Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Intelektikos pagrindai

(P176B101)

Laboratorinis darbas Nr. 1

Darbą atliko:

MGTM – 9/1 gr. studentas

Tauras Gaulia

Darbą priėmė:

lekt. **Germanas Budnikas**

doc. **Agnė Paulauskaitė-Tarasevičienė**

Kaunas, 2021

Duomenų rinkinio kokybės analizė

Pasirinktas duomenų rinkinys apibūdina tūkstančio JAV aukštųjų mokyklų studentų pasirodymą standartizuotuose testuose. Rinkinio atributai nurodo studentų lytį, tautybę, tėvų išsilavinimo lygį, gaunamus pietus, pasiruošimą testui bei jo rezultatus.

Atributas	Tipas	Prasmė	Pavyzdys
Gender	Kategorinis	Lytis	Male
Race/ethnicity	Kategorinis	Studento grupė pagal jo tautybę	Group A
Parental level of education	Kategorinis	Tėvų išsilavinimo lygis	Bachelor's degree
Lunch	Kategorinis	Studento gaunami pietūs	Standard
Test preparation course	Kategorinis	Pasirengimo testui kursas	Completed
Math score	Tolydinis	Rezultatas matematikos teste	72
Reading score	Tolydinis	Rezultatas skaitymo teste	72
Writing score	Tolydinis	Rezultatas rašymo teste	74
Average score	Tolydinis	Testo rezultatų vidurkis	72,67

Tolydinio tipo atributai:

Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Minimali reikšmė	Maksimali reikšmė	1-asis kvartilis	3-asis kvartilis	Vidurkis	Mediana	Standartinis nuokrypis
Math score	1000	0	81	0	100	57	77	66.089	66	15.16308
Reading score	1000	0	72	17	100	59	79	69.169	70	14.600192
Writing score	1000	0	77	10	100	57.75	79	68.054	69	15.195657
Average score	1000	0	194	9	100	58.33	77.67	67.77058	68.33	14.257311

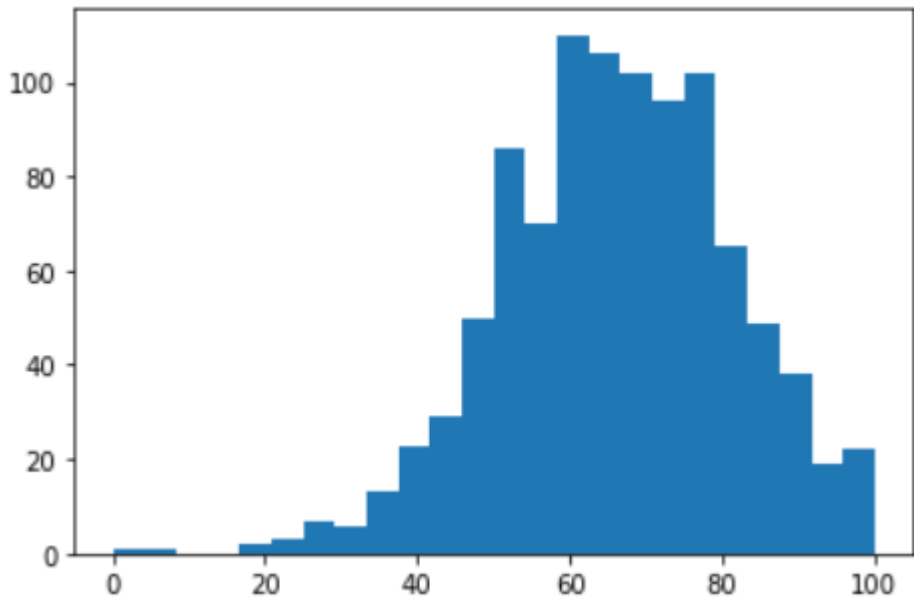
Kategorinio tipo atributai:

Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Moda	Modos dažnumas	Moda, %	2-oji Moda	2-osios Modos dažnumas	2-oji Moda, %
Gender	1000	0	2	female	518	51.8	male	482	48.2
Race/ethnicity	1000	0	5	group C	319	31.9	group D	262	26.2
Parental level of education	1000	0	6	some college	226	22.6	associate‘s degree	222	22.2
Lunch	1000	0	2	standard	645	64.5	free/reduced	355	35.5
Test preparation course	1000	0	2	none	642	64.2	completed	358	35.8

Histogramos

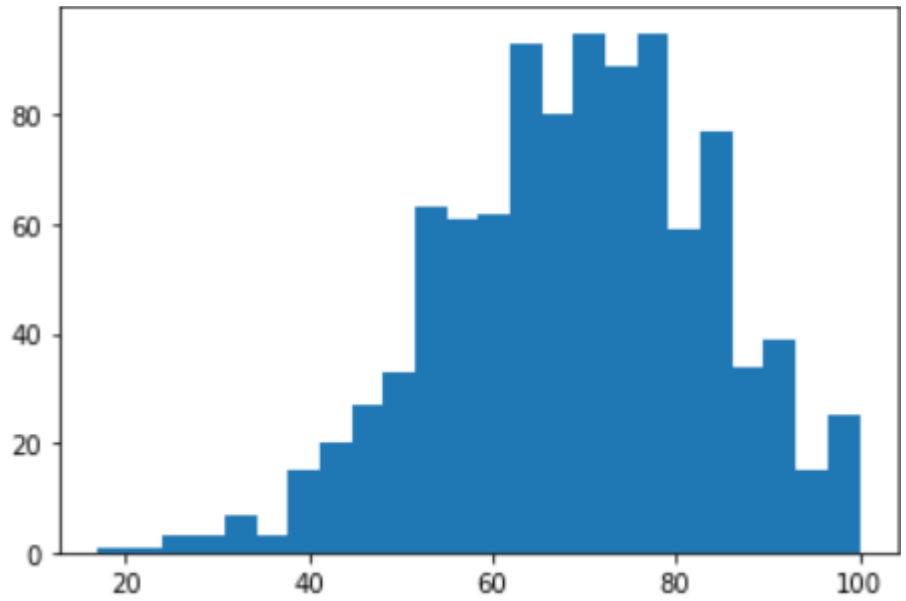
Tolydinio tipo atributai:

„Math score“ atributo histograma



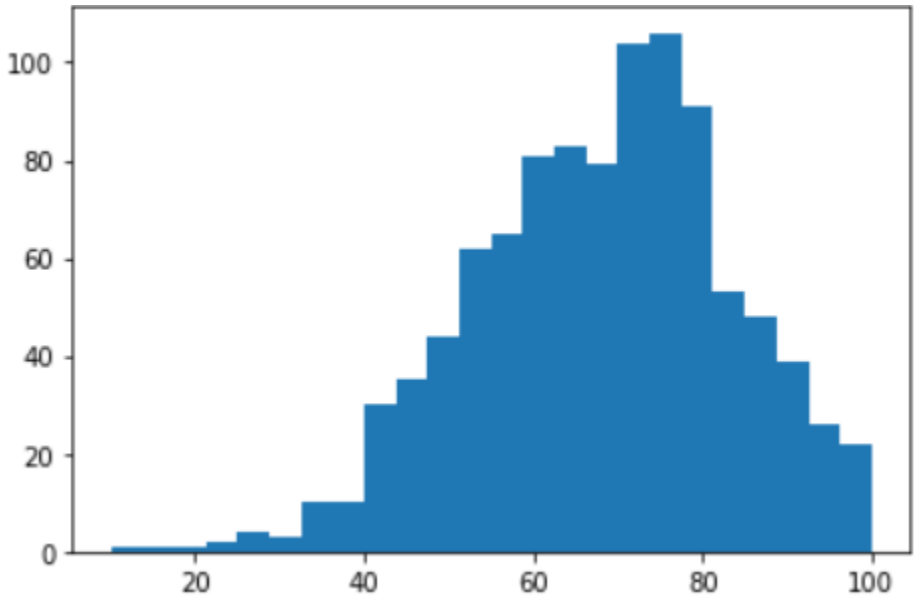
Atributo „Math score“ histograma rodo normalųjį pasiskirstymą šiek tiek pasislinkusį į dešinę.

„Reading score“ atributo histograma



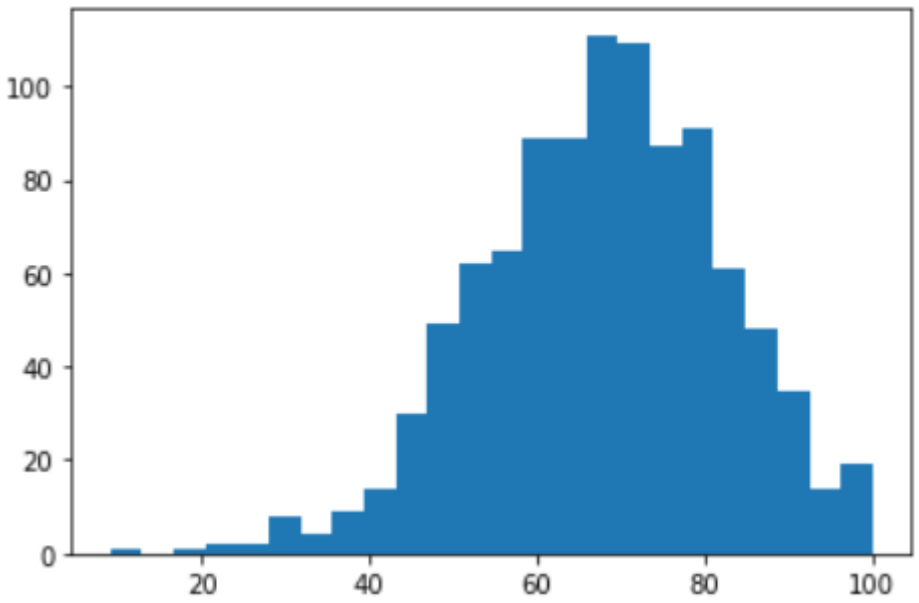
Atributo „Reading score“ histograma rodo normalųjį pasiskirstymą šiek tiek pasislinkusį į dešinę.

„Writing score“ atributo histograma



Atributo „Writing score“ histograma rodo normalųjį pasiskirstymą šiek tiek pasislinkusį į dešinę.

„Average score“ atributo histograma



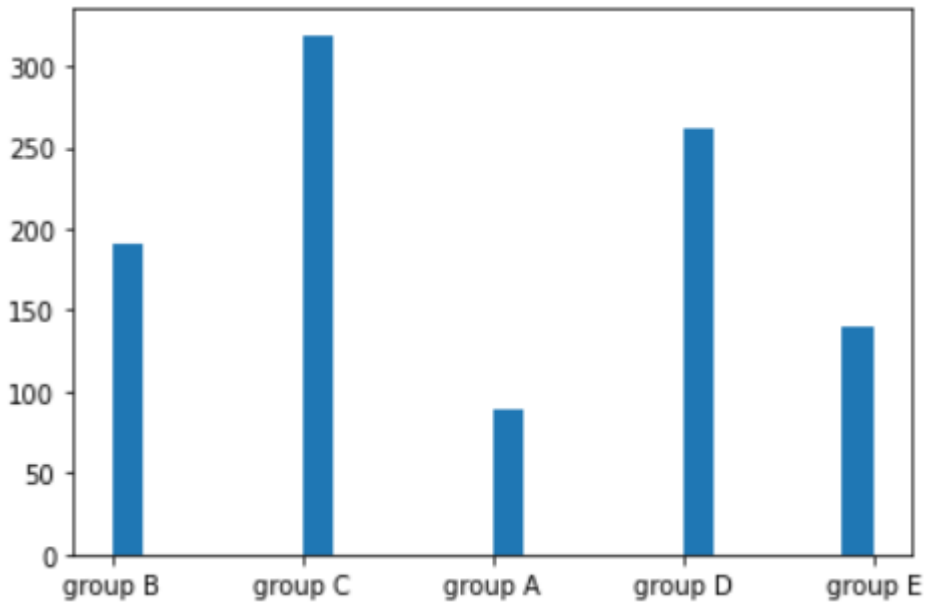
Atributo „Average score“ histograma rodo normalųjį pasiskirstymą šiek tiek pasislinkusį į dešinę. Taigi tolydinio tipo atributai yra normaliai pasiskirstę.

Kategorinio tipo atributai:

„Gender“ atributo histograma

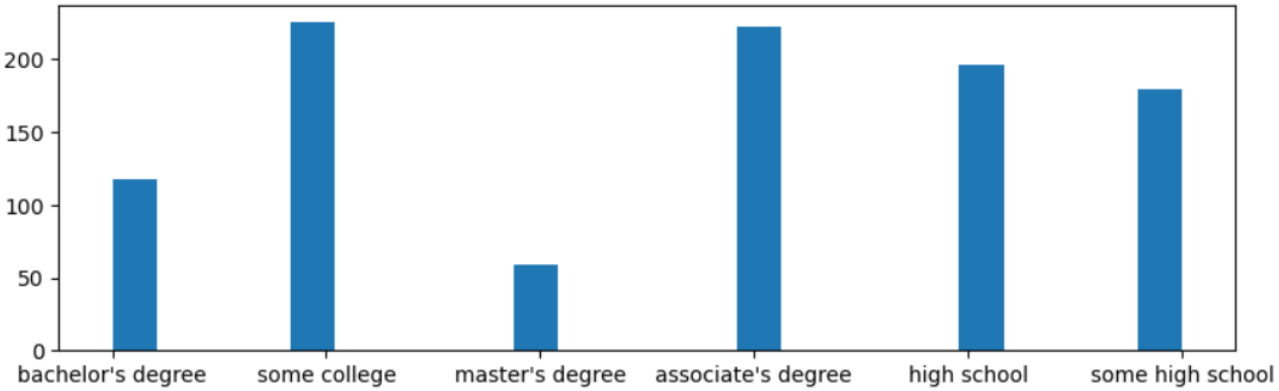
Atributo „Gender“ reikšmės pasiskirsčiusios gan tolygiai.

„Race/ethnicity“ atributo histograma



Atributo „Race/ethnicity“ reikšmės yra pasiskirsčiusios netolygiai. C ir D grupių reikšmių yra daugiausiai, o A grupės – mažiausiai.

„Parental level of education“ atributo histograma



Atributo „Parental level of education“ reikšmės yra pasiskirsčiusios netolygiai. „Some college“ bei „Associate‘s degree“ reikšmių yra daugiausiai, o „Master‘s degree“ – mažiausiai.

„Lunch“ atributo histograma

Iš šios histogramos matome, jog „standard“ reikšmių yra daugiau.

„Test preparation course“ atributo histograma

Iš šios histogramos matome, jog daugiau studentų nebuvo atlikę pasiruošimo testui kurso.

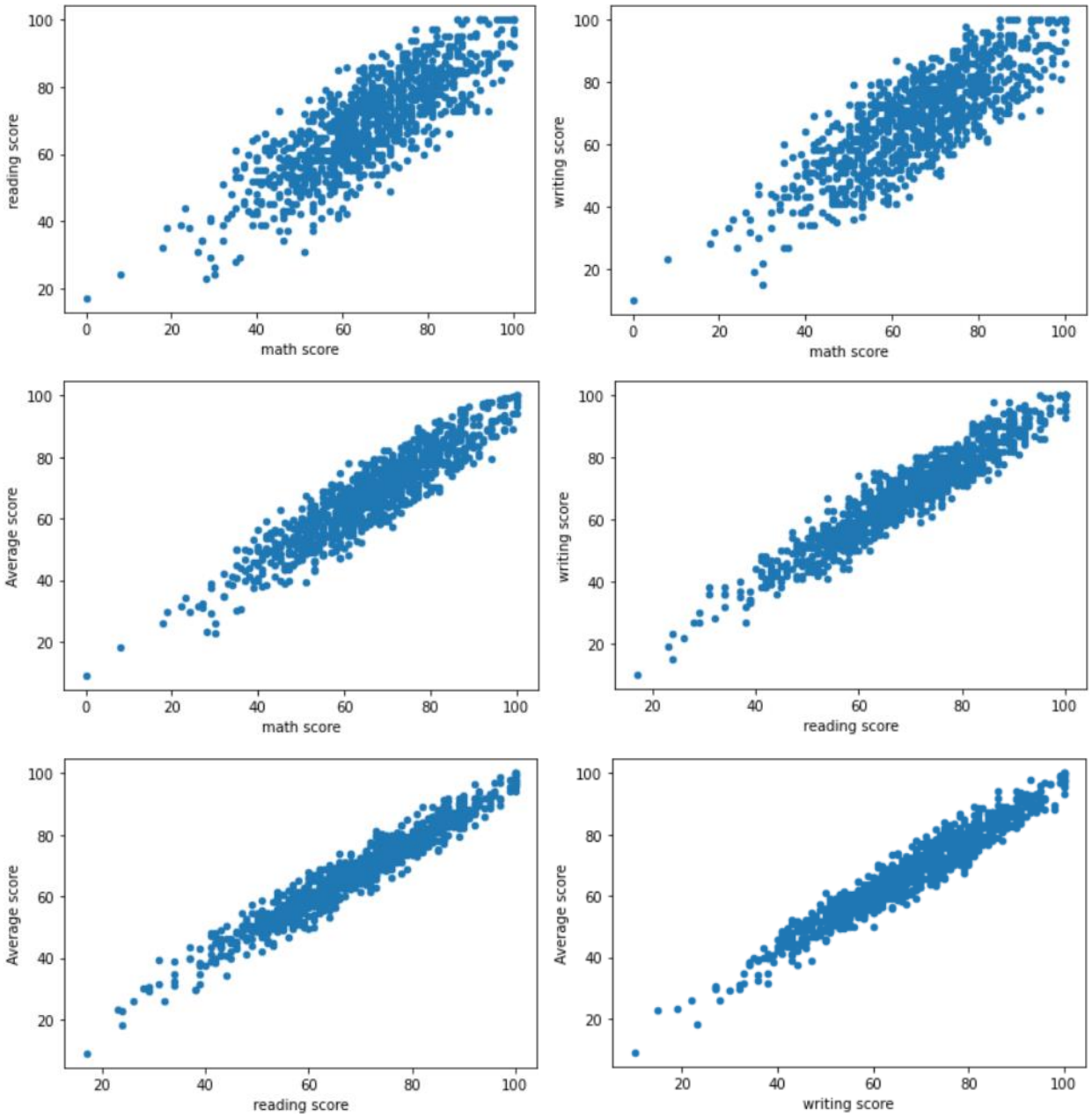
Duomenų kokybės problemų identifikacija

Duomenų rinkinys neturi trūkstamų reikšmių. Kategorinio tipo atributų kardinalumas yra mažas. Tolydinio tipo atributų kardinalumas nėra labai didelis lyginant su eilučių kiekiu, tačiau atributai yra intervale nuo 0 iki 100, tad kardinalumas nelabai ir gali būti didesnis.

Nors šiame duomenų rinkinyje ekstremalių reikšmių nebuvo, visvien tikriname, ar „Math score“, „Reading score“ bei „Writing score“ atributų reikšmės nėra didesnės nei 100 ir nėra mažesnės nei 0. Jei reikšmės nepatenka į intervalą, jos pakeičiamos į atitinkamo stulpelio modą.

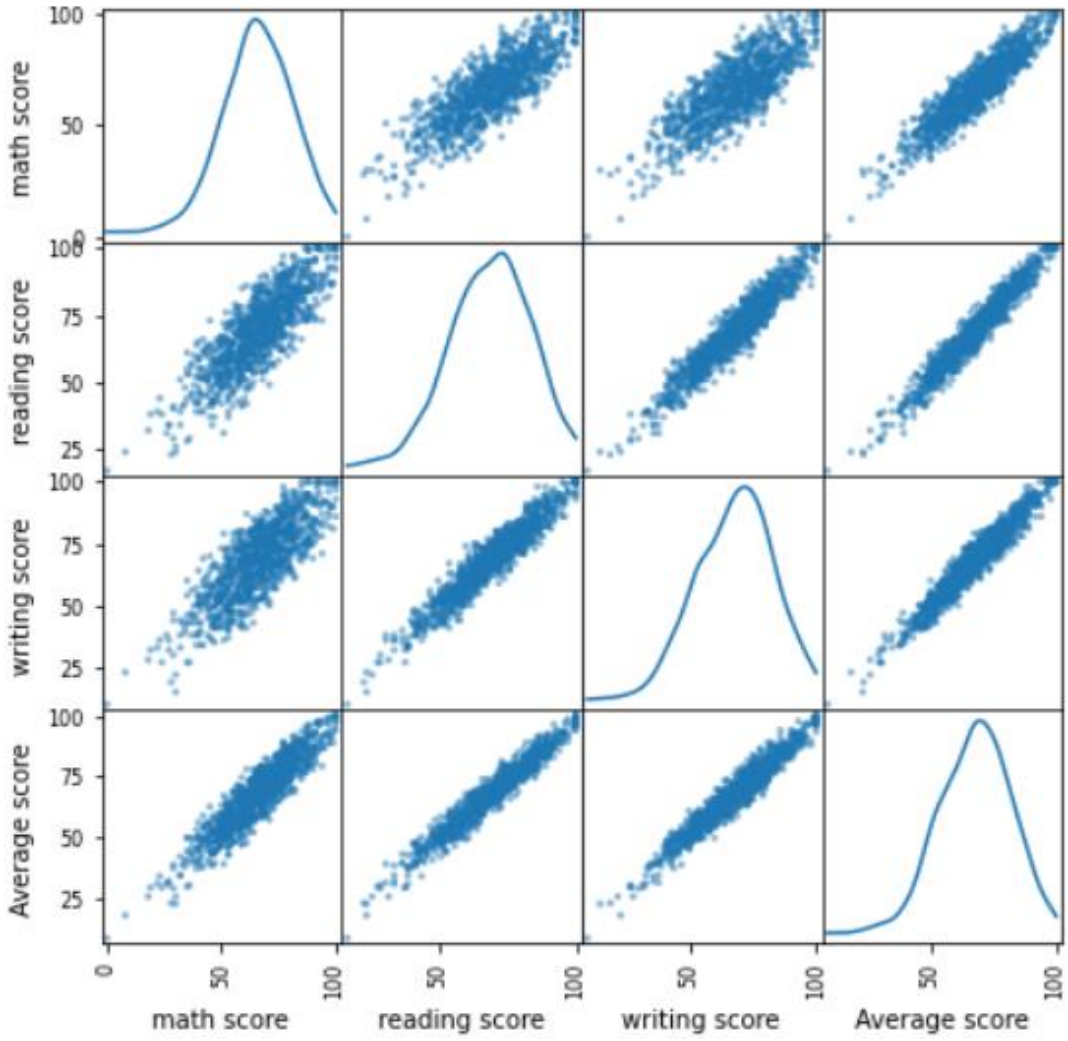
Sąryšių tarp atributų vizualizacija

Tolydinio tipo atributai:



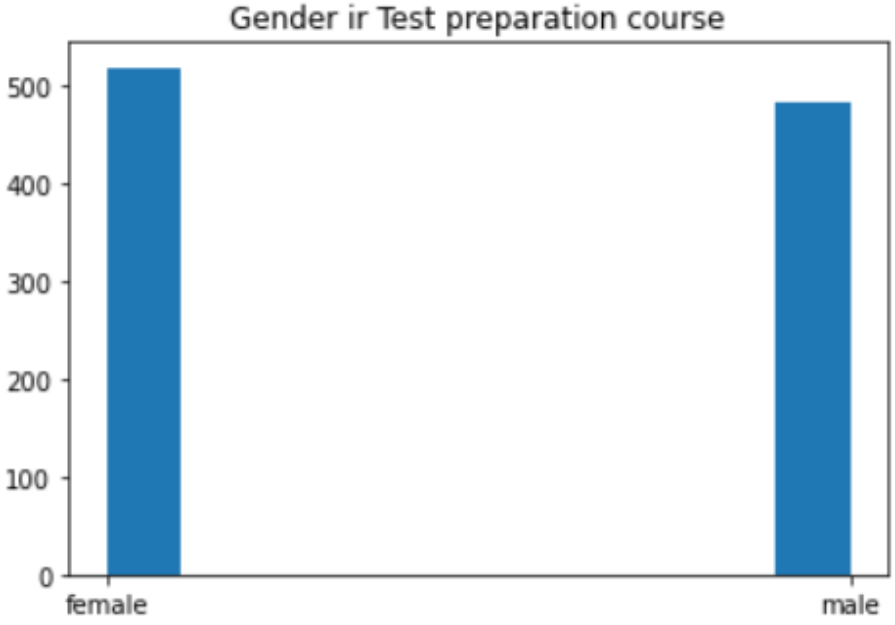
Kadangi tolydinio tipo atributai yra testų rezultatai, iš diagramų galima pastebėti stiprias tiesines priklausomybes. Taigi visi tolydinio tipo atributai yra koreliuoti.

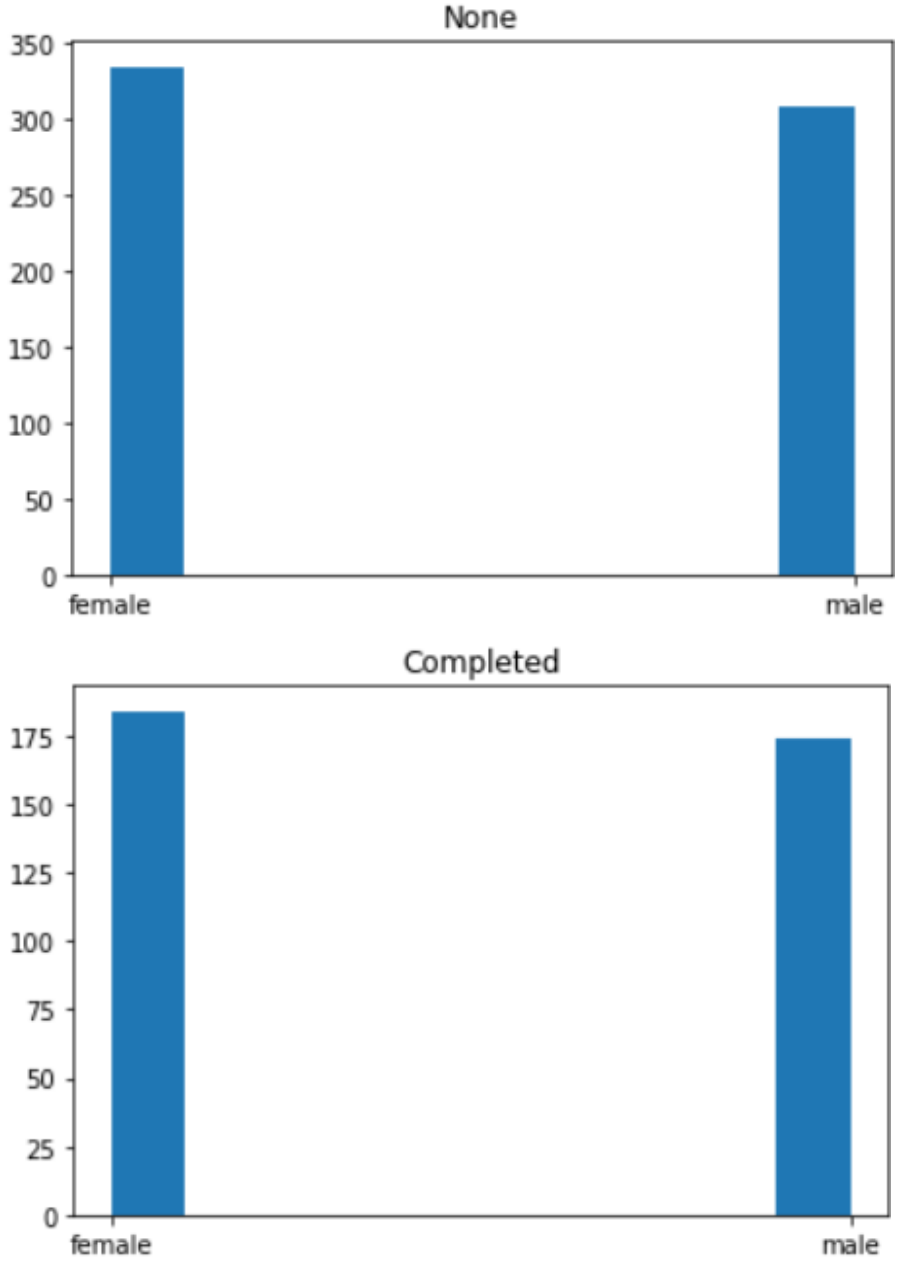
SPLOM diagrama:



Kategorinio tipo atributai:

Tiriamas „Gender“ ir „Test preparation course“ sąryšis

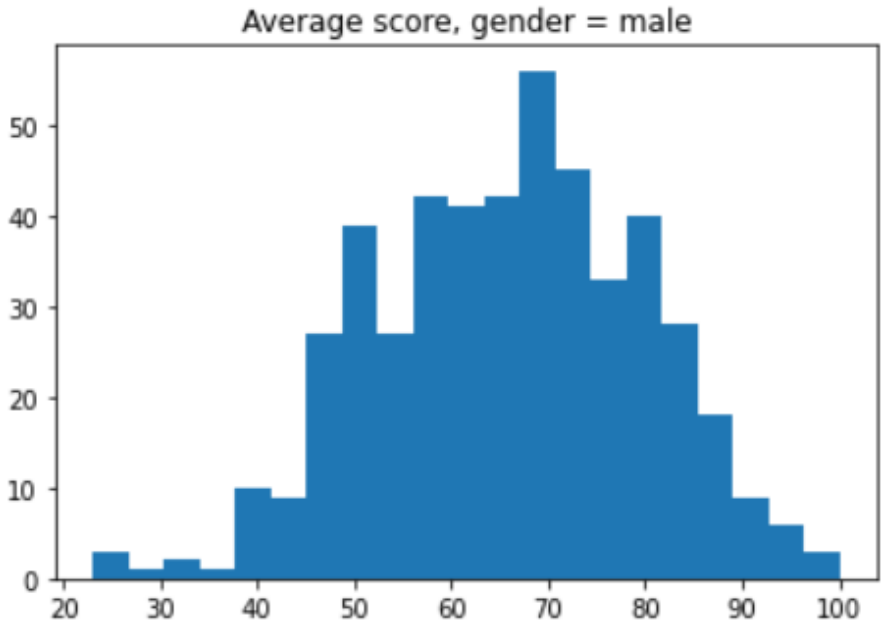


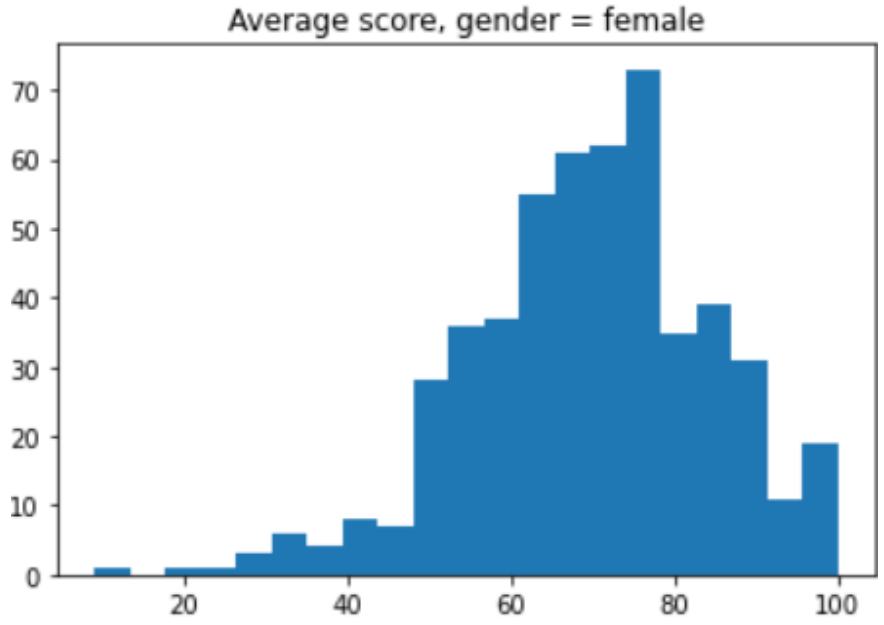


Kadangi abiejų atributų stulpeliai mažai kinta diagramose, ryšys tarp šių atributų yra silpnas. Testui ruošėsi ir merginos, ir vaikinai labai panašiu mastu.

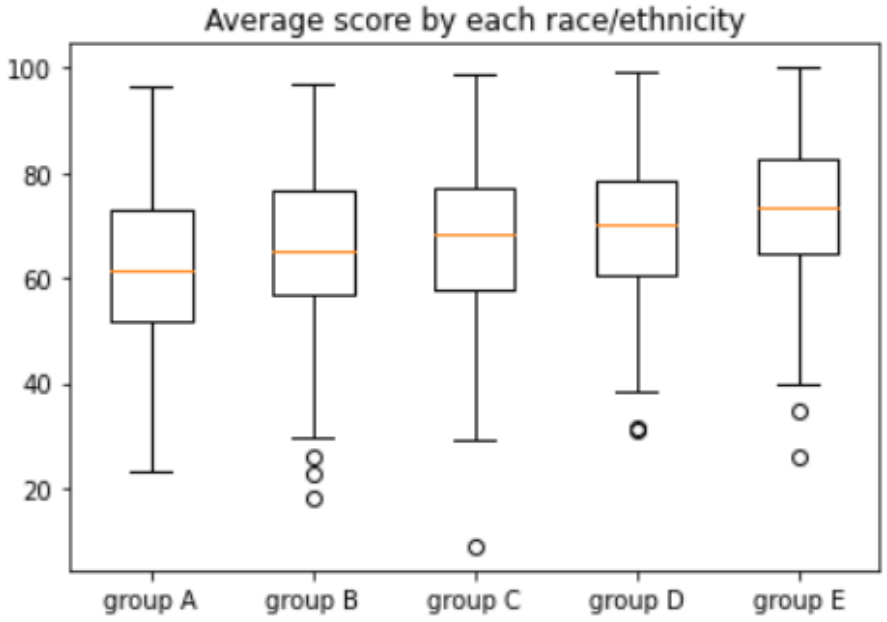
Sąryšiai tarp kategorinio ir tolydinio tipo kintamųjų:

Histogramos vaizduoja vyrų bei moterų testų rezultatų vidurkių pasiskirstymą. Iš šių histogramų galima matyti, jog moterų histograma yra labiau pasislinkus į dešinę, tad moterys testus atliko geriau.

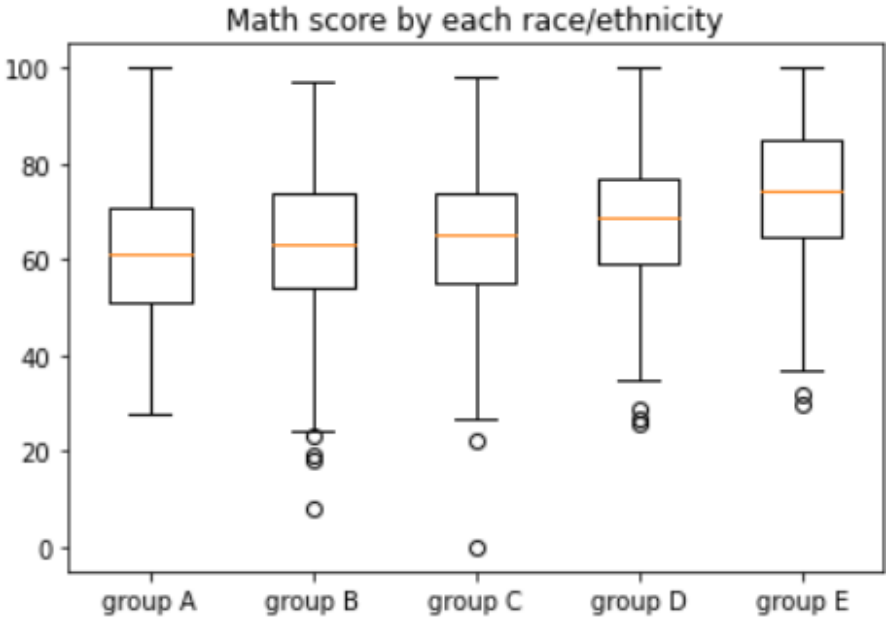




Diagramoje pavaizduota „Average score“ ir „Race/ethnicity“ atributų priklausomybė. Iš šios „box plot“ tipo diagramos galima pastebėti, jog E grupė testuose vidutiniškai yra surinkusi geriausius rezultatus, o A grupė – prasčiausius.



Diagramoje pavaizduota „Math score“ ir „Race/ethnicity“ atributų priklausomybė. Iš šios „box plot“ tipo diagramos galima pastebėti, jog, kaip ir praeitoje diagramoje, E grupė matematikos testuose vidutiniškai yra surinkusi geriausius rezultatus, o A grupė – prasčiausius. Stipraus ryšio tarp atributų abejose diagramose nėra.



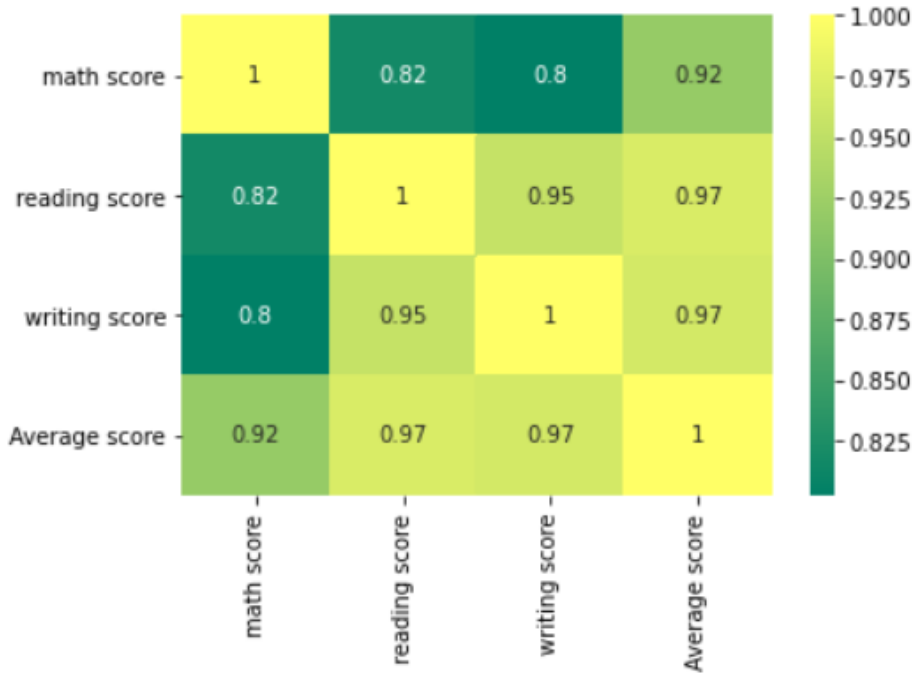
Kovariacija

	math score	reading score	writing score	Average score
math score	229.918998	180.998958	184.939133	198.618517
reading score	180.998958	213.165605	211.786661	201.983495
writing score	184.939133	211.786661	230.907992	209.211360
Average score	198.618517	201.983495	209.211360	203.270911

Koreliacija

	math score	reading score	writing score	Average score
math score	1.000000	0.817580	0.802642	0.918744
reading score	0.817580	1.000000	0.954598	0.970331
writing score	0.802642	0.954598	1.000000	0.965669
Average score	0.918744	0.970331	0.965669	1.000000

Koreliacijos matrica



Iš kovariacijos ir koreliacijos reikšmių galima pastebėti, jog visi tolydinio tipo atributai yra koreliuoti. Vadinasi, studentas atlikęs vieną testą puikiu rezultatu, gali tikėtis neblogų rezultatų ir iš kitų atsiskaitymų. Labiausiai priklausomi atributai yra „Writing score“ ir „Reading score“. Šiuose atributuose studentų rezultatai yra panašiausi.

Kategorinio tipo kintamųjų vertimas į tolydžiojo tipo atributus

Prieš atliekant duomenų normalizaciją kategorinio tipo atributai („Gender“, „Race/ethnicity“, „Parental level of education“, „Test preparation course“, „Lunch“) verčiami atrenkant unikalias reikšmes, jas sunumeruojant ir priskiriant vietoj buvusių tekstinių reikšmių. Pavyzdžiui, po keitimo „Gender“ unikalios reikšmės yra „0“ ir „1“.

Duomenų normalizacija

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	Average score
0	0.0	0.00	0.0	0.0	0.0	0.72	0.662651	0.711111	0.699670
1	0.0	0.25	0.2	0.0	1.0	0.69	0.879518	0.866667	0.805824
2	0.0	0.00	0.4	0.0	0.0	0.90	0.939759	0.922222	0.919451
3	1.0	0.50	0.6	1.0	0.0	0.47	0.481928	0.377778	0.443187
4	1.0	0.25	0.2	0.0	0.0	0.76	0.734940	0.722222	0.739890
...	...	...	...	...	...	...	...	...	...
995	0.0	1.00	0.4	0.0	1.0	0.88	0.987952	0.944444	0.934066
996	1.0	0.25	0.8	1.0	0.0	0.62	0.457831	0.500000	0.531099
997	0.0	0.25	0.8	1.0	1.0	0.59	0.650602	0.611111	0.615385
998	0.0	0.75	0.2	0.0	1.0	0.68	0.734940	0.744444	0.717912
999	0.0	0.75	0.2	1.0	0.0	0.77	0.831325	0.844444	0.813187

Po duomenų normalizacijos visų kintamųjų reikšmės priklauso intervalui [0;1]. Tai galime matyti iš normalizuoto duomenų rinkinio didžiausių ir mažiausių atributų reikšmių:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	Average score
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	0.482000	0.460750	0.537800	0.355000	0.358000	0.660890	0.628542	0.645044	0.645831
std	0.499926	0.342621	0.337231	0.478753	0.479652	0.151631	0.175906	0.168841	0.156674
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.250000	0.200000	0.000000	0.000000	0.570000	0.506024	0.530556	0.542088
50%	0.000000	0.250000	0.600000	0.000000	0.000000	0.660000	0.638554	0.655556	0.651978
75%	1.000000	0.750000	0.800000	1.000000	1.000000	0.770000	0.746988	0.766667	0.754615
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Išvados

Atlikta pasirinkto duomenų rinkinio kokybės analizė. Šis duomenų rinkinys neturėjo trūkstamų ar ekstremalių reikšmių. Tolydaus tipo atributų reikšmės priklausė intervalui [0;100]. Visi šie atributai buvo koreliuoti.

Dėl galimų klaidų buvo tikrinama ar tolydaus tipo atributų reikšmės buvo realios. Jei jos viršijo 100 arba buvo mažesnės nei 0, reikšmės buvo keičiamos į modą.

Sąryšiams tarp atributų tirti buvo braižomos „bar plot“ bei „box plot“ tipo diagramos.

Darbo pabaigoje kategorinio tipo kintamieji buvo paversti į tolydžiojo tipo atributus ir duomenys normalizuoti.