Tiriamoji duomenų analizė

II dalis

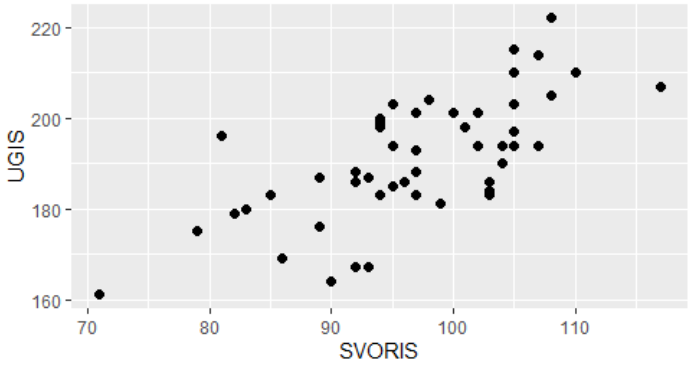
Tęsiame tiriamosios duomenų analizės temą. Šioje dalyje apžvelgsime kintamųjų tarpusavio sąveikos analizę ir pademonstruosime požymių inžinerijos pavyzdžius. Pavyzdžiuose remsimės ta pačia duomenų imtimi apie krepšininkus (1 pav.).

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

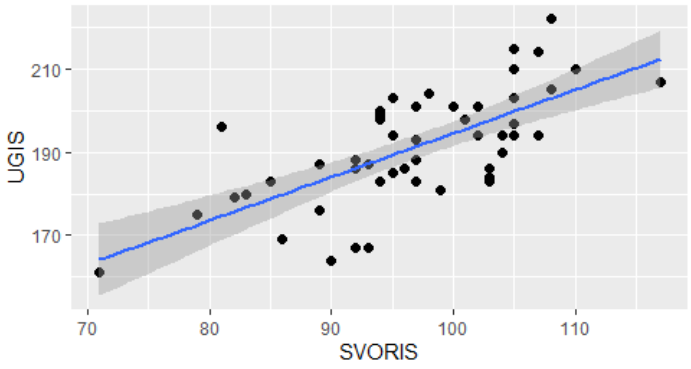
1 pav. *Duomenų imties fragmentas*

Pradėkime nuo tolydžiųjų kintamųjų porų. Viena dažniausiai naudojamų diagramų – sklaidos diagrama, gaunama ašyse atidedant tiriamus kintamuosius. Pavyzdžiui, atvaizduokime ūgio ir svorio sklaidos diagramą (2 pav.).



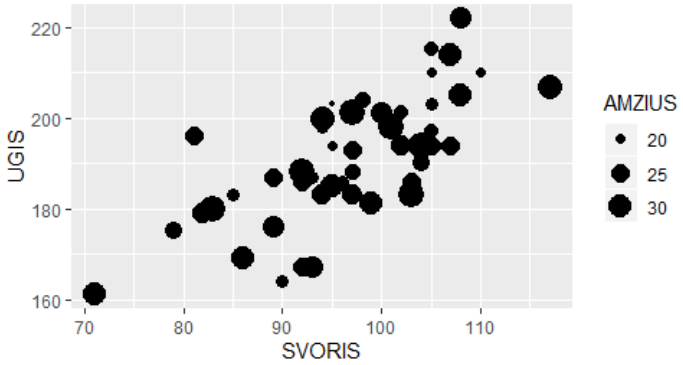
2 pav. *UGIS-SVORIS sklaidos diagrama*

Natūralu tikėtis, jog kuo didesnis ūgis, tuo ir svoris yra didesnis, t. y. kintamuosius sieja tiesinis ryšys (3 pav.).



3 pav. *UGIS-SVORIS tiesinė priklausomybė*

Sklaidos diagramoje galima atvaizduoti ir daugiau informacijos. Pavyzdžiui, taškams, kurie atvaizduoja stebėjimų poras, galima suteikti skirtingą dydį atsižvelgiant į tai, kokio amžiaus yra krepšininkas (4 pav.).



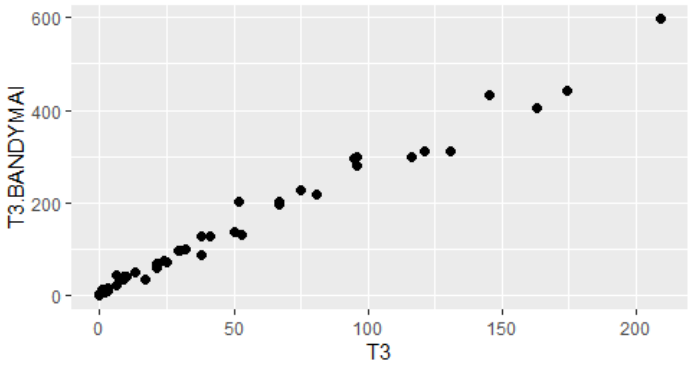
4 pav. *UGIS-SVORIS sklaidos diagrama su požymiu AMZIUS*

Panaudojant požymį REMEJAS, taškams diagramoje galima suteikti ir skirtingas spalvas (5 pav.).

5 pav. *UGIS-SVORIS sklaidos diagrama su požymiais AMZIUS ir REMEJAS*

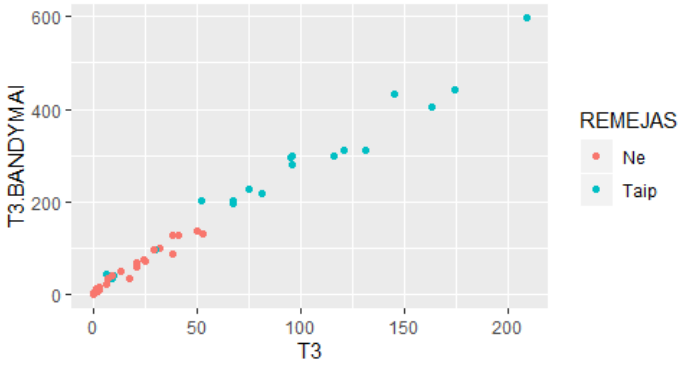
Perdengiant kelis informacijos sluoksnius siekiama atrasti duomenyse tam tikrų dėsningumų (angl. *pattern*). Galima pastebėti, jog amžiaus grupės yra gan atsitiktinai išsibarstę diagramoje, o žemesni krepšininkai ir turintys mažesnį svorį aktyviau dalyvauja kaip rėmėjai.

Panagrinėkime mestų ir pataikytų tritaškių sąryšį atvaizduojant sklaidos diagrama (6 pav.).



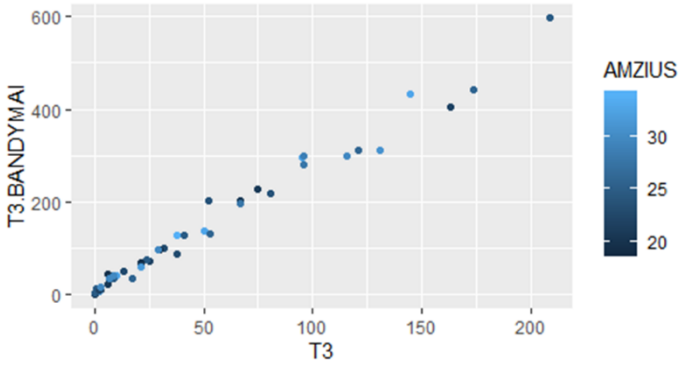
6 pav. *T3-T3.BANDYMAI sklaidos diagrama*

Tarp šių požymių yra stiprus tiesinis ryšys, kurį galima aprašyti tiese $\mathrm{T3} = 0{,}35 \cdot \mathrm{T3.BANDYMAI} + \mathrm{const}$. Iš formulės matyti, kad vidutiniškai reikia trijų bandymų, kad tritaškis būtų taiklus. Šiuo atveju taip pat galima įtraukti papildomos informacijos. Atvaizdavę REMEJAS spalvomis, matome, jog daugiau tritaškių prikaupę krepšininkai aktyviau dalyvauja kaip rėmėjai (7 pav.). Galime daryti prielaidą, jog tokie krepšininkai turi ilgesnę karjerą arba yra tiesiog drąsesni mesdami tritaškius, kas iš dalies gali sąlygoti jų kaip rėmėjų aktyvią veiklą.



7 pav. *T3-T3.BANDYMAI sklaidos diagrama su požymiu REMEJAS*

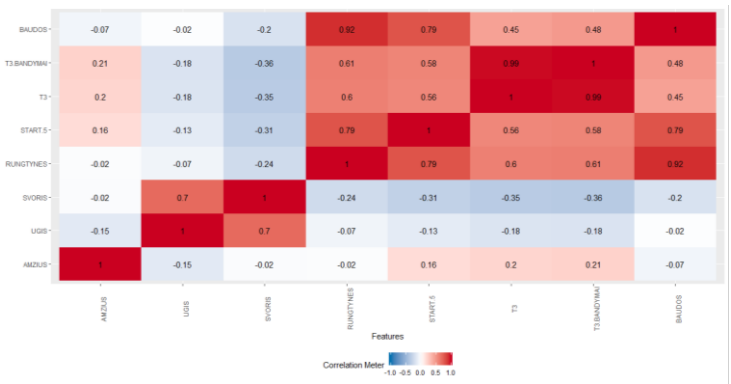
Panaginėkime, kaip amžius gali būti siejamas su tritaškių duomenimis (8 pav.). Šiuo atveju amžius neturi aiškaus dėsningumo.



8 pav. *T3-T3.BANDYMAI sklaidos diagrama su požymiu AMZIUS*

Diagramos leidžia vaizdžiai pamatyti, tačiau daug naudingiau yra kintamųjų tarpusavio priklausomybę įvertinti vienu skaičiumi, t. y. reikalingas priklausomybės matas, pagal kurį galėtume interpretuoti ryšio stiprumą ir kryptį. Populiariausias yra ***tiesinio ryšio stiprumą*** vertinantis Pearsono koreliacijos koeficientas $corr(X,Y)$, kuris apskaičiuojamas kovariaciją dalijant iš standartinių nuokrypių sandaugos. Pearsono koreliacijos koeficiento reikšmių sritis yra intervalas $[-1;1]$. 1 žymi nebeatsitiktinį tiesinį ryšį, o neigiama reikšmė žymi atvirkštinę priklausomybę, t. y. vienam didėjant, kitas mažėja. Norime atkreipti dėmesį į tai, jog 0 žymi tiesinės priklausomybės nebuvimą, tačiau tai gali reikšti, jog tarp ***kintamųjų yra tam tikra netiesinė priklausomybė***. Atskirais atvejais, netiesinė priklausomybė, aprašoma ranginiu (monotoniniu) sąryšiu, yra vertinama neparametriniais Spearmano koreliacijos koeficientu arba Kendallo τ koeficientu.

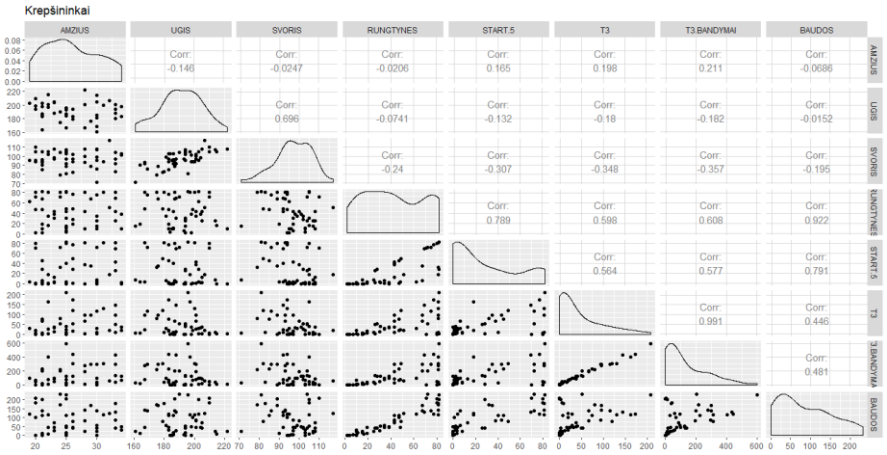
Apskaičiuotos Pearsono koreliacijos koeficiento reikšmės atvaizduotos koreliacijų matricoje (9 pav.).



9 pav. Koreliacijų matrica

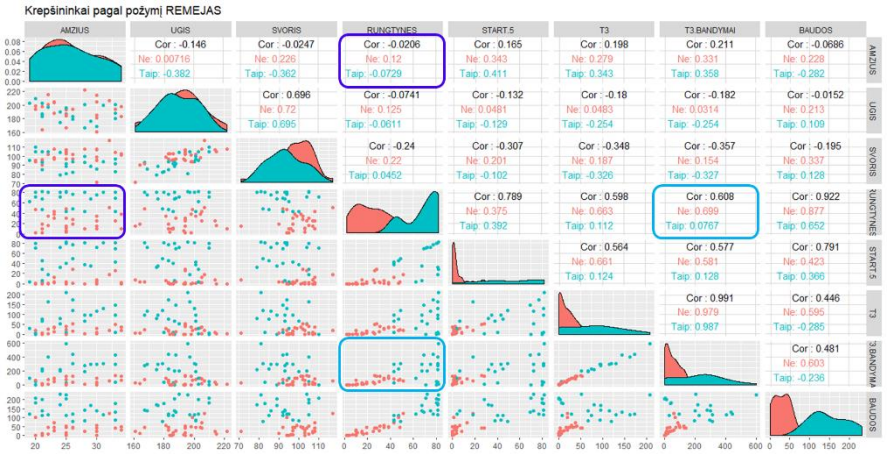
Mestų ir pataikytų tritaškių skaičius sieja ypač stipri tiesinė priklausomybė, įvertinta 0,99. Natūralu, jog tarp rungtynių ir požymių, susijusių su rungtynių statistika (startiniame penkete sužaistų rungtynių skaičius, asmeninės baudos, mestų ir pataikytų tritaškių skaičius), yra gaunama stipresnė nei vidutinė tiesinė priklausomybė. Galime pastebėti, jog krepšininko svorį ir tritaškių duomenis sieja atvirkštinė priklausomybė, t. y. kuo didesnis krepšininko svoris, tuo labiau jis vengia mesti nuo tritaškio linijos. Tačiau ši priklausomybė yra silpnesnė nei vidutinė, ir dėl to nereiktų vertinti jos reikšmingai.

Visas įmanomas kintamųjų poras galime atvaizduoti sklaidos diagramų matricoje (10 pav.). Viršutinis dešinysis trikampis yra atitinkamai užpildomas apskaičiuotu koreliacijos koeficientu tarp šių porų.



10 pav. *Sklaidos diagrama ir koreliacijos koeficientai*

Analizę galime dar labiau pagilinti ir vertinti priklausomybes tarp krepšininkų pagal jų kategorijas. Šiuo atveju yra pasirinkta kategorija REMEJAS ir atvaizduota skirtingomis spalvomis (11 pav.).



11 pav. *Sklaidos diagrama ir koreliacijos koeficientai su požymiu REMEJAS*

Tokia analizė leidžia atskleisti tam tikrus dėsningumus. Pavyzdžiui, tarp sužaistų rungtynių ir mestų tritaškių skaičiaus bendra koreliacija yra 0,6 (pažymėta žydra spalva). Tačiau jei krepšininkas yra rėmėjas, tai koreliacija tampa nereikšminga, t. y. 0,07, o jei krepšininkas nėra rėmėjas, tai koreliacija netgi sustiprinama iki 0,7. Panagrinėkime kontrapavyzdį. Kai apskaičiuota koreliacija apytiksliai artima nuliui, tai atitinkamai sklaidos diagramoje stebėjimų išsibarstymas stebimas atsitiktinai visoje diagramoje, ir nėra jokio aiškaus grupavimosi (pažymėta violetine spalva).

Pastebėsime tai, jog koreliacija gali būti apgaulinga (angl. *spurious*), t. y. apskaičiuoto koeficiento reikšmės arti vieneto, tačiau tai paaiškinama tik dėl sutampančių tendencijų, ir nėra jokio loginio ryšio tarp kintamųjų. Pavyzdžiui, margarino suvartojimo ir skyrybų skaičiaus priklausomybė ≈ 0,99 (https://www.bbc.com/news/magazine-27537142), mirčių nukrentant nuo laiptų ir „iPhone“ pardavimų priklausomybė ≈ 0,9 (https://hbr.org/2015/06/beware-spurious-correlations).

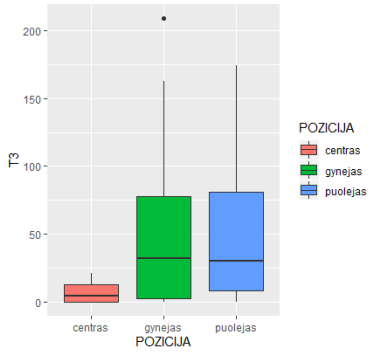
Kategorinių kintamųjų priklausomybei nusakyti labiau vartotinas terminas − ***asociacija*** (angl. *association*). Apie jos buvimą testuojama taikant Pearsono χ^2 testą (angl. *Pearson's Chi-Square Test*):

- ✓ H_0: dydžiai nepriklausomi;
- ✓ H_a: dydžiai priklausomi.

Galioja bendra visoms hipotezėms taikytina taisyklė. Jei testo tikimybė p_{value} > pasirinktas reikšmingumo lygmuo (5 % pagal nutylėjimą), tai priimama H_0; priešingu atveju, priimama H_a.

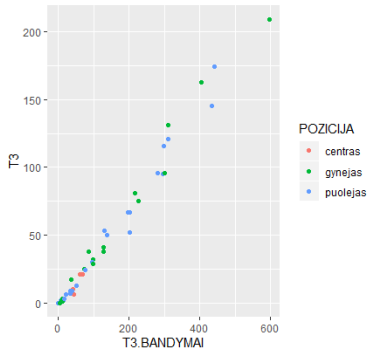
Naudodami demonstracinį pavyzdį patikrinkime priklausomybę tarp kintamųjų POZICIJA ir REMEJAS. Gauta χ^2 testo tikimybė ≈ 0,3, kas yra daugiau 0,05. Darome išvadą, jog tarp kintamųjų POZICIJA ir REMEJAS reikšmingo statistinio ryšio nėra.

Kategorinių ir tolydžiųjų kintamųjų tarpusavio sąryšių vertinimas yra kur kas sudėtingesnis. Todėl šį kartą pademonstruosime tai tik grafiniu būdu. Diagramoje atvaizduota, kaip mestų tritaškių skaičius priklauso nuo krepšininko standartinės pozicijos (12 pav.). Matome, jog gynėjo ir puolėjo tritaškių skaičiai beveik identiški. Tuo tarpu centro pozicijoje žaidžiančio krepšininko mestų tritaškių skaičius yra gerokai mažesnis.



12 pav. *T3 kintamojo stačiakampė diagrama pagal krepšininko poziciją*

Kitoje diagramoje (13 pav.) atvaizduota ta pati informacija, tik parinktas kitas vaizdavimo būdas. Šiuo atveju, vaizdumo ir aiškumo trūksta, todėl tinkamesnė yra kairiau esanti diagrama.



13 pav. *T3-T3.BANDYMAI sklaidos diagrama pagal krepšininko poziciją*

Pereinam prie paskutinio šios temos skyrelio – požymių inžinerijos.

Požymių inžinerija (angl. *feature engineering*) – tai būdas susikurti naujus požymius iš turimos, galimai šiek tiek apdorotos duomenų imties, ar visiškai „žalių“ duomenų, surinktų iš pirminių šaltinių. Pasak Scotto Locklino, „Feature engineering is another topic which doesn’t seem to merit any review

papers or books, or even chapters in books, but it is absolutely vital to ML success. […] Much of the success of machine learning is actually success in engineering features that a learner can understand“. Nors ir pateikta viena citata, tačiau ji atspindi daugelio analitikų nuomonę, jog požymių inžinerijos dėka nesudėtingi algoritmai dažnai aplenkia kur kas sudėtingesnius modelius.

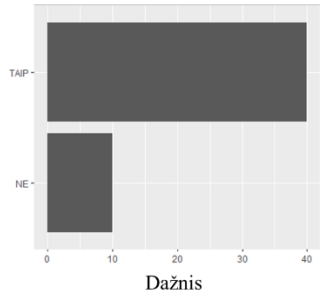
Pradėkime nuo to, kas nėra požymių inžinerija ? Tai:

- pradinių duomenų surinkimas;
- tikslinio kintamojo sukūrimas;
- duomenų kokybės užtikrinimas išvalant besikartojančius įrašus, trūkstamų reikšmių apdorojimas, išskirčių tvarkymas, standartizavimas ar normalizavimas, duomenų transformacijos;
- netgi kintamųjų svarbos parinkimas ar pagrindinių komponenčių (angl. *principal components*) konstravimas neturėtų būti priskiriamas prie požymių inžinerijos.

Kas gi galėtų būti požymių inžinerijos pavyzdžiai ?

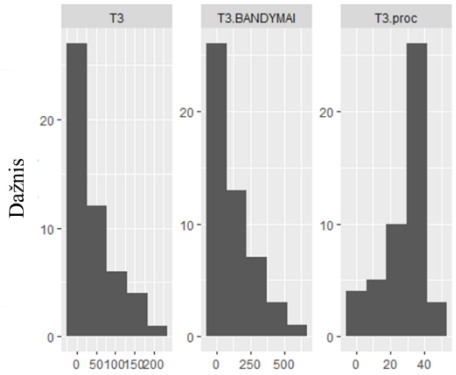
- įvairių indikatorių kūrimas;
- naujų požymių kūrimas iš jau egzistuojančių kintamųjų juos apjungiant ar tiesiog kitaip reprezentuojant jų tarpusavio sąryšį;
- turimos informacijos reprezentavimas kitu formatu;
- išorinių duomenų panaudojimas ir naujų požymių jų pagrindu kūrimas;
- ir dar daugelis kitų, ko neriboja analitiko fantazija.

Panagrinėkime kelis pavyzdžius. Pradėkime nuo indikatoriaus sukūrimo. Sukurtas naujas startiniame penkete žaistų rungtynių skaičiaus indikatorius, rodantis, ar krepšininkais yra žaidęs startinėje sudėtyje (14 pav.).



14 pav. *Startinio penketo START.5 indikatorius*

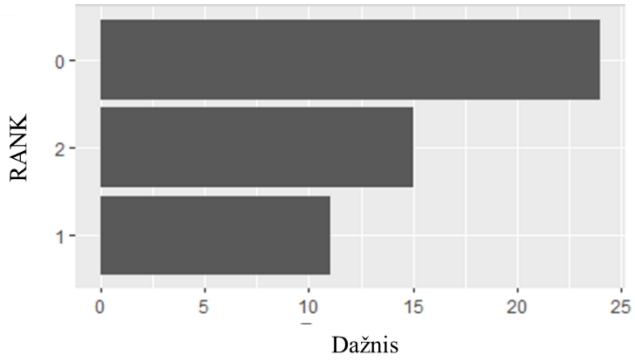
Kitas požymis – pataikytų tritaškių procentas (15 pav.). Galbūt prasmingiau atsisakyti dviejų priklausomų kintamųjų, ir naudoti šį naują požymį.



15 pav. *Pataikytų tritaškių procentas*

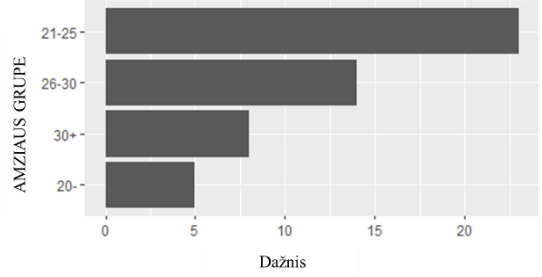
Naujas požymis RANK, sujungia du kintamuosius iš karto, t. y. sužaistų rungtynių skaičių ir pataikytų tritaškių skaičių (16 pav.). Tai yra sukuriamos trys kategorijos pagal pateiktas taisykles:

- 0 – tritaškių pataikymo procentas mažiau nei 30 proc.;
- 1 – sužaista mažiau rungtynių nei bendras vidurkis, o tritaškių pataikymo procentas daugiau nei 30 proc.;
- 2 – sužaista daugiau rungtynių nei bendras vidurkis, o tritaškių pataikymo procentas daugiau nei 30 proc.



16 pav. *Tritaškių pataikymo kategorijos*

Kitas pavyzdys – tolydusis kintamasis paverčiamas kategoriniu, t. y. įvedamas naujas požymis AMZIAUS GRUPE, kuris atspindi atitinkamai krepšininkų amžių pagal pateiktus intervalus: iki 20, (20; 25], (25; 30], virš 30 (17 pav.). Tai galėtume vadinti esamo požymio reprezentavimas kitu formatu.



17 pav. *Amžiaus grupės*

Apibendrinkime šią temos dalį.

- Duomenų imtyje svarbu suprasti ir įvertinti kintamųjų tarpusavio priklausomybę.
- Požymių tarpusavio priklausomybė gali būti vertinama atskirai tarp kategorinių ir atskirai tarp tolydžiųjų dydžių.
- Kur kas sudėtingiau įvertinti sąryšius tarp skirtingų tipų kintamųjų. Tačiau šiuo tikslu gali būti kuriami nauji požymiai, kurie atsižvelgtų į tolydaus ir kategorinio kintamųjų sąveiką.
- Požymių inžinerijos metu siekiama sukurti tokius požymius, kurie leistų pagerinti modelio veikimą atrandant reprezentatyvius požymius, galimai nekeičiant paties modelio algoritmo sudėtingumo (siekiant paprastumo ir greito veikimo).

Pabaigai keletas komentarų, kas dar svarbu dirbant su duomenimis ir formuojant įvestį į modelį. Atskiras svarbus klausimas yra dimensijos mažinimas (angl. *dimension reduction*) tuo atveju, kai turime pakankamai daug ar per daug duomenų. O kitas aktualus klausimas – kaip atsirinkti labiau ar mažiau svarbius kintamuosius (angl. *feature selection, importance*).