Rūta Šalkauskaitė

Sintetinės biologijos programavimo kurso baigiamasis projektas

Jėgos trikovės varžybų rezultatų analizė

Kadangi esu pametusi galvą dėl sporto ir 2019 m. lapkritį tapau Pasaulio svarsčių kilnojimo čempione, ieškojau duomenų susijusių su sunkiąja atletika. Duomenis radau *kaggle* svetainės pagalba. Tai OpenPowerlifting duomenų bazėje publikuojama informacija apie daugelio metų jėgos trikovės sporto varžybų rezultatus. Jėgos trikovė – trijų veiksmų su štanga (pitūpimo, spaudimo ir atkėlimo) sportas, kuriame laimi didžiausią svorį pagal griežtas taisykles įveikę atletai, besivaržydami savo svorio kategorijoje. Federacijoms paskelbus oficialius varžybų rezultatus, kiekvieno sportininko pasiekimai patalpinami šioje duomenų bazėje.

Pasaulio svarsčių kilnojimo čempionatas, 2019 m.

Šališkumas duomenyse gali atsirasti nebent dėl šiek tiek besiskiriančių taisyklių tarp jėgos trikovės federacijų ir skirtingo teisėjų akreditavimo. Duomenų rinkinyje kai kuriose vietose trūksta duomenų, pvz., 4 bandymas atlikti veiksmą egzistuoja tik nedaugelyje federacijų, todėl dažnai šiuose stulpeliuose trūksta reikšmių; kartais varžybos gali būti tik vieno iš trijų veiksmų, todėl kitų veiksmų bandymų verčių nebus ir kt. Nesant informacijos įrašyta „NaN“. Atliekant kintamojo analizę eilutes su „NaN“ pašalinsiu.

Vilniaus universiteto jėgos trikovės komanda

Aprašomoji statistika

Duomenų rinkinyje yra 37 kintamieji ir 1 423 354 stebėjimai, testams stebėjimų tikrai pakaks. Duomenyse yra įvairių tipų kintamųjų: tolydžiųjų (raudona), diskrečiųjų ordinalių (mėlyna) ir diskrečiųjų nominalių (žalia) (1 pav.).

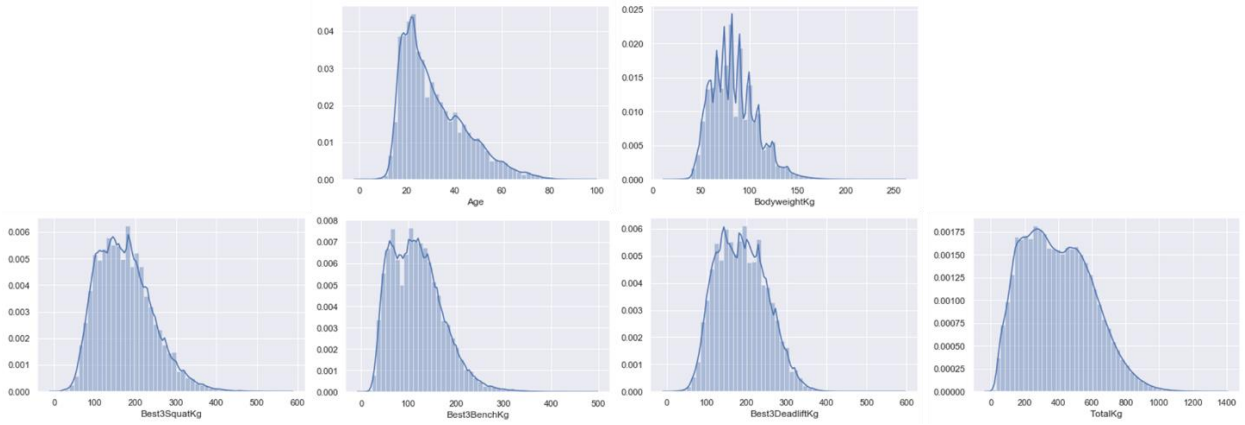
	Name	Sex	Event	Equipment	Age	AgeClass	Division	BodyweightKg	WeightClassKg	Squat1Kg	Squat2Kg	Squat3Kg	Squat4Kg
count	1423354	1423354	1423354	1423354	757527.0	786800	1415176	1406622.0	1410042	337580.0	333349.0	323842.0	3696.0
unique	412574	2	7	5	nan	16	4842	nan	224	nan	nan	nan	nan
top	Alan Aerts	M	SBD	Single-ply	nan	24-34	Open	nan	90	nan	nan	nan	nan
freq	214	1060189	1073237	787141	nan	244197	337927	nan	103156	nan	nan	nan	nan
mean	NaN	NaN	NaN	NaN	31.5	NaN	NaN	84.2	NaN	114.1	92.2	30.1	71.4
std	NaN	NaN	NaN	NaN	13.4	NaN	NaN	23.2	NaN	147.1	173.7	200.4	194.5
min	NaN	NaN	NaN	NaN	0.0	NaN	NaN	15.1	NaN	-555.0	-580.0	-600.5	-550.0
25%	NaN	NaN	NaN	NaN	21.0	NaN	NaN	66.7	NaN	90.0	68.0	-167.5	-107.8
50%	NaN	NaN	NaN	NaN	28.0	NaN	NaN	81.8	NaN	147.5	145.0	110.0	135.0
75%	NaN	NaN	NaN	NaN	40.0	NaN	NaN	99.2	NaN	200.0	205.0	192.5	205.0
max	NaN	NaN	NaN	NaN	97.0	NaN	NaN	258.0	NaN	555.0	567.0	560.0	505.5

	Best3SquatKg	Bench1Kg	Bench2Kg	Bench3Kg	Bench4Kg	Best3BenchKg	Deadlift1Kg	Deadlift2Kg	Deadlift3Kg	Deadlift4Kg	Best3DeadliftKg
count	1031450.0	499779.0	493486.0	478485.0	9505.0	1276181.0	363544.0	356023.0	339947.0	9246.0	1081808.0
unique	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
top	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
freq	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
mean	174.0	83.9	55.1	-18.5	24.8	116.5	162.7	130.2	13.0	78.9	187.3
std	69.2	105.2	130.3	144.2	165.6	54.8	108.7	162.7	215.1	192.6	62.3
min	-477.5	-480.0	-507.5	-575.0	-500.0	-522.5	-461.0	-470.0	-587.5	-461.0	-410.0
25%	122.5	57.5	-52.5	-140.0	-127.5	74.8	125.0	115.0	-210.0	-110.0	138.3
50%	167.8	105.0	95.0	-60.0	77.5	111.1	180.0	177.5	117.5	145.2	185.0
75%	217.5	145.0	145.0	117.5	157.5	150.0	226.8	230.0	205.0	210.0	230.0
max	575.0	467.5	487.5	478.5	487.6	488.5	450.0	460.4	457.5	418.0	585.0

	TotalKg	Place	Wilks	McCulloch	Glossbrenner	IPFPoints	Tested	Country	Federation	Date	MeetCountry	MeetState	MeetName
count	1313184.0	1423354	1304407.0	1304254.0	1304407.0	1273286.0	1093892	388884	1423354	1423354	1423354	941545	1423354
unique	nan	124	nan	nan	nan	nan	1	176	222	5367	96	111	11599
top	nan	1	nan	nan	nan	nan	Yes	USA	THSPA	2017-02-18	USA	TX	World Championships
freq	nan	541908	nan	nan	nan	nan	1093892	91333	290547	7001	856561	448753	32615
mean	395.6	NaN	288.2	296.1	271.8	485.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	201.1	NaN	123.2	125.0	117.6	113.3	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	2.5	NaN	1.5	1.5	1.4	2.2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	232.5	NaN	197.9	204.8	182.8	402.9	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	378.8	NaN	305.2	312.0	285.9	478.1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	540.0	NaN	374.6	383.8	355.3	559.7	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	1367.5	NaN	779.4	804.4	743.0	1245.9	NaN	NaN	NaN	NaN	NaN	NaN	NaN

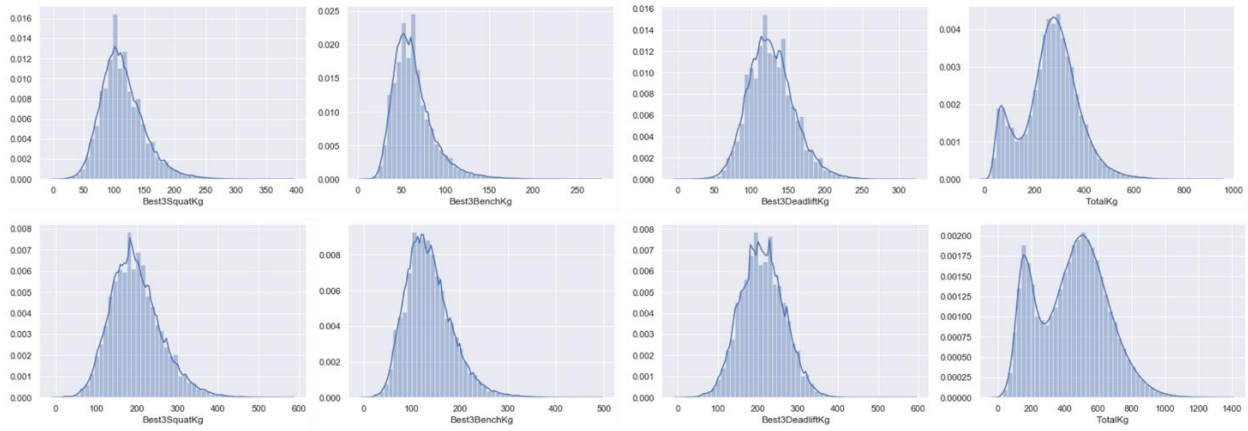
1 pav. Aprašomosios statistikos charakteristikos

Tolydžiųjų kintamųjų (raudona) centrinė statistika – vidurkis – nurodyta 1 pav. „mean” eilutėje. Kintamojo „Age” minimali vertė yra 0, bet tokio amžiaus būti negali, todėl prieš skaičiuojant šio kintamojo vidurkį reikėtų pašalinti nulines vertes. Nulines vertes pakeitus „NaN”, mažiausias amžius – 0,5 metų, kas irgi yra nerealu, todėl vidurkį apskaičiuosiu atmetus 0 – 14 metų vertes; gauta ~31,8 metų.



2 pav. Histogramos pasiskirstymui įvertinti

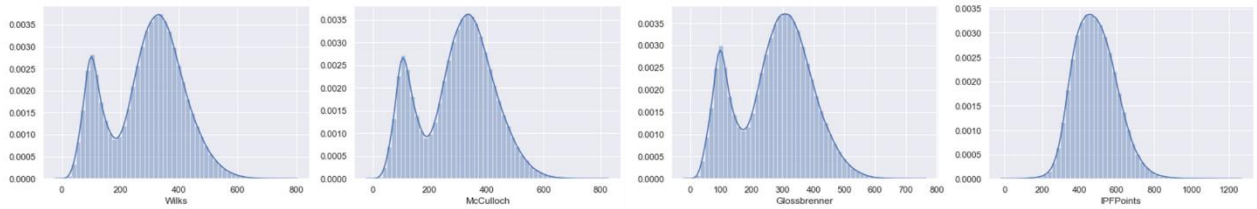
Kiekvieno veiksmo (Squat, Bench ir Deadlift) bandymų vidurkių vertės neatitinka realių, nes esant nesėkmingam bandymui prie kilogramų skaičiaus nurodomas minuso ženklas. Norint apskaičiuoti tikruosius vidurkius ir patikrinti ar duomenys normaliai pasiskirstę, reikia pašalinti neigiamas vertes. Tai atlikus, geriausio pritūpimo (Best3SquatKg) vidurkis – 174,5 kg; geriausio spaudimo (Best3BenchKg) – 116,96 kg; geriausio atkėlimo (Best3DeadliftKg) – 187,5 kg; sumos (TotalKg) – 395,6 kg. Ar duomenys pasiskirstę normaliai tikrinau vizualiai panaudodama histogramas (2 pav.), bet moterų ir vyrų rezultatai varžybose skiriasi, todėl siekiant įvertinti pasiskirstymą reikia atskirti duomenis pagal lytį (3 pav.). Šie duomenys pasiskirstę normaliai, išskyrus bendrą kilogramų sumą, kurios pasiskirstyme matyti du maksimumai, atsiradę dėl varžybų, kuriose atliekamas tik vienas iš veiksmų. Amžiaus ir kūno svorio pasiskirstymai nepanašūs į normalų (2 pav.).



3 pav. Histogramos pasiskirstymui nustatyti pagal lytis (viršuje – moterų, apačioje – vyrų)

Kiti kintamieji – taškai pagal įvairias skaičiuokles gaunami atsižvelgiant į lytis, todėl papildomai skirstyti nebereikia (4 pav.). Wilks, McCulloch ir Glossbrenner taškų pasiskirstymuose yra po du

maksimumus, nes skaičiuojant naudojama tik kilogramų suma ir neatsižvelgiama, jei atliekamas tik vienas veiksmas. IPF taškai pasiskirstę normaliai, nes skaičiuojant juos atsižvelgiama į varžybų pobūdį.



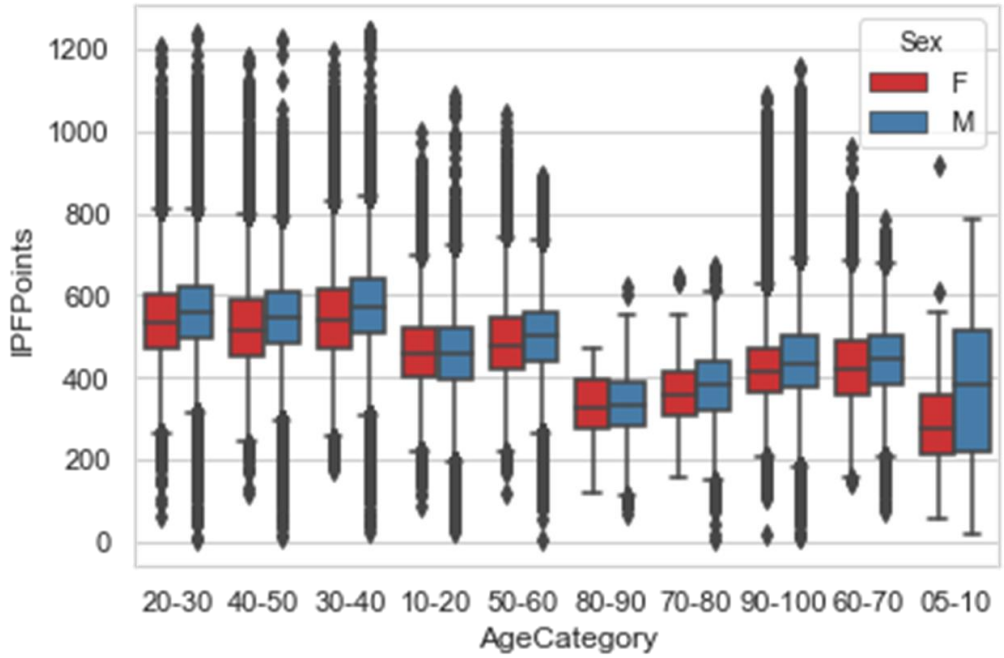
4 pav. Histogramos taškų pagal Wilks, McCulloch, Glossbrenner ir IPF pasiskirstymui nustatyti

Inferencinė statistika

Įdomu sužinoti ar skiriasi jaunesnių ir vyresnių moterų jėgos trikovės rezultatai. Šiai hipotezei patikrinti tiriamieji atskirti pagal lytį, o moterys į dvi grupes pagal amžių – iki 30 ir virš 30 metų. Moterų rezultatą lyginsiu remiantis IPF taškų skaičiumi (IPFPoints), nes šis rodmuo yra populiariausias ir tiksliausias vertinant jėgos trikovės sportininkų pajėgumą. Atlikus t-testą, nustatyta, kad šių amžiaus grupių IPF taškai statistiškai reikšmingai skiriasi (p vertė labai maža). Moterų iki 30 metų IPF taškų vidurkis – 510, o 30 ir vyresnių – 522.

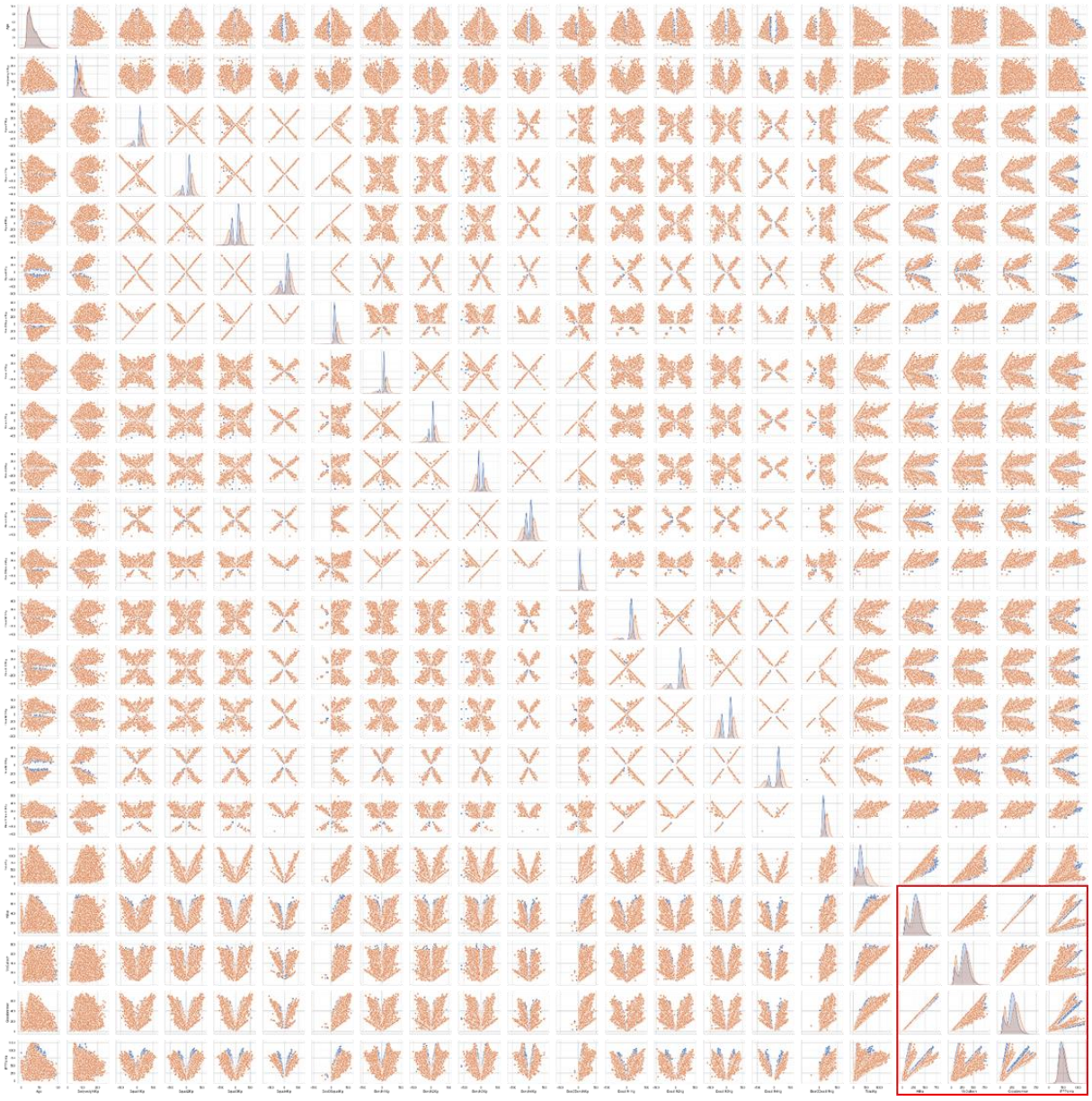
Suskirsčius atletus į amžiaus grupes nubraižytos IPF taškų stačiakampės diagramos atsižvelgiant ir į sportininko lytį (5 pav.).

Ieškant tarpusavyje koreliuojančių kintamųjų nubraižytos (su didelėm kompiuterio pastangom) taškinės diagramos (6 pav.). Nepaisant to, kad analizei trukdo nesėkmingi bandymai su neigiamomis kilogramų vertėmis, daugelyje grafikų matosi koreliacija. Stipriausiai koreliuoja atskiri to paties veiksmo bandymai, skirtingi taškų skaičiavimo metodai (raudonai apibrėžta, 6 pav.), kas yra aivaizdu, nes visi jie apskaičiuojami remiantis tais pačiais rezultatais. Pvz., Wilks ir Glossbrenner taškų koreliacijos koeficientas – 0,995.

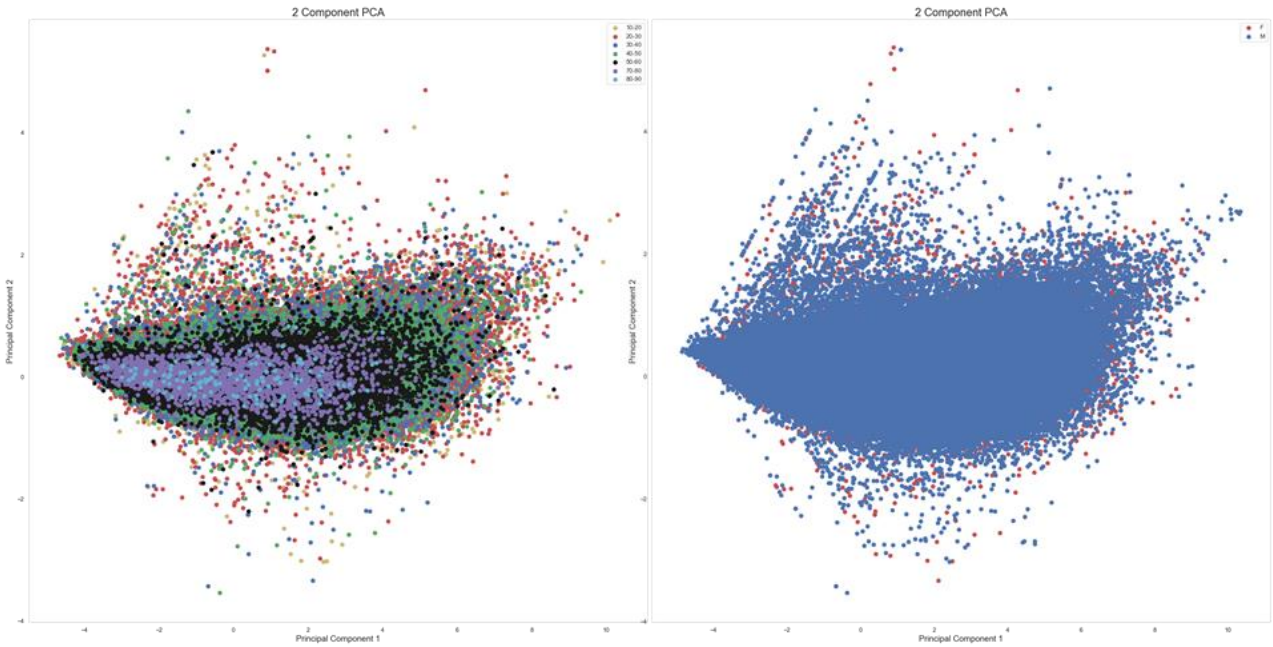


5 pav. IPF taškų pagal amžiaus kategoriją ir lytį stačiakampės diagramos

Rezultatų skirtumų tarp lyčių ir amžiaus grupių dar kartą ieškota pasitelkiant PCA, o ne stačiakampes diagramas. 4 dimensijos („Best3SquatKg“, „Best3BenchKg“, „Best3DeadliftKg“, „TotalKg“) suredukuotos iki dviejų („principal component 1“, „principal component 2“). Jas atvaizduojant suskirsčius į amžiaus grupes arba lytis nepastebėta išsiskiriančių klasterių – visi jie persidengia, arba nėra daug nutolę nuo kitų grupių (7 pav.)



6 pav. korelogramos (kategorinis kintamasis – „Sex”, geltona – „M”, mėlyna – „F“)



7 pav. PCA vizualizacijos. Kairėje – pagal amžiaus grupes, dešinėje – pagal lytį

Išvados

Pasirinkau duomenų rinkinį apie jėgos trikovės rezultatus, norėdama į šį sportą pažvelgti iš statistinės pusės bei skaičiais įrodyti, kad jėgos sportas – kiekvienam, tiek jaunuoliui, tiek seneliui. Atliekant aprašomąją statistinę analizę, kėblumų sukėlė nesėkmingi veiksmų bandymai su minuso ženklu, manau, reikalingas kitoks nesėkmių aprašymo būdas. Dar viena kliūtis – varžybos kuriose neatliekami visi 3 jėgos trikovės veiksmai. Tokiu atveju gaunamos mažos sumos ir taškų pagal Wilks, Glossbrenner ir McCulloch vertės ir sportininko pajėgumą galima vertinti tik pagal IPF taškų skaičių. Trūkstamos vertės duomenų rinkyje nesutrukdė atlikti statistinius testus, nes duomenų kiekis yra labai didelis, ką pajautė ir mano kompiuteris. Žvelgiant į koreliacijos analizės rezultatus pasitvirtina nerašyta taisyklė – kuo stipresnis atletas, tuo geresni rezultatai visuose trijuse veiksmuose. Taip pat stipri ir statistiškai reikšminga koreliacija stebėta tarp skirtingų taškų skaičiavimo formulių. Wilks ir Glossbrenner taškų koreliacijos koeficientas artimas 1, todėl abi sistemos vienodai tinkamos sportininkų rezultatams analizuoti. Ieškant įrodymų, kad jėgos sportas kiekvienam, palyginau moterų iki 30 ir vyresnių rezultatus remdamasi IPF taškų vertėmis. Mano nuostabai, vyresnės moterys netgi turėjo statistiškai reikšmingą pranašumą! Vargu, ar šis pranašumas iš esmės reikšmingas, bet aišku viena – vyresnės moterys nenusileidžia jaunesėms. Skirtumų tarp sportininkų amžiaus grupių ieškojau ir PCA pagalba, bet vizualiai jų pamatyti nepavyko. Žvelgiant į stačiakampes IPF taškų diagramas suskirstytas pagal amžiaus grupes, geriau išryškėja skirtumai tarp grupių, bet jie nėra drastiški.