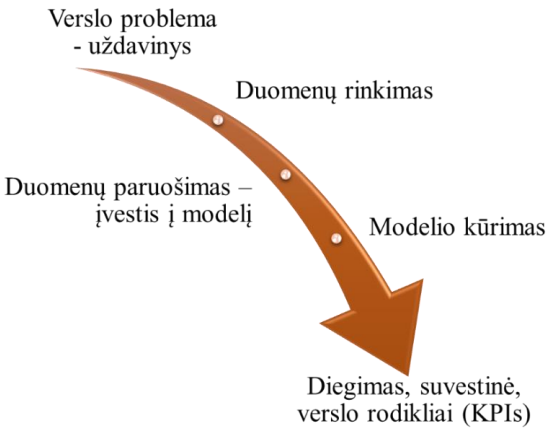
Tiriamoji duomenų analizė

I dalis

Verslas neišvengiamai susiduria su tam tikromis problemomis, kurias reikia spręsti, ar priešingai, tiesiog siekia veikti sėkmingiau, efektyviau, gerinti apsibrėžtus rodiklius. Abiem atvejais gali būti „įdarbinami“ duomenys. Tai apima duomenų rinkimą iš įvairių šaltinių ir sukauptų duomenų paruošimą analizei bei modelio kūrimui (1 pav.).



1 pav. *Duomenų paruošimas sprendžiant verslo problemą*

Verslo problemą transformuoti į spręstiną uždavinį nėra taip paprasta. Tam reikia apsibrėžti:

- Kokia yra verslo problema?
- Kokie yra tikslai (rodikliai), kuriuos verslas nori pasiekti?
- Kaip verslas veikia šiuo metu?
- Kokiu būdu modelio prognozės padės spręsti verslo problemą?

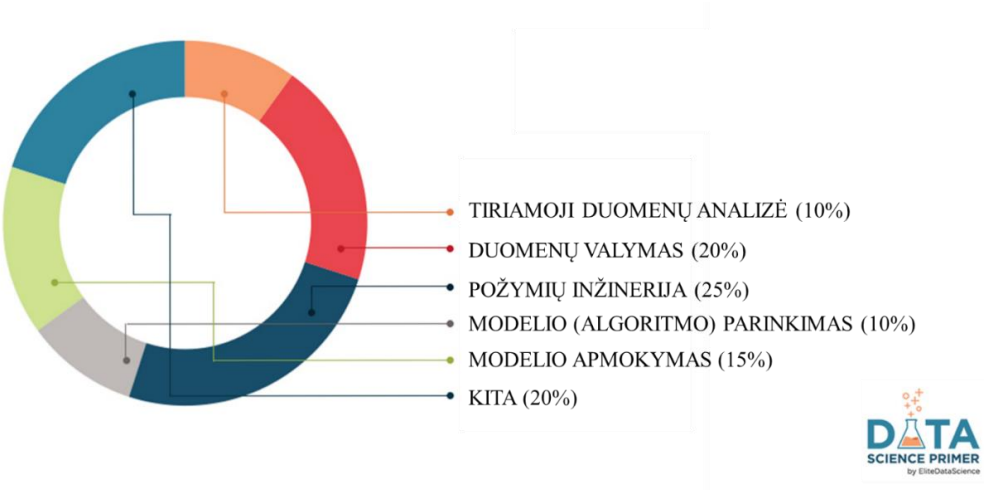
Pavyzdžiui, draudimo bendrovė siekia palengvinti apgaulingų draudiminių žalų skyriaus darbą ir siekia šį procesą kiek įmanoma automatizuoti įdiegdama dirbtinio intelekto modelį. Kuo ar kaip modelio prognozės gali būti naudingos? Pavyzdžiui, prognozės gali būti naudingos keliais lygmenimis:

- apgaulingų ir neapgaulingų žalų dengimo ieškinių prognozavimas;
- netipinės naujų klientų elgsenos prognozavimas;
- esamų klientų apgaulingų veiksmų prognozavimas.

Yra daug būdų, kaip panaudojant dabartinių informacinių sistemų, daviklių, daiktų internetą galima kaupti duomenis. Taip pat reiktų nepamiršti ir išorinių duomenų šaltinių (dalis jų jau tapo mokami), ir tokių duomenų kaip medija, socialiniai tinklai (2 pav.).

2 pav. Duomenų rinkimas sprendžiant verslo problemą

Kyla klausimas, o kas toliau? Pagal pateiktą grafiką, matome, jog su duomenimis susiję darbai – tiriamoji duomenų analizė, duomenų valymas, požymių inžinerija – užima daugiau nei pusę viso laiko; ir tai kur kas daugiau nei reikia laiko resursų modeliui sukurti (3 pav.). Priežastis ta, jog kurdami pati tobuliausią modelį naudojant nekokybiškus ar su sprendžiama problema menkai susijusius duomenis, patikimų ir naudingų prognozių neturime tikėtis, ir tai tebus veltui iššvaistytas laikas.



3 pav. *Duomenų paruošimas* (paruošta pagal šaltinį elitedatascience.com/birds-eye-view)

Ko gero, nėra vieno geriausio būdo duomenų imties kokybei apibendrinti ar griežtų reikalavimų. Dažniausiai yra sudaromos tam tikros lentelės ar braižomi grafikai, kuriuos nagrinėjant būtų galima įžvelgti duomenyse esančias problemas ar tiesiog gerai susipažinti su imtyje esančiais požymiais. Pavyzdžiui, dvi atskiros lentelės galėtų būti sudaromos atitinkamai tolydiesiems požymiams (

1 lentelė.) ir kategoriniams požymiams (2 lentelė.).

1 lentelė. *Tolydžiųjų požymių charakteristikos*

TOLYDIEJI POŽYMIAI (KINTAMIEJI)

Požymis	Kiekis	Trūkstamos reikšmės, %	Kardina-lumas	Vidurkis	Mediana	Min	Max	Q1	Q3	Standartinis nuokrypis	Asimetriš-kumas	…
…	…	…	…	…	…	…	…	…	…	…	…	…

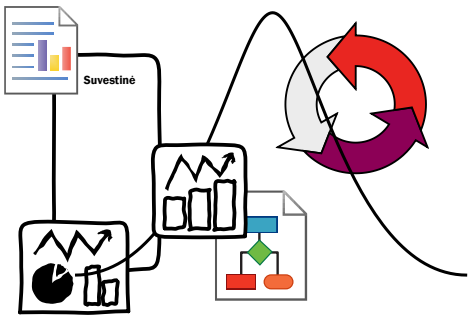
2 lentelė. *Kateogorinių požymių charakteristikos*

KATEGORINIAI POŽYMIAI (DISKRETIEJI KINTAMIEJI)

Požymis	Kiekis	Trūkstamos reikšmės, %	Kardina-lumas	Moda	Moda, %	2-oji moda	2-oji moda, %	…
…	…	…	…	…	…	…	…	…

Lentelėse dažniausiai pateikiamas duomenų kiekis, trūkstamų reikšmių procentas, stebėjimų centravimosi ir sklaidos charakteristikos. Nuo duomenų tipo priklauso skaičiuojamos skirtingos charakteristikos. Jei kintamieji yra tolydieji, tai centravimąsi nusako vidurkis, mediana; sklaidą charakterizuoja min, max ar skirtumas tarp jų, taip pat kvartiliai (Q1 ir Q3), standartinis nuokrypis. Kardinalumas nusako, kiek unikalių reikšmių yra imtyje. Asimetriškumas apibrėžia duomenų pasiskirstymo (a)simetriją. Kategoriniams kintamiesiems dažnai skaičiuojama moda, galimai ir 2-oji moda (sekantis požymis pagal pasikartojimo dažnį). Tokios sklaidos charakteristikos kaip kvartiliai irgi gali būti skaičiuojami, bet tai nulemia, kiek skirtingų kategorijų yra duomenyse.

Kitas būdas – apibendrinti duomenis grafikais (4 pav.). Iš tiesų yra svarbu gebėti parinkti tinkamą diagramą jos neperkraunant, bet tuo pačiu atskleidžiant reikšmingas tendencijas duomenų imtyje.



4 pav. *Scheminiai diagramų pavyzdžiai*

Šioje temoje demonstracijai naudosime tęstinį pavyzdį – tai duomenų imtis apie krepšininkus (5 pav.).

ID	POZICIJA	AMZIUS	UGIS	SVORIS	RUNGTYNES	START.5	T3	T3.BANDYMAI	BAUDOS	REMEJAS
K1	gynejas	25	194	105	31	2	41	127	53	Ne
K2	puolejas	28	222	108	10	0	2	15	24	Ne
K3	gynejas	22	201	102	34	1	25	74	45	Ne
K4	centras	25	187	89	80	80	0	2	204	Taip
K5	centras	21	203	105	82	28	3	15	203	Taip
K6	puolejas	21	164	90	19	3	6	23	13	Ne
K7	gynejas	25	193	97	7	0	0	4	4	Ne
K8	centras	33	180	83	81	81	10	42	179	Taip
K9	gynejas	21	186	96	10	1	3	12	7	Ne
K10	gynejas	23	204	98	38	2	32	99	47	Ne

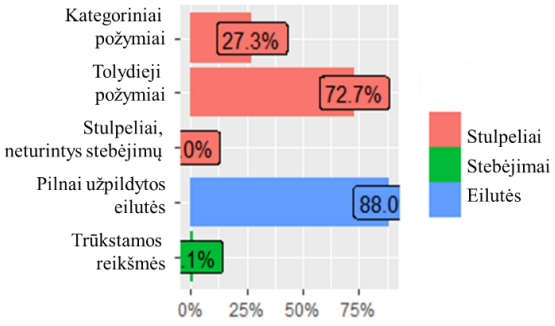
ID	Krepšininko kodas
POZICIJA	Krepšininko standartinė pozicija
AMZIUS	Krepšininko amžius, metais
UGIS	Krepšininko ūgis, cm
SVORIS	Krepšininko svoris, kg
RUNGTYNES	Sužaistų rungtynių skaičius
START.5	Sužaistų rungtynių skaičius startiniame penketuke
T3	Pataikytų tritaškių skaičius
T3.BANDYMAI	Iš viso mestų tritaškių skaičius
BAUDOS	Surinktos asmeninės baudos
REMEJAS	Ar krepšininkas remia tam tikras veiklas

5 pav. *Informacija apie krepšininkus, sukaupta per tam tikrą periodą*

Duomenų imtyje (matricoje) yra 11 stulpelių, kurių kiekvienas saugo tam tikrą informaciją apie krepšininką. Tad iš viso turime 11 požymių. Sąvokas „kintamieji“ ir „požymiai“ laikykime sinonimais, o pats jų vartojimas labiau priklauso nuo to, kokios srities yra tyrėjais. Stebėjimai (arba įrašai) apie konkretų krepšininką yra fiksuojami eilutėse. Iš turimų požymių krepšininko ID nėra svarbus analizei, prieš tai įsitikinus jog eilutėse jie yra unikalūs. Visų požymių prasmė pateikta 5 pav.

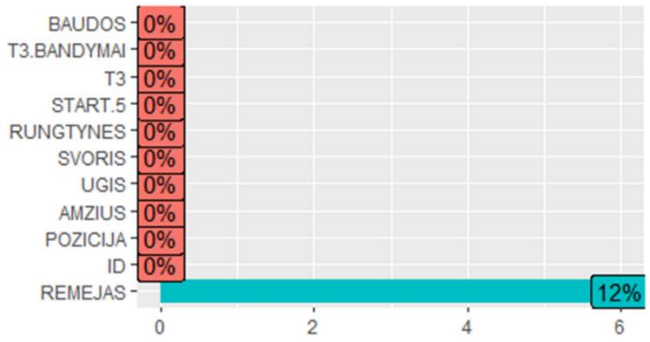
Pradėkime nuo visos imties apibendrinimo (6 pav.).

Rodiklis	Reikšmė
Eilučių skaičius	50
Stulpelių skaičius	11
Kategoriniai požymiai	3
Tolydieji požymiai	8
Stulpelių skaičius su trūkstamomis reikšmėmis	0
Trūkstamų stebėjimų skaičius	6
Pilnų eilučių skaičius	44
Stebėjimų skaičius	550



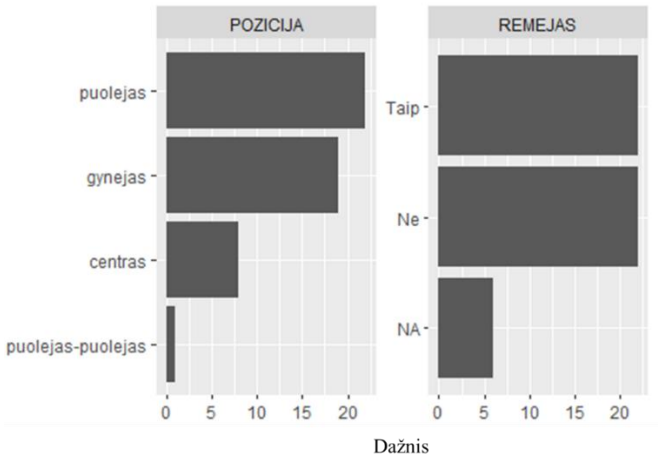
6 pav. *Duomenų imties apibendrinimas*

Matome, jog turime 50 įrašų, 3 požymiai yra kategoriniai ir 8 požymiai yra tolydieji, nėra visiškai tuščių stulpelių, tačiau yra 6 trūkstamos reikšmės, kurios sudaro 1 % visų stebėjimų. Atvaizdavę trūkstamas reikšmes kiekvienam stulpeliui atskirai, įsitikiname, jog jų trūksta būtent požymiui „REMEJAS“ (7 pav.).



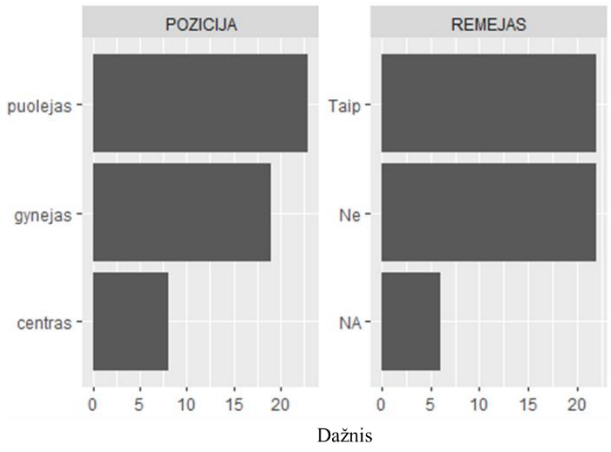
7 pav. *Trūkstamos reikšmės*

Kaip jau minėjome, ID nėra svarbus analizei, todėl koncentruokimės į likusius du kategorinius požymius (8 pav.)



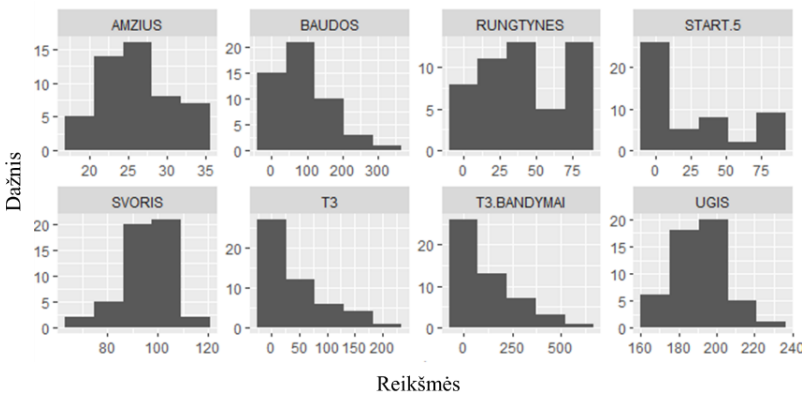
8 pav. *Kategorinių požymių stulpelinė diagrama*

Požymis POZICIJA turi keturias skirtingas kategorijas – „puolejas“, „gynejas“, „centras“, „puolejas-puolejas“. Darome prielaidą, jog įrašas „puolejas-puolejas“ yra klaidingas ir keičiame įrašu „puolejas“ (9 pav.). Požymis REMEJAS turi dvi prasmingas kategorijas „Taip“ ir „Ne“. Trūkstamos reikšmės pažymėtos sutartiniu kodu NA (angl. *not available, not aplicalble*) ir diagramoje atvaizduotos kaip atskira kategorija.



9 pav. Atnaujintų kategorinių požymių stulpelinė diagrama

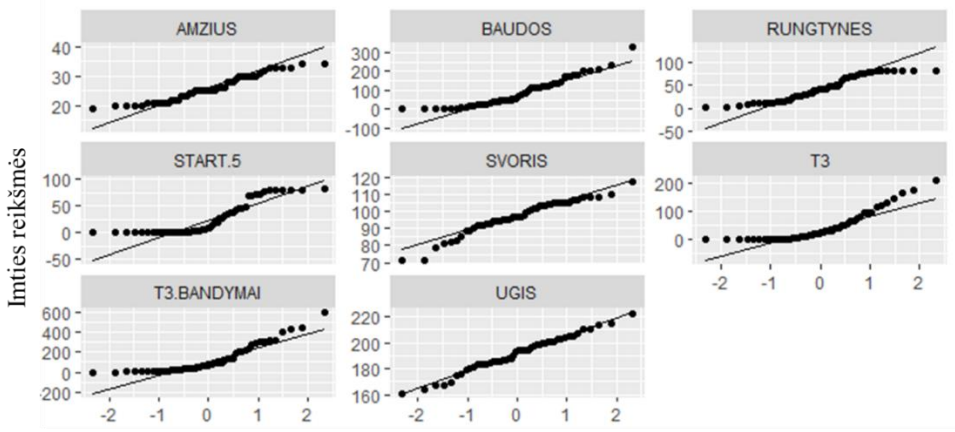
Imties tolydieji požymiai grafiškai yra apibendrinami intervalinių dažnių ar normalizuotų tikimybių histograma dėl to, kad beveik visos reikšmės yra skirtingos ir jos atvaizduojamos grupuojant intervalais (10 pav.).



10 pav. *Intervalinių dažnių histograma*

Kintamieji BAUDOS, START.5, T3 ir T3.BANDYMAI yra gan asimetriškai pasiskirstę, o kintamieji AMZIUS, SVORIS ir UGIS daugiau mažiau yra simetriški ir primena normalųjį (Gauso) pasiskirstymą.

Būtų idealu turėti visus požymius, pasiskirsčiusius pagal normalųjį dėsnį. Tai patikrinti galima grafiniu būdu braižant Q-Q sklaidos diagramą (11 pav.), o tikimybiškai – pagal Shapiro-Wilko normalumo testą (3 lentelė.).



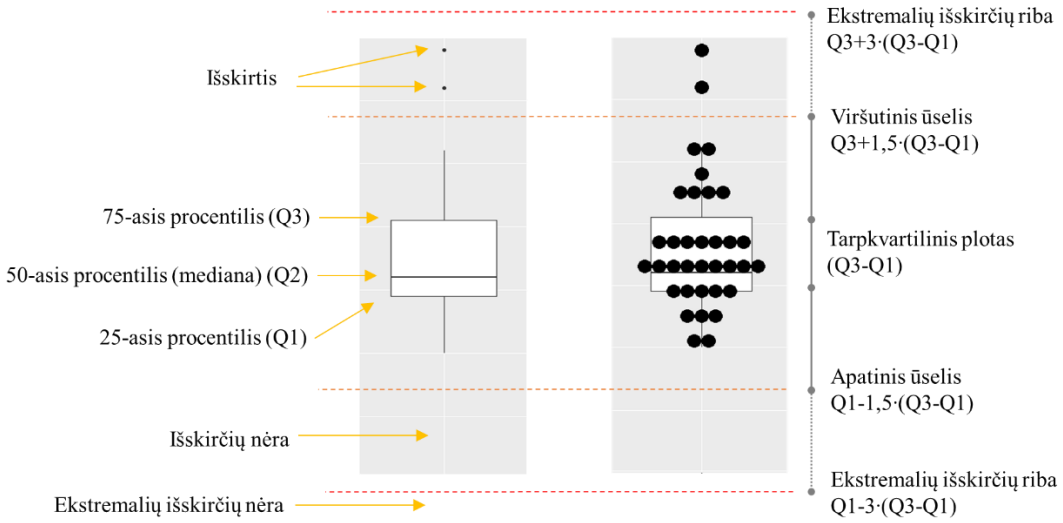
11 pav. *Q-Q sklaidos diagramos*

3 lentelė. Shapiro-Wilko testo rezultatai

AMZIUS	p-value = 0.00979
UGIS	p-value = 0.8431
SVORIS	p-value = 0.2944
RUNGTYNES	p-value = 0.001176
START.5	p-value = 3.851e-07
T3	p-value = 1.262e-06
T3.BANDYMAI	p-value = 3.416e-06
BAUDOS	p-value = 0.002043

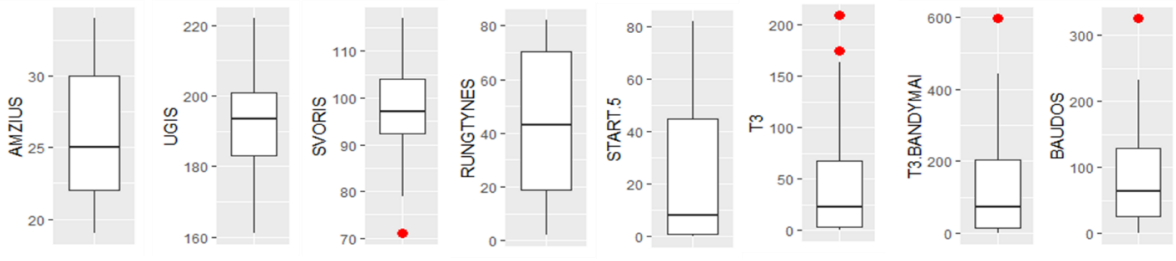
Q-Q sklaidos diagramoje normalumas stebimas tuomet, kai stebėjimai (atvaizduoti taškais) išsidėsto beveik tiesėje. Pagal gautas testo tikimybes (3 lentelė.) matome, jog normalumo prielaidos požymiams UGIS ir SVORIS atmesti negalime, nes gauta tikimybė viršija 0,05 (reikšmingumo lygmuo dažniausiai taikomas pagal nutylėjimą).

Stačiakampė diagrama (angl. *boxplot*) atvaizduoja tokias skaitines charakteristikas, kaip pirmasis, antrasis ir trečiasis kvartiliai. Jų pagrindu apskaičiuojamas tarpkvartilinis plotis ir viršutinis bei apatinis ūseliai. Stebėjimai esantys ūselių išorėje yra laikomi netipiniais stebėjimais ar išskirtimis, o ypač nutolę stebėjimai – ekstremaliomis išskirtimis. Visa tai apskaičiuojama pagal paveiksle pateiktas formules (12 pav.).



12 pav. *Stačiakampės diagramos interpretacija*

Demonstracinės duomenų imties atveju požymiai SVORIS, T3, T3.BANDYMAI ir BAUDOS turi išskirtis (13 pav.). Atsižvelgiant į situaciją šie įrašai (eilutės) gali būti šalinami, ekstremalios reikšmės keičiamos atitinkamai gretimo ūselio reikšme arba pašalinus pačias stebėjimų reikšmes taikomi duomenų atstatymo (angl. *imputation*) metodai. Be to, pagal stačiakampę diagramą galima vaizdžiai vertinti kintamųjų sklaidą, apskaičiuojamą kaip tarpkvartilinis plotis, bei asimetriją.



13 pav. *Stačiakampės diagramos*

Tiriamosios duomenų analizės metu būtina atlikti klaidingų įrašų patikrą. Tokie įrašai nebūtinai turi būti pavienės (dažniausiai įvedimo) klaidos, kurios nepatenka į požymio apibrėžimo sritį. Jei tik įmanoma, kontrolę reiktų atlikti ir tarp požymių. Pavyzdžiui, startiniame penketuke sužaistų rungtynių skaičius negali viršyti visų sužaistų rungtynių skaičiaus (14 pav.).

ID	POZICIJ	AMZIUS	UGIS	SVORIS	RUNGT	5-STAR		3T	3T_BAN	BAUDO	REMEJA
K15	puolejas	25	186	92	43	45	KLAIDA	9	34	112	Taip
K41	puolejas	26	194	102	28	38	KLAIDA	53	130	43	Ne

43
28

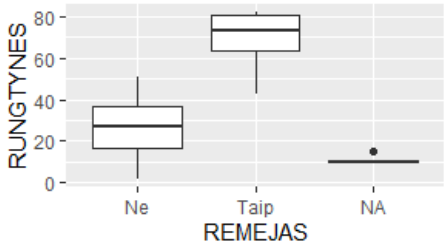
14 pav. *Klaidingi įrašai duomenų imtyje*

Demonstraciniame pavyzdyje, kaip jau esame nustatę, požymis REMEJAS turi trūkstamas reikšmes (15 pav.).

ID	POZICIJ	AMZIUS	UGIS	SVORIS	RUNGT	START.5	T3	T3.BAN	BAUDO	REMEJAS
K17	puolejas	30	161	71	15	8	7	34	23	NA
K18	puolejas	30	185	95	10	0	2	6	2	NA
K2	puolejas	28	222	108	10	0	2	15	24	NA
K22	puolejas	34	183	103	10	2	21	64	32	NA
K33	gynejas	25	186	103	11	0	1	9	16	NA
K9	gynejas	21	186	96	10	1	3	12	7	NA

15 pav. Trūkstamų reikšmių identifikavimas

Įrašai su trūkstamomis reikšmėmis gali būti šalinami arba ignoruojami analizės metu tol, kol neaktualus pats požymis. Arba įrašai su trūkstamomis reikšmėmis gali būti atstatomi pagal duomenų imtyje esančių įrašų panašumą. Šiuo atveju įrašai nėra šalinami – trūkstamas reikšmes pakeisime reikšme „taip“ arba „ne“. Galima būtų pastebėti, jog jei sužaista rungtynių mažiau nei 40, tai REMEJAS įgyja reikšmę „Ne“ (16 pav.).



16 pav. *Požymio REMEJAS stačiakampė diagrama*

Todėl šiuo atveju buvo priimtas sprendimas jas pakeisti reikšme „Ne“ (17 pav.).

ID	POZICI	AMZIUS	UGIS	SVORIS	RUNGT	START.5	T3	T3.BAN	BAUDO	REMEJAS	
K1	gynejas	25	194	105	31	2	41	127	53	Ne	
K10	gynejas	23	204	98	38	2	32	99	47	Ne	
K12	gynejas	26	183	97	19	1	17	36	46	Ne	
K16	puolejas	30	214	107	25	8	9	40	25	Ne	
K17	puolejas	30	161	71	15	8	7	34	23	NA	Ne
K18	puolejas	30	185	95	10	0	2	6	2	NA	Ne
K19	centras	20	194	95	3	0	0	0	0	Ne	
K2	puolejas	28	222	108	10	0	2	15	24	NA	Ne
K21	puolejas	21	198	94	2	0	0	0	0	Ne	
K22	puolejas	34	183	103	10	2	21	64	32	NA	Ne
K26	puolejas	33	194	104	26	26	50	139	44	Ne	
K3	gynejas	22	201	102	34	1	25	74	45	Ne	
K32	gynejas	25	194	107	15	0	1	13	18	Ne	
K33	gynejas	25	186	103	11	0	1	9	16	NA	Ne
K34	gynejas	25	184	103	4	0	0	4	2	Ne	
K35	gynejas	22	215	105	16	0	2	9	11	Ne	
K38	gynejas	34	198	101	38	0	38	128	50	Ne	
K41	puolejas	26	194	102	28	38	53	130	43	Ne	
K43	puolejas	23	188	97	30	3	13	52	29	Ne	
K45	gynejas	30	181	99	34	6	29	98	56	Ne	
K50	puolejas	30	205	108	26	2	3	17	42	Ne	
K6	puolejas	21	164	90	19	3	6	23	13	Ne	
K7	gynejas	25	193	97	7	0	0	4	4	Ne	
K9	gynejas	21	186	96	10	1	3	12	7	NA	Ne

17 pav. *Trūkstamų reikšmių pakeitimas reikšme „Ne“*

Tolydžiųjų požymių charakteristikos, apskaičiuotos krepšininkų duomenų imčiai, pateiktos 4 lentelėje. Atsižvelgiant į situaciją skaičiuojamų charakteristikų sąrašas gali būti koreguojamas.

4 lentelė. *Apskaičiuotos tolydžiųjų požymių skaitinės charakteristikos*

	Vidurkis	Standartas	Mediana	Min	Max	Asimetriškumas
AMZIUS	25,86	4,37	25,00	19	34	0,31
UGIS	191,40	13,64	193,50	161	222	-0,16
SVORIS	97,02	9,05	97,00	71	117	-0,50
RUNGTYNES	43,28	27,00	43,00	2	82	0,11
START.5	26,24	30,52	8,00	0	82	0,78
T3	44,10	52,87	22,50	0	209	1,34
T3.BANDYMAI	127,66	141,92	72,00	0	596	1,30
BAUDOS	85,12	68,51	64,00	0	232	0,54

Norėdami požymius palyginti tarpusavyje arba „paslėpti“ tikrąsias reikšmes galime jas normalizuoti į tam tikrą intervalą $[a; b]$ pagal formulę

$$Z = \frac{x_i - \min(X)}{\max(X) - \min(X)}(b - a) + a\,.$$

Lentelėje keturiems požymiams atliktas normalizavimas į intervalą $[0; 1]$ (5 lentelė.). Šiuo atveju matome, jog didžiausia sklaida būdinga požymiui START.5. Ne mažiau nei 50 proc. požymių START.5, T3 ir T3.BANDYMAI stebėjimų įgyja pakankamai mažas reikšmės , t. y. mažesnes nei apskaičiuota mediana.

5 lentelė. *Apskaičiuotos normalizuotų tolydžiųjų požymių skaitinės charakteristikos*

	Vidurkis	Standartas	Mediana	Min	Max
RUNGTYNES	0,52	0,34	0,51	0	1
START.5	0,32	0,37	0,10	0	1
T3	0,21	0,25	0,11	0	1
T3.BANDYMAI	0,21	0,24	0,12	0	1

Duomenų standartizavimas atliekamas pagal kitą formulę:

$$Z = \frac{x_i - \text{vidurkis}(X)}{\text{standartas}(X)}.$$

Ją pritaikius reikšmės centruojasi apie nulį, o jų standartinis nuokrypis lygus vienam (6 lent.).

6 lentelė. *Apskaičiuotos standartizuotų tolydžiųjų požymių skaitinės charakteristikos*

	Vidurkis	Standartas	Mediana	Min	Max
RUNGTYNES	0	1	-0,01	-1,53	1,43
START.5	0	1	-0,60	-0,86	1,83
T3	0	1	-0,41	-0,83	3,12
T3.BANDYMAI	0	1	-0,39	-0,90	3,30

Kategoriniams požymiams skaičiuojamų charakteristikų sąrašas kur kas trumpesnis. Demonstracinio pavyzdžio atveju požymiams POZICIJA ir REMEJAS apskaičiuotas kardinalumas, moda ir antroji moda (7 lentelė.).

7 lentelė. *Apskaičiuotos kategorinių požymių skaitinės charakteristikos*

	Kardinalumas	Moda	Moda,%	2-oji moda	2-oji moda, %
POZICIJA	3	puolejas	46	gynejas	38
REMEJAS	2	Ne	56	Taip	44

Kardinalumas rodo, kiek skirtingų reikšmių (kategorijų) įgyja kintamasis ir ar tai suderinta su šio požymio apibrėžimo sritimi. Pagal gautą modos reikšmę matome, kiek dažnai stebėjimo reikšmė kartojasi. Antrosios modos reikšmė leidžia įvertinti pagal dažnį sekančios kategorijos atotrūkį nuo pirmosios, arba kitaip tariant, įvertinti, ar pirmoji kategorija nėra dominuojanti. Be to, verta apsvarstyti, gal kategorijas su mažu dažniu verta sujungti į vieną.

Apibendrinkime šią temos dalį:

- tiriamoji duomenų analizė yra labiau būdas (angl. *approach*) nei duomenų imties apibendrinimo metodas (angl. *method*);
- šioje dalyje buvo pademonstruoti tipiniai tiriamosios duomenų analizės pavyzdžiai visam duomenų rinkiniui ir atskirai kintamiesiems. Analizė atliekama grafiniu būdu arba skaičiuojant tam tikras charakteristikas;
- atskiras dėmesys skirtas išskirtims, trūkstamoms reikšmėms ar klaidingiems įrašams nustatyti.

Kita šios temos dalis skirta požymių (kintamųjų) tarpusavio analizei ir naujų požymių kūrimo pavyzdžiams pademonstruoti.