KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS

Intelektikos pagrindai (P176B101)

Laboratorinis darbas Nr. 1

Duomenų apdorojimas ir analizė

Atliko:

IFF-8/12 gr. studentas

Jokūbas Akramas

2021m. kovo 3 d.

Priėmė:

doc. Agnė Paulauskaitė - Tarasevičienė

lekt. Germanas Budnikas

KAUNAS 2021

TURINYS

1. Duomenų rinkinys

Pasirinktas duomenų rinkinys vaizduoja įvairius prisijungimų prie serverio parametrus, o išvesties kintamasis – *boolean* požymį, ar serveris buvo atakuotas, ar ne. Pavyzdinė duomenų rinkinio lentelė:

1 lentelė. Duomenų rinkinio ištrauka

RE-GION	OS	PRO-TOCOL	X_1	X_2	X_3	X_4	X_5	X_6	X_7	MALICIOUS_O-FFENSE
	Linux	HTTP	36	3000	5	6	29	174	92	0
	Win-dows	HTTP	37	0	11	17	142	236	103	1
	Linux		3	3	1	0	93	174	110	1
	Linux		33	2	7	1	29	249	72	1
	Linux		33	2	8	3	29	174	112	1
	Linux		45	10	1	0	62	303	72	1
	Linux		30	7	7	1	29	174	112	1
			8	7	9	8	62	316	72	1
			49	6	8	3	14	316	103	1
			4	6	15	10	29	145	103	0
			4	6	15	10	130	174	103	1
			33	2	5	6	93	174	72	1
			4	6	5	6	29	174	42	1
			4	6	15	10	133	174	18	1
			7	7	2	7	62	316	72	1
			36	2	4		29	249	92	1
			25	9	4	2	29	0	72	1
			36	2	1	0	142	0	92	1
			37	0	11	17	115	174	103	1
		HTTP	3	3	1	0	29	249	2	1
		HTTP	21	4	5		53	174	18	1
		HTTP	36	2	1	0	93	174	92	1
		HTTP	47	7	4		62	316	98	1
		HTTP	22	7	13	18	29	316	92	1
		HTTP	49	6	5	6	29	249	67	1
		HTTP	4	6	3	5	80	174	72	1
Asia		HTTP	4	6	3	5	142	0	103	1

Duomenų rinkinys turi 1000 įrašų (500 iš kurių išvesties kintamasis lygus 1, o 500 likusių – 0). Duomenų rinkinyje taip pat yra trūkstamosios reikšmės (*ang. Missing values*) ir triukšmai (*ang. Outliers*), kurie atitinkamai apdoroti programinio kodo algoritmo ir bus paminėti sekančiuose skyreliuose.
Duomenų rinkinį sudaro šie atributai:

2 lentelė. Duomenų rinkinio atributai

Atributas	Tipas	Prasmė	Pavyzdys
REGION	Kategorinis	Pasaulio regionas, iš kurio jungtasi prie serverio	Asia
OS	Kategorinis	Operacinė sistema, iš kurios jungtasi prie serverio	Linux
PROTOCOL	Kategorinis	Komunikavimo protokolas tarp kliento ir serverio	SSH
X_1	Tolydinis	Prisijungimo parametras	6
X_2	Tolydinis	Prisijungimo parametras	2
X_3	Tolydinis	Prisijungimo parametras	8
X_4	Tolydinis	Prisijungimo parametras	18
X_5	Tolydinis	Prisijungimo parametras	115

X_6	Tolydinis	Prisijungimo parametras	236
X_7	Tolydinis	Prisijungimo parametras	72
MALICIOUS_OFFENSE	Kategorinis	Požymis, ar serveris buvo atakuojamas	1

2. Duomenų rinkinio kokybės analizė

2.1. Tolydinio tipo atributai

Kodo sugeneruotas dokumentas *output/tolydinių_duomenų_analizė.csv*

3 lentelė Tolydinių atributų analizė

Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Minimali reikšmė	Maksimali reikšmė	1-asis kvartilis	3-asis kvartilis	Vidurkis	Mediana	Standartinis nuokrypis
X_1	1000	1.6	47	2	52	6	37	25.38114	25	15.58241
X_2	1000	2.7	10	0	12	2	6	4.432297	4	2.937442
X_3	1000	2.3	14	1	14.5	2	7	5.831494	5	4.120817
X_4	1000	12.1	15	0	14.5	2	7	4.914744	5	3.581023
X_6	1000	1.8	68	61.5	330	174	249	188.1489	174	76.46123
X_7	1000	1.8	35	25.5	116	72	103	85.88917	98	25.5432

Šioje duomenų rinkinio analizėje nebėra *X_5* tolydinio atributo. Taip yra todėl, nes pastarojo atributo trūkstamosios reikšmės viršijo leistiną limitą (60 %) ir jis buvo automatiškai pašalintas ir nebenagrinėjamas. Plačiau apie tai problemų sprendimo plano skyrelyje.

2.2. Kategorinio tipo atributai

Kodo sugeneruotas dokumentas *output/kategorinių_duomenų_analizė.csv*

4 lentelė Kategorinių duomenų analizė

Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Moda	Modos dažnumas	Moda, %	2-oji moda	2-osios Modos dažnumas	2-oji Moda, %
REGION	1000	6.3	4	Asia	361	36.20862588	Africa	278	27.88365095
OS	1000	7.8	3	Linux	826	82.84854564	Windows	130	13.03911735
PROTOCOL	1000	6.3	4	HTTP	964	96.69007021	FTP	21	2.106318957

2.3. Duomenų kokybės problemos

Pagrindinė problema, su kuria susiduriama analizuojant didelius duomenų rinkinius – trūkstamosios reikšmės. Šiai problemai spręsti programa atlieka keletą etapų:

- Nuskaičius duomenis ir dalinai juos paanalizavus jie nukreipiami metodui *handle_missing_values()*. Šis metodas toliau kviečia dviejų tipų duomenų šalinimo metodus.
- Pirmasis metodas – horizontaliojo šalinimo (*horizontal_removal()*). Skenuojama kiekviena eilutė ir jeigu daugiau nei 60% (kriterijų galima keisti laisva nuožiūra) eilutės atributų yra tušti, eilutė šalinama.

- Antrasis metodas – vertikalaus šalinimo (vertical_removal()). Skenuojamas kiekvienas stulpelis ir jeigu daugiau nei 60% (kriterijų galima keisti laisva nuožiūra) eilučių yra tuščios, stulpelis (atributas) šalinamas.
- Apdorojus tuščias reikšmes masiniams stulpeliams/eilutėms, likusios tuščios reikšmės užpildomos atitinkamai:
    - Tolydinės – atributo vidurkiu;
    - Kategorinės – atributo moda;

Ištrauka iš programos vykdymo konsolės:

```
---------------------------------------Horizontalus šalinimas--------------------
------------------
Eilutės:  {'REGION': 'Africa', 'OS': '', 'PROTOCOL': '', 'X_1': '', 'X_2': '',
'X_3': '', 'X_4': '', 'X_5': '', 'X_6': '', 'X_7': '', 'MALICIOUS_OFFENSE': '1'}
tuščiosios reikšmės viršija nustatytą limitą ( 60.0 %), todėl eilutė PAŠALINAMA.
Eilutės:  {'REGION': '', 'OS': '', 'PROTOCOL': '', 'X_1': '', 'X_2': '', 'X_3':
'5', 'X_4': '6', 'X_5': '', 'X_6': '', 'X_7': '103', 'MALICIOUS_OFFENSE': '1'}
tuščiosios reikšmės viršija nustatytą limitą ( 60.0 %), todėl eilutė PAŠALINAMA.
Eilutės:  {'REGION': '', 'OS': '', 'PROTOCOL': '', 'X_1': '', 'X_2': '', 'X_3':
'', 'X_4': '', 'X_5': '', 'X_6': '316', 'X_7': '112', 'MALICIOUS_OFFENSE': '1'}
tuščiosios reikšmės viršija nustatytą limitą ( 60.0 %), todėl eilutė PAŠALINAMA.
---------------------------------------Vertikalus šalinimas----------------------
----------------
Tolydinio atributo:  X_5 tuščiosios reikšmės viršija nustatytą limitą ( 60.0 %),
todėl atributas PAŠALINAMAS.
```

Kita problema – triukšmai, ekstremalios reikšmės. Ši problema galioja tik tolydiniams atributams, todėl šios problemos apdorojimas (*handle_noise()*) bus kviečiamas tolydinių atributų analizės metu (*analyze_continuous_data()*). Sprendimo būtas – kiekvienam atributui apskaičiuoti 1 ir 3 kvartilius ir iš jų išvesti formulę, ekstremalių reikšmių mažinimui:

```
lower = q1 - 1.5 * (q3 - q1)
upper = q3 + 1.5 * (q3 - q1)
```

Kiekvienam atributui, esančiam virš viršutinio rėžio bus priskirtas viršutinis rėžis, analogiškai apatinis – atributams, esančiams žemiau apatinio rėžio:

```
for u in data:  # Triukšmai priskiriami viršutiniams ir apatiniams rėžiams duomenų
faile
    if int(u[header]) < lower:
        u[header] = lower
    elif int(u[header]) > upper:
        u[header] = upper
```

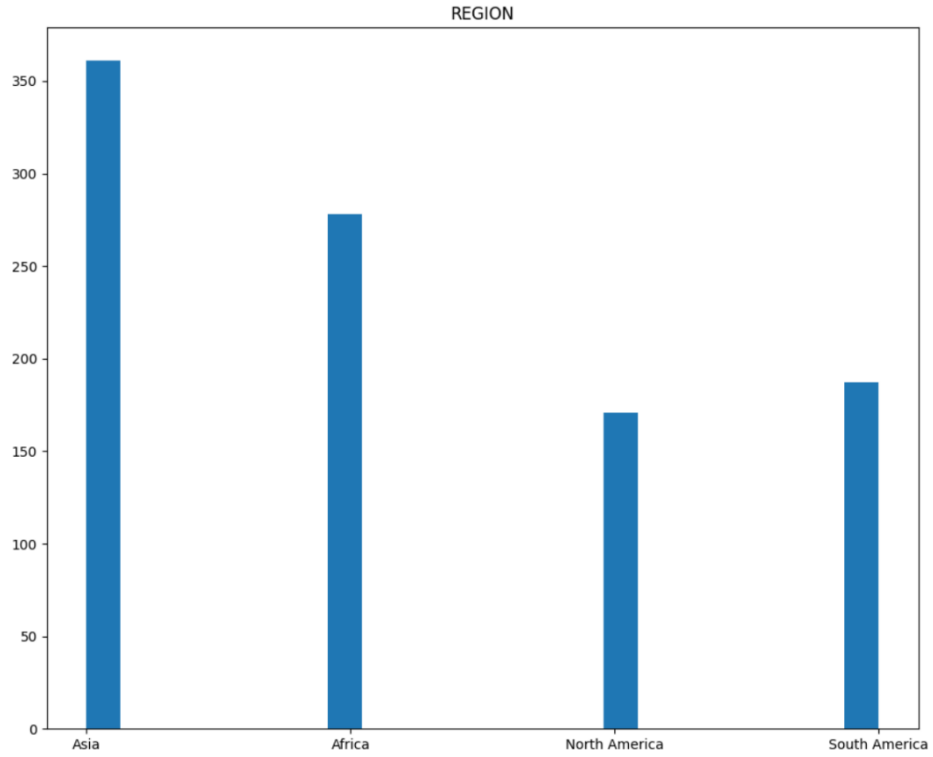
Apdorojus duomenis, šie išvedami faile *output/apdoroti_duomenys.csv*

REGION	OS	PROTO-COL	X_1	X_2	X_3	X_4	X_6	X_7	MALICIOUS_OF-FENSE
Asia	Linux	HTTP	36	12	5	6	174	92	0
Asia	Win-dows	HTTP	37	0	11	14.5	236	103	1
Asia	Linux	HTTP	3	3	1	0	174	110	1
Asia	Linux	HTTP	33	2	7	1	249	72	1
Asia	Linux	HTTP	33	2	8	3	174	112	1
Asia	Linux	HTTP	45	10	1	0	303	72	1
Asia	Linux	HTTP	30	7	7	1	174	112	1
Asia	Linux	HTTP	8	7	9	8	316	72	1
Asia	Linux	HTTP	49	6	8	3	316	103	1
Asia	Linux	HTTP	4	6	14.5	10	145	103	0
Asia	Linux	HTTP	4	6	14.5	10	174	103	1
Asia	Linux	HTTP	33	2	5	6	174	72	1
Asia	Linux	HTTP	4	6	5	6	174	42	1
Asia	Linux	HTTP	4	6	14.5	10	174	25.5	1
Asia	Linux	HTTP	7	7	2	7	316	72	1
Asia	Linux	HTTP	36	2	4	5	249	92	1
Asia	Linux	HTTP	25	9	4	2	61.5	72	1
Asia	Linux	HTTP	36	2	1	0	61.5	92	1
Asia	Linux	HTTP	37	0	11	14.5	174	103	1
Asia	Linux	HTTP	3	3	1	0	249	25.5	1
Asia	Linux	HTTP	21	4	5	5	174	25.5	1
Asia	Linux	HTTP	36	2	1	0	174	92	1
Asia	Linux	HTTP	47	7	4	5	316	98	1
Asia	Linux	HTTP	22	7	13	14.5	316	92	1
Asia	Linux	HTTP	49	6	5	6	249	67	1
Asia	Linux	HTTP	4	6	3	5	174	72	1
Asia	Linux	HTTP	4	6	3	5	61.5	103	1
Asia	Linux	HTTP	15	3	4	2	174	116	1
Africa	Linux	HTTP	36	2	1	5	316	92	1
Africa	Linux	HTTP	21	4	4	2	174	111	1
North America	Linux	HTTP	37	0	11	5	127	103	1
Asia	Linux	HTTP	49	6	2	7	249	72	1

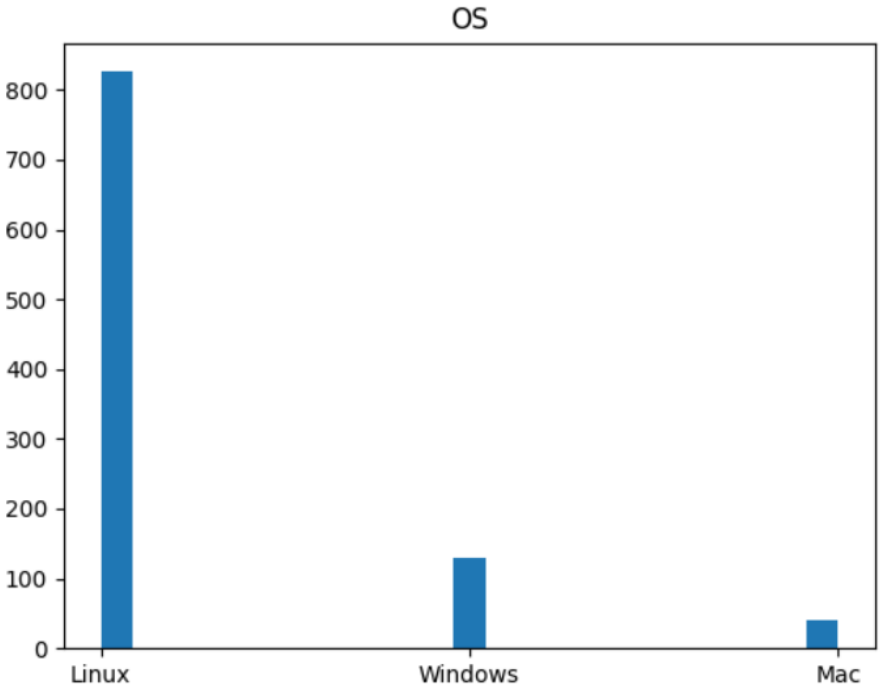
3. Duomenų vaizdavimas

3.1. Atributų histogramos

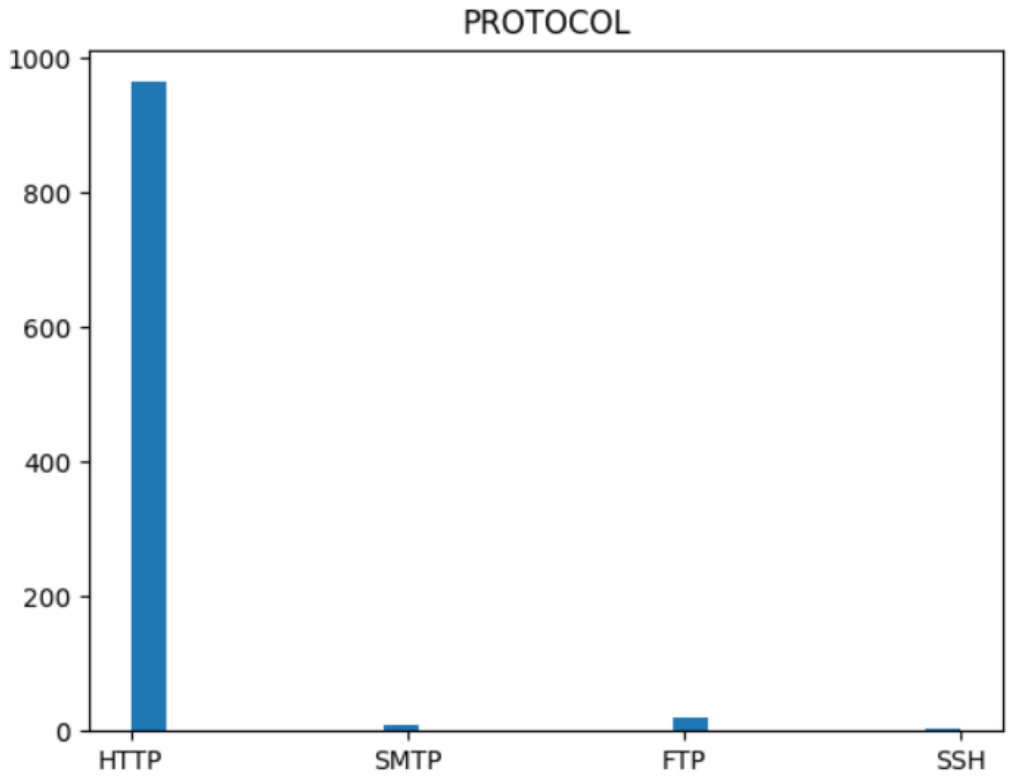
Kiekvienam atributui stulpelių skaičius paskaičiuojamas pagal formulę: $1 + 3.22 \cdot \log_e n$, kur n – imties dydis.



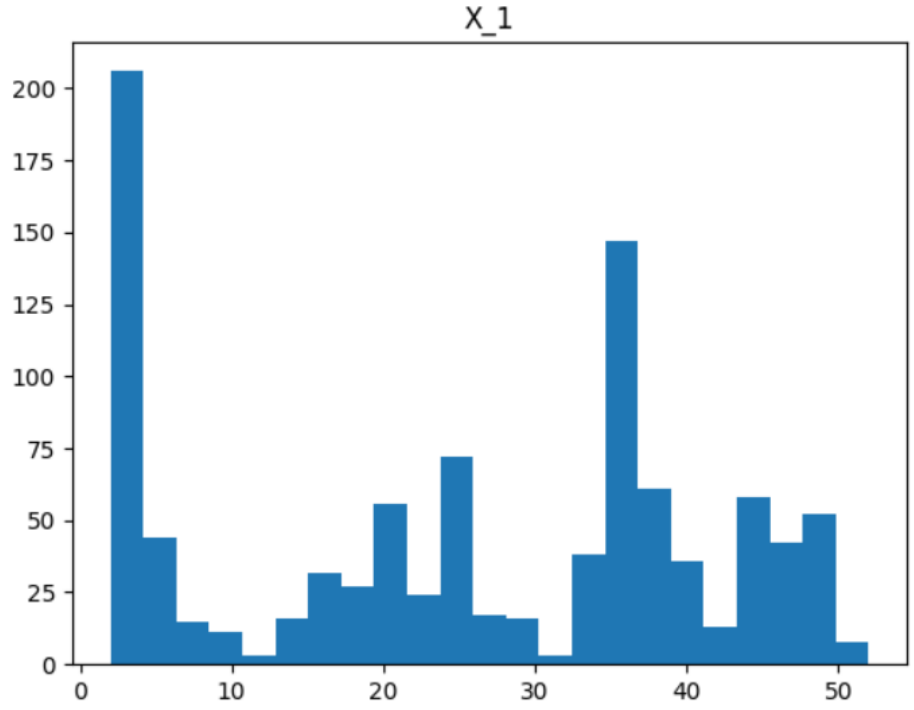
Atributo „REGION“ pasiskirstymas atvirkštinis eksponentinis (sukeitus NA su SA). Galima daryti prielaidą, kad daugiausia prisijungimų prie serverio padaryta iš 3-ųjų pasaulio šalių. Bandymus įsilaužti siekiant ekonominės naudos pateisintų nestabili ekonominė situacija šalyje.



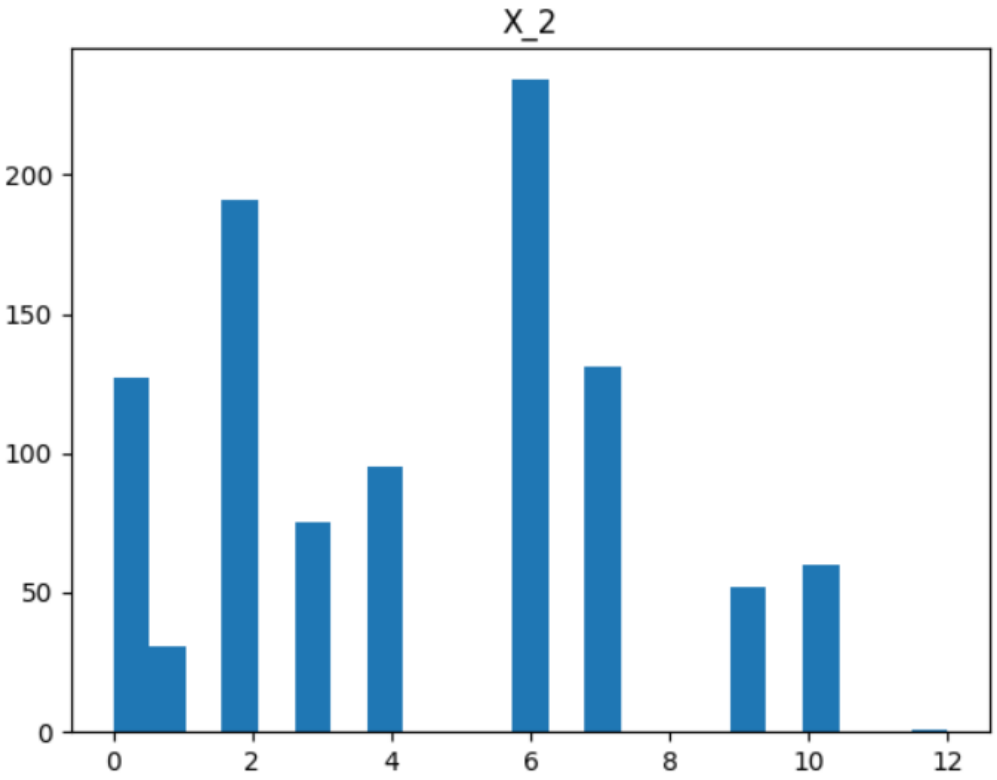
Atributo „OS“ pasiskirstymas atvirkštinis eksponentinis. Daugiausia prisijungimų padaryta iš Linux operacinės sistemos, kuri turi gerus įrankius serverių sukompromitavimui.



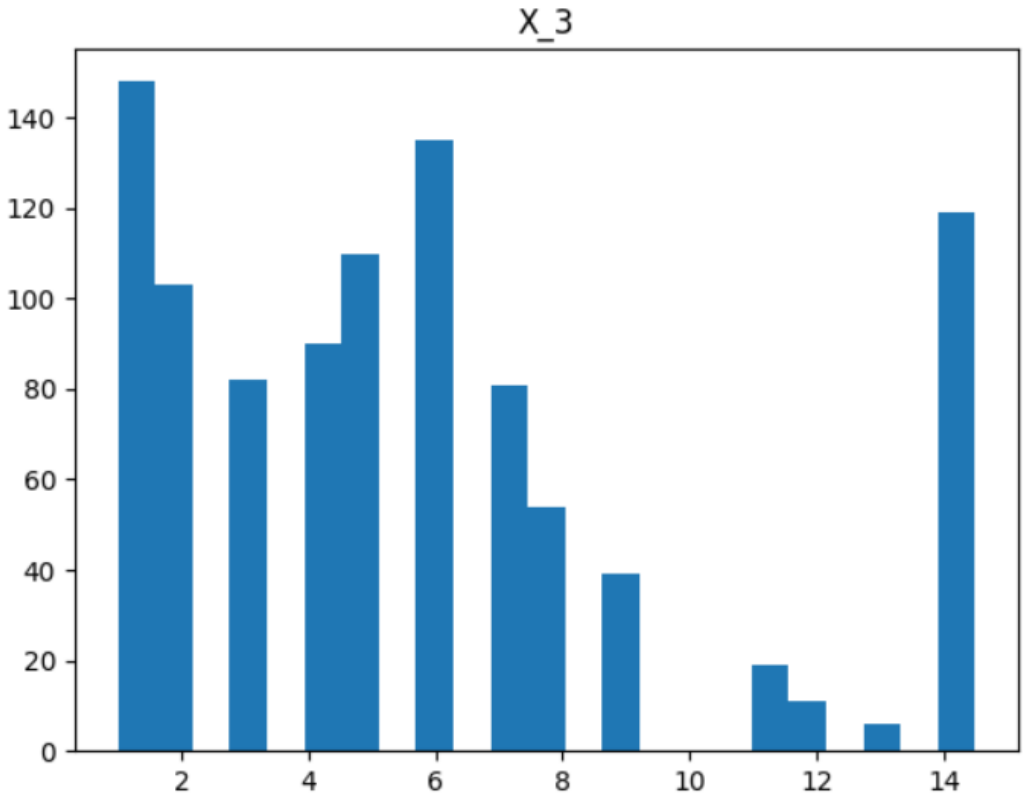
Atributo „PROTOCOL“ pasiskirstymas netolygus, nes daugiausia prisijungimų padaryta per http protokolą, kuris naudojamas svetainių atvaizdavimui.



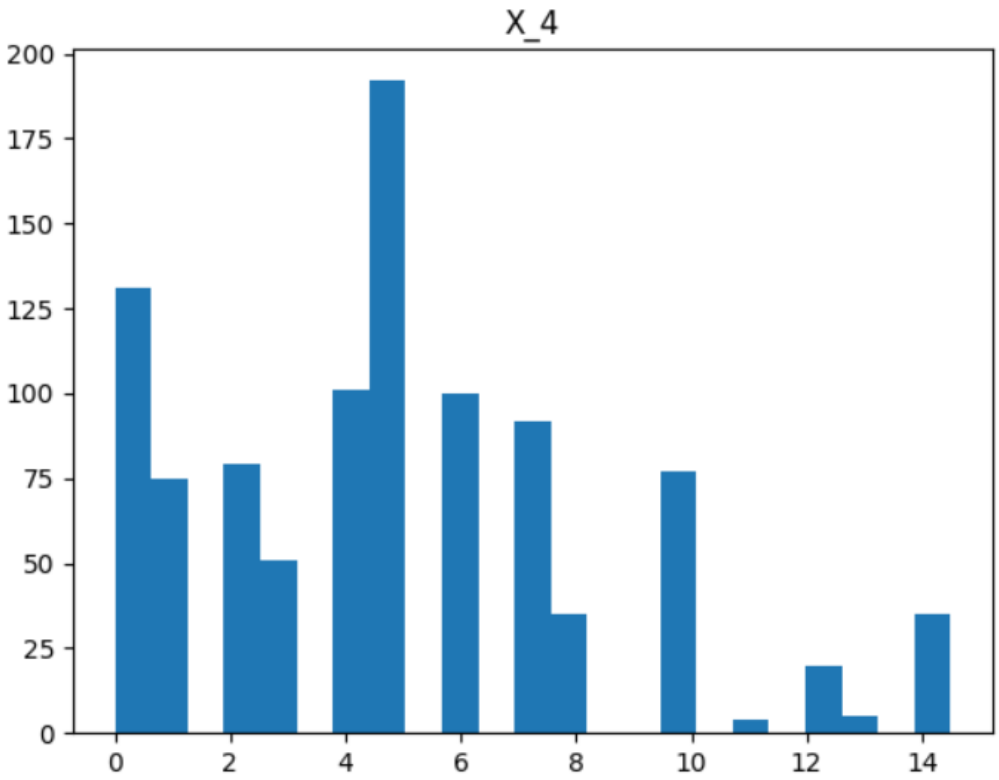
Atributo „X_1“ pasiskirstymas netolyginis.



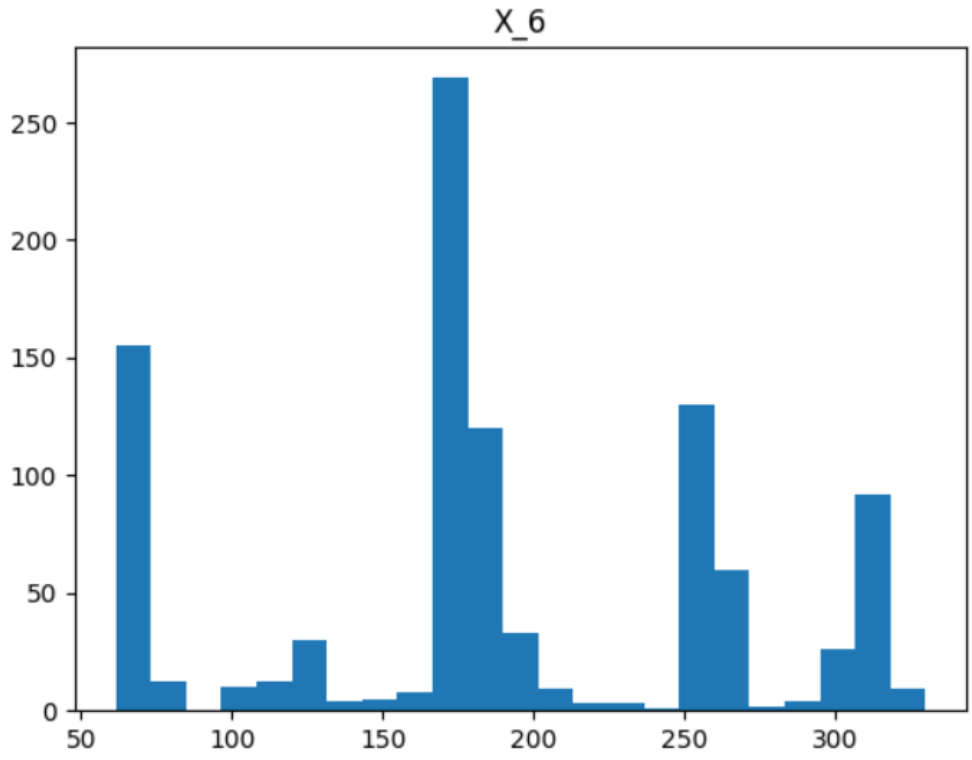
Atributo „X_2“ pasiskirstymas netolyginis.



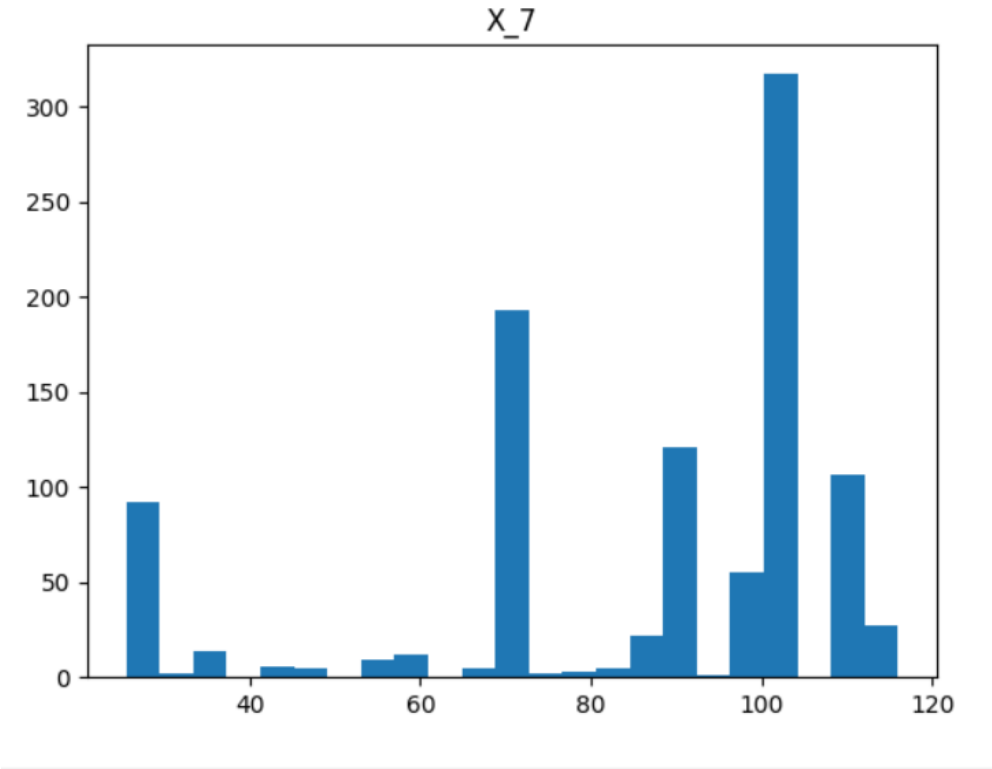
Atributo „X_3“ pasiskirstymas netolyginis.



Atributo „X_4“ pasiskirstymas netolyginis.

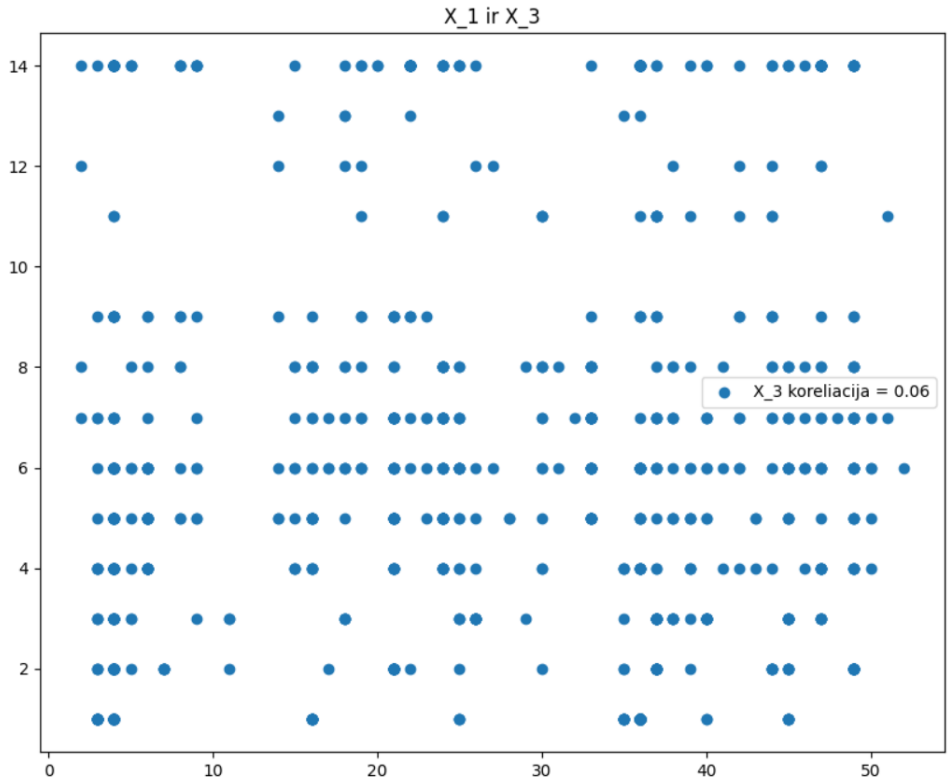


Atributo „X_6“ pasiskirstymas netolyginis
(nors sumažinus mastelį galima įžvelgti Gauso skirstinį)

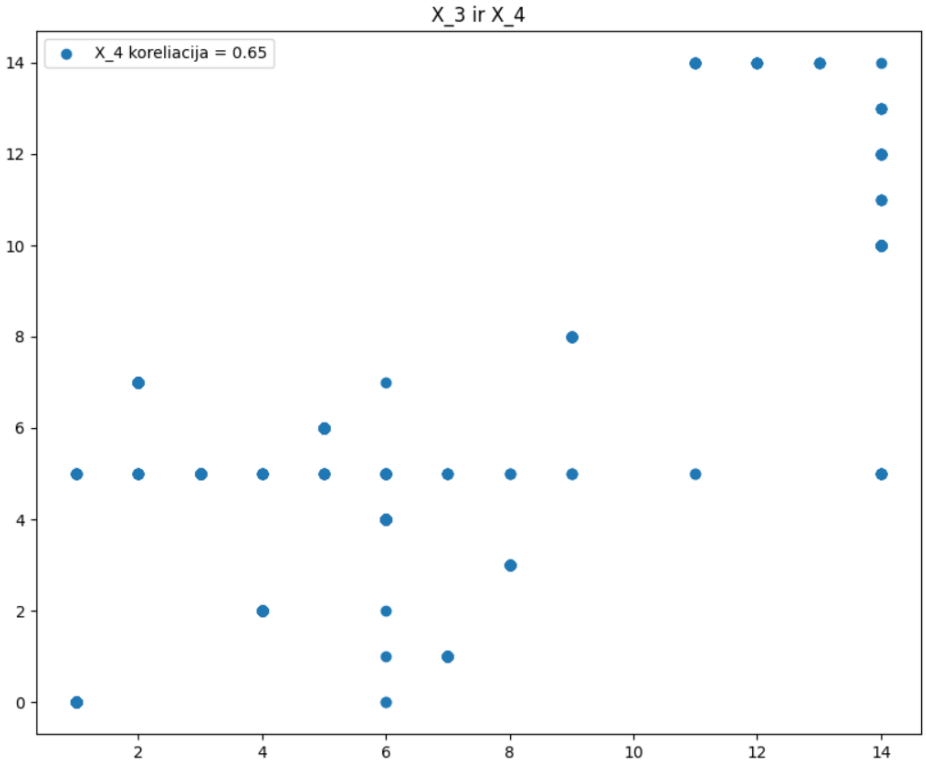


Atributo „X_7“ pasiskirstymas netolyginis.

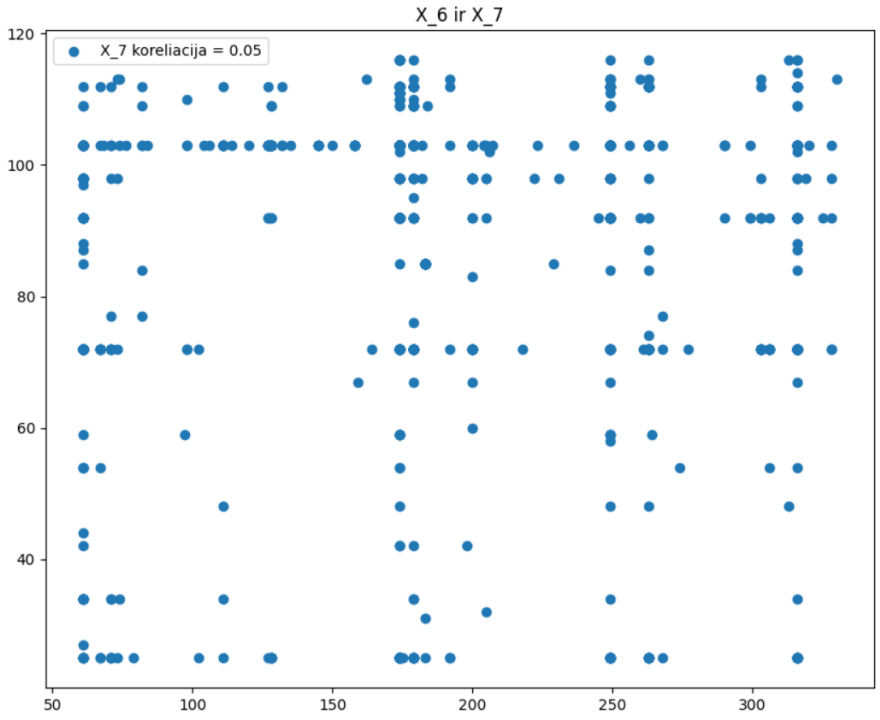
3.2. Tolydinių atributų priklausomybės



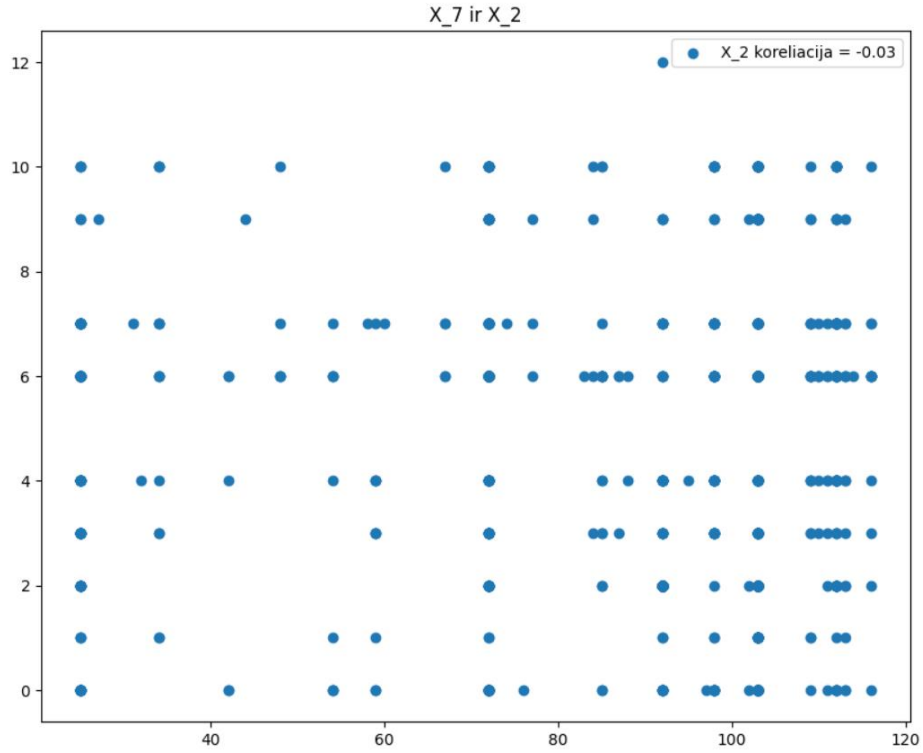
Priklausomybė tarp atributų „X_1“ ir „X_3“ yra labai silpna teigiama , koreliacijos koeficientas = 0.06



Priklausomybė tarp atributų „X_3“ ir „X_4“ stipri (stipriausia tarp visų atributų), koreliacijos koeficientas = 0.65

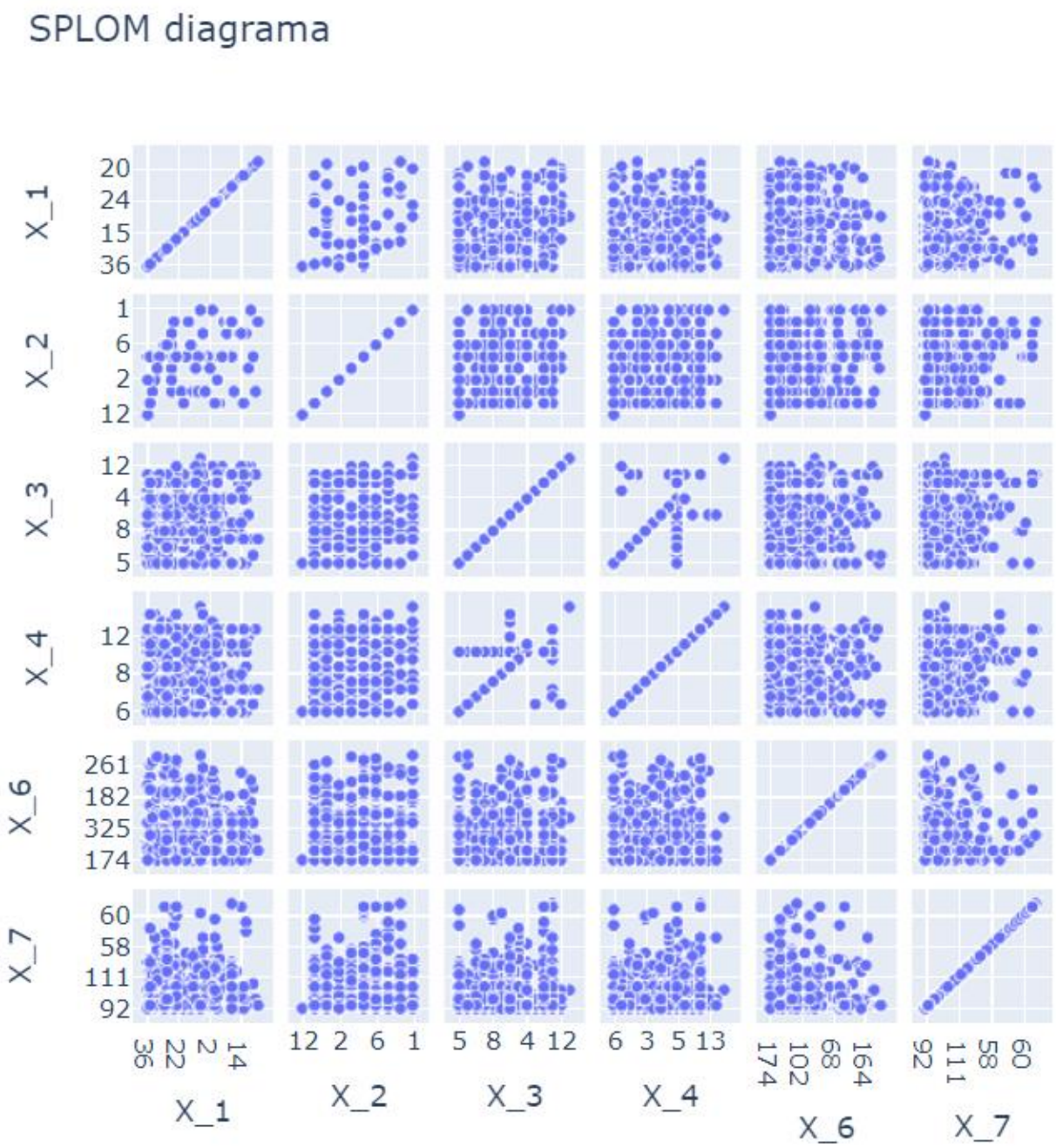


Priklausomybė tarp atributų „X_6“ ir „X_7“ yra labai silpna teigiama, koreliacijos koeficientas = 0.05



Priklausomybė tarp atributų „X_7“ ir „X_2“ yra labai silpna neigiama, koreliacijos koeficientas = -0.03

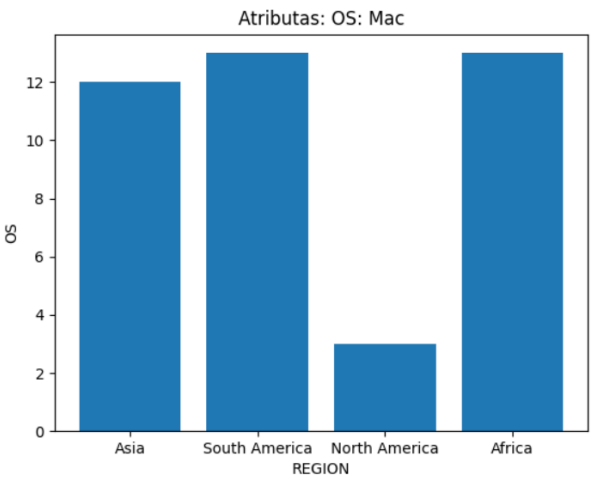
3.3. SPLOM diagrama

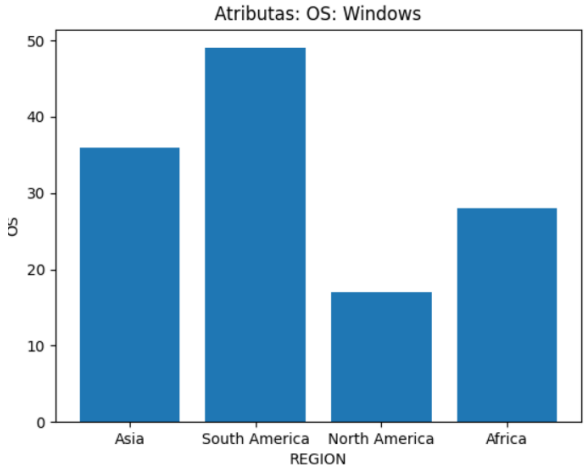


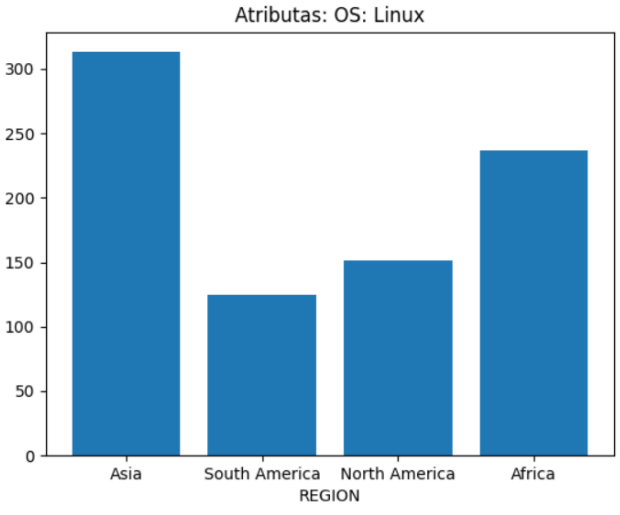
Iš pateiktos SPLOM diagramos matosi, kad visi atributai yra mažai priklausomi, išskyrus „X_3“ ir „X_4“, kurių koreliacijos koeficientas lygus 0.65.

3.4. Kategorinių atributų priklausomybės

Atributo “REGION” priklausomybė nuo “OS”



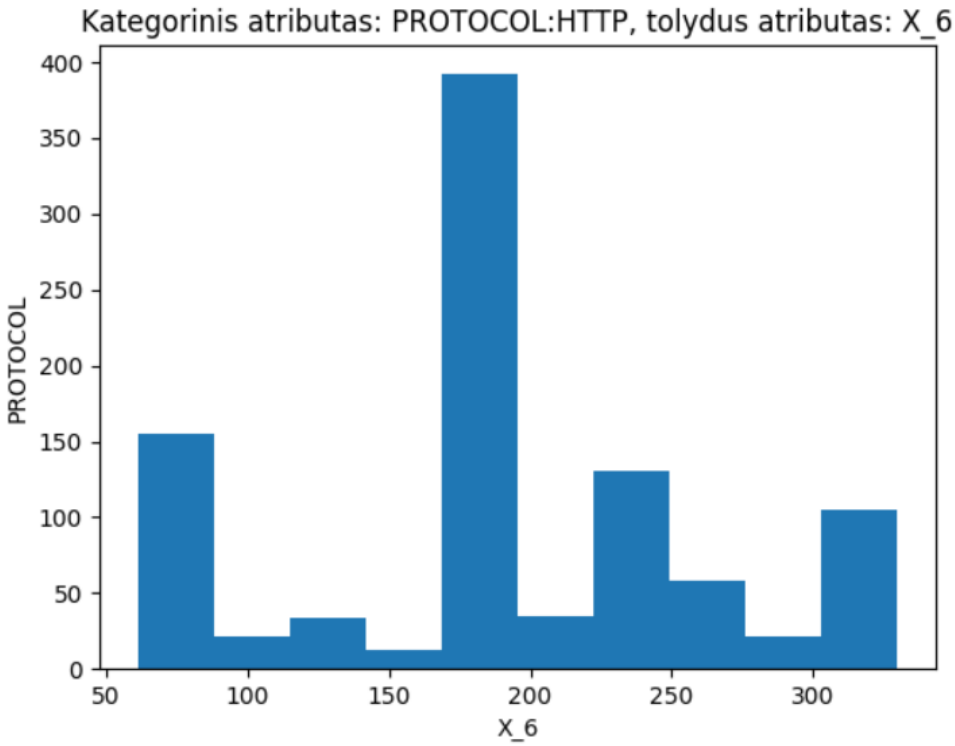




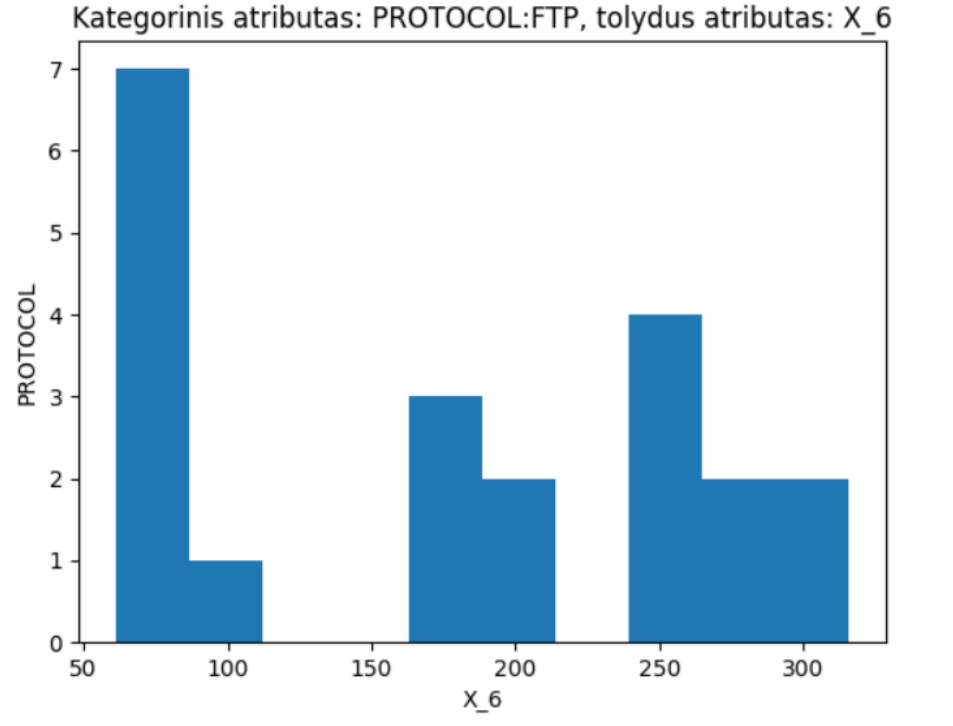
Matomas skirtingas operacinių sistemų naudojimas pagal kiekvieną regioną – pavyzdžiui „Mac“ populiari Pietų Amerikos ir Afrikos regionuose, „Windows“ – Pietų Amerikos ir Azijos, o „Linux“ – stipriai pirmauja Azijoje, palyginti su kitomis operacinėmis sistemomis.

3.5. Kategorinių ir tolydinių atributų priklausomybės

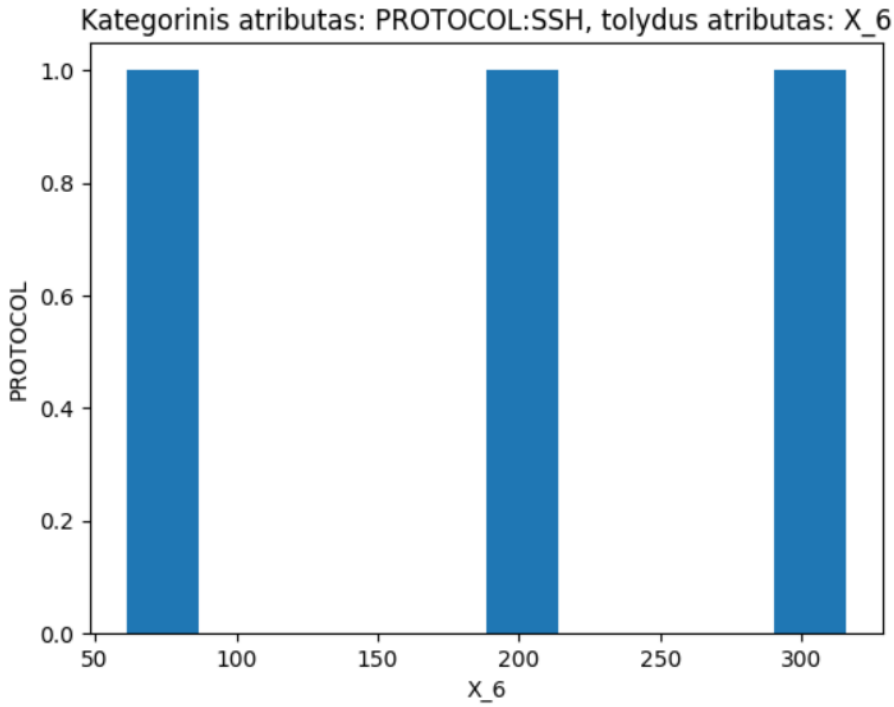
Atributo „X_6“ priklausomybė nuo „PROTOCOL“



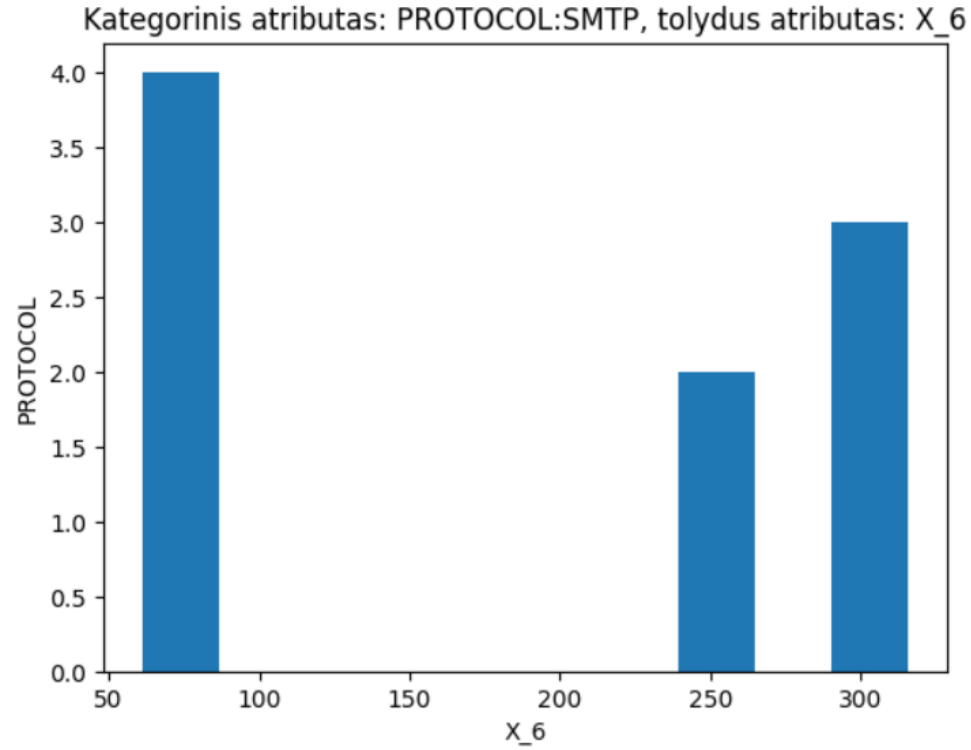
Diagramos skirstinys pasiskirstęs netolygiai (nors sumažinus mastelį galima įžvelgti normalųjį Gauso pasiskirstymą).



Diagramos skirstinys pasiskirstęs netolygiai. Taip yra todėl, nes atributas „PROTOCOL“ įrašą „FTP“ turi labai ribotą kiekį kartų (iš viso 21), todėl pastebėti tendencijas sunku.

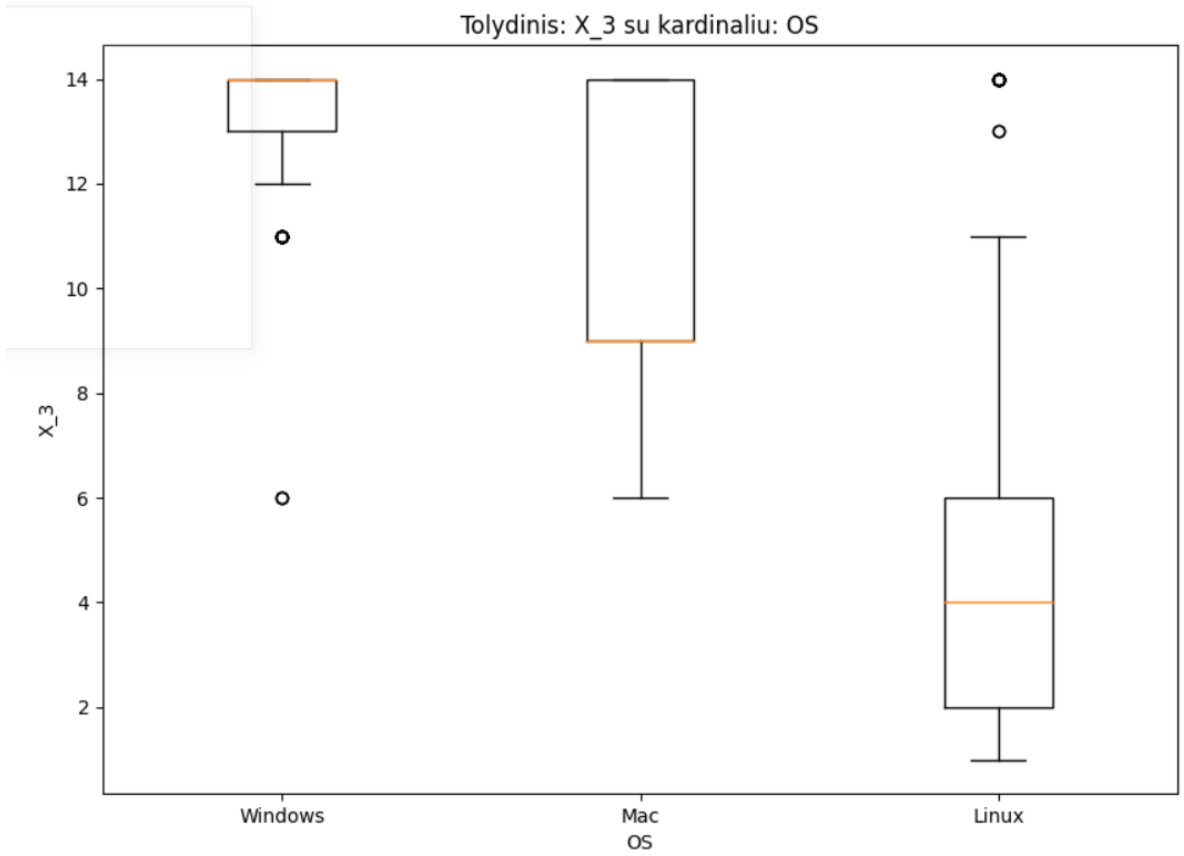


Diagramos skirstinys pasiskirstęs netolygiai. Taip yra todėl, nes atributas „PROTOCOL“ įrašą „SSH“ turi labai ribotą kiekį kartų (iš viso 3), todėl pastebėti tendencijas sunku.



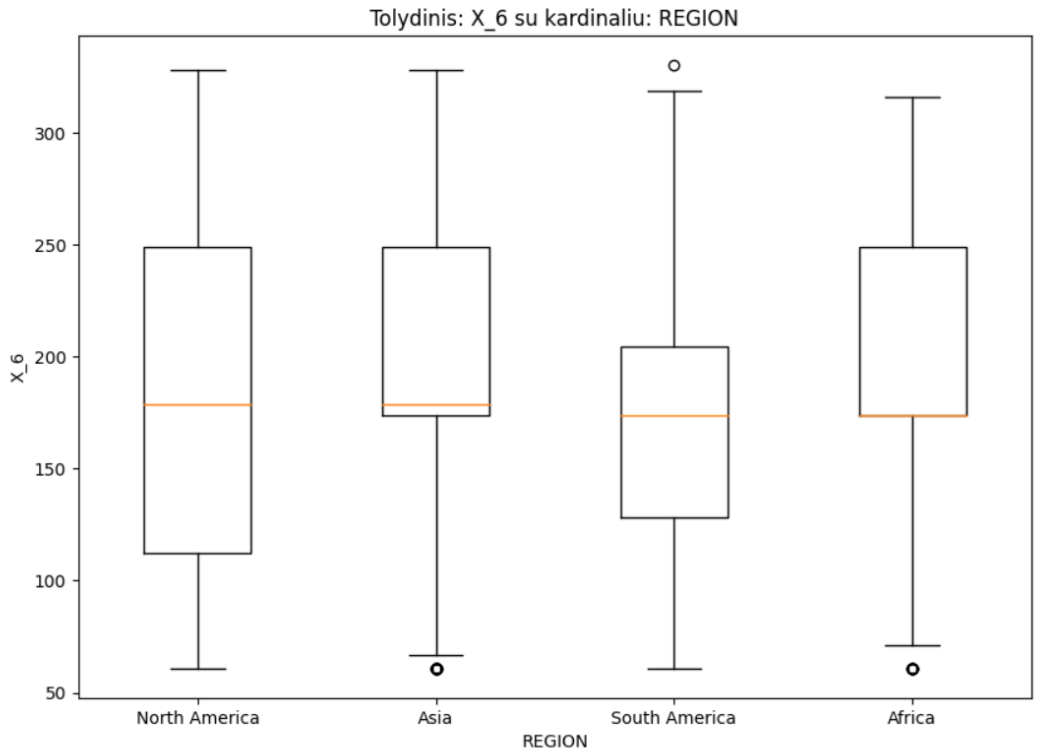
Diagramos skirstinys pasiskirstęs netolygiai. Taip yra todėl, nes atributas „PROTOCOL“ įrašą „SMTP“ turi labai ribotą kiekį kartų (iš viso 9), todėl pastebėti tendencijas sunku.

Atributo „OS“ priklausomybė nuo „X_3“



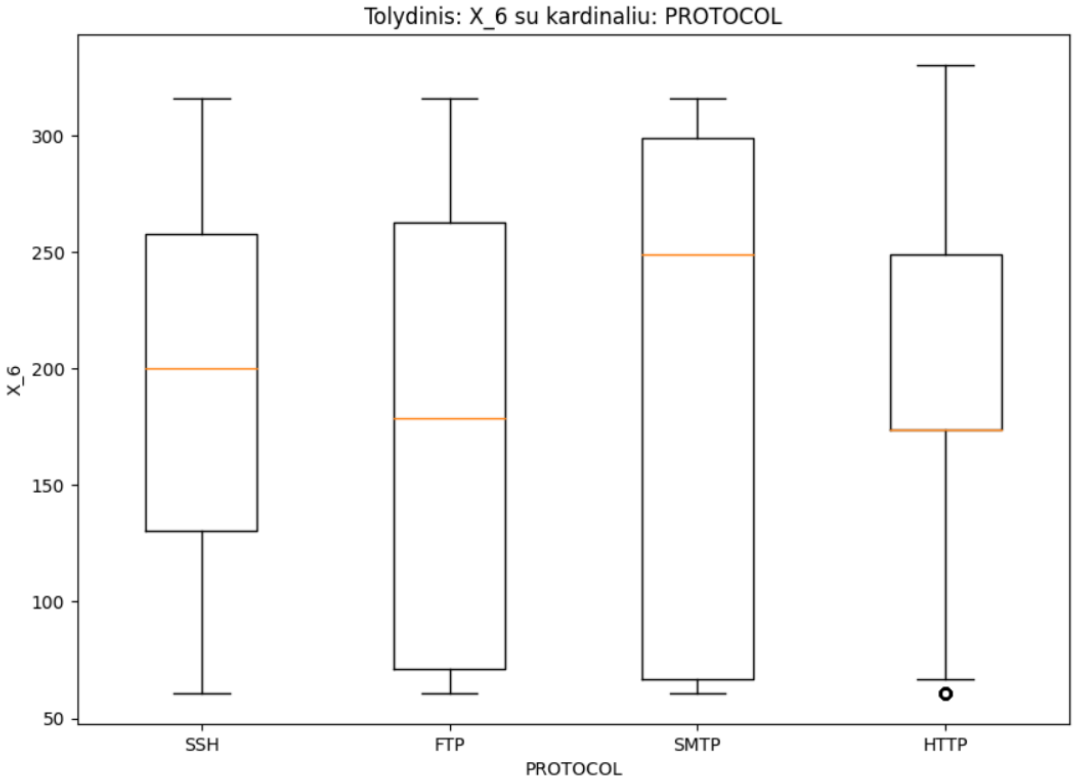
Atributas „OS“ „X_3“ atžvilgiu turi stiprų ryšį, nes prie tam tikrų tolydinio atributo „X_3“ reikšmių naudojamos skirtingos operacinės sistemos.

Atributo „REGION“ priklausomybė nuo „X_6“



Atributas „REGION“ „X_6“ atžvilgiu turi silpną ryšį, nes prie tam tikrų tolydinio atributo „X_6“ reikšmių naudojamos tos pačios regionų reikšmės ir tendencijų nepastebėta.

Atributo „PROTOCOL“ priklausomybė nuo „X_6“



Atributas „PROTOCOL“ „X_6“ atžvilgiu turi silpną ryšį, nes prie tam tikrų tolydinio atributo „X_6“ reikšmių naudojamos tos pačios protokolų reikšmės ir tendencijų nepastebėta.

3.6. Kovariacija ir koreliacija

Kodo sugeneruotas dokumentas *output/kovariacijos_matrica.csv*

5 lentelė. Kovariacijos matrica

	X_1	X_2	X_3	X_4	X_6	X_7
X_1	243.0554	-3.71915	3.956358	0.362347	41.75291	40.55132
X_2	-3.71915	8.637229	0.949424	1.220628	-15.0368	-2.50776
X_3	3.956358	0.949424	16.99818	9.616142	-1.4676	9.180101
X_4	0.362347	1.220628	9.616142	12.8366	9.262209	1.464186
X_6	41.75291	-15.0368	-1.4676	9.262209	5852.189	92.46507
X_7	40.55132	-2.50776	9.180101	1.464186	92.46507	653.1099

Kodo sugeneruotas dokumentas *output/koreliacijos_matrica.csv*

6 lentelė. Koreliacijos matrica

	X_1	X_2	X_3	X_4	X_6	X_7
X_1	1	-0.08117	0.061552	0.006487	0.035009	0.101779
X_2	-0.08117	1	0.078356	0.115923	-0.06688	-0.03339
X_3	0.061552	0.078356	1	0.65099	-0.00465	0.087127
X_4	0.006487	0.115923	0.65099	1	0.033793	0.015991
X_6	0.035009	-0.06688	-0.00465	0.033793	1	0.047296
X_7	0.101779	-0.03339	0.087127	0.015991	0.047296	1

Koreliacijos matrica atspindima grafiškai

1pav. Koreliacijos matrica

Koreliacijos matrica atvaizduota intervale [-1; 1], kad būtų galima matyti ir neigiamas priklausomybės. Iš grafiko aišku, kad dauguma atributų yra labai silpnai susiję (koreliacija iki |0.2|). Vieninteliai atributai „X_3“ ir „X_4“ koreliuoja vidutiniškai – koreliacijos koeficientas lygus 0,65.

4. Duomenų apdorojimas

4.1. Kategorinio tipo atributų vertimas į tolydinio tipo

Šiai užduočiai spręsti kode kviečiamas metodas *convert_cat_to_cont()*, kuriame išrenkamos unikalios kategorinės reikšmės kiekvienam kategoriniam atributui ir priskiriami sveikieji skaičiai, intervale [0; n] – kur n kategorinių reikšmių skaičius.

4.2. Duomenų normalizacija

Siekiant įgyvendinti šį reikalavimą kviečiamas metodas *normalize_values()*, kuriame nagrinėjamas kiekvienas atributas atskirai. Kadangi triukšmus pašalinome jau ankstesniame žingsnyje, šiame žingsnyje naudosime metodą, kuris yra jautrus jiems, bet patogus, dėl intervalo [0; 1] paprastumo.

```
def normalize_values(data, headers):
    for head in headers:
        sublist = list(map(lambda x: float(x[head]), data))
        min = np.min(sublist)
        max = np.max(sublist)
        for i in range(len(data)):
            data[i][head] = (float(data[i][head]) - min) / max
```

Kiekvienam atributui apskaičiuojamos min ir max reikšmės ir einant per kiekvieną elementą, nauja jo reikšmė bus lygi formulės išraiškai => $x = (x - min) / max$

Apdorojus duomenis išvedamas galutinis apdorotų duomenų formatas faile *output/apdoroti_normalizuoti_duomenys.csv*

REGION	OS	PRO-TO-COL	X_1	X_2	X_3	X_4	X_6	X_7	MALI-CIOUS_OF-FENSE
0	0	0	0.653846	1	0.275862	0.413793	0.340909	0.573276	0
0	1	0	0.673077	0	0.689655	1	0.528788	0.668103	1
0	0	0	0.019231	0.25	0	0	0.340909	0.728448	1
0	0	0	0.903846	0.5	0.275862	0.413793	0.568182	0.357759	1
0	0	0	0.038462	0.5	0.137931	0.344828	0.340909	0.400862	1
0	0	0	0.038462	0.5	0.137931	0.344828	0	0.668103	1
0	0	0	0.25	0.25	0.206897	0.137931	0.340909	0.780172	1
1	0	0	0.653846	0.166667	0	0.344828	0.771212	0.573276	1
1	0	0	0.365385	0.333333	0.206897	0.137931	0.340909	0.737069	1
0.666667	0	0	0.673077	0	0.689655	0.344828	0.198485	0.668103	1
0	0	0	0.903846	0.5	0.068966	0.482759	0.568182	0.400862	1
0	0	0	0.038462	0.5	0.413793	0.068966	0.771212	0.400862	1
0.666667	0	0	0.269231	0	0.482759	0.206897	0	0	1
0.333333	0	0	0.384615	0.583333	0.931034	0.827586	0	0.073276	1

1	0	0	0.653846	0.166667	0.344828	0.344828	0.771212	0.573276	1
0.666667	0	1	0.673077	0	0.689655	1	0.719697	0.668103	1
0.333333	0	0	0.173077	0.583333	0.137931	0.344828	0.771212	0	1
0	0	0	0.038462	0.5	0.275862	0.413793	0.340909	0.668103	1
1	0	0	0.653846	0.166667	0	0	0.568182	0.573276	1
1	0	0	0.653846	0.166667	0.206897	0.137931	0.719697	0.573276	1
0	0	0	0.038462	0.5	0.206897	0.137931	0.568182	0.400862	1
1	1	0	0.653846	0.166667	0.931034	0.689655	0.771212	0.573276	1
1	0	0	0.884615	0.333333	0.413793	0.068966	0.292424	0.668103	0
0	0	0	0.711538	0.5	0.206897	0.137931	0.568182	0.668103	1

5. Išvados

Duomenų apdorojimas – atsakingas procesas. Prieš pradedant analizuoti grafines atributų išraiškas svarbu tinkamai apdoroti trūkstamas reikšmes, triukšmus, kardinalumo problemas. O pradėjus analizuoti svarbu žinoti, ko ieškoti, nes dauguma priklausomybių gali būti itin nereikšmingos. Nagrinėjant įvairius įrankius, tokius kaip atvirojo kodo interpretuojama programavimo kalbe „Python“ bei jos bibliotekos „matplotlib“, „numpy“, „pandas“ ir t.t. galima palengvinti darbą ir optimaliau atvaizduoti duomenų priklausomybes – kas šiame darbe ir buvo atlikta braižant histogramas, „scatterplot (matrix)“, „bar plot“, „box plot“ tipo diagramas.