

Body Performance Analysis

407474, Barbara DONCER, poniedziałek 14⁴⁰
AGH, Wydział Informatyki Elektroniki i Telekomunikacji
Rachunek prawdopodobieństwa i statystyka 2020/2021

Kraków, 27 stycznia 2022

Ja, niżej podpisany(na) własnoręcznym podpisem deklaruję, że przygotowałem(łam) przedstawiony do oceny projekt samodzielnie i żadna jego część nie jest kopią pracy innej osoby.

.....Barbara..Doncer.....

1 Streszczenie raportu

Raport powstał w oparciu o analizę danych dotyczących ilości/jakości ćwiczeń wykonywanych przez osoby o różnej płci, wieku oraz sylwetce.

2 Opis danych

Dane do projektu pochodzą ze strony <http://www.kaggle.pl>. Składają się z 13393 rekordów o 12 cechach.

```
> dim(body_perf)  
[1] 13393     12
```

Każdy rekord reprezentuje jedną osobę, dla której podane są następujące dane:

- age - wiek
- gender - płeć
- height - wzrost podany w cm
- weight - waga podana w kg
- fat - procent zawartości tłuszczy w organizmie
- diastolic - ciśnienie tętnicze rozkurczowe
- systolic - ciśnienie tętnicze skurczowe
- grip force - siła chwytu
- sit and bend forward - głębokość skłonu do prostych nóg podana w cm
- sit-ups - ilość przysiadów zrobionych w jednej serii
- broad jump - długość skoku w dal podana w cm
- class - klasa (A - najlepsza, D - najgorsza)

Przykładowe wiersze wyglądają w ten sposób:

```
> head(body_perf)
```

	age	gender	height	weight	fat	diastolic	systolic	grip.force
1	27	M	172.3	75.24	21.3	80	130	54.9
2	25	M	165.0	55.80	15.7	77	126	36.4
3	31	M	179.6	78.00	20.1	92	152	44.8
4	32	M	174.5	71.10	18.4	76	147	41.4
5	28	M	173.8	67.70	17.1	70	127	43.5
6	36	F	165.4	55.40	22.0	64	119	23.8

	sit.and.bend.forward	sit.ups	broad.jump	class
1	18.4	60	217	C
2	16.3	53	229	A
3	12.0	49	181	C
4	15.2	53	219	B
5	27.1	45	217	B
6	21.0	27	153	B

Funkcja

```
colSums(is.na(body_perf))
```

wzwróciła dla wszystkich wierszy 0, więc dalszą analizę można prowadzić w oparciu o wszystkie 13393 rekordy (wszystkie rekordy zawierają pełne dane).

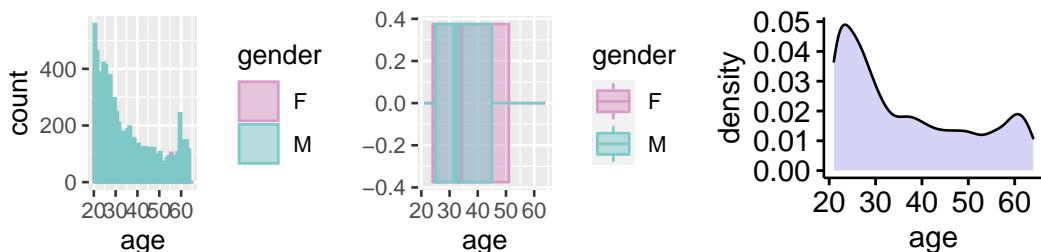
3 Analiza danych

3.1 Wydobywanie podstawowych informacji z danych

3.1.1 Cecha age

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
21.00	25.00	32.00	36.78	48.00	64.00

	srednia	wariancja	odchyl_stand	wsp_skośności	kurtoza	wsp_wyostrzenia
1	36.77511	185.6581	13.62564	0.5998283	1.98226	-1.01774

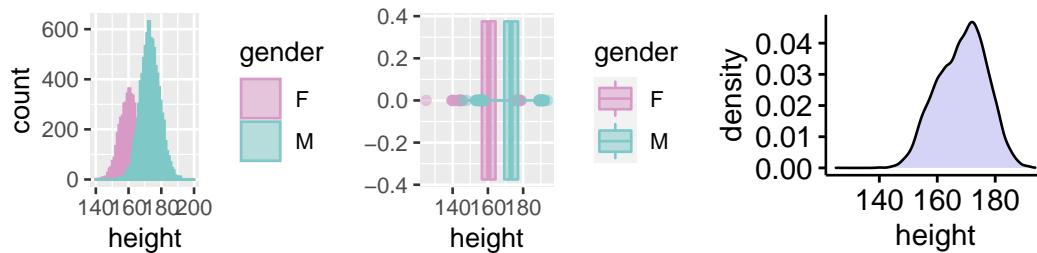


Cecha age pochodzi z rozkładu dodatnio skośnego, platykurtycznego. Gęstość oraz kurtoza sugerują, że nie pochodzi z rozkładu normalnego. Widać, że dane były w większości zbierane od ludzi młodych z przewagą mężczyzn.

3.1.2 Cecha height

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
125.0	162.4	169.2	168.6	174.8	193.8

```
srednia wariancja odchyl_stand wsp_skośności kurtoza wsp_wyostrzenia
1 36.77511 71.00729     8.426583    -0.1868614 2.56666      -0.4333398
```

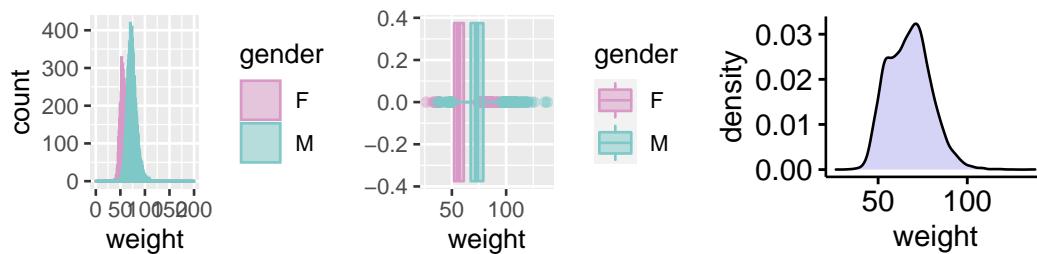


Cecha height pochodzi z rozkładu ujemnie skośnego, platykurtycznego. Gęstość oraz kurtoza sugerują, że może pochodzić z rozkładu normalnego. Widać, że wzrosty mężczyzn są większe od kobiecych.

3.1.3 Cecha weight

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	26.30	58.20	67.40	67.45	75.30	138.10

```
srednia wariancja odchyl_stand wsp_skośności kurtoza wsp_wyostrzenia
1 36.77511 142.7945     11.94967    0.3497654 3.171094      0.171094
```

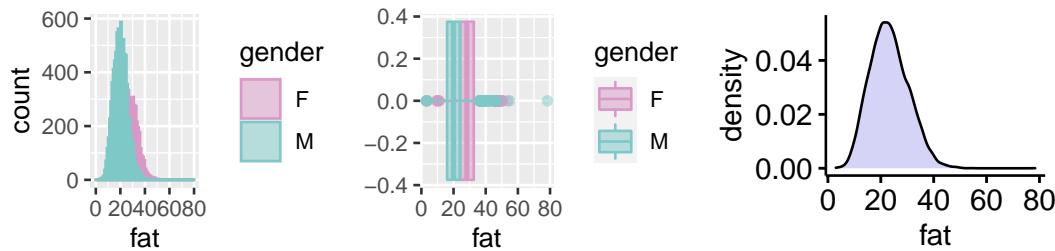


Cecha weight pochodzi z rozkładu dodatnio skośnego, lepokurtycznego. Gęstość oraz kurtoza sugerują, że może pochodzić z rozkładu normalnego. Widać, że wagi mężczyzn są większe od kobiecych.

3.1.4 Cecha fat

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	3.00	18.00	22.80	23.24	28.00	78.40

```
srednia wariancja odchyl_stand wsp_skośności kurtoza wsp_wyostrzenia
1 36.77511 52.66179     7.256844    0.3610918 3.128216      0.1282162
```

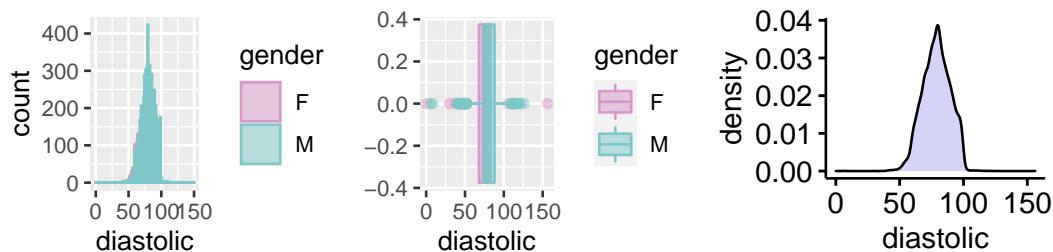


Cecha fat pochodzi z rozkładu dodatnio skośnego, lepokurtycznego. Gęstość oraz kurtoza sugerują, że może pochodzić z rozkładu normalnego. Widać, że procent tkanki tłuszczowej jest większy u kobiet niż u mężczyzn.

3.1.5 Cecha diastolic

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.0	71.0	79.0	78.8	86.0	156.2

srednia wariancja odchyl_stand wsp_skośności kurtoza wsp_wystrzenia
 1 36.77511 115.3913 10.74203 -0.1596193 3.362941 0.3629409

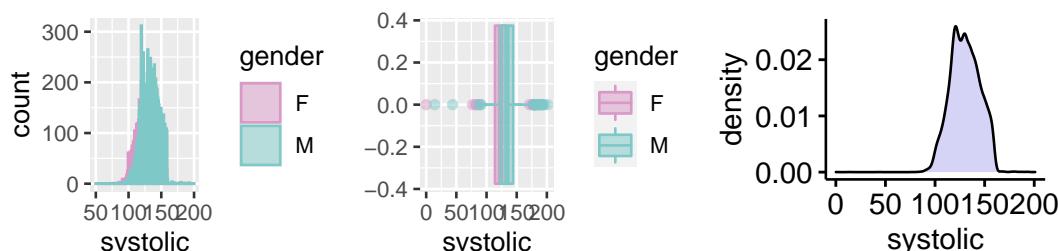


Cecha diastolic pochodzi z rozkładu ujemnie skośnego, lepokurtycznego. Gęstość oraz kurtoza sugerują, że może pochodzić z rozkładu normalnego.

3.1.6 Cecha systolic

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.0	120.0	130.0	130.2	141.0	201.0

srednia wariancja odchyl_stand wsp_skośności kurtoza wsp_wystrzenia
 1 36.77511 216.5004 14.71395 -0.04864816 3.379695 0.3796949

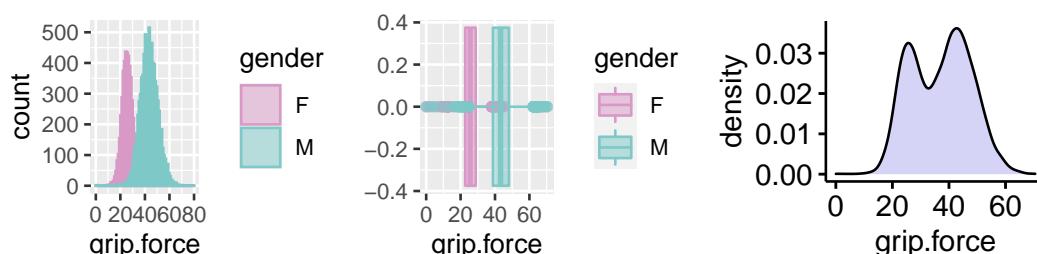


Cecha systolic pochodzi z rozkładu ujemnie skośnego, lepokurtycznego. Gęstość oraz kurtoza sugerują, że może pochodzić z rozkładu normalnego.

3.1.7 Cecha grip force

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.00	27.50	37.90	36.96	45.20	70.50

srednia wariancja odchyl_stand wsp_skośności kurtoza wsp_wystrzenia
 1 36.77511 112.8877 10.62486 0.01845443 2.177659 -0.8223412

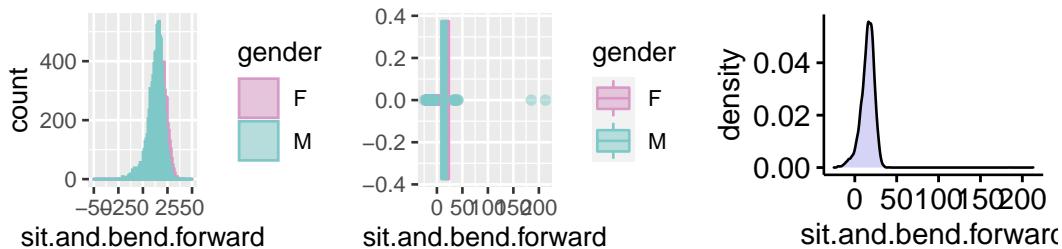


Cecha grip force pochodzi z rozkładu dodatnio skośnego, platykurytycznego. Gęstość oraz kurtoza sugerują, że nie pochodzi z rozkładu normalnego. Widać, że wyniki mężczyzn są większe od kobiecych.

3.1.8 Cecha sit and bend forward

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-25.00	10.90	16.20	15.21	20.70	213.00

srednia wariancja odchyl_stand wsp_skośności kurtoza wsp_wyostrzenia
 1 36.77511 71.51539 8.456677 0.785404 38.20726 35.20726

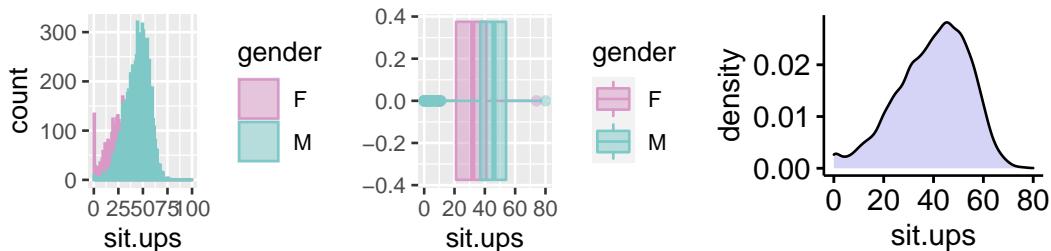


Cecha sit and bend forward pochodzi z rozkładu dodatnio skośnego, lepokurytycznego. Gęstość oraz kurtoza sugerują, że nie pochodzi z rozkładu normalnego.

3.1.9 Cecha sit-ups

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	30.00	41.00	39.77	50.00	80.00

srednia wariancja odchyl_stand wsp_skośności kurtoza wsp_wyostrzenia
 1 36.77511 203.8241 14.2767 -0.4677775 2.843285 -0.1567155

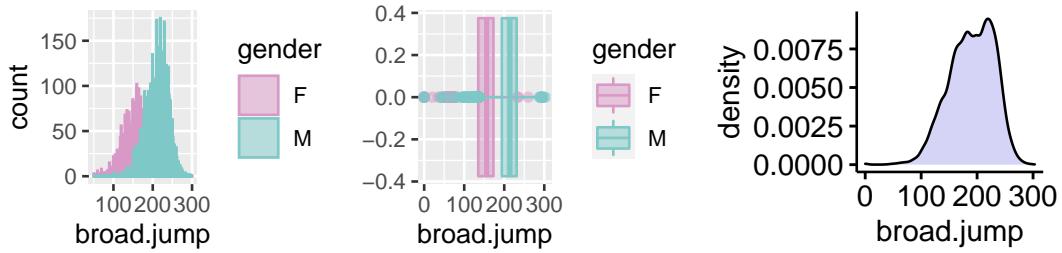


Cecha sit-ups pochodzi z rozkładu ujemnie skośnego, platykurytycznego. Gęstość oraz kurtoza sugerują, że może pochodzić z rozkładu normalnego. Widać, że wyniki mężczyzn są większe od kobieczych.

3.1.10 Cecha broad jump

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	162.0	193.0	190.1	221.0	303.0

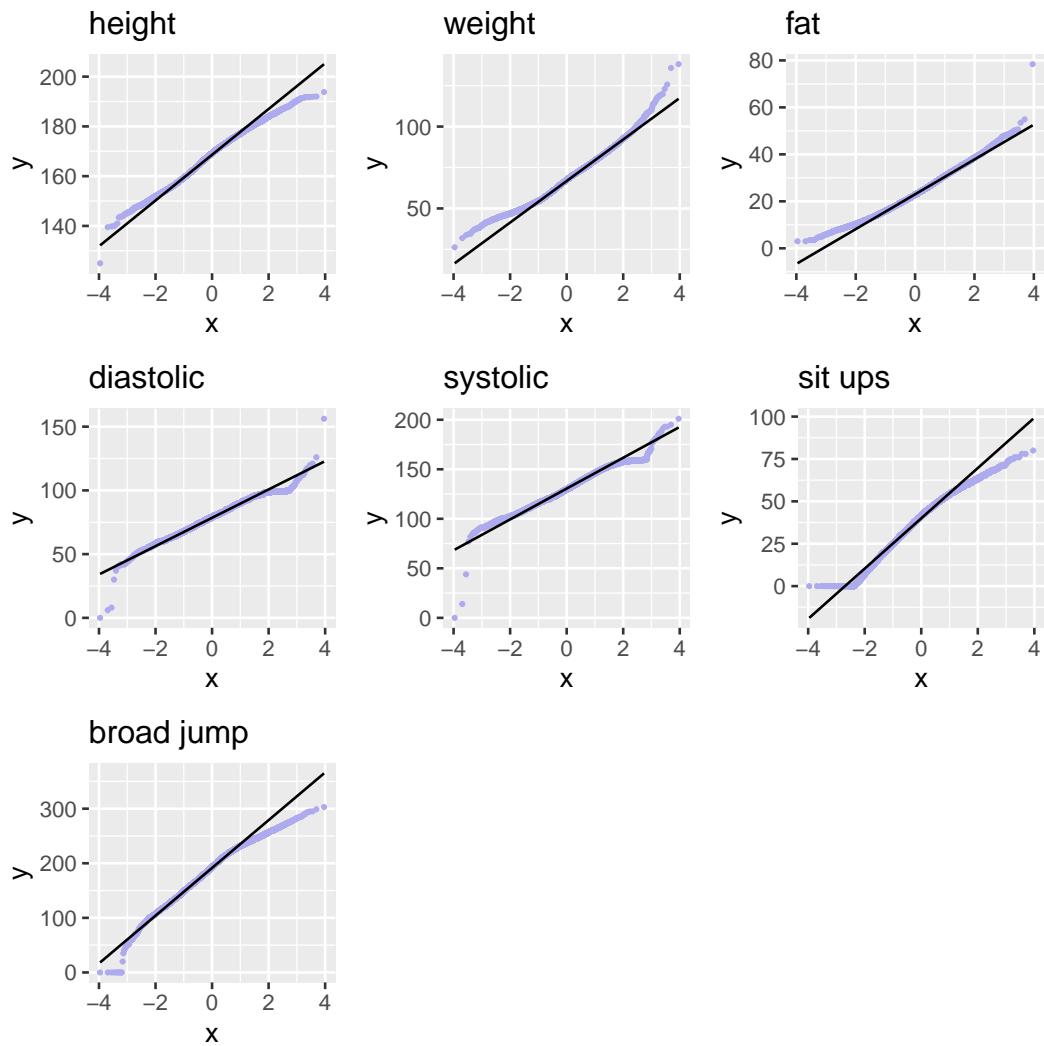
srednia wariancja odchyl_stand wsp_skośności kurtoza wsp_wyostrzenia
 1 36.77511 1589.457 39.868 -0.4225752 3.001948 0.001947645



Cecha broad jump pochodzi z rozkładu ujemnie skośnego, lepokurtycznego. Gęstość oraz kurtoza sugerują, że może pochodzić z rozkładu normalnego. Widać, że wyniki mężczyzn są większe od kobieczych.

3.2 Badanie rozkładu

Kurtoza (bliska 3) oraz wykres gęstości każdej cechy numerycznej (oprócz age, grip force i sit and bend forward) pozwalają przypuszczać, że jej rozkład jest zbliżony do normalnego, dlatego dla każdej takiej cechy został narysowany QQ-plot.



Wszystkie krzywe są zbliżone do prostej, więc wykonano dla wszystkich powyższych cech test Sharpia-Wilk, który w każdym przypadku zwrócił $p > 0.05$, co w połączeniu z wcześniejszą analizą pozwala sądzić, że height, weight, fat, diastolic, systolic, sit-ups i broad jump pochodzą z rozkładów normalnych. Przykładowe wywołanie testu:

```
> shapiro.test(sample(body_perf$weight, 50))
```

```
Shapiro-Wilk normality test
```

```
data: sample(body_perf$weight, 50)
W = 0.97759, p-value = 0.4555
```

3.3 Przedziały ufności

Dla liczbowych cech z rozkładem normalnym wyznaczono przedziały ufności dla średniej oraz wariancji. Przy wyznaczaniu przedziału ufności 0,95 dla średniej skorzystano z rozkładu t-Studenta:

```
> estMean <- function(x){
+   return (c(mean(x) - qt(1-0.05/2, length(x)-1)*sd(x)/sqrt(length(x)),
+             mean(x) + qt(1-0.05/2, length(x)-1)*sd(x)/sqrt(length(x))))
+ }
+ }
```

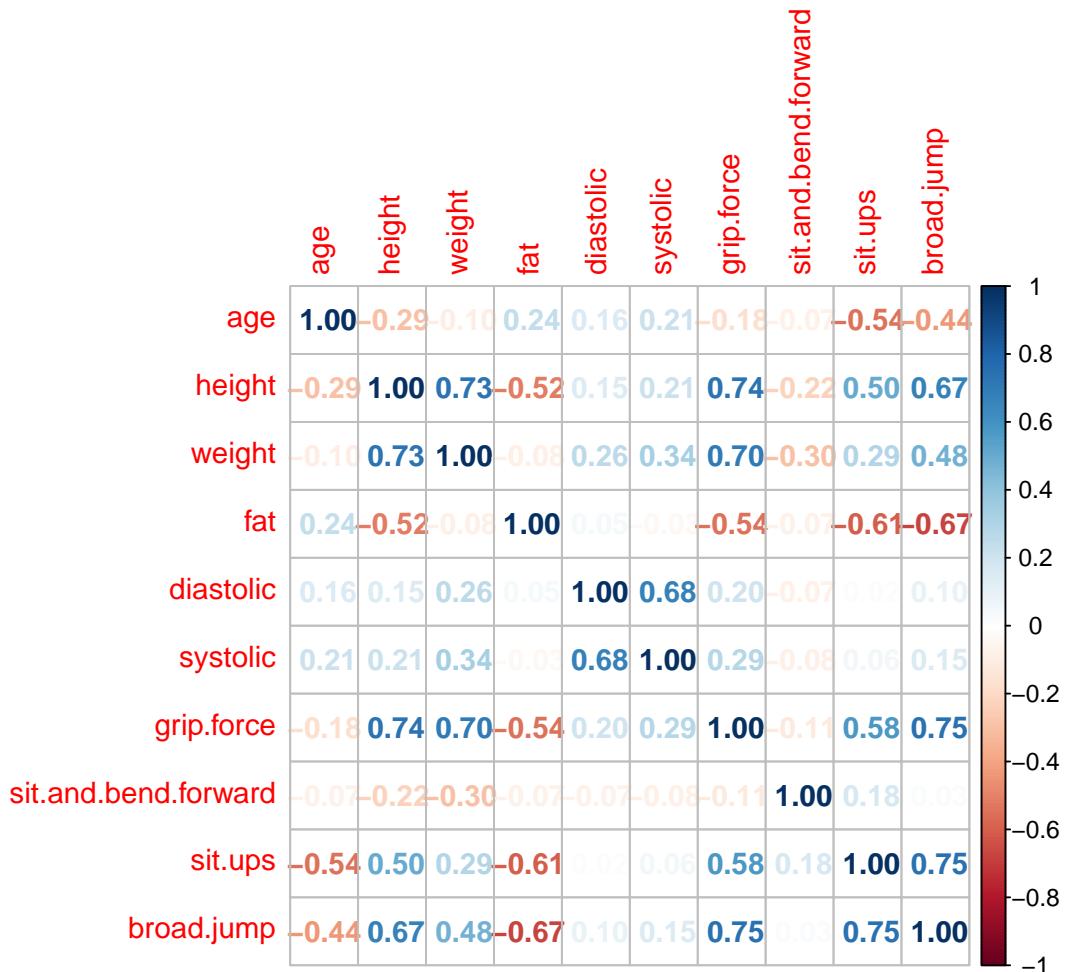
Przy wyznaczaniu przedziału ufności 0,95 dla wariancji skorzystano z rozkładu chi kwadrat:

```
> estVar <- function(x){
+   return (c(var(x)*(length(x)-1)/qchisq(1-0.05/2, length(x)-1),
+             var(x)*(length(x)-1)/qchisq(0.05/2, length(x)-1)))
+ }
+ }
```

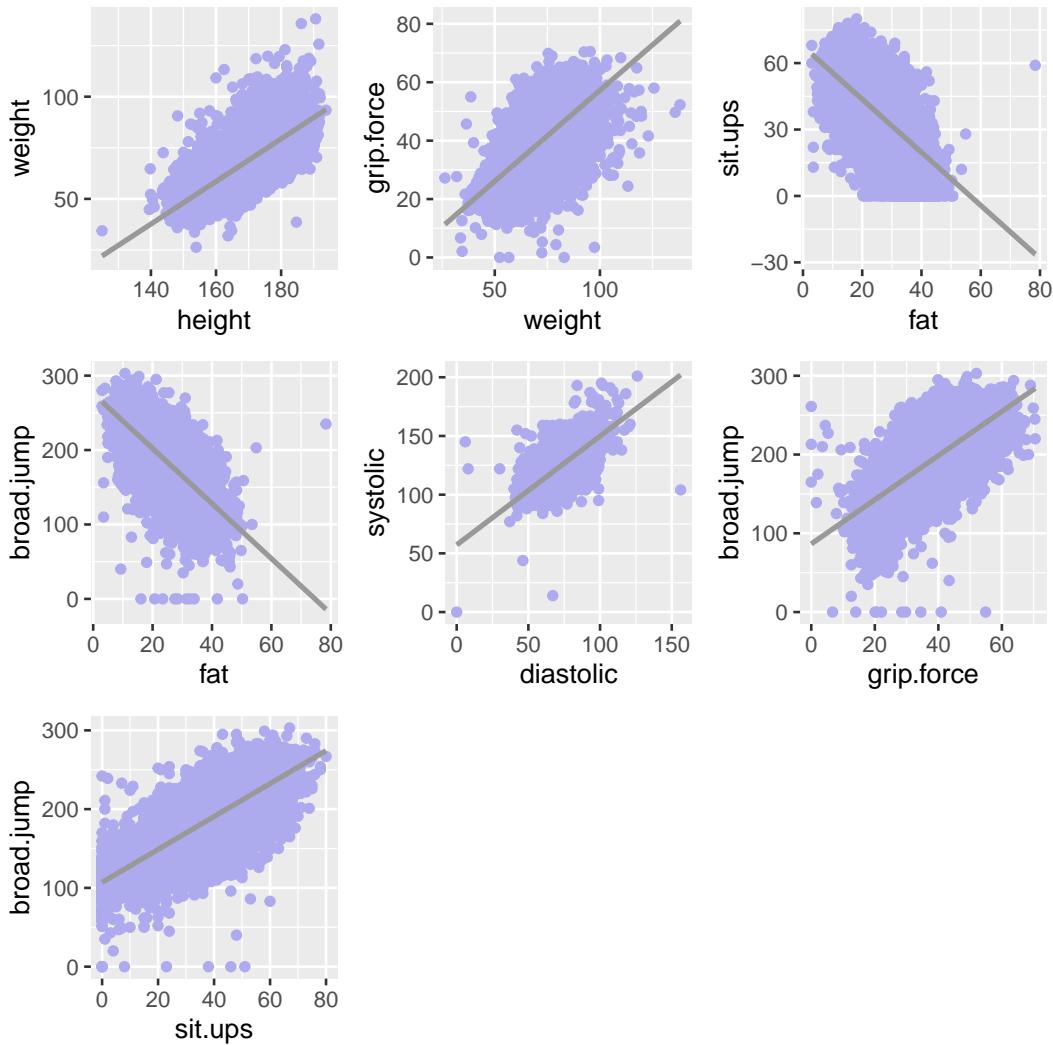
	przedzial_ufnosti_srednia	przedzial_ufnosti_wariancja
height-start	168.41708	69.33676
height-end	168.70253	72.73923
weight-start	67.24492	139.43511
weight-end	67.64971	146.27741
fat-start	23.11725	51.42285
fat-end	23.36308	53.94625
diastolic-start	78.61490	112.67655
diastolic-end	78.97878	118.20577
systolic-start	129.98560	211.40699
systolic-end	130.48403	221.78107
sit-ups-start	39.52941	199.02890
sit-ups-end	40.01303	208.79557
broad jump-start	189.45436	1552.06348
broad jump-end	190.80489	1628.22575

3.4 Korelacja między danymi

Macierz korelacji między cechami liczbowymi wygląda następująco:



Dla par cech z największym współczynnikiem korelacji stworzono osobne wykresy:



- im większy wzrost tym większa waga
- im większa waga tym większa siła chwytu
- im mniejszy procent tkanki tłuszczowej tym więcej wykonanych przysiadów
- im mniejszy procent tkanki tłuszczowej tym lepszy wynik skoku
- im większe ciśnienie tętnicze rozkurczowe tym większe ciśnienie tętnicze skurczowe
- im większa siła chwytu tym lepszy wynik skoku
- im większa liczba wykonanych przysiadów tym lepszy wynik skoku

3.5 Regresja

Dokładniejsza analiza zależności między liczbą zrobionych przysiadów a długością skoku wygląda następująco:

x - sit-ups
y - broad jump

$$y = a \cdot x + b + \varepsilon$$

Call:

```
lm(formula = broad.jump ~ sit.ups, data = body_perf)
```

Residuals:

Min	1Q	Median	3Q	Max
-213.593	-17.503	0.974	17.855	134.975

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	107.02493	0.67647	158.2	<2e-16 ***							
sit.ups	2.08957	0.01601	130.5	<2e-16 ***							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'..'	0.1	' '	1

Residual standard error: 26.45 on 13391 degrees of freedom
Multiple R-squared: 0.5599, Adjusted R-squared: 0.5599
F-statistic: 1.704e+04 on 1 and 13391 DF, p-value: < 2.2e-16

R^2 jest równe 0.5599, więc jest dość oddalone od 1, co oznacza, że poziom dopasowania nie jest bardzo wysoki. Wartość p-value sugeruje jednak, że istnieje znacząca relacja między zmiennymi sit-ups i broad jump.

4 Wnioski

Wnioski płynące z przeprowadzonej analizy są następujące:

- większość rozkładów cech jest zbliżona do rozkładów normalnych
- zazwyczaj osoby szczupłe wykonują lepiej ćwiczenia sportowe
- dane znacząco różnią się w zależności od płci