KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS

Intelektikos pagrindai (P176B101)

Laboratorinis darbas Nr.1

Atliko:

IFF-7/14 gr. studentas

Eigijus Kiudys

Priėmė:

lekt. Nečiūnas Audrius

doc. Paulauskaitė- Tarasevičienė Agnė

KAUNAS 2020

Turinys

Įvadas

Laboratorinis darbas Nr.1. Duomenų apdorojimas rinkinio analizė

Laboratorinio darbo užduotis

Pasirinkti (susikurti) duomenų rinkinį, su kuriuo atliksite šį ei sekančius laboratorinius darbus. Jūsų pasirinkimą turi patvirtinti vienas iš laboratorinių darbų dėstytojų. Duomenų rinkinio reikalavimai:

- Turi egzistuoti skaitinės (integer ir real tipo) ir /arba kategorinės reikšmės. Duomenų rinkinys kuriame yra tik kategorinio tipo atributai yra netinkamas.
- Duomenų rinkinyje įrašų (eilučių) m turi būti ne mažiau nei 500, t.y., $\infty > m \geq 500$ ir atributų n nemažiau nei 8 (stulpeliai) $\infty > n \geq 8$. Jeigu atributų n pasirinktame duomenų rinkinyje yra mažiau, privalote pridėti išvestinius (sukurtus) atributus

Užduotys:

1. Atlikti duomenų rinkinio kokybės analizę (žr. 2 pav.). Kiekvienam tolydinio tipo atributui paskaičiuoti:
    - bendrą reikšmių skaičių,
    - trūkstamų reikšmių procentą,
    - kardinalumą,
    - minimalią (min) ir maksimalią (max) reikšmes,
    - 1-ąją ir 3- ją kvartilius,
    - vidurkį,
    - medianą,
    - standartinį nuokrypį.

2. Kiekvienam **kategorinio** tipo atributui paskaičiuoti:
    - bendrą reikšmių skaičių,
    - trūkstamų reikšmių procentą,
    - kardinalumą,
    - modą,
    - modos dažnumo reikšmę,
    - modos procentinę reikšmę,
    - 2-ąją modą,
    - 2-osios modos dažnumo reikšmę,
    - 2-osios modos procentinę reikšmę.

3. Nupaišyti atributų histogramas. Pateikti aprašymus, koks tai pasiskirstymas ir kokias išvadas pagal tai galima formuluoti.
4. Identifikuoti duomenų kokybės problemas:
    - trūkstamas reikšmes
    - kardinalumo problemas
    - triukšmus– ekstremalias reikšmes (angl. outliers).
    - Pateikti šių problemų sprendimo planą, kuris bus realizuotas programiškai.

5. Nustatyti sąryšius tarp atributų panaudojant vizualizacijos būdus:
    - **Tolydinio tipo atributams**: naudojant „scatter plot“ tipo diagramą pateikti kelis (2-3) pavyzdžius su stipria tiesine atributų priklausomybe bei kelis pavyzdžius su tarpusavyje nekoreliuojančiais (silpnai koreliuojančiais) atributais. Pakomentuoti rezultatus.
    - Pateikti SPLOM diagramą (Scatter Plot Matrix).
    - **Kategorinio tipo atributams**: naudojant „bar plot“ tipo diagramą pateikti keletą (2-3) atributų priklausomybės pavyzdžių ir pakomentuoti rezultatus.
    - Pateikti keletą (2-3) histogramų ir „box plot“ diagramų pavyzdžių, vaizduojančių sąryšius tarp **kategorinio** ir **tolydinio** tipo kintamųjų

6. Paskaičiuoti kovariacijos ir koreliacijos reikšmes tarp tolydinio tipo atributų ir grafiškai atvaizduoti koreliacijos matricą. Rezultatus pakomentuoti.
7. Atlikti duomenų normalizaciją (režiai [0;1] arba [-1;1]).
8. Kategorinio tipo kintamuosius paversti į tolydinio tipo kintamuosius.

Pasirinkti darbo įrankiai:

Pasirinkta programavimo kalba: Python.

Pasirinkta integruota kūrimo aplinka: PyCharm.

Duomenų rinkinys: Internetinių vaizdo įrašų charakteristikos ir perkodavimo laikas (angl. Online Video Characteristics and Transcoding Time)

Duomenų rinkinys	
Įrašų kiekis	68785
Atributų kiekis	20
Naudojamų atributų kiekis	19
Tolydinio tipo atributų kiekis	17
Kategorinio tipo atributų kiekis	2
Tolydus atributai	
Trukmė (angl. Duration)	
Plotis (angl. Width)	
Aukštis (angl. Height)	
Pralaidumas (angl. Bitrate)	
Kadrų dažnis (angl. Framerate)	
I (angl. I)	
P (angl. P)	
Kadrai (angl. Frames)	
I_dydis (angl. I_size)	
P_dydis (angl. P_size)	
Dydis (angl. Size)	
Išėigos pralaidimas (angl. O_bitrate)	
Išeigos kadrų dažnis (angl. O_framerate)	
Išeigos plotis (angl. O_width)	
Išeigos aukštis (angl. O_height)	
Sunaudotas RAM kiekis (angl. Umem)	
Užtruktas laikas (angl. Utime)	
Kategoriniai atributai	
Kodekas (angl. Codec)	
Konvertuotas Kodekas (angl. O_codec)	

Užduočių sprendimai

Atmečiau šį atributą: ID.

1. Atlikti duomenų rinkinio kokybės analizę. Kiekvienam tolydinio tipo atributui.

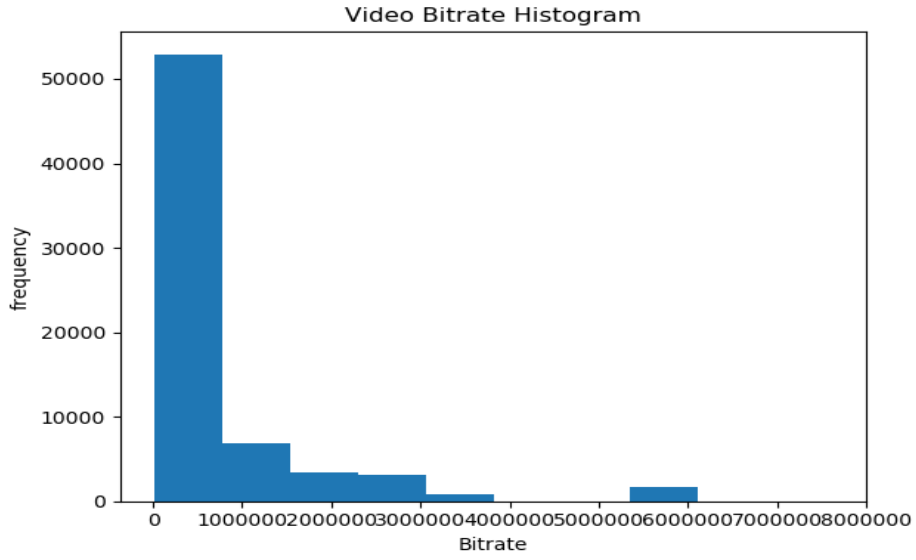
Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Minimali reikšmė	Maksimali reikšmė	1-asis kvartilis	3-asis kvartilis	Vidurkis	Mediana	Standartinis nuokrypis
duration	68784	0	1086	31.08	25844.086	106.765	379.32	286.4139214	239.14166	287.2555618
width	68784	0	6	176	1920	320	640	624.9341707	480	463.1657021
height	68784	0	6	144	1080	240	480	412.5722261	360	240.6137227
bitrate	68784	0	1095	8384	7628466	134334	652967	693701.5	291150	1095619.591
framerate	68784	0	261	5.7057524	48	15	29	23.24132053	25.02174	7.224795437
i	68784	0	306	7	5170	39	138	100.8683124	80	84.76417462
p	68784	0	1042	175	304959	2374	9155	6531.69221	5515	6075.827577
frames	68784	0	1044	192	310129	2417	9232	6641.708377	5628	6153.297723
i_size	68784	0	1099	11648	90828552	393395	3392479	2838986.702	945865	4325105.154
p_size	68784	0	1099	33845	768996980	1851539	15155062	22180569.3	6166260	50972690.79
size	68784	0	1099	191879	806711069	2258222	19773349	25022942.37	7881069	54143621.81
o_bitrate	68784	0	7	56000	5000000	109000	3000000	1395035.953	539000	1749338.79
o_framerate	68784	0	5	12	29.97	15	25	21.19086168	24	6.668654311
o_width	68784	0	6	176	1920	320	1280	802.3363573	480	609.9553631
o_height	68784	0	6	144	1080	240	720	503.8255408	360	315.9681412
umem	68784	0	9395	22508	711824	216820	219656	228224.7179	219480	97430.17013
utime	68784	0.004361479	10961	0	224.574	2.096	10.433	9.996128242	4.408	16.10760703

2. Atlikti duomenų rinkinio kokybės analizę. Kiekvienam kategorinio tipo atributui.

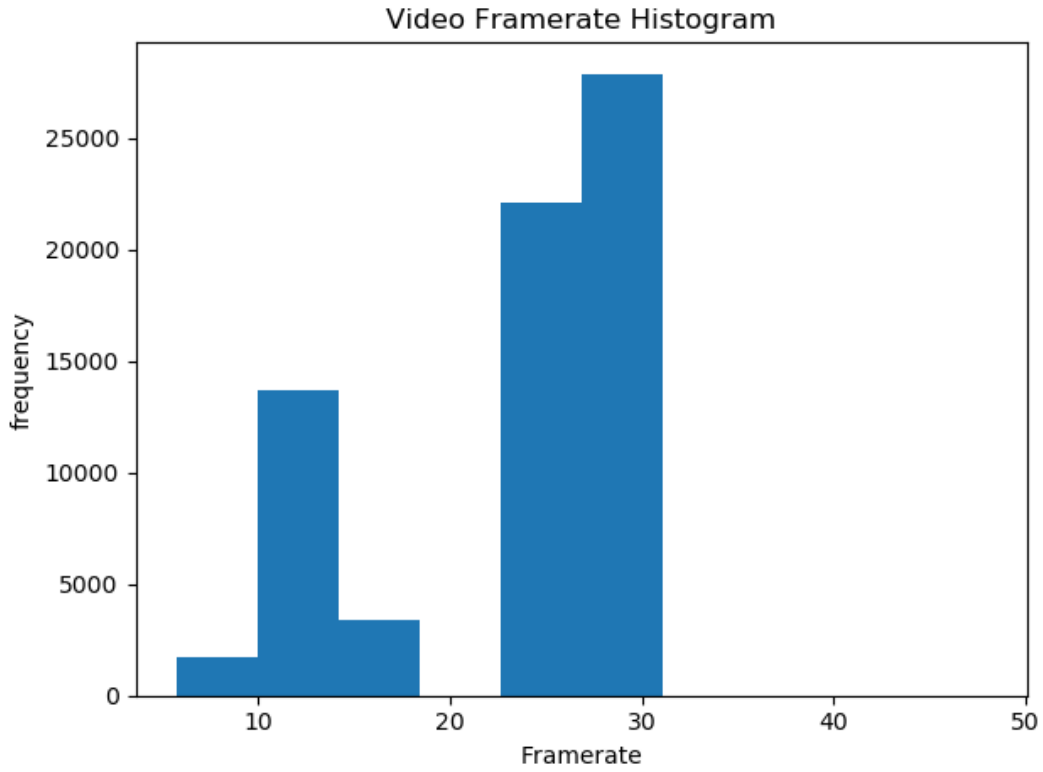
Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Moda	Modos dažnumas	Moda, %	2-oji Moda	2-osios Modos dažnumas	2-oji Moda, %
codec	68784	0	4	h264	31545	45.860956	vp8	18387	26.73150733
o_codec	68784	0	4	mpeg4	17291	25.1381135	vp8	17277	25.11775994

3. Nupaišyti atributų histogramas.

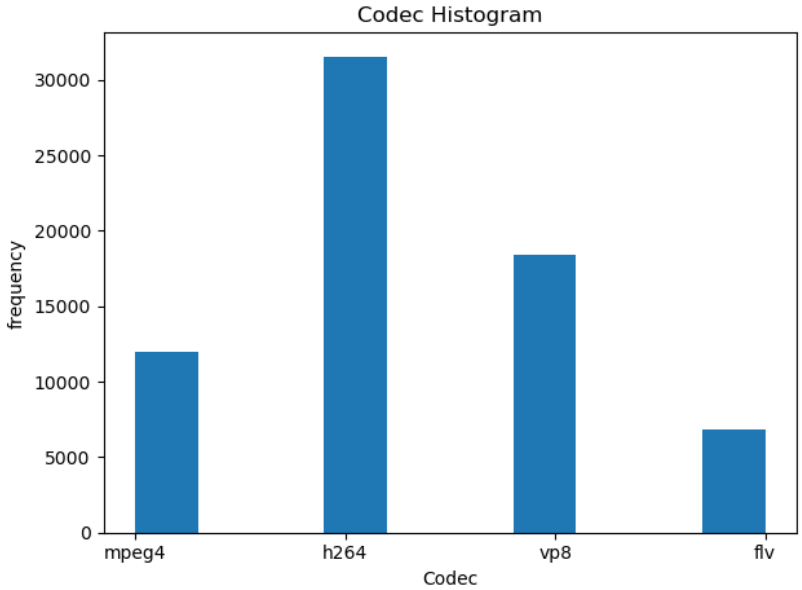
Tolydinio atributo **Bitrate** histograma. Iš histogramos galime spręsti jog dominuoja vaizdo įrašai su mažesnio bitų perdavimo sparta. Tai reiškia jog vidutinio vaizdo įrašo kokybė yra prasta.



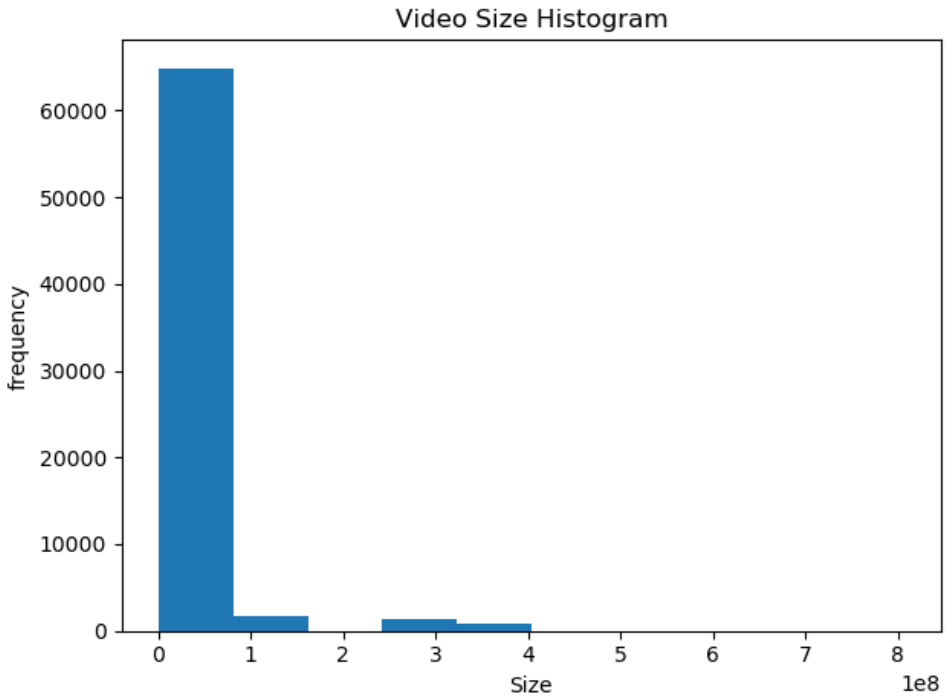
Tolydinio atributo **Framerate** histograma. Histogramoje matosi jog ištirtų vaizdo įrašų kadrų kiekis per sekundę yra nuo 25 iki 30. Jeigu tirčiau atnaujintus duomenis pamatytume, kad 60 kadrų per sekundę vaizdo įrašai vis labiau populiarėja nors šitoje diagramoje nematome nei vieno vaizdo įrašo su 60 FPS (kadru kiekis per sekundę).



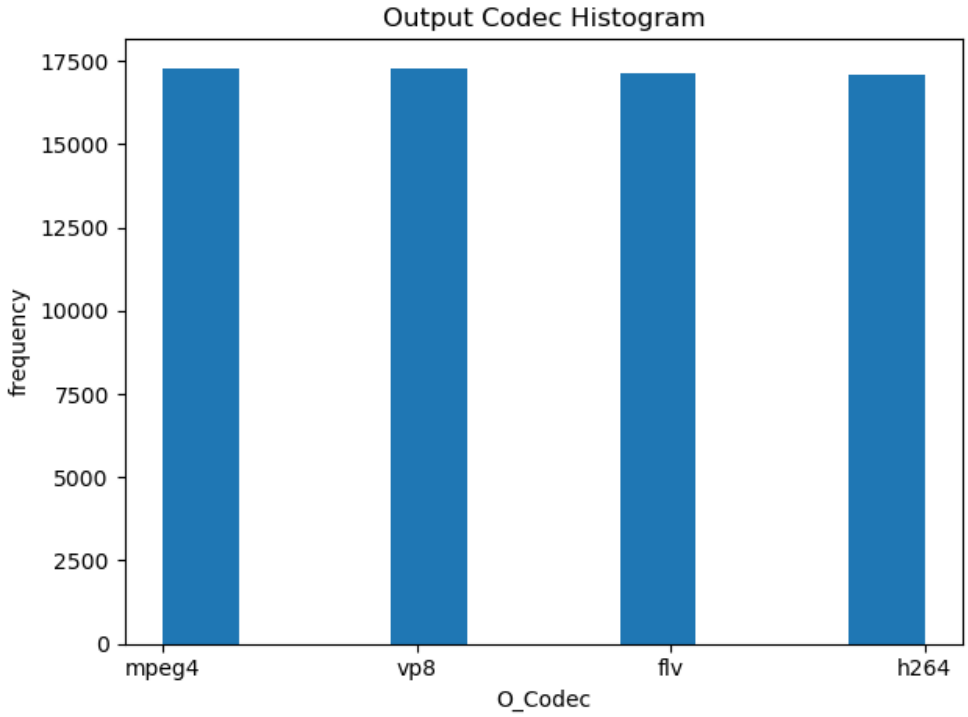
Kategorinio atributo **Codec** histograma. Iš histogramos galime spręsti jog dažniausiai naudojamas codec spaudimo būdas yra H264. Šitas codec būdas yra efektyviausias norint sumažinti vaizdo įrašo dydį išlaikant geros kokybės vaizdo įrašą.



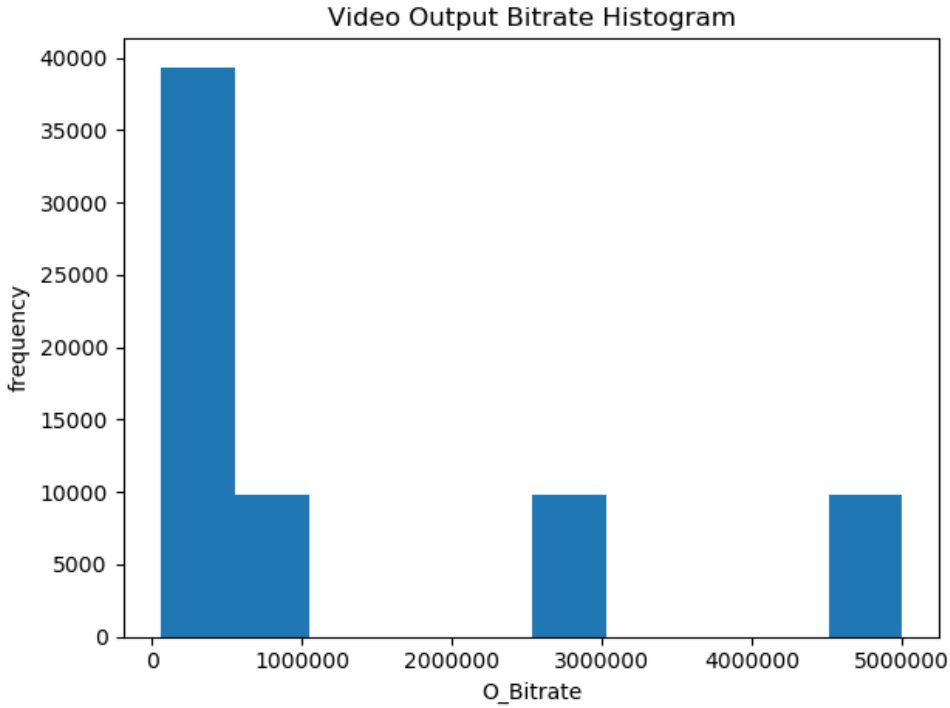
Tolydinio atributo **Size** histograma. Histogramoje matome jog vaizdo įrašų dydžiai dažniausia yra tarp 10000 ir 100000000 bitų dydžio. Kitais atvejai jeigu vaizdo įrašo dydžiai yra didesni negu vidutinio vaizdo įrašo dydis šitie vaizdo įrašai yra dažniausiai neapdoroti ir labai didelės kokybės.



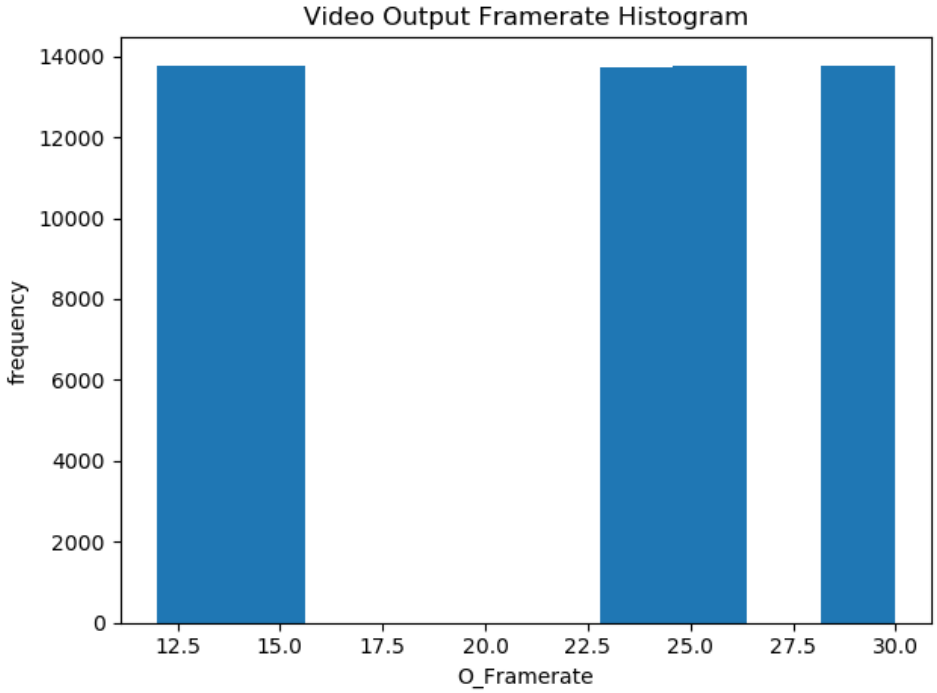
Kategorinio atributo **O_Codec** histograma. Iš histogramos galime spręsti jog konvertuojant, codec pasiskirstymas išsilygina.



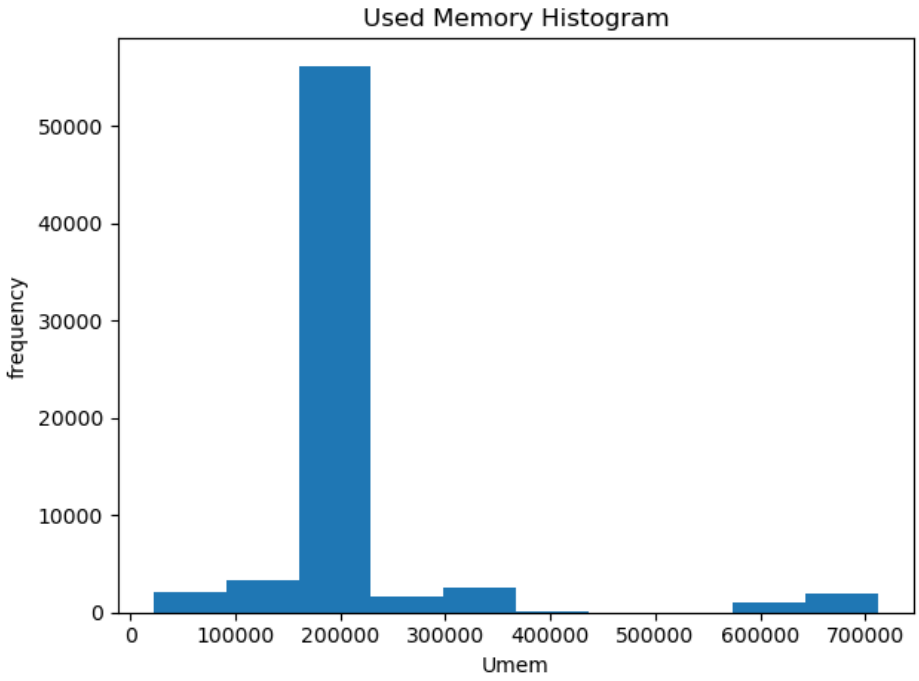
Tolydinio atributo O_**Bitrate** histograma. Histograma parodė jog vidutiniškai konvertuoto vaizdo įrašo bitu kiekis per sekundę išliko toks pat, bet padidėjo ir didesnio bitų kiekio per sekundę vaizdo įrašų kiekis. Todėl galime teigti jog konvertuojant padidėjo didesnės kokybės vaizdo įrašų kiekis.



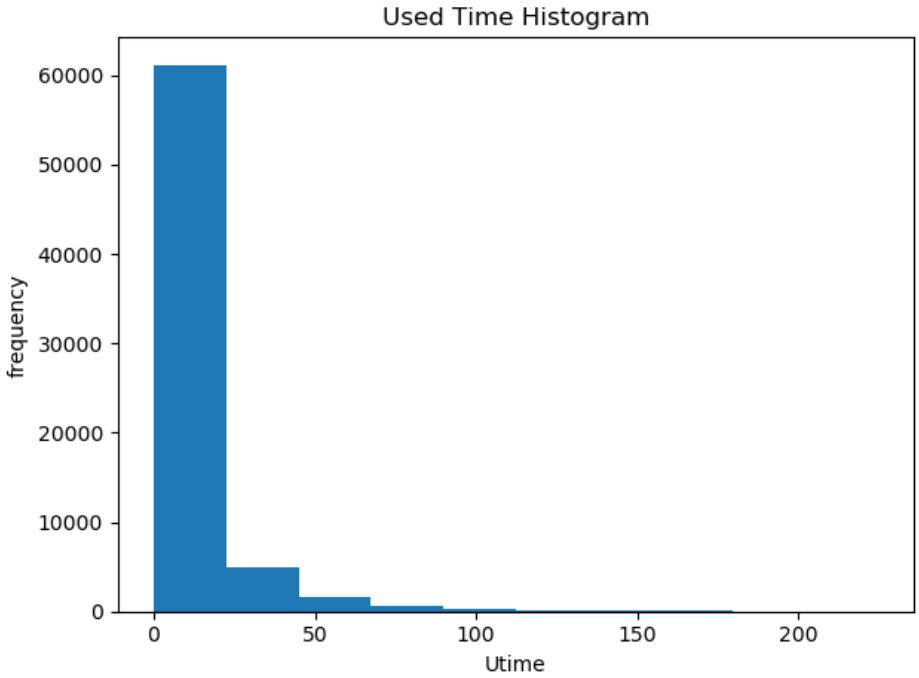
Tolydinio atributo **O_Framerate** histograma. Iš Histogramos grafiko matome jog konvertuojant kadrų kiekis per sekundę pasiskirstė daugmaž vienodai. Kaip ir akstesnėje diagramoje nematome jog būtu naudota 60 FPS (kadru kiekis per sekundę).



Tolydinio atributo **Umem** histograma. Histogramoje matome kad vidutinis atminties sunaudojimas konvertuojant vaizdo įrašą yra apie 2GB. Jeigu vaizdo įrašo kokybė yra geresnė arba vaizdo įrašas yra ilgesnis, tada panaudos daugiau atminties konvertuojant vaizdo įrašą.



Tolydinio atributo **Utime** histograma. Histogramoje matome kad vidutinis laikas užtrunkant konvertuoti vaizdo įrašą yra nuo 0.6 min iki 20 min . Jeigu vaizdo įrašo kokybė yra geresnė arba vaizdo įrašas yra ilgesnis, tada užtruks ilgiau konvertuoti vaizdo įrašą.



4. Identifikuoti duomenų kokybės problemas:

Gauti pradiniai duomenys buvo tvarkingi, be jokių trūkstamų reikšmių, tačiau programa buvo realizuota taip, jog surastu reikšmes su trūkstamomis reikšmėmis ir atspausdintu jų procentą, taip pat skaičiavimuose ir grafikų paišymuose, laukai su trūkstamomis reikšmėmis yra ignoruojami.

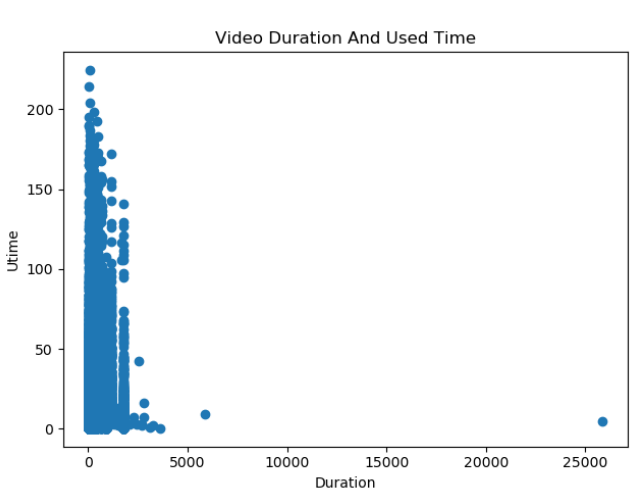
Ekstremalūs duomenys yra koreguojami į artimiausią tinkamą reikšmę.

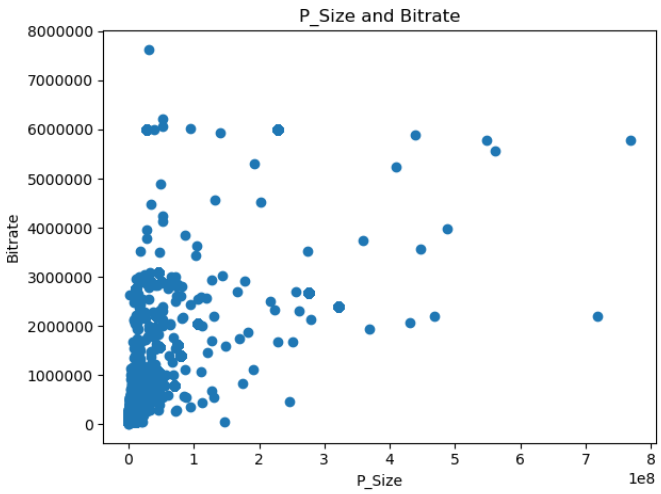
Pavyzdžiui Bitrate galimos reikšmės yra tarp 0 – 80000, jeigu duomenyse atsirastu tarkim skaičius 90000, tada jis bus pakoreguotas į 80000. Jeigu bus neigiama reikšmė amžiuje, ji bus pakeista į nulį.

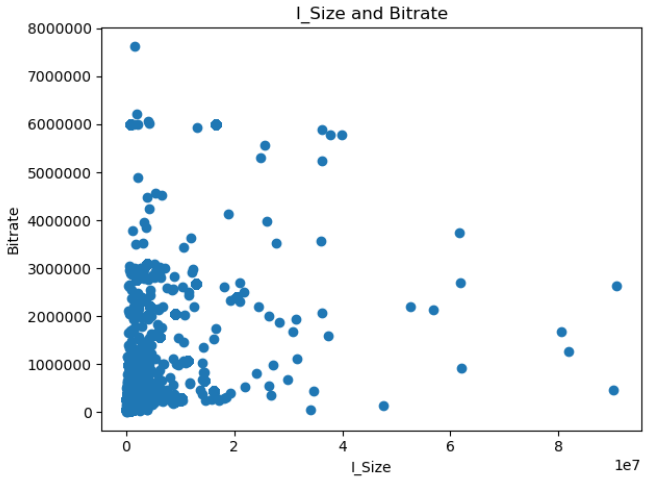
5. Nustatyti sąryšius tarp atributų panaudojant vizualizacijos būdus

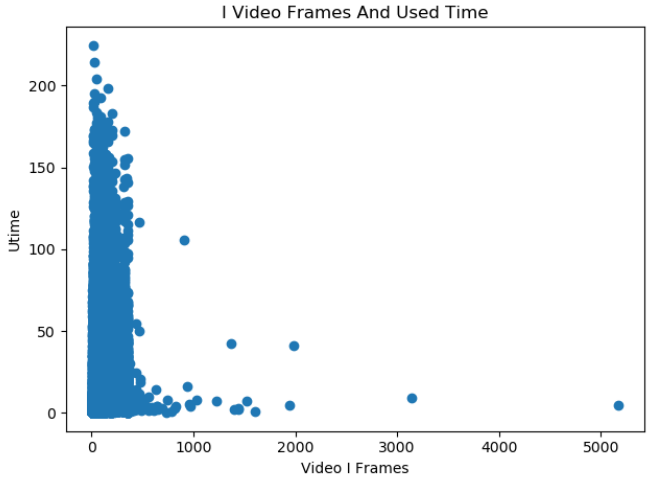
a. Tolydinio tipo atributams, naudojant „scatter plot“

Koreliuojantys duomenys:







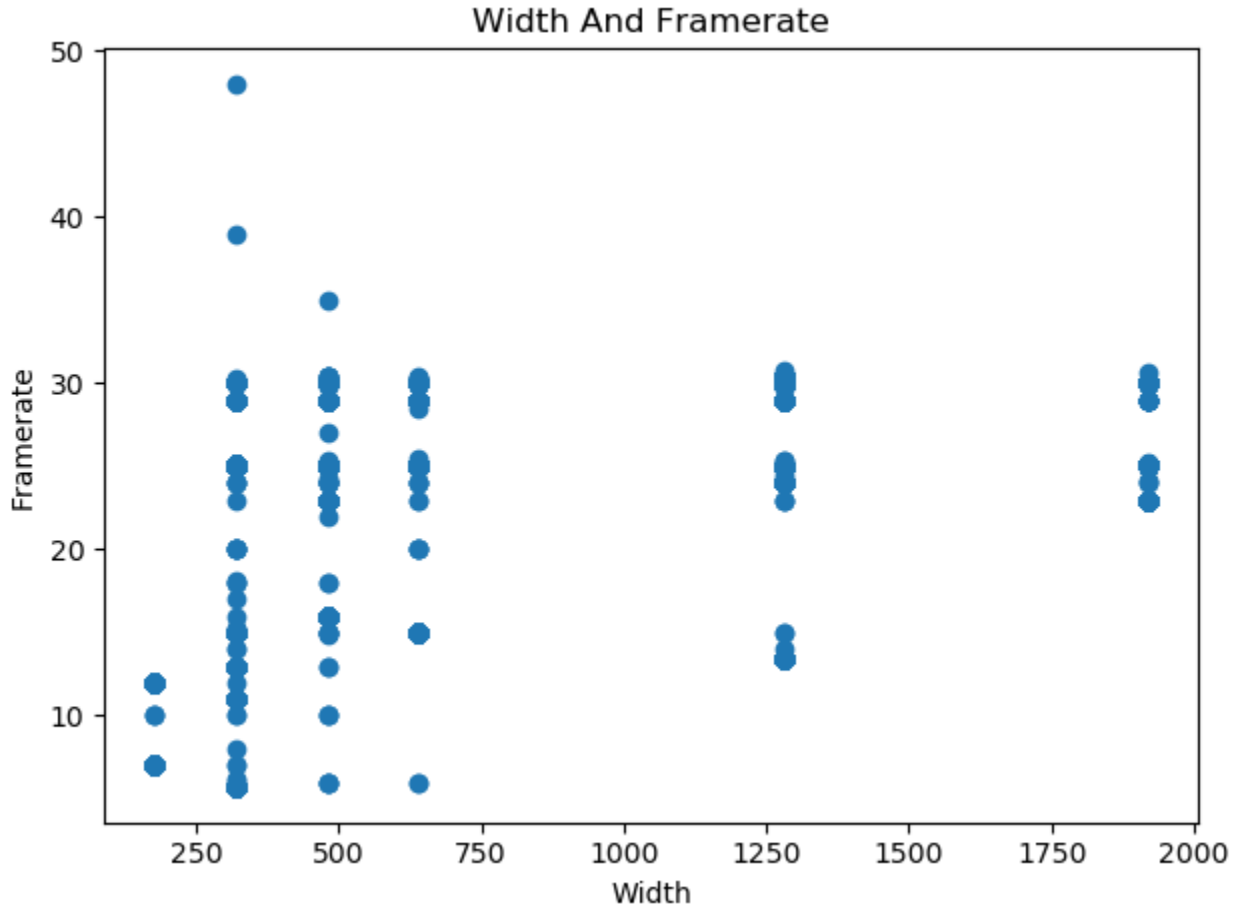


Iš gautų diagramų matome, kad:

- Matome jog pirmojoje diagramoje kuo trumpesnis vaizdo įrašas tuo ilgiau užtruks konvertuoti. Na žinoma vaizdo įrašo konvertavimo ilgis priklauso nuo daugiau nei dviejų parametrų todėl diagrama gali būti netiksli.

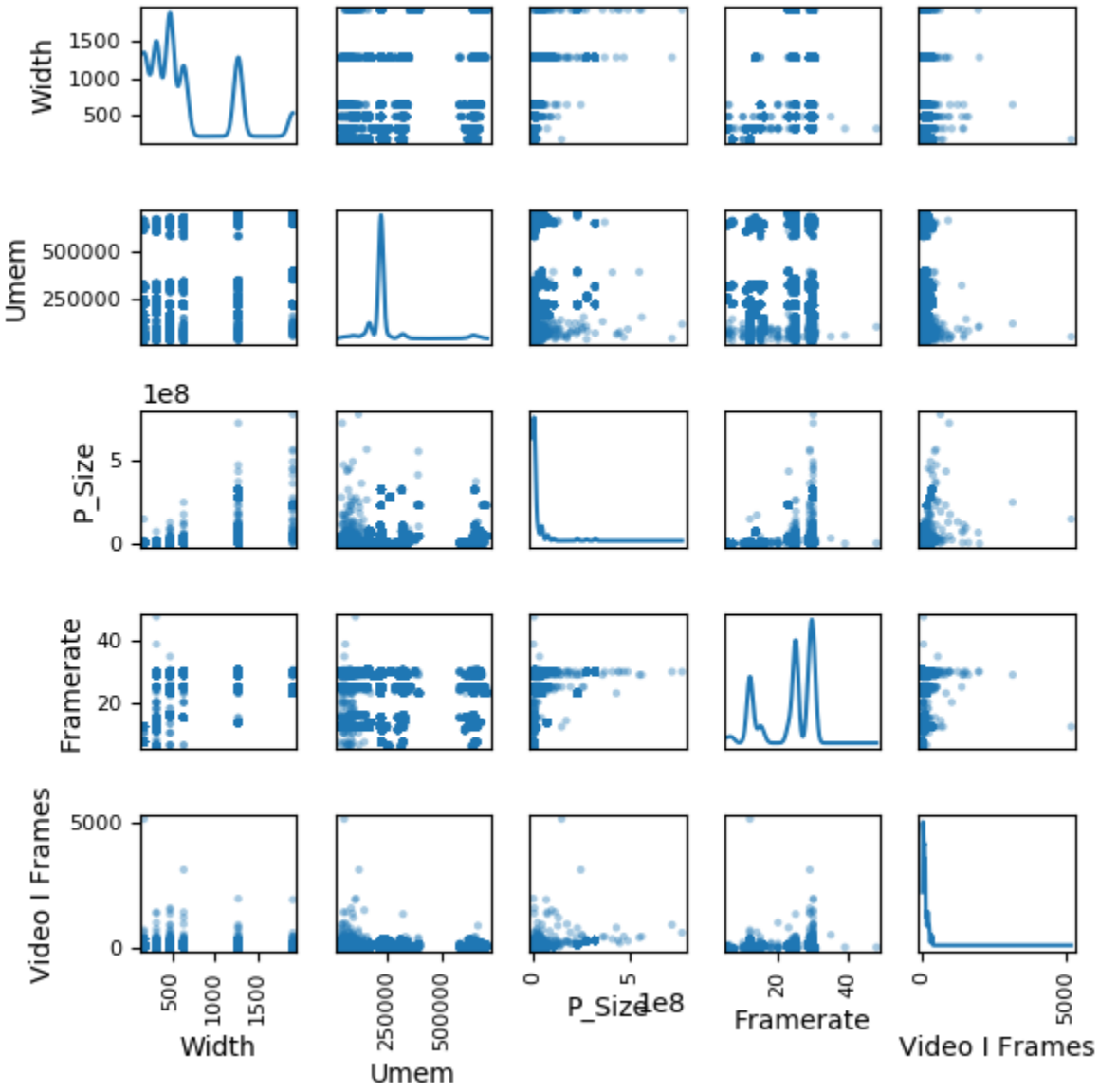
- Antroje Diagramoje galime pastebėti kad P tipo vaizdo įrašų dydis priklauso nuo bitu kiekio per sekundę. Kuo didesnis dydis tuo didesnis bitų kiekis per sekundę.
- Trečioje kaip ir antroje diagramoje galime pastebėti kad i tipo vaizdo įrašų dydis priklauso nuo bitu kiekio per sekundę. Kuo didesnis dydis tuo didesnis bitų kiekis per sekundę.
- Ketvirtoje diagramoje matome jog kuo daugiau vaizdo įrašo i FPS tuo trumpiau konvertuos vaizdo įrašą. Na žinoma vaizdo įrašo konvertavimo ilgis priklauso nuo daugiau nei dviejų parametrų todėl diagrama gali būti netiksli.

Nekoreliuojantys Duomenys:

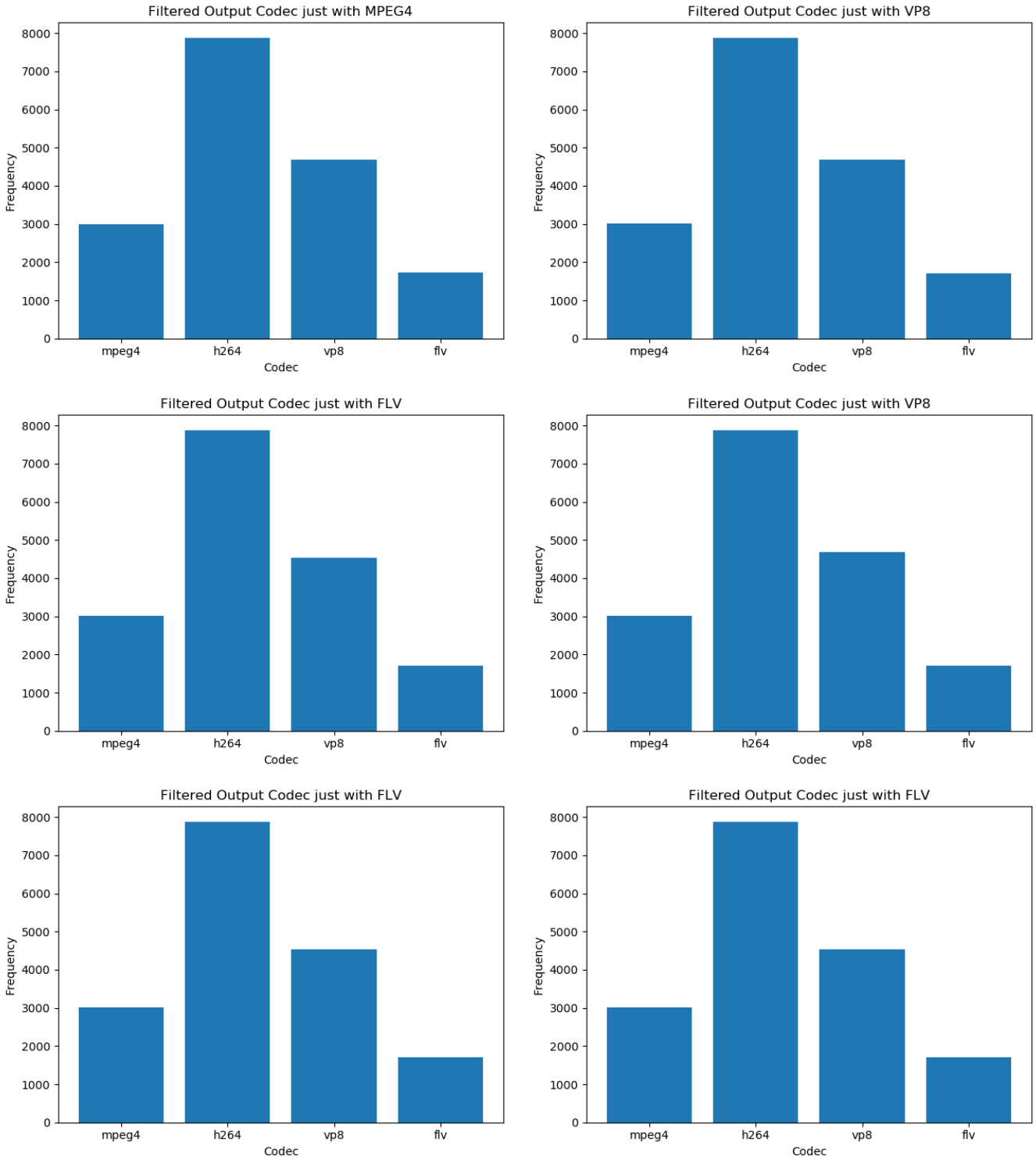


Matome jog šitoje diagramoje kadrų kiekis per sekundę nekoreliuoja su vienu iš rezoliucijos duomenų.

b. Pateikti SPLOM diagramą (Scatter Plot Matrx)



c. Kategorinio tipo atributų priklausomybė.

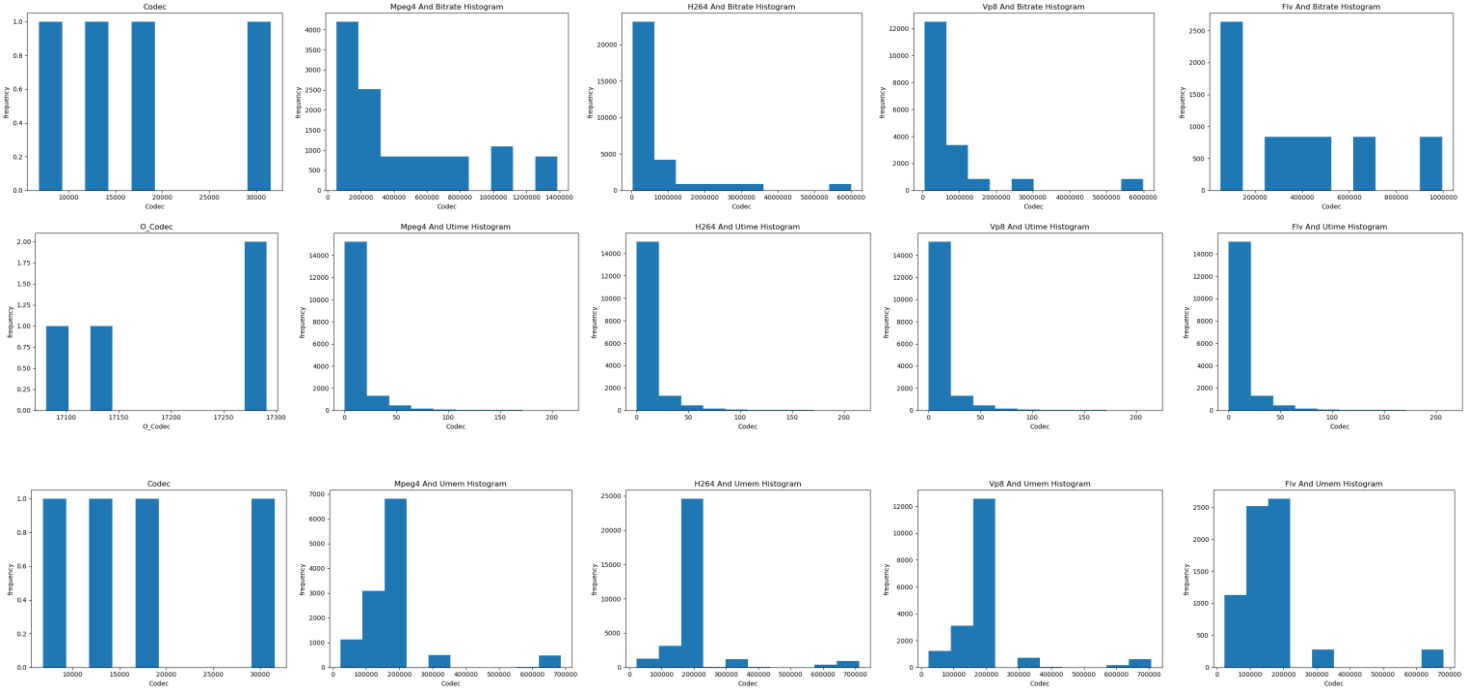


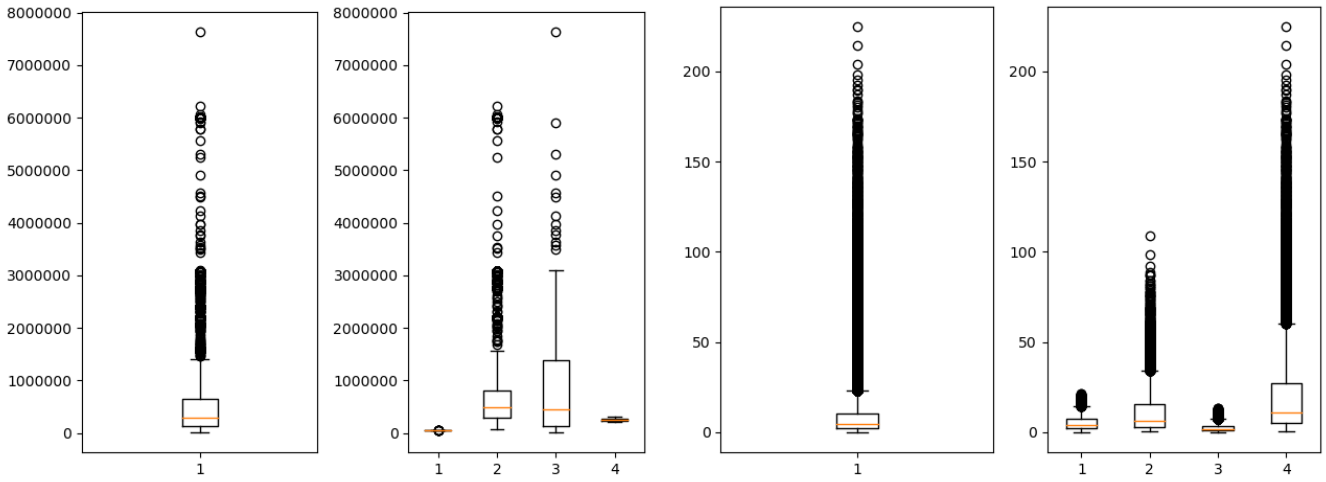
Iš diagramų matome, kad:

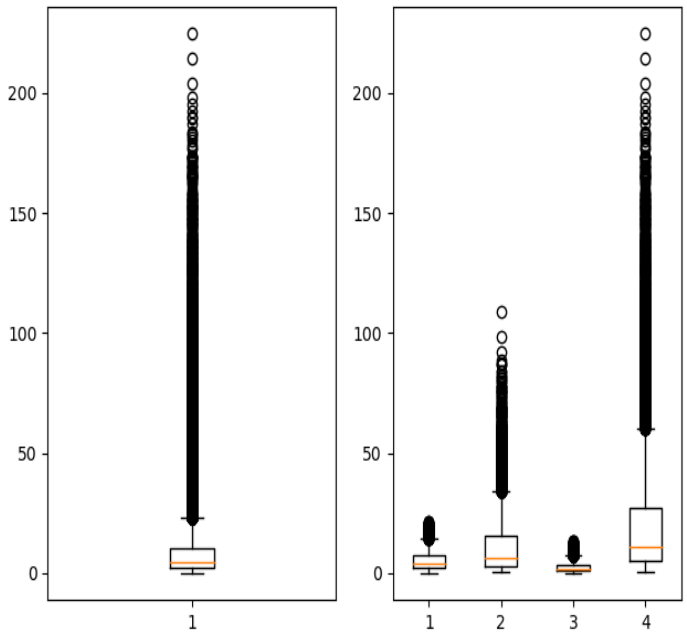
- Kad primos dvi diagramos yra vienodos tai reiškia kad vaizdo įrašų kiekis nesiskyrė konvertuojant į MPEG4 ar VP8 formatus.

- Palyginus antrą ir trečią diagramas matome jog konvertuojant iš VP8 į FLV formatą vaizdo įrašų kiekis yra mažesnis negu konvertuojant iš VP8 į VP8 formatą.
- Palyginus ketvirtą ir penktą diagramas matome jog konvertuojant iš bet kokio formato į FLV formatą vaizdo įrašų kiekis yra mažesnis negu palyginus visus kitus fomatus.

d. Pateikti histogramų ir „box plot“ diagramų pavyzdžių, vaizduojančių sąryšius tarp kategorinio ir tolydinio tipo kintamųjų.



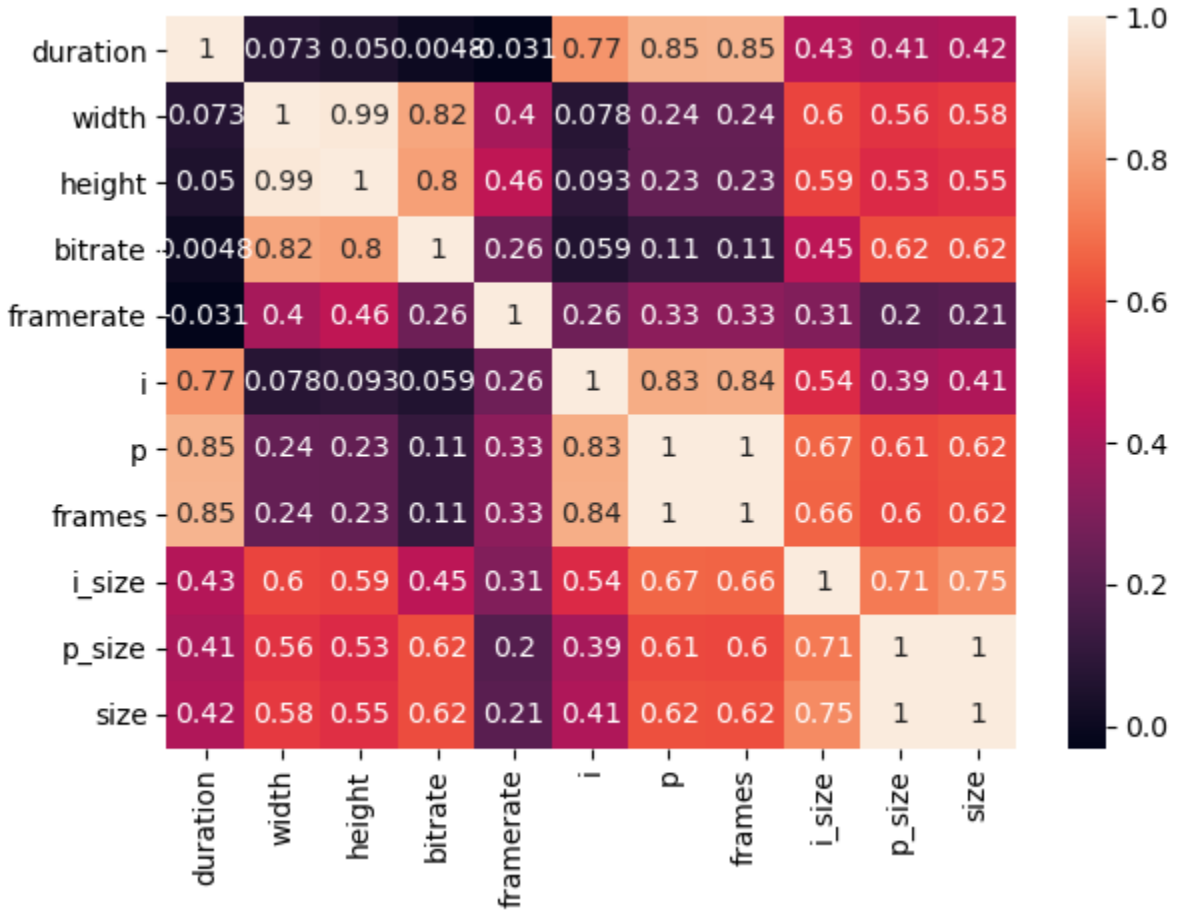




Iš gautų rezultatų matome, kad:

- Didesnio bitu kiekis per sekunde vaizdo įrašai dažniausia naudoja h264 ir vp8 kodekus
- Konvertavimo laikas didėja kai konvertuojama į h264 arba flv vaizdo įrašus
- Konvertavimo naudojami RAM didėja kai vaizdo įrašai yra konvertuojami iš h264 ir flv kodeku

6. Paskaičiuoti kovariacijos ir koreliacijos reikšmes tarp tolydinio tipo atributų ir grafiškai atvaizduoti koreliacijos matricą.



Iš matricos galime spręsti, kad:

- vaizdo įrašo ilgumas (angl. duration) priklauso nuo kadrų (angl. frames),
- matome jog kadrai per sekundę nepriklauso nuo vaizdo įrašo ilgumo,
- taip pat bitų perdavimo sparta priklauso nuo rezoliucijos (pločio ir aukščio).

7. Atlikti duomenų normalizaciją.

Duomenų normalizaciją realizavau, panaudojęs šią formulę:

$$z = \frac{x - min(x)}{[\max(x) - min(x)]} \times (High - Low) + Low$$

8. Kategorinio tipo kintamuosius paversti į tolydinio tipo kintamuosius.

Kategorinio tipo kintamuosius su Python paverčiau į tolydinio tipo kintamuosius naudodamas Dictionary mapping pvz. („MPEG4“=1, “H264”=2, “VP8”=3, “FLV”=4).

Išvados

Pasirinkus vaizdo įrašų charakteristikų duomenų rinkinį ir atlikus jam analizę sužinojome, kad

- Vaizdo įrašo ilgumas visiškai nepriklauso nuo kadru kiekio per sekundę,
- Vaizdo įrašo kokybė priklauso nuo rezoliucijos, kadrų kiekio per sekundę ir bitu spartos (angl. bitrate),
- Vaizdo įrašo kodekas prieš formatavimą yra dažniausiai naudojamas H264 vaizdo įrašai kurie naudoja šį formatą yra aukštos kokybės,
- Pastebėjau, kad skirtingų kodekų naudojimas išsilygina konvertuojant.
- Galime pamatyti tendenciją, kad įvesties ir išvesties charakteristikos beveik nėra susijusios,
- Konvertuojant vaizdo įrašo konvertavimo ilgis priklauso daugiausia nuo kodeko, bitų spartos (angl. bitrate), rezoliucijos (aukščio ir pločio) ir kadrų kiekio per sekundę.