Akademia Górniczo-Hutnicza im. Stanisława Staszica

Wydział Zarządzania

Kierunek: Informatyka i Ekonometria

Przedmiot: Statystyczna analiza danych, ćwiczenia projektowe

Analiza osiągnięć szczypiornistek podczas Igrzysk Olimpijskich 2020

Wiktoria Mróz, 402108

Kraków, 07.11.2021r.

Spis treści

1. Wstęp

Analizowanie wyników sportowych jest powszechnym zjawiskiem. Nie jest ono jedynie motywowane przez fanów danej dyscypliny. Bardzo często również z punktu widzenia klubów lub reprezentacji, takie zestawienia są kluczowe, bo pozwalają na wybór najlepszych zawodników do drużyny. W tym projekcie rozważane będą dane zebrane podczas Igrzysk Olimpijskich 2020, z reprezentacji kobiet w piłce ręcznej. Utworzony zostanie ranking zawodniczek, przeprowadzone zostanie grupowanie oraz skalowanie wielowymiarowe.

2. Opis danych

2.1 Pochodzenie danych

Dane pozyskane zostały z oficjalnej strony Międzynarodowej Federacji Piłki Ręcznej[1], zawodniczki zostały losowo wybrane z bazy danych i zestawione w tabeli. Każda szczypiornistka posiada informacje o 5 wybranych zmiennych: age - wiek, nominanta; goals - liczba goli, stymulanta; efficiency - skuteczność czyli liczba goli na liczbę rzutów, stymulanta; shots_7m - liczba oddanych karnych, stymulanta oraz steals - liczba przechwyceń piłki, stymulanta.

2.2 Podstawowe statystyki

Tabela 1: Podstawowe statystyki

	age	goals	efficiency	shots_7m	steals
Min	24	7,000	0,280	0,000	0,000
1st Qu	26	11,500	0,588	0,000	1,000
Median	29	18,500	0,665	0,000	1,000
Mean	28,670	21,000	0,631	4,167	2,600
3rd Qu	30	26,250	0,698	5,000	3,750
Max	36	52,000	0,880	28,000	11,000
Sd	3,262	11,759	0,128	7,961	3,001

Dla wieku zawodniczek średnia wartość to około 29 lat. Obserwacje różnią się od średniej o 3. Różnica pomiędzy najmłodszą zawodniczą a najstarszą wynosi 12 lat. Dla liczby goli średnio zawodniczka zdobyła ich 21. Obserwacje różnią się od średniej o 12.
Dla skuteczności, średnio wynosiła 63%. Obserwacje różnią się od średniej o 0,13. Dla liczby karnych średnia wynosi 4. Obserwacje różnią się od średniej o 8. Można zauważyć wyraźnie rozkład prawostronny.

[1] https://www.ihf.info/competitions/women/307/olympic-games-tokyo-2020---womens-tournament/20353/statistics/top-steals, dostęp: 07.11.2021r.

Dla liczby przechwyceń średnia wynosi około 3 na zawodniczkę. Obserwacje różnią się od średniej o 3.

2.3 Korelacja i współczynnik zmienności

Tabela 2: Korelacje pomiędzy zmiennymi

	age	goals	efficiency	shots_7m	steals
age	1	0,327	0,389	0,403	0,292
goals	0,327	1	0,310	0,674	0,246
efficiency	0,389	0,310	1	0,307	0,059
shots_7m	0,403	0,674	0,307	1	-0,152
steals	0,292	0,246	0,059	-0,152	1

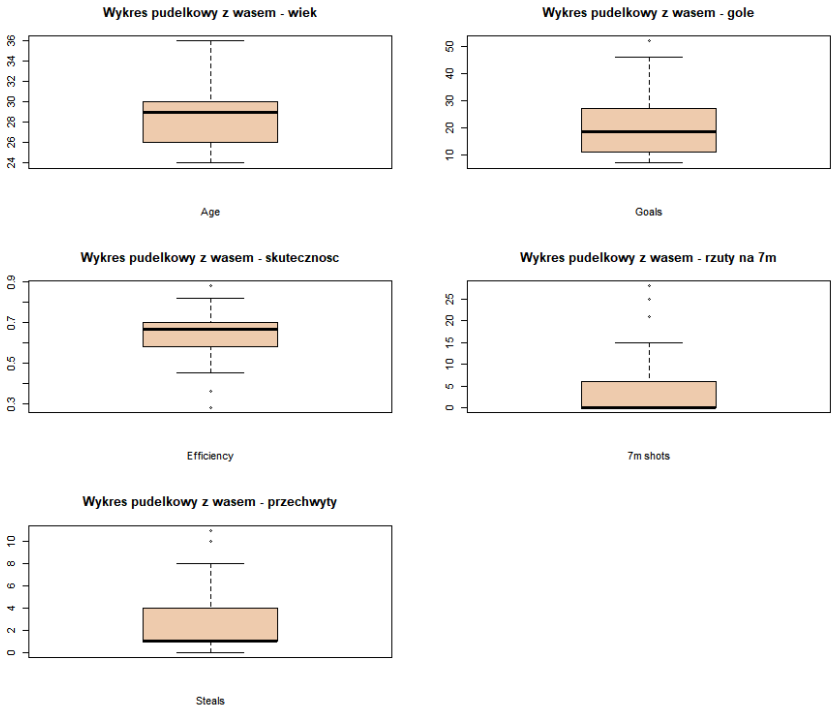
Macierz korelacji pokazuje, że część zmiennych wykazuje istotne korelacje. Jednak przy zmiennej steals występuje niska korelacja ze zmiennymi efficiency i shots_7m.

Tabela 3: Współczynniki zmienności

Zmienna	Age	goals	efficiency	Shots_7m	Steals
Wsp. Zmienności	11,381	55,996	20,275	191,075	115,429

Wszystkie zmienne posiadają pożądany współczynnik zmienności większy od 10.

2.4 Wartości odstające



Rysunek 1: Wykresy pudełkowe z wąsem dla zmiennych

Na rysunku 1. można zauważyć kilka wartości odstających jednak ze względu na charakter badania – ranking zawodniczek, nie będą one usuwane.

2.5 Zamiana zmiennych na stymulanty

Zmienna age odpowiadająca za wiek szczypiornistek jest nominatą. Młody wiek wskazuje na brak doświadczenia, zaś starszy na mniejszą odporność ciała przy wysiłku fizycznym. Wartość nominalna została wybrana z użyciem mody, która dla zestawu danych wynosi 30 lat. Z pomocą wzoru (2.5.1) dane zostały przekształcone by mogły zostać użyte w dalszej analizie.

$$z_{i,j} = \begin{cases} 1 & dla\ x_{ij} = N_j \\ \dfrac{1}{x_{ij} - N_j + 1} & dla\ x_{ij} > N_j \\ \dfrac{-1}{x_{ij} - N_j - 1} & dla\ x_{ij} < N_j \end{cases} \qquad 2.5.1$$

gdzie:
$z_{i,j}$ – obserwacja po zmianie,
x_{ij} – stara obserwacja,
N_j – wartość nominalna.

3. Porządkowanie liniowe

3.1 Metoda Standaryzowanych Sum

Pierwszą metodą wykorzystaną w tym projekcie do porządkowania liniowego jest metoda bezwzorcowa – metoda standaryzowanych sum. Polega ona na sumowaniu wartości otrzymanych po zmianie danych na stymulanty, w przypadku tego projektu zmienna age oraz standaryzacji według wzoru (3.1).

$$s_i = \frac{s_i - \min(s_i)}{\max\{s_i - \min(s_i)\}} \qquad 3.1$$

gdzie: s_i jest sumą oszacowań w ramach obiektu.

Poniższy ranking jest wynikiem zastosowanej metody:

Tabela 4: Ranking - MSS

imię	Wskaźnik	Imię	wskaźnik	imię	wskaźnik
Nora Mork	1	Adriana Cardoso	0,458	Pauline Coatanea	0,229

Ekaterina Ilina	0,810	Camilla Herrem	0,454	Marit Jacobsen	0,204
Jovanka Radicevic	0,744	Elin Hansson	0,417	Liliana da Silva Venancio	0,186
Migyeong Lee	0,689	Laura Flippes	0,406	Kyungmin Kang	0,161
Anna Vyakhireva	0,556	Lois Abbingh	0,402	Igner Smits	0,149
Nathalie Hagman	0,550	Marta Lopez Herrero	0,383	Ema Ramusović	0,123
Estelle Nze Minko	0,532	Shio Fuji	0,354	Jinyi Kim	0,091
Polina Kuznetsova	0,530	Danick Snelder	0,337	Itana Grabić	0,080
Stine Bredal Oftedal	0,512	Lara Gonzalez Ortega	0,328	Haruno Sasaki	0,010
Veronica Kristiansen	0,477	Reka Bordas	0,292	Dijana Mugosa	0

3.2Metoda Hellwiga

Kolejna zastosowana metoda należy do grupy metod wzorcowych, opiera się na obliczaniu odległości od idealnego obiektu, tj. takiego, który ma najlepsze wartości w każdej zmiennej. Następnie należy określić odległość możliwie daleką jako sumę średniej z odległości obiektów z odchyleniem standardowym pomnożonym przez 2. Wartość miary dla obiektu wyznaczona zostaje według wzoru:

$$s_i = 1 - \frac{d_i}{d_0}$$

gdzie:

s_i – wartość miary,
d_i – odległość obiektu od wzorca,
d_0 – odległość możliwie daleka.

Poniższy ranking jest wynikiem zastosowanej metody:

Tabela 5: Ranking - Metoda Hellwiga

imię	wskaźnik	Imię	wskaźnik	imię	wskaźnik
Nora Mork	0,807	Stine Bredal Oftedal	0,372	Igner Smits	0,252

Ekaterina Ilina	0,774	Danick Snelder	0,351	Reka Bordas	0,249
Jovanka Radicevic	0,660	Anna Vyakhiereva	0,348	Kyungmin Kang	0,244
Migyeong Lee	0,590	Elin Hansson	0,336	Camilla Herrem	0,242
Nathalie Hagman	0,499	Lura Flippes	0,330	Liliana da Silva Venancio	0,236
Lois Abbingh	0,443	Lara Gonzalez Ortega	0,321	Ema Ramusović	0,214
Shio Fuji	0,432	Estelle Nze Minko	0,312	Itana Grabić	0,196
Veronica Kristiansen	0,427	Pauline Coatanea	0,291	Jinyi Kim	0,169
Marta Lopez Herrero	0,383	Polina Kuznetsova	0,276	Dijana Mugosa	0,096
Adriana Cardoso	0,375	Marit Jacobsen	0,274	Haruno Sasaki	0,044

3.3 Porównanie

By sprawdzić czy obie metody dają podobne wyniki, użyte zostanie grupowanie według średniej.

Tabela 6: Porównanie metod porządkowania

Grupa	MSS	Hellwig
Gr 1	Nora Mork, Ekaterina Ilina, Jovanka Radicevic, Migyeong Lee	Nora Mork, Ekaterina Ilina, Jovanka Radicevic, Migyeong Lee
Gr 2	Anna Vyakhiereva, Nathalie Hagman, Estelle Nze Minko, Polina Kuznetsova, Stine Bredal Oftedal, Veronica Kristiansen, Adriana Cardoso, Camilla Herrem, Elin Hansson, Lura Flippes, Lois Abbingh, Marta Lopez Herrero	Nathalie Hagman, Lois Abbingh, Shio Fuji, Veronica Kristiansen, Marta Lopez Herrero, Adriana Cardoso, Stine Bredal Oftedal
Gr 3	Shio Fuji, Danick Snelder, Lara Gonzalez Ortega, Reka Bordas, Pauline Coatanea, Marit Jacobsen, Liliana da Silva Venancio, Kyungmin Kang, Igner Smits	Danick Snelder, Anna Vyakhiereva, Elin Hansson, Lura Flippes, Lara Gonzalez Ortega, Estelle Nze Minko, Pauline Coatanea, Polina Kuznetsova, Marit Jacobsen, Igner Smits, Reka Bordas,

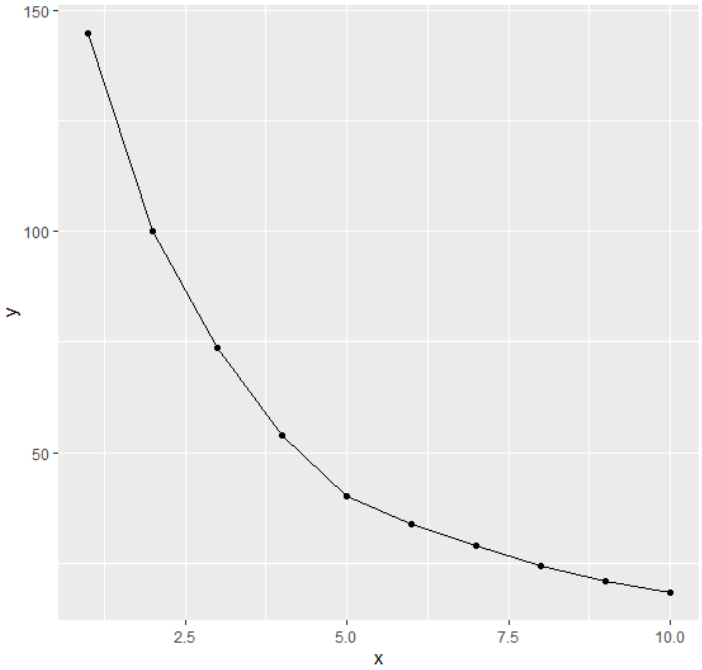
		Kyungmin Kang, Camilla Herrem, Liliana da Silva Venancio, Ema Ramusović, Itana Grabić
Gr 4	Ema Ramusović, Jinyi Kim, Itana Grabić, Haruno Sasaki, Dijana Mugosa	Jinyi Kim, Dijana Mugosa, Haruno Sasaki

Na czerwono zaznaczone zostały zawodniczki, które w zależności od metody znajdują się w innej grupie. Największe różnice są pomiędzy grupą 2 i 3, które są liczniejsze lub nie. Jednak niezależnie od metody wyłaniane są te same 4 najlepsze zawodniczki.
Stabilność porządkowania została sprawdzona przy pomocy współczynnika τ-Kendalla, który wskazuje na wysokie skorelowanie obu rankingów (0,81).

4. Analiza skupień

4.1Metoda podziałowa

Przy metodzie podziałowej przed użyciem algorytmu należy wybrać liczbę grup. Do tego została użyta jedna z metod: łokciowa. Jej istota sprowadza się do wybranie tej liczby grup, po której na wykresie zależności liczby grup (x) od sumy odległości od centroidy (y) następuje najniższy spadek.



Rysunek 2: Wykres zależności liczby grup od sumy odległości od centroidy

Na wykresie (rys. 2) można zauważyć, że ów spadek jest niski począwszy od wartości 5.

4.1.1 Metoda k-średnich

Metoda k-średnich polega na obliczaniu odległości między punktem a centroidami i przyporządkowywanie go do najbliższego skupienia. Ze względu na losowość początkowego przyporządkowywania wyniki w zależności od liczby iteracji mogą się różnić.

Metoda k-średnich dla 5 grup:

Tabela 7: K-means dla 5 grup

Grupa1	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**Max**
Age	7	0,860	1,000	0,240	0,500	1,000
Goals	7	23,140	23,000	7,450	11,000	32,000
Efficiency	7	0,640	0,630	0,040	0,580	0,710
Shots_7m	7	5,140	6,000	5,400	0,000	15,000
Steals	7	3,000	3,000	2,080	1,000	7,000
Grupa2	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**Max**
Age	8	0,300	0,220	0,170	0,140	0,500
Goals	8	12,880	12,000	5,640	7,000	22,000
Efficiency	8	0,720	0,690	0,080	0,640	0,880
Shots_7m	8	1,620	0,000	4,210	0,000	12,000
Steals	8	0,620	0,500	0,740	0,000	2,000
Grupa3	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**Max**
Age	3	0,730	1,000	0,460	0,200	1,000
Goals	3	44,330	46,000	8,620	35,000	52,000
Efficiency	3	0,750	0,720	0,060	0,710	0,810
Shots_7m	3	24,670	25,000	3,510	21,000	28,000
Steals	3	2,000	0,500	2,000	0,000	4,000
Grupa4	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**Max**
Age	7	0,230	0,200	0,080	0,140	0,330
Goals	7	14,290	16,000	5,150	9,000	20,000
Efficiency	7	0,450	0,450	0,100	0,280	0,560
Shots_7m	7	0,290	0,000	0,760	0,000	2,000
Steals	7	1,000	1,000	1,000	0,000	3,000
Grupa5	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**Max**
Age	5	0,190	0,200	0,010	0,170	0,200
Goals	5	26,400	27,000	11,080	14,000	43,000

Efficiency	5	0,660	0,697	0,030	0,610	0,700
Shots_7m	5	0,000	0,000	0,000	0,000	0,000
Steals	5	7,800	8,000	2,770	5,000	11,000

Grupa 1 charakteryzuje się średnią bliską 1 dla zmiennej age, co pozwala wnioskować, że w grupie znajduje się sporo zawodniczek w okolicach 30 lat, średnia goli i karnych również świadczy, że są to osoby doświadczone, jednak ze względu na skuteczność nie są to najlepsze zawodniczki.

Grupa 2, porównując średnie, zawiera przeciętne wyniki. Analizując pozycje zawodniczek możemy zauważyć, że trafiły tutaj wszystkie szczypiornistki grające na pozycji obrotowego.

Grupa 3 w większości zmiennych ma najlepsze wyniki, oprócz przechwyceń. Do tej grupy zakwalifikowały się najlepsze zawodniczki z zestawu danych.

Grupa 4 charakteryzuje się dość niską średnią wieku, jednak patrząc na inne zmienne możemy wnioskować, że dalej są to aktywne na boisku zawodniczki

Grupa 5 ma najniższą średnią wieku, znajdują się tu najmłodsze jak i najstarsze zawodniczki. Szczególnie wskazuje na to brak karnych. Pod względem średniej liczby przechwytów w tej grupie jest ona największa z zestawu. Analizując pozycje zawodniczek, możemy zauważyć, że większość to skrzydłowe.

Metoda k-średnich dla 6 grup:

Tabela 8:K-means dla 6 grup

Grupa 1	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**Max**
Age	7	0,230	0,200	0,080	0,140	0,330
Goals	7	14,290	16,000	5,150	9,000	20,000
Efficiency	7	0,450	0,450	0,100	0,280	0,560
Shots_7m	7	0,290	0,000	0,760	0,000	2,000
Steals	7	1,000	1,000	1,000	0,000	3,000
Grupa 2	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**Max**
Age	9	0,320	0,250	0,170	0,140	0,500
Goals	9	14,110	13,000	6,450	7,000	24,000
Efficiency	9	0,710	0,680	0,080	0,630	0,880
Shots_7m	9	02,220	0,000	4,320	0,000	12,000
Steals	9	0,670	1,000	0,710	0,000	2,000
Grupa 3	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**Max**
Age	5	1,000	1,000	0,000	1,000	1,000
Goals	5	21,000	22,000	7,660	11,000	32,000

Efficiency	5	0,640	0,620	0,050	0,580	0,710
Shots_7m	5	5,600	6,000	6,190	0,000	15,000
Steals	5	3,400	3,000	2,300	1,000	7,000
Grupa 4	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**Max**
Age	4	0,270	0,200	0,160	0,170	0,500
Goals	4	32,750	30,500	7,140	27,000	43,000
Efficiency	4	0,660	0,660	0,030	0,610	0,680
Shots_7m	4	0,250	0,000	0,500	0,000	1,000
Steals	4	5,250	5,000	2,060	3,000	8,000
Grupa 5	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**Max**
Age	2	0,200	0,200	0,000	0,200	0,200
Goals	2	16,500	16,500	3,540	14,000	19,000
Efficiency	2	0,690	0,690	0,020	0,670	0,700
Shots_7m	2	0,000	0,000	0,000	0,000	0,000
Steals	2	10,500	10,500	0,710	10,000	11,000
Grupa 6	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**Max**
Age	3	0,730	1,000	0,460	0,200	1,000
Goals	3	44,330	46,000	8,620	35,000	52,000
Efficiency	3	0,750	0,720	0,060	0,710	0,810
Shots_7m	3	24,670	25,000	2,510	21,000	28,000
Steals	3	2,000	2,000	2,000	0,000	4,000

Grupa 1 charakteryzuje się dość młodymi zawodniczkami, które jeszcze nie mają dużego doświadczenia. Głównie są to skrzydłowe.

Grupa 2 to zawodniczki, które nie zdobywają wielu goli jednak są wysoce skuteczne. Trafiły tutaj min. zawodniczki grające na pozycji obrotowego.

Grupa 3: zawodniczki mające 30 lat. Grupa charakteryzuje się sporą średnią goli i karnych, jednak nie są to najlepsze zawodniczki ze względu na skuteczność.

Grupa 4 posiada zawodniczki młode, średnia wieku jest oddalona od nominalnej wartości. Duża średnia przechwyceń, więc są to aktywne na boisku zawodniczki

Grupa 5 Starsze, doświadczone zawodniczki, które grają na lewym skrzydle, stąd dość niska średnia goli, ale wysoka średnia przechwytów.

Grupa 6: praktycznie wszystkie zawodniczki mają 30 lat, wysoka średnia goli i karnych świadczy, że znalazły się w niej najlepsze zawodniczki.

Grupa 1 i grupa 2 są do siebie zbliżone pod względem statystyk. Sugeruje to, że nie ma powodu dzielić obserwacji na 6 grup.

4.1.2 Algorytm PAM

Algorytm technicznie jest podobny do metody k-średnich, jednak tutaj zamiast centroidy, która nie jest odporna na wartości odstające, rozważana jest modoida. Ze względu na wcześniej zauważone pokrywanie się dwóch grup, zostanie wybrane 5 skupień.

Tabela 9: Algorytm PAM dla 5 grup

Grupa 1	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**max**
Age	3	0,73	1	0,46	0,2	1
Goals	3	44,33	46	8,62	35	52
Efficiency	3	0,75	0,72	0,06	0,71	0,81
Shots_7m	3	24,67	25	3,51	21	28
Steals	3	2	2	2	0	4
Grupa 2	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**max**
Age	5	0,41	0,20	0,35	0,17	1
Goals	5	30,8	29	7,56	23	43
Efficiency	5	0,65	0,66	0,03	0,61	0,68
Shots_7m	5	0,2	0	0,45	0	1
Steals	5	5	5	1,87	3	8
Grupa 3	**Liczba obserwacji**	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**max**
Age	4	0,66	0,75	0,42	0,14	1
Goals	4	25	23	4,76	22	32
Efficiency	4	0,66	0,65	0,04	0,62	0,71
Shots_7m	4	10,25	9,5	3,95	7	15
Steals	4	1,25	1	1,26	0	3
Grupa 4	**Liczba obserwacji**	**średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**max**
Age	10	0,38	0,29	0,26	0,14	1
Goals	10	17,2	17,5	2,35	13	20
Efficiency	10	0,62	0,62	0,11	0,45	0,82
Shots_7m	10	0,3	0	0,67	0	2
Steals	10	3,3	1	4,32	0	11
Grupa 5	**Liczba obserwacji**	**średnia**	**Mediana**	**Odchylenie standardowe**	**Min**	**max**
Age	8	0,3	0,2	0,29	0,14	1
Goals	8	8,88	9	1,55	7	11
Efficiency	8	0,58	0,64	0,2	0,28	0,88
Shots_7m	8	0,75	0	2,12	0	6
Steals	8	1,12	1	0,99	0	3

1 grupa PAM pokrywa się statystycznie z 3 grupą z k-średnich znajdują się tu najlepsze zawodniczki.

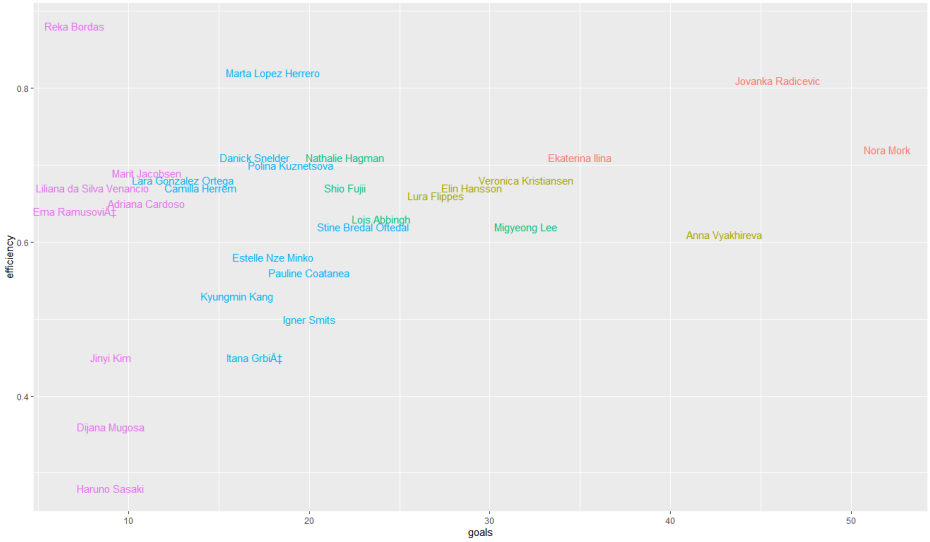
Grupa 2 charakteryzuje się dużą średnią oddanych goli oraz sporą ilością przechwytów.

Grupa 3 to głównie rozgrywające, doświadczone zawodniczki. Wskazuje na to średnia liczba goli jak i średnia liczba oddanych karnych.

W grupie 4 znajdują się przeciętne zawodniczki, które nie wyróżnia nic spośród reszty.

Na grupę 5 składają się głównie młode zawodniczki, które ogólnie mają gorsze wyniki od reszty grup.

Występują różnice w przyporządkowywaniu zawodniczek do grupy w odniesieniu do metody k-średnich i skutkuje to ogólnym mniejszym odchyleniem od średniej w grupie. Obiekty są dość bliskie sobie w obrębie skupienia, co jest pożądaną cechą w analizie skupień.



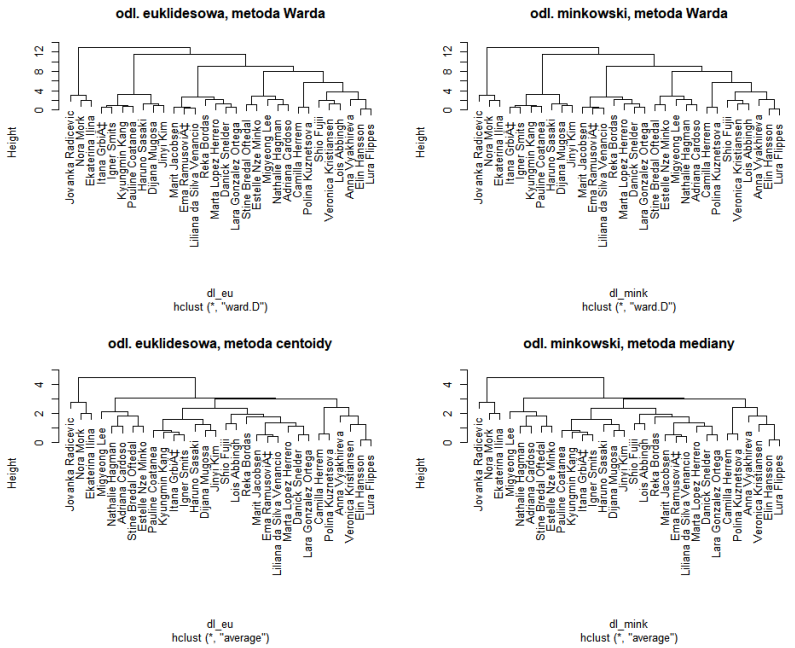
Rysunek 3: Zależność goli od skuteczności

Na wykresie przedstawiającym liczbę goli od skuteczności wyraźnie widać jak przebiega grupowanie. Z dużą liczbą goli i skutecznością są najlepsze zawodniczki. Stopniowo, w ramach przekątnej, usytuowane są słabsze i mniej efektywne szczypiornistki, z wyjątkiem Reki Bordas.

4.2 Grupowanie hierarchiczne

Sprowadza się ono do założenia, że każdy obiekt jest osobną grupą, które następnie za pomocą funkcji odległości zostają połączone do momentu uzyskania jednego skupienia.

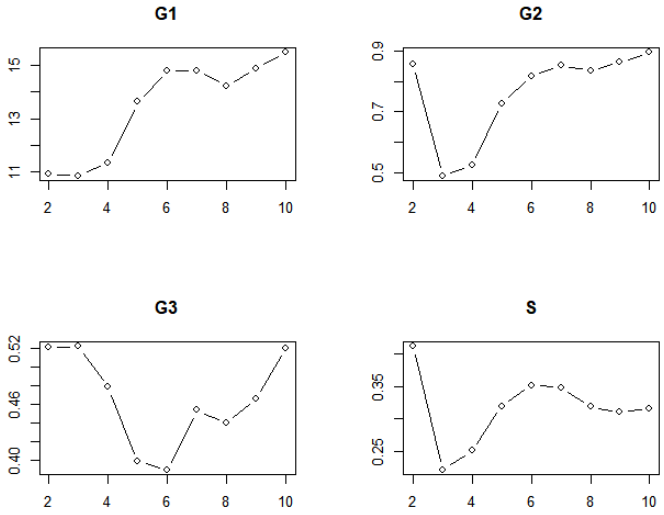
W przypadku tego projektu rozważane odległości to: euklidesowa i Minkowskiego oraz metody: Warda i z użyciem wielowymiarowej średniej.



Rysunek 4: Grupowanie hierarchiczne - dendrogramy

Rysunek 4 przedstawia brak zależności pomiędzy wybranymi odległościami. Na wygląd dendrogramów mają jedynie wpływ użyte metody.

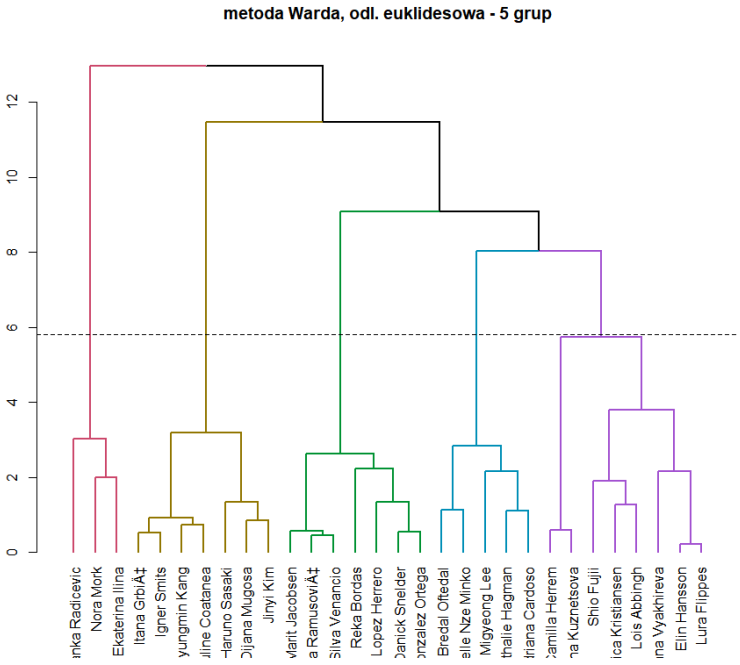
Do określania najlepszego wyboru liczby grup dla odległości euklidesowej i metody Warda można użyć indeksów. Na poniższych wykresach zostały zaprezentowane wartości indeksów w zależności od liczby grup.



Rysunek 5: Wykresy indeksów

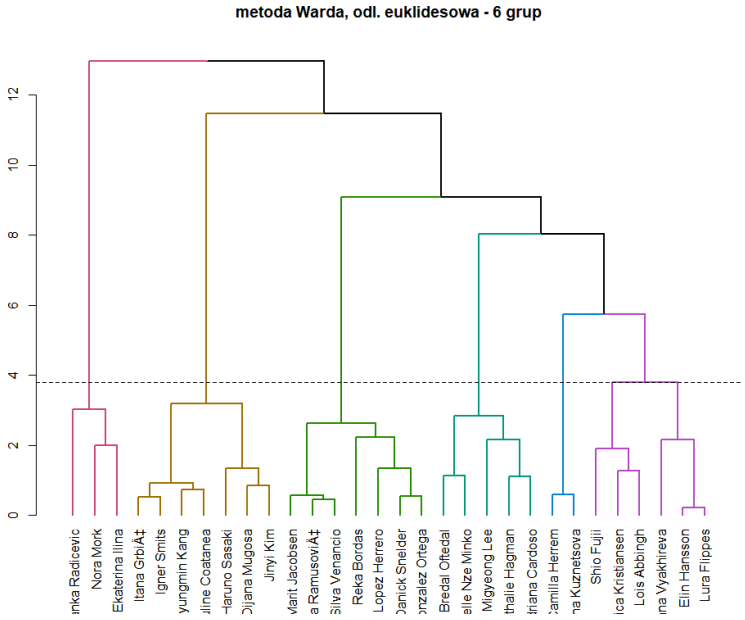
Indeksy nie pozwalają wybrać liczby grup, ponieważ wskazują na różne liczby. Gdyby pominąć skrajne wartości, to dla G1, G2 i S funkcja osiąga maksimum dla liczby 6, w G3 minimum również wskazuje na 6.

Bazując na tych informacjach grupowanie zostanie rozważane dla pięciu (wybranej liczby przy porządkowaniu podziałowym) i sześciu grup.



Rysunek 6: dendrogram- metoda Warda, odl. euklidesowa, 5 grup

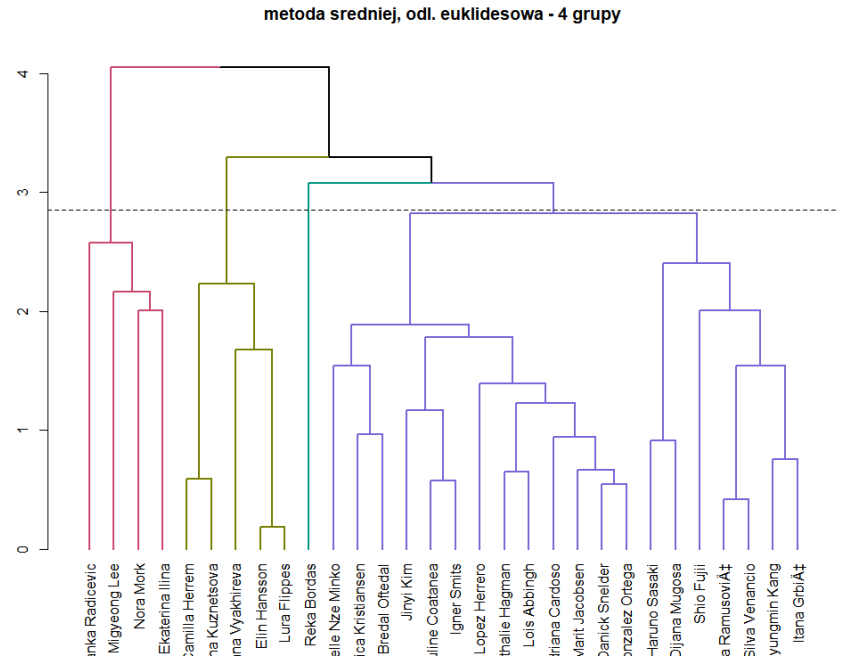
Wykres dla odległości Euklidesowej i metody Warda przedstawia obserwacje podzielone na 5 grup.



Rysunek 7:dendrogram- metoda Warda, odl. euklidesowa, 6 grup

Wykres dla odległości Euklidesowej i metody Warda przedstawia obserwacje podzielone na 6 grup.

Analizując dendrogramy dla odległości euklidesowej i metody średniej, podział na 4 grupy wydaje się być optymalnym wyborem, jednak odnosi się on konkretnie do tej metody grupowania.



Rysunek 8:dendrogram- metoda średniej, odl. euklidesowa, 4 grupy

Dzieląc dane na 4 grupy, spora część obserwacji przyporządkowana zostaje do zielonej grupy. Pożądaną przez analizę skupień cechą jest posiadanie grup o zbliżonej liczności. Z tego powodu należy odrzucić powyższy dendrogram.

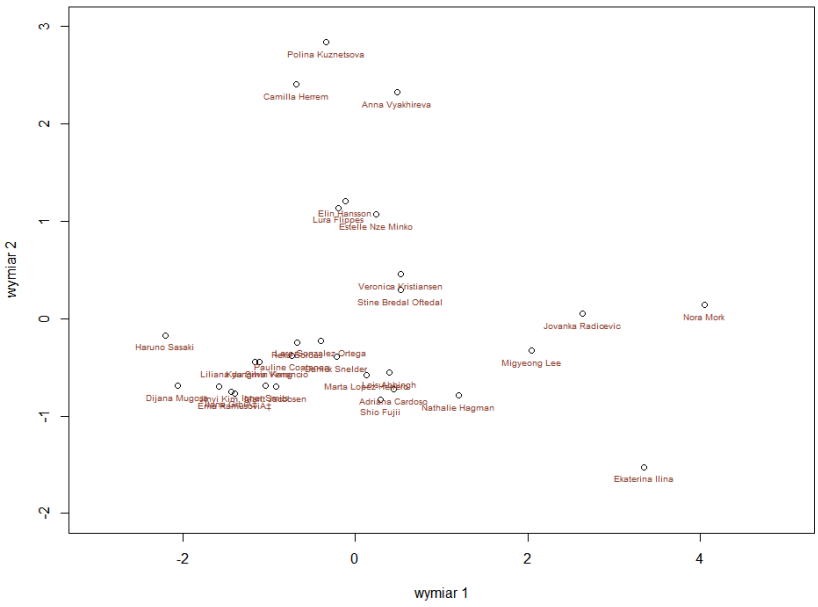
5. Skalowanie wielowymiarowe

Idea tej metody polega na przedstawieniu obiektów wielowymiarowych w przestrzeni o mniejszej liczbie wymiarów. Poszukiwana jest funkcja, która przekształca rzeczywiste odległości na skalowane, starając się przy tym utracić jak najmniej informacji o obiektach.

5.1 Klasyczne skalowanie wielowymiarowe

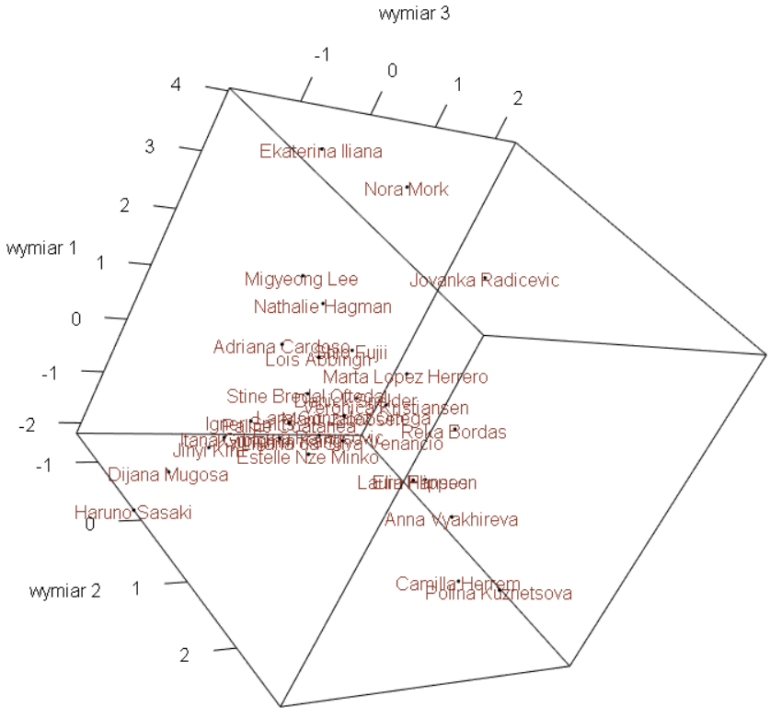
Klasyczne skalowanie wielowymiarowe używa odległości euklidesowej. Jest to metoda jednokrokowa oraz liniowa. Rozpoczyna się ją od obliczenia macierzy odległości.

Przeskalowanie obserwacji na 2 wymiary klasycznym skalowaniem zaprezentowane zostało na poniższym rysunku.



Rysunek 9: Klasyczne skalowanie wielowymiarowe: 2 wymiary

By ocenić jakość rzutowania, użyta została funkcja STRESS. Dla danych na 2 wymiarach wynosi ona 0.3 co nie jest zadowalającym wynikiem. Należy zwiększyć liczbę wymiarów.

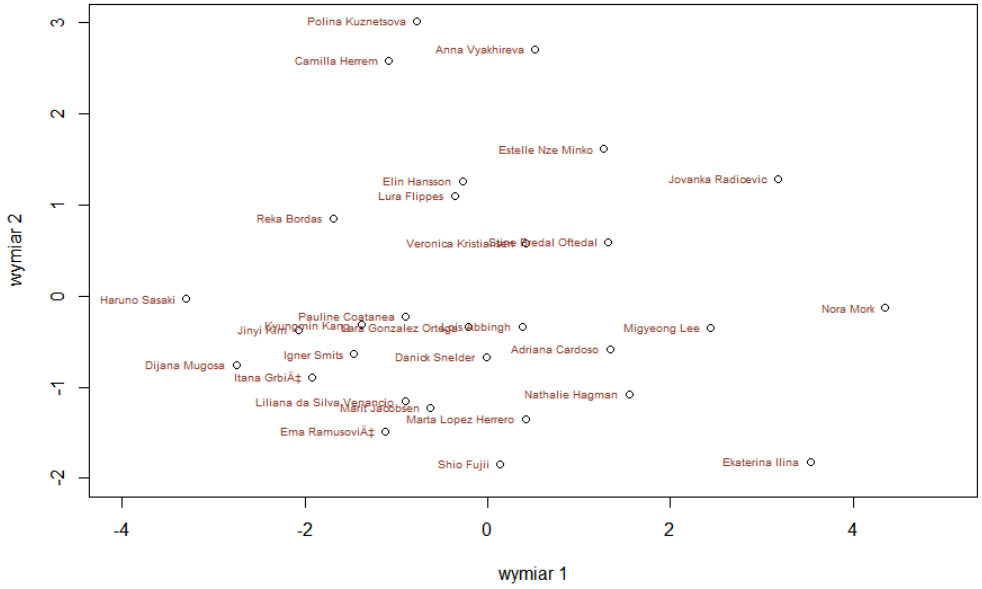


Rysunek 10: Klasyczne skalowanie wielowymiarowe: 3 wymiary

Wartość funkcji STRESS zmniejszyła się do 0,15 jednak dalej jest to niezadowalający wynik.

5.2 Metoda Sammona

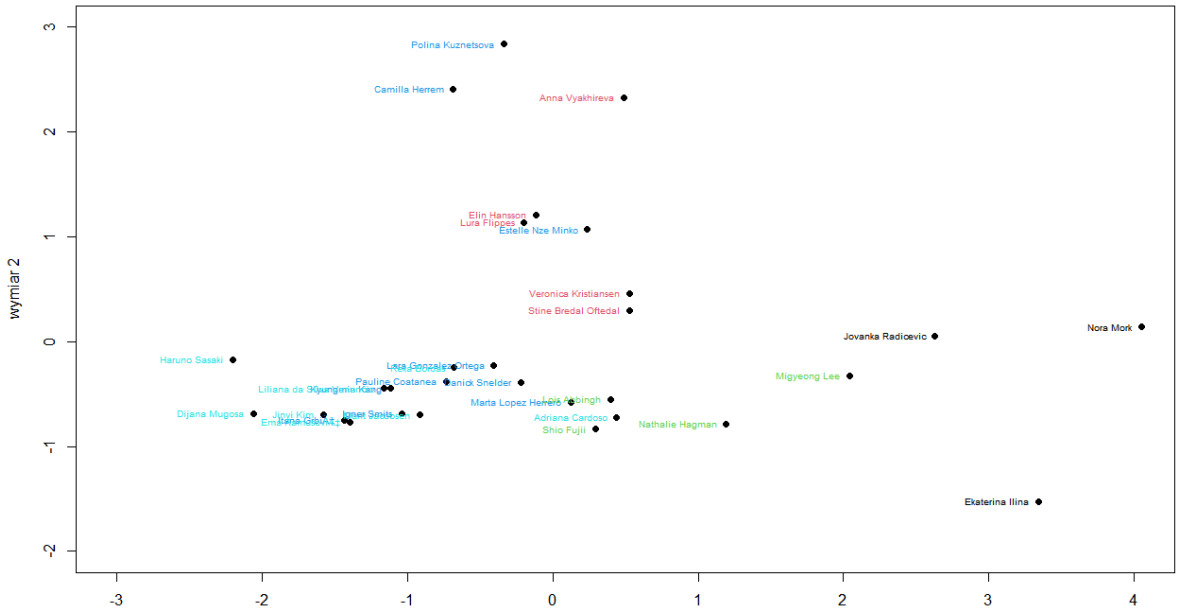
Oprócz klasycznej metody, dość popularna jest również nieliniowa, iteracyjna metoda Sammona. Skupia ona uwagę na porządku rangowym niżeli odległości pomiędzy obiektami.



Rysunek 11: Metoda Sammona: 2 wymiary

Wartość funkcji STRESS dla 2 wymiarów wynosi 0,04 co jest bardzo dobrym wynikiem.

5.3 Skalowanie wielowymiarowe a wyniki grupowania



Rysunek 12: Skalowanie wielowymiarowe i wyniki grupowania

Wymiar 1 można zinterpretować jako technika i doświadczenie zawodniczek, wymiar 2 jako aktywność na boisku, szczególny wpływ na nią ma liczba przechwytów.
Na powyższym wykresie kolorami zostały zaznaczone grupy uzyskane poprzez algorytm PAM. Kolorem czarnym zostały zaznaczone najlepsze zawodniczki, zielonym grupa 3 czyli głównie rozgrywające, doświadczone zawodniczki z wysoką średnią liczbą goli jak i średnią liczbą oddanych karnych. Na różowo grupa 2 charakteryzująca się dużą średnią oddanych goli oraz sporą ilością przechwytów. Na ciemny niebieski zaznaczona została grupa 4 czyli przeciętne zawodniczki, które nie wyróżnia nic spośród reszty. Na jasno niebieski grupa 5 składająca się głównie z młodych zawodniczki, które ogólnie mają gorsze wyniki od reszty grup.

Oprócz grupy najlepszych zawodniczek, trudno jest jasno wskazać miejsca przebiegu granic grup. Sugeruje to, że choć w praktyce podział wydawał się dobry, to niekoniecznie sprawdza się przy skalowaniu wielowymiarowym.

6. Podsumowanie

Metody porządkowania liniowego pozwoliły wyłonić 3 najlepsze zawodniczki, są nimi: Nora Mork, Ekaterina Iliana, Javaka Radicievic. Porównawszy rankingi, które zostały stworzone metodą bezwzorcową i wzorcową, otrzymany został dość wysoki współczynnik τ-Kendalla. Oznacza to, że obie metody w podobny sposób ułożyły zawodniczki w rankingu.

Najlepsze wyniki grupowania podziałowego zostały uzyskane z pomocą algorytmu

PAM. Z jego pomocą zbiór szczypiornistek został podzielony na 5 grup. Grupowanie hierarchiczne i stworzone dendrogramy pokazały, że dla danych należało użyć odległości euklidesowej oraz metody Warda, przy podziale zbioru na 5 grup.

Funkcja STRESS, która ocenia jakość rzutowania danych wielowymiarowych na mniejszą liczbę wymiarów dała najmniejszy wynik przy użyciu metody Sammona i 2 wymiarach, które można interpretować jako technika i doświadczenie zawodniczek oraz aktywność na boisku.