Projekt 2

PW

2 grudnia 2018

1. Wprowadzenie i przygotowanie danych

1.1 Wprowadzenie

Przedmiotem projektu będzie przeprowadzenie analizy skupień (grupowania podziałowego i hierarchicznego) oraz analiza wyników. Analiza Skupień jest dziedziną zajmującą się organizowaniem obserwacji w pewne stryktury - grupy, które cechować mają się jak największym podobieństwem wśród obserwacji jednej grupy oraz jak największymi różnicami pomiędzy poszczególnymi grupami. Analiza skupień pozwala na wykrycie odpowiednich struktur, lecz bez wyjaśnienia dlaczego one występują i dlaczego w takiej właśnie formie. Jest metodą uczenia maszynowego bez nadzoru.

1.2 Opis danych

W ramach projektu użyję znalezione przeze mnie dane Przedstawiają one statystyki meczowe dla poszczególnych piłkarzy z pola występujących w czołowych zespołach Premier League. Statystyki dotyczą sezonu 2017/2018, pochodzą ze strony https://www.premierleague.com/home. W zebranych danych uwzględniłem jedynie zawodników występujących w klubach regularnie walczących o europejskie puchary, którzy rozegrali co najmniej 25 meczów w danym sezonie.

Dane przedstawiają się następująco:

	Goals	Offsides	Shots	Assists	Passer per match	Big chances created	Crosses	Interceptions	Clearences
Alexandre Lacazette	14	23	68	4	22.56	5	16	7	4
Andreas Christensen	0	1	12	0	49.96	0	0	38	104
Chris Smalling	4	2	13	0	39.21	0	0	63	159
David Silva	9	4	54	11	83.76	14	74	18	6
Dejan Lovren	2	1	14	1	62.17	1	4	38	142
Delle Alli	9	5	69	10	34.75	16	18	19	18

Zmienne w kolumnach opisują:

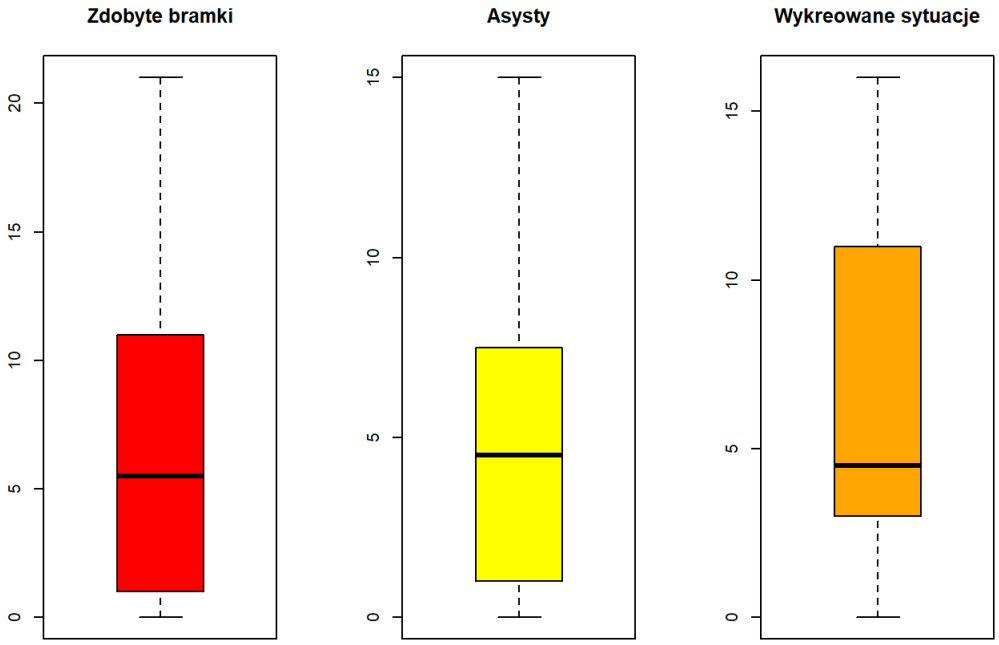
- **Goals** - liczba bramek zdobyta przez zawodnika,
- **Offsides** - liczba spalonych danego zawodnika,
- **Shots** - liczba oddanych strzałów,
- **Passer per match** - liczba podań wykonywana przez danego zawodnika średnio w jednym meczu,
- **Big Chances Created** - liczba sytuacji stuprocentowych jakie dany zawodnik wykreował kolegom z drużyny,
- **Crosses** - liczba dośrodkowań w pole karne przeciwnika,
- **Interceptions** - liczba odbiorów/przechwyceń piłki,
- **Clearences** - liczba wybić piłki, pozwalających na oddalenie zagrożenia od bramki.

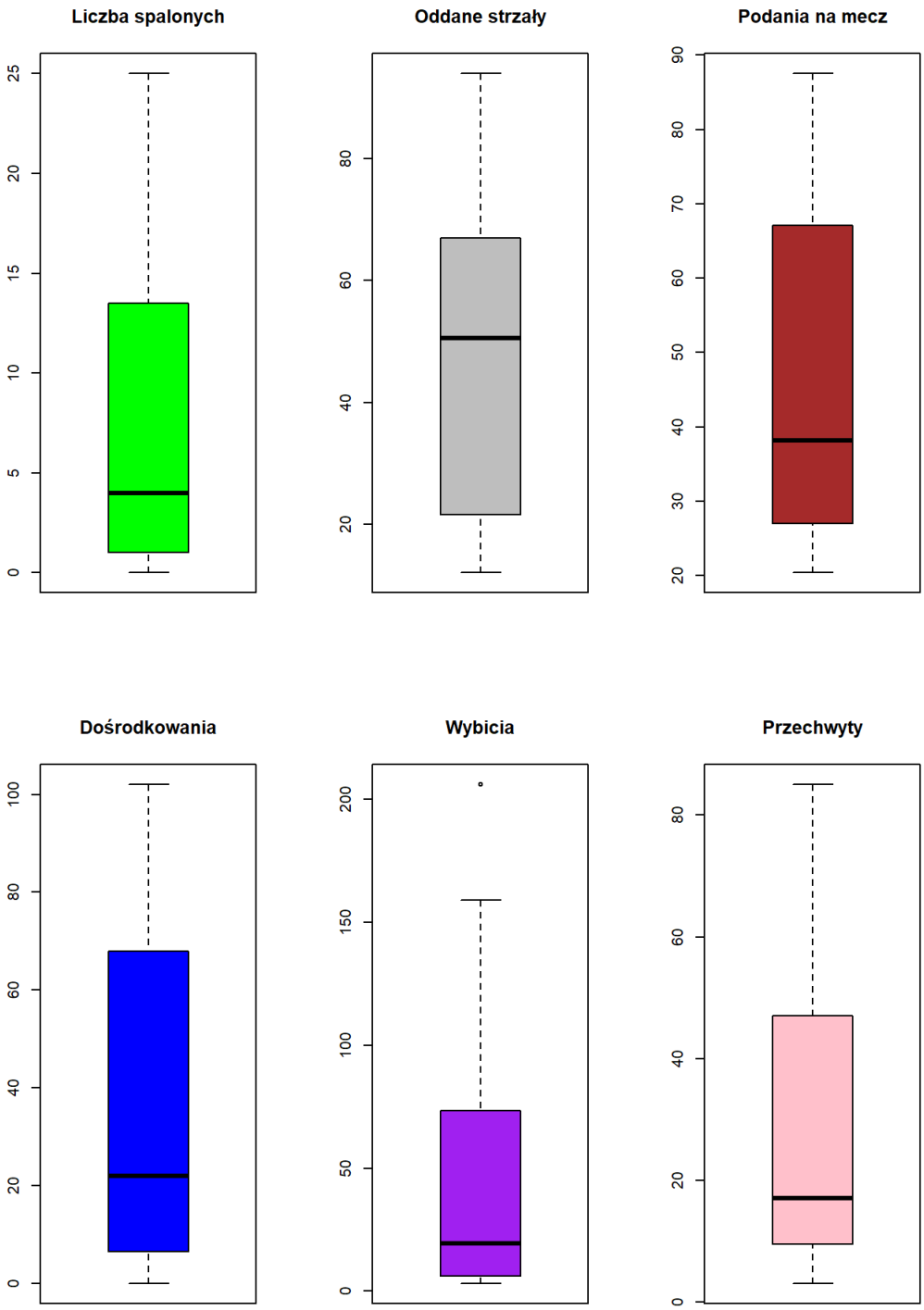
1.2 Analiza danych

W celu wykrycia potencjalnych **outlier'ów** oraz sprawdzenia jak przedstawiają się **podstawowe statystyki opisowe** zebranych danych za pomocą funkcji `summary` oblicze je dla poszczególnych zmiennych.

```
##      Goals           Offsides          Shots          Assists
##  Min.   : 0.000   Min.   : 0.000   Min.   :12.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:23.25   1st Qu.: 1.00
##  Median : 5.500   Median : 4.000   Median :50.50   Median : 4.50
##  Mean   : 6.958   Mean   : 7.375   Mean   :47.25   Mean   : 5.00
##  3rd Qu.:10.500   3rd Qu.:12.750   3rd Qu.:66.50   3rd Qu.: 7.25
##  Max.   :21.000   Max.   :25.000   Max.   :94.00   Max.   :15.00
##  Passer per match Big chances created    Crosses       Interceptions
##  Min.   :20.38    Min.   : 0.00       Min.   :  0.00   Min.   : 3.00
##  1st Qu.:27.60    1st Qu.: 3.00       1st Qu.:  7.75   1st Qu.: 9.75
##  Median :38.13    Median : 4.50       Median : 22.00   Median :17.00
##  Mean   :47.10    Mean   : 6.75       Mean   : 37.96   Mean   :28.12
##  3rd Qu.:66.61    3rd Qu.:11.00       3rd Qu.: 65.00   3rd Qu.:47.00
##  Max.   :87.50    Max.   :16.00       Max.   :102.00   Max.   :85.00
##    Clearences
##  Min.   :  3.00
##  1st Qu.:  6.00
##  Median : 19.50
##  Mean   : 50.25
##  3rd Qu.: 69.25
##  Max.   :206.00
```

Widzimy, że w niektórych zmiennych wartość 1 kwantyla jest stosunkowo bliska wartości minimalnej. Poszczególne zmienne ciężko porównywać ze sobą ze względu na to że przyjmują różne wartości - np. liczba bramek jest z zakresu 0-21, podczas gdy liczba przechwytów jest z zakresu 3-85. W kolejnych etapach badania zdecyduję się zestandaryzować zmienne, aby były porównywalne i miały porównywalny wpływ na wyniki grupowania.
Aby lepiej zobrazować dane oraz sprawdzić możliwość występowania ewentualnych outlier'ów narysuję wykresy pudełkowe dla poszczególnych zmiennych. Wykorzystam funkcję `boxplot`.





Widzimy, że na podstawie wykresów pudełkowych jedynym sugerowanych outlierem jest wartość maksymalna liczby wybić. **Ze względu na to, że to jedyny taki przypadek oraz nie jest to wartość znacząco większa od wartości zaznaczonej jako granica (koniec) wąsa gónego, decyduję się zostawić obserwację w dalszymi etapie badania.** Ponieważ algorytmy analizy skupień są bardzo czułe na występowanie outlier'ów, to w razie gdyby jego pozostawienie spowodowało problemy z wynikami grupowania, usunę go, po czym przeprowadzę ponowne grupowanie.

1.3 Sprawdzenie danych pod kątem formalnym i standaryzacja

Ważnym aspektem przygotowania danych do badania jest sprawdzenie wymogów formalnych dotyczących formalnego doboru zmiennych do analizy. **Użyte w grupowaniu zmienne powinny charakteryzować się zmiennością na poziomie wyższym niż 0.1, a także współliniowością mniejszą niż 0.9.**

Wyniki przedstawiają się następująco:

	Goals	Offsides	Shots	Assists	Passer per match	Big chances created	Crosses	Interceptions	Clearences

	Goals	Offsides	Shots	Assists	Passer per match	Big chances created	Crosses	Interceptions	Clearences
Goals	1.00	0.82	0.87	0.49	-0.63	0.58	0.06	-0.74	-0.62
Offsides	0.82	1.00	0.71	0.33	-0.64	0.39	0.06	-0.68	-0.53
Shots	0.87	0.71	1.00	0.65	-0.45	0.66	0.32	-0.74	-0.72
Assists	0.49	0.33	0.65	1.00	-0.18	0.83	0.69	-0.64	-0.66
Passer per match	-0.63	-0.64	-0.45	-0.18	1.00	-0.17	0.05	0.58	0.30
Big chances created	0.58	0.39	0.66	0.83	-0.17	1.00	0.44	-0.59	-0.65
Crosses	0.06	0.06	0.32	0.69	0.05	0.44	1.00	-0.45	-0.52
Interceptions	-0.74	-0.68	-0.74	-0.64	0.58	-0.59	-0.45	1.00	0.65
Clearences	-0.62	-0.53	-0.72	-0.66	0.30	-0.65	-0.52	0.65	1.00

Wśród otrzymanych wyników nie ma korelacji wyższej niż graniczna wartość 0.9. Są zmienne bardzo skorelowane, lecz skoro spełniają warunek o którym wspomniałem, to mimo wszystko decyduję się je zostawić w badaniu.

Teraz sprawdzę współczynnik zmienności:

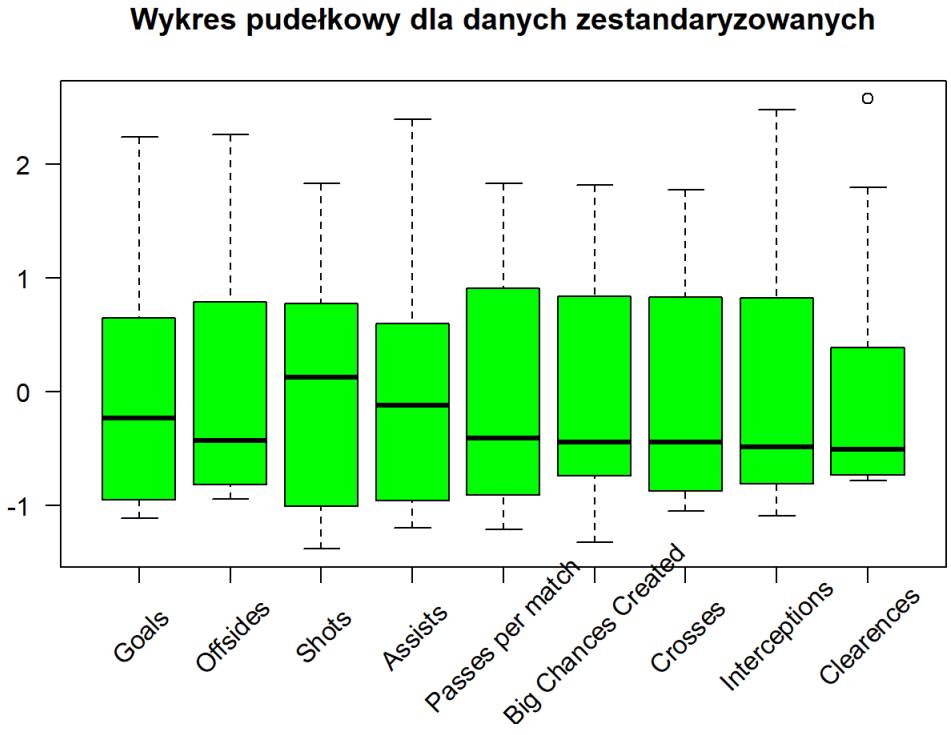
	x
Goals	0.900
Offsides	1.057
Shots	0.540
Assists	0.836
Passer per match	0.469
Big chances created	0.753
Crosses	0.952

Widzimy, że wszystkie wartości są większe od wymaganego poziomu zmienności, więc zmienne **nadają się do dalszej części analizy.**

Jak wspomniałem wcześniej, aby uzyskać porównywalność i podobny wpływ zmiennych na wynik, należy je zestandaryzować. Używam do tego funkcji `scale`.

```
dane_st <- scale(dane)
```

Tak przygotowane i sprawdzone dane mogą być już użyte w procesie grupowania, do którego teraz przechodzę. Zanim jednak dokonam grupowania, przedstawiam wykres pudełkowy zmiennych po standaryzacji, żeby zobrazować jak przedstawiają się zmienne po zestandaryzowaniu.



2. Przeprowadzenie analizy skupień

W tej części projektu zajmę się grupowaniem zebranych danych. Przeprowadzone zostanie zarówno grupowanie metodą podziałową jak i metodą hierarchiczną. Po przeprowadzonych badaniach omówię otrzymane wyniki oraz wybiorę moim zdaniem najlepszy wynik.

2.1 Grupowanie podziałowe

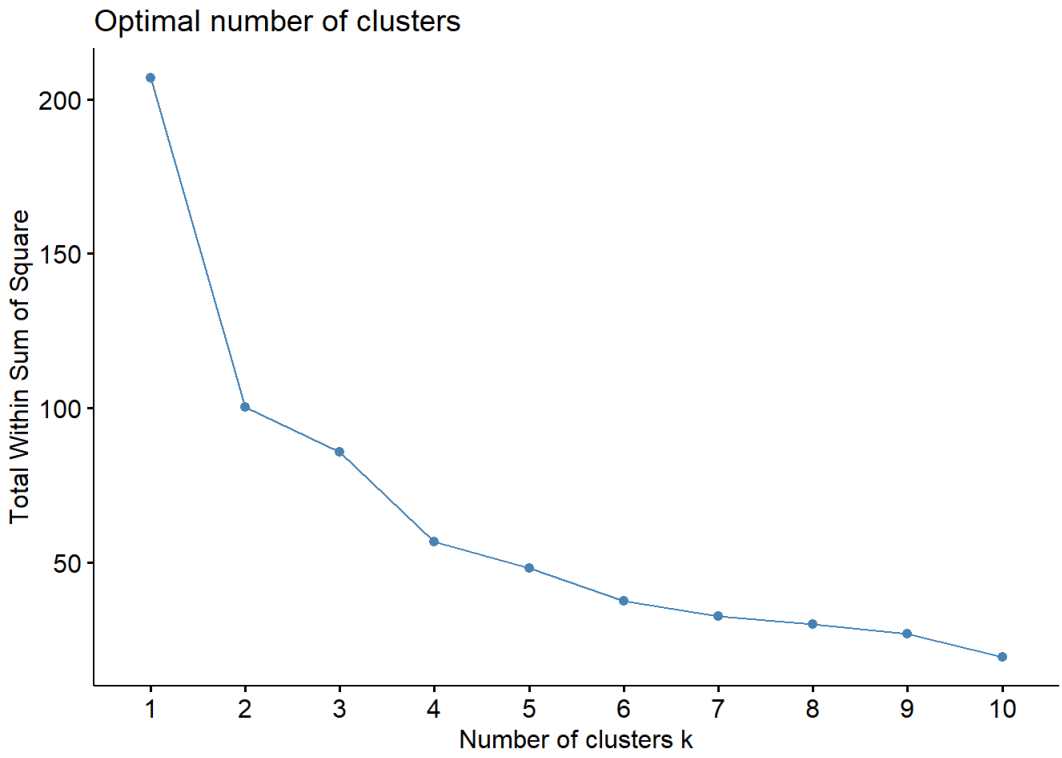
Metoda grupowania podziałowego polega na wyselekcjonowaniu spośród danych obserwacji k-grup (k-skupień). Co ważne, owe skupienia są rozłączne, a ich liczba (k) musi być określona przed początkiem badania.
Do przeprowadzenia grupowania podziałowego wykorzystam metody k-średnich oraz algorytm PAM, będący odmianą metody k-medoid. W obu przypadkach, w algorytmach użyję najpopularniejszy i najczęściej stosowany rodzaj odległości - odległość Euklidesową.

2.1.1 Wybór liczby klastrów

Do wyboru liczby grup wykorzystam tzw. metodę łokciową. W grupowaniu skład grupy dobierany jest tak, aby całkowita wariancja wewnątrz klastra była możliwie najmniejsza. Metoda łokciowa bada jaka jest całkowita suma kwadratów wewnątrz klastra. Naszym celem jest, aby była jak najmniejsza, (im więcej grup tym ta suma jest mniejsza, bo grupy różnią się od siebie coraz bardziej) lecz jednocześnie pamiętać musimy o sensownym wyborze liczby grup, tak aby był on interpretowalny i możliwy do uzasadnienia ‘życiowo’.
Aby sprawdzić tą metodą optymalną liczbę klastrów stworzę tzw. wykres łokciowy, za pomocą funkcji `fviz_nbclust` z pakietu `factoextra`. Na jego podstawie, za optymalną liczbę klastrów uważa się liczbę, dla której wykres ‘załamuje się’, przypominając tym samym łokieć.

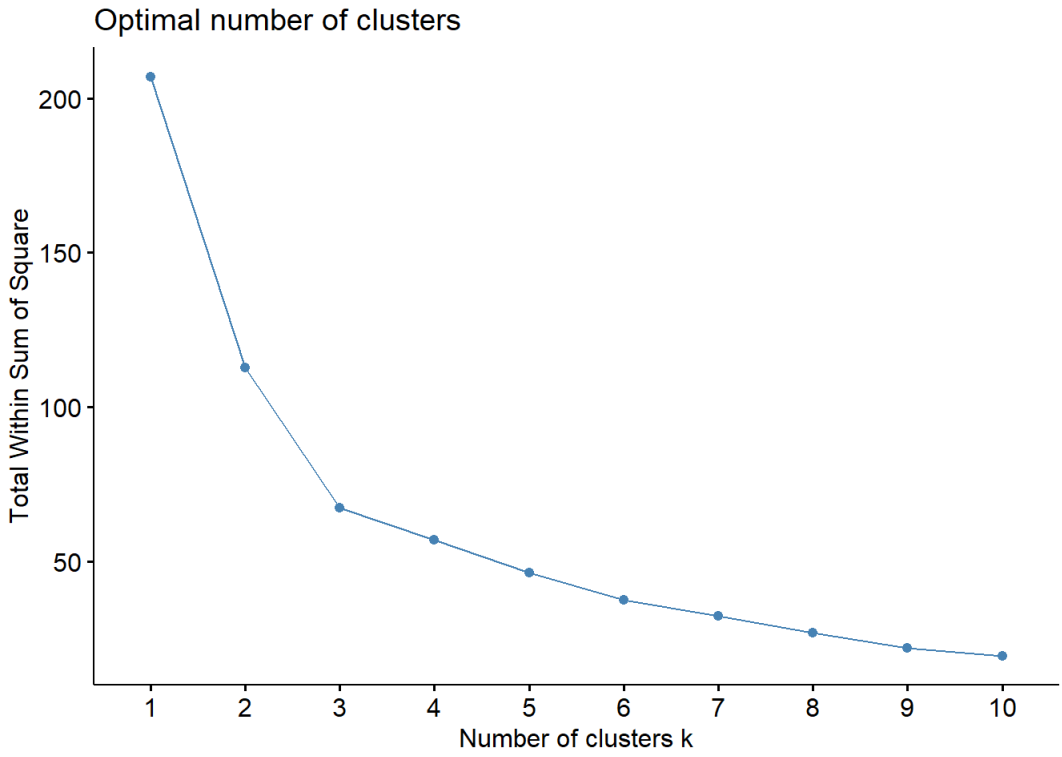
Dla metody k-średnich:

```
library(factoextra)
fviz_nbclust(dane_st, kmeans, method = "wss")
```



Dla metody PAM:

```
fviz_nbclust(dane_st, pam, method = "wss")
```



Widzimy, że dla metody k-średnich skłaniać można się ku liczbie klastrów równej 4, a dla metody PAM - 3. Sprawdzę więc obie możliwości i wtedy zadecyduję, która opcja jest lepsza i przy użyciu której metody. Wracając jednak do liczby klastrów - wyniki na podstawie wykresu wydają się mieć sens również w praktyce. W 3 klastrowym podziale, możemy się spodziewać grup odpowiadających pozycji zawodników na boisku - obrońców, pomocników i napastników. 4 klastry wiązałyby się prawdopodobnie z podzieleniem jednej z grup na 2 mniejsze - na podstawie intuicji i własnej orientacji w tym temacie spodziewałbym się podziału pomocników na tych o usposobieniu bardziej defensywnym oraz tych, o usposobieniu zdecydowanie bardziej ofensywnym.

2.1.2 Przeprowadzenie analizy i omówienie wyników

Za pomocą funkcji `kmeans` oraz `pam` dokonuję grupowania za pomocą tych metod. Ze względu na to, że metoda k-średnich nie znajduje minimum globalnego, a lokalne ustawiam liczbę wykonań procedury na 10 - powinno to zapewnić wybór najlepszego grupowania. Wyniki zapisuję, sortuję za pomocą funkcji `order` i wyświetlam: (analogicznie będę postępował przy podziale na 4 klastry)

3 klastry

```
kM <- kmeans(dane_st, centers=3,nstart = 10)
pM <- pam(dane_st, k=3)
wynik <- cbind(kM$cluster,pM$clustering)
colnames(wynik) <- c("k-means","PAM")
wynik <- wynik[order(wynik[,1]),]
wynik %>% kable() %>% kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive")) %>
% scroll_box(width = "100%", height = "250px")
```

	k-means	PAM
Andreas Christensen	1	2
Chris Smalling	1	2
Dejan Lovren	1	2
Eric Dier	1	2
Fernandinho	1	2
Michael Keane	1	2

Widzimy, że obie procedury pogrupowały zawodników tak samo. Skład poszczególnych grup nie różni się w obu algorytmach. Aby zobaczyć główne różnice w grupach obliczam średnie i odchylenia standardowe dla każdej z grup:

ŚREDNIE

kM2	Goals	Offsides	Shots	Assists	Passer per match	Big chances created	Crosses	Interceptions	Clearences
1	1.44	1.33	21.33	1.00	60.84	2.56	8.33	53.67	110.78
2	6.67	5.44	52.89	8.22	48.14	9.44	75.11	16.56	17.67
3	15.67	19.33	77.67	6.17	24.91	9.00	26.67	7.17	8.33

ODCHYLENIA

kM2	Goals	Offsides	Shots	Assists	Passer per match	Big chances created	Crosses	Interceptions	Clearences
1	1.88	1.22	11.60	1.00	15.90	2.24	9.07	14.95	58.80
2	3.32	3.32	11.19	3.70	23.39	5.08	28.50	8.38	20.03
3	3.39	3.93	14.09	2.79	5.49	4.34	18.26	2.48	6.56

Widzimy, że poszczególne grupy różnią się od siebie. Grupa pierwsza charakteryzuje się małą liczbą strzelonych bramek, oddanych strzałów, spalonych, asyst, stwarzanych okazji i dośrodkowań, natomiast dużą liczbą przechwyceń piłki i wybić. Wskazywało by to na grupę zawodników stricte defensywnych, szczególnie obrońców i defensywnych pomocników. Patrząc po nazwiskach przyporządkowanych piłkarzy - nie mam wątpliwości, że tak jest. Do tej grupy zaklasyfikowani zostali obrońcy i defensywni pomocnicy.
Grupa 2 posiada wszystkie ofensywne statystyki na poziomie znacznie wyższym niż grupa 1, a defensywne - na niższym. Podobnie grupa 3. Wskazuje to na fakt, że są to grupy zawodników ofensywnych. Czym się jednak różnią ?
Od razu w w oczy rzuca się większa liczba bramek, strzałów i spalonych w grupie 3. Widać również prawie 2x mniejsze wartości statystyk defensywnych (Interceptions, Clearences). Grupa 2 natomiast ma lepsze liczby pod względem asyst, podań wykonanych średnio w meczu, tworzonych sytuacji oraz dośrodkowań. Dzięki temu mogę z pewnością stwierdzić, że grupa 2 jest grupą zawodników zajmujących się w większym stopniu kreowaniem gry, a 3 - zdobywaniem bramek. Grupa 2 więc to pomocnicy ofensywni, a trzecia to napastnicy. Rzut oka na nazwiska w tych grupach pozwala ze stuprocentową pewnością stwierdzić poprawność moich przypuszczeń.
Podział na 3 grupy jest dobrym rozwiązaniem. Jednak na podstawie wykresów łokciowym rozsądnym rozwiązaniem będzie sprawdzić również podział na 4 klastry. Rzut oka na odchylenia w poszczególnych klastrach pozwala mi domyślać się, że podział na 4 klastry będzie zawierał w sobie niezmienione grupy 2 i 3, a także grupę 1 podzieloną na 2 inne. Skłaniają mnie do tego stosunkowo dużę wartości odchylenia standardowego dla zmiennych Interceptions i Clearences w grupie 1. Jak wyżej wspomniałem, do tej grupy zostali przyporządkowani zawodnicy występujący na pozycji obrońcy oraz defensywnego pomocnika. Prawdopodobnie więc, nastąpi podział grupy numer 1 na te dwie grupy.

4 klastry - wyniki

	k-means	PAM
Alexandre Lacazette	1	1
Gabriel Jesus	1	1
H.M.Son	1	1
Raheem Sterling	1	1
Romelu Lukaku	1	1
Sergio Aguero	1	1
David Silva	2	3
Delle Alli	2	3

Wyniki algorytmu PAM w 100% pokrywają się z moimi przewidywaniami. Inaczej sytuacja ma się jednak jeśli chodzi o wyniki algorytmu k-średnich. Przyjrzyjmy się więc bliżej średnim i odchyleniom w grupach. Najpierw te powstałe w wyniku zastosowania algorytmu k-średnich.

ŚREDNIE

kM4	Goals	Offsides	Shots	Assists	Passer per match	Big chances created	Crosses	Interceptions	Clearences
1	15.67	19.33	77.67	6.17	24.91	9.00	26.67	7.17	8.33
2	7.00	6.40	51.00	10.60	51.97	13.40	73.40	15.40	8.00
3	6.25	4.25	55.25	5.25	43.35	4.50	77.25	18.00	29.75
4	1.44	1.33	21.33	1.00	60.84	2.56	8.33	53.67	110.78

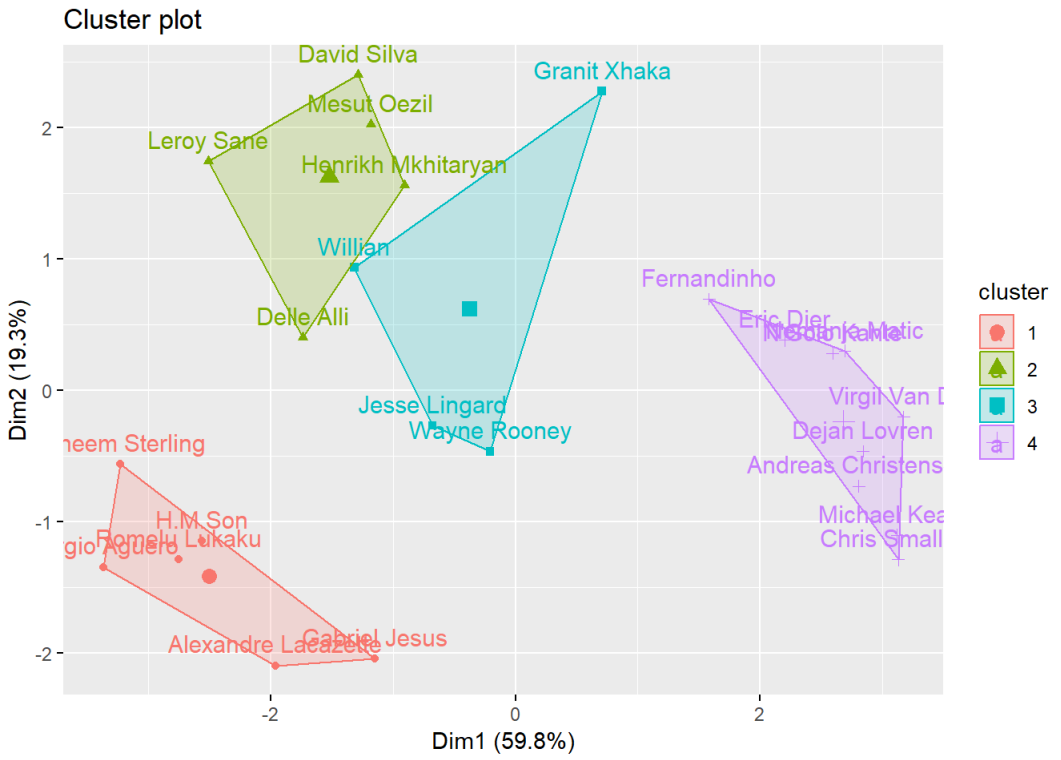
ODCHYLENIA

kM4	Goals	Offsides	Shots	Assists	Passer per match	Big chances created	Crosses	Interceptions	Clearences
1	3.39	3.93	14.09	2.79	5.49	4.34	18.26	2.48	6.56
2	3.24	3.36	13.58	2.70	23.13	1.82	32.29	3.05	5.61
3	3.86	3.30	8.62	2.36	26.26	2.38	27.66	13.04	26.03
4	1.88	1.22	11.60	1.00	15.90	2.24	9.07	14.95	58.80

Na podstawie różnic między grupa 2 i 3, można by stwierdzić, że zawodnicy z grupy 3 mają słabsze statystyki w ofensywie, bo grają bliżej swojej bramki. Wskazywałyby na to istotnie większa liczba wybić piłki, mniejsza liczba spalonych i zdecydowanie mniejsza liczba wykreowanych sytuacji. Mogłoby to sugerować, że z grupy ofensywnych pomocników powstały dwie nowe - jedna składająca się z zawodników występujących na pozycji '8' - rozgrywających, mających więcej zadań w defensywie i wyprowadzeniu piłki, mniej kreatywnych. Druga grupa zawierałaby wtedy skrzydłowych lub zawodników grający na tzw. dziesiątce - za plecami zawodnika, kreatywnych zawodników potrafiących 1 zagraniem stworzyć sytuację koledze z drużyny. Jednak rzut oka na nazwiska nie napawa mnie optymizmem w związku z takim podziałem - w grupie 3 znaleźli się Lingard, Rooney i Willian. Zawodnicy, którzy na boisku ustawiani są zdecydowanie bardziej tuż za plecami napastnika/na skrzydle. Jedynie Xhaka pasowałby do pozycji numer 8. Może to wynikać, ze słabszych liczb w ofensywie jakie zanotowali Ci zawodnicy w tamtym sezonie czy też faktu że np. Rooney występuje w Evertonie - prawdopodobnie najsłabszym klubie spośród tych, których zawodnicy znaleźli się w moich danych.

Dla lepszego zobrazowania wyniku i późniejszej różnicy miedzy wynikami algorytmu k-średnich i algorytmu PAM, dołączam wykres klastrów stworzony za pomocą funkcji `fviz_cluster` z wspomnianego wcześniej pakietu `factoextra`. Funkcja ta prezentuje grupy powstałe w wyniku danego podziału na dwuwymiarowym wykresie. Automatycznie przestałca ona zmienne wielowymiarowe na dwuwymiarowe, tak aby można było je łatwo i czytelnie przedstawić na wykresie. Są one liniową kombinacją zmiennych podstawowych. Nowe zmienne mają za zadanie w jak największym stopniu zawierać informacje, które niosły ze sobą wcześniejsze zmienne. Procenty zawarte na osiach informują w jakim stopniu nowe zmienne wyjaśniają/opisują informacje zawarte w pierwotnych zmiennych. Proces ten zwany jest 'analizą głównych składowych'. Nie będe sie jednak w niego bardziej zagłębiał, ponieważ w tym wypadku głównym motywem do użycia tej funkcji była chęć przedstawienia grup na wykresie i temu ma służyć jej zastosowanie. Wykresie, który prezentuje się następująco:

```
fviz_cluster(kM, data = dane_st)
```



Biorąc pod uwagę aspekty, o których wspomniałem wcześniej oraz wykres na którym widać, że grupy 2 i 3 są bardzo blisko siebie - odrzucam ten podział. Spójrzmy jednak na podział, który proponuje metoda PAM.

ŚREDNIE

pM4	Goals	Offsides	Shots	Assists	Passer per match	Big chances created	Crosses	Interceptions	Clearences
1	15.67	19.33	77.67	6.17	24.91	9.00	26.67	7.17	8.33
2	1.20	1.60	14.00	0.40	51.60	1.40	1.80	47.80	153.60
3	6.67	5.44	52.89	8.22	48.14	9.44	75.11	16.56	17.67
4	1.75	1.00	30.50	1.75	72.40	4.00	16.50	61.00	57.25

ODCHYLENIA

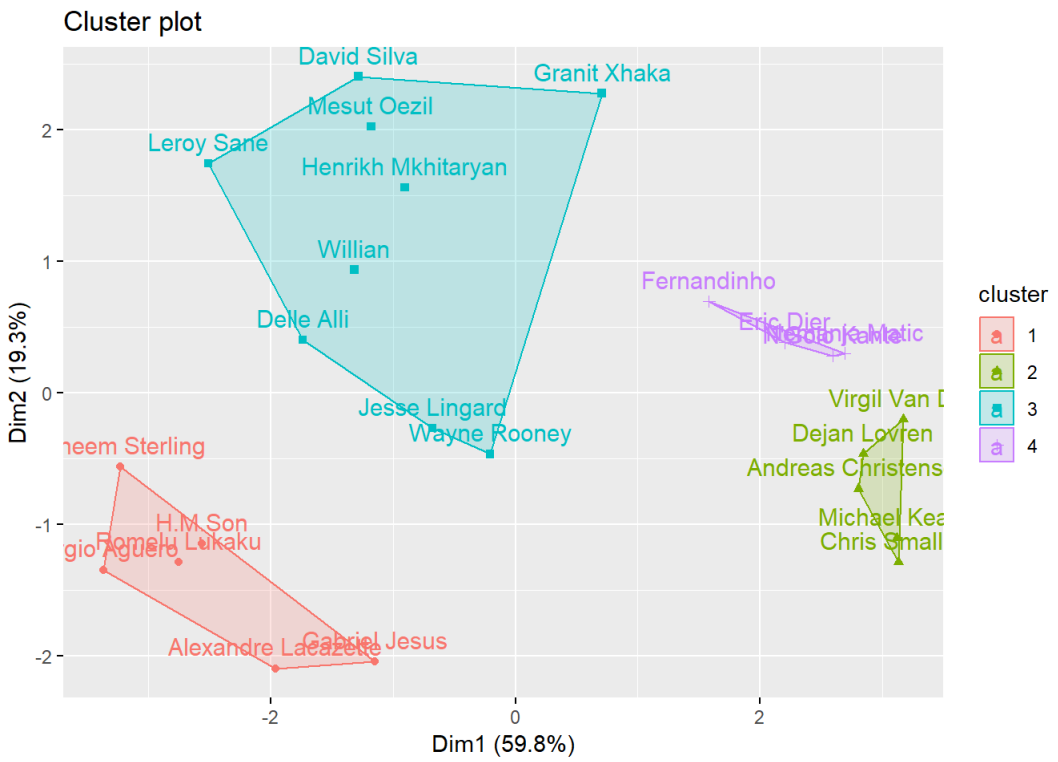
pM4	Goals	Offsides	Shots	Assists	Passer per match	Big chances created	Crosses	Interceptions	Clearences
1	3.39	3.93	14.09	2.79	5.49	4.34	18.26	2.48	6.56
2	1.79	1.52	1.58	0.55	13.32	1.67	2.05	10.62	36.68
3	3.32	3.32	11.19	3.70	23.39	5.08	28.50	8.38	20.03
4	2.22	0.82	12.40	0.96	10.82	2.16	7.33	17.80	23.46

Widzimy, że grupa 1 to grupa napastników z podziału na 3 klastry. Grupa 3 - pomocnicy. Podzielona została grupa zawodników defensywnych. Grupa oznaczona jako 2 charakteryzuje się zdecydowanie największą liczbą wybić piłki oddalających zagrożenie od własnej bramki - co jest charakterystyczną cechą środkowych obrońców. Bardzo mała liczba bramek, asysty, strzałów, spalonych i wykreowanych sytuacji tylko potwierdza to przypuszczenie. Czym jednak różni się grupa 4, skoro wcześniej te grupy tworzyły jedną większą?

Zdecydowanie mniejsza liczba wybić sugerować może, że grają oni wyżej na boisku. Wyższa liczba przechwytów również. Podobnie jak niewiele wyższe statystyki liczby bramek czy asyst. Zdecydowanie więcej tworzą oni jednak sytuacji kolegom, wykonują zdecydowanie (8x!) więcej dośrodkowań, a także oddają 2 razy więcej strzałów. Te statystyki potwierdzają moje przypuszczenia i grupa ta prawdopodobnie będzie grupą środkowych defensywnych pomocników, zawodników występujących na pozycji numer ‘6’. Rzut oka na zawodników przyporządkowanych do tej grupy potwierdza tę tezę - Kante, Matic, Fernanhinho i Dier to jedni z najlepszych zawodników grających na ‘szóstce’ w Premier League.

Żeby zobrazować wyniki dla algorytmu PAM przedstawiam analogiczny wykres jak w przypadku metody k-średnich:

```
fviz_cluster(pM,dane_st)
```



Ten wykres i grupy na nim przedstawione sprawia już lepsze wrażenie niż grupy z podziału metodą k-średnich. Różnice między grupami wydają się być większa. Co więcej, obserwacje z grupy 4 są do siebie bardzo zbliżone, co jest kolejnym plusem tego podziału.
Biorąc pod uwagę przeprowadzone grupowania i ich wyniki, za najlepszy spośród grupowania podziałowego uznaję ten ostatni - otrzymany przy zastosowaniu algorytmu PAM przy podziale na 4 klastry.

2.2 Grupowanie hierarchiczne

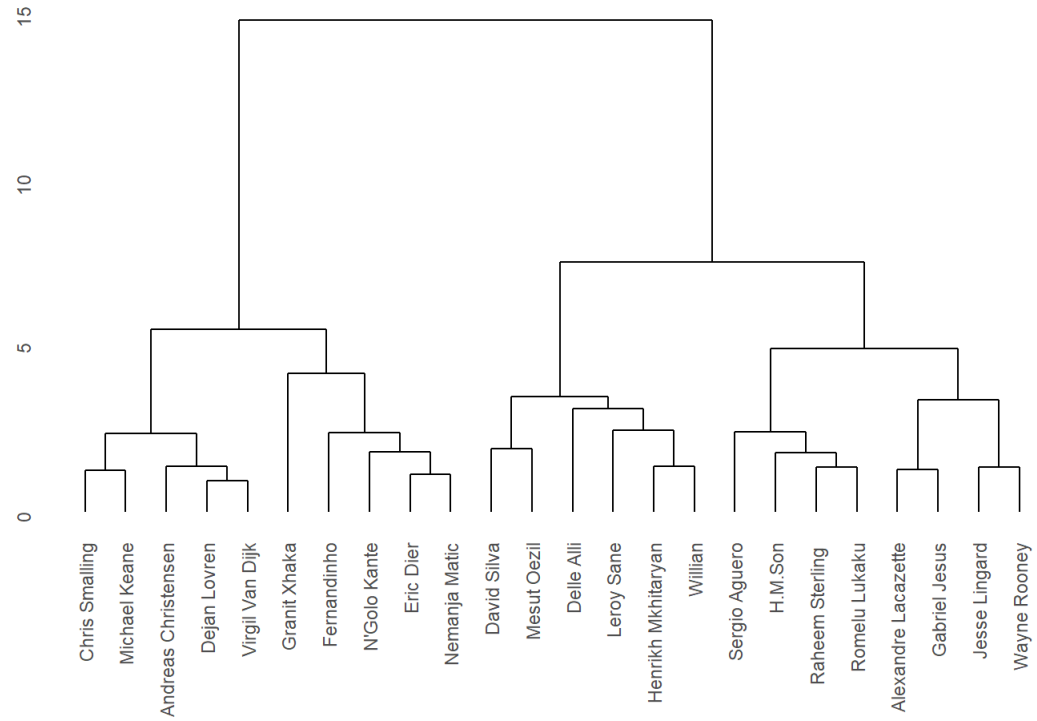
Polega na łączeniu elementów w coraz większe grupy (na podstawie funkcji odległości) do momentu, aż uzyskamy skupienie zawierające wszystkie elementy. Do przeprowadzenia tego badania można użyć kilku funkcji odległości. W moim badaniu pojawią się metody centroidalna, mediany oraz Warda. Metod najbliższego i najdalszego sąsiada nie sprawdzam, ponieważ w praktyce nie są one stosowane ze względu na swoje wady.

2.2.1 Wyniki

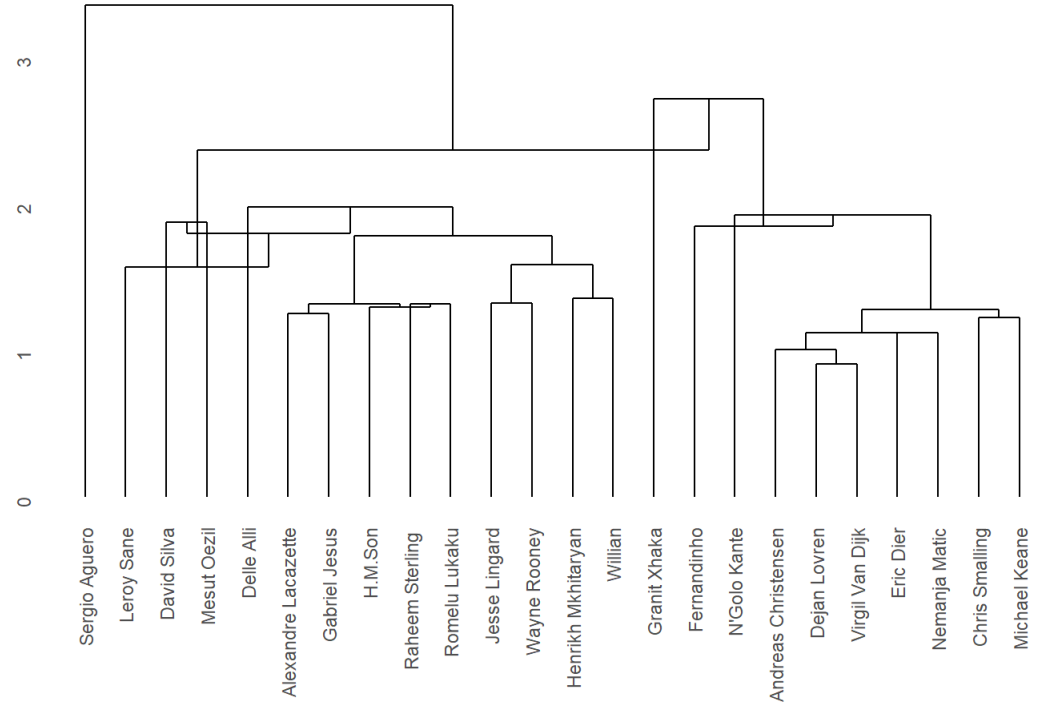
Wyniki poszczególnych metod przedstawię za pomocą dendrogramów, stworzonych z użyciem funkcji `ggdendrogram` i pakietu `ggdendro`:

Metoda Warda

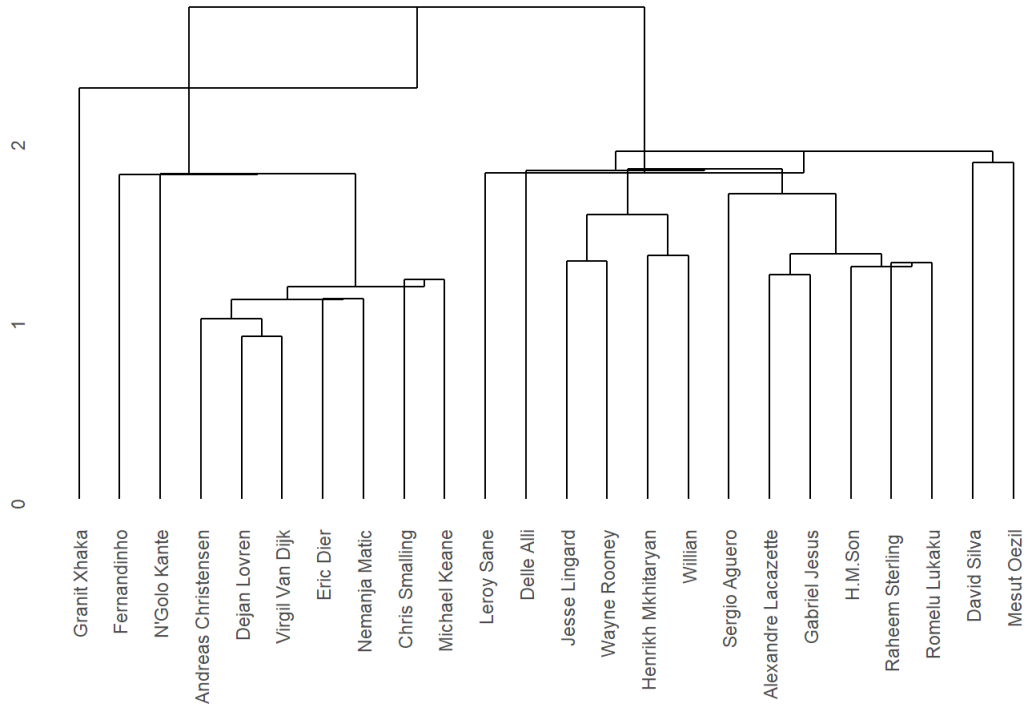
```
library(ggdendro)
odl <- dist(dane_st)
podzialW <- hclust(odl, method = "ward.D2")
ggdendrogram(podzialW)
```



Metoda mediany



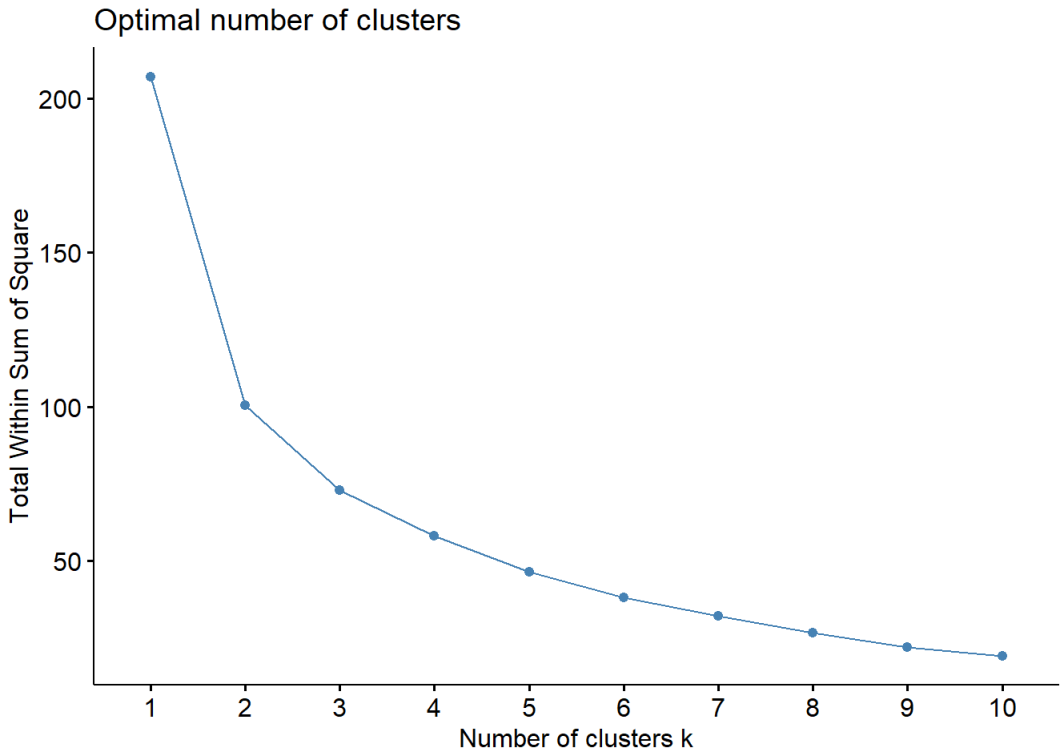
Metoda Centroidalna



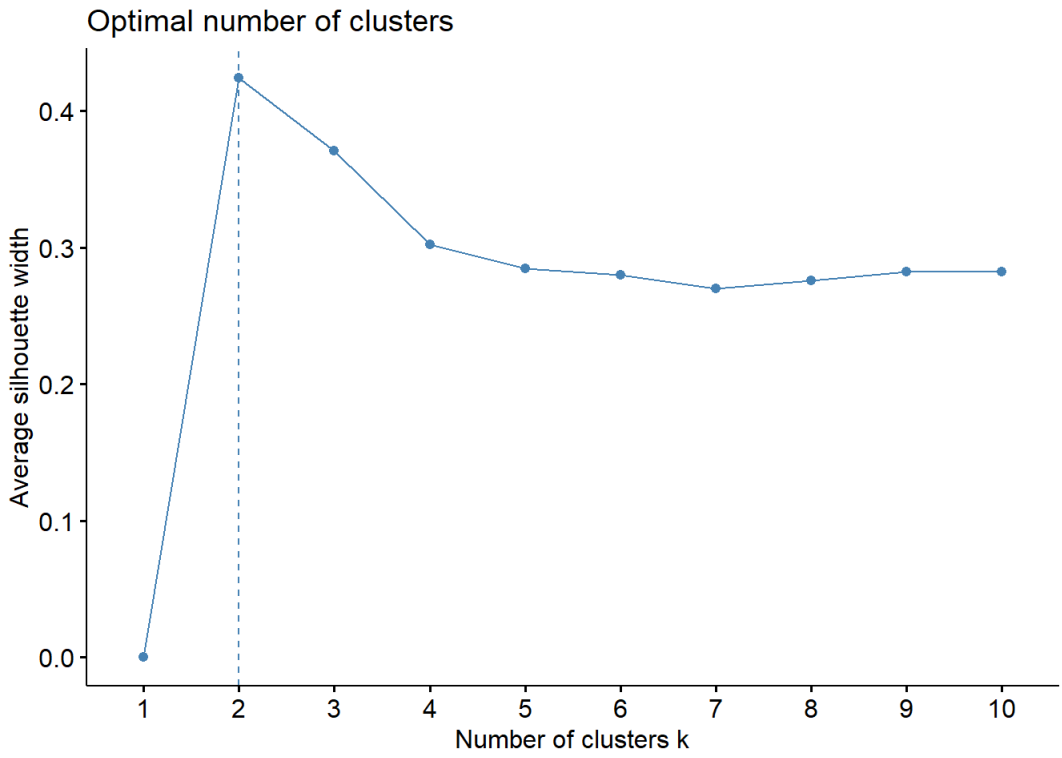
Zdecydowanie najbardziej czytelny jest dendrogram dla metody Warda. Ponadto, patrząc na to jak dobierani są do siebie poszczególnie piłkarze, metoda Warda również i pod tym względem wydaje się sensowniejsza niż pozostałe. Tak więc to dla niej będę wykonywał kolejne etapy badania.

2.2.2 Wybór liczby klastrów

Do wyboru liczby klastrów dla grupowania hierarchicznego metodą Warda posłużę się ponownie metodą łokciową - analogicznie jak w wypadku metod k-średnich i PAM.



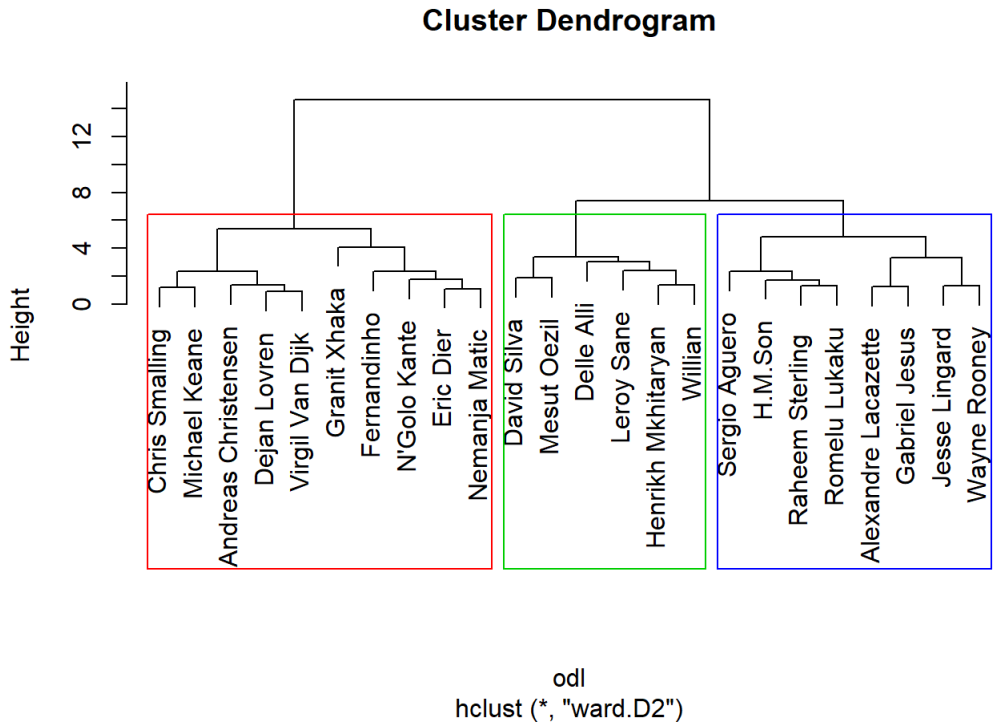
Metoda nie wskazuje jednoznacznie na jakiś wynik. Można się zastanawiać na 2 / 3 grupami, a także 4 jak w najlepszym grupowaniu podziałowym. Sprawdzę jeszcze jaką liczbę klastrów podpowiada metoda ‘silhouette’. Określa ona jak dobrze każdy z obiektów pasuje do klastra, do którego został przypisany. Za optymalną liczbę skupień przyjmuje się tą, która przyjmuje wartość maksymalną.



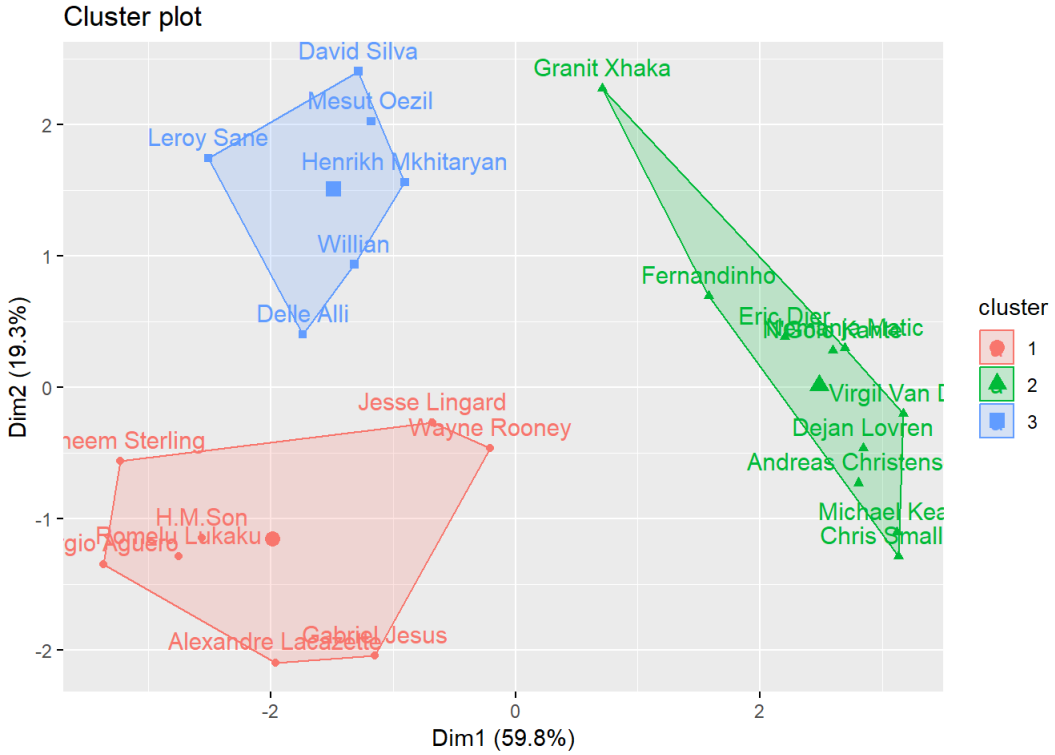
Z wykresu wynika, że wg tej metody najlepszym rozwiązaniem byłoby zastosowanie 2-klastrowego podziału. Jest to wg mnie jednak zbyt obszerny podział. Z tego względu decyduję się sprawdzić jak przedstawiają się trzy- i cztero-klastrowe podziały metodą Warda.

3 klastry

Wydzielenie klastrów na dendrogramie:



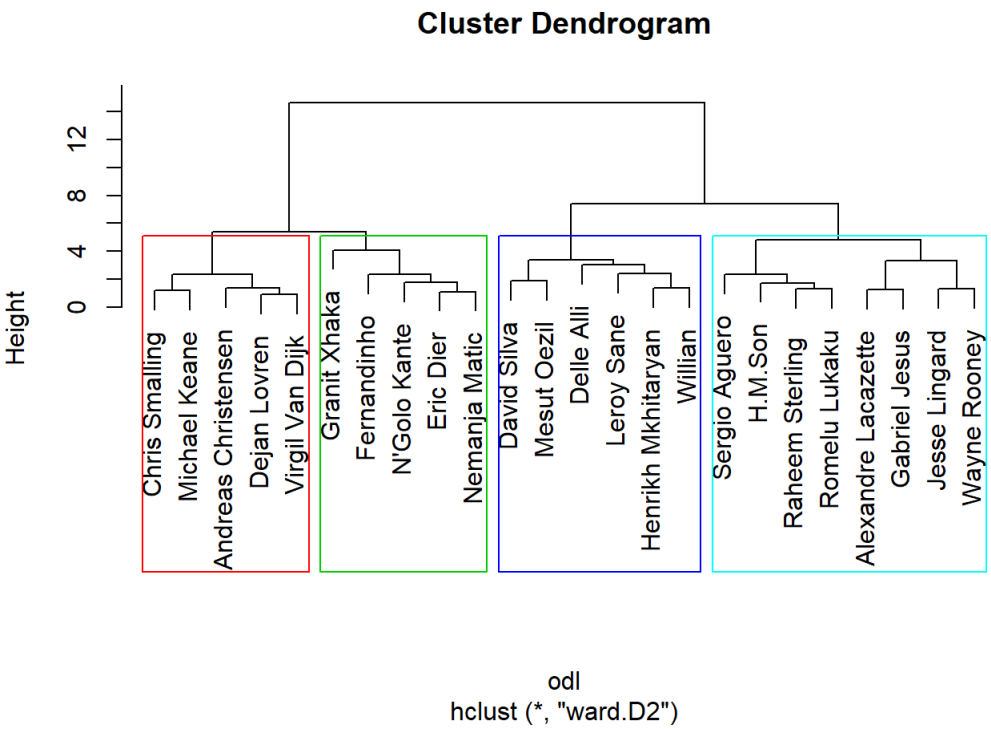
Przedstawienie wyniku na wykresie:



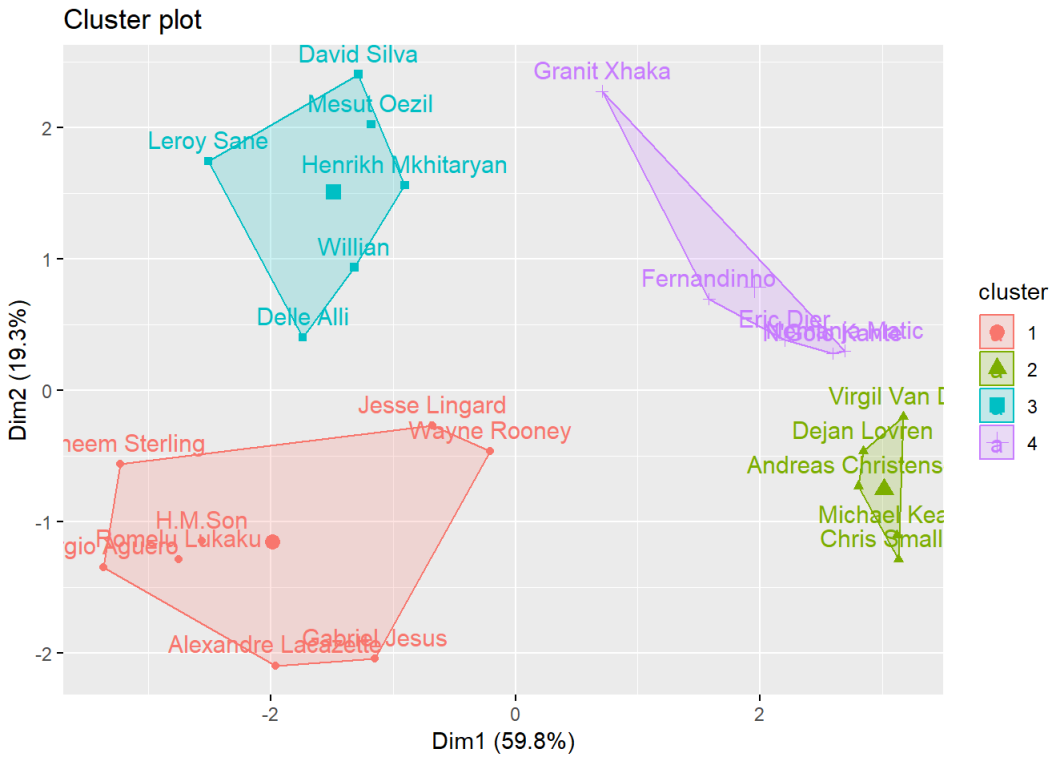
Grupa 1 przedstawia napastników, do grona których włączono jednak tym razem 2 kolejnych zawodników (Rooney'a i Lingarda) z czym nie do końca można się zgodzić. Również sam fakt, że Ci zawodnicy są bardzo blisko grupy 3 na wykresie jest niepokojący. Zobaczmy jak przedstawiają się wyniki dla 4 klastrów.

4 klastry

Wydzielenie klastrów na dendrogramie:



Przedstawienie wyniku na wykresie:



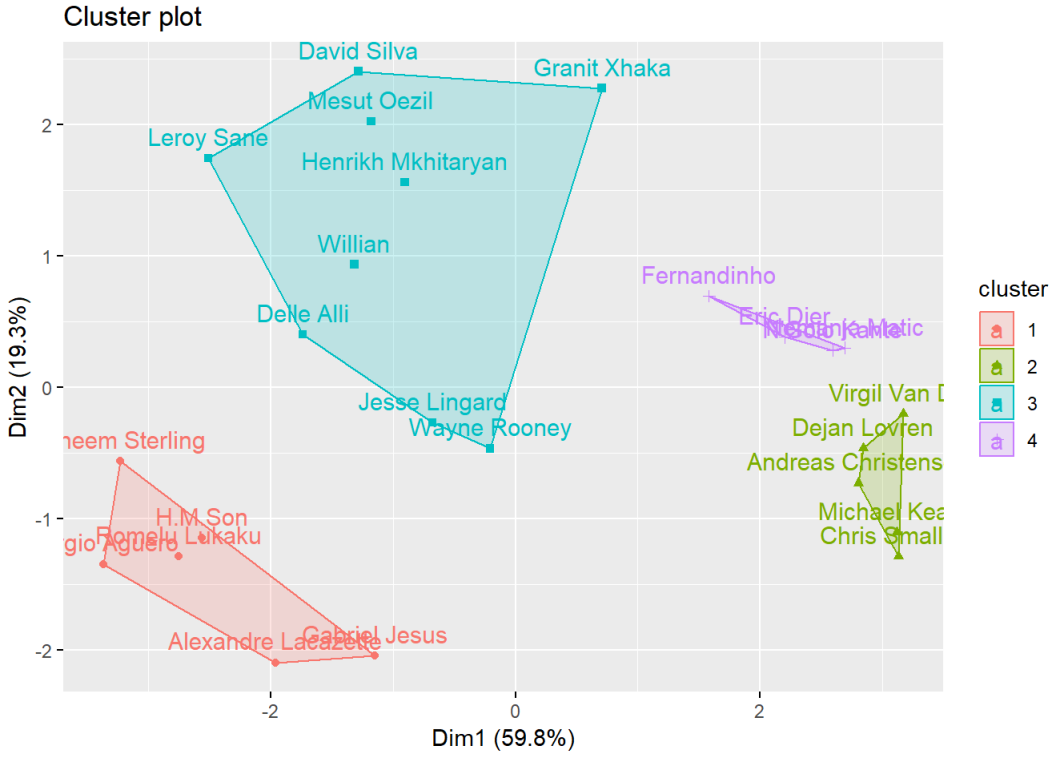
Nastąpił podział grupy 2 na 2 części. Analogicznie jak w grupowaniu podziałowym metodą PAM - defensywni pomocnicy odłączyli się od obrońców. Wątpliwości dotyczące zaklasyfikowania 2 zawodnikóW o których wspomniałem nadal pozostały.
Zobaczę jak przedstawiają się statystyki średniej dla tego podziału, który ponownie uważam za lepszy niż 3 klastrowy:

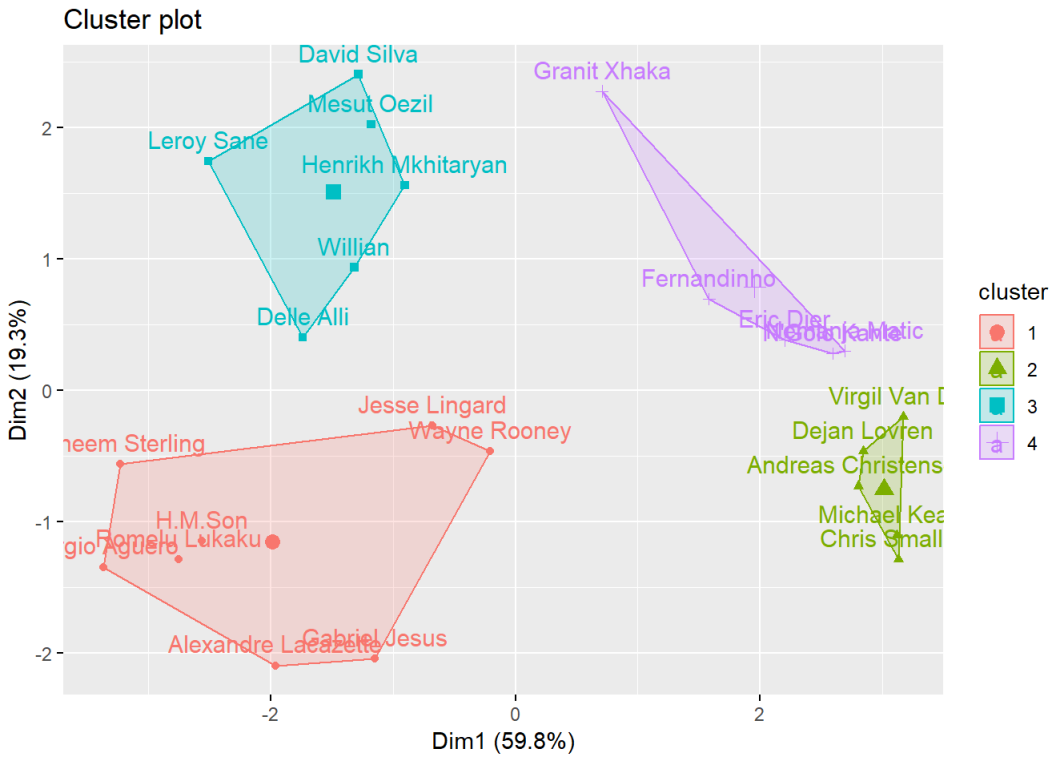
pW	Goals	Offsides	Shots	Assists	Passer per match	Big chances created	Crosses	Interceptions	Clearences
1	14.00	16.00	70.88	5.5	26.90	7.62	33.50	8.5	12.00
2	1.20	1.60	14.00	0.4	51.60	1.40	1.80	47.8	153.60
3	6.83	6.17	51.50	10.0	47.59	12.50	77.67	14.5	8.17
4	1.60	0.80	37.60	2.8	74.32	3.80	33.60	56.2	58.60

Jest to podział, który z pewnością się wybroni - ma sens, co potwierdzają powyższe średnie. Odpowiednie grupy nadal mają poszczególne statystyki na odpowiednim poziomie i są od siebie różne, na tyle, że jasno możemy określić, która grupa zawiera jakich piłkarzy i czym poszczególne grupy się od siebie różnią.

2.3 Podsumowanie i wybór najlepszego wyniku.

Podsumowując: za najlepszy wynik grupowania podziałowego uznano podział 4 klastrowy za pomocą metody PAM. Jeśli chodzi o grupowanie hierarchiczne, tu również najlepszy okazał się wg mnie 4 klastrowy podział, za pomocą metody Warda. Przedstawiają się one następująco:





Powyższe wyniki nie różnią się znacząco od siebie i oba zdecydowanie można uznać za satysfakcjonujące. Obie metody sa najczęściej stosowanymi w praktyce, tak więc fakt, że to właśnie te podziały uznałem za najlepsze nie powinien szczególnie zaskakiwać. Jako lepszy wybrałbym grupowanie podziałowe metodą PAM, ze względu na lepsze zaklasyfikowanie Rooney'a i Lingarda - w moich oczach są to zawodnicy, których profil i pozycja na boisku w tamtym sezonie bardziej odpowiadała grupie ofensywnych pomocników. Jednak plusem metody Warda jest w tym wypadku klasyfikacja Granita Xhaki jako defensywnego pomocnika - co jest moim zdaniem zgodne z rzeczywistością. Przekładam jednak większe odległości między grupami i 2 lepiej zaklasyfikowanych zawodników nad 1, **więc za najlepszy podział, uznaję grupowanie podziałowe metodą PAM.**
Zaznaczam przy tym jednak, że zarówno zaklasyfikowanie Rooney'a i Lingarda jako napastników, jak i Xhaki jako ofensywnego pomocnika, nie jest błędem i w rzeczywistości zawodnicy Ci, ze względu na swoje wszechtronne umiejętności mogą występować (i występują) na obu rozważanych pozycjach, a zwolennicy klasyfikowania ich inaczej niż ja z pewnością zaleźliby swoje argumenty.

3 Funkcja tworząca ranking obiektów

Drugim elementem mojego projektu będzie funkcja, która będzie tworzyć renking obiektów za pomocą metody sum standaryzowanych. Przedstawia się ona następująco:

```
mss <- function(dane, rodzaj_zm, wagi = NULL){
  n_row <- nrow(dane)
  n_col <- ncol(dane)

  #zamiana zmiennych na stymulanty
  for (i in 1:n_col){
    if (rodzaj_zm[i] == "D") {dane[,i] <- -dane[,i]}
    else if (rodzaj_zm[i] != "S") {
      dane[,i] <- -abs(dane[,i] - as.double(rodzaj_zm[i]))
    }
  }

  #standaryzacja zmiennych
  dane <- scale(dane)

  if (!is.null(wagi)) {
    for (i in 1:n_col){
      dane[,i] <- dane[,i]*wagi[i]
    }
  }

  #obliczam wektor będący sumą po zmiennych
  s_rang <- NULL
  for (i in 1:n_row){
    s_rang <- c(s_rang,sum(dane[i,1:n_col]))
  }

  #zmienne pomocnicze do obleczenia wskaźnika ciut szybciej
  minn <- min(s_rang)
  pom <- s_rang - minn
  mianownik <- max(pom)

  #obliczam wskaźnik i zapisuje go do zmiennej wynik
  wynikk <- NULL
  for (i in 1:n_row){
    a <- (s_rang[i]-minn)/mianownik
    wynikk <- rbind(wynikk,a)
  }

  #nadaje nazwy wierszy takie jak w danych wejściowych
  rownames(wynikk) <- rownames(dane)

  #sortowanie od najlepszego obiektu
  wynikk <- wynikk[order(wynikk[,1], decreasing = TRUE),]
}
```

Argumentami funkcji są dane, dla których chcemy stworzyć ranking (nie zestandaryzowane) oraz wektor rodzaj_zm - ma on przyjmować wartości “S” jeśli zmienna jest stymulantą, “D” jeśli jest destymulantą oraz podawać konkretną wartość jesli zmienna jest nominantą. Odpowiednie miejsce w wektorze ma odpowiadać odpowiedniej kolumnie w danych tj. 1 miejsce w wektorze odpowieda rodzajowi zmiennej zawartej w 1 kolumnie, 10 miejsce w wektorze odpowiada 10tej kolumnie itd. Do funkcji można również wprowadzić opcjonalnie trzeci argument, czyli system wag - wektor, zawierający wagi poszczególnych zmiennych, działający na dokładnie takiej samej zasadzie jak wektor zawierający informacje o rodzajach zmiennych.