ atributų: aukštesnio reitingo programos šiek tiek labiau linkusios sulaukti daugiau atsiliepimų.



15 pav. Price ir Rating bei Price ir Installs sklaidos grafikas

Dydžio („Price“) ir reitingo („Rating“) bei parsisiuntimų „Installs“ atributai nesisieja: reitingo reikšmės pjūviuose pagal kainą išsidėsčiusios visame grafike labai panašiai, nes vos keletas kainos reikšmių didesnės, nei 0.

Scatter plot matrica:



16 pav. Tolydinių atributų sklaidos matrica

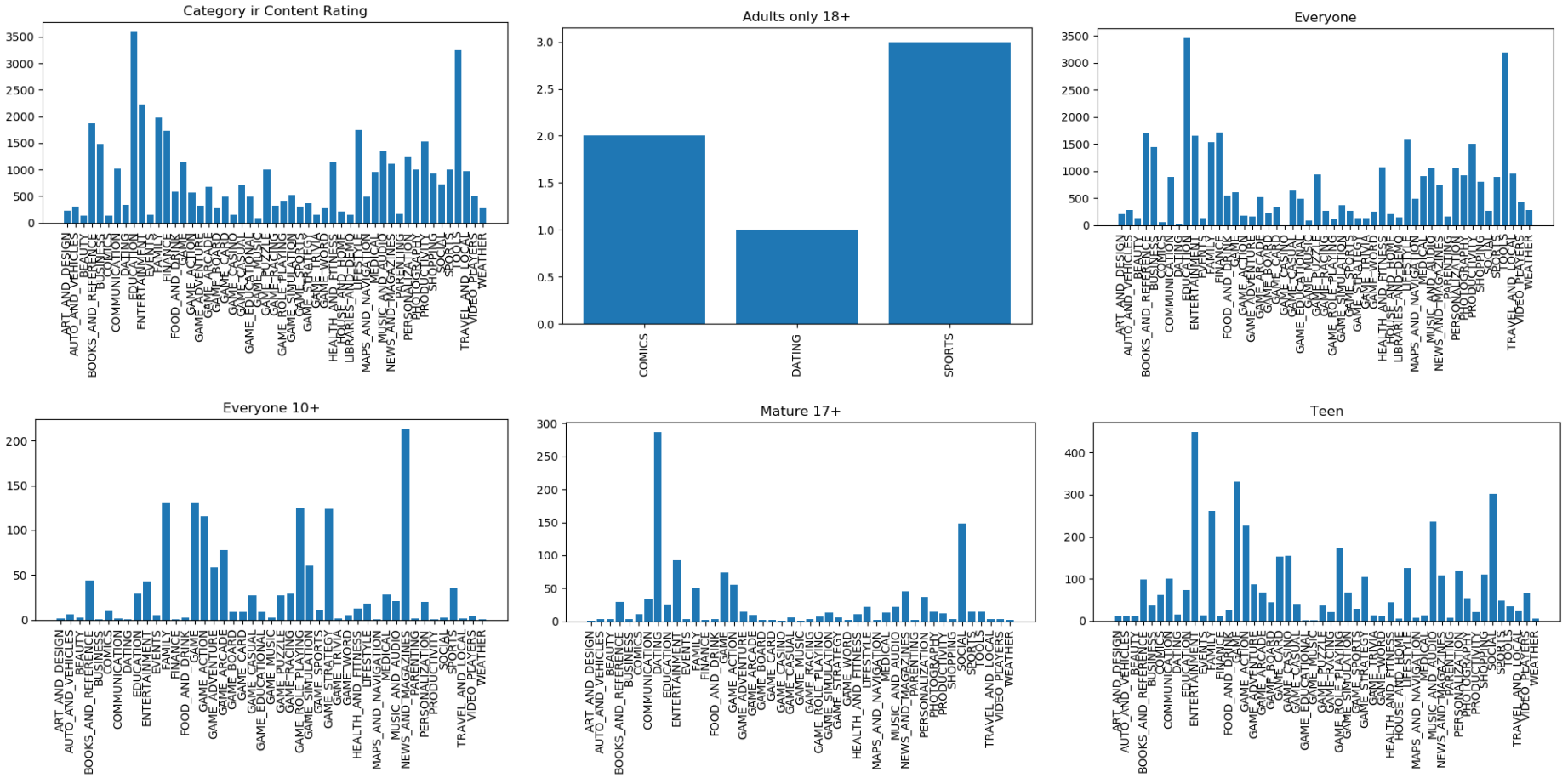
Kaip matoma iš scatter plot matrix (16 pav.), stipriai susietų tolydinių atributų duomenų rinkinyje nėra, išskyrus „Installs“ ir „Reviews“

4.2 Sąryšiai tarp kategorinio tipo atributų

Šio tipo atributų sąryšis pastebimas, jei stulpelinėse diagramose stulpelių pasiskirstymas neatsikartoja.

„Category“ ir „Content Rating“ atributų sąryšio tyrimas:

(1) (2) (3)



(4) (5) (6)

17 pav. „Category“ ir „Content Rating“ atributų priklausomybės diagrama

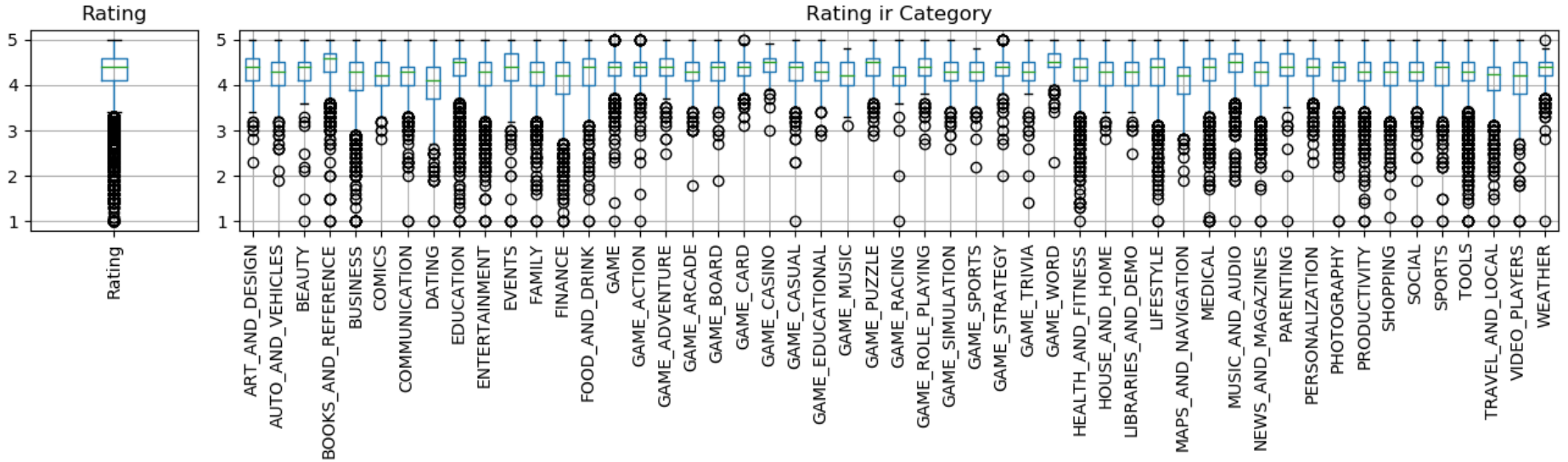
Matomas sąryšis tarp programėlės kategorijos ir turinio reitingavimo atributų: tik suaugusiems nuo 18 metų skirtoms programėlėms priskirtos komiksų, susitikinėjimo ir sporto kategorijos. Programėlių, kurios skirtos visiems – be amžiaus apribojimų, pasiskirstymas per kategorijas beveik atitinka originalią kategorijų pasiskirstymo diagramą. Toks sąryšis racionalus, kadangi kaip ir (1) grafike, taip ir (3) amžiaus grupė nėra specifikuota.

Visiems, kuriems daugiau nei 10 metų, skirtos programėlės daugiausia yra naujienų ir žaidimų kategorijos. Subrendusiems, nuo 17 metų amžiaus, skirtos programėlės daugiausia yra susitikinėjimo programėlės. Į paauglius daugiausia orientuoti žaidimai, pramogų ir socialinės tematikos programėlės.

Daugiau sąryšių duomenų rinkinyje tarp kategorinių atributų neaptikta.

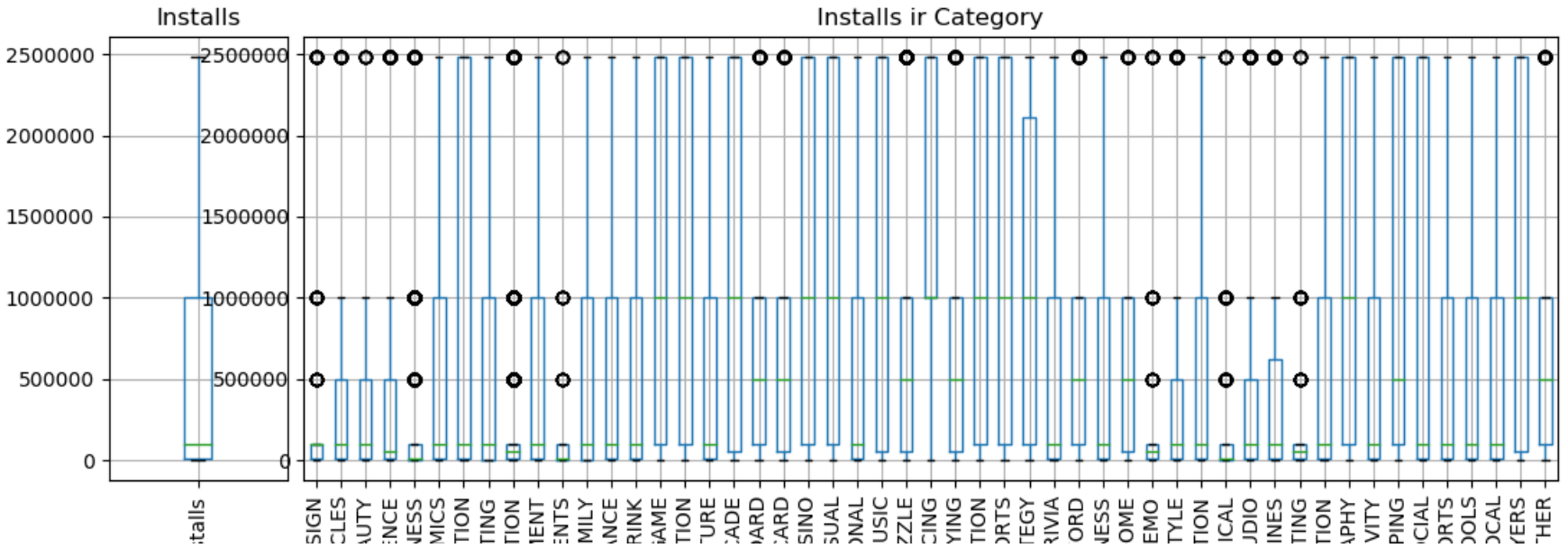
4.3 Sąryšiai tarp tolydinio ir kategorinio tipų atributų

Šio tipo sąryšiai tirti „box plot“ diagramomis, kuriose „dėžutė“ apima reikšmes, esančias intervale nuo Q1 ir Q1, „uodegėlės“ rodo sąlyginį minimumą ir maksimumą (Q1-1.5*IQR, Q3+1.5*IQR), o taškai už jų – ekstremalias reikšmes. Jei „dėžutės“ nepersidengia, laikoma, kad rastas sąryšis tarp atributų.



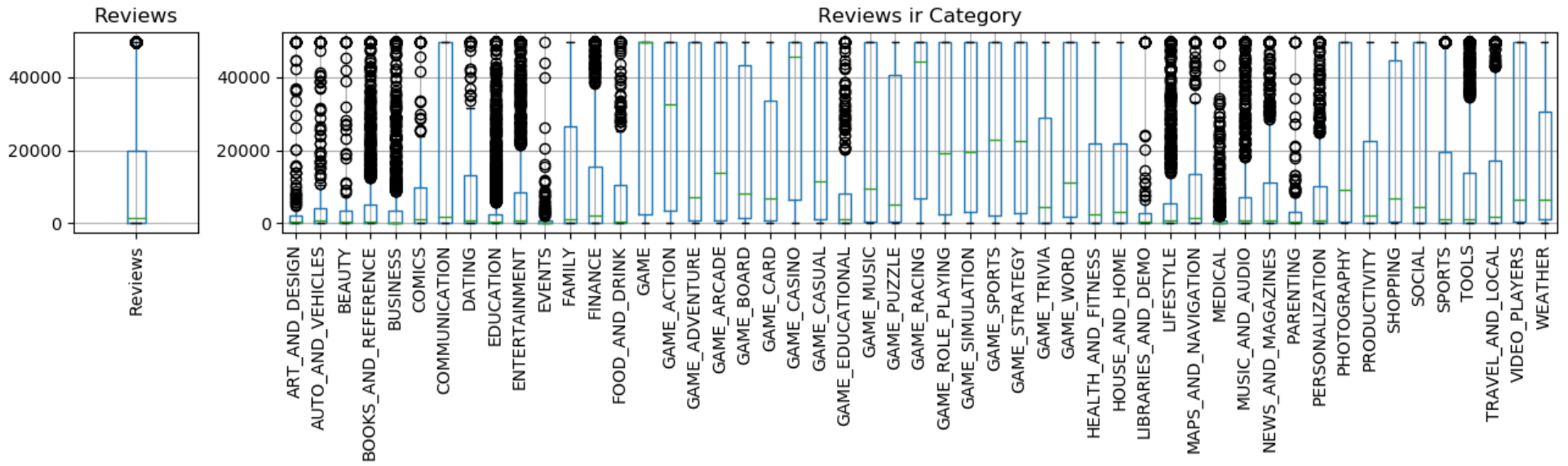
18 pav. „Rating“ ir „Category“ box plot

Reitingo intervalas išlieka panašus kiekvienoje kategorijoje, taigi sąryšio tarp kategorijos ir reitingo nėra.



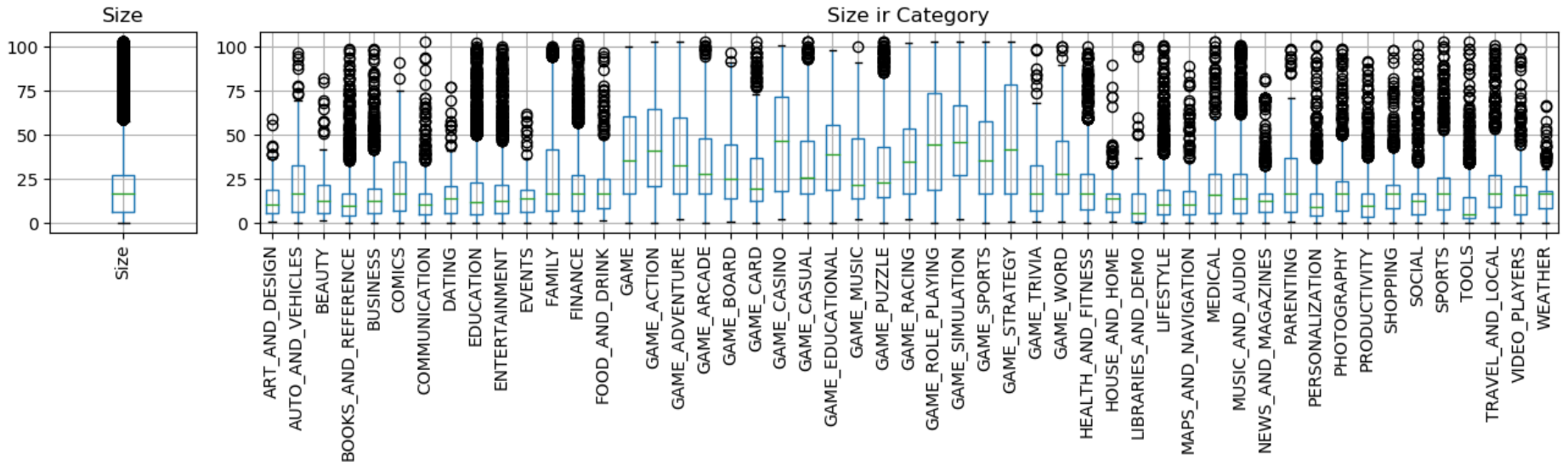
19 pav. „Installs“ ir „Category“ box plot

Sąryšis silpnas, beveik nepastebimas, tačiau matoma, jog keletoje kategorijų parsisiuntimų skaičiaus intervalas didesnis



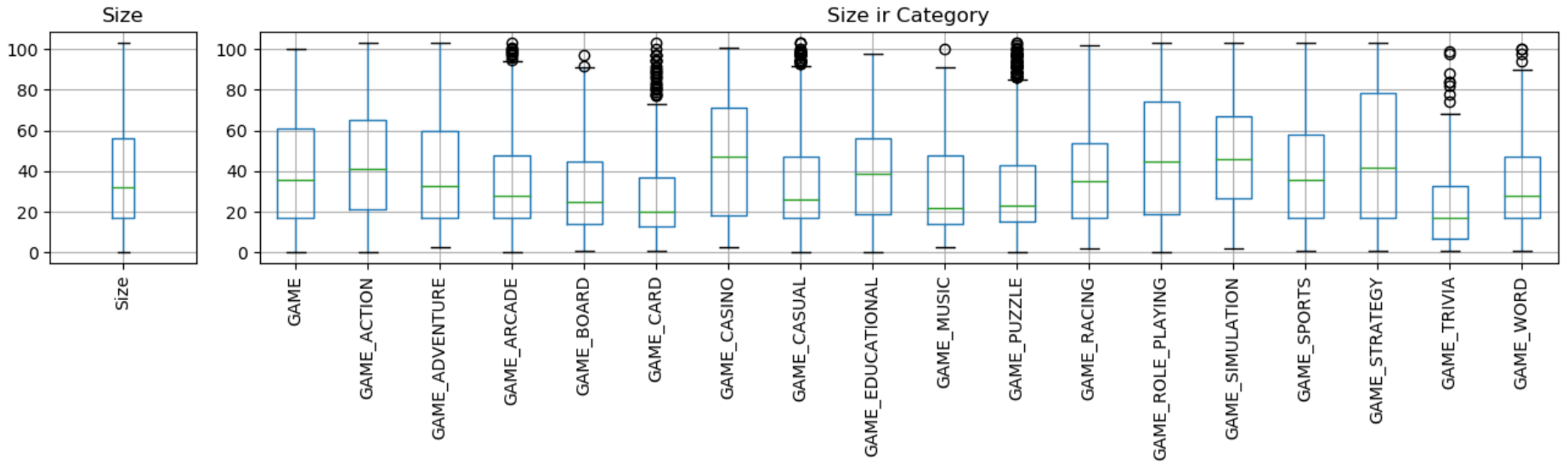
20 pav. „Reviews“ ir „Category“ box plot

Matoma, kad ryšio tarp „Reviews“ ir „Category“ nėra.

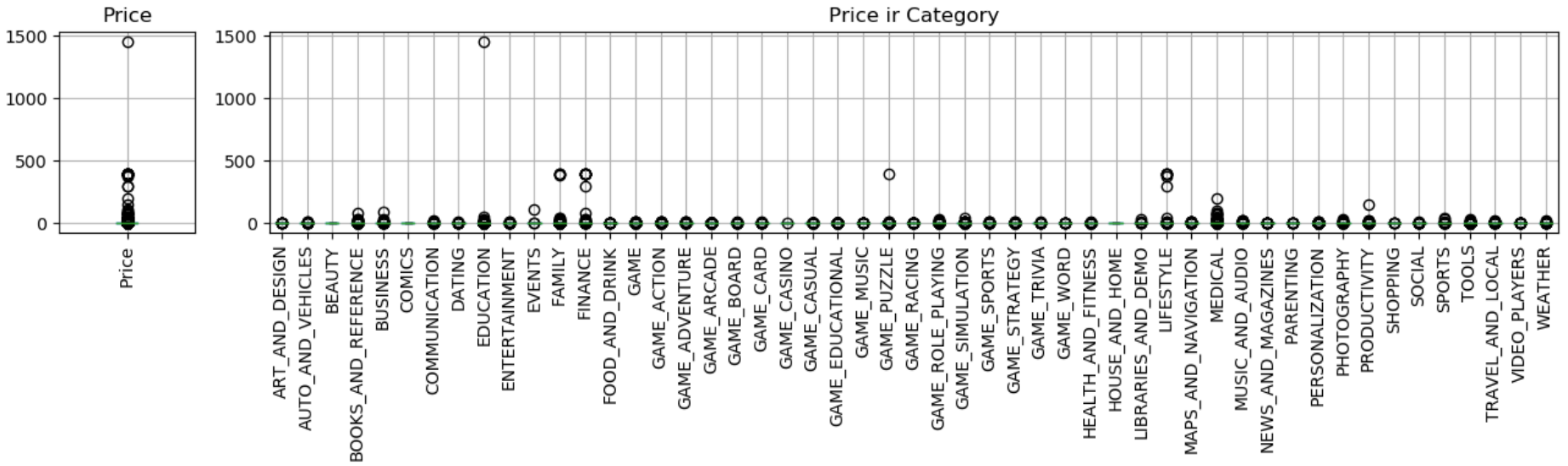


21 pav. „Size“ ir „Category“ box plot

Matoma labai nežymi sąsaja tarp dydžio ir kategorijos atributų. Žaidimų kategorijos programų dydis kiek didesnis, nei likusiųjų kategorijų.

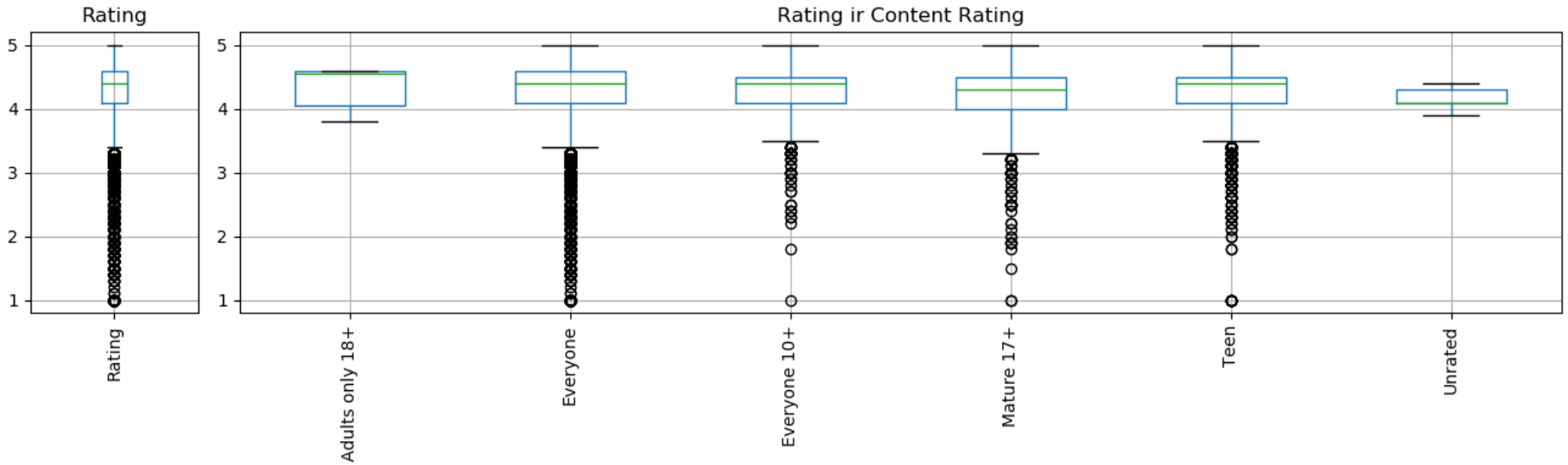


Ištirtas sąryšis tarp dydžio ir kategorijos atskirai žaidimų kategorijose, tikintis pastebėti skirtumą tarp paprastų, pavyzdžiui, kortų, arkados, žodžių ir pan., žaidimų ir sudėtingesnio įgyvendinimo reikalaujančių, pavyzdžiui, strateginių, nuotykių, lenktynių ir pan., žaidimų, tačiau toks sąryšis neaptiktas



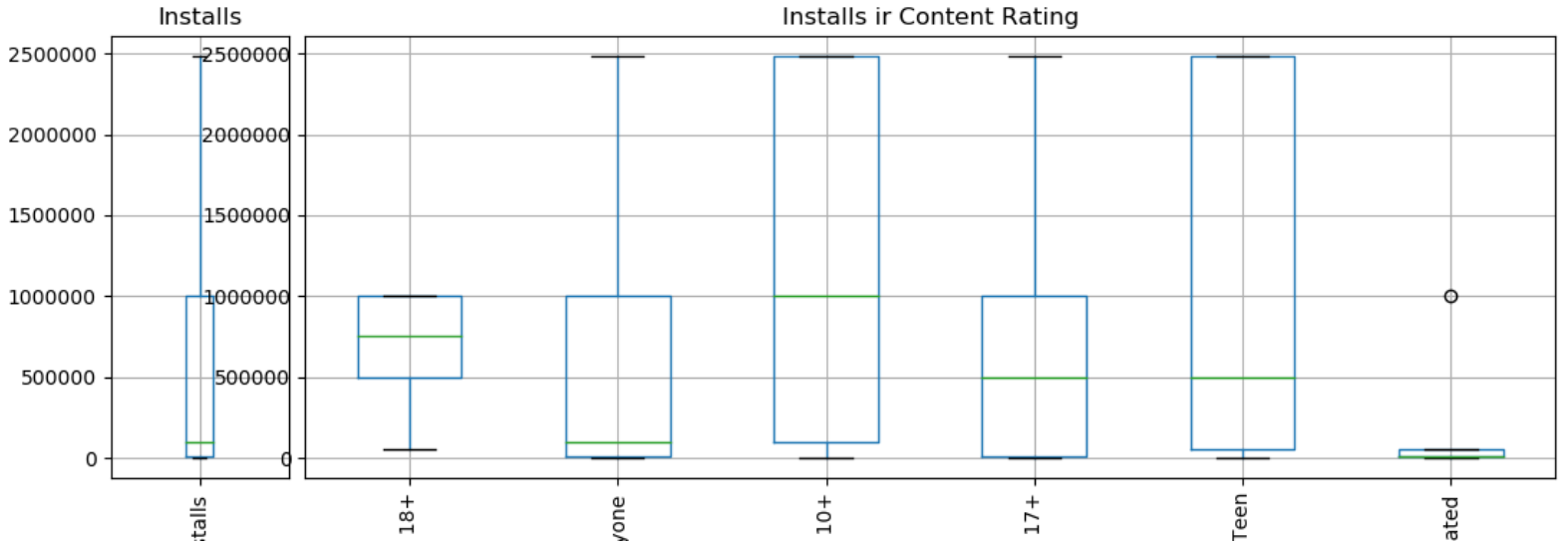
22 pav „Price“ ir „Category“ box plot

Tarp kainos ir kategorijos nematomas sąryšis – beveik visos programėlės, kad ir kokiai kategorijai priklausytų, yra nemokamos.



23 pav „Rating“ ir „Content Rating“ boc plot

Reitingų reikšmės, išskirstytos pagal amžiaus grupes, kurioms skirtos programos, išsidėsčiusios labai panašiuose intervaluose, nepriklausomai nuo kategorijos, taigi, sąryšio tarp šių dviejų atributų nėra.



Matoma, jog parsisiuntimų skaičiaus pasiskirstymas pagal amžiaus kategorijas, kurioms skirta programa, skirtingas kiekvienoje kategorijoje, taigi, galima teigti, jog sąryšis, nors ir ne stiprus, kadangi „dėžutės“ persidengia, yra.

5 Kovariacija ir koreliacija

KOVARIACIJOS

7 lentelė. Kovariacijos

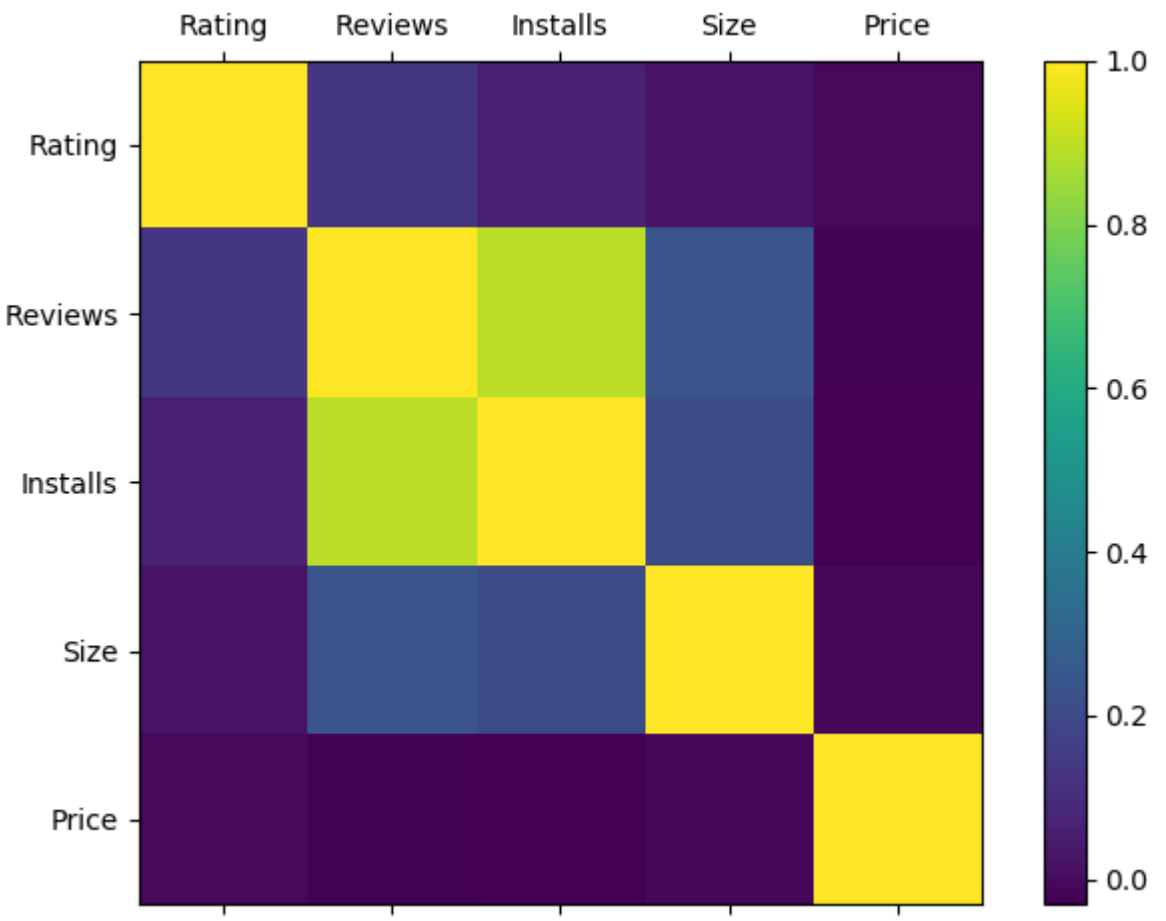
	Rating	Reviews	Installs	Size	Price
Rating	0.247	1259.402	29178.155	0.274	-0.023
Reviews	1259.402	360808350.67	15116000899.64	99093.411	-5278.173
Installs	29178.155	15116000899.64	791201680829.127	3997815.97	-297460.131
Size	0.274	99093.411	3997815.97	466.394	-1.553
Price	-0.023	-5278.173	-297460.131	-1.553	122.289

Kovariacija – ryšio tarp tolydinio atributų matas. Jos reikšmes sunku vertinti, kadangi reikšmė pateikiama tais pačiais vienetais kaip ir atributas. Atlikti koreliacijos – normalizuotos kovariacijos – skaičiavimai:

8 lentelė. Koreliacijos

	Rating	Reviews	Installs	Size	Price
Rating	1.0	0.13	0.07	0.03	-0.0
Reviews	0.13	1.0	0.89	0.24	-0.03
Installs	0.07	0.89	1.0	0.21	-0.03
Size	0.03	0.24	0.21	1.0	-0.01
Price	-0.0	-0.03	-0.03	-0.01	1.0

Pastebima, kad stipraus sąryšio, išskyrus „Reviews“ ir „Installs“ atributų ryšį, nėra. Rezultatai vizualizuoti koreliacijos matrica:



24 pav. Koreliacijos matrica

Koreliacijos matricoje spalvų reikšmės išdėstytos ne [-1;1], o [min(visos koreliacijos);max(visos koreliacijos)] intervale, kad aiškiau matytųsi sąryšiai tarp atributų. Matomas stiprus sąryšis tarp parsisiuntimų skaičiaus („Installs“) ir atsiliepimų skaičiaus („Reviews“); nestiprų programėlės dydžio („Size“) sąryšį su atsiliepimų skaičiumi ir parsisiuntimų skaičiumi. Matoma, kad kaina („Price“) beveik nekoreliuoja su kitais atributais. Taip gali būti todėl, kad didesnės nei 0 kainos reikšmės yra retos ir nemokamų programų skaičius nustelbia kainos koreliaciją su kitais parametrais.

6 Kategorinio tipo atributų skaitmeninimas

Kategorinio tipo atributai „Category“, „Content Rating“ suskaitmeninami tokiu būdu: išskiriamos unikalios reikšmės, jos sunumeruojamos, o tuomet šie numeriai priskiriami vietoje buvusių tekstinių reikšmių. Kategorinio tipo atributų skaitmeninimas reikalingas paprastam jų apdorojimui, skaičiavimams, prognozavimui.

7 Duomenų normalizacija

Duomenų normalizacija atlikta tolydinio tipo bei skaitmenintiems kategorinio tipo atributams, pasirinkus standartizavimo metodą:

$$standartizuota\ reikšmė = \frac{reikšmė - vidurkis}{standartinis\ nuokrypis}$$

Pasirinkta standartizacija, nes įprastinė normalizacija (vertinant atskiras reikšmės su atributo reikšmių intervalu (nuo min. iki maks. reikšmės) yra jautri ekstremalioms reikšmėms, kurios išsaugotos kai kuriiems atributams dėl jų galimos įtakos parsisiuntimų skaičiui, pavyzdžiui, kainos ir parsisiuntimų skaičiaus koreliacija, tikėtina, gali būti neigiama (antrame laboratoriniame darbe ketinama susintetinti didesnės nei 0 kainos reikšmių).

8 Išvados

1. Atlikta paisirinkto duomenų rinkinio kokybės analizė. Pastebėta, kad duomenų rinkinyje buvo daug klaidų: trūkstamų, ekstremalių reikšmių; reikšmių, neatitinkančių numatyto formato.
2. Atliktas duomenų apdorojimas. Pasirinktos įvairios duomenų rinkinio problemų sprendimo metodikos. Sprendžiant trūkstamų reikšmių problemą pirmiausia pasirinktas vidurkio, medianos įstatymo metodas, tačiau pastebėta, jog tai iškraipo duomenų rinkinį. Galiausiai išvystytas sprendimas įterpia atsitiktinę reikšmę iš visų jau egzistuojančių atributo reikšmių su jų pasirodymo atributo reikšmių rinkinyje tikimybe, tokiu būdu išlaikytas neiškraipytas atributų reikšmių pasiskirstymas.
3. Atliktas duomenų rinkinio atributų sąryšių tyrimas: tolydinio tipo atributų sąryšiai vertinti pagal koreliacijos reikšmes, kategorinio tipo – pagal stulpelines diagramas, kategorinio ir tolydinio tipo – pagal „box plot“ diagramas. Sąryšiai ypatingai ieškoti tarp parsisiuntimų skaičiaus ir kitų atributų, kadangi planuota parsisiuntimų skaičių pasirinkti kaip tikslo reikšmę antrame laboratoriniame darbe. Pastebėta, jog nėra stipraus sąryšio tarp tolydinio tipo atributų, išskyrus parsisiuntimų ir atsiliepimų skaičių. Kategorinio tipo atributų sąryšis pastebėtas tik tarp amžiaus kategorijų, kurioms skirtos programos, ir programos kategorijos. Kategorinio ir tolydinio tipo atributų nežymūs sąryšiai matomi tarp parsisiuntimų skaičiaus bei amžiaus kategorijos, kuriai skirta programa ir tarp parsiuntimų skaičiaus bei programos kategorijos.