Kauno technologijos universitetas

Informatikos fakultetas

Laboratorinis darbas Nr. 1

Laboratorinio darbo ataskaita

Lukas Kuzmickas IFF-1/6

Studentas

Nakrošis Arnas

Dėstytojas

Kaunas, 2024

Turinys

1. Įvadas

Laboratorinio darbo tikslas yra surasti, apdoroti ir išanalizuoti tinkamą duomenų rinkinį. Darbo tikslo atlikimo eiga yra:

1. Pasirinkti tinkamą duomenų rinkinį;
2. Atlikti duomenų rinkinio kokybės analizę;
3. Nupaišyti ir aprašyti duomenų rinkinio atributų histogramas;
4. Nustatyti sąryšius tarp atributų;
5. Paskaičiuoti kovariacijos ir koreliacijos reikšmes tarp tolydinio tipo atributų ir grafiškai atvaizduoti koreliacijos matricą;
6. Atlikti duomenų normalizaciją;
7. Kategorinio tipo kintamuosius paversti į tolydinio tipo kintamuosius.

2. Duomenų rinkinys

Darbui atlikti reikalinga pasirinkti duomenų rinkinį, kuris turėtų daugiau nei 500 įrašų ir būtų sudarytas iš 8 stulpelių, kurie būtų tolydinio ir kategorinio tipo (po 2).

Duomenų rinkinį sudaro:

- Youtube platformos rankingas (tolydinis);
- Youtube kūrėjo kanalo pavadinimas;
- Youtube sekėjų skaičius (tolydinis);
- Youtube vaizdo įrašo peržiūrų skaičius (tolydinis);
- Kategorija (kategorinis);
- Pavadinimas;
- Įkeltų vaizdo įrašų skaičius (tolydinis);
- Šalis (kategorinis);
- Šalies sutrumpinimas (kategorinis);
- Kanalo tipas (kategorinis);
- Rankingas pagal vaizdo įrašo peržiūras (tolydinis);
- Šalies rankingas (tolydinis);
- Kanalo tipo rankingas (tolydinis);
- Vaizdo įrašo peržiūros per paskutinias 30 dienų (tolydinis);
- Mažiausios mėnesio pajamos (tolydinis);
- Didžiausios mėnesio pajamos (tolydinis);
- Mažiausios metinės pajamos (tolydinis);
- Didžiausios metinės pajamos (tolydinis);
- Sekėjų skaičius per paskutinias 30 dienų (tolydinis);
- Sukūrimo data (metai);
- Sukūrimo data (data).

3. Duomenų rinkinio kokybės analizė

Kadangi duomenų rinkinyje yra įmanomos anomalijos svarbu atlikti duomenų rinkinio kokybės analizę.

3.1. Tolydinis tipas

Tolydiniam tipui kokybės analizei reikia apskaičiuoti:

- Bendrą reikšmių skaičių;
- Trūkstamų reikšmių procentą;
- Kardinalumą;
- Minimalią ir maksimalią reikšmes;
- 1-ąją ir 3-ąją kvartilius;
- Vidurkį;
- Medianą;
- Standartinį nuokrypį,

1 lentelė. Tolydinio tipo atributų kokybės analizės lentelė

Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %
youtube_rank	992	0.3015075376884391
subscribers	995	0.0
video_views	995	0.0
uploads	995	0.0
video_views_rank	994	0.10050251256281673
country_rank	879	11.65829145728643
channel_type_rank	962	3.316582914572863
video_views_for_the_last_30_days	939	5.628140703517593
lowest_monthly_earnings	995	0.0
highest_monthly_earnings	995	0.0
lowest_yearly_earnings	995	0.0
highest_yearly_earnings	995	0.0
subscribers_for_last_30_days	658	33.86934673366834
Kardinalumas	**Minimali reikšmė**	**Maksimali reikšmė**
993	1.0	995.0
289	12300000.0	245000000.0
988	0.0	228000000000.0
777	0.0	301308.0
954	1.0	4057944.0
247	1.0	7741.0
287	1.0	7741.0
909	1.0	6589000000.0
557	0.0	850900.0
736	0.0	13600000.0
757	0.0	10200000.0

419	0.0	163400000.0
54	1.0	8000000.0
1-asis kvartilis	**3-iasis kvartilis**	**Vidurkis**
251.75	747.25	499.3094758064516
14500000.0	24600000.0	0
4288145410.0	13554701853.0	22982412.06030151
194.5	2667.5	11039537052.03819
323.0	3584.5	0
11.0	123.0	0
27.0	139.75	9187.125628140704
20137500.0	168826500.0	0
2700.0	37900.0	0
43500.0	606800.0	0
32650.0	455100.0	554248.9044265593
521750.0	7300000.0	386.0534698521047
100000.0	400000.0	745.7193347193347
Mediana	**Standartinis nuokrypis**	
499.5	286.6689707331033	
0	17517296.038876075	
17700000.0	14103751717.735266	
7760819588.0	34134.18645607931	
0	1362096.5332542644	
0	1231.5436111169406	
729.0	1943.3757024340916	
0	416156393.532454	
0	71822.60510325353	
0	1148045.1361556875	
915.5	860783.2232396941	
51.0	13790102.687342707	
65.5	613888.4280797704	

3.2. Kategorinis tipas

Kategoriniam tipui kokybės analizei reikia apskaičiuoti:

- Bendrą reikšmių skaičių;
- Trūkstamų reikšmių procentą;
- Kardinalumą;
- Modą;
- Modos dažnumo reikšmę;
- Modos procentinę reikšmę;
- 2-ają modą;
- 2-osios modos dažnumo reikšmę;
- 2-osios modos procentinę reikšmę.

2 lentelė. Kategorinio tipo atributų kokybės analizės lentelė

Atributo pavadiniams	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas
category	949	4.623115577889447	18
Country	873	12.261306532663319	49
Abbreviation	873	12.261306532663319	49
channel_type	965	3.015075376884424	14
Moda	**Modos dažnumas**	**Moda, %**	**2-oji Moda**
Entertainment	241	25.40%	Music
United States	313	35.85%	India
US	313	35.85%	IN
Entertainment	304	31.50%	Music

2-osios Modos dažnumas	2-oji Moda, %
202	28.53%
168	30.00%
168	30.00%
216	32.68%

3.3. Duomenų modifikavimas

Kadangi svarbu, kad neegzistuotų tuščių laukų duomenų rinkinio įrašuose, buvo nuspręsta ištrinti kiekvieną įrašą, kurio vienas iš atributų yra tuščias. Nuspręsta ištrinti, o ne įrašyti reikšmę dėl to, kad neaišku kaip reikšmių keitimas nulems galutinį rezultatą (duomenų kokybę).

3 lentelė. Tolydinio tipo atributų kokybės analizės lentelė po modifikavimo

Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %
youtube_rank	551	0
subscribers	551	0
video_views	551	0
uploads	551	0
video_views_rank	551	0
country_rank	551	0
channel_type_rank	551	0
video_views_for_the_last_30_days	551	0
lowest_monthly_earnings	551	0
highest_monthly_earnings	551	0
lowest_yearly_earnings	551	0
highest_yearly_earnings	551	0
subscribers_for_last_30_days	551	0
Kardinalumas	**Minimali reikšmė**	**Maksimali reikšmė**
551	1.0	995.0
236	12300000.0	245000000.0
551	2634.0	228000000000.0
505	1.0	301308.0
549	1.0	4054962.0
179	1.0	7683.0
186	1.0	7670.0
549	3.0	6589000000.0
427	0.0	576000.0
441	0.0	9200000.0
471	0.0	6900000.0
258	0.0	110600000.0
37	1.0	8000000.0
1-asis kvartilis	**3-iasis kvartilis**	**Vidurkis**
210.0	707.5	466.0
14850000.0	26450000.0	0
5056593575.5	15052173669.0	24503085.299455535
437.0	4110.0	13028662242.00363
223.5	1578.0	0
10.0	105.0	0
19.0	114.0	14828.662431941924
49179500.0	236143000.0	0
11650.0	58550.0	0

186650.0	936450.0	0
140050.0	702350.0	123062.3393829401
2200000.0	11200000.0	181.56987295825772
100000.0	400000.0	229.57350272232304

Mediana	**Standartinis nuokrypis**
450.0	288.9412741217618
0	19461523.525200784
18700000.0	17289604344.141457
8984089026.0	44318.484314075286
0	661980.0280565019
0	733.350656389331
1294.0	912.9645542216995
0	486975536.613558
0	76180.47216453808
0	1217570.3390879042
570.0	912746.1983166751
41.0	14626369.743877957
48.0	588923.4010863579

4 lentelė. Kategorinio tipo atributų kokybės analizės lentelė po modifikavimo

Atributo pavadiniams	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %
category	551	0
Country	551	0
Abbreviation	551	0
channel_type	551	0

Kardinalumas	Moda	Modos dažnumas
17	Entertainment	145
41	United States	178
41	US	178
14	Entertainment	138

Moda, %	2-oji Moda	2-osios Modos dažnumas	2-oji Moda, %
26.32%	Music	109	26.85%
32.30%	India	138	37.00%
32.30%	IN	138	37.00%
32.85%	Music	123	33.24%

4. Atributų grafikai

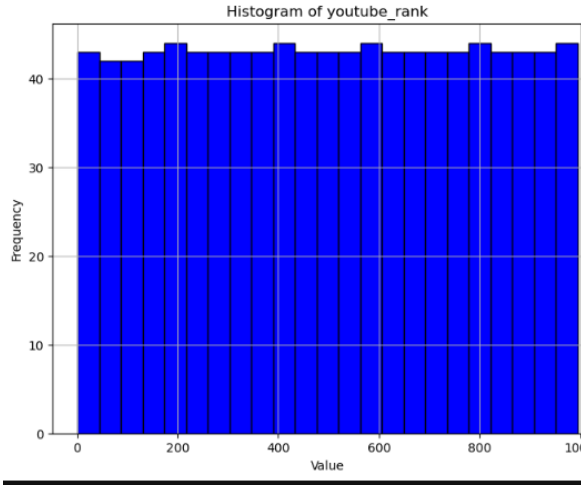
Grafikai buvo sukurti su programavimo kalba „Python“ naudojant biblioteką „Python Pandas“.

4.1. Tolydinio tipo histogramos

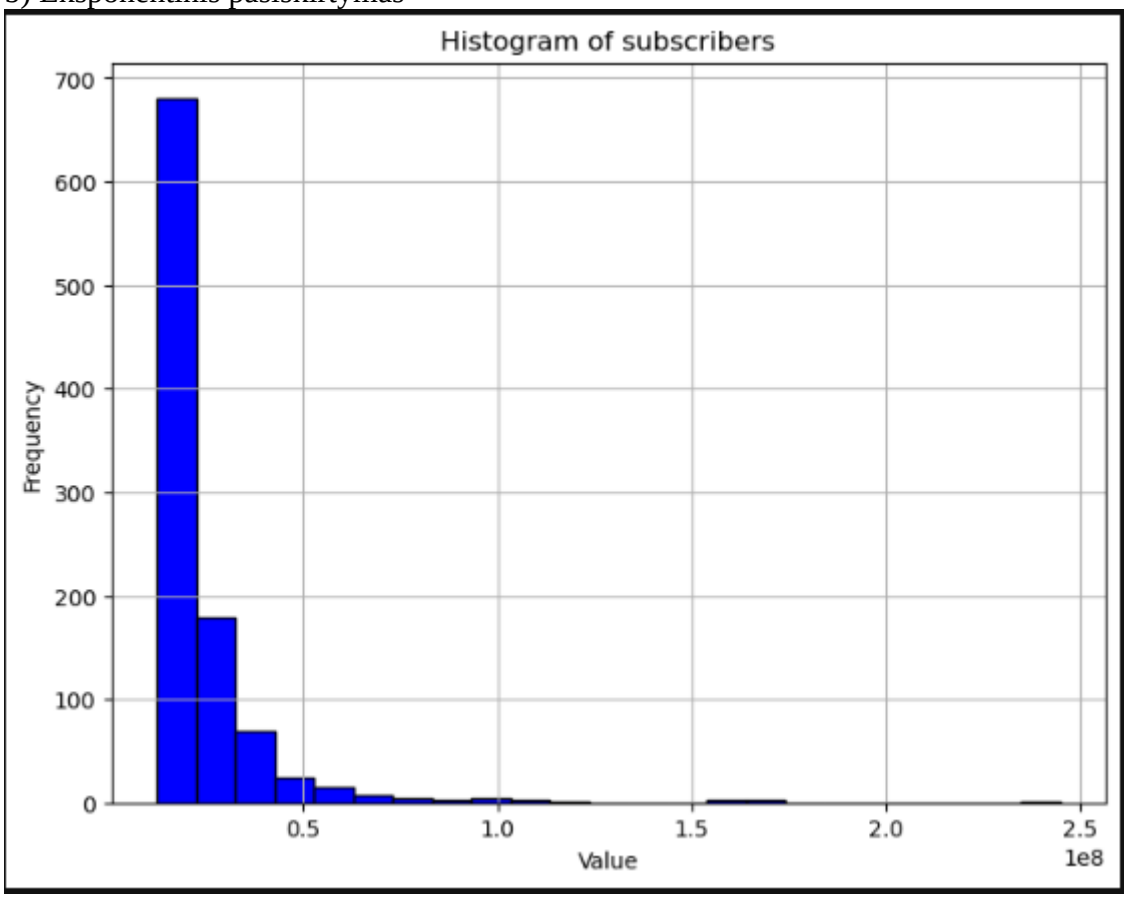
Tolydinio tipo atributais atvaizduoti buvo naudojamos histogramos. Histogramoms reikia suskaičiuoti kiek reikės naudoti stulpelių. Tam naudojama formulė: $1 + 3{,}22 \bullet log_e^n$, kur n yra duomenų rinkinio įrašų kiekis. Panaudojus formulę gaunama, kad reikia naudoti ~7 stulpelius esant 551 duomenų kiekiui. Sukurtos histogramos pateiktos 1 paveiksle.

Galima pastebėti, kad kiekvienoje histogramoje reikšmės yra įvairaus pasiskirstymo tipo, tačiau dominuoja eksponentinis, nes tik viena reikšmė arba intervalas yra didžiausias arba turi daugiausiai reikšmių.

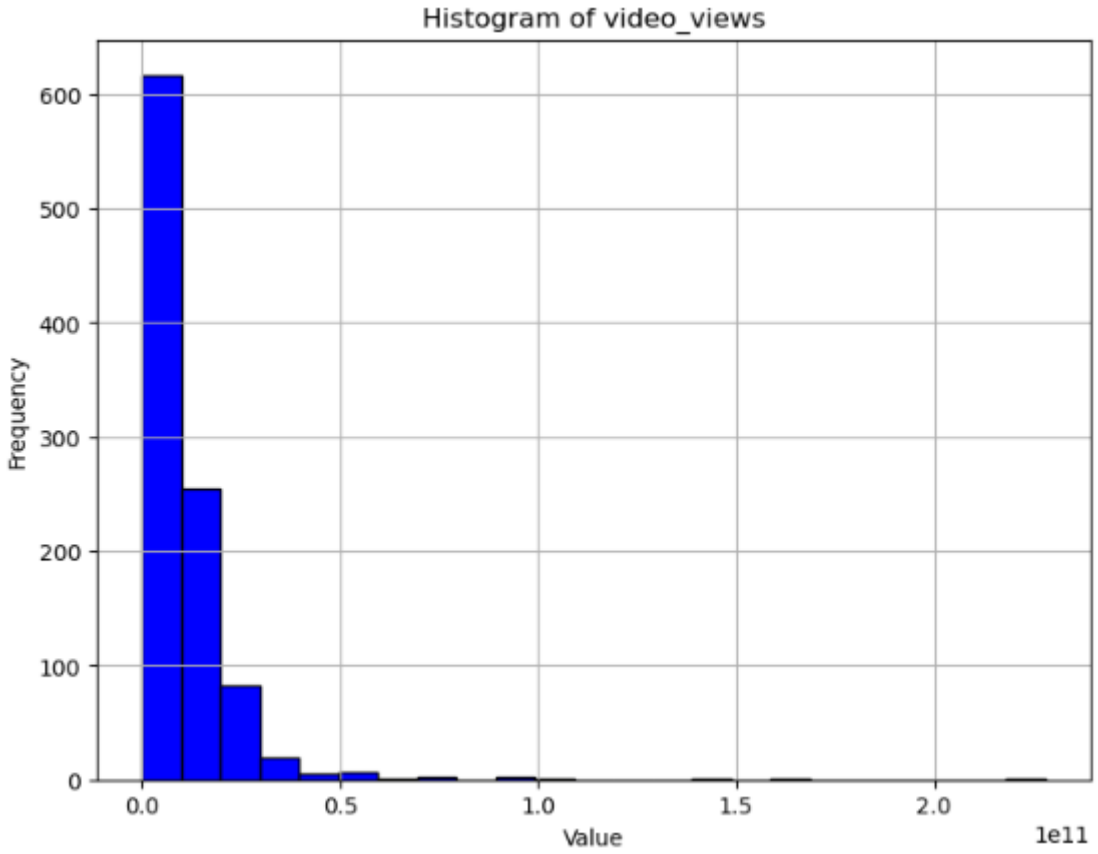
a) Normalusis pasiskirstymas



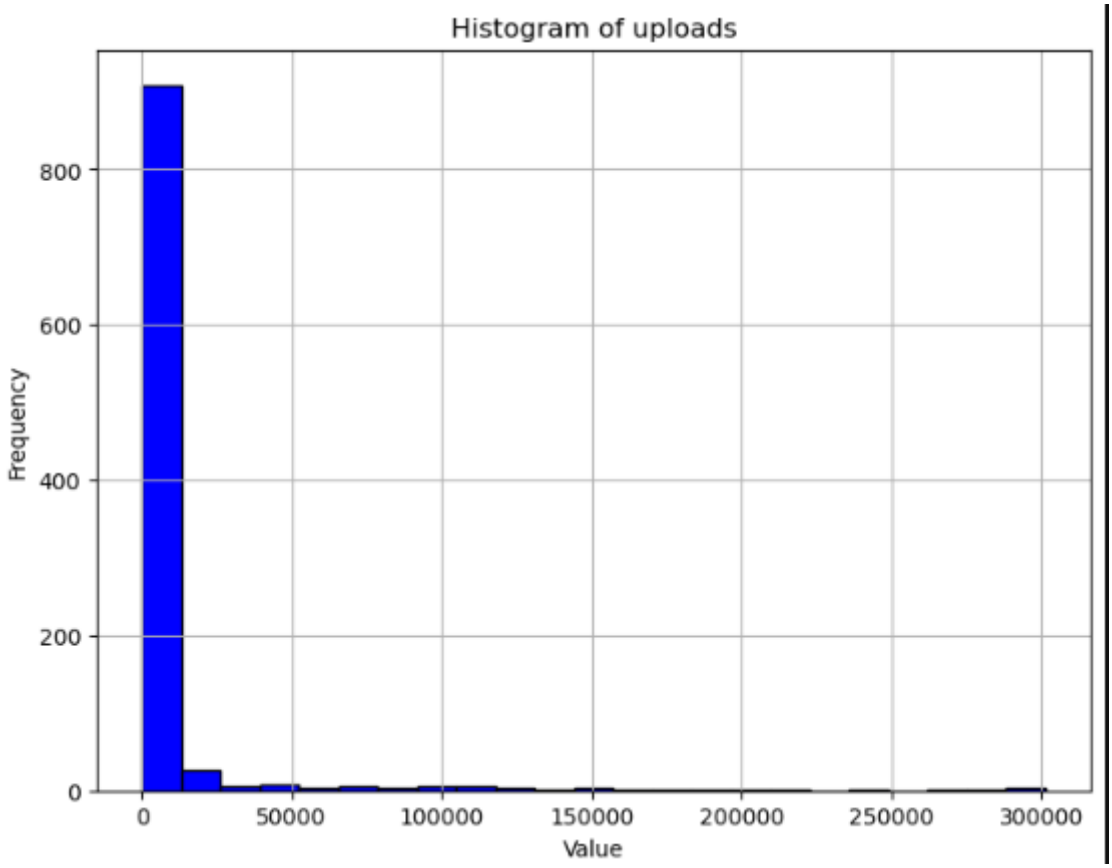
b) Eksponentinis pasiskirtymas



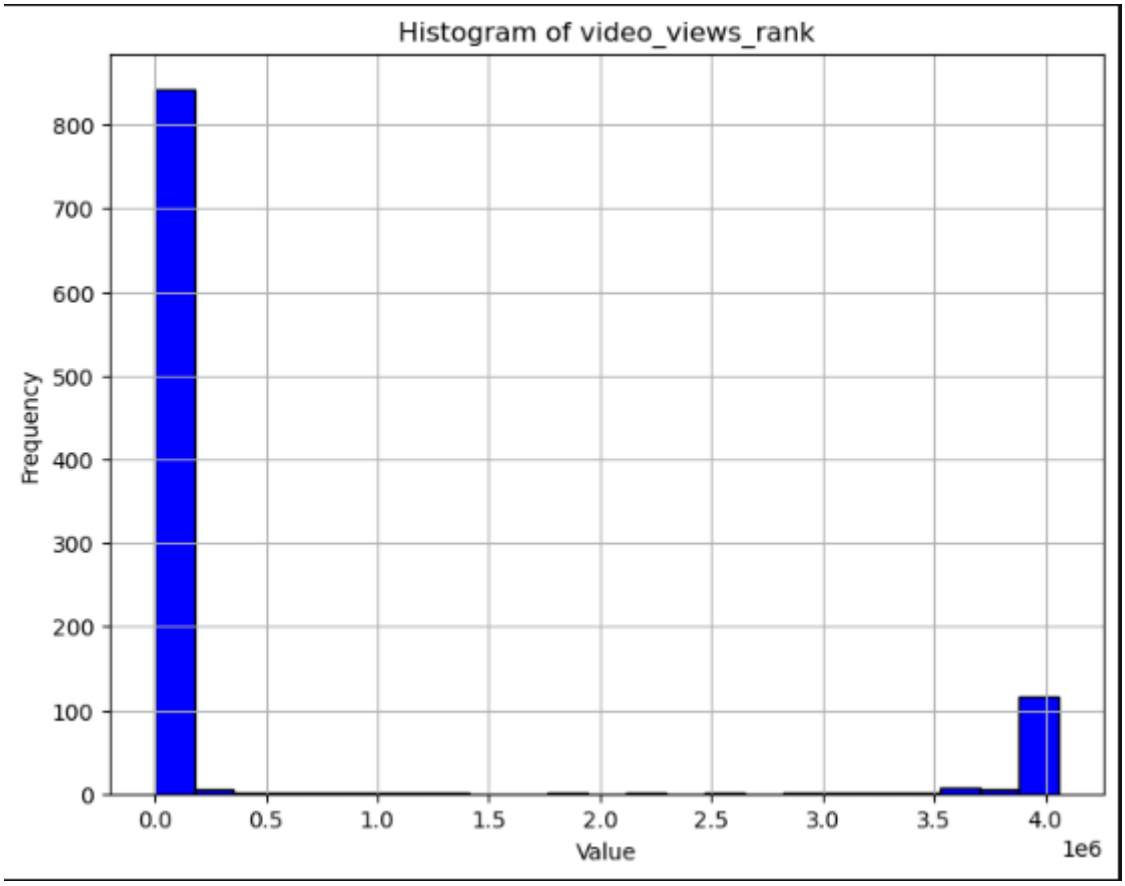
c) Eksponentinis pasiskirtymas



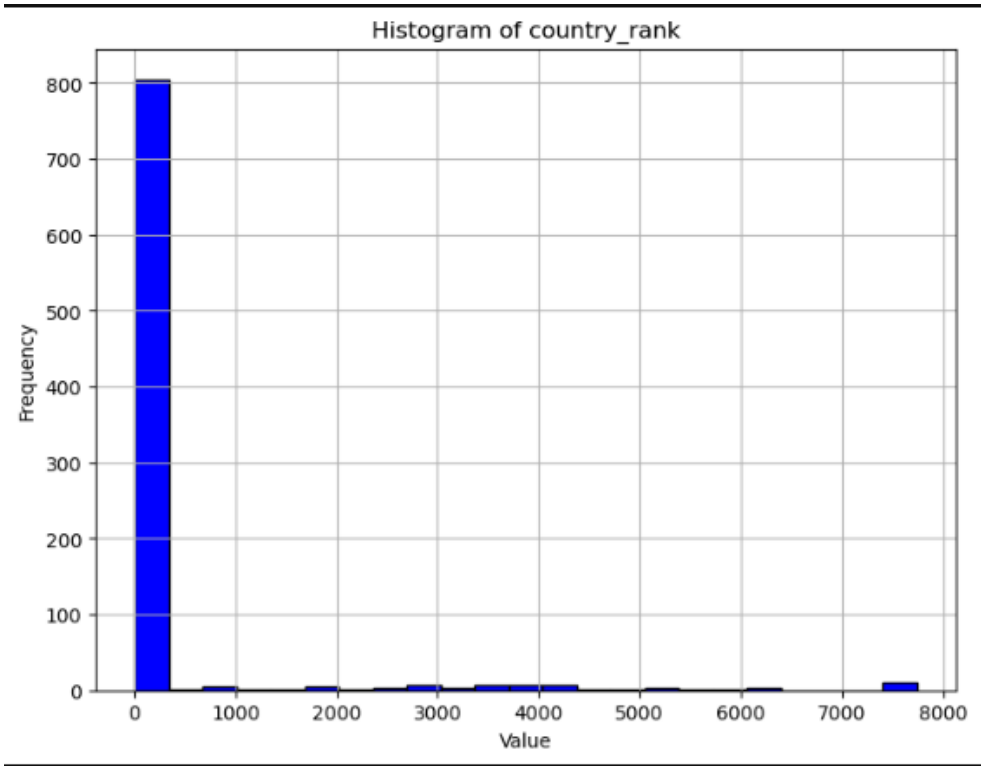
d) Eksponentinis pasiskirtymas



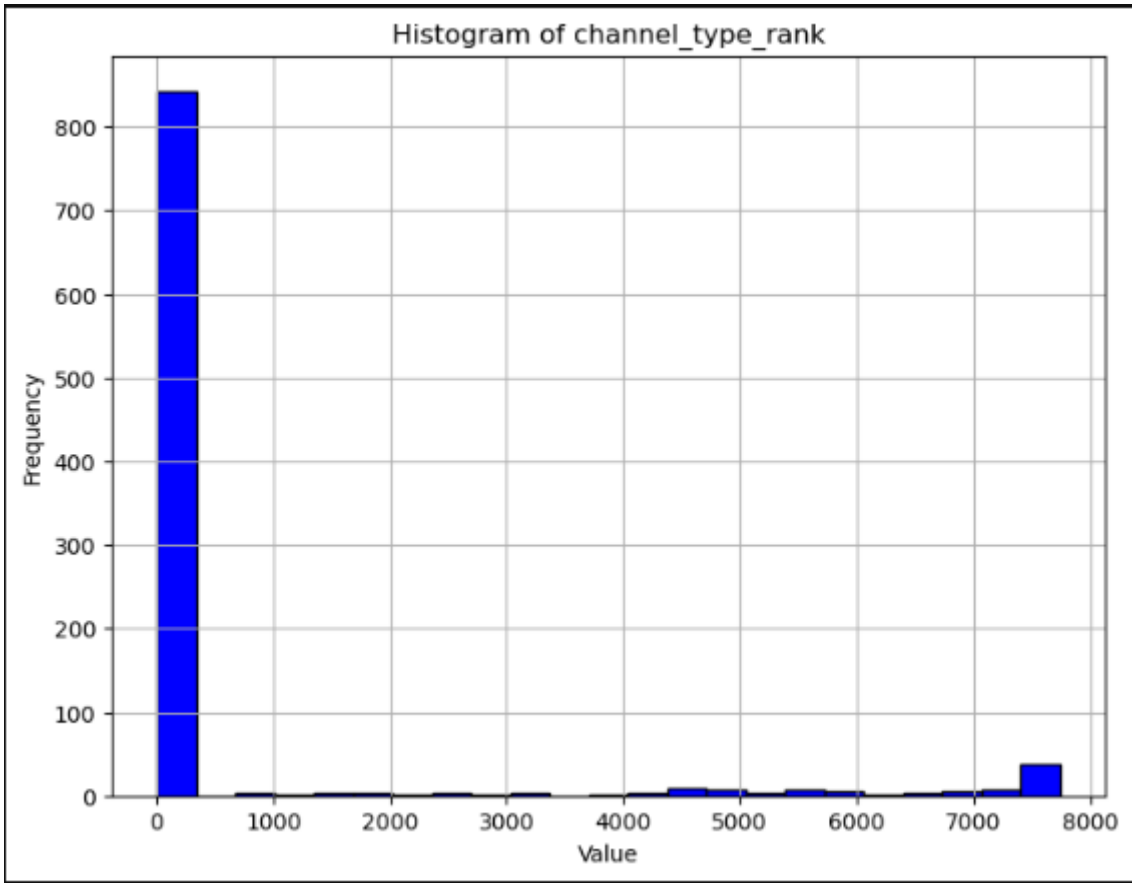
e) Eksponentinis pasiskirtymas



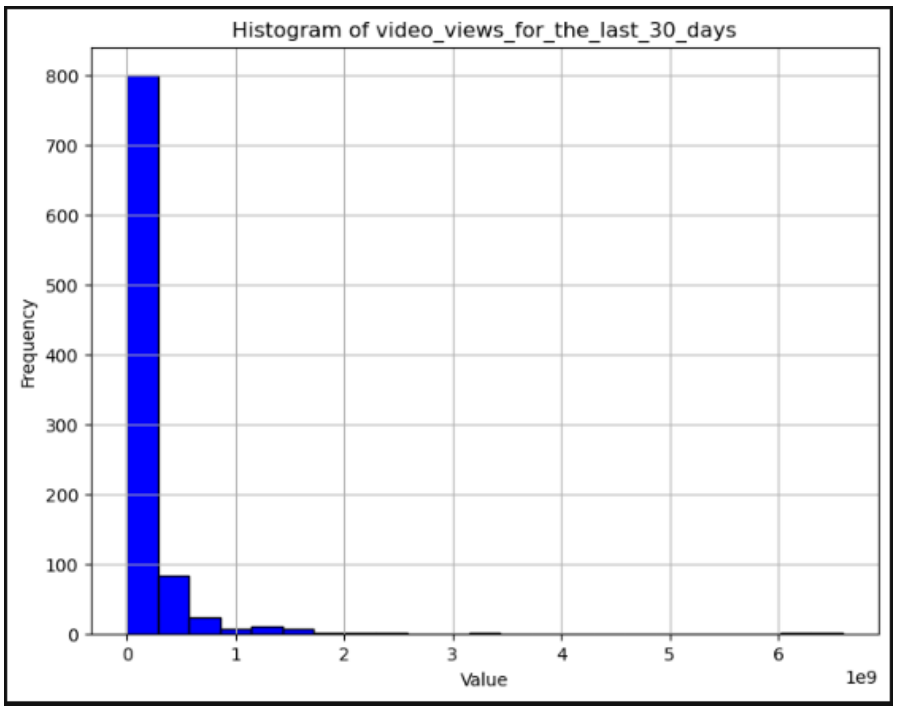
f) Eksponentinis pasiskirtymas



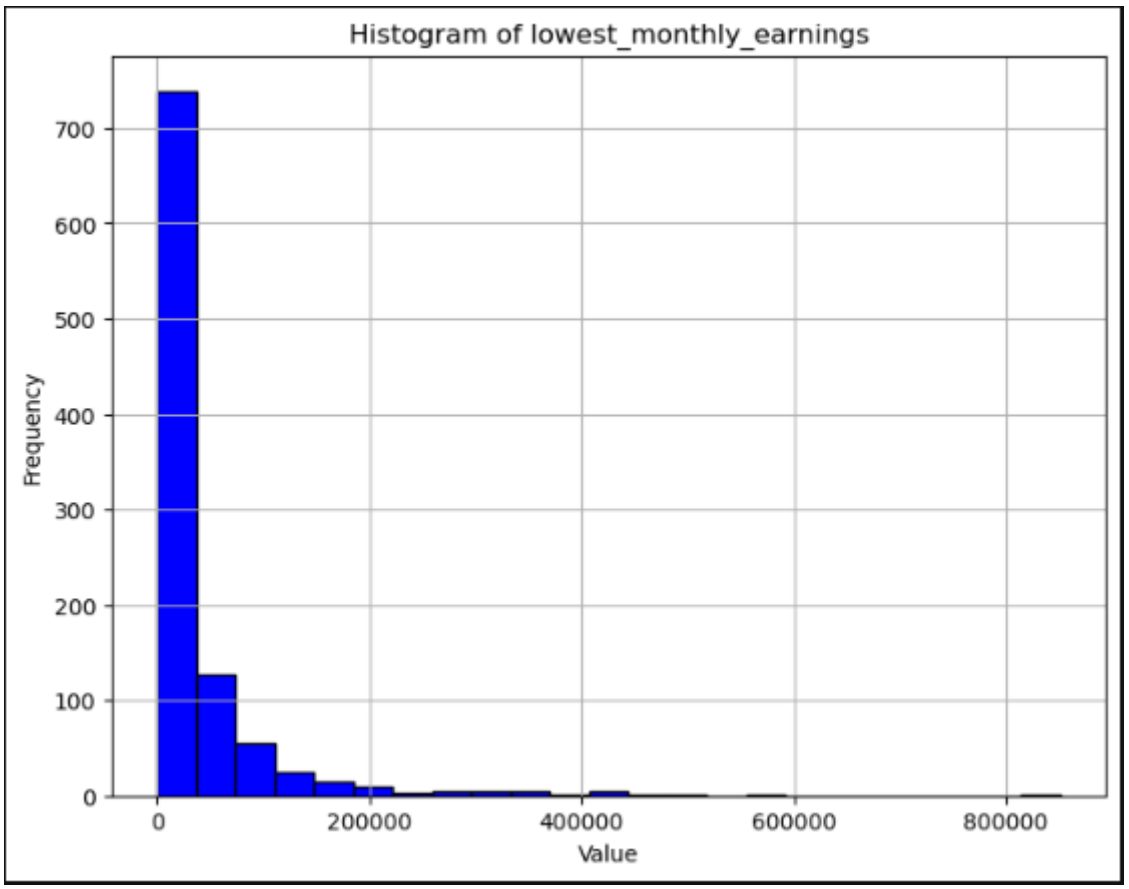
g) Eksponentinis pasiskirtymas



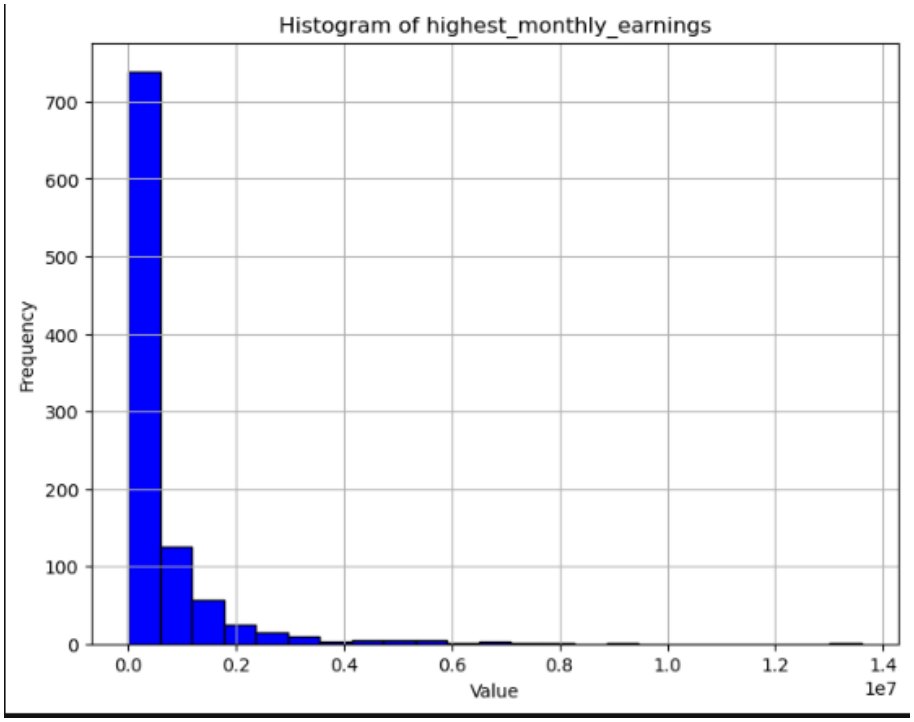
h) Eksponentinis pasiskirtymas



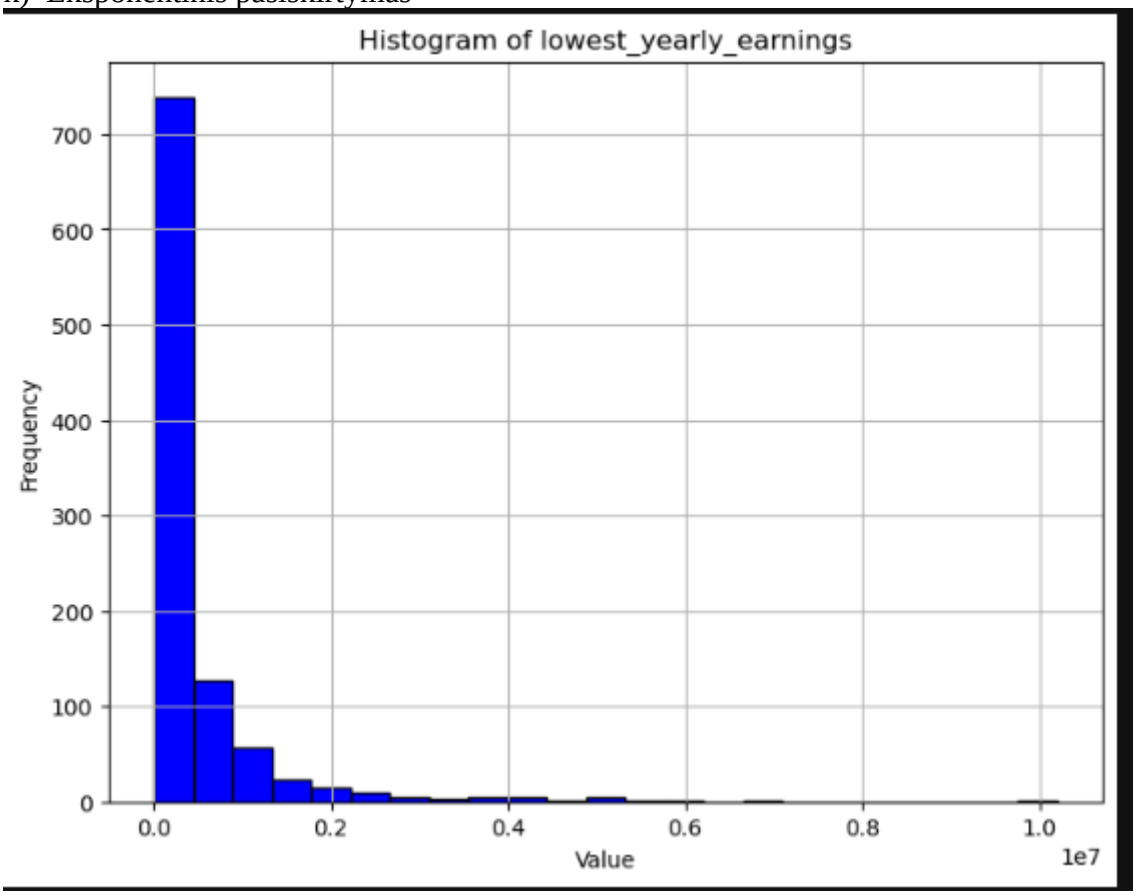
i) Eksponentinis pasiskirtymas



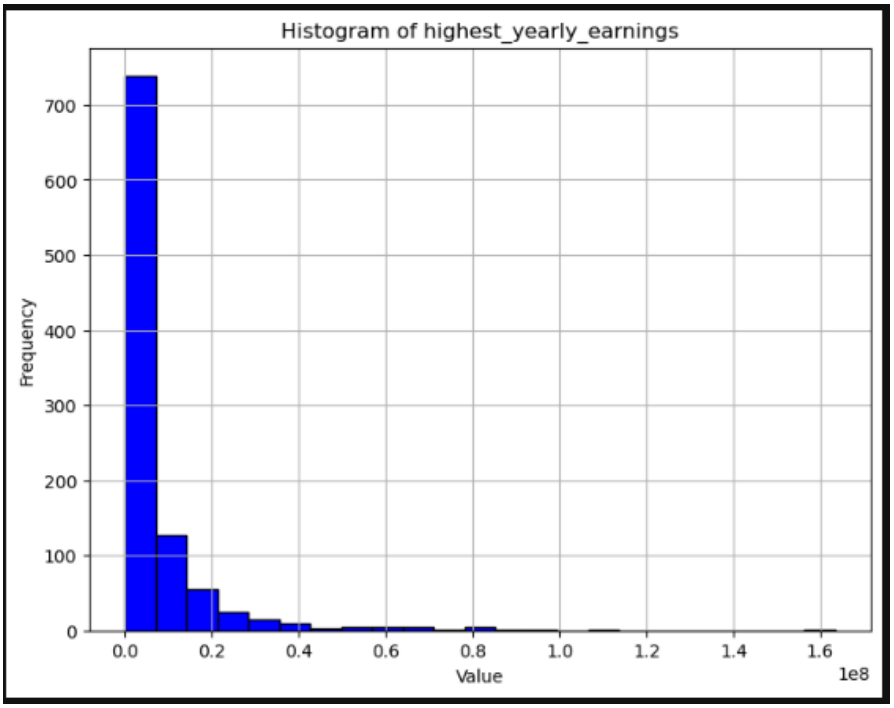
j) Eksponentinis pasiskirtymas



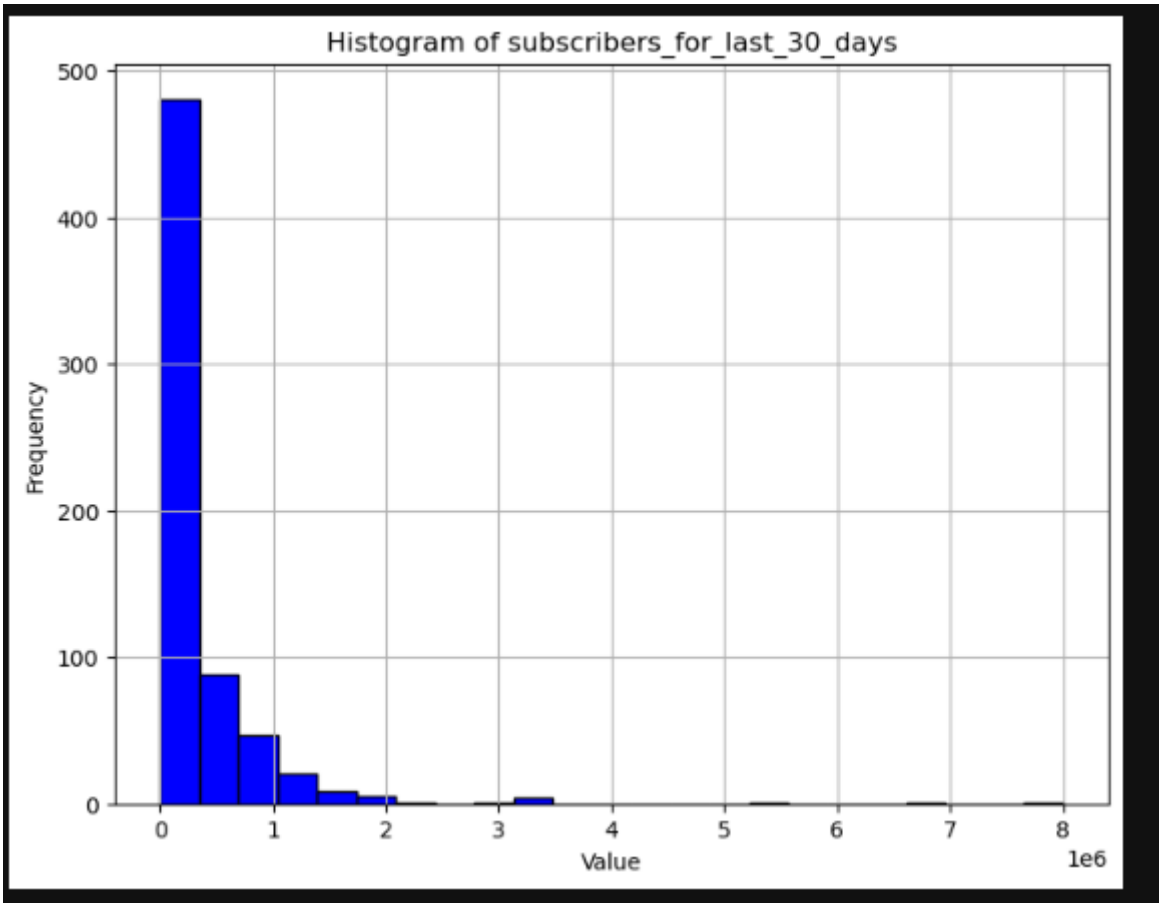
k) Eksponentinis pasiskirtymas



l) Eksponentinis pasiskirtymas



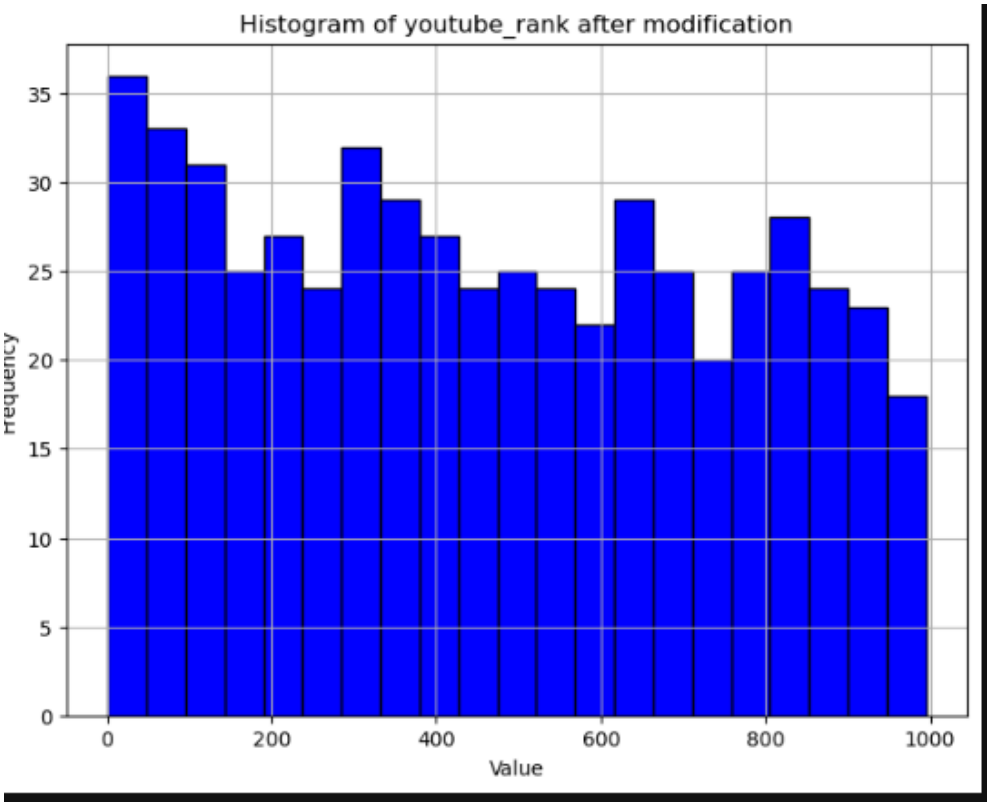
m) Eksponentinis pasiskirtymas



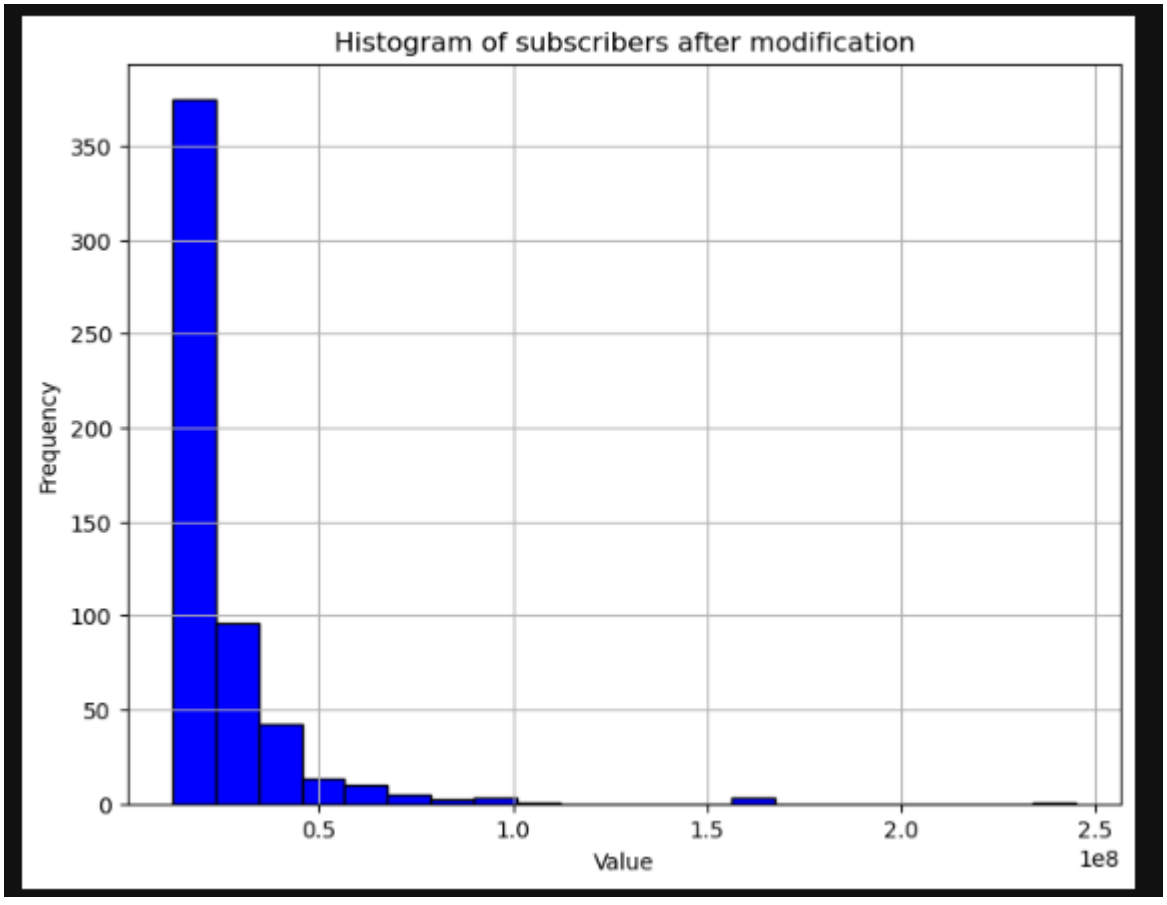
1 pav. tolydinių atributų histogramos prieš modifikavimą a), b) , c) , d) , e) , f) , g) , h) , i) , j), k), l), m)

Modifikuojame mūsų tolydžiuosius kintamuosius ir gauname tokias histogramas:

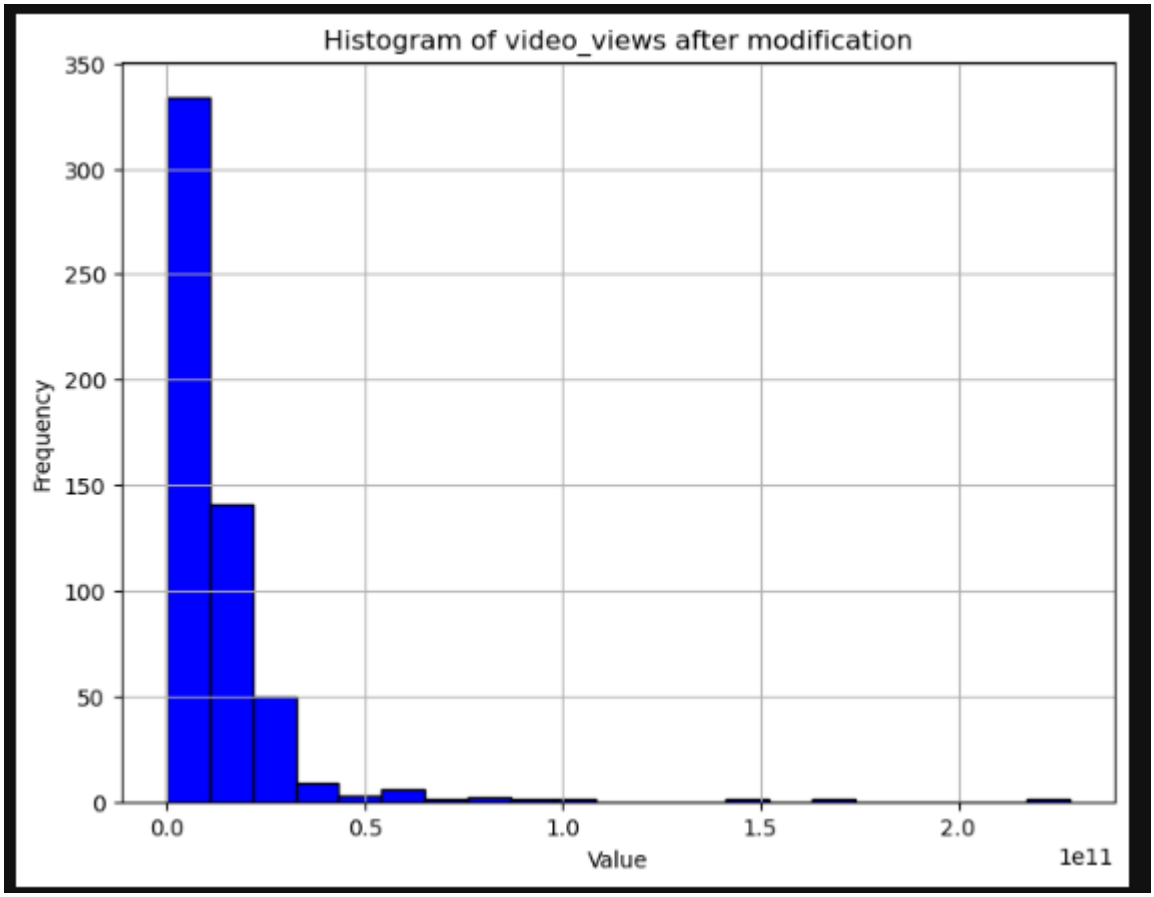
a) Normalusis pasiskirstymas



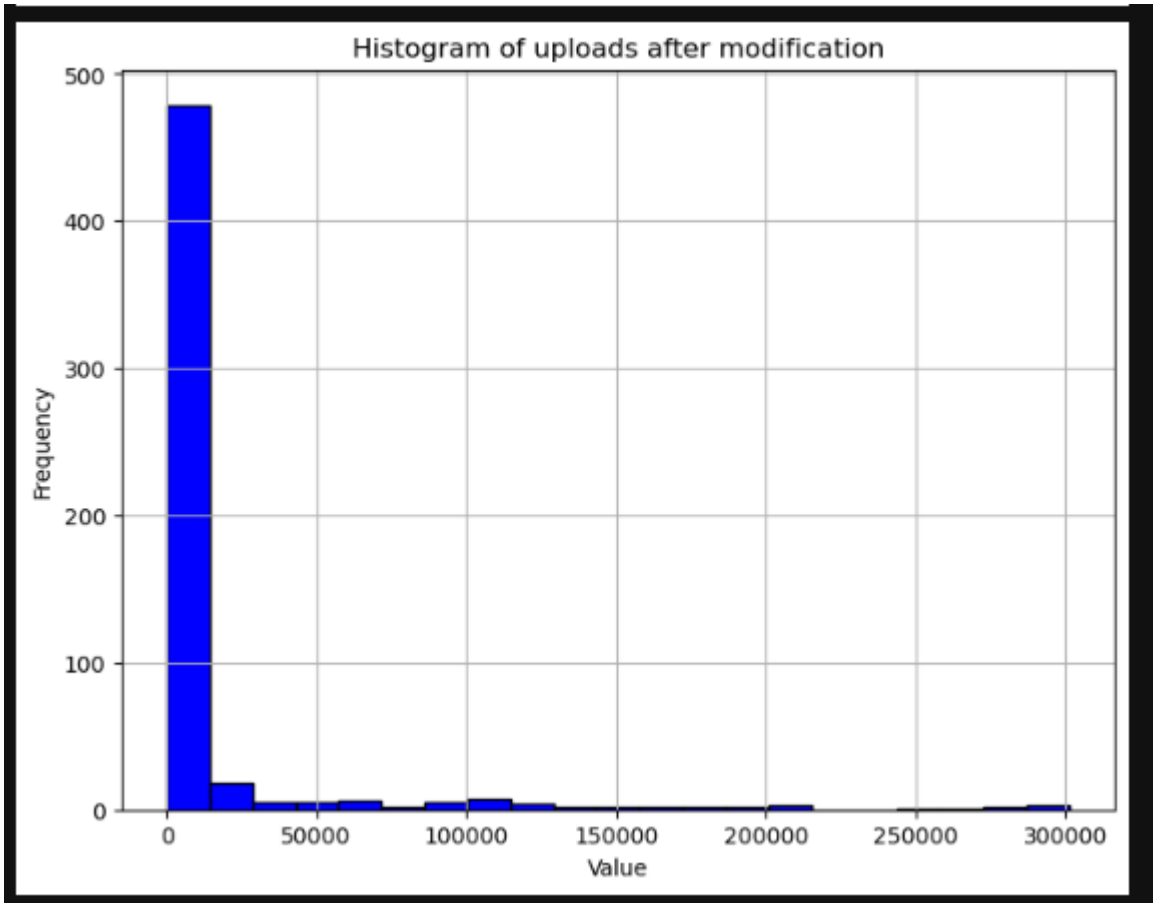
b) Eksponentinis pasiskirtymas



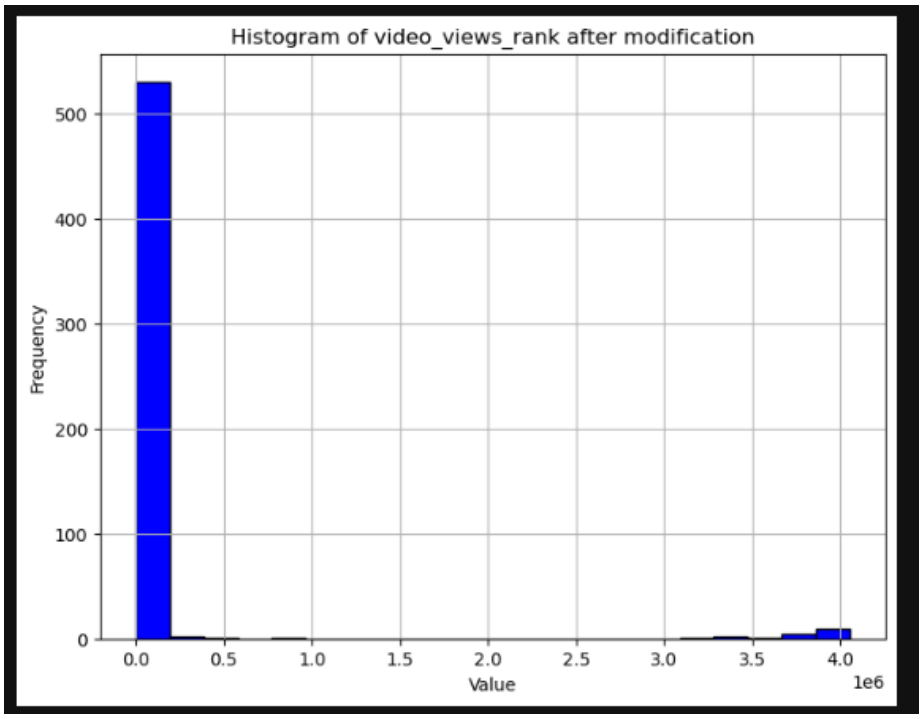
c) Eksponentinis pasiskirtymas



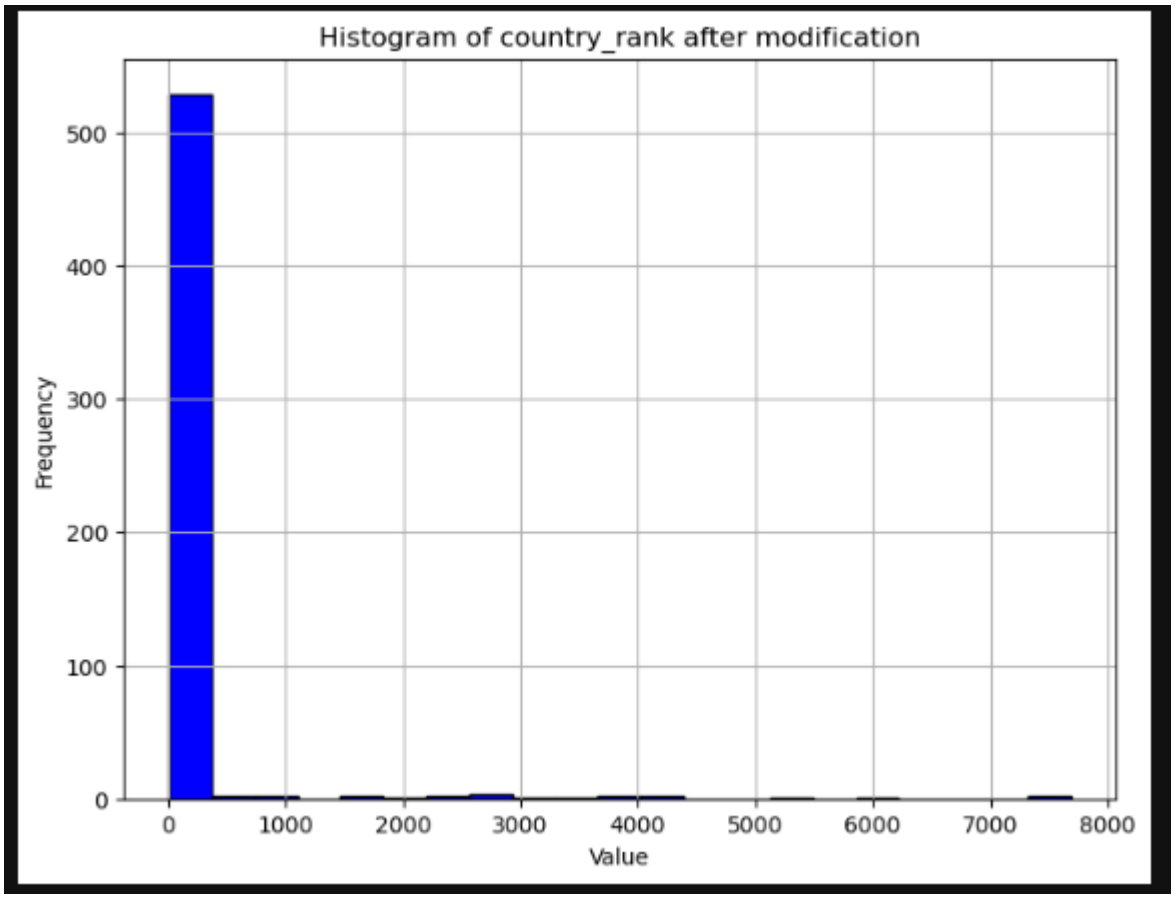
d) Eksponentinis pasiskirtymas



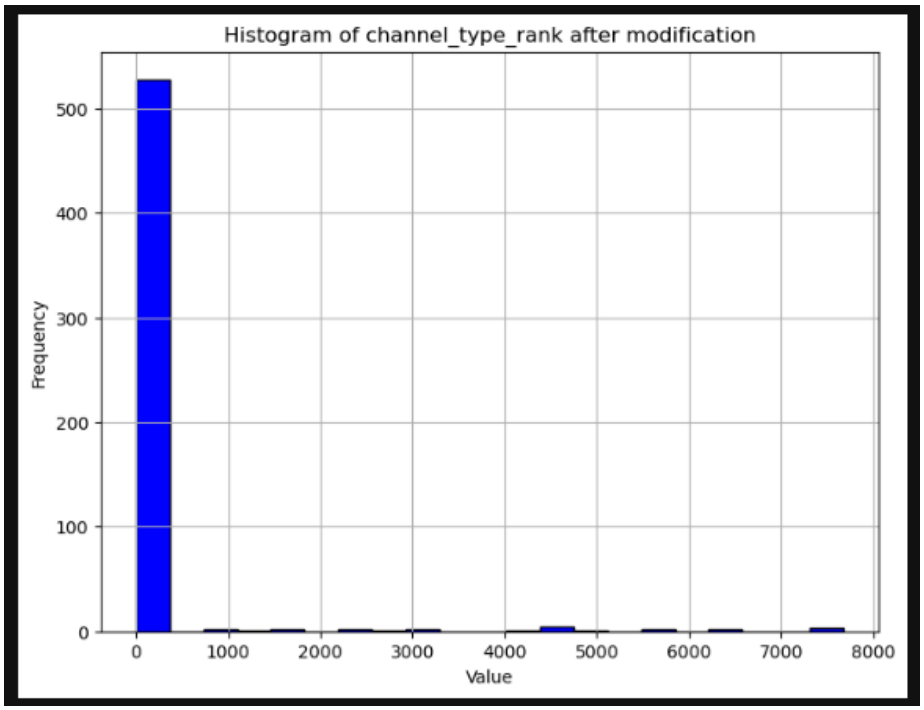
e) Eksponentinis pasiskirtymas



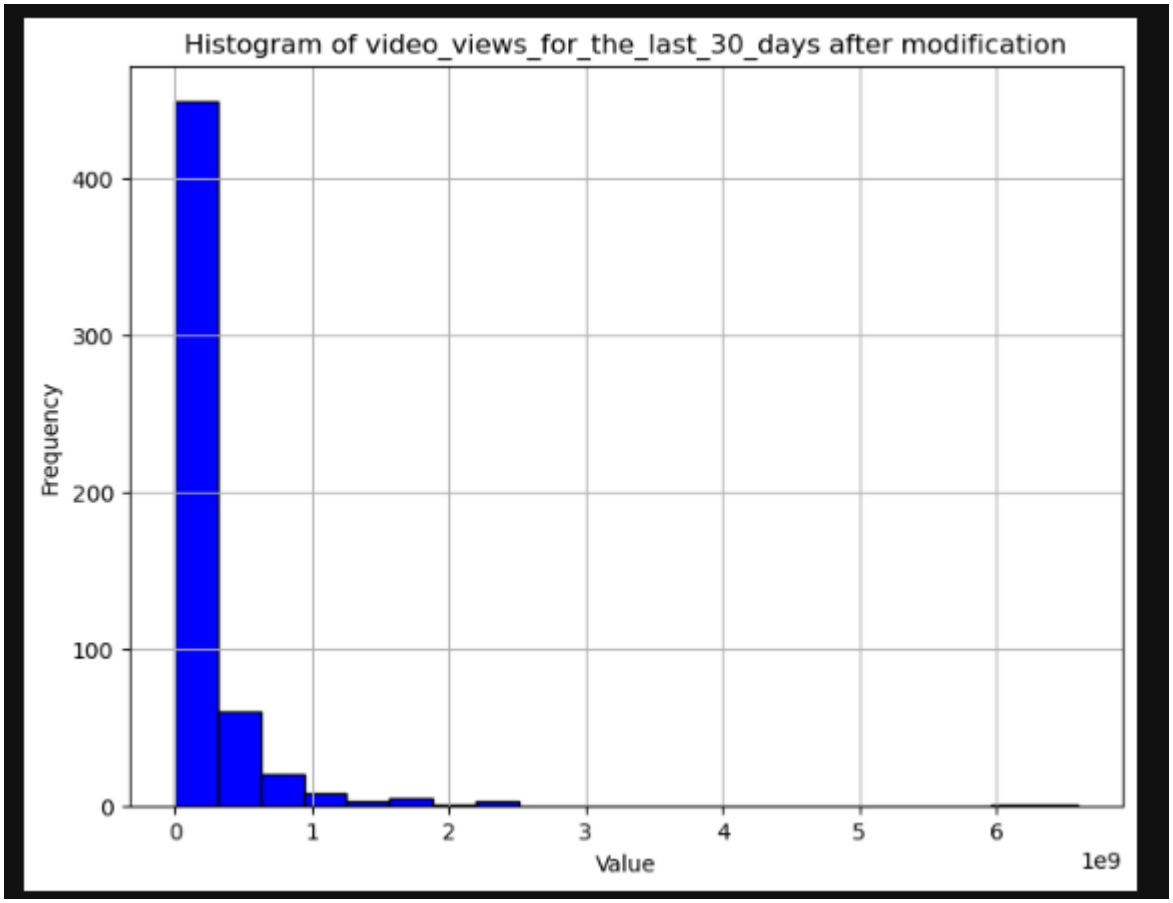
f) Eksponentinis pasiskirtymas



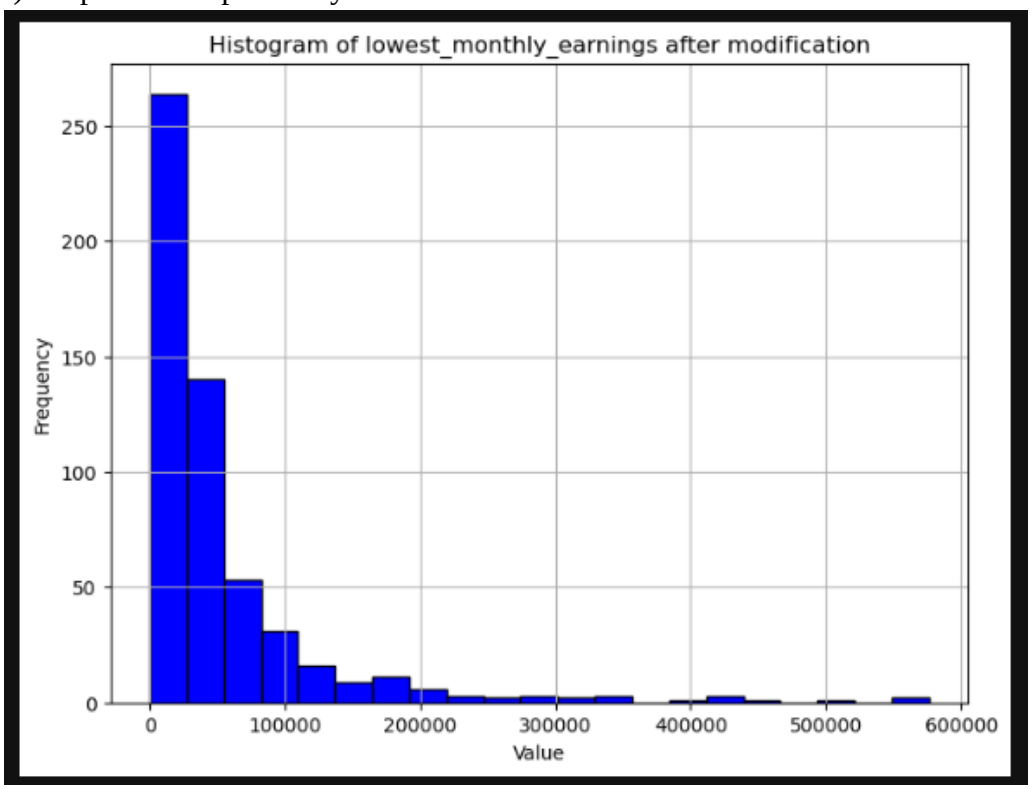
g) Eksponentinis pasiskirtymas



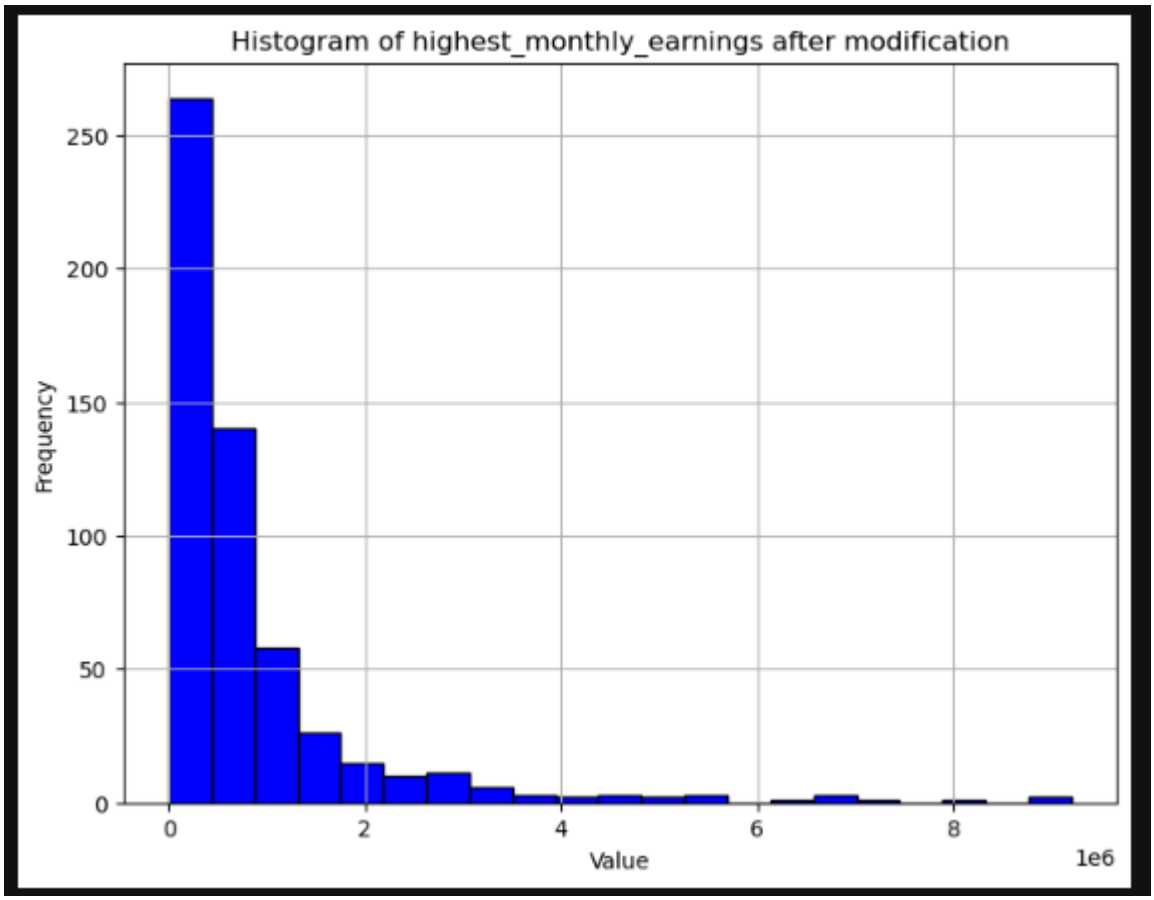
h) Eksponentinis pasiskirtymas



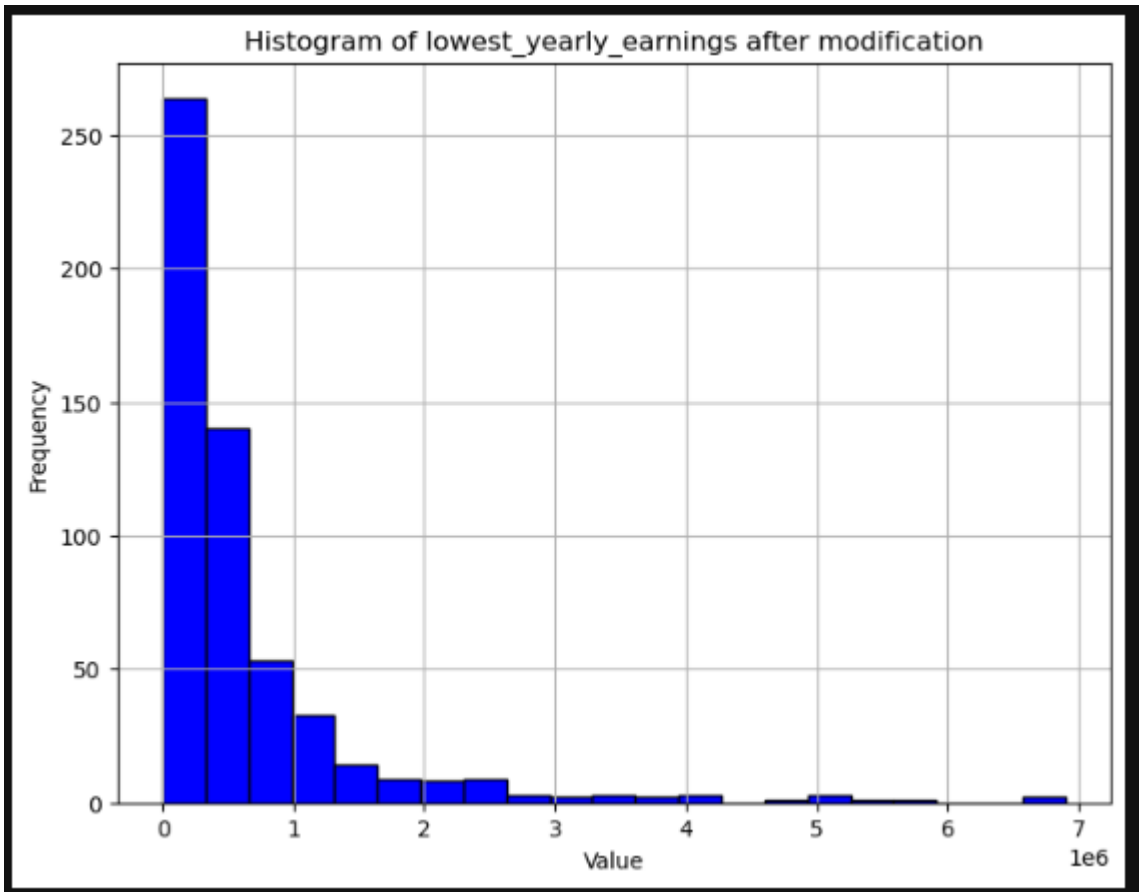
i) Eksponentinis pasiskirtymas



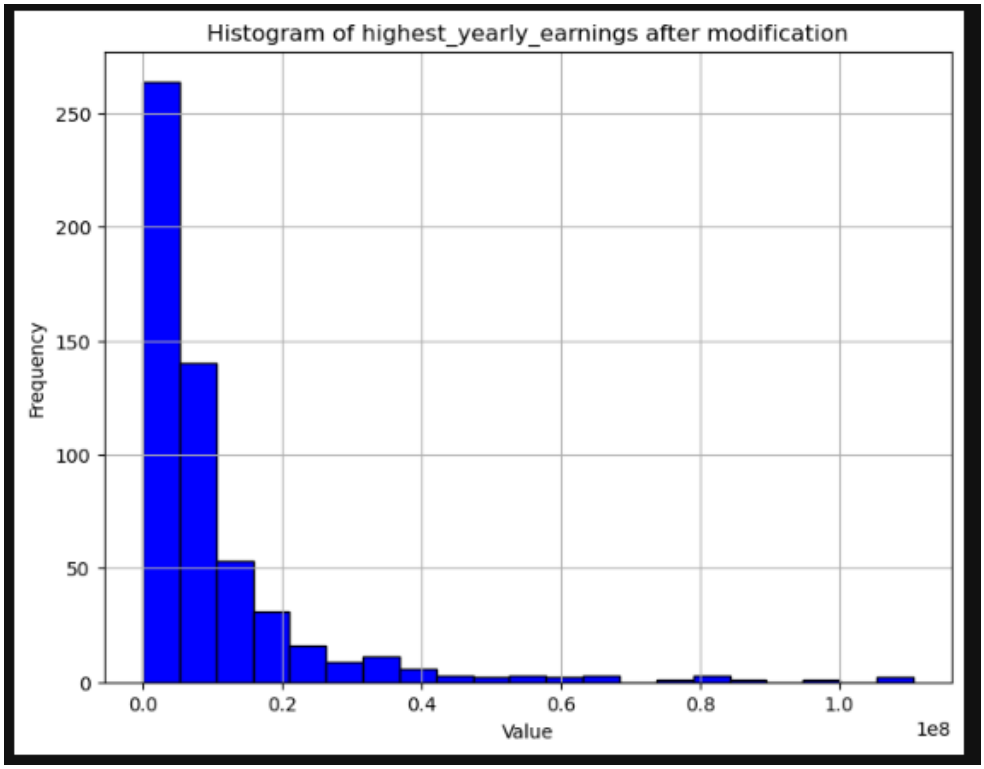
j) Eksponentinis pasiskirtymas



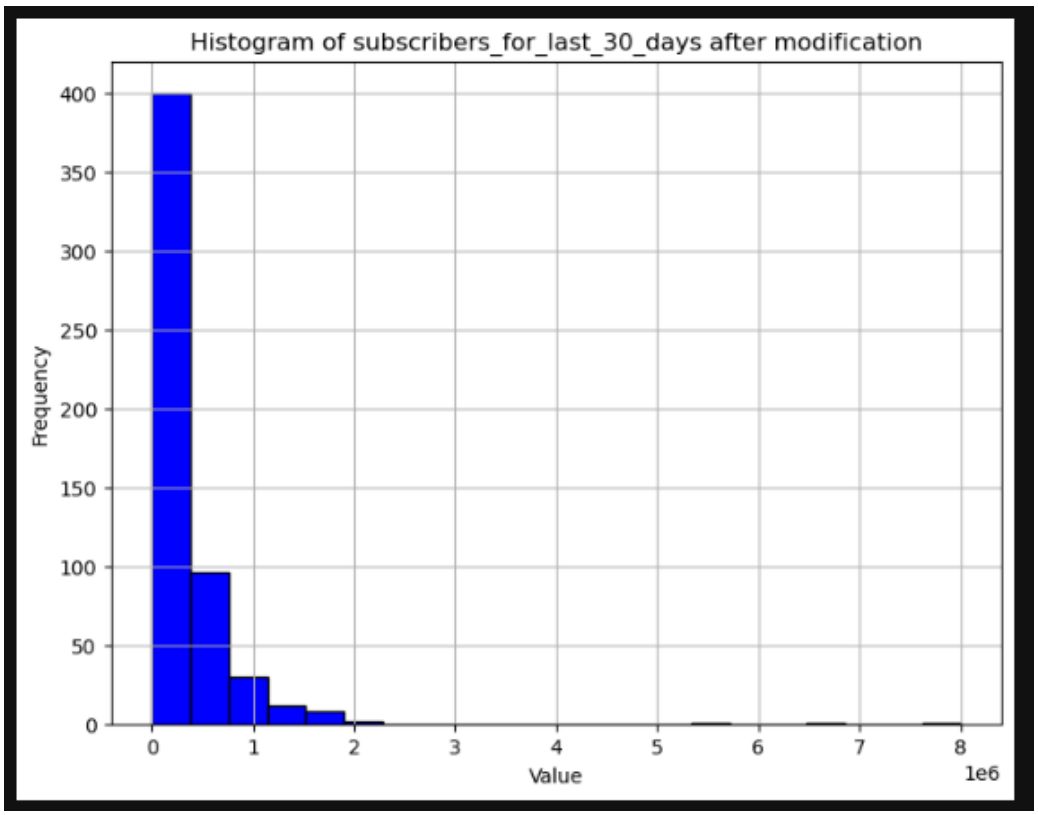
k) Eksponentinis pasiskirtymas



l) Eksponentinis pasiskirtymas



m) Eksponentinis pasiskirtymas

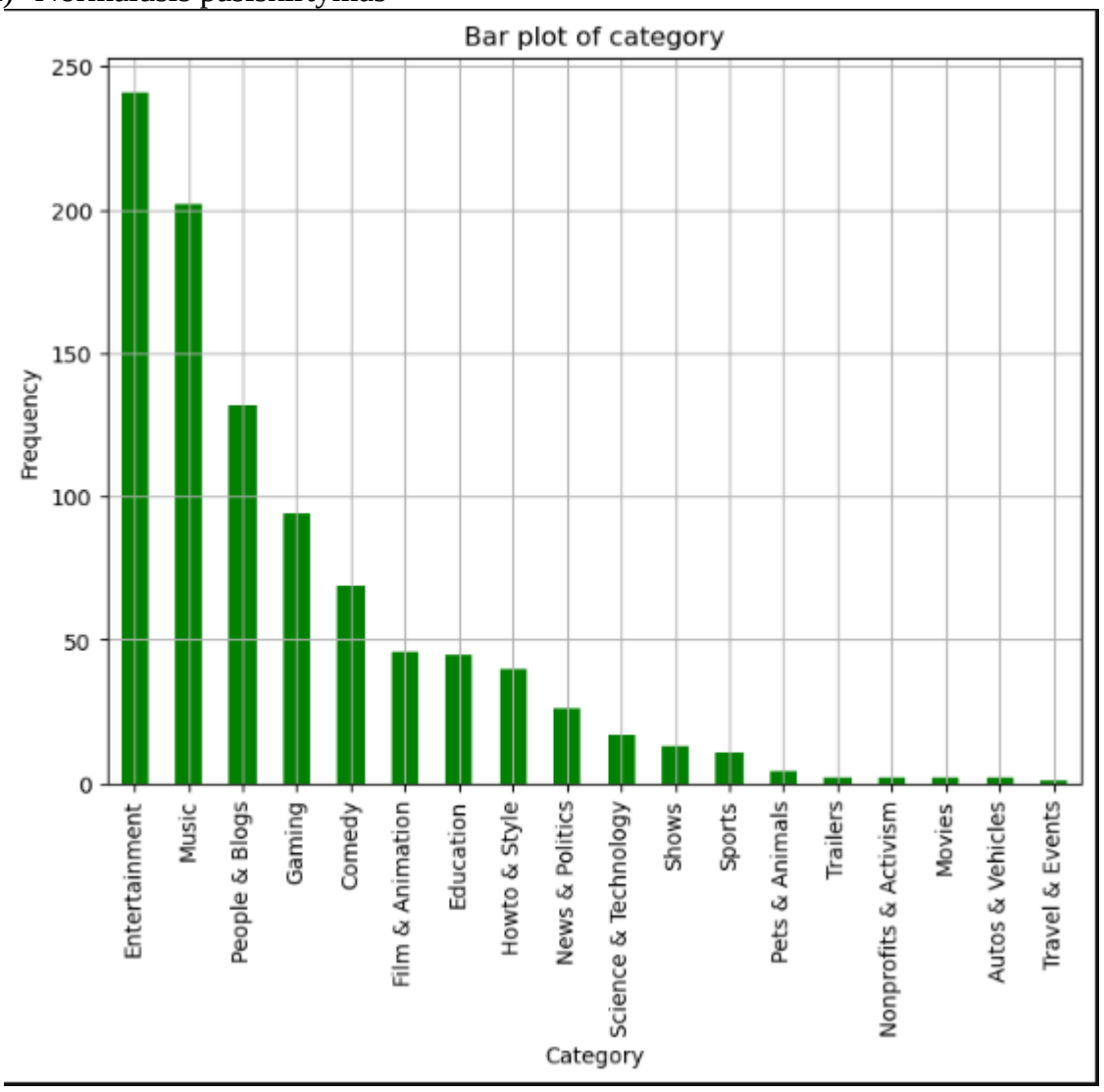


2 pav. tolydinių atributų histogramos prieš modifikavimą a), b) , c) , d) , e) , f) , g) , h) , i) , j), k), l), m)

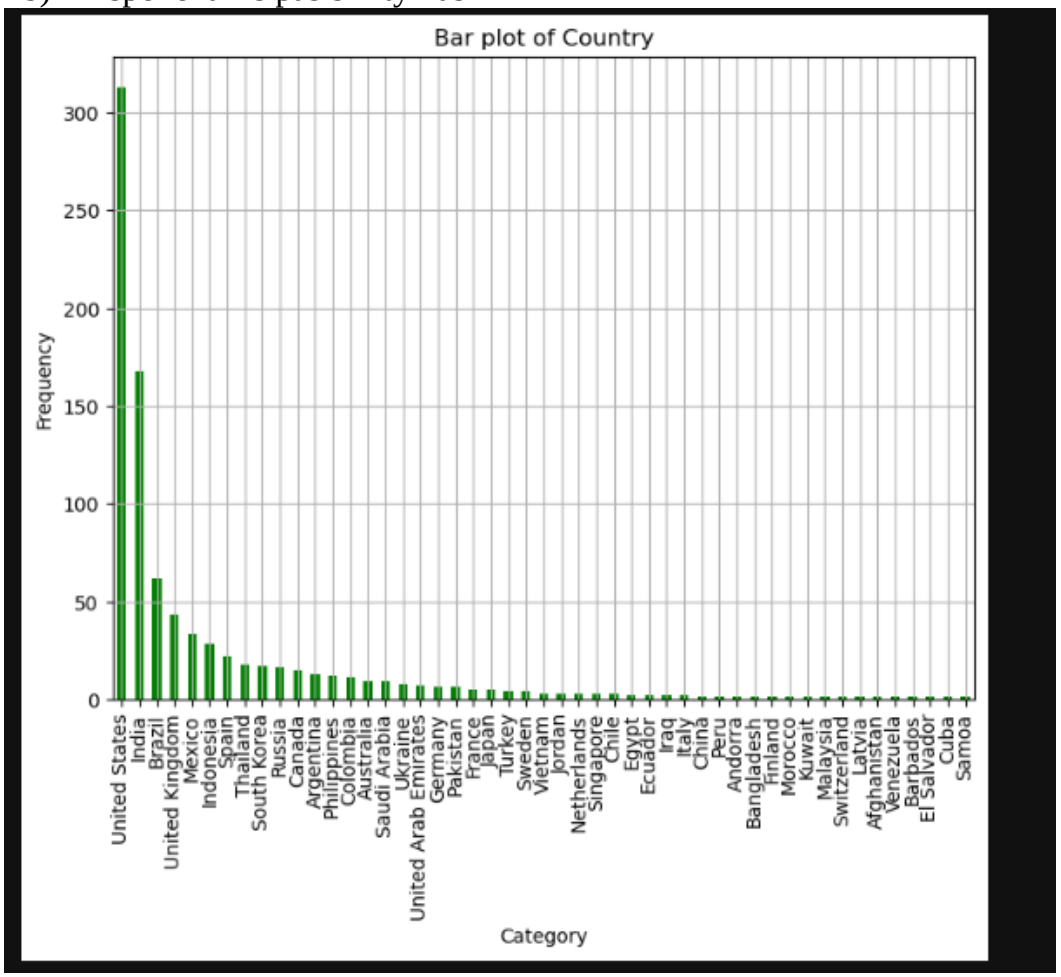
4.2. Kategorinio tipo stulpelinės diagramos

Kategorinio tipo atributams atvaizduoti bus naudojamos stulpelinės diagramos. Stulpelinė diagrama yra 3 paveiksle.

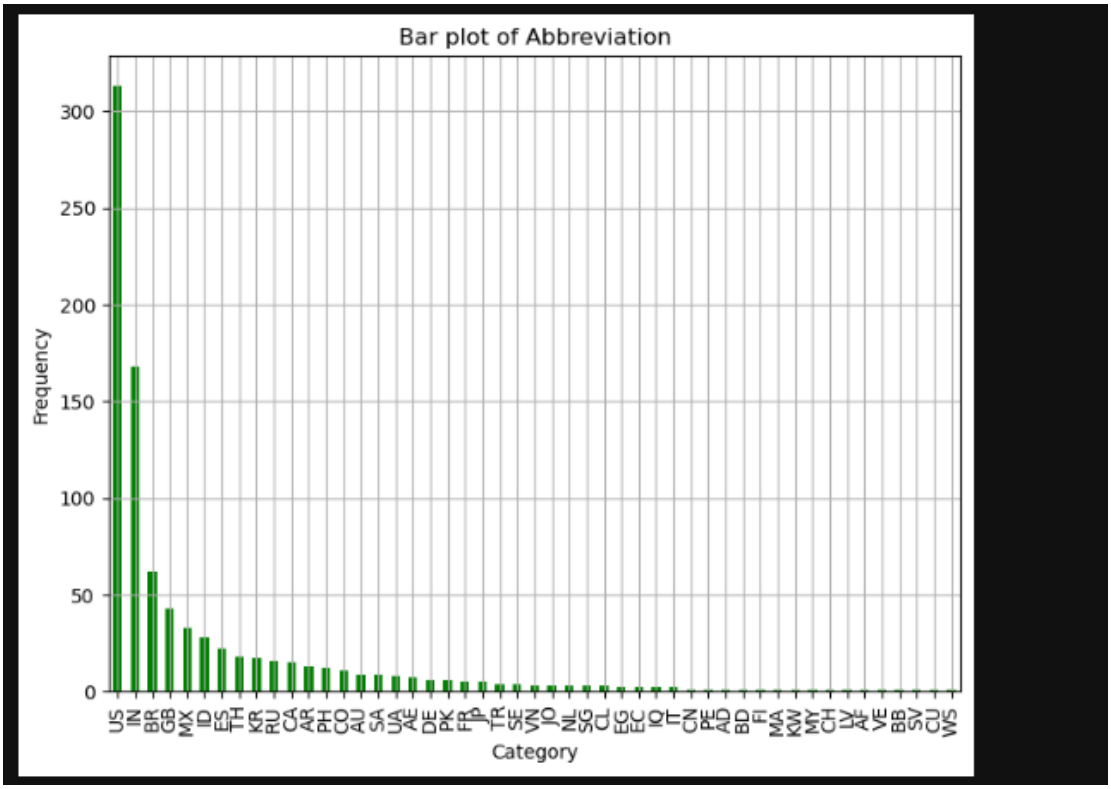
a) Normalusis pasiskirtymas



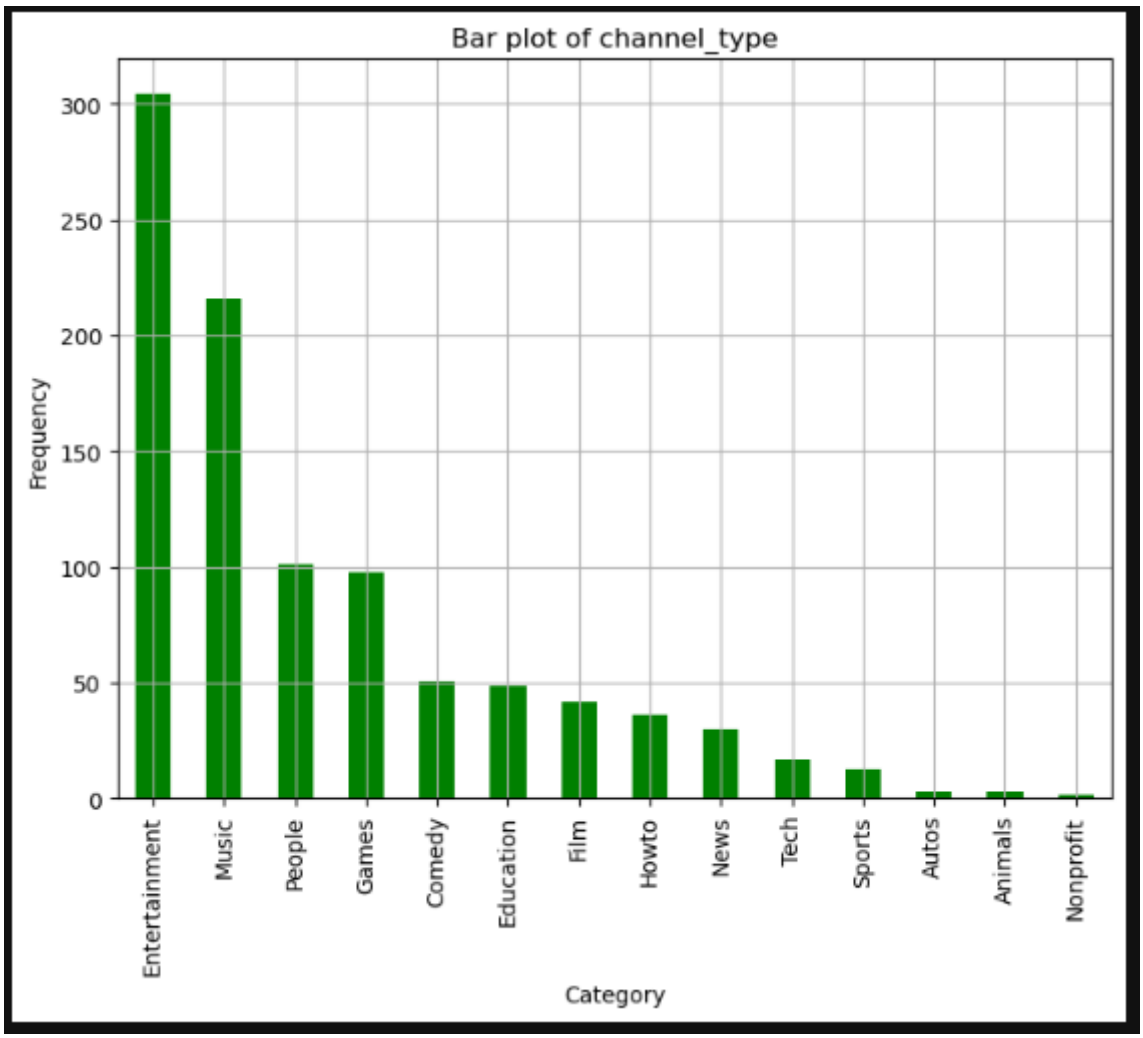
b) Eksponentinis pasiskirtymas



c) Eksponentinis



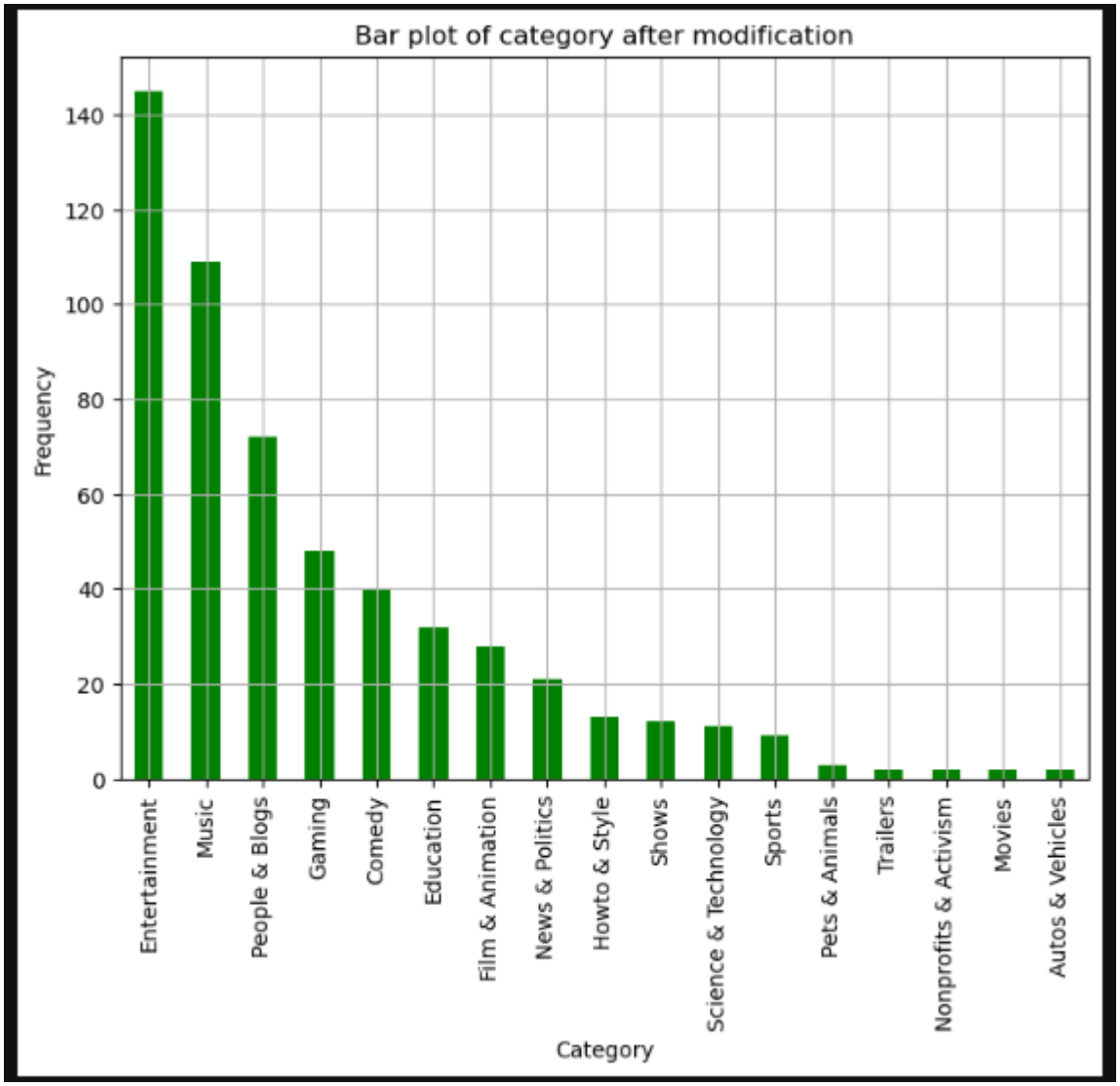
d) Eksponentinis



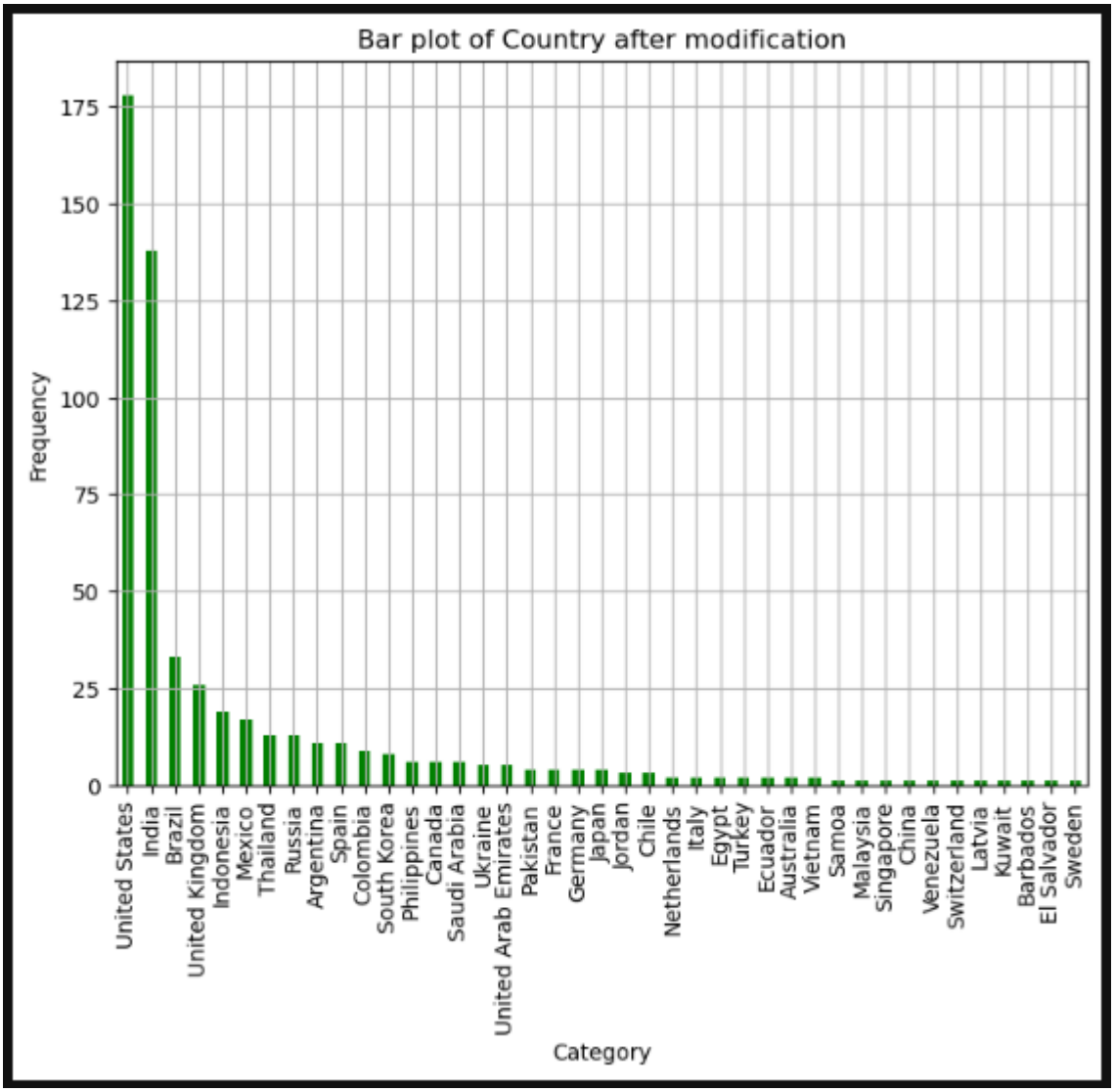
3 pav. Kategorinių atributų stulpelinės diagramos a), b), c), d)

Po duomenų modifikavimo, gauname tokius rezultatus:

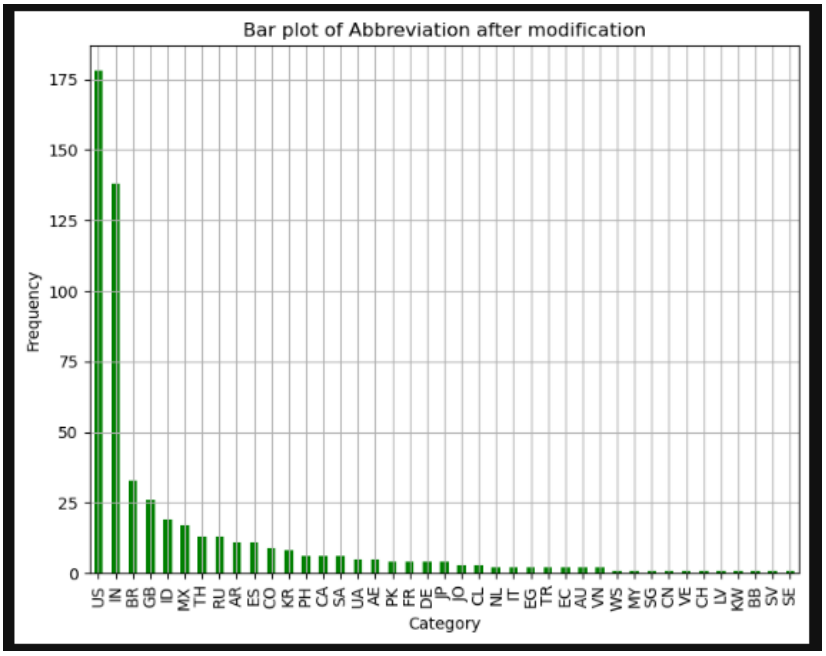
a) Normalusis pasiskirstymas



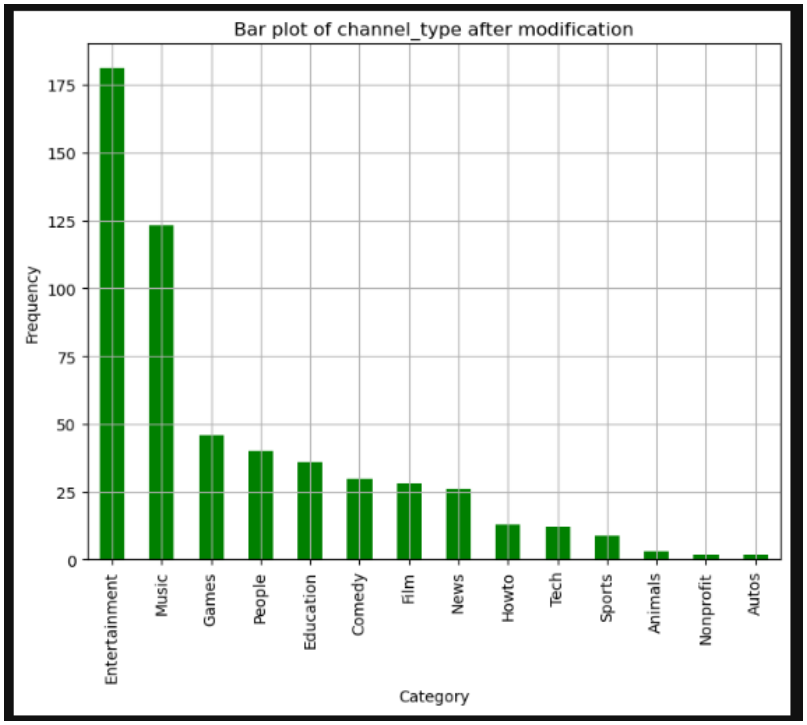
b) Eksponentinis



c) Eksponentinis



d) Eksponentinis

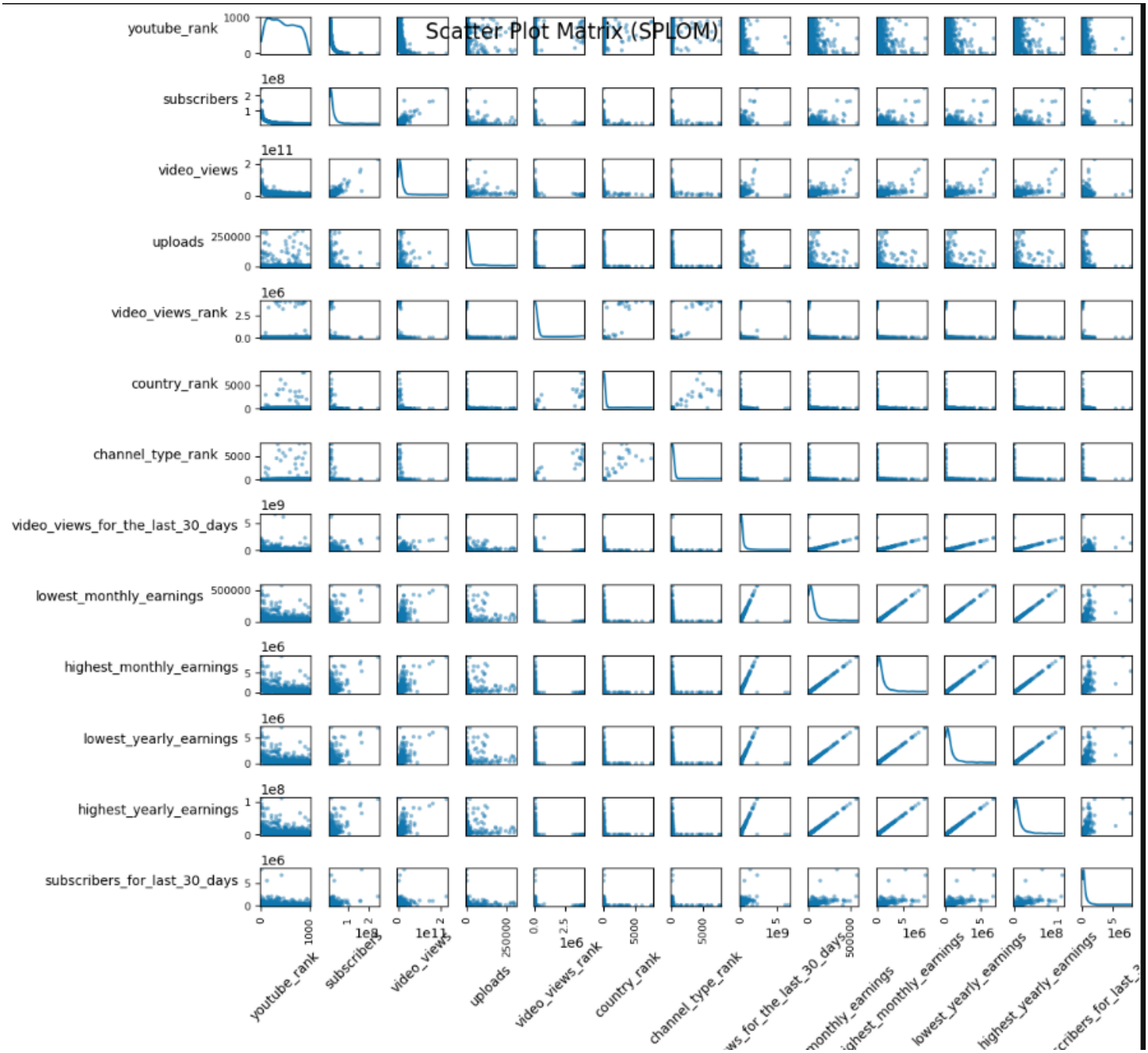


4 pav. Kategorinių atributų stulpelinės diagramos a), b), c), d)

4.3. „Scatter plot“ ir SPLOM diagrama, Box-plot, histogramos

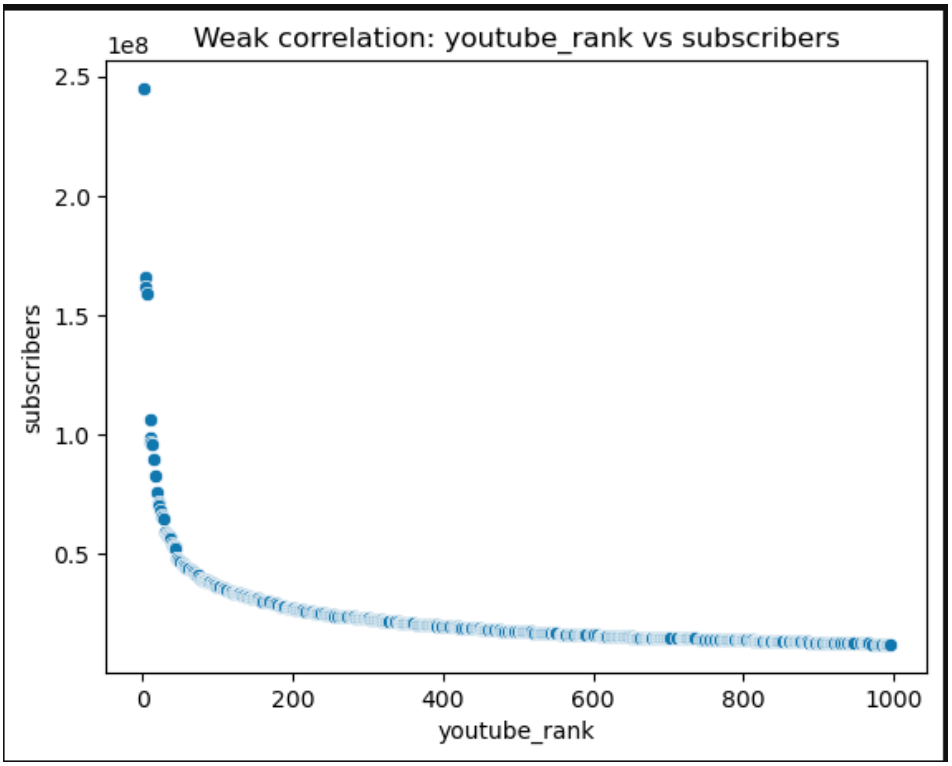
Norint sužinoti, ar yra sąryšis tarp dviejų tolydinių atributų, reikia kurti „scatter plot“ grafikus. SPLOM diagrama – visų galimų atributų „scatter plot“ grafikai – yra atvaizduoti 5 paveiksle.

Galima pastebėti, kad nėra nei vienos dviejų atributų poros „scatter plot“ grafiko, kuris turėtų pastebimą priklausomybę tarp vieno ir kito. Dėl šitos priežasties galima spėti, kad atributai nekoreliuoja ir nėra tiesioginio sąryšio tarp atributų. Tačiau reikia apskaičiuoti tikslias kovariacijos ir koreliacijos reikšmes, kad tai patvirtintume.

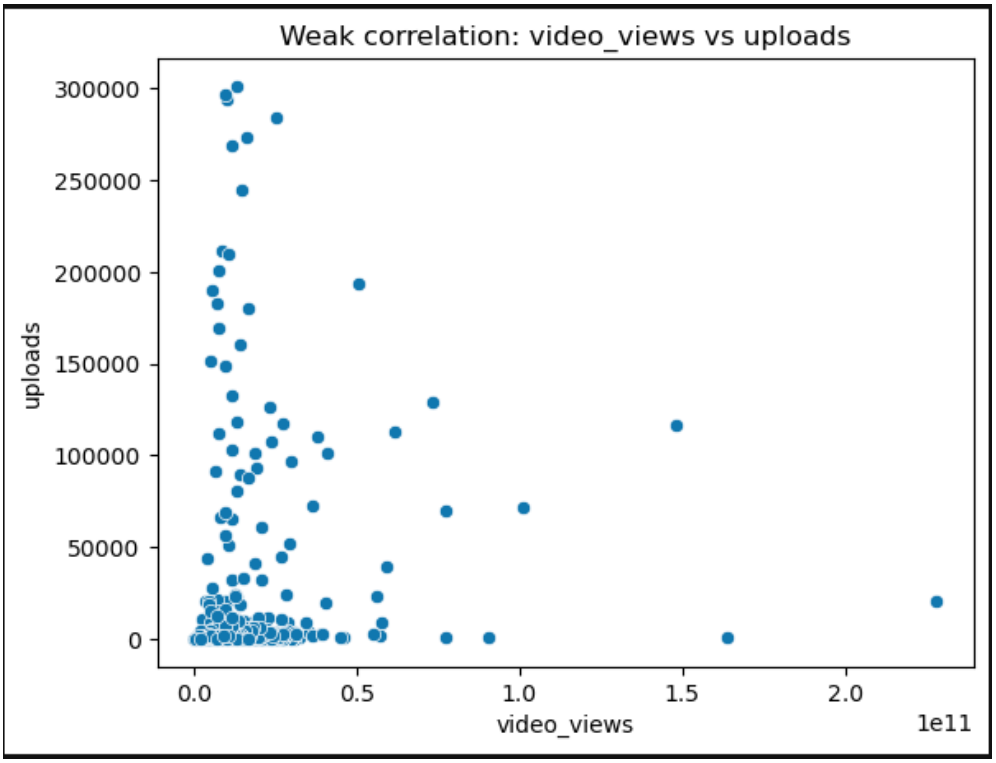


5 pav. SPLOM diagrama

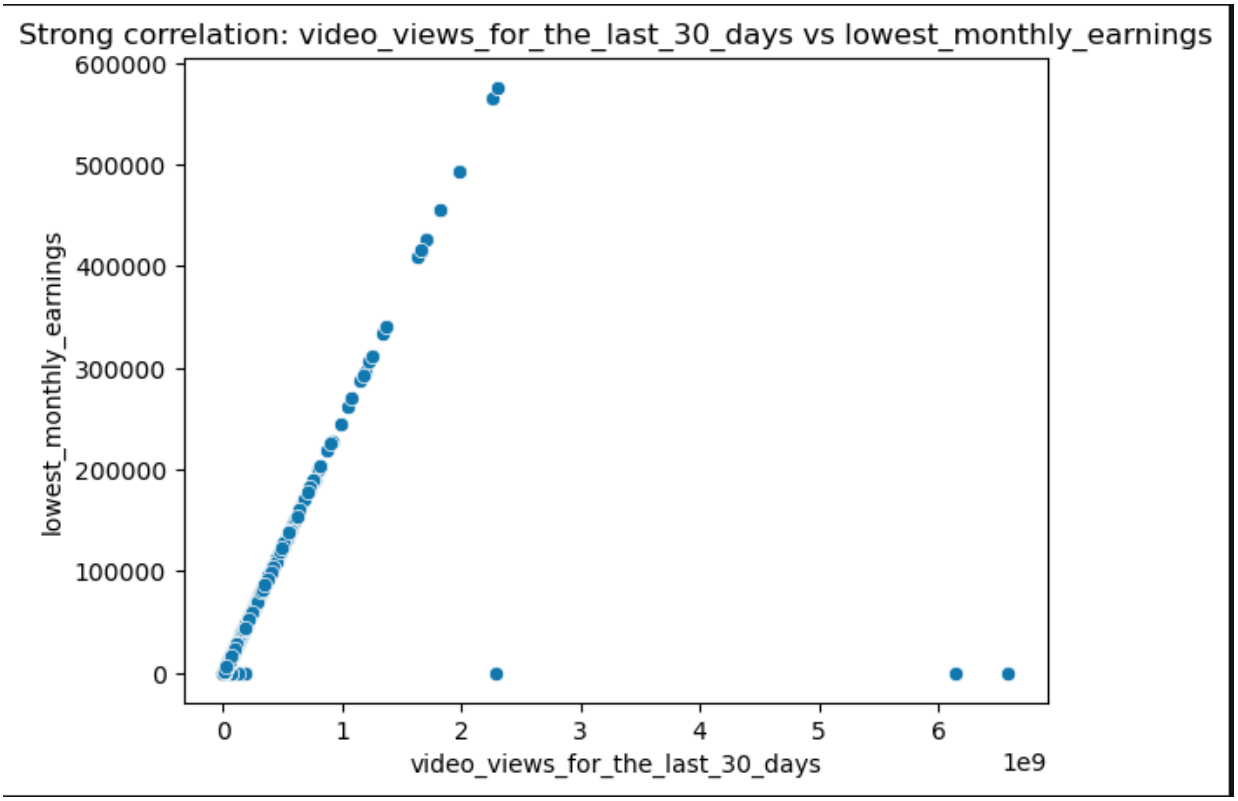
Iš šios SPLOM diagramos, galime pavaizduoti, kelis stiprios priklausomybės ryšius ir kelis silpnos priklausomybės ryšius tarp tolydžiųjų atributų.



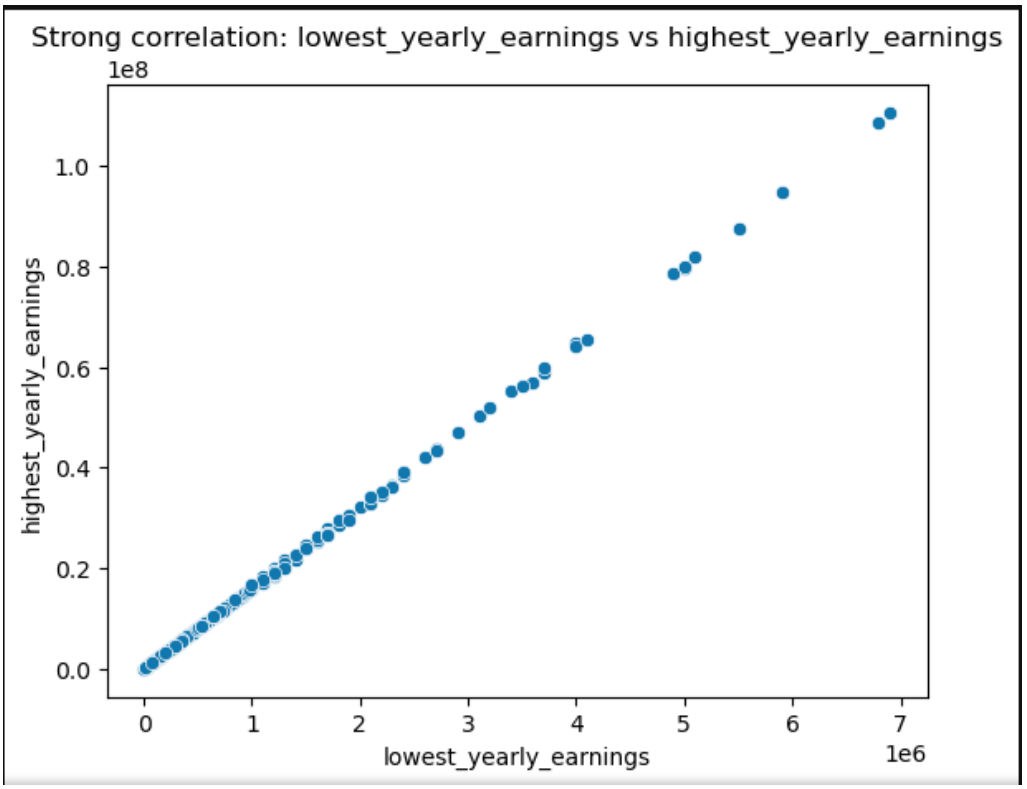
6 pav. Silpnos priklausomybės ryšys.



7 pav. Silpnos priklausomybės ryšys.

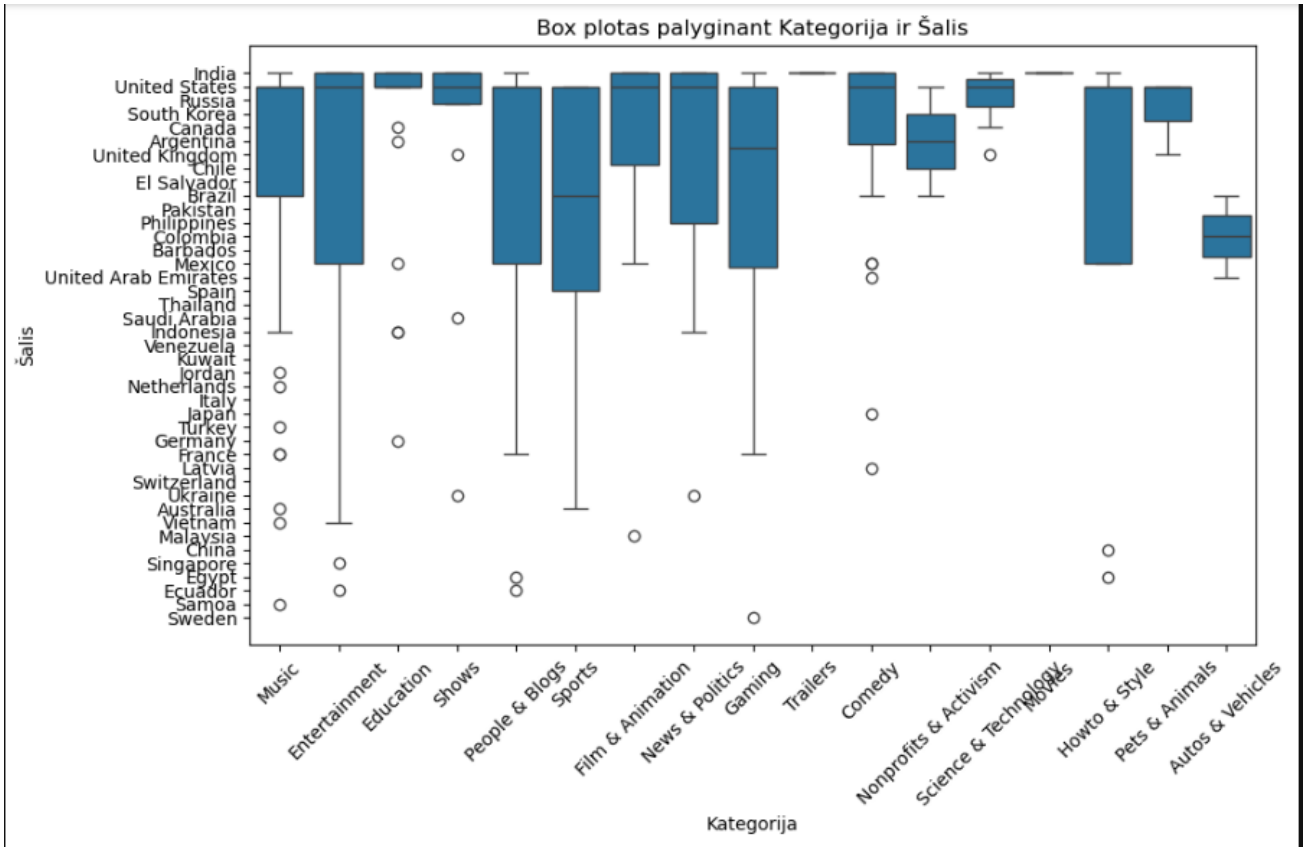


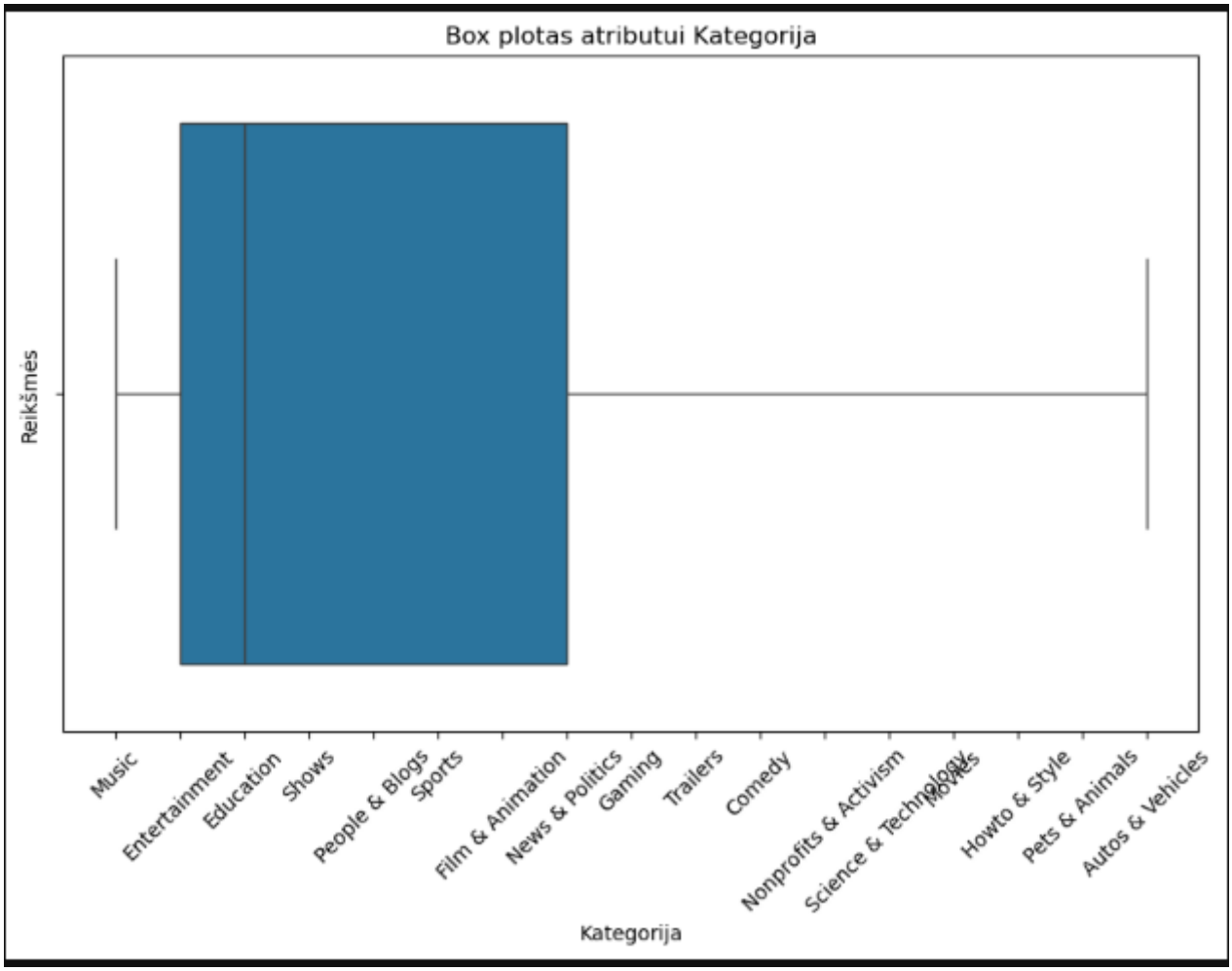
8 pav. Stiprios priklausomybės ryšys.



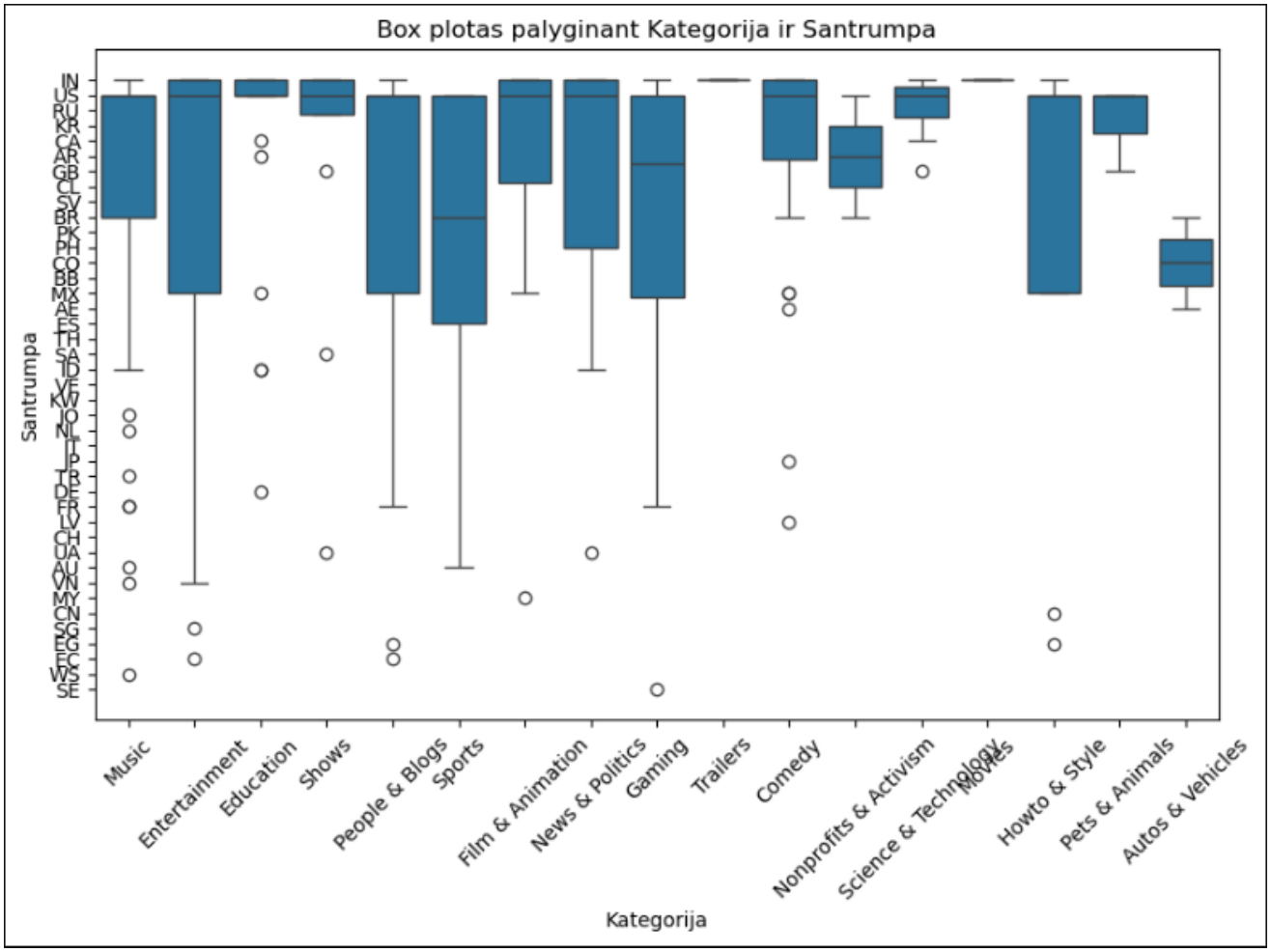
9 pav. Stiprios priklausomybės ryšys.

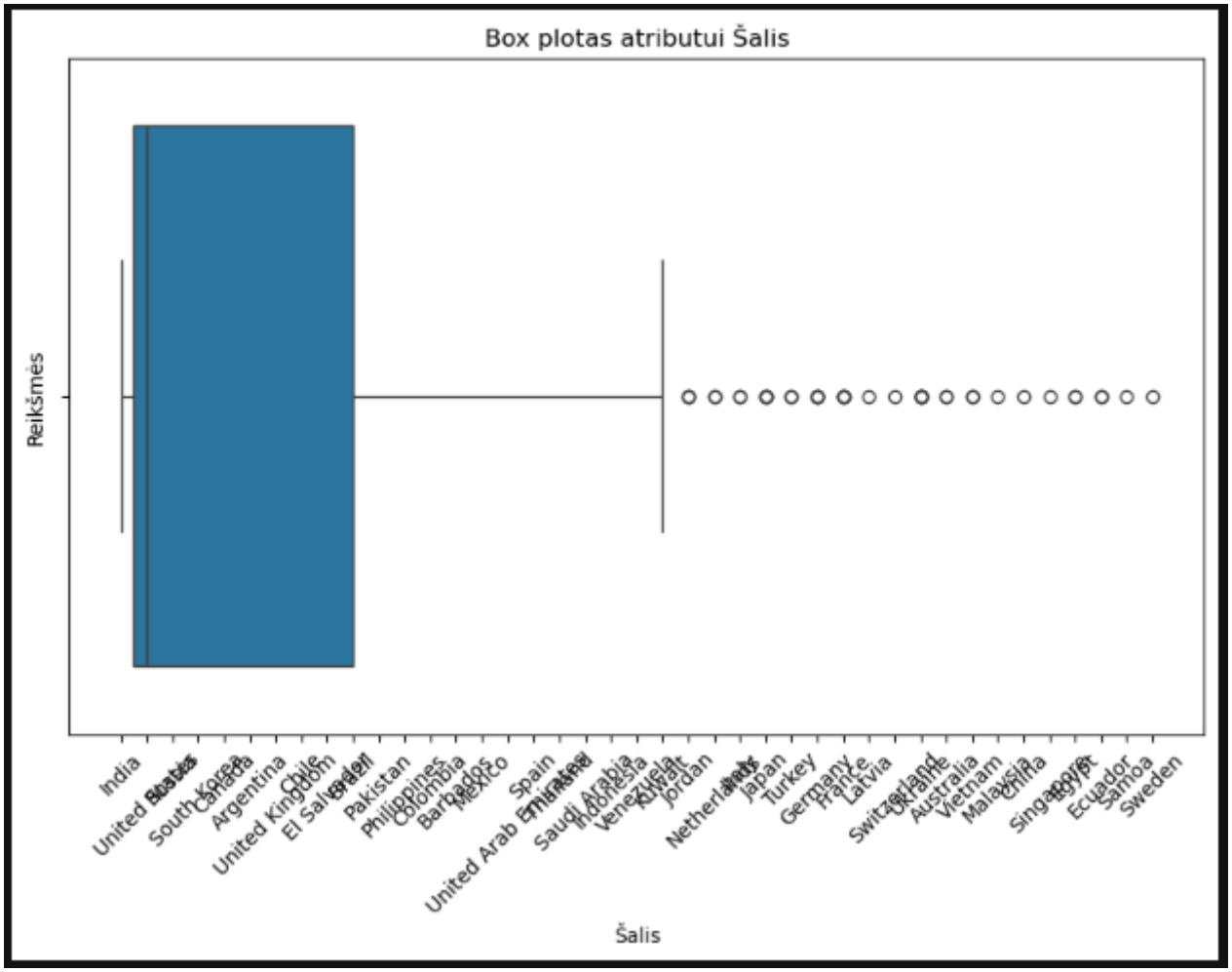
Pavaizduojame, „bar plot“ tipo diagramas, keliems iš mūsų kategorinių atributų.





9 pav. Box plot grafikas.

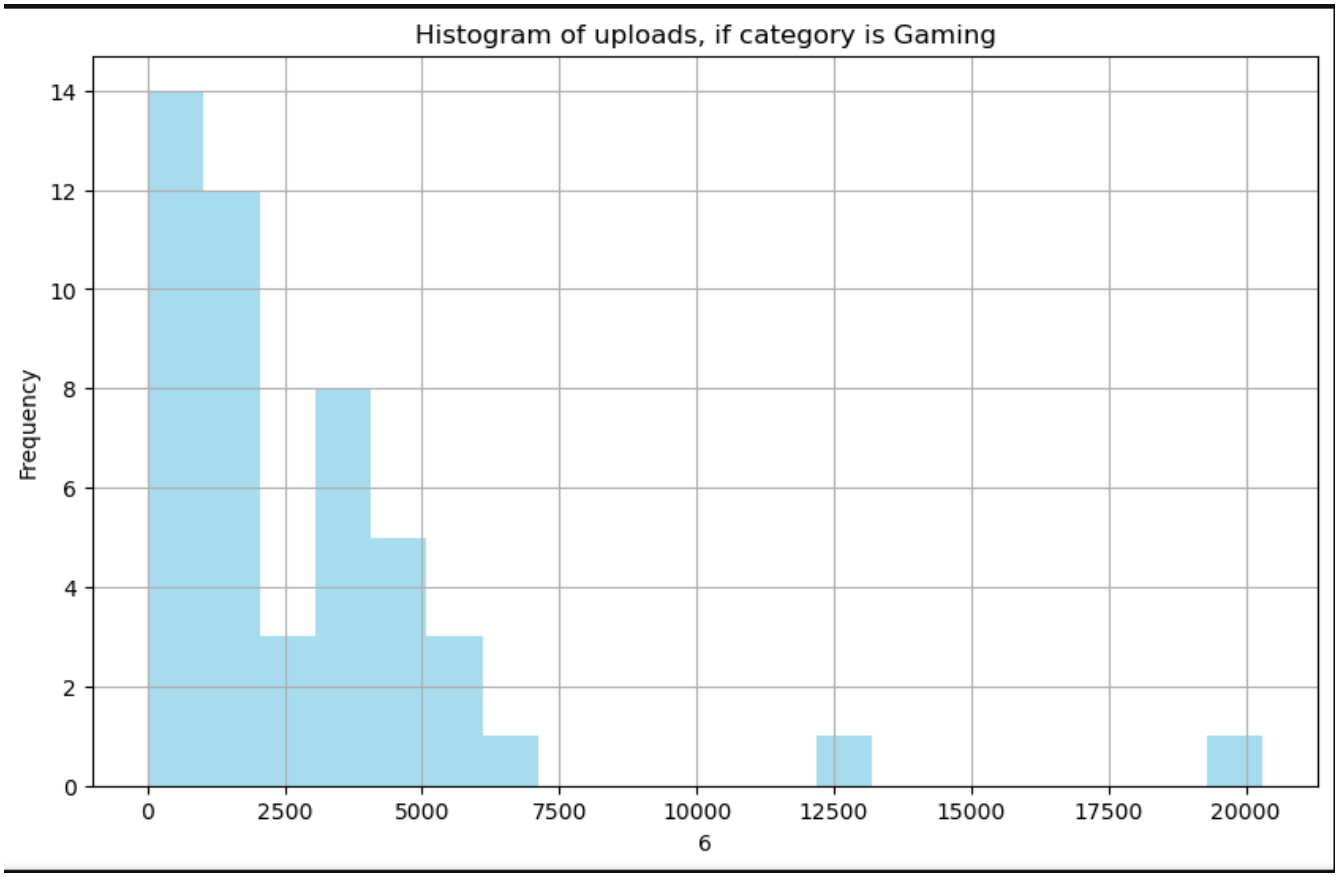

Box plotas palyginant Kategorija ir Santrumpa
Santrumpa
Kategorija
Music
Entertainment
Education
Shows
People & Blogs
Sports
Film & Animation
News & Politics
Gaming
Trailers
Comedy
Nonprofits & Activism
Science & Technology
Howto & Style
Pets & Animals
Autos & Vehicles



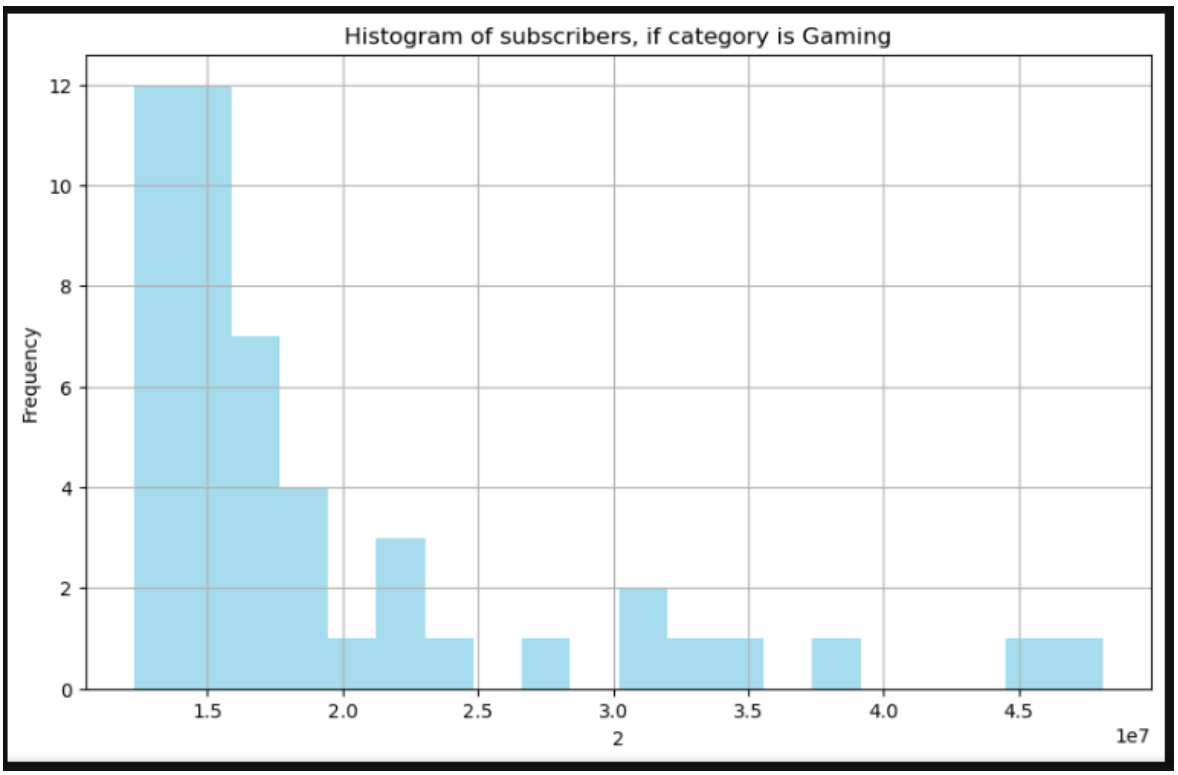
10 pav. Box plot grafikas.

Matome iš „Box plot“ grafikų, kad kategoriniai atributai, nėra tolygiai pasiskirstę, vieni turi daugiau reikšmių negu kiti, dominuoja tokios šalys, kaip Indija ir JAV.

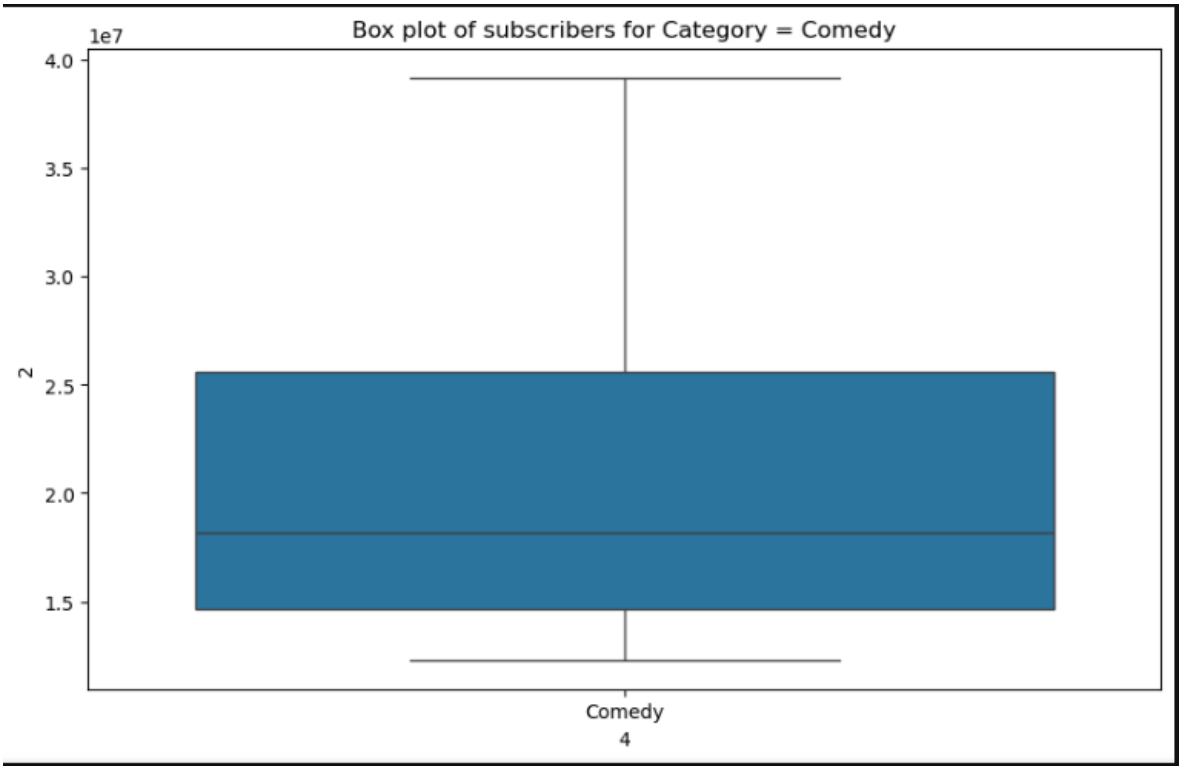
Pavaizduojame dar kelias histogramas ir „box plot“ grafikus, kurie atvaizduoja ryšį tarp tolydaus ir kategorinio atributo.



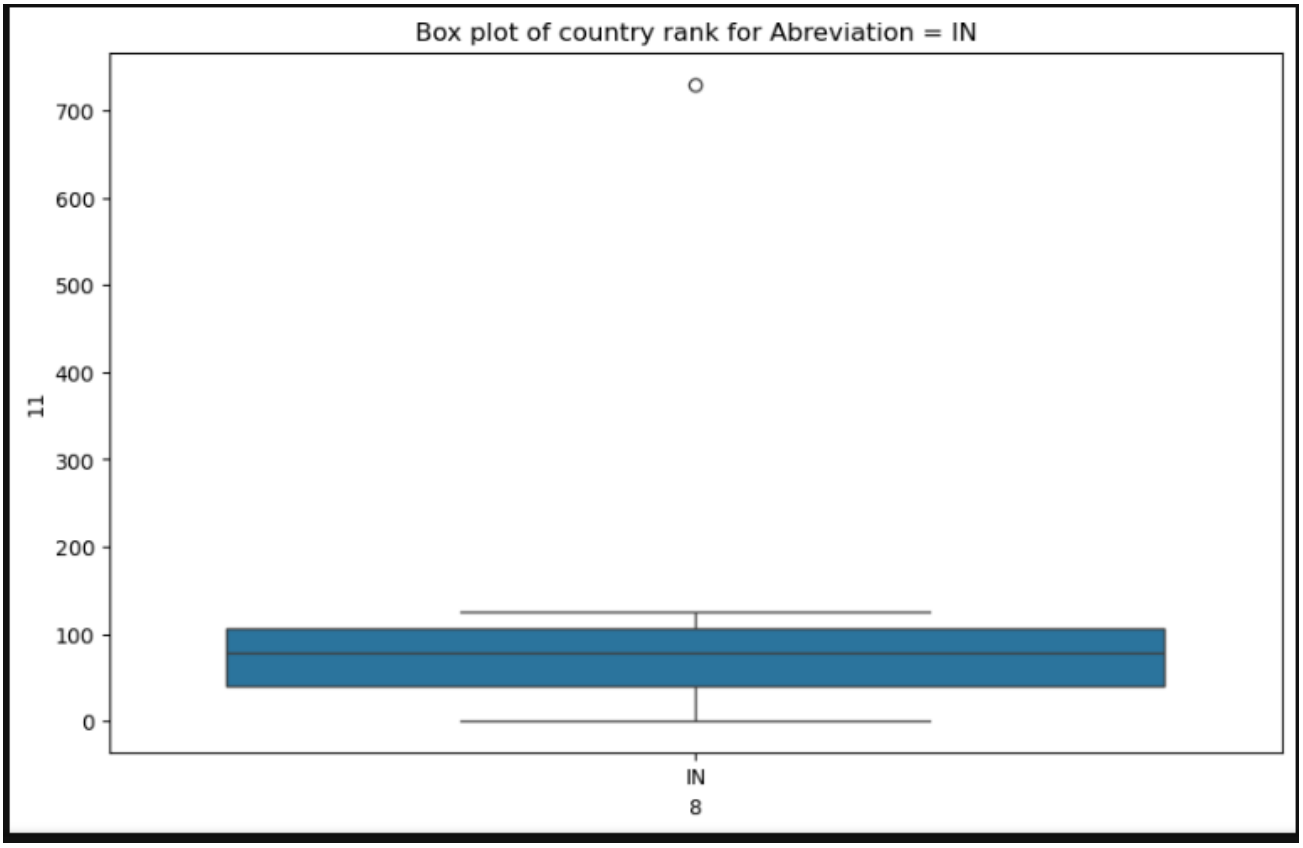
11 pav. Histograma, kuri parodo vaizdo įrašų paskelbimo skaičiaus priklausomybę, kai vaizdo įrašų kategorija yra Gaming.



12 pav. Histograma, kuri parodo Youtube sekėjų priklausomybę, kai kategorija yra Gaming.



13 pav. Box plot, kuris parodo Youtube sekėjų priklausomybę, kai kategorija yra Comedy.



14 pav. Box plot, kuris parodo šalies vertinimo priklausomybę, kai šalies sutrumpinimas yra IN.

Darome išvadą, kad priklausomybės tarp kategorinių ir tolydžių atributų yra labai silpnos arba dominuoja tik keli kategoriniai atributai, kurie turi daug tolydžių reikšmių.

5. Kovariacijos ir koreliacijos apskaičiavimas

Įmanoma tiksliai ir matematiškai parodyti, ar yra stiprus ryšys tarp dviejų atributų. Šis ryšys yra kovariacija ir koreliacija. Abiem apskaičiuoti reikalingos skirtingos formulės.

5.1. Kovariacija

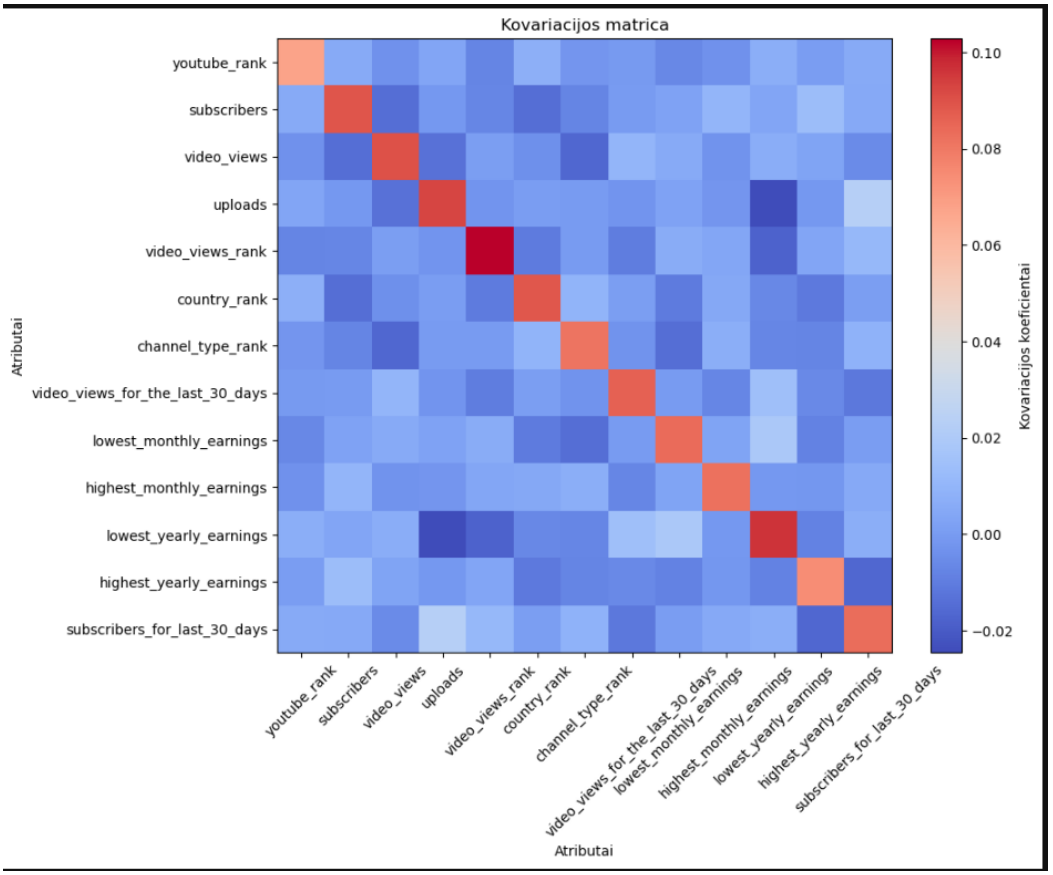
Kovariacija yra vienas iš būdų suskaičiuoti dviejų atributų ryšio stiprumą ir palyginti su kitais ryšiais. Kovariacija apskaičiuojama formule:

$$cov(a,b) = \frac{1}{n-1}\sum_{i=1}^{n}((a_i - \bar{a}) \bullet (b_i - \bar{b}))$$

Čia a – pirmasis atributas, b – antrasis atributas, $\bar{a}$ – pirmojo atributo vidurkis, $\bar{b}$ – antrojo atributo vidurkis, n – duomenų rinkinio įrašų kiekis.

Taikant minėtą formulę įmanoma apskaičiuoti kovariacijas tarp kiekvieno atributo. Rezultatai pateikti 15 paveikslėlyje.

Pagal gautas kovariacijos reikšmes įmanoma atspėti, kurios atributų poros ryšys yra stipresnis už kitus. Tačiau matome, kad tarp mūsų tolydžiųjų atributų, nėra labai didelis kovariacijos koeficientas, nebent tarp savęs.



15 pav. Kovariacijos matrica.

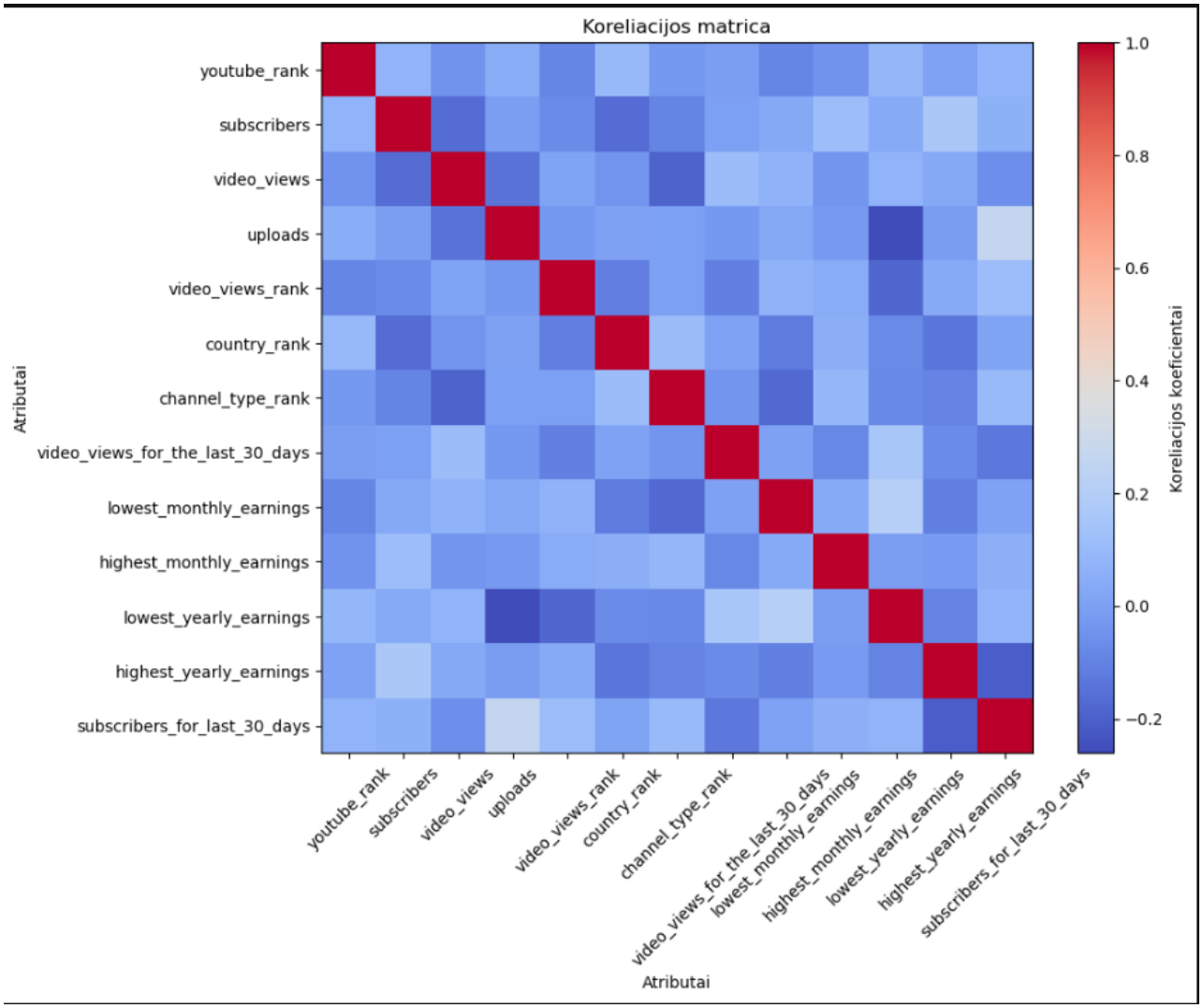
5.2. Koreliacija

Koreliacija, taip pat kaip kovariacija, skirta suskaičiuoti dviejų atributų ryšio stiprumą, tačiau yra normalizuota. Koreliacijos rezultatas yra patogesnis, kadangi šis yra intervale [-1;1]. Koreliacijos suskaičiavimo formulė:

$$corr(a,b) = \frac{cov(a,b)}{sd(a) \bullet sd(b)}$$

Čia *a* – pirmasis atributas, *b* – antrasis atributas, *cov*(*a*,*b*) – pirmojo ir antrojo atributo kovariacija, *sd*(*a*/*b*) – pirmojo/antrojo atributo standartinis nuokrypis.

Su formule įmanoma apskaičiuoti koreliacijos matricą. Koreliacijos matrica yra pateikta 16 paveiksle. Tarp atributų koreliacija yra labai silpna, tarp 0.2 arba -0.2. Tai darome išvadą, kad ryšiai tarp atributų yra labai silpni, t.y. nekoreliuoja.



16 pav. Koreliacijos matrica.

6. Duomenų normalizacija

Dažnai pasitaiko didelių reikšmių duomenys, kurių analizę ir supratimą gali palengvinti duomenų normalizacija. Savo duomenų rinkinio normalizavimui buvo naudojama formulė:

$$z = \frac{x - \min(X)}{\max(X) - \min(X)}$$

Čia X – duomenų aibė, x – iš duomenų aibės X išrinkta reikšmė, z – normalizuota x reikšmė.

Panaudojus formulę su kiekviena duomenų reikšme buvo sukurtas duomenų rinkinys, kurios reikšmės yra intervale [0;1].

9 dalis
Table 1:

	youtube_rank	subscribers	video_views	uploads
0	0.5424426318353484	0.6043051123735979	0.5427626248387982	0.441896288514817
1	0.5134419755842668	0.262593420593866266	0.7735478825792041	0.018331897106155348
2	0.6546101765103708	0.4664388217079326	0.2404254569160365	0.663128511789185
3	0.5615733104078844	0.5239659886023397	0.0889989839586712	0.32152329895767684
4	0.7044814501983029	0.817525932354298	0.3934569757632504	0.8910899756388424
5	0.916424451336336	0.4356775465501762	0.8919737811155186	0.100695945328525855
6	0.7372958901594224	0.5217316301224786	0.049148719463265184	0.8020602244874048
7	0.19678287380113071	0.053927296530114306	0.4316050201479594	0.3813798377008863
8	0.1518505147672478	0.6899077865503251	0.21132663980670371	0.25616036580665774
9	0.7752978303882909	0.908379468848297	0.0784151105653194	0.9722138840517425
10	0.2673407912676114	0.6195750064337493	0.007790471941795459 5	0.9923333284286109
11	0.7325678437210796	0.7738611201093162	0.14233670984899738	0.6790256430080589
12	0.2312069380068863	0.5889842708665468	0.2689994065515405	0.4644184351859218
13	0.9008243378101965	0.7269671552899222	0.3959923183980213	0.7069137550244788
14	0.23041268486853988	0.028946609610434253	0.9594526522210763	0.015842583519859465
15	1.0	0.40858894889899827	0.15844463060173306	1.0
16	0.9943165893890281	0.294925609342437	0.9926877907753284	0.9611050933267146
17	0.7045844654554785	1.0	0.4491189583288364	0.15341711096821906
18	0.3108841585068224	0.8665493479041622	0.043465696748174604	0.10514704135770558
19	0.6214021309660807	0.2072199560261507	0.11089784790782023	0.6789339442314679
20	0.617740221269325	0.986036157236964	0.8717101856510193	0.5468713061841927
21	0.017898857566916126	0.7428462616130443	0.8981511871590325	0.3072399598521175
22	0.5878425185955797	0.396669376839603	1.0	0.6441682294702045
23	0.361246679327648	0.687327490367245	0.6553412819675418	0.5027710070369791
24	0.6131172817727691	0.8262240418325607	0.6519795855658509	0.11106284264534458
25	0.35359005468799604	0.0766632754791255	0.00831541208186475	0.26738474570082654
26	0.768798885729008	0.501551266350684	0.37321454649279767	0.5009439207366799
27	0.3521949274128385	0.2455480469154945	0.31469589535381435	0.448976943753215
28	0.06234627944649663	0.30264721357196317	0.0703013020882867	0.5748135771928746
29	0.26255453027124975	0.10888903366474556	0.20823409820065833	0.7530438290162628
30	0.40187209040257493	0.10361841672617443	0.8169229598229177	0.7409640521612582
31	0.8916992938066924	0.9035715203991126	0.18114925271540536	0.31947384089333547
32	0.8844201040032074	0.40088379380754646	0.41376926977961725	0.213189910527220 67
33	0.5487143830296856	0.7283876182743415	0.03326925736371686	0.9127580097079999
34	0.12681537244778304	0.17695495849114304	0.3421378493013523	0.8626584728198154

17 pav. Normalizuoti duomenys.

Table 2:

	video_views_rank	country_rank	channel_type_rank
0	0.7944216611411867	0.5297317350569563	0.5744051370584508
1	0.9473613247333857	0.6829284185379035	0.3616030100708525
2	0.8238663255145182	0.09719463667219648	0.8498268997150265
3	0.2890739831397714	0.18343295122734185	0.5932512679929157
4	0.9597712581974536	0.6450251834877971	0.42726633239753686
5	0.14827231914700456	0.8695469546610992	0.1695582456300428
6	0.70658458536467	0.03181992289233005	0.1628044918517579
7	0.6814120889919509	0.45440418518502407	0.5422961861115978
8	0.20664393228787592	0.42534285812247896	0.37656499206145755
9	0.014451607770963198	0.9310559370978193	0.6783608402913233
10	0.36872990329135585	0.137062157750358	0.8336751920146124
11	0.8636627841805278	0.7282222479419815	0.27059899024731197
12	0.519886381875666	0.3072645332542433	0.5840978074428292
13	0.23939039169589624	0.16074140897587041	0.807422721335235
14	0.09813847430111394	0.9468553100998849	0.8820149247433452
15	0.2407971697263582	0.6635714334995648	0.24583786493372053
16	0.7330378264865625	0.8831645157949339	0.27275103056913397
17	0.3830391166518895	0.896853717054161	0.9823318469677716
18	0.624060012154385	0.7116779512463408	0.20385688824396936
19	0.6919352187331378	0.5896794647710354	0.9772047255264622
20	0.9232515583775912	0.036922033471103734	0.1730884745951279
21	0.3784487249986383	0.9811396346189478	0.17297253183956543
22	0.767094817680984	0.6993768829914392	0.33710217750392697
23	0.862686696120428	0.058561880270172534	0.47498874948313224
24	0.028053826137793989	0.738449807184377	0.10677165527444818
25	0.16102060877668744	0.5695722384597848	0.7840176534167249
26	0.11601095454013935	0.16004667922779833	0.042505380399021936
27	0.995793849148979	0.37731610687930106	0.9806509279772476
28	0.9796908774328213	0.9340930444108491	0.39455249696540634
29	0.4258896014268961	0.06358937423969022	0.20725455725144146
30	0.8923434572712429	0.7384741862052732	0.0
31	0.9184105037095505	0.7329302629848122	0.7371700740704377
32	0.11876736269200008	0.33497589595958704	0.17370081385197095
33	0.476379598112876	0.8213605092240314	0.2990282018750964
34	0.11009123865286866	0.45968627979322907	0.32366827286225913
35	0.010167458655249744	0.3233115532334665	0.2290047162052923 4
36	0.5901111157677683	0.754463208371678	0.23539967805645123
37	0.6539794968096321	0.854641253198823	0.47978884220332485
38	0.6936004727522431	0.10901742098618844	0.2648044495722356
39	0.2654318214351791	0.8294986746314331	1.0

18 pav. Normalizuoti duomenys.

Table 3:

	video_views_for_the_last_30_days	lowest_monthly_earnings	highest_monthly_earnings
0	0.9483439224836451	0.06770106090903663	0.08337519024677519
1	0.4384785978099338	0.6991965475435975	0.056164102496524156
2	0.08268089365233368	0.980204205330744	0.469253719784753
3	0.0033769517057602757	0.8315322944796896	0.0
4	0.6152237386793259	0.015452758835570906	0.30026045935790483
5	0.6247897270666525	0.12089778743584369	0.8529426939603044
6	0.6309666161349142	0.5778523197873846	0.23586026783743566
7	0.918157478361783	0.9941934487735181	0.2146246756921355
8	0.4661793213116038	0.2759093871303382	0.5887358483387218
9	0.8017770739873014	0.2800428619874311	0.5883573952346897
10	0.18051772635139707	0.5114272611727694	0.22212945100924003
11	0.11960798801816408	0.05191683897160968	0.30029354914413636
12	0.983655851230043	0.6467284482807506	0.03101715111993156
13	0.9833774840897238	0.4578312052512514	0.5929836256953483
14	0.45635590088645533	0.3253653363787923	0.23065278189586896
15	0.6772822605543576	0.517463637573496	0.42418264260171223
16	0.37797592013683146	0.3733328616686139	0.7525897113777247
17	0.5531209588111797	0.27308231285786716	0.5942441103268798
18	0.33898835571190816	0.6776405613549487	0.8845258091047518
19	0.0	0.6980387982863442	0.8182210986621706
20	0.3884930763226032	0.9556984811414152	0.29870590927961777
21	0.3246805026769069	0.6817788579062823	0.05918030287844332
22	0.13651718222440515	0.05923531231354049	0.23991492882236973
23	0.10327692742109525	0.4567426925074773	0.9864035986765098
24	0.615134948197119	0.7048265362963778	0.6372859680641826
25	0.4587010422736579	0.15067777248473957	0.19712627187715703
26	0.9954465323726744	0.0	0.17587011032256591
27	0.8087970031266908	0.6770828993138288	0.2429368219865077
28	0.23512994276622987	0.24846593493688882	0.48416439871227224
29	0.9554375746077315	0.21319220541747247	0.8633900566652118
30	0.7068147739754257	0.9220879632533852	0.7138194876150584
31	0.2849459617148759	0.5783369159230423	0.7833279294713149
32	0.1033442056272478	0.9030136976037054	0.052777656168773356
33	0.13990816645368487	0.3289593797649897	0.8184248917570267
34	0.31269242122566776	0.4824677974714311	0.7334125745953243
35	0.5113541228070975	0.738725252538083	0.09484265686685838
36	0.6299454538242736	0.6407339298577852	0.9546227596748463
37	0.9938544488576111	0.2638194112376967	0.008913870386262 8
38	1.0	0.6405732142290 45	0.5218744106572912
39	0.7999442370633373	0.5191581903366199	0.06207968313499523
40	0.32724000258168073	0.6394921566089325	0.27931758530708856
41	0.14517106526471962	0.42544343947407964	0.8981520541082268
42	0.8870474140363472	0.8186429977303834	0.917109803036646

19 pav. Normalizuoti duomenys.

7. Kategorinių kintamųjų keitimas tolydžiaisiais.

- Mapiname kiekvieną, unikalią kategorijos reikšmę ir priskiriame unikalią skaitinę vertę.

Gauname tokius rezultatus:

10 dalis

Category	Country	Abbreviation	Channel-type
0	0	0	0
1	1	1	1
2	1	1	2
3	0	0	1
4	2	2	3
1	1	1	1
0	0	0	0
5	1	1	4
4	3	3	0
3	0	0	1
0	3	3	0
0	4	4	0
0	3	3	0
1	0	0	1
2	1	1	2
2	0	0	2
0	0	0	0
3	0	0	1
5	1	1	4
6	1	1	5
0	5	5	0
0	0	0	0
7	0	0	6
0	0	0	0
0	1	1	0
1	1	1	0
6	0	0	0
0	6	6	0
2	0	0	2

20 pav. Kategorinių duomenų keitimas tolydžiais.

8. Išvados

Atlikus uždavinius suprasta, kad duomenų rinkinys, kuris buvo pasirinktas, nebuvo pats geriausias. Jame trūksta nemažos dalies reikšmių. Buvo pasirinkta per daug tolydžių reikšmių, kurios buvo analizuojamos. Modifikuojus duomenys buvo gautas žymiai mažesnis duomenų rinkinys, apie (40-50%) dingo po pašalinimo.

- Duomenų rinkinio nereikalingos lentelės;
- Per mažas duomenų kiekis;
- Pašalinimas nebuvo geriausias pasirinkimas.