KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS

Intelektikos pagrindai 2020

Laboratorinio darbo nr. 1 ataskaita

Atliko:

IFF-7/2 gr. studentas

Justas Milišiūnas

2020-02-29

Dėstytojai:

lekt. Audrius Nečiūnas

doc. Agnė Paulauskaitė-Tarasevičienė

KAUNAS 2020

TURINYS

1. Užduotis

1. Pasirinkti duomenų rinkinį
2. Atlikti duomenų rinkinio kokybės analizę
3. Nupaišyti atributų histogramas
4. Identifikuoti duomenų kokybės problemas. Pateikti šių problemų sprendimo būdą.
5. Nustatyti sąryšius tarp atributų panaudojant vizualizacijos būdus:
    1. Tolydinio tipo atributams: naudojant „scatter plot“ tipo diagramą pateikti kelis (2-3) pavyzdžius su stipria tiesine atributų priklausomybe (tiesioginė arba atvirkštinė koreliacija) bei kelis pavyzdžius su tarpusavyje nekoreliuojančiais (silpnai koreliuojančiais) atributais. Pakomentuoti rezultatus.
    2. Kategorinio tipo atributams: naudojant „bar plot“ tipo diagramą pateikti keletą (2-3) atributų priklausomybės pavyzdžių ir pakomentuoti rezultatus.
    3. Pateikti keletą (2-3) histogramų ir „box plot“ diagramų pavyzdžių, vaizduojančių sąryšius tarp kategorinio (pavyzdys pateiktas pav.3) 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

2. Sprendimas

2.1. Duomenų rinkinys

Duomenų rinkinį sudaro 11 stulpelių. 54 tūksnt. įrašu.

Stulpeliai:

- Id – įrašo numeracija
- Carat – deimanto karatai
- Cut (Fair, Good, Very Good, Premium, Ideal) – deimanto pjūvis
- Color (From J (worst) to D (best)) – deimanto spalva
- Clarity (I1 (worst), SI2, SI1, VS1, VS2, VVS2, VVS1, IF (best)) – deimanto skaidrumas
- Depth percentage – deimanto gylis padalintas iš deimanto pločio
- Table – deimanto viršaus plotis reliatyvus plačiausiai vietai
- Price – deimanto kaina
- Length – deimanto ilgis
- Width – deimanto plotis
- Depth – deimanto gylis

2.2. Duomenų rinkinio kokybės analizė

Tolydžių atributų analizė:

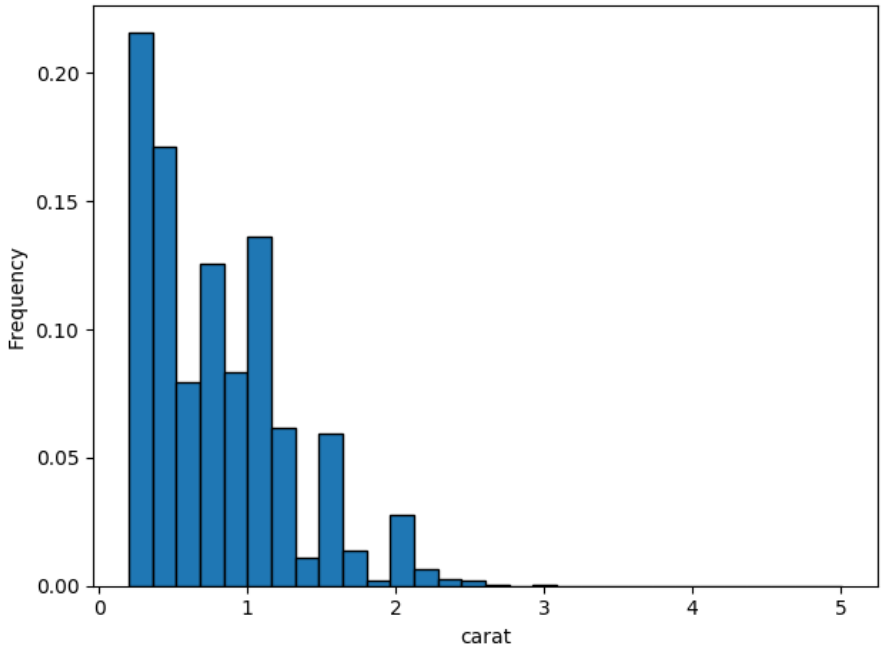
Atributo pav.	Kiekis	Trūks. Reikšmės %	Kardinalumas	Min. reikšmė	Max. reikšmė	1 kvartilis	3 kvartilis	Vidurkis	Mediana	Standart. nuokrypis
carat	53940	0 %	273	0.2	5.01	0.4	1.04	0.80	0.80	0.47
depth_pct	53940	0 %	184	43	79	61	62.5	61.75	61.75	1.43
table	53940	0 %	127	43	95	56	59	57.46	57.46	2.23
price	53940	0 %	11602	326	18823	950	5325	3932.80	3932.80	3989.44
length	53940	0 %	554	0	10.74	4.71	6.54	5.73	5.73	1.12
width	53940	0 %	552	0	58.9	4.72	6.54	5.73	5.73	1.14
depth	53490	0 %	375	0	31.8	2.91	4.04	3.54	3.54	0.71

Kategorinių atributų analizė:

Atributo pav.	Kiekis	Trūks. Reikšmės %	Kardinalumas	Moda	Modos dažnumas	Moda %	2 Moda	2 Modos dažnumas	2 Moda %
cut	53490	0 %	5	Ideal	21551	39.95 %	Premium	13791	25.57 %
color	53490	0 %	7	G	11292	20.93 %	E	9797	18.16 %
clarity	53490	0 %	8	SI1	13065	24.22 %	VS2	12258	22.73 %

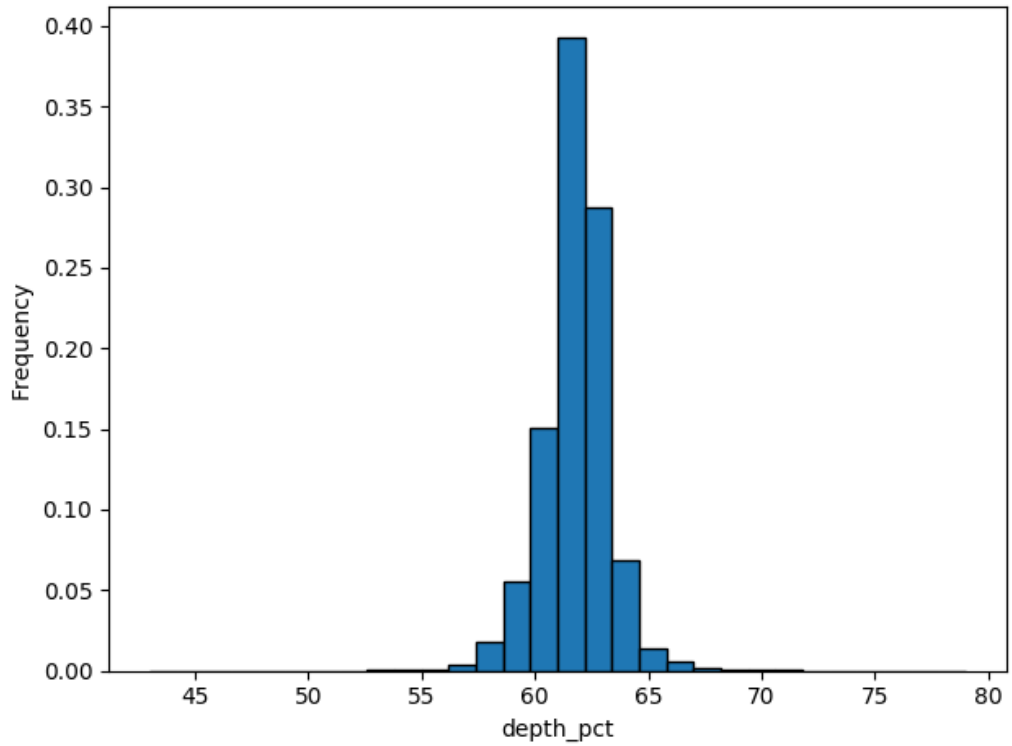
2.3. Atributų histogramos

Carat atributo histograma:



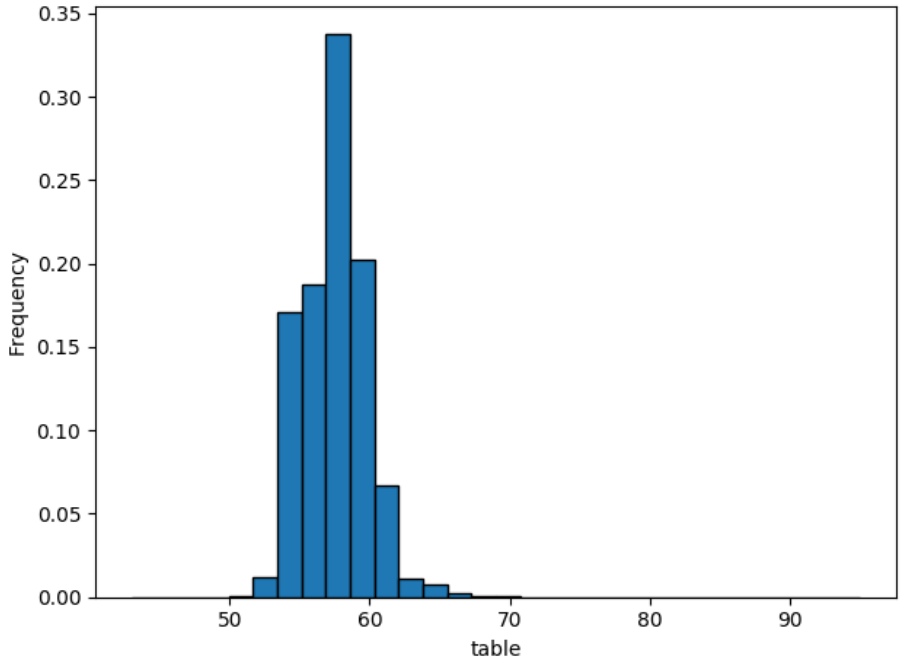
Iš šios histogramos matome, kad stulpelio ,,cut“ dumenys pasiskirstę eksponentiškai.

depth_pct atributo histograma:



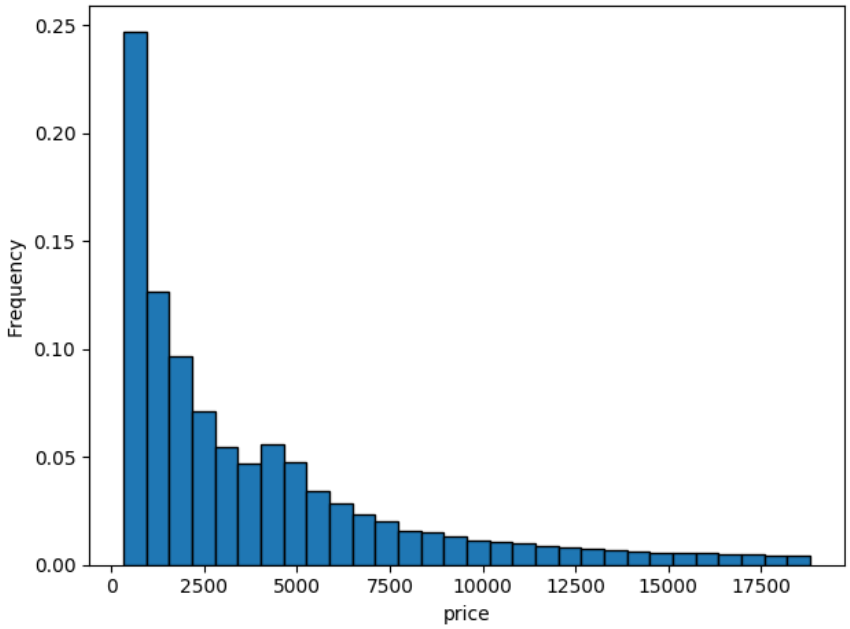
Iš šios histogramos matome, kad stulepio ,,depth_pct“ duomenys pasiskirstę normaliai.

Table atributo histograma:



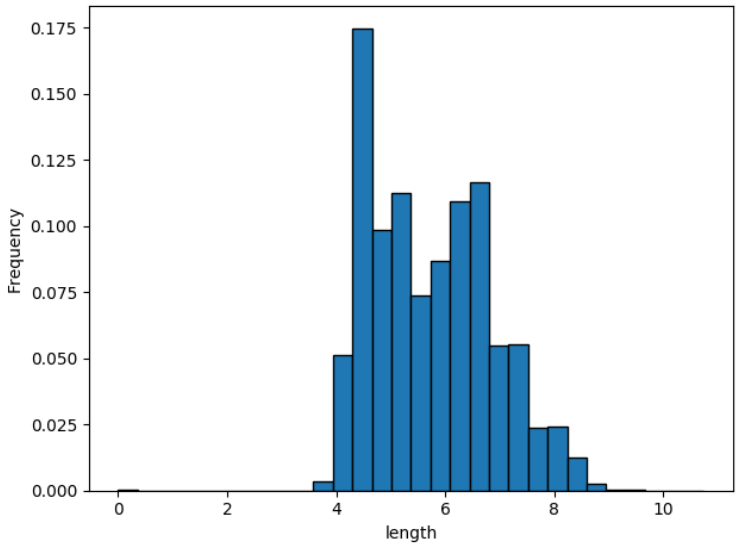
Iš šios histogramos matome, kad atributo ,,table“ duomenys pasiskirstę normaliai

Price atributo histograma:



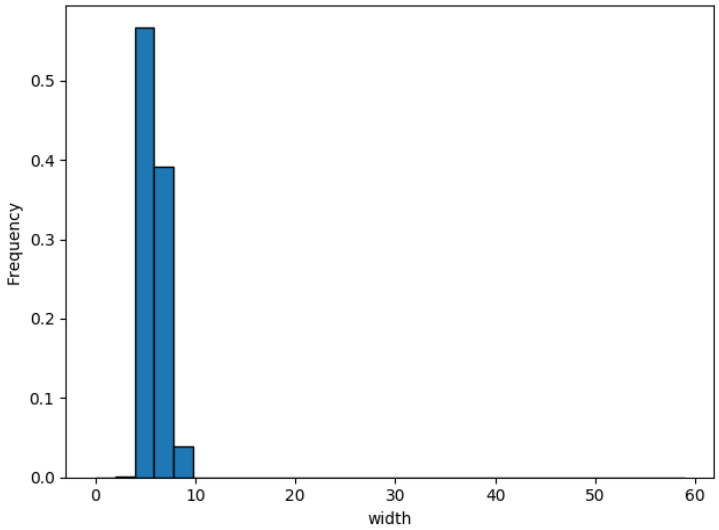
Iš šios histogramos matome, kad atributo ,,price“ duomenys pasiskirstę eksponentiškai.

Length atributo histograma:



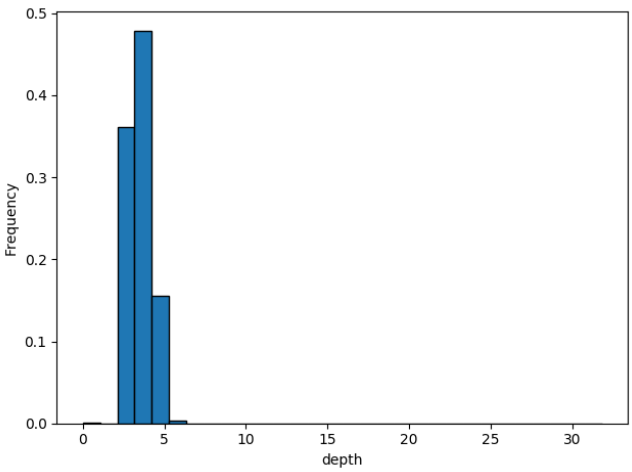
Iš šios histogramos matome, kad atributo ,,length“ duomenys pasiskirstę normaliai.

Width atributo histograma:



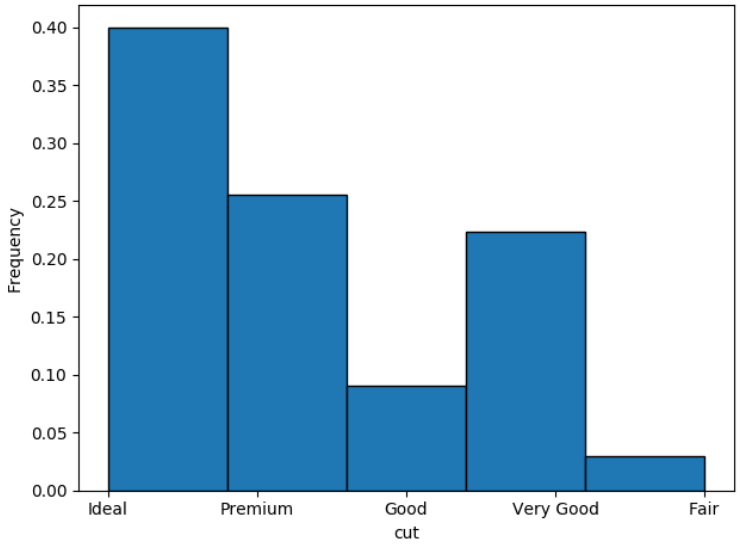
Iš šios histogramos matome, kad atributo ,,width“ duomenys pasiskirstę normaliai.

Depth atributo histograma:



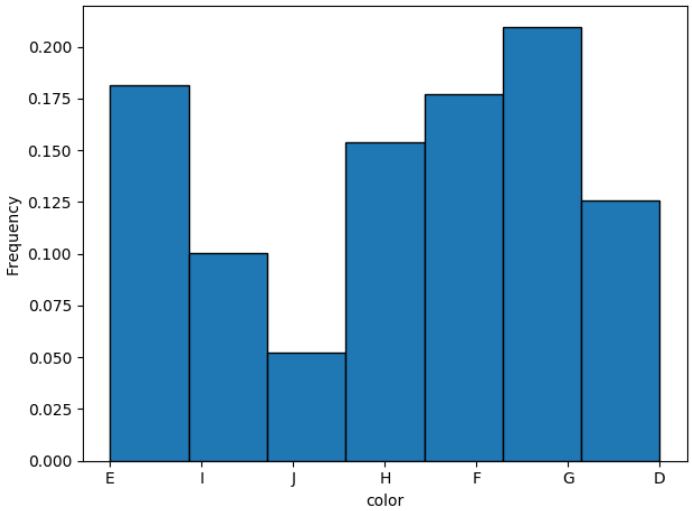
Iš šios histogramos matome, kad atributo ,,depth“ duomenys pasiskirstę normaliai.

Cut atributo histograma:



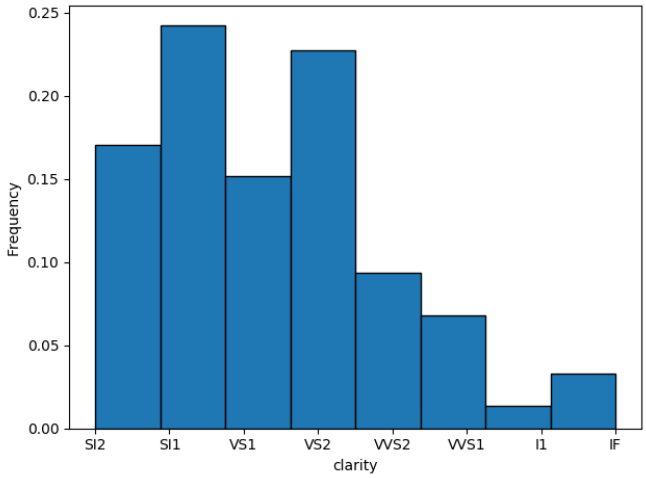
Iš šios histogramos matome, kad ,,cut“ atributo duomenys pasiskirstę bimoduliai.

Color atributo histograma:



Iš šios histogramos matome, kad atributo ,,color“ duomenys pasiskirstę bimoduliai.

Clarity atributo histograma:



Iš šios histogramos matome, kad atributo ,,clarity“ duomenys pasiskirstę left-skewed.

2.4. Problemų identifikavimas

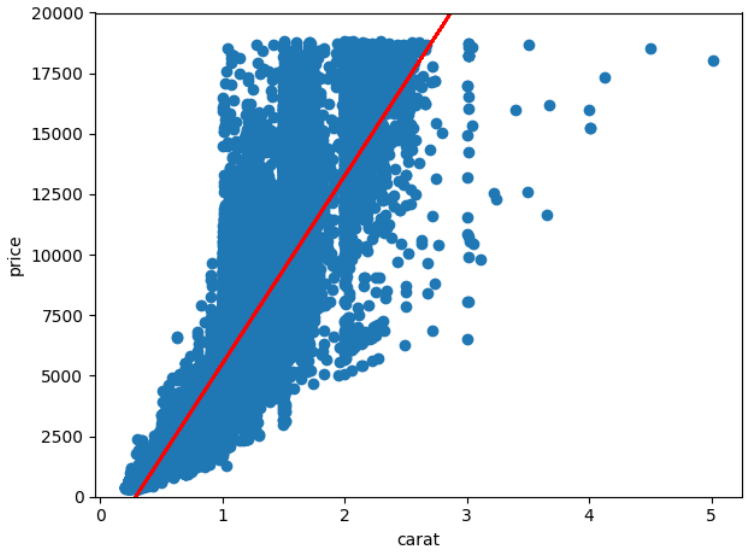
Iš histogramų matome, kad atributai: ,,depth_pct“, ,,table“, ,,length“ turi ekstremalias reikšmes. Taip pat atributai ,,length“, ,,width“ ir ,,depth“ turi reikšmių lygių 0. Šie įrašai bus išmetami.

Iš viso pašalinta tik 20 įrašų.

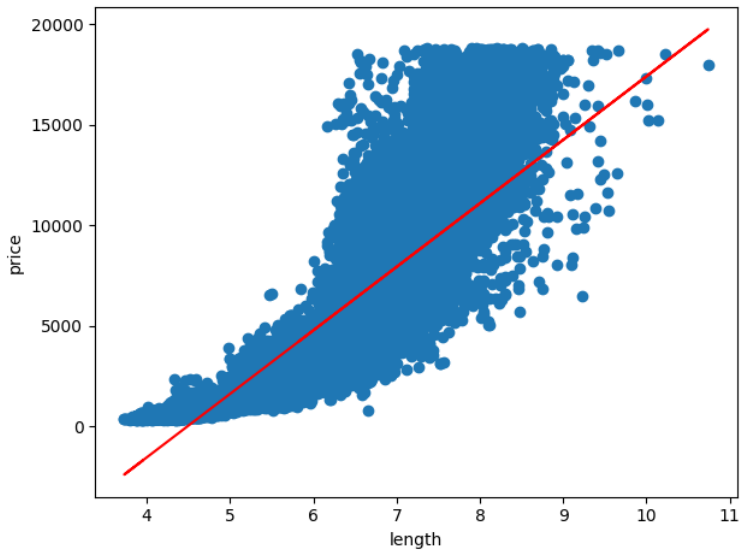
2.5. Nustatyti sąryšius tarp atributų panaudojant vizualizacijos būdus

Tolydinio tipo atributams:

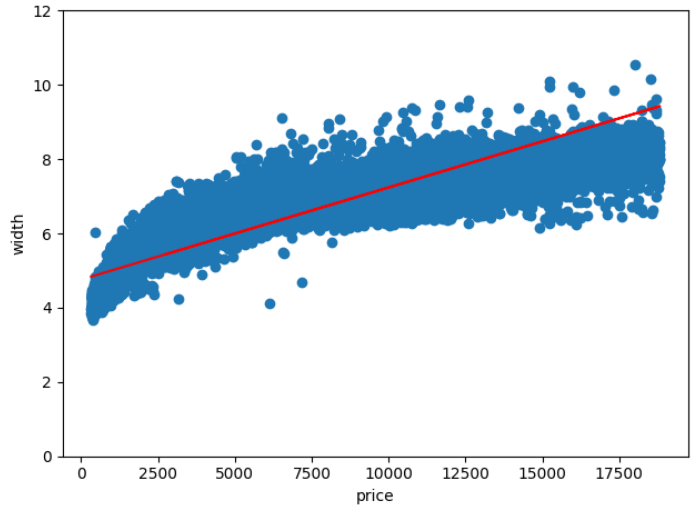
Tarp ,,carat“ ir ,,price“:



Tarp ,,length“ ir ,,price“:

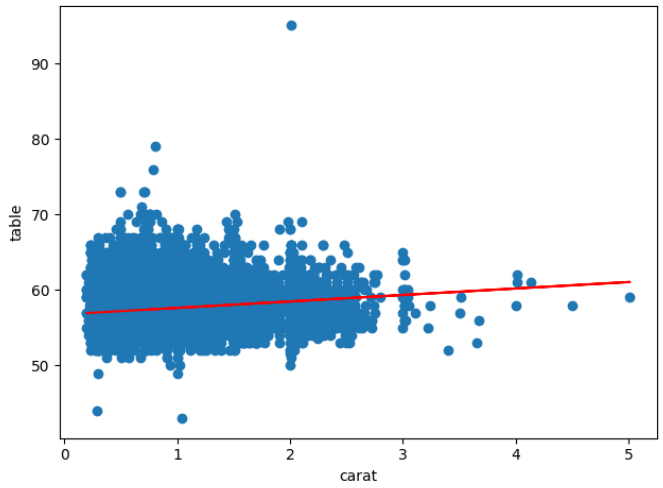


Tarp ,,width“ ir ,,price“:

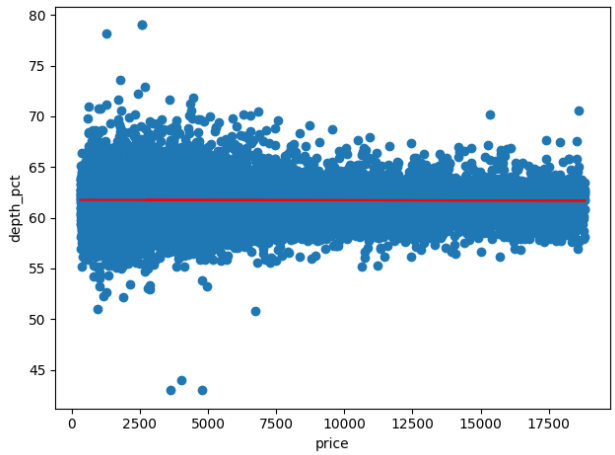


Iš šių histogramų matome, kad atributas ,,price“ labai priklauso nuo ,,,carat“, ,,length“ ir ,,width“ reikšmių. Labiausiai susiję yra ,,price“ ir ,,width“.

Tarp ,,table“ ir ,,carat“

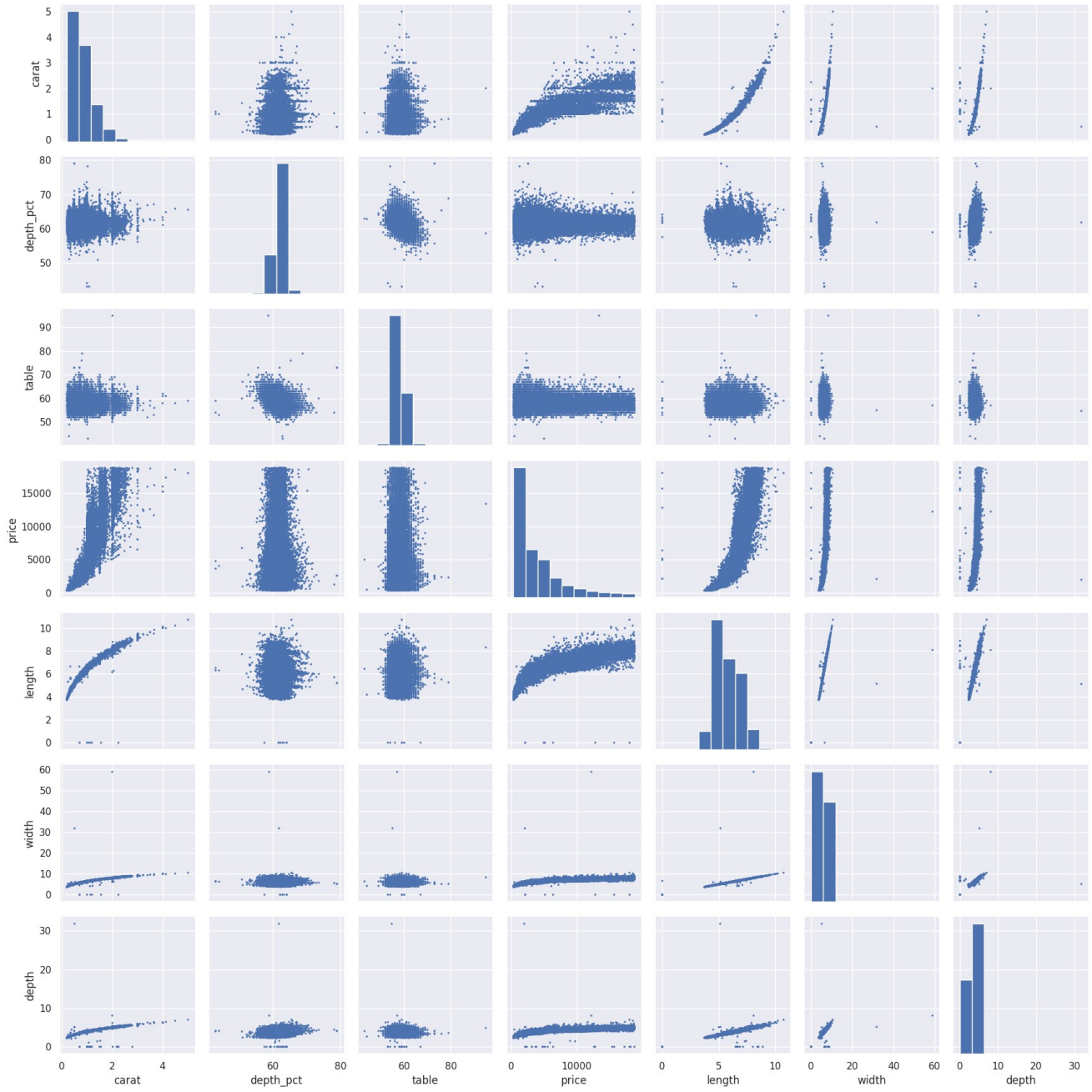


Tarp ,,depth_pct“ ir ,,price“:



Matome, kad atributai ,,table“ ir ,,carat“, ,,depth_pct“ ir ,,price“ yra silpnai tarpusavyje koreliuojantys.

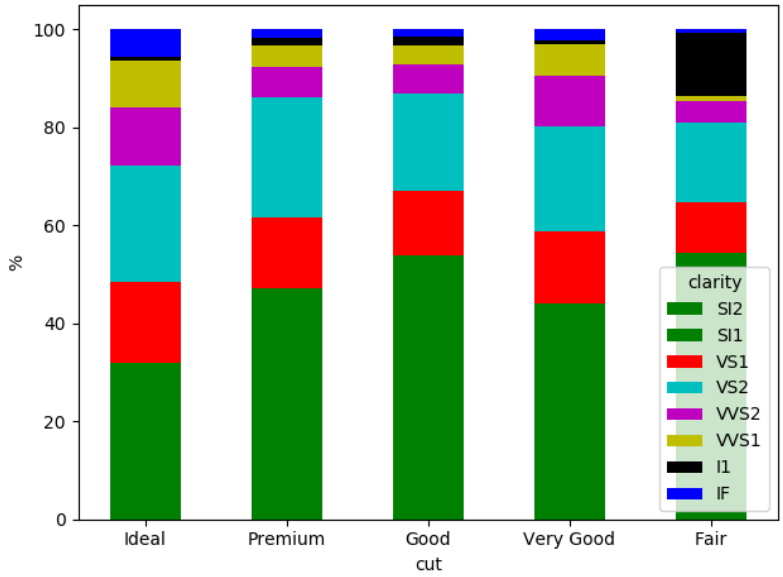
SPLOM diagrama:



Kategorinių atributų priklausomybė:

Tarp ,,cut“ ir „clarity“:

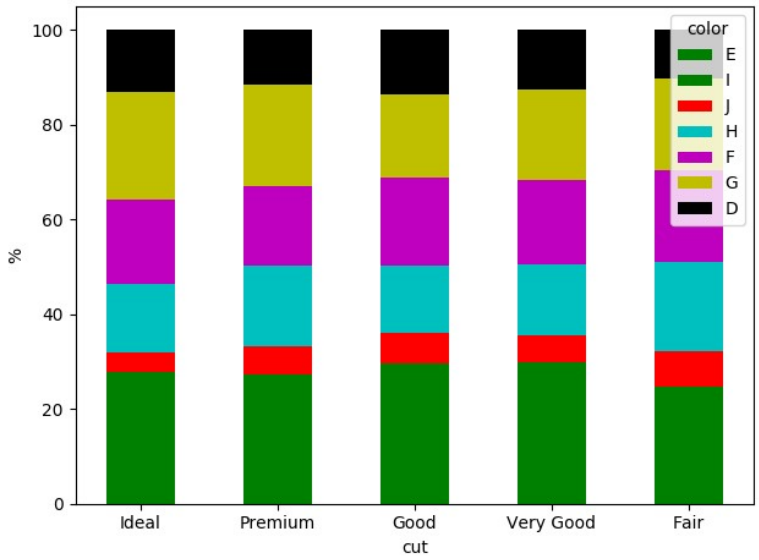
1 etapas:



Iš šios diagramos matome, kad atributo ,,cut“ reikšmė kai yra ,,Ideal“ turi įtakos geriausiam skaidrumui (IF).

O esant reikšmei ,,Fair“ taip pat smarkiai įtakoją blogiausią skaidrumą (I1).

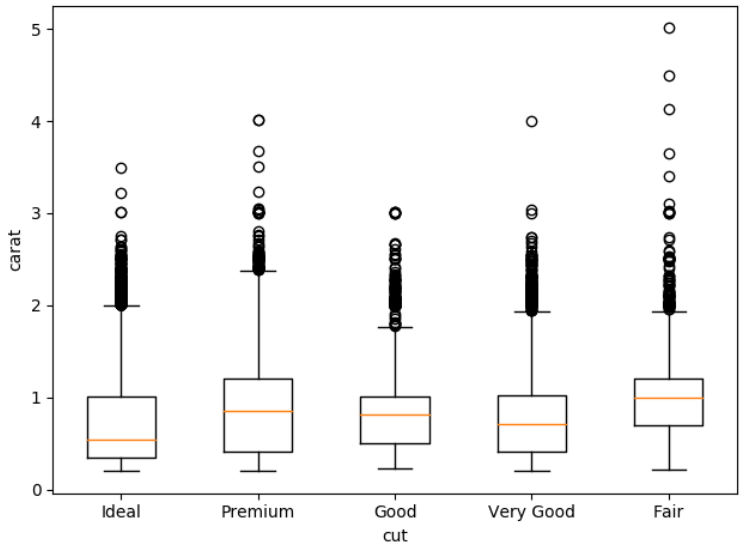
Tarp ,,cut“ ir „color“:



Iš diagramos matome, kad šie atributai tarpusavyje nekoreliuoja.

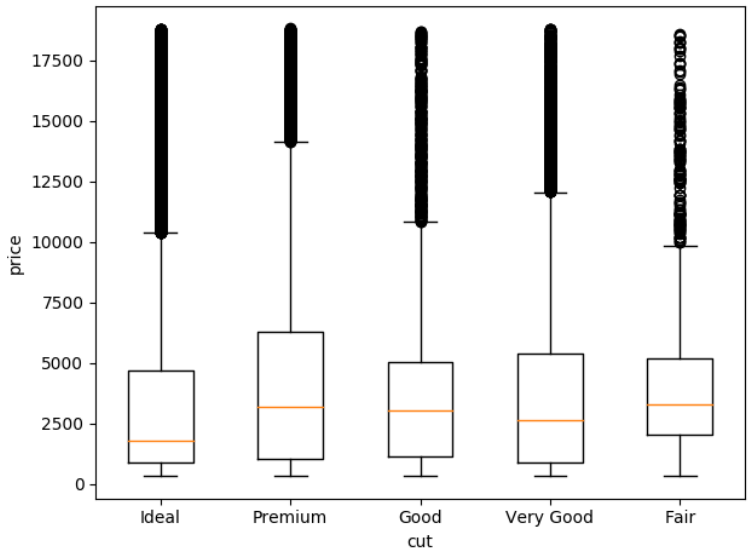
Priklausomybė tarp tolydžių ir kategorinių atributų:

Tarp ,,carat“ ir ,,cut“:



Iš šios diagramos matome, kad ryšis tarp atributų ,,carat“ ir ,,cut“ yra labai silpnas.

Tarp ,,price“ ir ,,cut“:



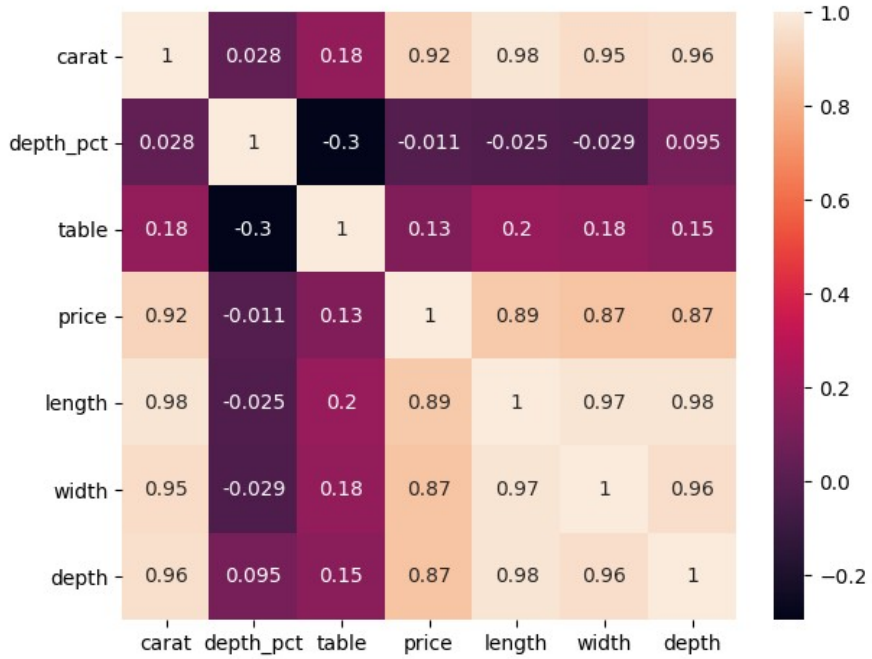
Iš šios diagramos matome, kad tarp šių atributų yra tik silpnas ryšis. Taip pat matome, kad premium ir fair tipo deimantai yra paprastai brangiausi.

2.6. Paskaičiuoti kovariacijos ir koreliacijos reikšmes. Grafiškai atvaizduoti koreliacijos matricą

Kovariacijos matrica:

	carat	depth_pct	table	price	length	width	depth
carat	1	0.02	0.19	1742.77	0.52	0.52	0.32
depth_pct	0.02	1	-0.95	-60.85	-0.04	-0.05	0.10
table	0.19	-0.95	1	1133.32	0.49	0.47	0.24
price	1742.77	-60.85	1133.32	1	3958.02	3943.27	2424.71
length	0.52	-0.04	0.49	3958.02	1	1.25	0.77
width	0.52	-0.05	0.47	3943.27	1.25	1	0.77
depth	0.32	0.10	0.24	2424.71	0.77	0.77	1

Koreliacijos matricos grafikas:



Iš šios koreliacijos matricos diagramos matome, kad kainą labiausiai įtakoją atributai: „carat“, „length“, „width“, „depth“. Labai silpnas ryšys su „depth_pct“ ir „table“. Taigi matome, kad deimanto kainų prognozei svarbiausi deimanto išmatavimai ir karatų skaičius.

2.7. Atlikti duomenų normalizaciją

Tolydus atributai buvo normalizuoti į [0; 1] rėžius.