Pozyskiwanie Wiedzy
Projekt - Część II

Justyna Domańska i Julia Lenczewska

31 stycznia 2020

Spis treści

1 Cel raportu **2**

2 Analiza skupień **2**
2.1 k-means . . . 2
2.1.1 Dla danych przeskalowanych . . . 2
2.1.2 Dla danych bez skalowania . . . 11
2.2 PAM . . . 14
2.3 AGNES . . . 20
2.3.1 average linkage . . . 20
2.3.2 single linkage . . . 23
2.3.3 complete linkage . . . 26
2.4 DIANA . . . 28

3 Redukcja wymiaru **33**
3.1 PCA . . . 33
3.2 MDS . . . 33

4 Klasyfikacja po MDS **35**

5 Klasteryzacja po MDS **38**
5.1 k-means . . . 38
5.2 PAM . . . 40
5.3 AGNES . . . 41
5.3.1 average linkage . . . 41
5.3.2 single linkage . . . 42
5.3.3 complete linkage . . . 44
5.4 DIANA . . . 45
5.5 Ocena . . . 47

6 Podsumowanie **51**

1 Cel raportu

W tym raporcie zajmiemy się dalszą analizą zbioru danych *German Credit*. Najpierw przeprowadzimy analizę skupień - przeanalizujemy algorytmy takie jak k-means, PAM, AGNES i DIANA. W kolejnym etapie zastosujemy redukcję wymiaru i porównamy wyniki analizy skupień oraz klasyfikacji, otrzymane dla danych przed i po zastosowaniu redukcji.

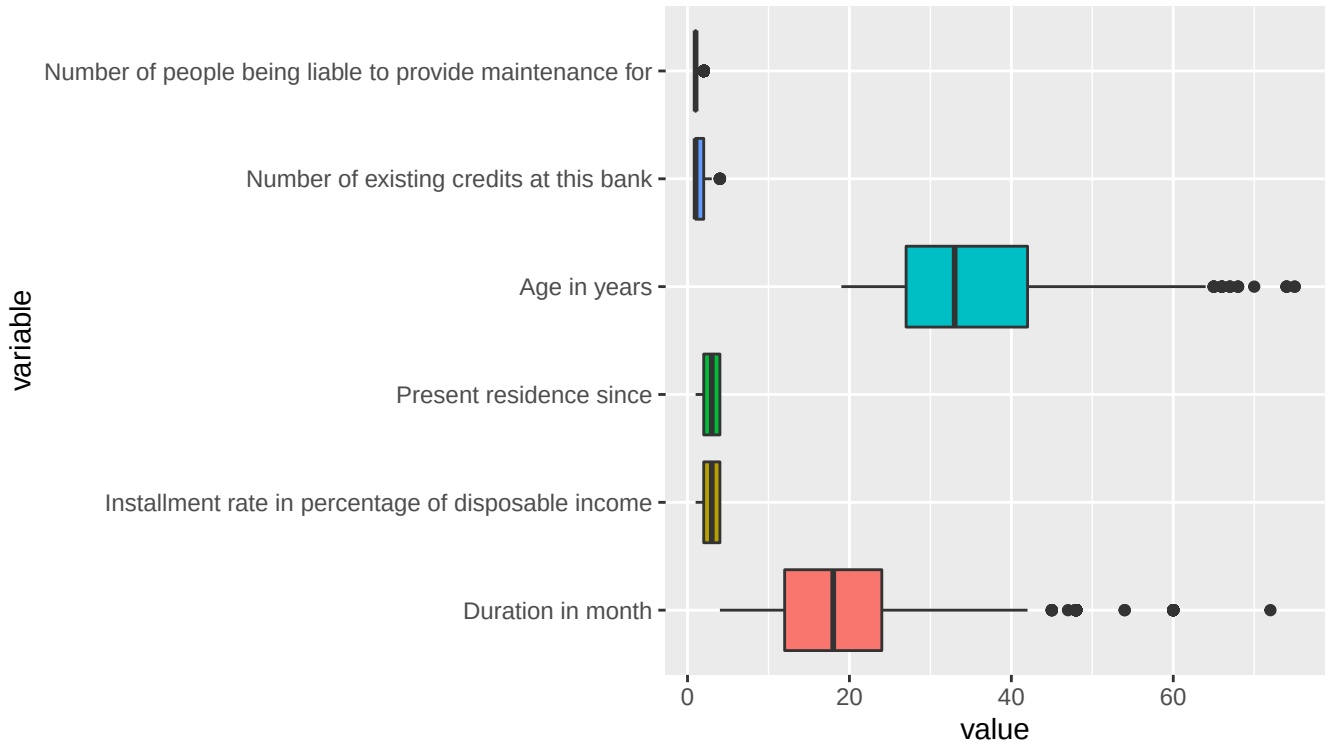
2 Analiza skupień

2.1 k-means

Jednym z głównych i najczęściej stosowanych algorytmów analizy skupień jest algorytm k-means. Nie możemy go jednak zastosować do wszystkich danych jakimi dysponujemy, gdyż algorytmu tego nie stosuje się do danych jakościowych lub mieszanych typów. W związku z tym do przeprowadzenia analizy bazującej na algorytmie k-średnich wykorzystamy jedynie zmienne numeryczne. Wybierzemy te same zmienne, które w pierwszej części raportu użyte zostały m.in. przy algorytmie k-NN. Będą to: Duration in month, Credit amount, Installment rate in percentage of disposable income, Present residence since, Age in years, Number of existing credits at this bank, Number of people being liable to provide maintenance for.

Ważnym elementem przy stosowaniu algorytmu k-means, jest wybór optymalnej wartości k. Porównamy wartości wskaźników oceniających jakość grupowania dla różnej liczby klastrów.

2.1.1 Dla danych przeskalowanych



Rysunek 1: Box-ploty dla zmiennych numerycznych z pominięciem Credit amount

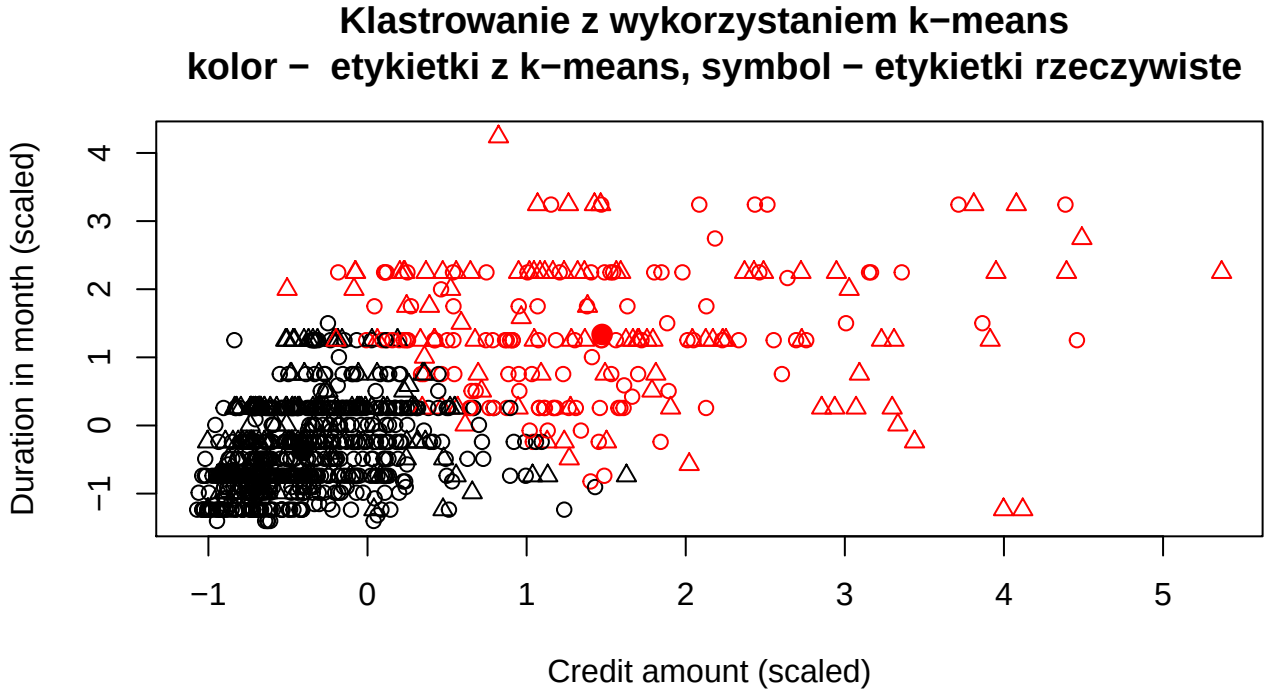
Na Rysunku 1 widzimy, że zakresy wartości jakie przyjmują zmienne znacząco się różnią. Zauważmy, że rysunek nie zawiera box-plotu dla zmiennej Credit amount. Maksymalna kwota kredytu wynosi bowiem 18424 marek

niemieckich, zatem umieszczenie jej na rysunku tylko zamazałoby obraz.
W związku z powyższym przed zastosowaniem algorytmu k-średnich przeskalujemy dane wykorzystując w tym celu funkcję *scale*. Jako, że znamy liczbę grup, na które w rzeczywistości podzielone są nasze dane, rozpocznijmy od wybrania $k = 2$ i sprawdźmy co w takim wypadku otrzymujemy. Porównamy wartości Total Within Sum of Squares dla *nstart*=1 , 10 oraz 20.

nstart=1	nstart=10	nstart=20
6013.39	5859.47	5859.47

Tabela 1: Total Within Sum of Squares dla różnych wartości *nstart*

W Tabeli 1 widzimy, że dla domyślnej wartości parametru nstart otrzymujemy wyższą wartość Total Within Sum of Squares niż dla np. 10. Dla *nstart* = 10 i 20 otrzymane wyniki są takie same.
Zilustrujemy teraz wyniki analizy skupień w zestawieniu z rzeczywistymi etykietkami jedynie dla dwóch wybranych zmiennych. Wybierzemy zmienne Credit amount oraz Duration in month. Wyniki są przedstawione na Rysunku 2.

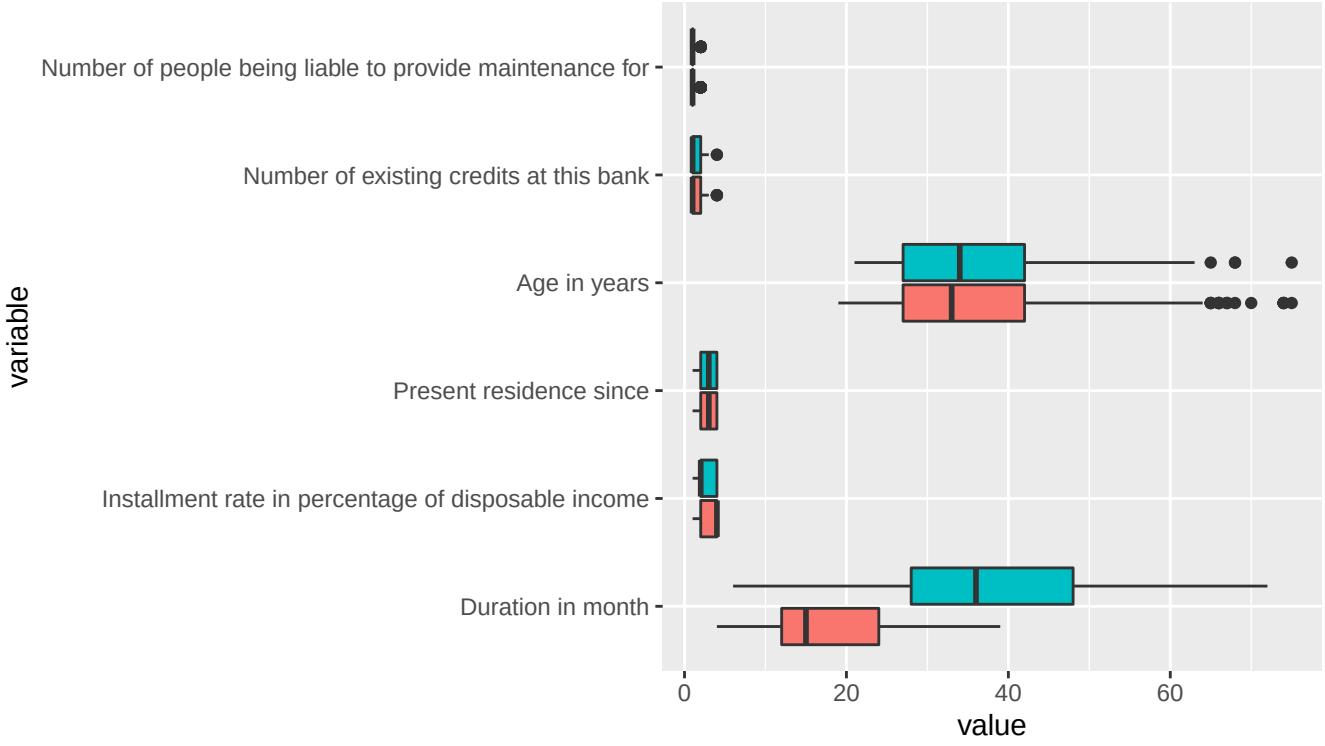


Rysunek 2: Porównanie etykietek po zastosowaniu k-means dla k = 2

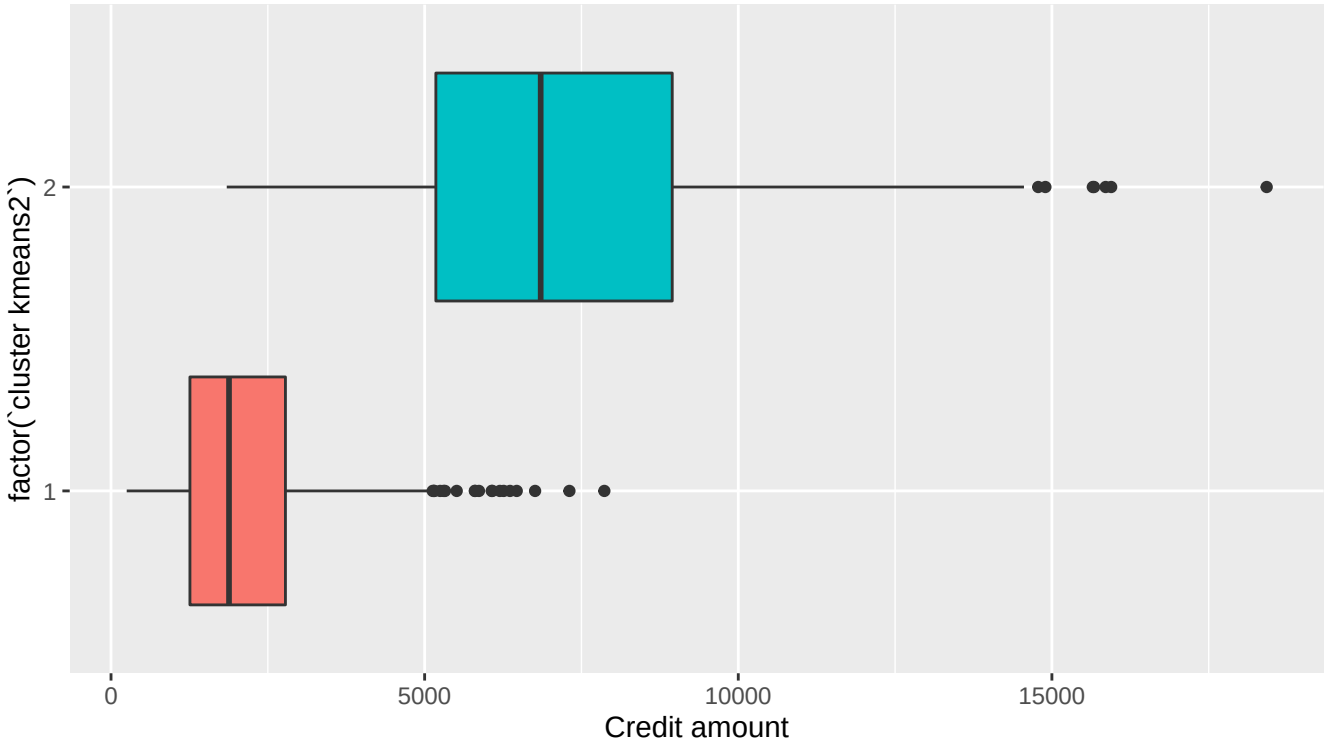
Na Rysunku 2 w jednym klastrze znajdują się obserwacje z wysokimi wartościami zmiennych Credit amount oraz Duration in month, zaś w drugim z niskimi wartościami wymienionych zmiennych.
Przeanalizujemy teraz wartości zmiennych w poszczególnych klastrach. Tabela 2 zawiera średnie wartości zmiennych, zaś Rysunki 3 i 4 przedstawiają ich box-ploty.

Cluster	Duration in month	Credit amount	Installment rate in percentage of disposable income	Present residence since	Age in years	Number of existing credits at this bank	Number of people being liable to provide maintenance for
1	16.44	2117.93	3.07	2.83	35.43	1.38	1.15
2	37.01	7432.81	2.62	2.91	35.96	1.49	1.19

Tabela 2: Średnie wartości zmiennych w klastrach



Rysunek 3: Box-ploty zmiennych numerycznych z podziałem na klastry



Rysunek 4: Box-plot zmiennej Credit amount z podziałem na klastry

Z Tabeli 2 możemy odczytać, że do pierwszego klastra zaliczono osoby, które brały większe kredyty, rozłożone na dłuższy okres czasu. Średni czas spłacania kredytu w pierwszym klastrze wynosi 16.44 miesiąca, natomiast w drugim jest ponad dwukrotnie większy i wynosi 37.01. Podobnie, średnie kwoty kredytu w drugim klastrze są znacznie większe niż te w pierwszym (wynoszą one odpowiednio 7432.81 i 2117.93). Takie grupowanie sugeruje, że decyzja o przyznaniu kredytu zależy od tego czy bierze się mały kredyt na krótki okres czasu, czy duży kredyt na długi. Także na Rysunkach 3 oraz 4 widzimy, że obserwacje zostały przydzielone do klastrów głównie na podstawie czasu trwania kredytu oraz jego kwoty.
Zobaczymy, jak porównują się etykietki rzeczywiste i przypisane przez algorytm k-means. Ich zestawienie znajduje się w Tabeli 3.

	1	2
1	120	97
2	580	203

Tabela 3: Macierz kontyngencji - k-means

Korzystając z funkcji *matchClasses* uzyskujemy procentową zgodność klas rzeczywistych i wskazanych przez algorytm k-means.

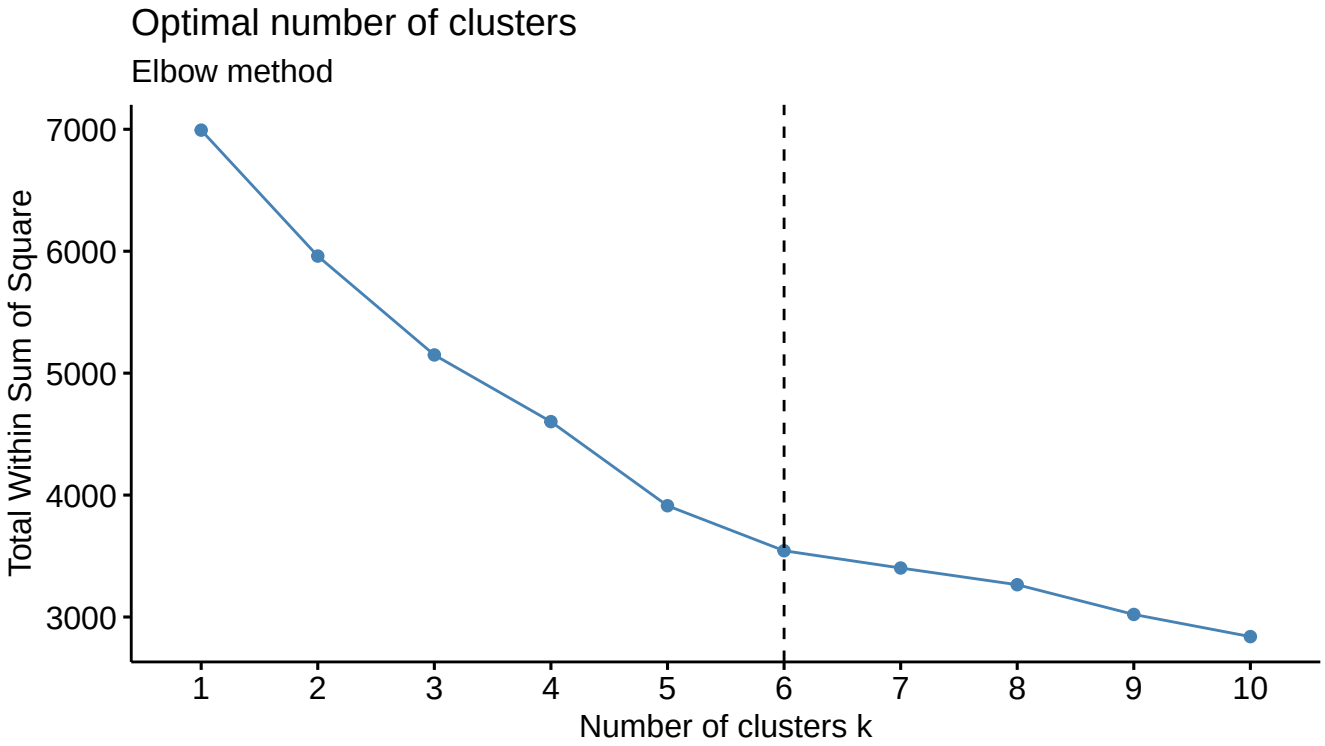
```
## Direct agreement: 1 of 2 pairs
## Iterations for permutation matching: 1
## Cases in matched pairs: 67.7 %
## 1 2
## 2 1
```

Otrzymana wartość jest dość wysoka i wynosi 67,7%.
Kolejnym etapem analizy będzie porównanie wyników algorytmu k-means dla różnych wartości k. Optymalną

liczbę klastrów spróbujemy wybrać wykorzystując m.in. funkcję *NbClust* (szczegółowe informacje można znaleźć a artykule *NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set*, Journal of Statistical Software, Volume 61, Issue 6). Wyliczanych jest z jej użyciem 26 różnych wskaźników, m.in. Silhouette, Dunn, Gap Statistic. Skorzystamy także z tzw. "elbow method", która pomaga w wyborze liczby k, na podstawie wartości Total Within Sum of Squares. Przypomnijmy, że wartość Silhouette wyraża się wzorem:

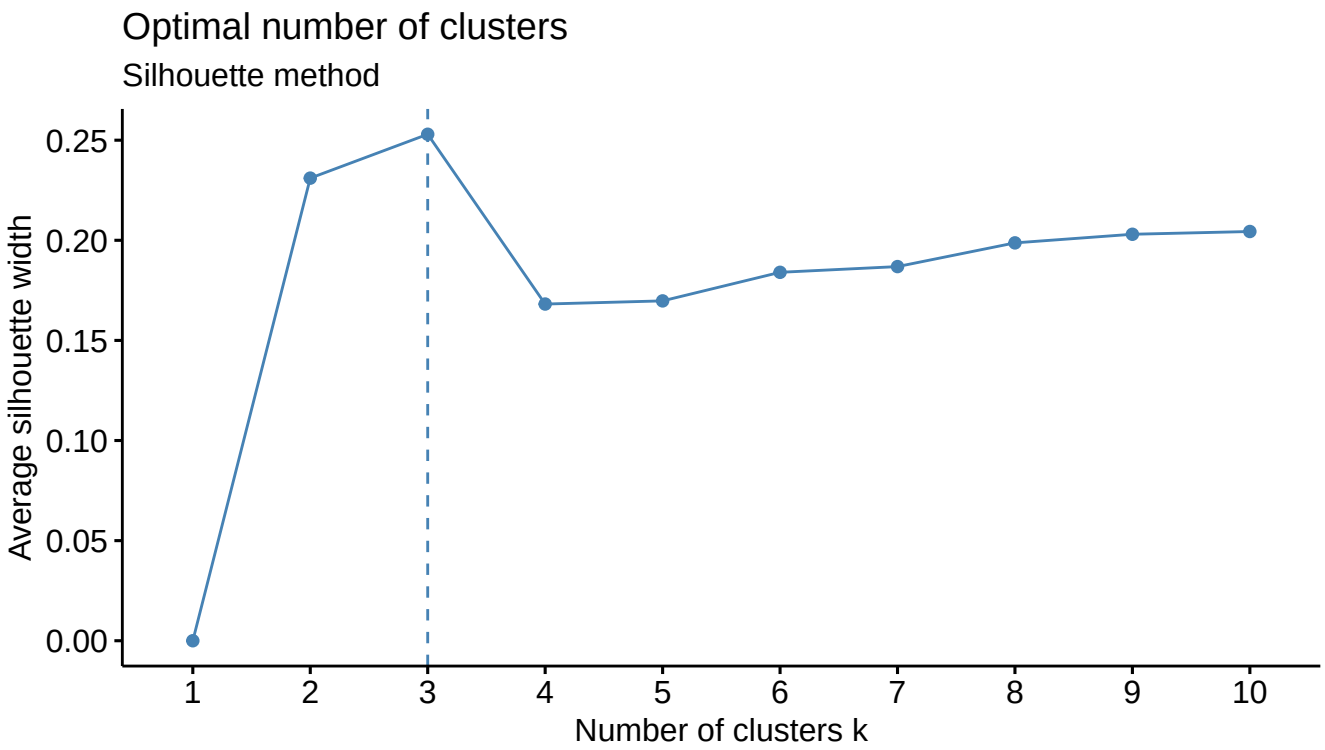
$$Silhouette = \frac{1}{n}\sum_{i=1}^{n} S(i) \in [-1, 1]\,,$$

gdzie $S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, $a(i) = \frac{1}{n_r - 1}\sum_{j \in C_r / id_{ij}}$, $b(i) = \min_{s \neq r} d_{iC_s}$, $d_{iC_s} = \frac{1}{n_s}\sum_{j \in C_s} d_{ij}$.
Wskaźniki stabilności oceniają spójność klasteryzacji, porównując wyniki z klastrami otrzymanymi po usunięciu jednej z kolumn. Rysunek 5 przedstawia wartości Total Within Sum of Squares w zależności od k.



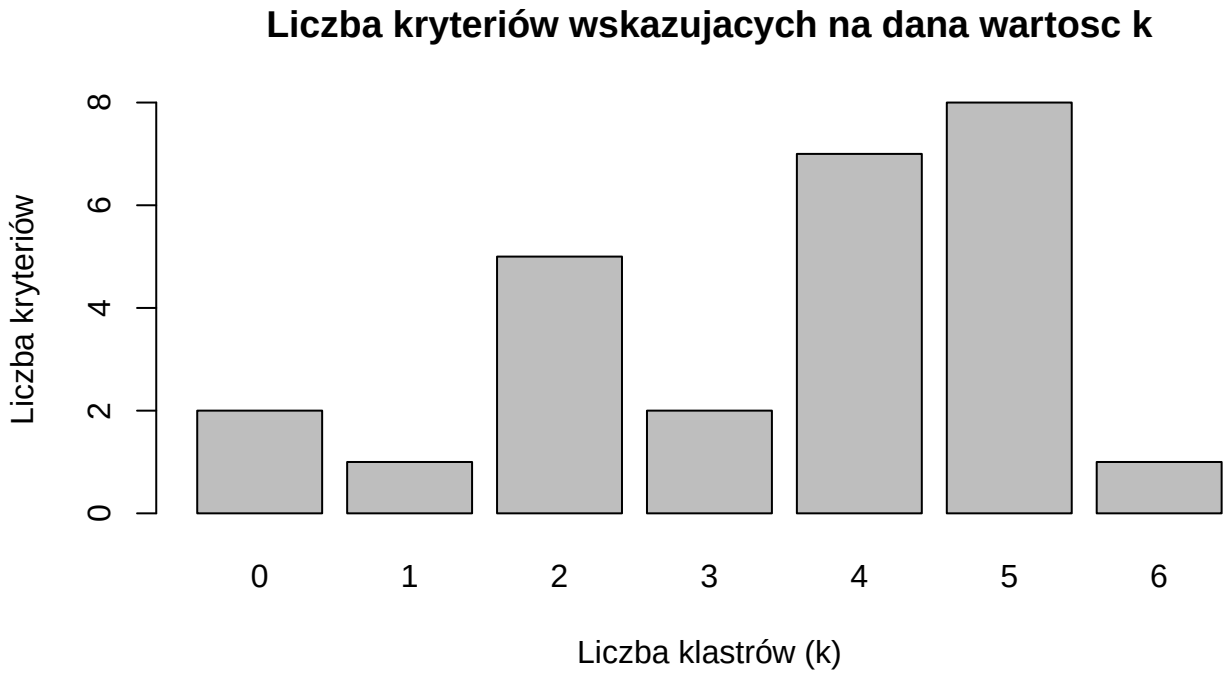
Rysunek 5: Wykres wartości Total Within Sum of Squares

Korzystając z "elbow method" wybieramy taką liczbę klastrów, przy której wykres zaczyna się wypłaszczać. Analizując powyższy rysunek trudno wybrać jedno optymalne k, może to być np. $k = 6$.



Rysunek 6: Wykres wartości Silhouette w zależności od liczby klastrów

Rysunek 6 wskazuje na to, że najwyższą wartość wskaźnika Silhouette otrzymujemy dla $k = 3$. Na Rysunku 7 zostało przedstawione, ile kryteriów z funkcji NbClust wybrało daną wartość k.



Rysunek 7: Liczba kryteriów wybierających daną liczbę klastrów

Z powyższego rysunku odczytujemy, że większość wskaźników przyjmowała optymalne wartości dla $k = 5$. Zwróćmy jednak uwagę na to, że funkcja ta nie bierze pod uwagę wskaźników stabilności takich jak: APN, AD, ADM, FOM.
W poniższej tabeli porównamy wartości kryteriów: Connectivity, Dunn i Silhouette w zależności od liczby klastrów. W tym celu wykorzystamy funkcję clValid.

	k=2	k=3	k=4	k=5	k=6
Connectivity	141.981	290.990	280.543	273.449	278.788
Dunn	0.033	0.031	0.023	0.019	0.024
Silhouette	0.191	0.199	0.203	0.196	0.187

Tabela 4: Wskaźniki wewnętrzne dla różnych liczb klastrów

	Score	Method	Clusters
Connectivity	141.981	kmeans	2
Dunn	0.033	kmeans	2
Silhouette	0.203	kmeans	4

Tabela 5: Optymalne liczby klastrów w zależności od kryterium

Analizując Tabelę 4 zauważamy, że wartości Silhouette nie różnią się znacząco dla różnych liczb klastrów. Z Tabeli 5 odczytujemy, że najwyższą wartość mamy dla $k = 4$. Zwróćmy uwagę na to, że z Rysunku 6 wynika, że optymalne jest $k = 3$. Zatem prawdopodobne jest, że wykorzystane funkcje mogą obliczać wartość Silhouette w różny sposób. Najmniejszą, czyli najbardziej pożądaną wartość Connectivity mamy dla $k = 2$, także wskaźnik Dunn wskazuje na taką liczbę klastrów (najwyższa wartość wskaźnika). Przejdziemy teraz do porównania wskaźników stabilności.

	k=2	k=3	k=4	k=5	k=6
APN	0.272	0.259	0.319	0.256	0.297
AD	3.484	3.209	3.048	2.862	2.800
ADM	1.116	0.936	0.996	0.835	0.961
FOM	0.999	0.996	0.986	0.979	0.963

Tabela 6: Wskaźniki stabilności dla różnych liczb klastrów

	Score	Method	Clusters
APN	0.256	kmeans	5
AD	2.800	kmeans	6
ADM	0.835	kmeans	5
FOM	0.963	kmeans	6

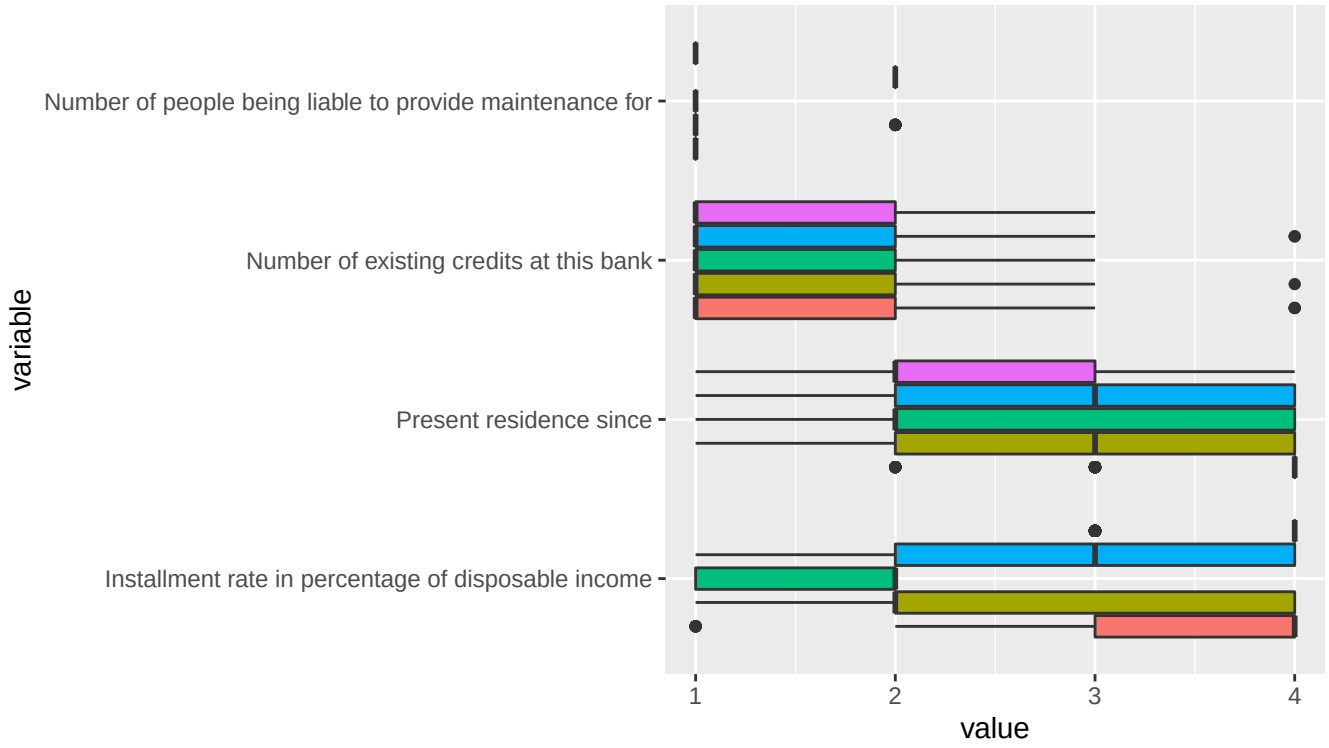
Tabela 7: Optymalne liczby klastrów w zależności od kryterium

Tabela 6 zawiera wartości wskaźników stabilności takich jak APN, AD, ADM oraz FOM dla $k \in \{2, 3, 4, 5, 6\}$. Ponownie widzimy, że wartości są zbliżone dla różnej liczby klastrów. Przykładowo, dla $k = 3$ wartość APN jest jedynie o 0.003 większa niż dla $k = 5$. W Tabeli 7 widzimy, że jako optymalne k wybierano 5 i 6. Na podstawie przeprowadzonych analiz nie możemy więc jednoznacznie wybrać liczby klastrów. Jednak w związku z tym, że na $k = 5$ wskazała większość wskaźników z funkcji NbClust oraz część wskaźników stabilności, sprawdzimy jak wyglądają rozkłady zmiennych w odpowiednich klastrach.

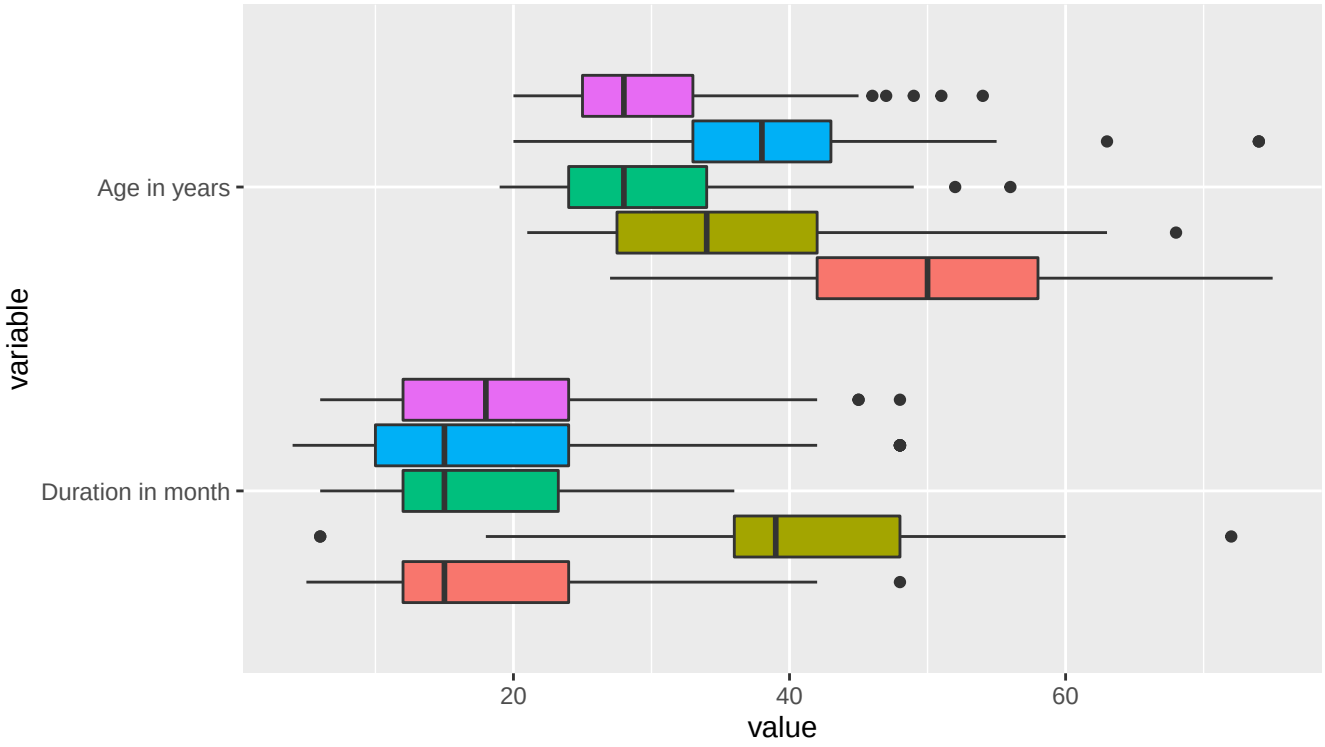
Podobnie jak po zastosowaniu k-means z $k = 2$, przeanalizujemy teraz wartości zmiennych w zależności od przypisanego klastra. Tabela 8 zawiera średnie wartości zmiennych, zaś Rysunki 8-10 przedstawiają ich box-ploty.

Cluster	Duration in month	Credit amount	Installment rate in percentage of disposable income	Present residence since	Age in years	Number of existing credits at this bank	Number of people being liable to provide maintenance for
1	16.83	2203.78	3.39	3.74	50.49	1.57	1.00
2	40.76	8664.14	2.63	2.82	35.73	1.45	1.12
3	16.20	2880.61	1.60	2.67	29.93	1.32	1.00
4	17.77	2730.96	2.81	2.93	38.69	1.55	2.00
5	19.04	2083.57	3.78	2.44	29.43	1.29	1.00

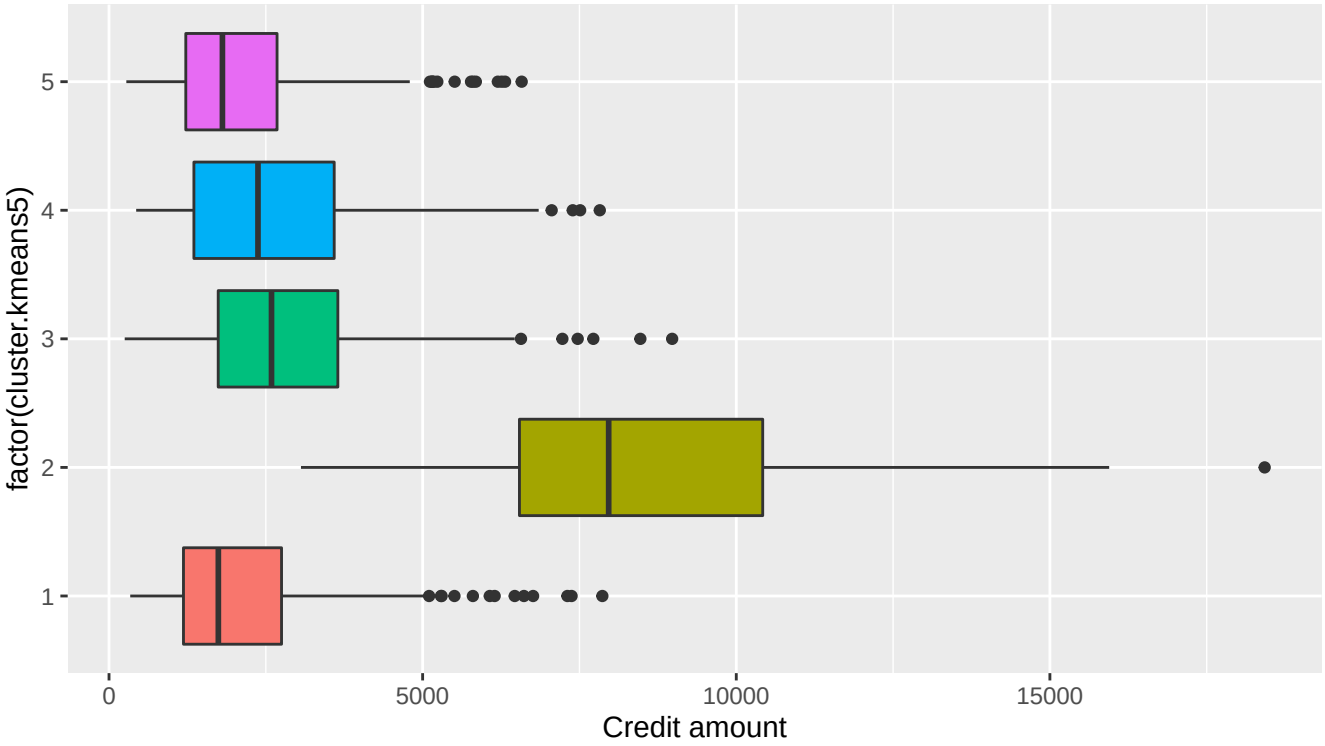
Tabela 8: Średnie wartości zmiennych w klastrach



Rysunek 8: Box-ploty zmiennych numerycznych z podziałem na klastry



Rysunek 9: Box-ploty zmiennych numerycznych z podziałem na klastry



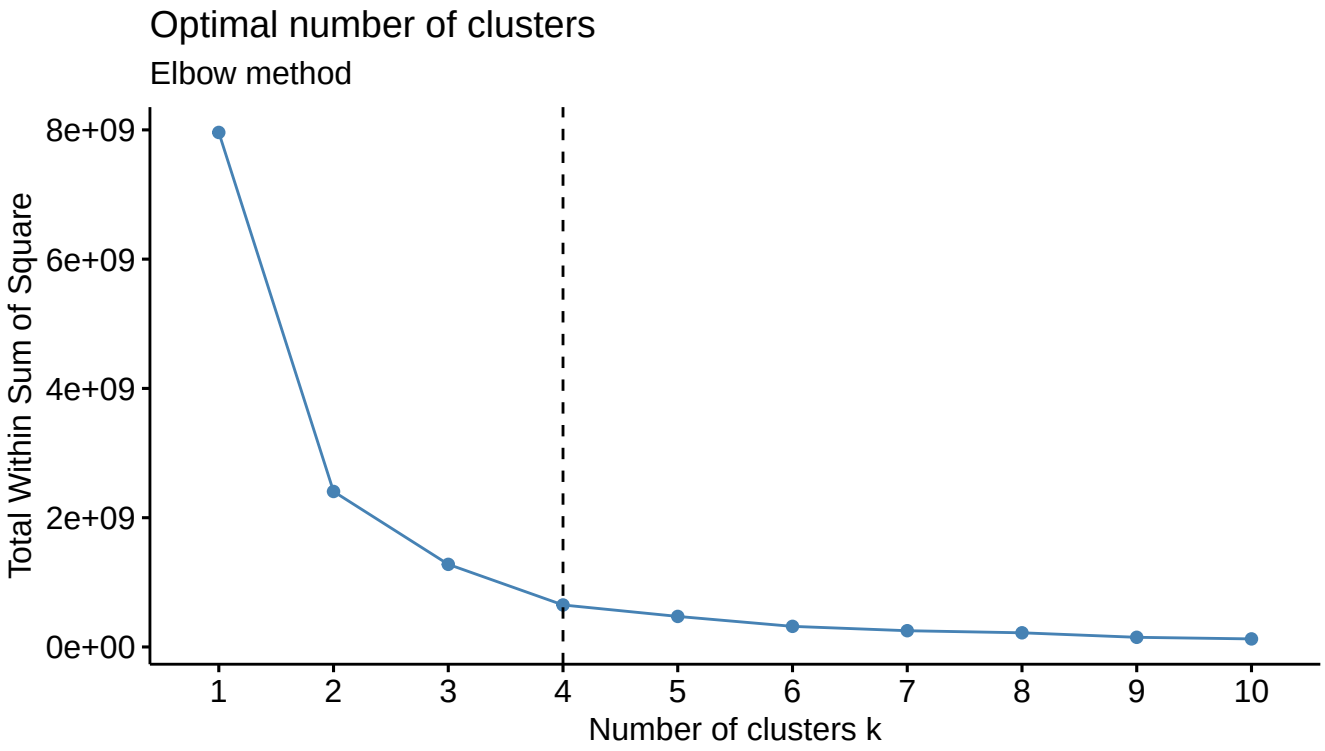
Rysunek 10: Box-plot zmiennej Credit amount z podziałem na klastry

Na powyższych rysunkach i w tabeli widzimy, że do jednego z klastrów trafiły obserwacje z wysokimi wartościami zmiennych Credit amount i Duration in month. W innym klastrze znajdują się klienci starsi niż w

pozostałych. Trudno jednak zrozumieć, dlaczego optymalną liczbą klastrów miałaby być liczba 5.

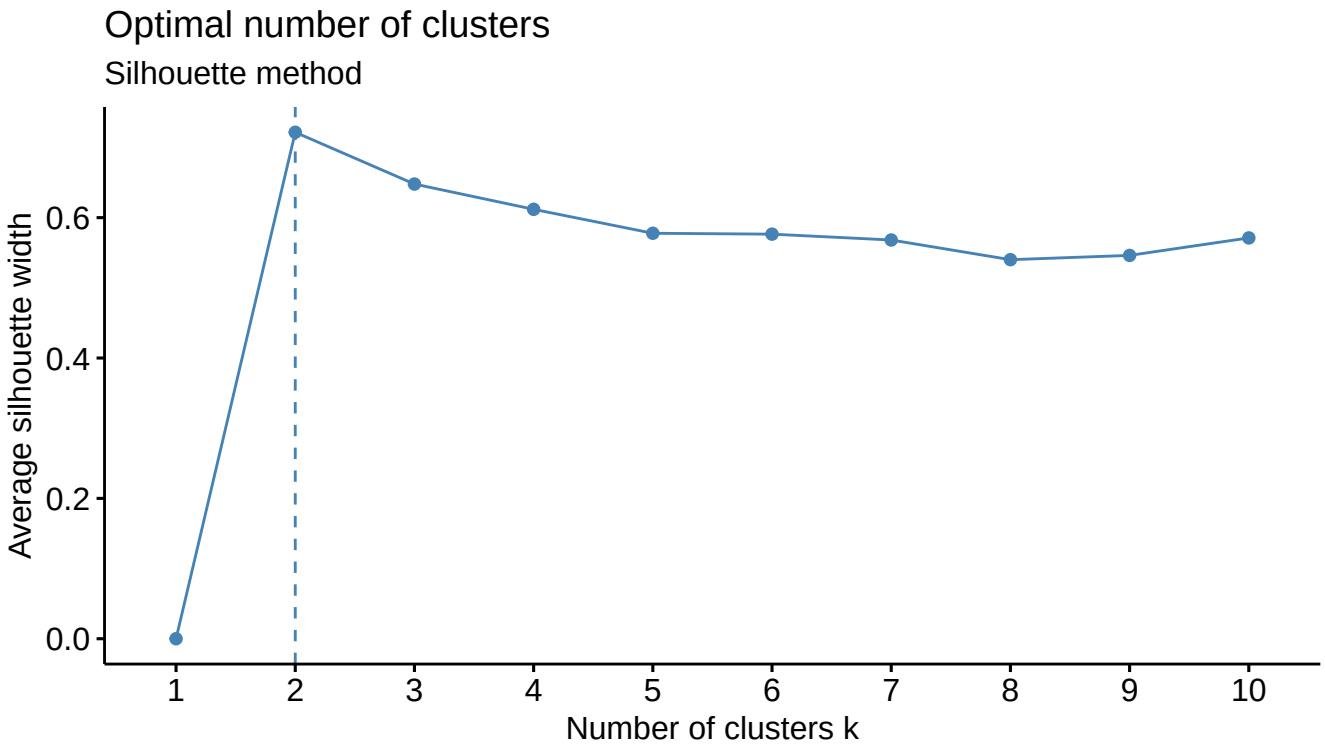
2.1.2 Dla danych bez skalowania

Zobaczymy, jak zmieniłyby się wyniki, gdybyśmy nie wykonywali skalowania danych.



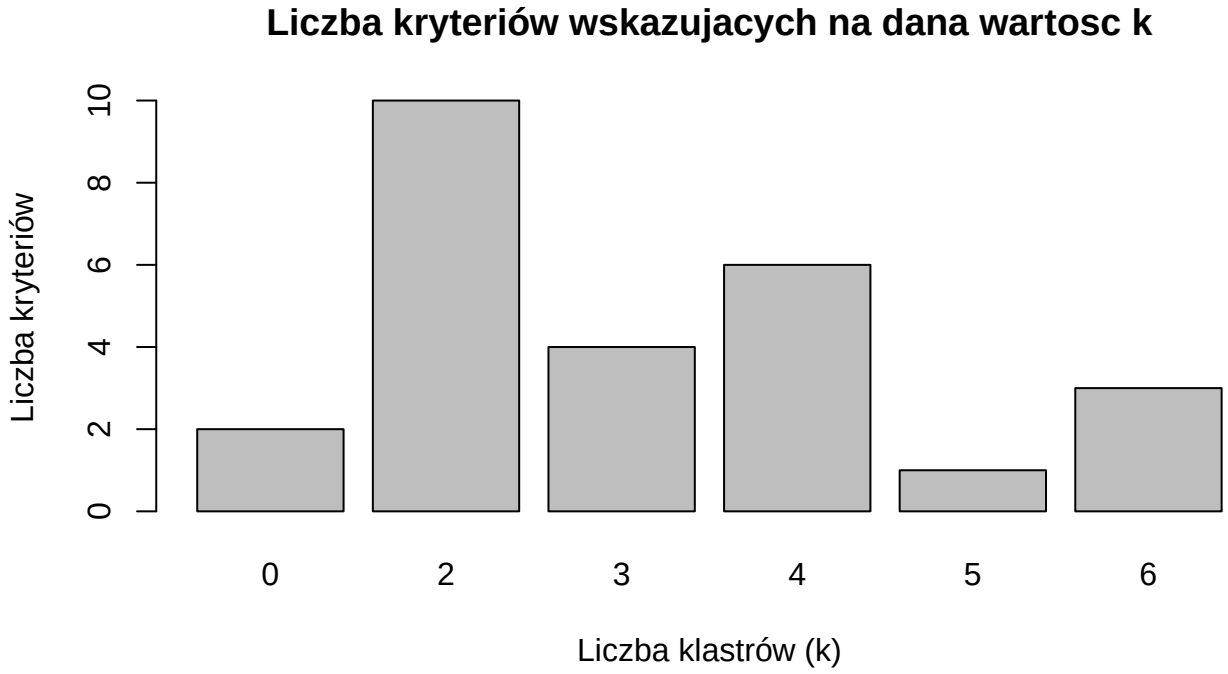
Rysunek 11: Wykres wartości Total Within Sum of Squares

Na Rysunku 11 widzimy, że przy użyciu "elbow method" wybieramy inne k niż w przypadku zastosowania algorytmu do danych przeskalowanych. Powyższy wykres wskazuje na $k = 4$. Warto zauważyć, że wartości Total Within Sum of Squares są znacznie wyższe niż te, które uzyskaliśmy wcześniej - po skalowaniu.



Rysunek 12: Wykres wartości Silhouette w zależności od liczby klastrów

Rysunek 12 zawiera informacje o wartościach wskaźnika Silhouette. Są one wyższe niż te które otrzymaliśmy dla danych przeskalowanych.



Rysunek 13: Liczba kryteriów wybierających daną liczbę klastrów

Najwięcej kryteriów wybiera jako optymalne $k = 2$. Przejdźmy do oceny jakości grupowania.

	k=2	k=3	k=4	k=5	k=6
Connectivity	6.866	7.703	16.737	17.958	22.331
Dunn	0.002	0.003	0.001	0.003	0.003
Silhouette	0.722	0.648	0.612	0.598	0.574

Tabela 9: Wskaźniki wewnętrzne dla różnych liczb klastrów

	Score	Method	Clusters
Connectivity	6.866	kmeans	2
Dunn	0.003	kmeans	6
Silhouette	0.722	kmeans	2

Tabela 10: Optymalne liczby klastrów w zależności od kryterium

Wskaźniki Silhouette i Connectivity wskazują na $k = 2$.

	k=2	k=3	k=4	k=5	k=6
APN	0.058	0.072	0.078	0.079	0.095
AD	1752.421	1377.658	1044.577	964.501	880.786
ADM	254.689	201.401	236.008	237.221	232.309
FOM	406.848	340.600	337.273	334.646	335.052

Tabela 11: Wskaźniki stabilności dla różnych liczb klastrów

	Score	Method	Clusters
APN	0.058	kmeans	2
AD	880.786	kmeans	6
ADM	201.401	kmeans	3
FOM	334.646	kmeans	5

Tabela 12: Optymalne liczby klastrów w zależności od kryterium

Zwróćmy uwagę na wartości wskaźników stabilności. W porównaniu do wyników uzyskanych po skalowaniu danych, otrzymane wartości są bardzo duże. Świadczy to o niskiej jakości grupowania. Ponadto nie otrzymujemy potwierdzenia, że $k = 2$ jest optymalne, gdyż każde z kryteriów wybiera inną wartość k. Wykorzystamy teraz macierz kontyngencji.

	1	2
1	102	73
2	598	227

Tabela 13: Macierz kontyngencji - k-means

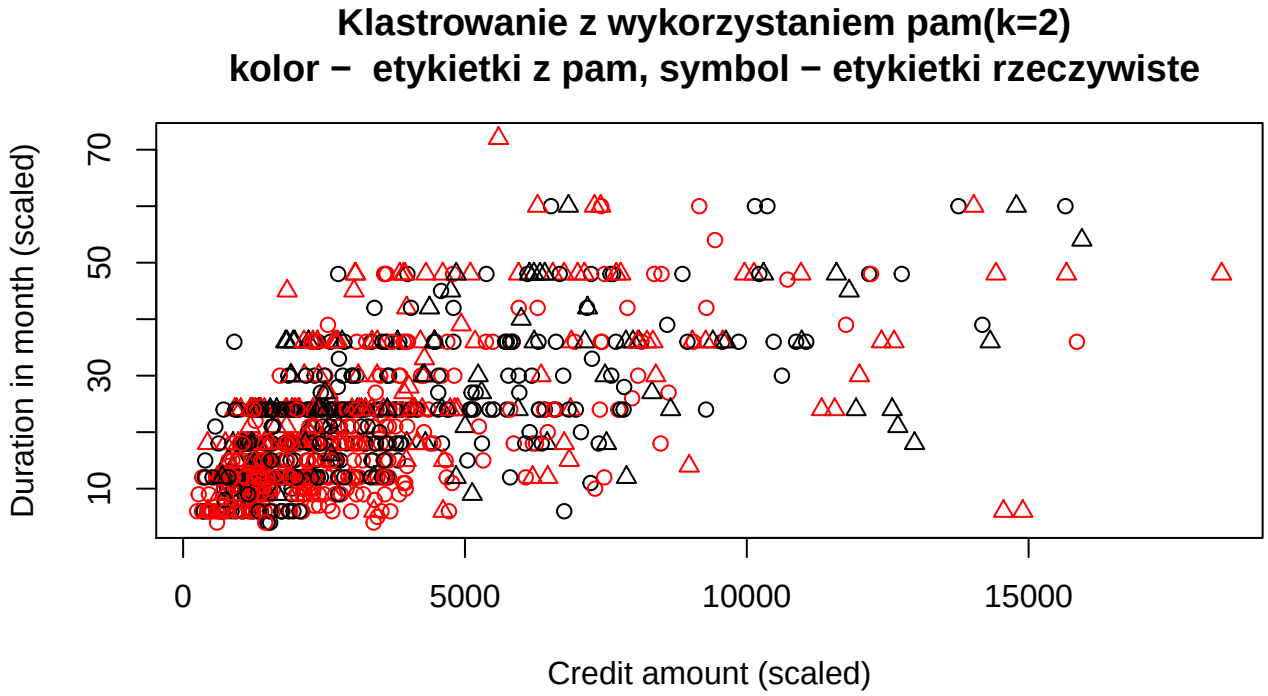
```
## Direct agreement: 1 of 2 pairs
## Iterations for permutation matching: 1
## Cases in matched pairs: 67.1 %
## 1 2
## 2 1
```

Otrzymany procent zgodności klasy przypisanej i rzeczywistej jest dość wysoki i wynosi 67.1%.

2.2 PAM

Następnym algorytmem analizy skupień, który zastosujemy do danych, jest algorytm PAM. Jego zaletą jest to, że możemy go zastosować do danych mieszanego typu. Pierwszym krokiem jest wyznaczenie tzw. macierzy niepodobieństwa. Wykorzystujemy w tym celu funkcję *daisy*. Jednym z jej argumentów jest *metric*, który odpowiada za miarę odmienności. W przypadku kiedy w danych występują zmienne o typie innym niż *numeric* automatycznie wybierana jest miara Gowera. Najpierw rozważymy $k = 2$.

Zilustrujmy na początek wyniki analizy skupień dla dwóch wybranych zmiennych. Tak samo jak w przypadku k-means weźmiemy zmienne Credit amount oraz Duration in month.

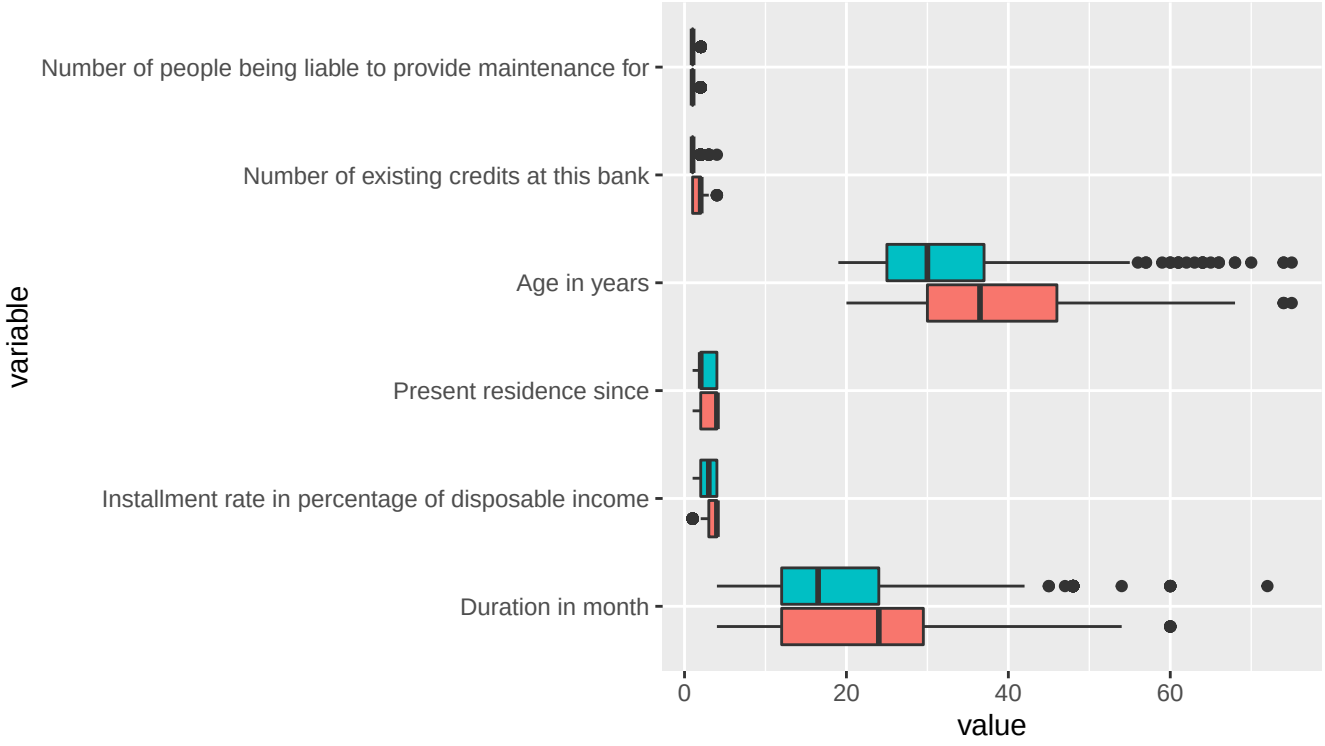


Rysunek 14: Porównanie etykietek po zastosowaniu pam dla k = 2

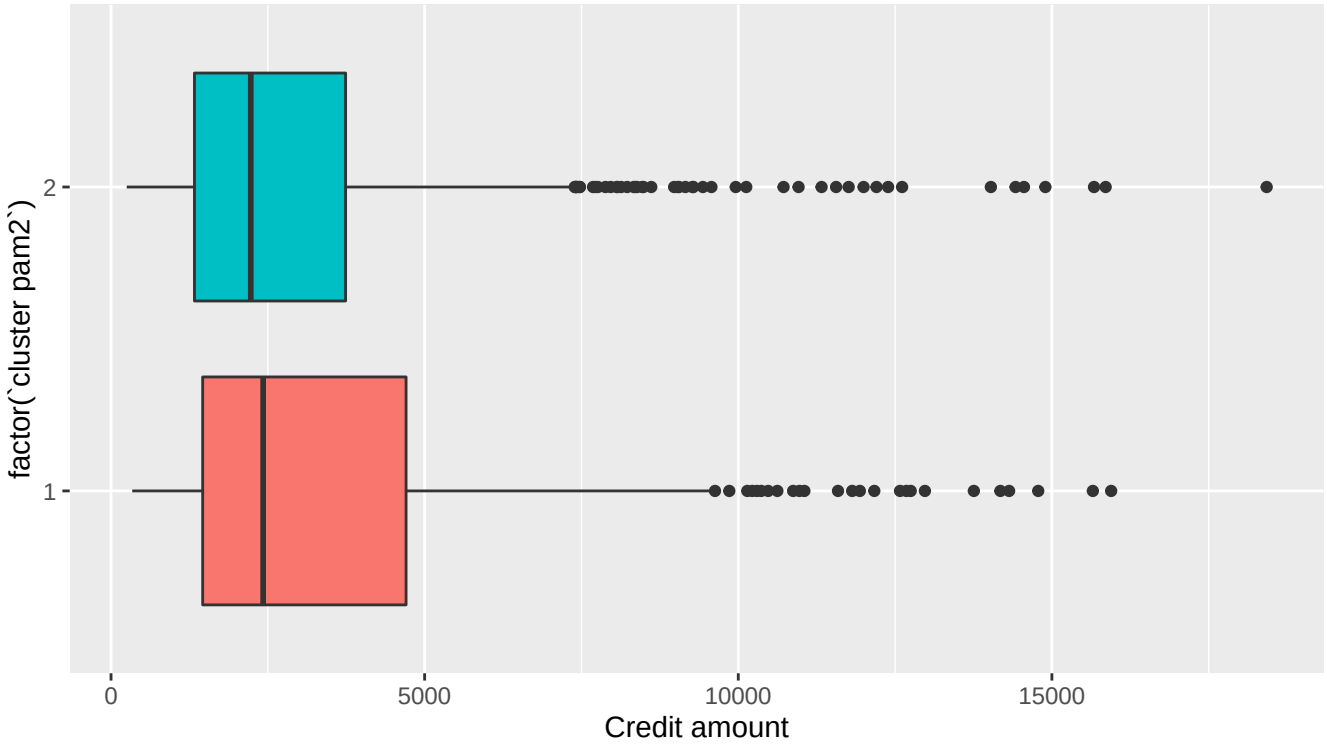
Analizując Rysunek 14 możemy zauważyć, że klastry nie zostały podzielone ze względu na wartości tych dwóch zmiennych tak jak to było w przypadku klastrów powstałych przy użyciu algorytmu 2-means.
Przeanalizujemy teraz wartości zmiennych w zależności od przypisanego klastra. Tabela 14 zawiera średnie wartości zmiennych, zaś Rysunki 15 i 16 przedstawiają ich box-ploty.

Cluster	Duration in month	Credit amount	Installment rate in percentage of disposable income	Present residence since	Age in years	Number of existing credits at this bank	Number of people being liable to provide maintenance for
1	22.68	3521.88	3.34	3.22	38.88	1.63	1.16
2	19.54	3079.09	2.69	2.55	32.99	1.23	1.15

Tabela 14: Średnie wartości zmiennych numerycznych w klastrach



Rysunek 15: Box-ploty zmiennych numerycznych z podziałem na klastry



Rysunek 16: Box-plot zmiennej Credit amount z podziałem na klastry

Rysunek 15 potwierdza to co zilustrowaliśmy na Rysunku 14, tzn. wartości przyjmowane przez zmienne Credit Amount oraz Duration in Month są porównywalne. Podobnie jest dla pozostałych zmiennych. Zmienne numeryczne przyjmują podobne wartości w obu klastrach.

Status of existing checking account	Credit history	Purpose	Savings account	Present employment since	Personal status and sex	Other debtors
A11: 64	A30: 16	A43 :154	A61:228	A71: 30	A91: 14	A101:406
A12:104	A31: 15	A40 : 90	A62: 50	A72: 53	A92:112	A102: 15
A13: 21	A32:138	A41 : 63	A63: 29	A73: 75	A93:279	A103: 13
A14:245	A33: 46	A49 : 43	A64: 24	A74: 79	A94: 29	
	A34:219	A42 : 38	A65:103	A75:197		
		A46 : 22				
		(Other): 24				

Tabela 15: Statystyki opisowe dla zmiennych jakościowych w klastrze 1

Status of existing checking account	Credit history	Purpose	Savings account	Present employment since	Personal status and sex	Other debtors
A11:210	A30: 24	A40 :144	A61:375	A71: 32	A91: 36	A101:501
A12:165	A31: 34	A42 :143	A62: 53	A72:119	A92:198	A102: 26
A13: 42	A32:392	A43 :126	A63: 34	A73:264	A93:269	A103: 39
A14:149	A33: 42	A49 : 54	A64: 24	A74: 95	A94: 63	
	A34: 74	A41 : 40	A65: 80	A75: 56		
		A46 : 28				
		(Other): 31				

Tabela 16: Statystyki opisowe dla zmiennych jakościowych w klastrze 2

Status of existing checking account	Credit history	Purpose	Savings account	Present employment since	Personal status and sex	Other debtors
A11:274	A30: 40	A43 :280	A61:603	A71: 62	A91: 50	A101:907
A12:269	A31: 49	A40 :234	A62:103	A72:172	A92:310	A102: 41
A13: 63	A32:530	A42 :181	A63: 63	A73:339	A93:548	A103: 52
A14:394	A33: 88	A41 :103	A64: 48	A74:174	A94: 92	
	A34:293	A49 : 97	A65:183	A75:253		
		A46 : 50				
		(Other): 55				

Tabela 17: Statystyki opisowe dla zmiennych jakościowych dla danych przed klasteryzacją

Property	Other installment plans	Housing	Job	Telephone	foreign worker
A121: 72	A141: 58	A151: 59	A171: 6	A191:151	A201:430
A122: 80	A142: 22	A152:312	A172: 55	A192:283	A202: 4
A123:195	A143:354	A153: 63	A173:275		
A124: 87			A174: 98		

Tabela 18: Statystyki opisowe dla zmiennych jakościowych w klastrze 1

Property	Other installment plans	Housing	Job	Telephone	foreign worker
A121:210	A141: 81	A151:120	A171: 16	A191:445	A201:533
A122:152	A142: 25	A152:401	A172:145	A192:121	A202: 33
A123:137	A143:460	A153: 45	A173:355		
A124: 67			A174: 50		

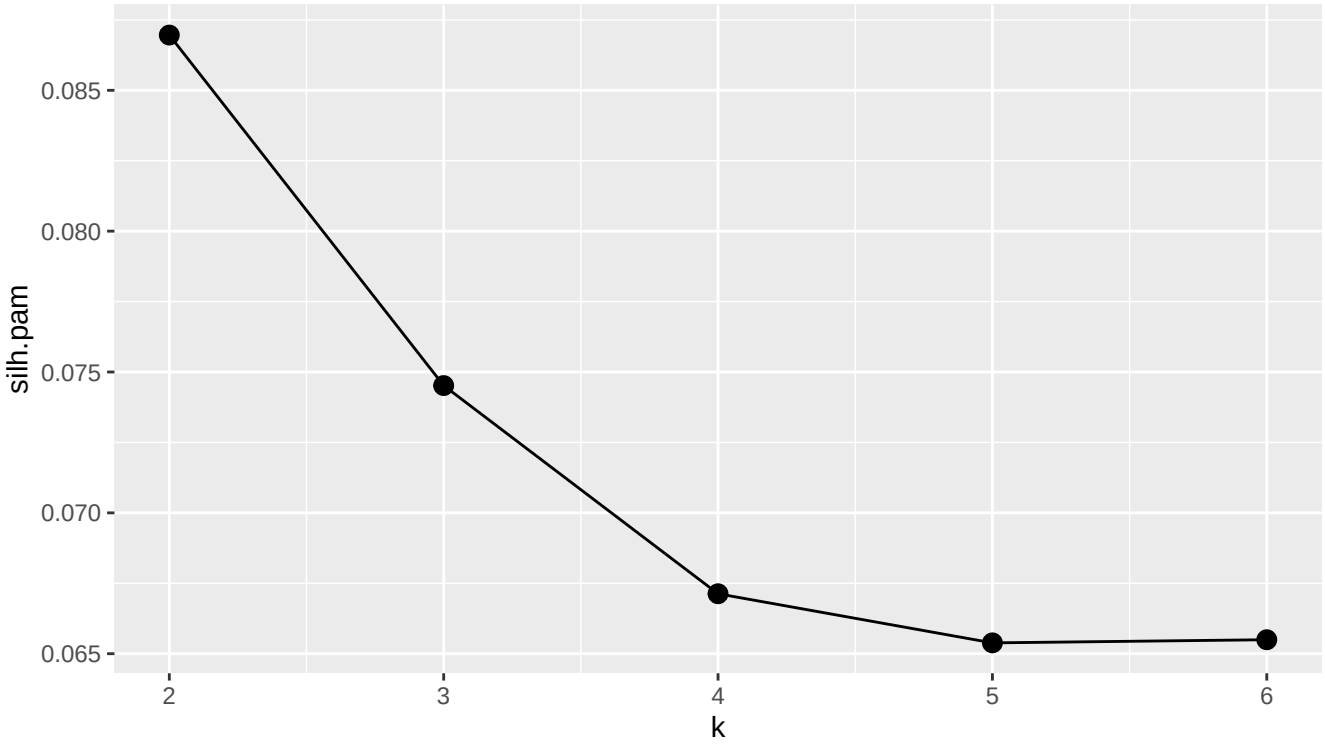
Tabela 19: Statystyki opisowe dla zmiennych jakościowych w klastrze 2

Property	Other installment plans	Housing	Job	Telephone	foreign worker
A121:282	A141:139	A151:179	A171: 22	A191:596	A201:963
A122:232	A142: 47	A152:713	A172:200	A192:404	A202: 37
A123:332	A143:814	A153:108	A173:630		
A124:154			A174:148		

Tabela 20: Statystyki opisowe dla zmiennych jakościowych dla danych przed klasteryzacją

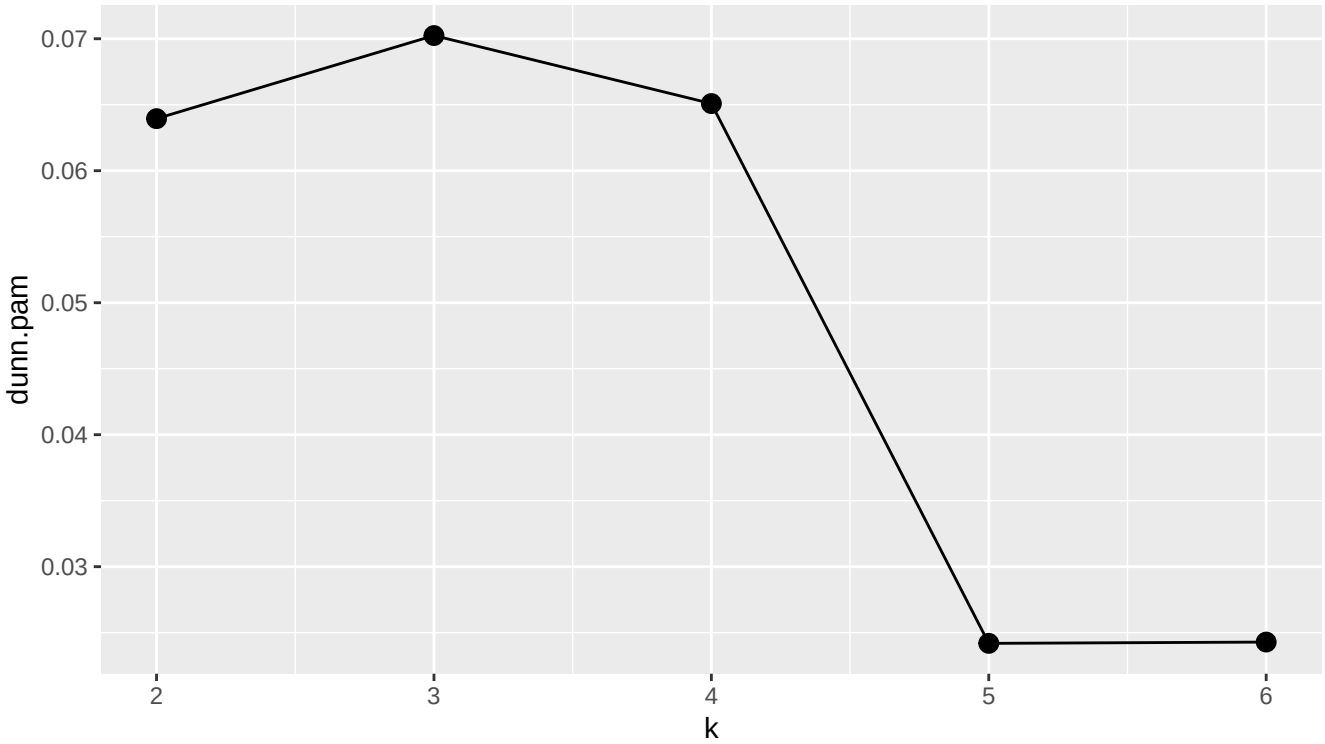
Po przeanalizowaniu Tabel 15- 20 możemy zauważyć m.in. że w klastrze drugim stosunek osób, które nie posiadają telefonu (oznaczenie A191) do wszystkich przypisanych do tego klastra wynosi 78%. Dla danych z przed podziału na klastry udział ten wynosi około 60%. Wydaje się więc, że to czy posiada się telefon czy nie, może mieć znaczący wpływ na przydział do klastra. W pozostałych przypadkach nie widać wyraźnie zwiększonego udziału danych poziomów w klastrach.

Zobaczymy, jaka liczba k jest wybierana przez wskaźniki oceniające grupowanie.



Rysunek 17: Wartości wskaźnika Silhouette w zależności od liczby klastrów

Z powyższego rysunku odczytujemy, że największą wartość Silhouette mamy dla $k = 2$.



Rysunek 18: Wartości wskaźnika Dunn dla różnej liczby klastrów

Na powyższym rysunku widzimy, że wskaźnik Dunn jest najwyższy dla $k = 3$.

Do oceny zastosujemy teraz tzw. miary zewnętrzne, które porównują przypisane klasy z rzeczywistymi. Najprostszym sposobem jest wykorzystanie macierzy kontyngencji.

	1	2
1	338	96
2	362	204

Tabela 21: Macierz kontyngencji - PAM

```
## Direct agreement: 0 of 2 pairs
## Iterations for permutation matching: 2
## Cases in matched pairs: 54.2 %
## 1 2
## 1 2
```

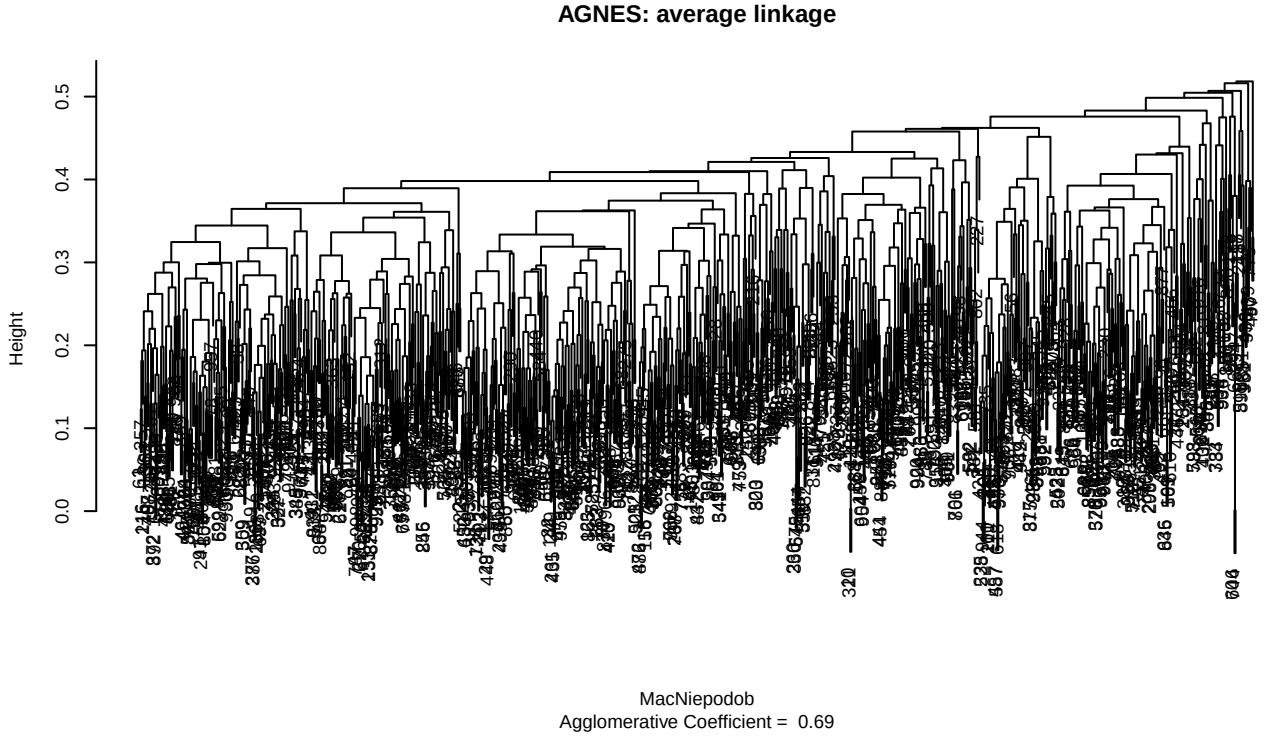
Otrzymany procent zgodności klas rzeczywistych i przypisanych nie jest szczególnie wysoki - wynosi $54,2\%$.

2.3 AGNES

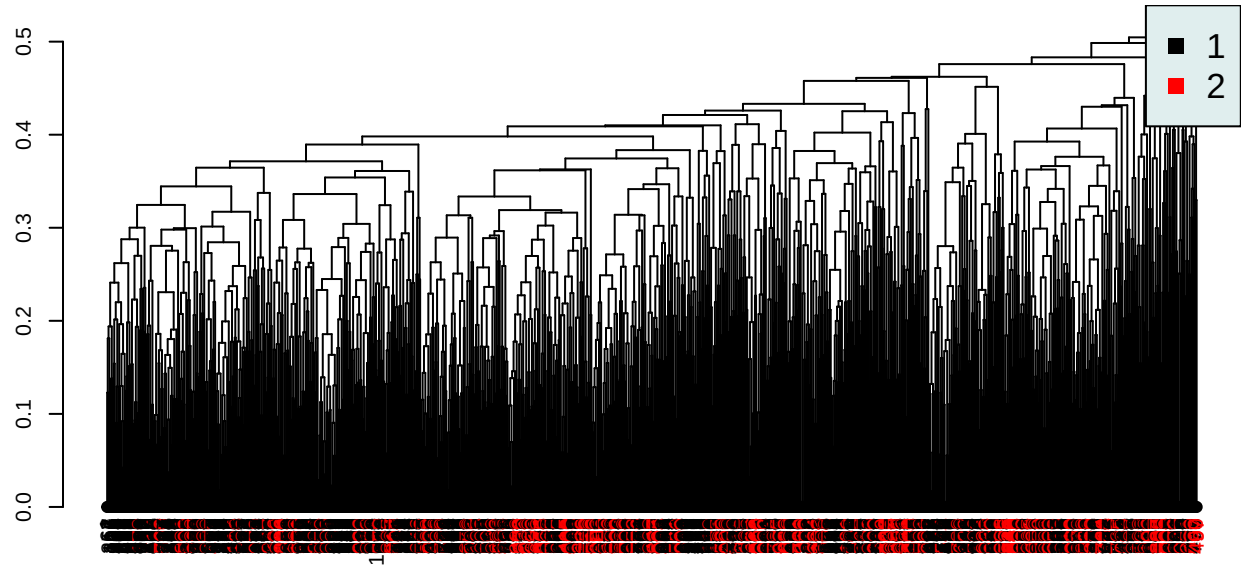
Przejdziemy teraz do zastosowania metod hierarchicznych. Jako pierwszą zastosujemy metodę aglomeracyjną AGNES. Porównamy trzy metody łączenia klastrów: average, single, complete.

2.3.1 average linkage

Rysunek 19 przedstawia dendrogram, zaś na Rysunku 20 znajduje się dendrogram z zaznaczonymi klasami rzeczywistymi.



Rysunek 19: Dendrogram - AGNES - average linkage



Rysunek 20: Dendrogram - AGNES - average linkage z etykietkami rzeczywistymi

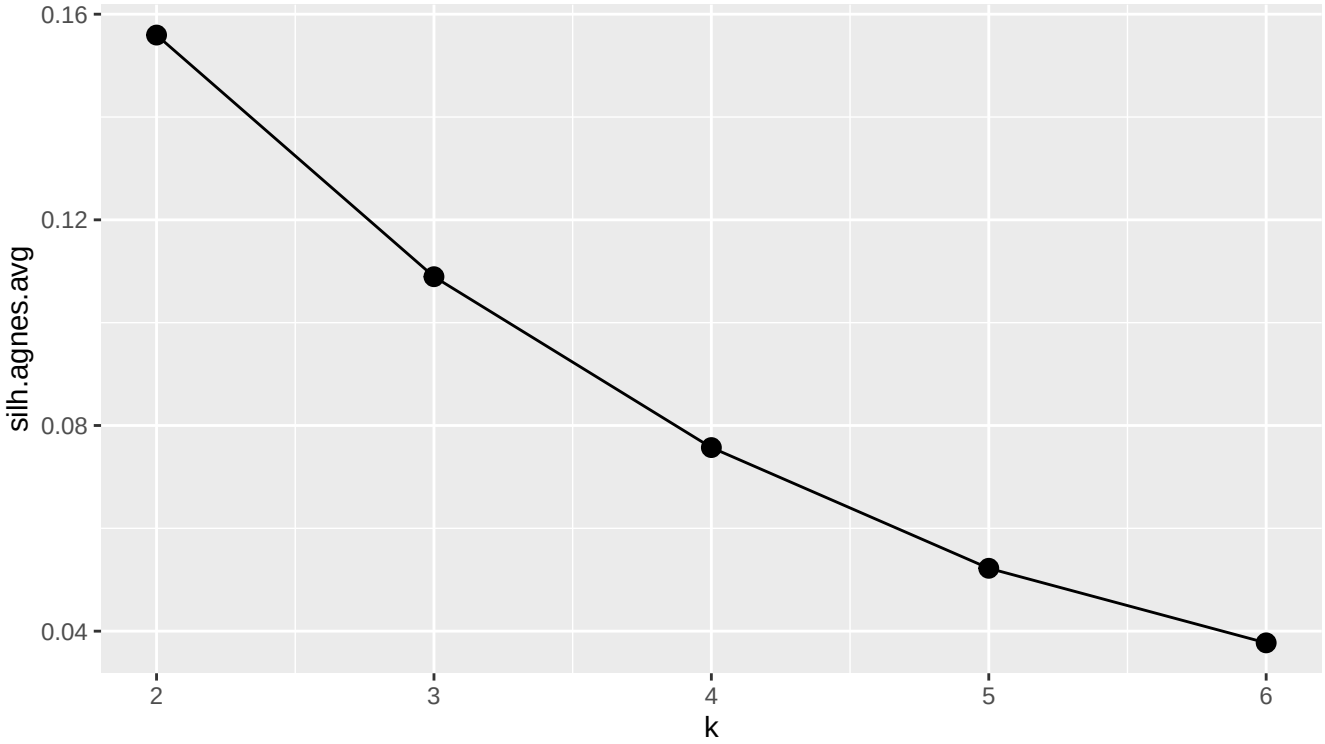
Poniższa tabela przedstawia macierz kontyngencji dla $k = 2$.

	1	2
1	700	298
2	0	2

Tabela 22: Macierz kontyngencji - AGNES - average linkage

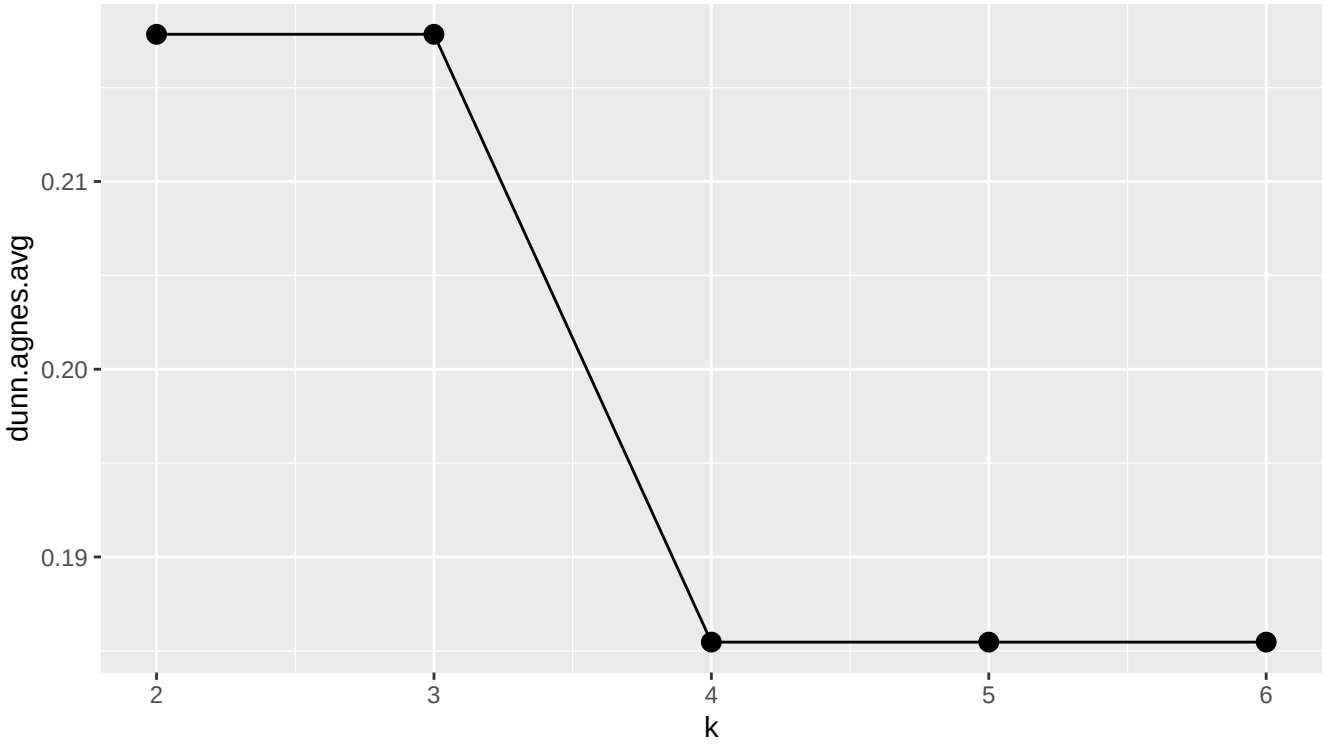
```
## Direct agreement: 1 of 2 pairs
## Iterations for permutation matching: 1
## Cases in matched pairs: 70.2 %
## 1 2
## 1 2
```

Zauważmy, że po klasteryzacji otrzymaliśmy zaledwie dwie obserwacje w jednym klastrze. W związku z tym, dalsza analiza traci sens. Również w przypadku wybrania $k = 3$, prawie wszystkie obserwacje należą do jednego klastra.



Rysunek 21: Wartości wskaźnika Silhouette w zależności od liczby klastrów

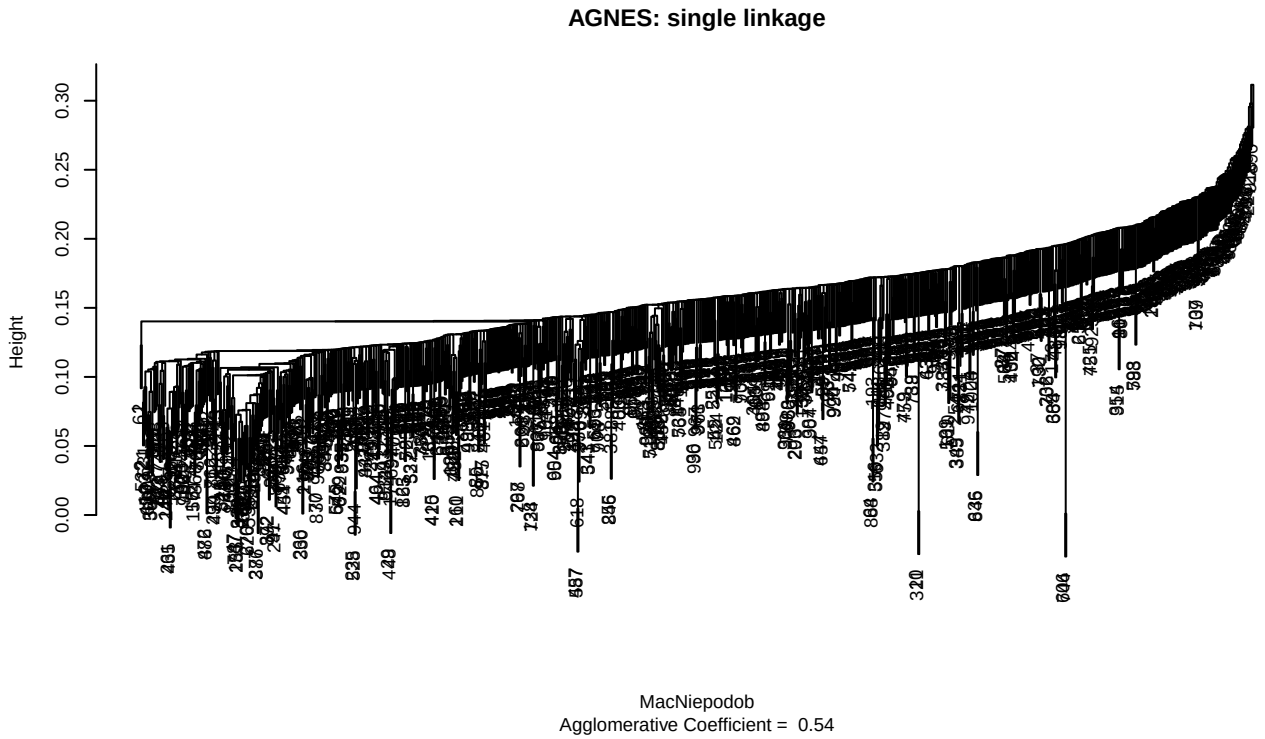
Z Rysunku 21 odczytujemy, że najwyższą wartość Silhouette mamy dla $k = 2$, jednak wartość ta jest stosunkowo niska. Dla wyższych k wartość coraz bardziej zbliża się do 0. Na tę samą liczbę klastrów wskazuje także wskaźnik DUNN. Dla większych k widzimy tendencję malejącą.



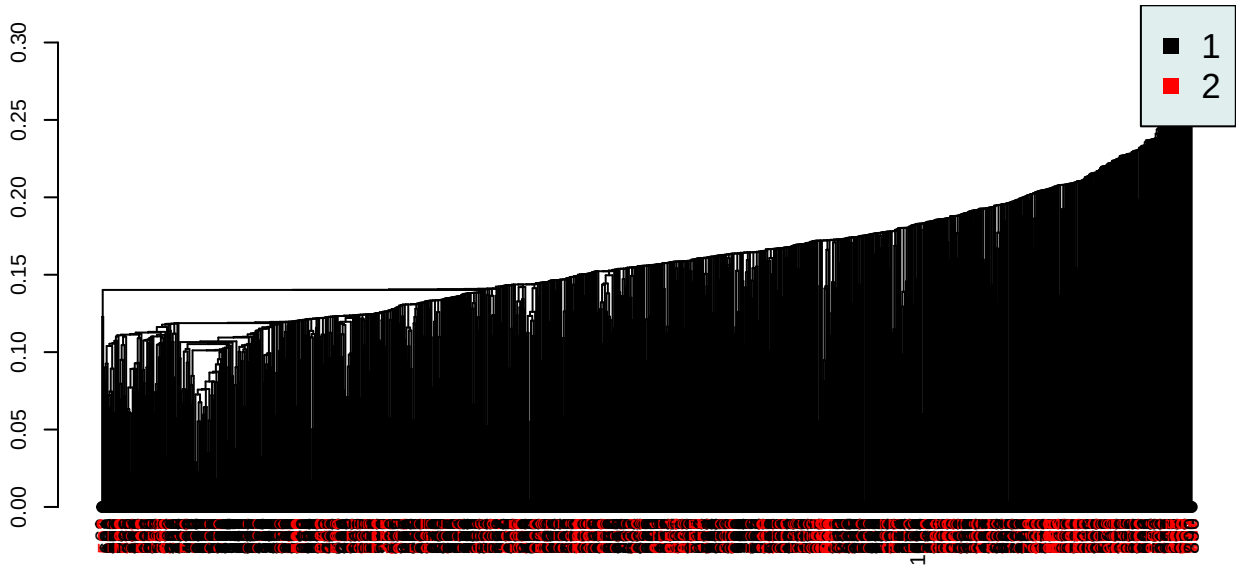
Rysunek 22: Wartości wskaźnika Dunn dla różnej liczby klastrów

2.3.2 single linkage

Dendrogram oraz dendrogram z etykietkami rzeczywistymi dla metody łączenia single zostały przedstawione na Rysunkach 23 i 24.



Rysunek 23: Dendrogram - AGNES - single linkage



Rysunek 24: Dendrogram - AGNES - single linkage z etykietkami rzeczywistymi

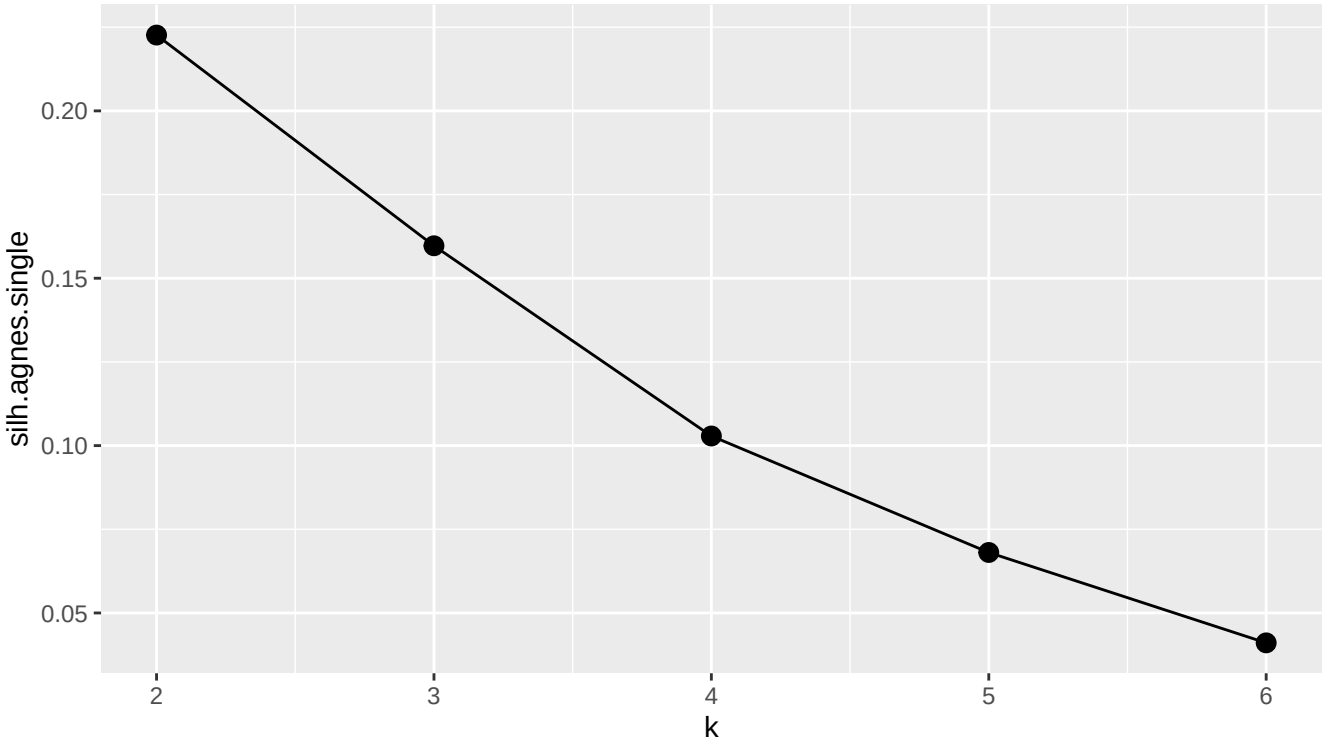
Macierz kontyngencji dla $k = 2$ znajduje się w Tabeli 45.

	1	2
1	699	300
2	1	0

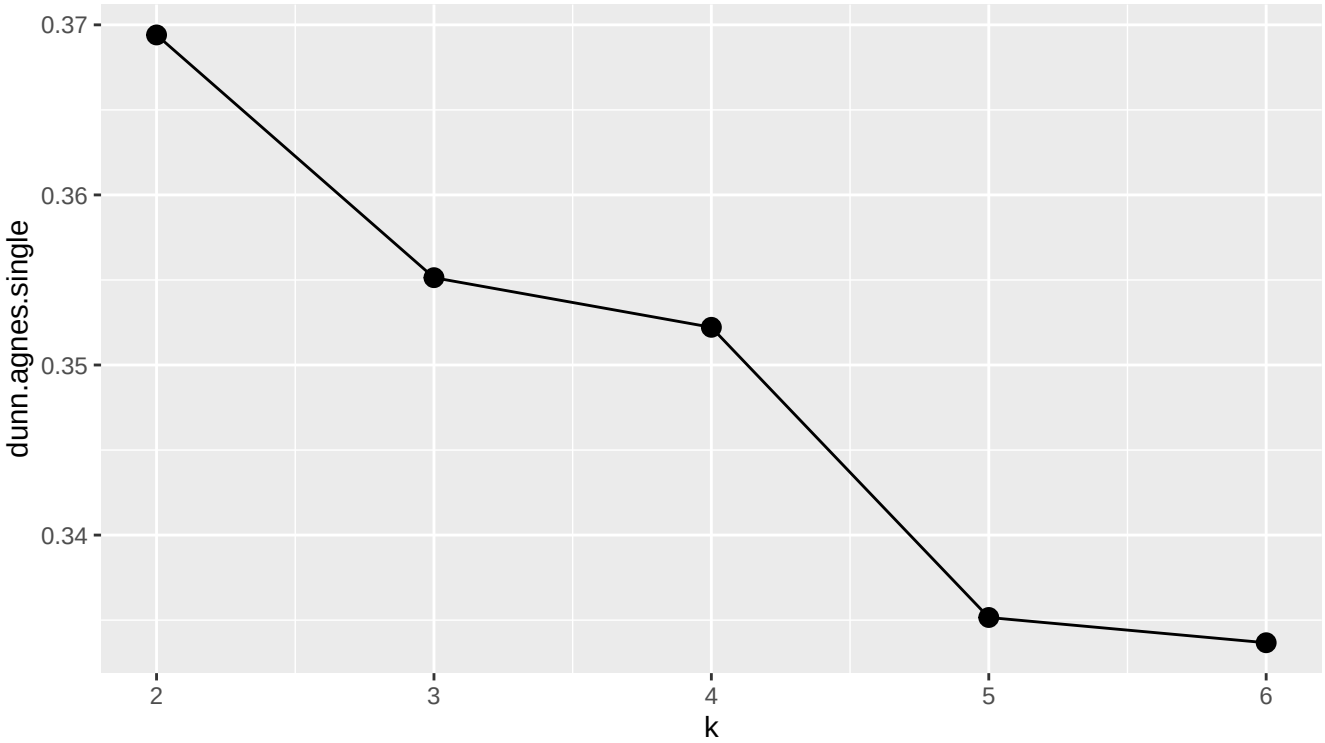
Tabela 23: Macierz kontyngencji - AGNES - single linkage

```
## Direct agreement: 1 of 2 pairs
## Iterations for permutation matching: 1
## Cases in matched pairs: 69.9 %
## 1 2
## 1 2
```

W przypadku tej metody łączenia otrzymujemy podobne wnioski jak poprzednio. Po klasteryzacji otrzymaliśmy zaledwie jedną obserwację w jednym z klastrów. Ponownie więc dalsza analiza traci sens. Również w przypadku wybrania $k = 3$, prawie wszystkie obserwacje należą do jednego klastra.



Rysunek 25: Wartości wskaźnika Silhouette w zależności od liczby klastrów

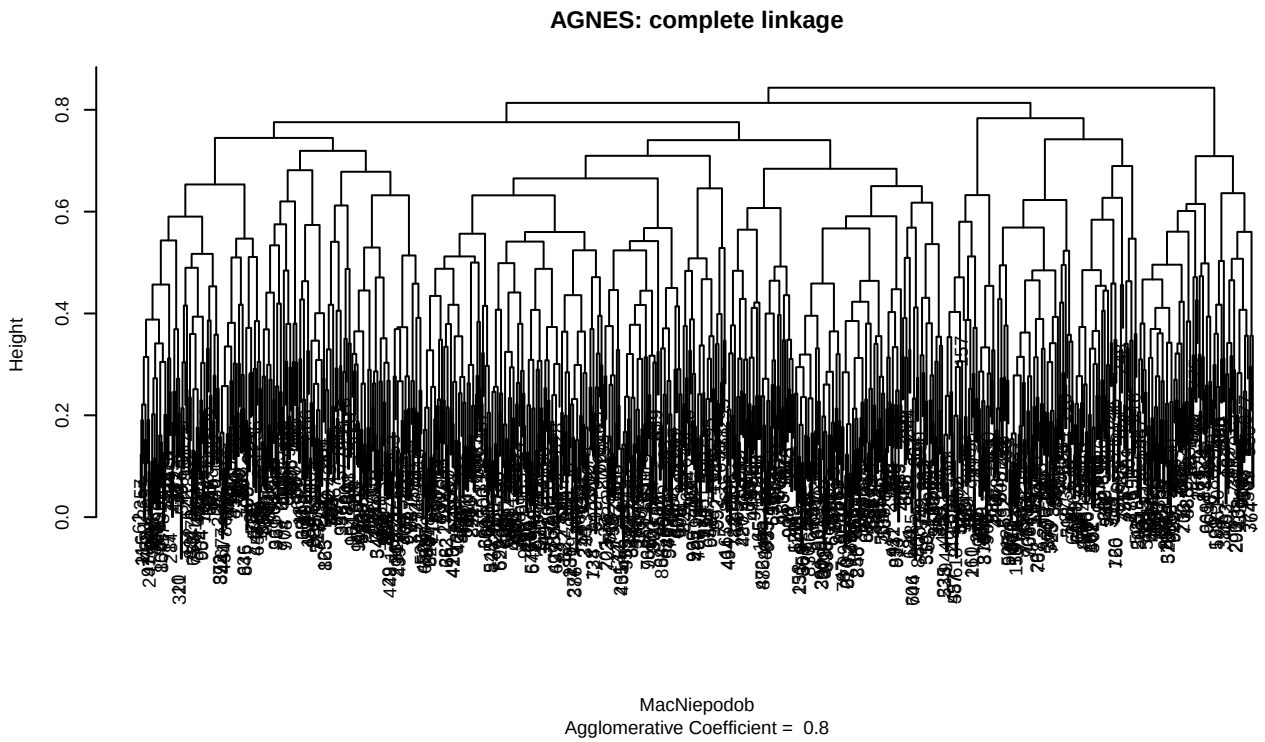


Rysunek 26: Wartości wskaźnika Dunn dla różnej liczby klastrów

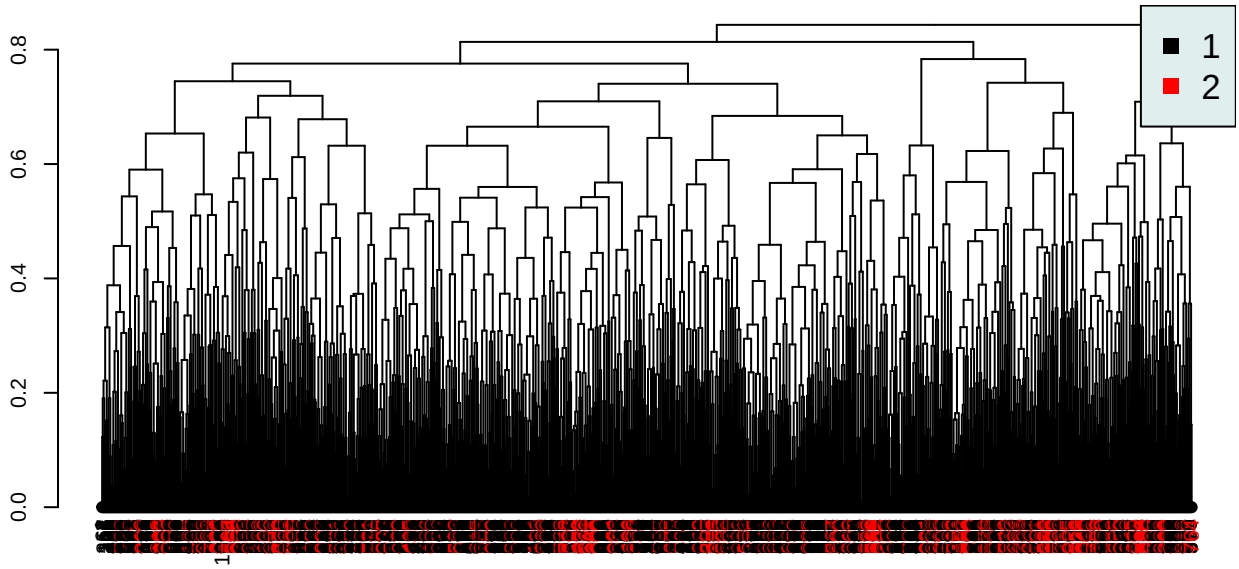
Ponownie jak w przypadku average linkage najwyższą wartość Silhouette otrzymujemy dla $k = 2$. Wartość ta jest nieznacznie wyższa niż dla poprzedniej metody łączenia.

2.3.3 complete linkage

Na Rysunkach 27 i 28 znajdują się odpowiednio - dendrogram i dendrogram z klasami rzeczywistymi.



Rysunek 27: Dendrogram - AGNES - complete linkage



Rysunek 28: Dendrogram - AGNES - complete linkage z etykietkami rzeczywistymi

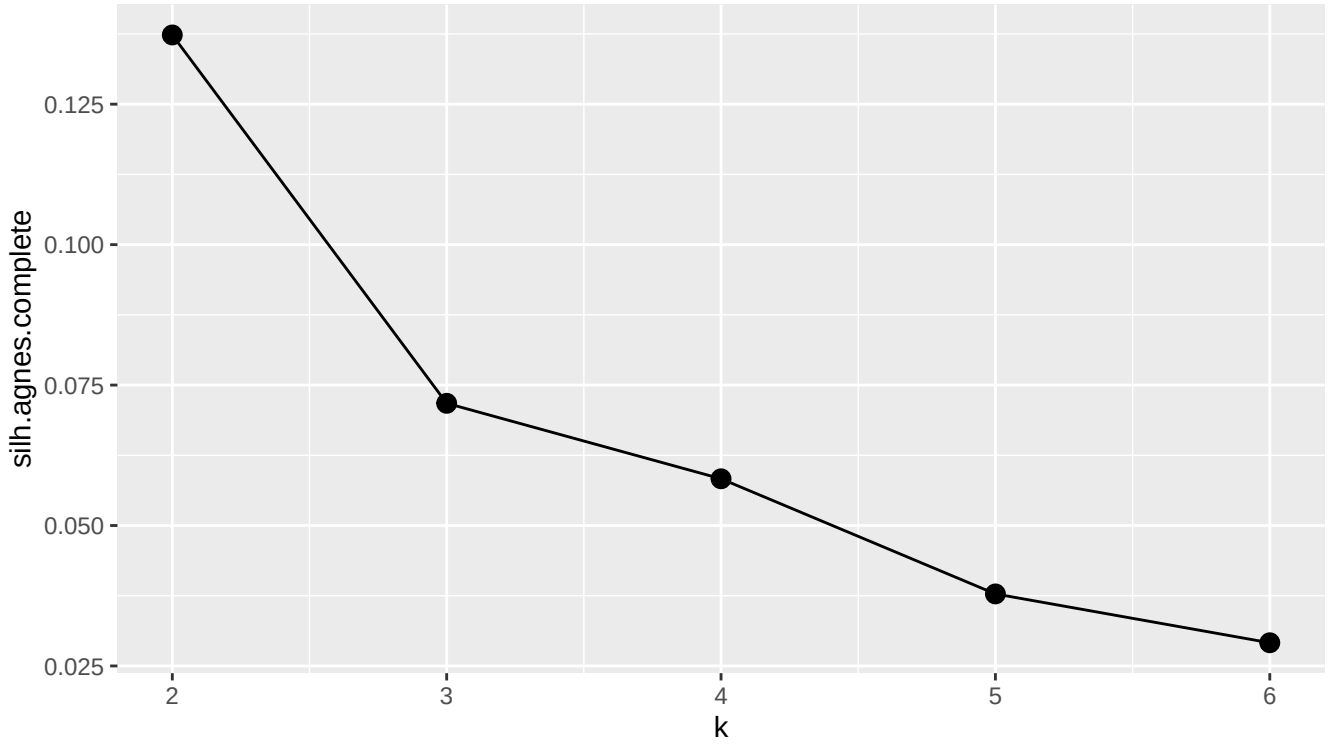
Poniżej przedstawiamy macierz kontyngencji dla $k = 2$.

	1	2
1	642	253
2	58	47

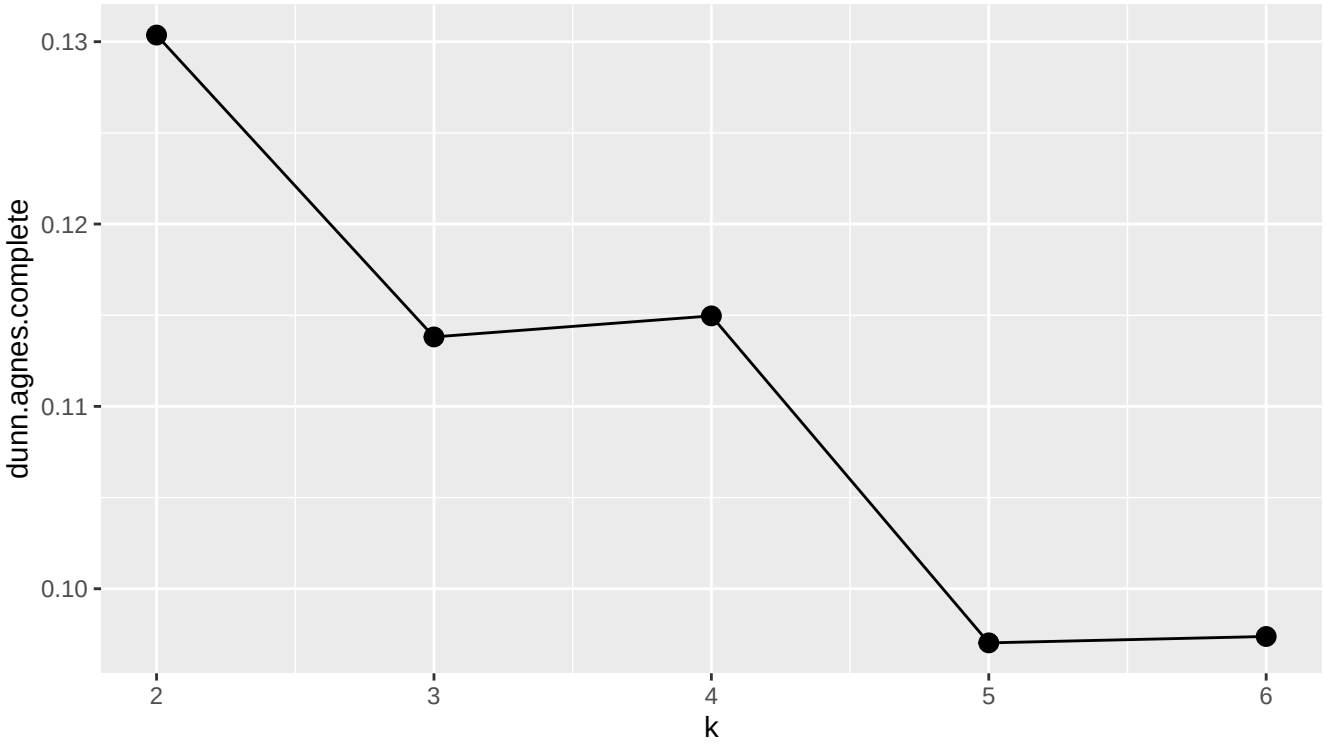
Tabela 24: Macierz kontyngencji - AGNES - complete linkage

```
## Direct agreement: 1 of 2 pairs
## Iterations for permutation matching: 1
## Cases in matched pairs: 68.9 %
## 1 2
## 1 2
```

W przypadku tej metody łączenia uzyskujemy lepsze wyniki niż dla poprzednich. Mniej liczny klaster zawiera 105 obserwacji, a zgodność klas rzeczywistych i przypisanych wynosi 68.9%. Jednak nie zamieszczamy informacji o rozkładzie wartości poszczególnych zmiennych w klastrach, ponieważ podobnie jak w przypadku PAM, trudno jest z nich wysnuć jednoznaczne wnioski.



Rysunek 29: Wartości wskaźnika Silhouette w zależności od liczby klastrów

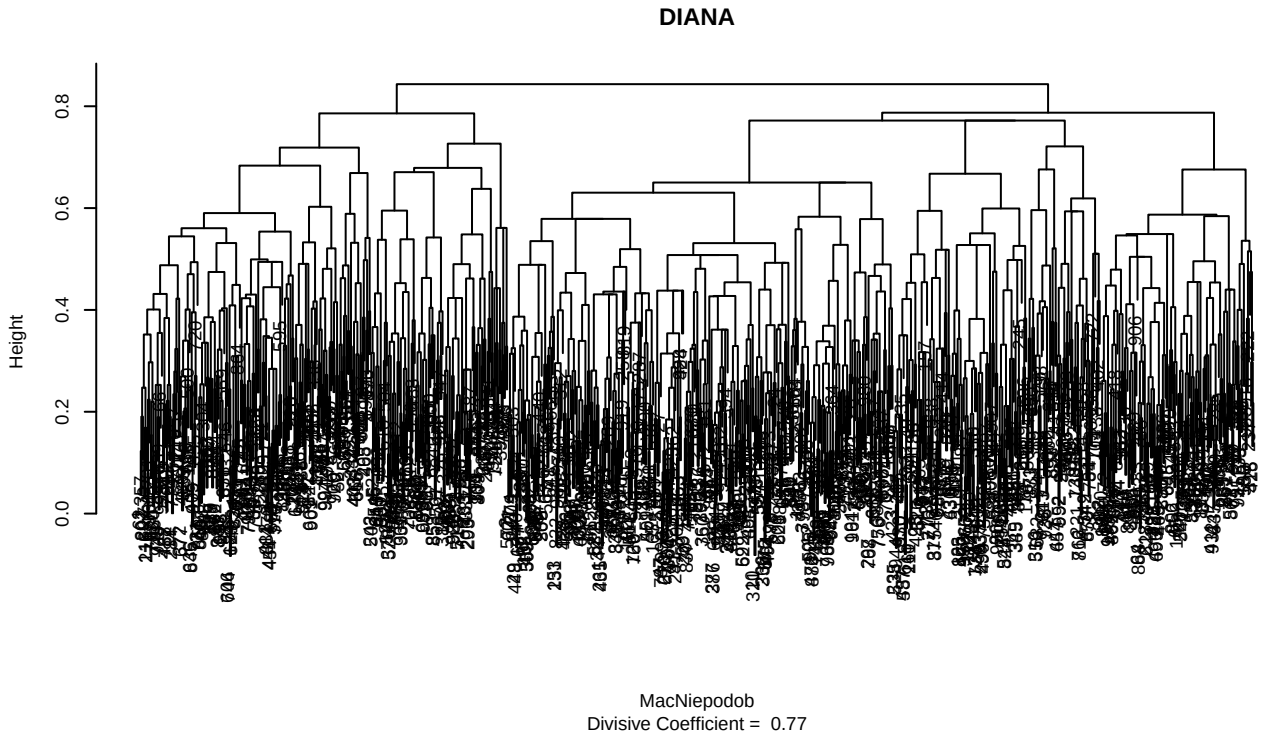


Rysunek 30: Wartości wskaźnika Dunn dla różnej liczby klastrów

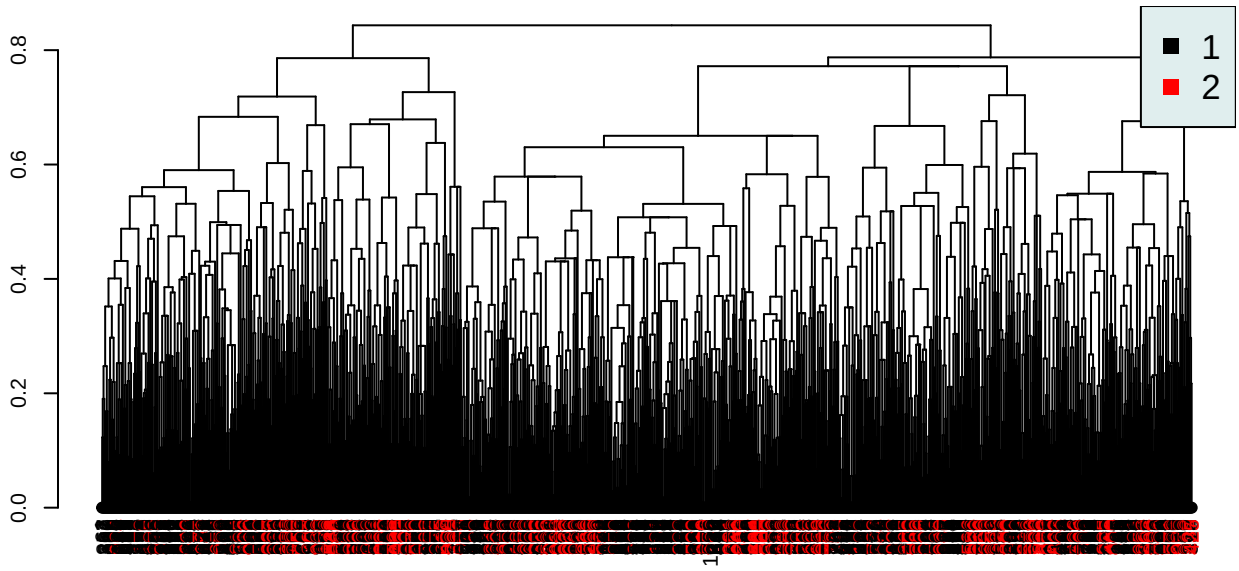
Na Rysunku 29 widzimy, że najwyższą wartość Silhouette otrzymujemy dla $k = 2$. Dla większej liczby klastrów obserwujemy znaczne obniżenie wartości rozważanego wskaźnika. Zastosowanie wyłącznie tej metody sugerowałoby więc, że najlepszym wyborem jest podział na 2 klastry. Również wskaźnik Dunn wskazuje na podział na 2 klastry. Ze względu na występujące w danych zmienne typu factor, nie udało się obliczyć wskaźników oceniających stabilność. Funkcja, która wykorzystana została wcześniej, (clValid) działa bowiem prawidłowo jedynie dla zmiennych numerycznych. Także wyznaczenie tych wskaźników przy użyciu funkcji stability nie powiodło się ze względu na bardzo długi czas obliczeń.

2.4 DIANA

Zastosujemy kolejną metodę hierarchiczną, tym razem dzielącą. Będzie to algorytm DIANA. Rysunek 31 przedstawia dendrogram, zaś na Rysunku 32 zaznaczyliśmy na nim klasy rzeczywiste.

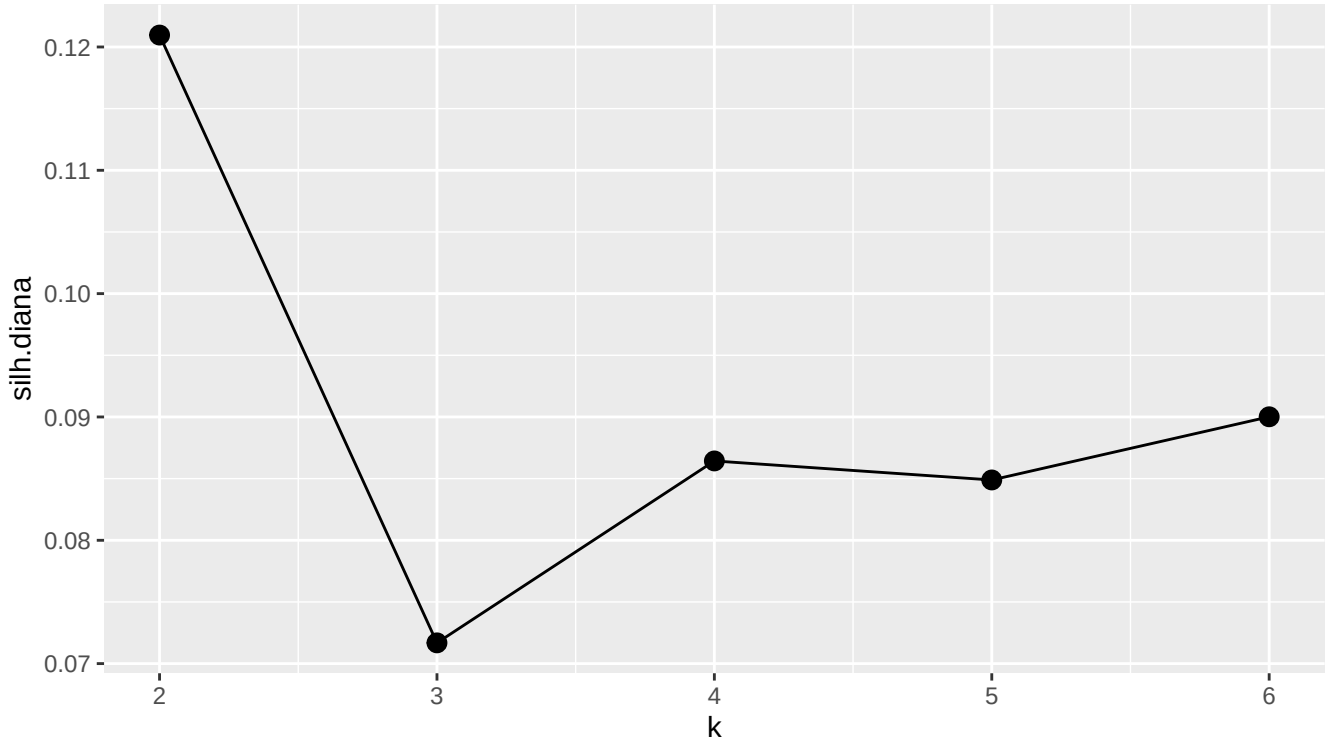


Rysunek 31: Dendrogram - DIANA

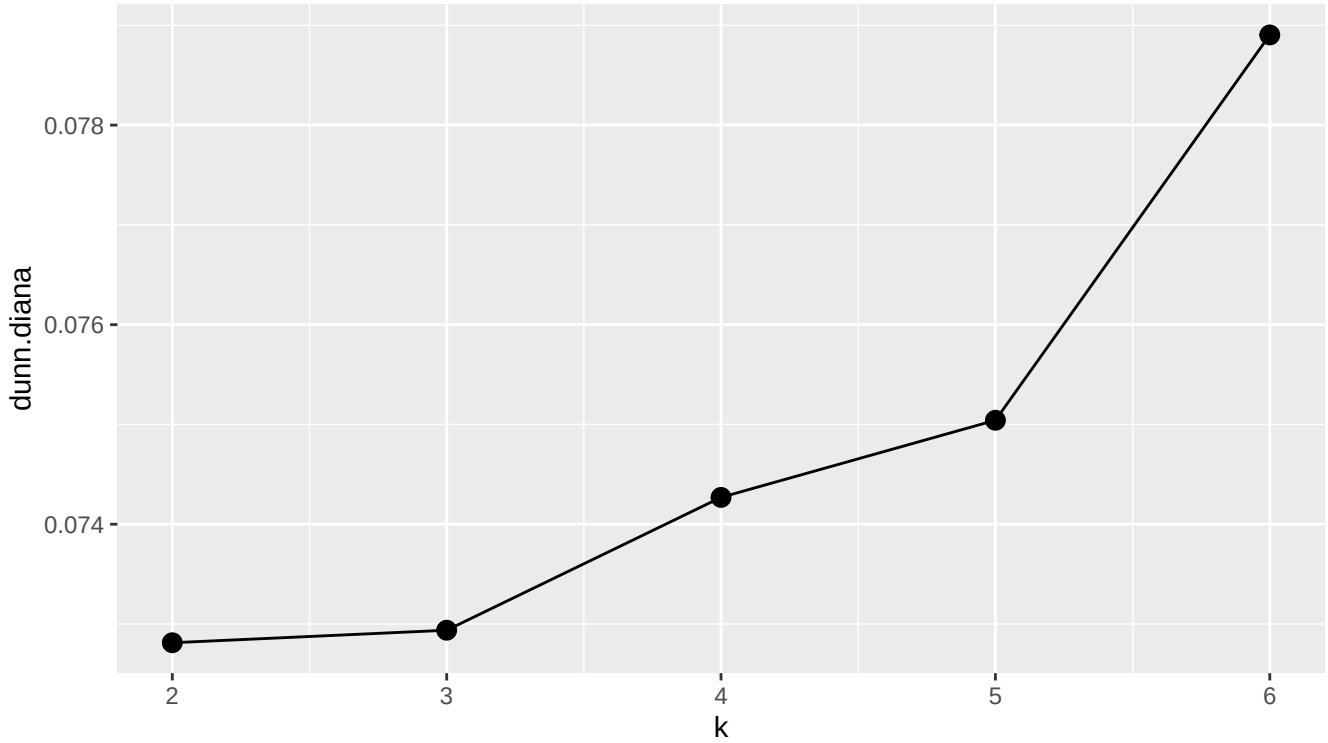


Rysunek 32: Dendrogram - DIANA z etykietkami rzeczywistymi

Na Rysunkach 33 i 34 znajdują się wykresy wartości Silhouette i Dunn w zależności od liczby klastrów.



Rysunek 33: Wartości wskaźnika Silhouette w zależności od liczby klastrów



Rysunek 34: Wartości wskaźnika Dunn dla różnej liczby klastrów

Najwyższą wartość Silhouette obserwujemy dla $k = 2$. Dla $k = 3$ widoczny jest nagły spadek wartości, po czym dla kolejnych k możemy zaobserwować wzrost wartości. W przypadku wskaźnika Dunn, najwyższą jego

wartość mamy dla $k = 6$. Nie możemy jednak wykluczyć, że dla większej liczby klastrów nie uzyskalibyśmy wyższych wyników, gdyż widać tendencję rosnącą.

Tabela 47 przedstawia macierz kontyngencji dla $k = 2$.

	1	2
1	234	96
2	466	204

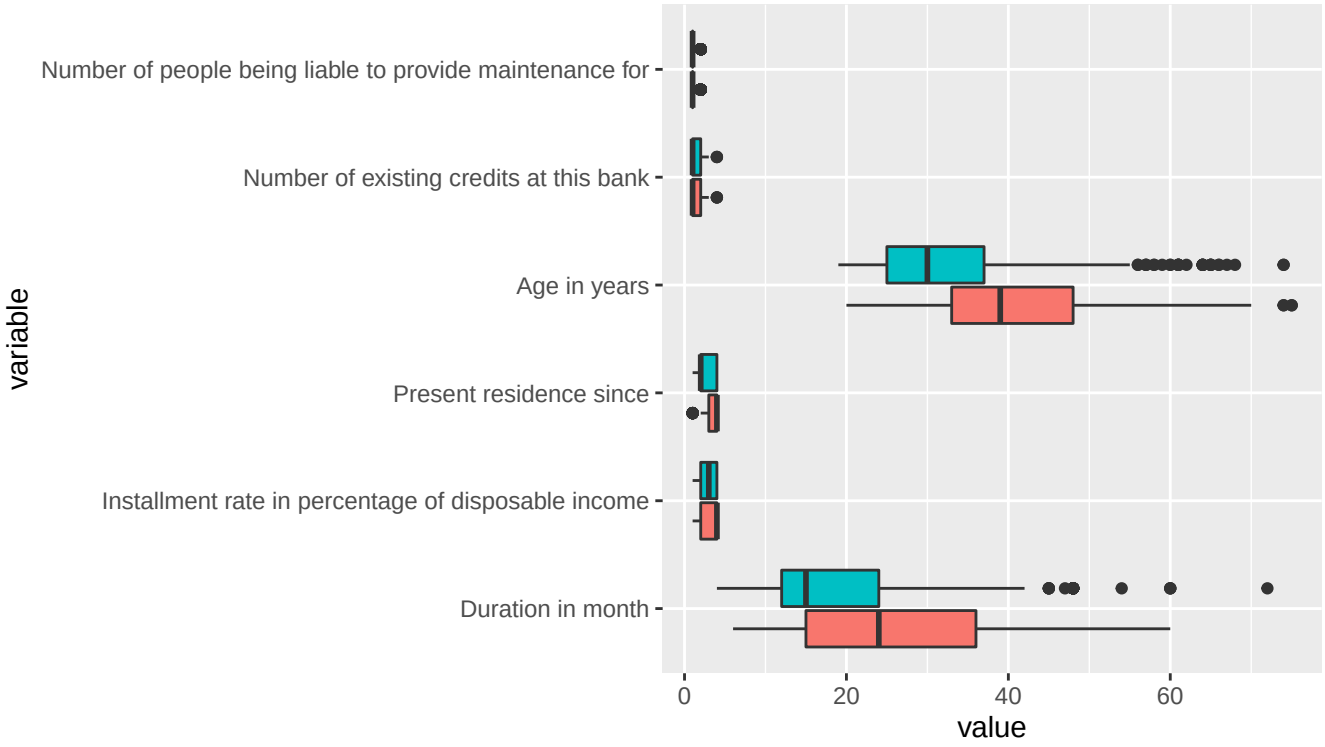
Tabela 25: Macierz kontyngencji - DIANA

```
## Direct agreement: 1 of 2 pairs
## Iterations for permutation matching: 1
## Cases in matched pairs: 56.2 %
## 1 2
## 2 1
```

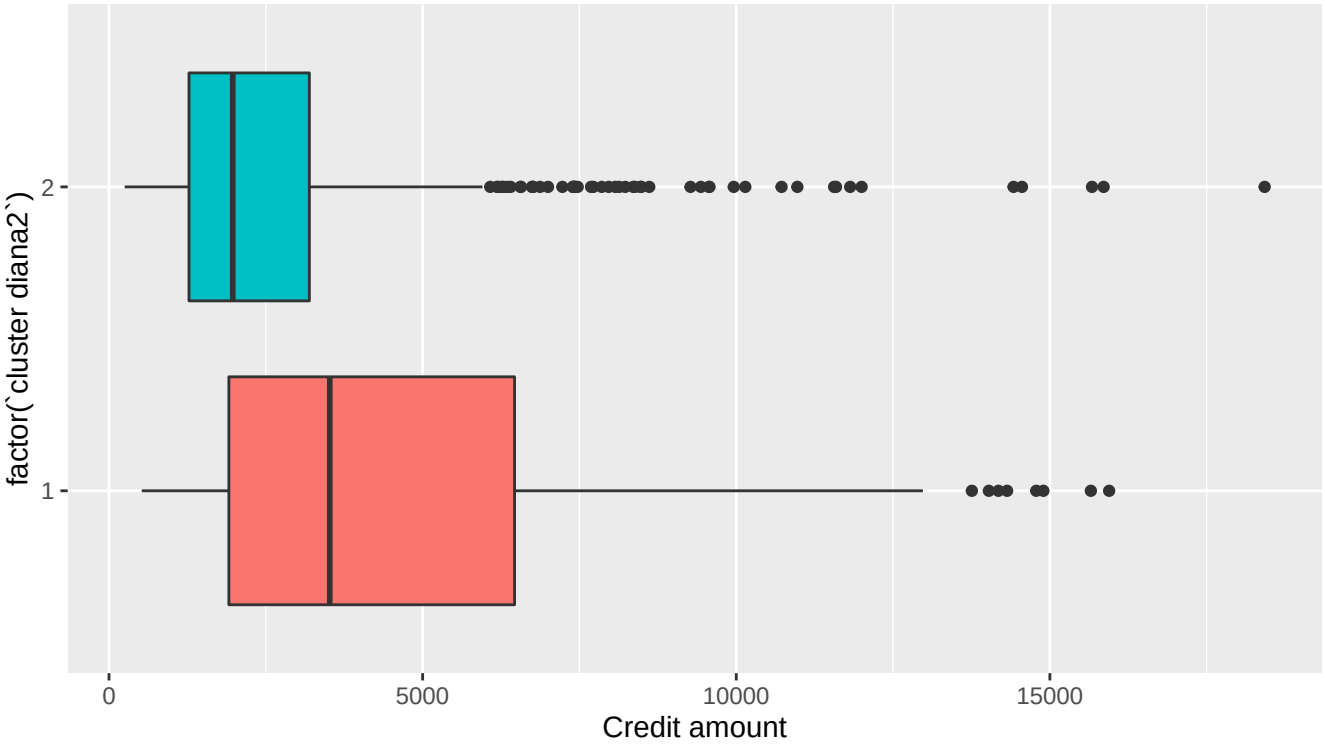
Uzyskany wyżej wynik jest niższy niż wyniki otrzymane dla większości omawianych wcześniej algorytmów. Sprawdzimy, jak rozkładają się wartości poszczególnych zmiennych w klastrach. Tabela 26 zawiera średnie wartości zmiennych, zaś Rysunki 35 i 36 przedstawiają ich box-ploty.

Cluster	Duration in month	Credit amount	Installment rate in percentage of disposable income	Present residence since	Age in years	Number of existing credits at this bank	Number of people being liable to provide maintenance for
1	25.23	4553.26	3.16	3.34	41.10	1.54	1.23
2	18.77	2639.82	2.88	2.60	32.81	1.34	1.12

Tabela 26: Średnie wartości zmiennych numerycznych w klastrach



Rysunek 35: Box-ploty zmiennych numerycznych z podziałem na klastry



Rysunek 36: Box-plot zmiennej Credit amount z podziałem na klastry

Na podstawie powyższych analiz nie widać znacznych różnic w wartościach zmiennych numerycznych w klastrach.

3 Redukcja wymiaru

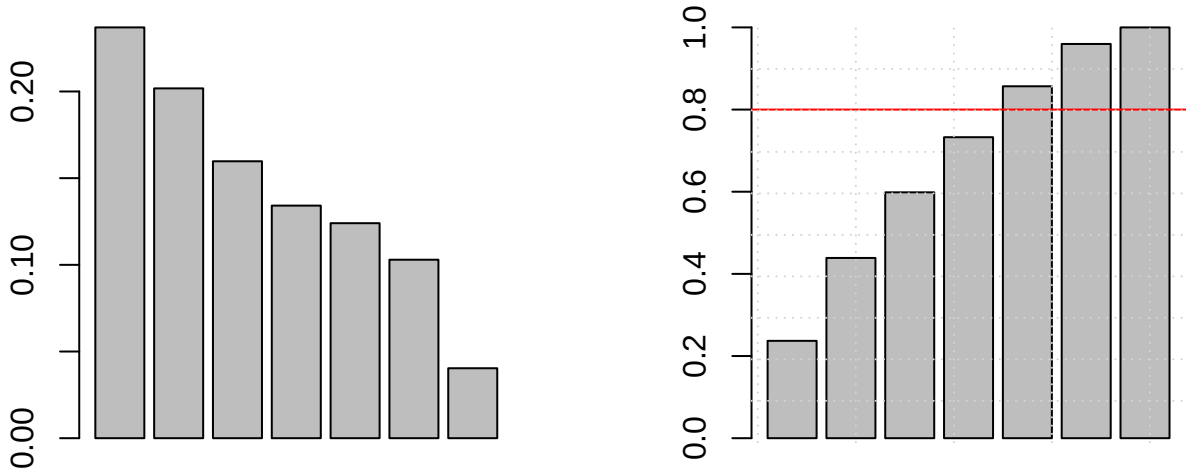
Przejdziemy teraz do redukcji wymiaru. Wykorzystamy w tym celu dwie różne metody: PCA i MDS.

3.1 PCA

Analizę składowych głównych możemy przeprowadzić jedynie dla zmiennych numerycznych. Ponownie więc wykorzystamy jedynie 7 zmiennych, których typ jest określony jako numeric. Z uwagi na duże różnice w wariancji poszczególnych zmiennych, zastosujemy PCA do przeskalowanych danych.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.2879	1.1885	1.0574	0.9691	0.9318	0.8489	0.5315
Proportion of Variance	0.2370	0.2018	0.1597	0.1342	0.1240	0.1029	0.0403
Cumulative Proportion	0.2370	0.4388	0.5985	0.7327	0.8567	0.9597	1.0000

Tabela 27: Udział wyjaśnionej wariancji

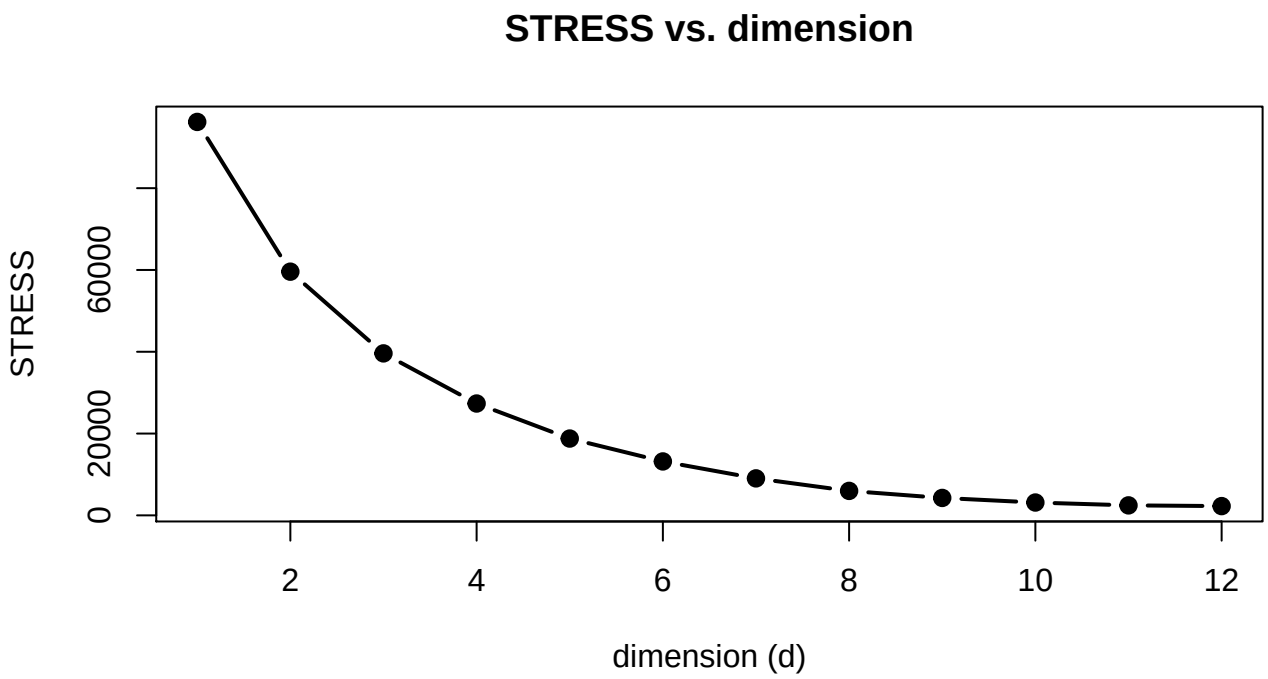


Rysunek 37: Scree plot

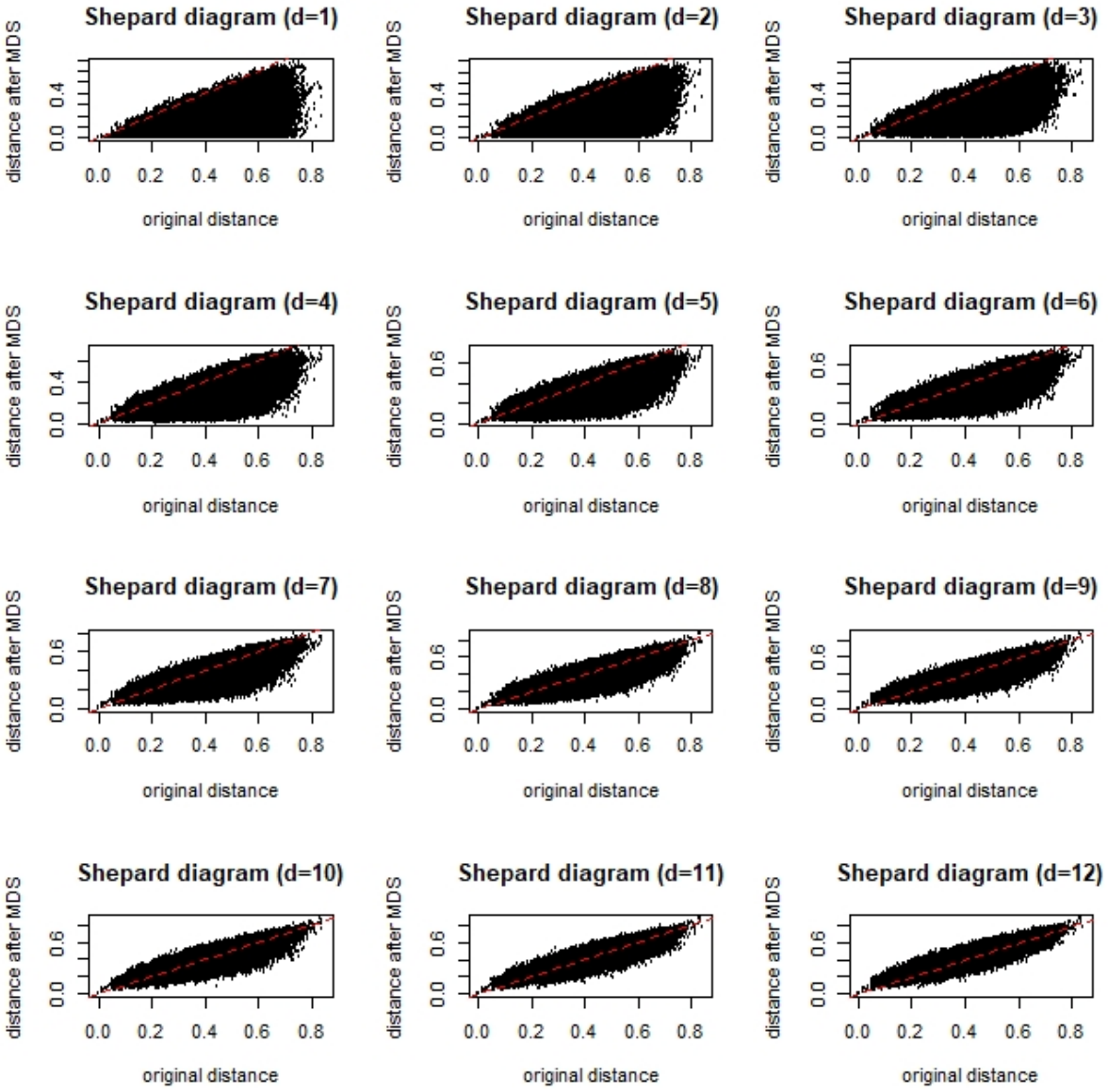
Na Rysunku 37 oraz w Tabeli 27 widzimy, że pięć pierwszych składowych głównych wyjaśnia łącznie ponad 85% całkowitej wariancji.

3.2 MDS

Zastosujemy inną metodę redukcji wymiaru, a mianowicie skalowanie wielowymiarowe. Zaletą tej metody jest to, że możemy zastosować ją także do zmiennych mieszanych typów, wyznaczając macierz odmienności. Do oceny jakości MDS oraz wybrania wymiaru zastosujemy kryterium STRESS i diagram Sheparda. Ich zależność od wymiaru została przedstawiona odpowiednio na Rysunkach 38 i 39.



Rysunek 38: Wartość kryterium STRESS w zależności od wymiaru



Rysunek 39: Diagramy Sheparda dla różnych wymiarów

Kierując się momentem, w którym wykres wartości STRESS zaczyna się wypłaszczać, oraz diagramami Sheparda, wybieramy $k = 8$.

4 Klasyfikacja po MDS

Ponieważ analizowane dane są mieszanego typu, nie będziemy prowadzić dalszej analizy danych otrzymanych po PCA. Dane, które otrzymujemy po zastosowaniu MDS, wykorzystamy jako dane wejściowe w klasyfikacji. Podobnie jak w pierwszej części projektu, do danych dopasujemy następujące modele: regresji logistycznej, regresji logistycznej z metodą krokową, regresji logistycznej z kategoryzacją, liniowej analizy dyskryminacyjnej i kwadratowej analizy dyskryminacyjnej. Ponieważ znacznie lepsze były modele, które uczyniliśmy cost-sensitive za pomocą zmiany thresholdu (na $p = 1/6$), tym razem nie będziemy już budować modeli cost-insensitive.
W celu uzyskania wiarygodniejszych wyników, dane podzielimy na dwa zbiory: treningowy i testowy. Najpierw sprawdzimy działanie modeli na zbiorze treningowym, przeprowadzając w tym celu 5-krotną walidację krzyżową. Następnie, najlepsze modele zbudujemy dla całego zbioru treningowego i porównamy wyniki na zbiorze

testowym. Ponieważ analizowane dane mają nierówny rozkład klas, zastosujemy tzw. stratified sampling. W każdej iteracji walidacji wyznaczymy wskaźniki modelu takie jak: dokładność(ACC), czułość(TPR), specyficzność(TNR), F1 oraz średni koszt złej klasyfikacji(MMC).

Wyniki zostały przedstawione w Tabelach 28-32.

	ACC	TPR	TNR	F1	MMC
1	0.62	0.86	0.52	0.61	0.54
2	0.57	0.89	0.44	0.68	0.56
3	0.55	0.91	0.41	0.71	0.55
4	0.60	0.89	0.48	0.65	0.53
5	0.63	0.93	0.51	0.64	0.45
średnia	0.60	0.89	0.47	0.66	0.53

Tabela 28: Tabela wskaźników dla 5-fold cross-validation w modelu regresji logistycznej z p=1/6 po MDS

	ACC	TPR	TNR	F1	MMC
1	0.64	0.79	0.58	0.55	0.60
2	0.55	0.82	0.44	0.66	0.66
3	0.63	0.86	0.54	0.60	0.53
4	0.60	0.89	0.48	0.65	0.53
5	0.63	0.79	0.56	0.56	0.61
średnia	0.61	0.83	0.52	0.61	0.59

Tabela 29: Tabela wskaźników dla 5-fold cross-validation w modelu regresji logistycznej z kategoryzacją z p=1/6 po MDS

	ACC	TPR	TNR	F1	MMC
1	0.62	0.86	0.52	0.61	0.54
2	0.57	0.89	0.44	0.68	0.56
3	0.55	0.91	0.41	0.71	0.55
4	0.60	0.89	0.48	0.65	0.53
5	0.63	0.88	0.52	0.62	0.51
średnia	0.59	0.88	0.48	0.66	0.54

Tabela 30: Tabela wskaźników dla 5-fold cross-validation w modelu regresji logistycznej z metodą krokową z p=1/6 po MDS

	ACC	TPR	TNR	F1	MMC
1	0.60	0.86	0.50	0.64	0.56
2	0.55	0.89	0.41	0.71	0.59
3	0.54	0.91	0.39	0.73	0.57
4	0.59	0.89	0.47	0.66	0.54
5	0.61	0.93	0.49	0.66	0.47
średnia	0.58	0.89	0.45	0.68	0.54

Tabela 31: Tabela wskaźników dla 5-fold cross-validation w LDA z p=1/6 po MDS

	ACC	TPR	TNR	F1	MMC
1	0.61	0.79	0.54	0.58	0.63
2	0.59	0.89	0.47	0.66	0.54
3	0.58	0.86	0.47	0.66	0.58
4	0.57	0.82	0.47	0.64	0.64
5	0.62	0.84	0.53	0.60	0.57
średnia	0.60	0.84	0.50	0.63	0.59

Tabela 32: Tabela wskaźników dla 5-fold cross-validation w modelu QDA z p=1/6 po MDS

Wszystkie modele rokują dość dobrze. W każdym z przypadków średnia wartość $F1$ jest wyższa niż 0.5. Dla żadnej z metod nie uzyskujemy MMC wyższego niż 0.6. Także czułość jest na wysokim poziomie - dla każdego z modeli powyżej 0.8.
Teraz wszystkie modele zbudujemy ponownie, tym razem dla całego zbioru treningowego i porównamy ich skuteczność na zbiorze testowym. Aby skontrolować, czy modele nie są przeuczone, sprawdzimy również jak wygląda ich predykcja na zbiorze treningowym. Wyniki znajdują się w Tabelach 33-37.

	ACC	TPR	TNR	F1	MMC
test	0.61	0.87	0.48	0.65	0.57
train	0.58	0.89	0.46	0.68	0.54

Tabela 33: Tabela wskaźników dla LR z p=1/6 po MDS

	ACC	TPR	TNR	F1	MMC
test	0.62	0.88	0.50	0.64	0.54
train	0.61	0.89	0.50	0.64	0.52

Tabela 34: Tabela wskaźników dla LR(category) z p=1/6 po MDS

	ACC	TPR	TNR	F1	MMC
test	0.61	0.87	0.48	0.65	0.57
train	0.58	0.89	0.46	0.68	0.54

Tabela 35: Tabela wskaźników dla LR(step) z p=1/6 po MDS

	ACC	TPR	TNR	F1	MMC
test	0.60	0.88	0.46	0.67	0.56
train	0.58	0.90	0.45	0.69	0.54

Tabela 36: Tabela wskaźników dla LDA z p=1/6 po MDS

	ACC	TPR	TNR	F1	MMC
test	0.60	0.87	0.47	0.66	0.57
train	0.60	0.88	0.49	0.65	0.53

Tabela 37: Tabela wskaźników dla QDA z p=1/6 po MDS

Analizując wyniki zawarte w Tabelach 33-37 nie widzimy znacznych różnic pomiędzy wartościami wskaźników dla zbioru testowego oraz treningowego. Oznacza to, że modele nie są przeuczone.
Zbiorcze porównanie wyników na zbiorze testowym przedstawia Tabela 38.

	ACC	TPR	TNR	F1	MMC
LR	0.61	0.87	0.48	0.65	0.57
LR(category)	0.62	0.88	0.50	0.64	0.54
LR(step)	0.61	0.87	0.48	0.65	0.57
LDA	0.60	0.88	0.46	0.67	0.56
QDA	0.60	0.87	0.47	0.66	0.57

Tabela 38: Tabela wskaźników dla wszystkich modeli na zbiorze testowym z p =1/6 po MDS

Wskaźniki F1 i MMC są bardzo zbliżone dla wszystkich metod. Najniższy MMC otrzymujemy dla regresji logistycznej z kategoryzacją - wynosi on 0.54. Najwyższą wartość F1, równą 0.67, uzyskujemy dla liniowej analizy dyskryminacyjnej.
Wyniki, które otrzymaliśmy w pierwszej części projektu, dla danych bez zastosowania redukcji wymiaru, znajdują się w Tabeli 39.

	ACC	TPR	TNR	F1	MMC
LR	0.61	0.77	0.53	0.58	0.69
LR(category)	0.65	0.87	0.54	0.60	0.53
LR(step)	0.66	0.81	0.59	0.55	0.60
LDA	0.61	0.86	0.49	0.64	0.58
QDA	0.67	0.77	0.62	0.51	0.64

Tabela 39: Tabela wskaźników dla wszystkich modeli na zbiorze testowym z p =1/6

Porównując dwie powyższe tabele widzimy, że po zastosowaniu MDS poprawił się wskaźnik F1 w przypadku wszystkich modeli klasyfikacyjnych. W większości przypadków mamy też niższe wartości MMC. Nieznacznie wyższą wartość otrzymujemy dla modelu regresji liniowej z kategoryzacją. Po zastosowaniu MDS wynosi on 0.54, zaś przed redukcją wymiaru był równy 0.53. Z modeli budowanych przed MDS najwyższe F1 miała liniowa analiza dyskryminacyjna, zaś najniższy MMC - regresja liniowa z kategoryzacją. Wśród modeli stworzonych po MDS jest tak samo.

5 Klasteryzacja po MDS

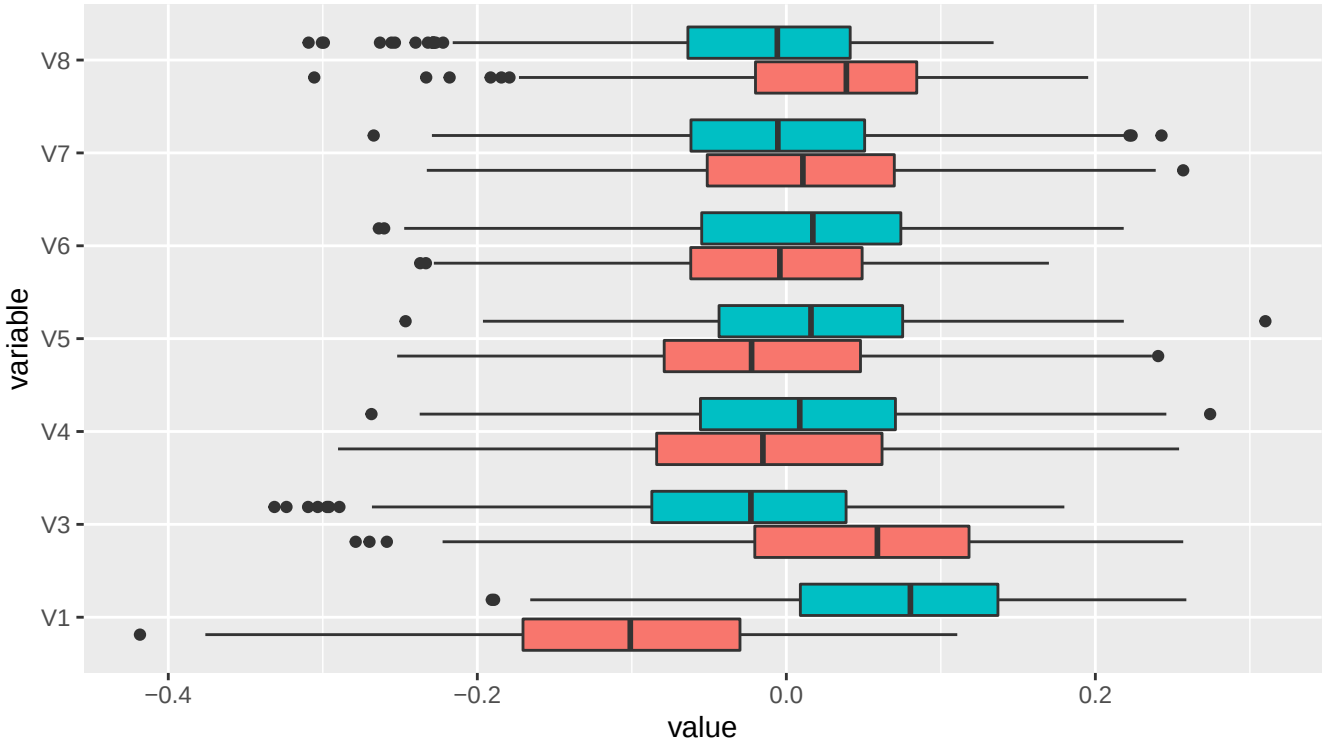
Przeprowadzimy analizę skupień dla danych po zastosowaniu MDS. Mając na uwadze to, że w tym przypadku wartości zmiennych są numeryczne, ocenę jakości grupowania przedstawimy na sam koniec, uwzględniając wszystkie z omawianych algorytmów oraz różne wartości k.

5.1 k-means

Sprawdzimy jedynie jakie wyniki otrzymujemy dla k=2. Przeanalizujemy wartości zmiennych po MDS w zależności od przypisanego klastra. Tabela 40 zawiera średnie wartości zmiennych, zaś Rysunek 40 przedstawia ich box-ploty.

Cluster	V1	V2	V3	V4	V5	V6	V7	V8
1	-0.10	0.05	0.04	-0.01	-0.02	-0.01	0.01	0.03
2	0.07	-0.04	-0.03	0.01	0.01	0.01	-0.01	-0.02

Tabela 40: Średnie wartości zmiennych po MDS w klastrach



Rysunek 40: Box-ploty zmiennych po MDS z podziałem na klastry

Z tabeli możemy odczytać, że największą różnicę w średnich wartościach zmiennych w klastrach mamy dla V1. Dla pozostałych zmiennych różnice tę są mniejsze niż 0.1. Potwierdzenie wyników z tabeli otrzymujemy na box-plotach. Widzimy, że zakresy przyjmowane przez zmienną V1 są wyraźnie różne w obu klastrach (w jednym są znacznie wyższe niż w drugim). Podobną sytuację obserwowaliśmy przed zastosowaniem MDS. Wówczas największe różnice miały miejsce dla zmiennych Credit Amount i Durarion in Month.

	0	1
1	405	186
2	295	114

Tabela 41: Macierz kontyngencji - k-means

Z powyższej tabeli możemy odczytać, że w pierwszym klastrze jest 591 obserwacji, w drugim natomiast 409.

```
## Direct agreement: 0 of 2 pairs
## Iterations for permutation matching: 2
## Cases in matched pairs: 51.9 %
##   1   2
## "0" "1"
```

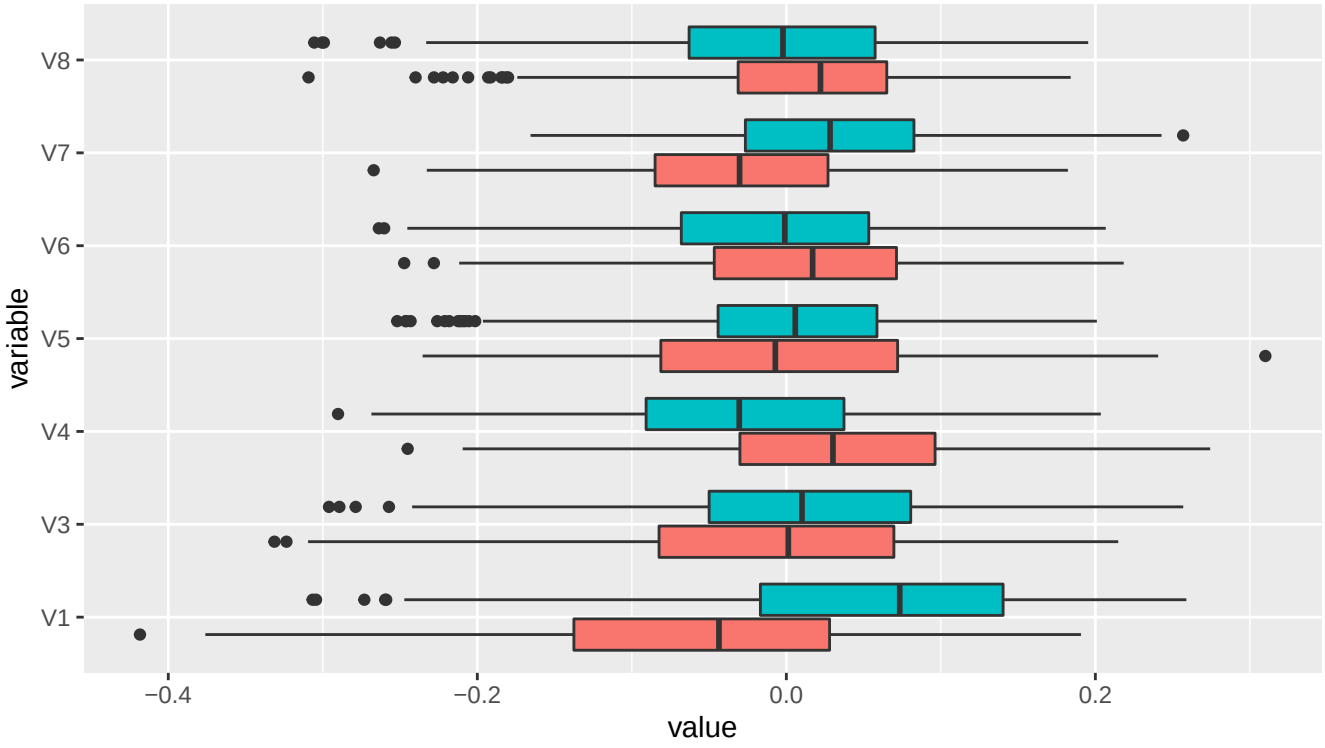
Wynik, który otrzymaliśmy przy użyciu funkcji matchClasses jest niższy od tego, który uzyskaliśmy dla algorytmu 2-means dla danych przeskalowanych, przed redukcją wymiaru.

5.2 PAM

Sprawdzimy jedynie jakie wyniki otrzymujemy dla k=2. Przeanalizujemy wartości zmiennych po MDS w zależności od przypisanego klastra. Tabela 42 zawiera średnie wartości zmiennych, zaś Rysunek 41 przedstawia ich box-ploty.

Cluster	V1	V2	V3	V4	V5	V6	V7	V8
1	-0.06	-0.05	-0.01	0.03	-0.00	0.01	-0.03	0.01
2	0.05	0.05	0.01	-0.03	0.00	-0.01	0.03	-0.01

Tabela 42: Średnie wartości zmiennych po MDS w klastrach



Rysunek 41: Box-ploty zmiennych po MDS z podziałem na klastry

Podobnie jak w przypadku algorytmu k-means, największe różnice w średnich wartościach zmiennych otrzymujemy dla V1. Jednak teraz wartości tej zmiennej nie różnią się tak wyraźnie. Zakresy wartości pozostałych zmiennych niewiele różnią się pomiędzy klastrami. Przed zastosowaniem redukcji wymiaru również nie było znacznych różnic w zakresie wartości zmiennych w poszczególnych klastrach.

	0	1
1	377	113
2	323	187

Tabela 43: Macierz kontyngencji - PAM

Do pierwszego klastra zaliczono 490 obserwacji, zaś do drugiego 510.

```
## Direct agreement: 0 of 2 pairs
## Iterations for permutation matching: 2
## Cases in matched pairs: 56.4 %
##   1   2
## "0" "1"
```

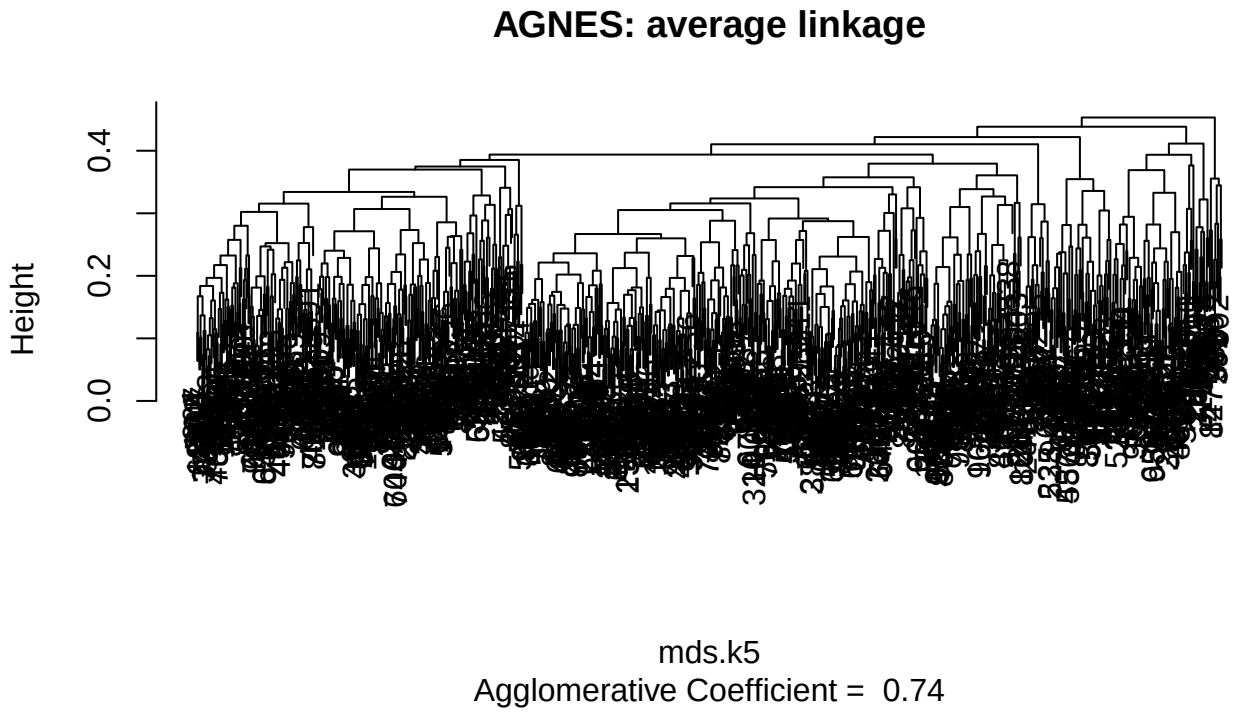
Procent zgodnych klas rzeczywistych i przypisanych wynosi $56,4\%$.

5.3 AGNES

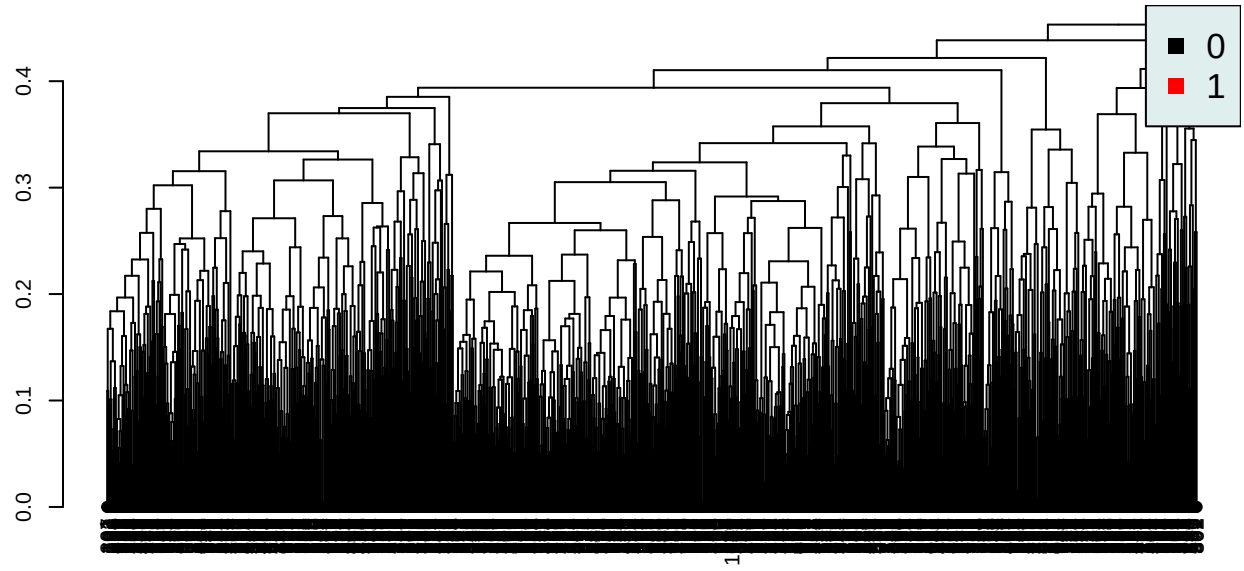
Zastosujemy teraz metodę AGNES dla różnych metod łączenia klastrów z $k = 2$.

5.3.1 average linkage

Rysunek 42 przedstawia dendrogram, zaś na Rysunku 43 zaznaczyliśmy klasy rzeczywiste na dendrogramie.



Rysunek 42: Dendrogram - AGNES - average linkage



Rysunek 43: Dendrogram - AGNES - average linkage z etykietkami rzeczywistymi

Poniżej przedstawiamy macierz kontyngencji dla $k = 2$.

	0	1
1	693	294
2	7	6

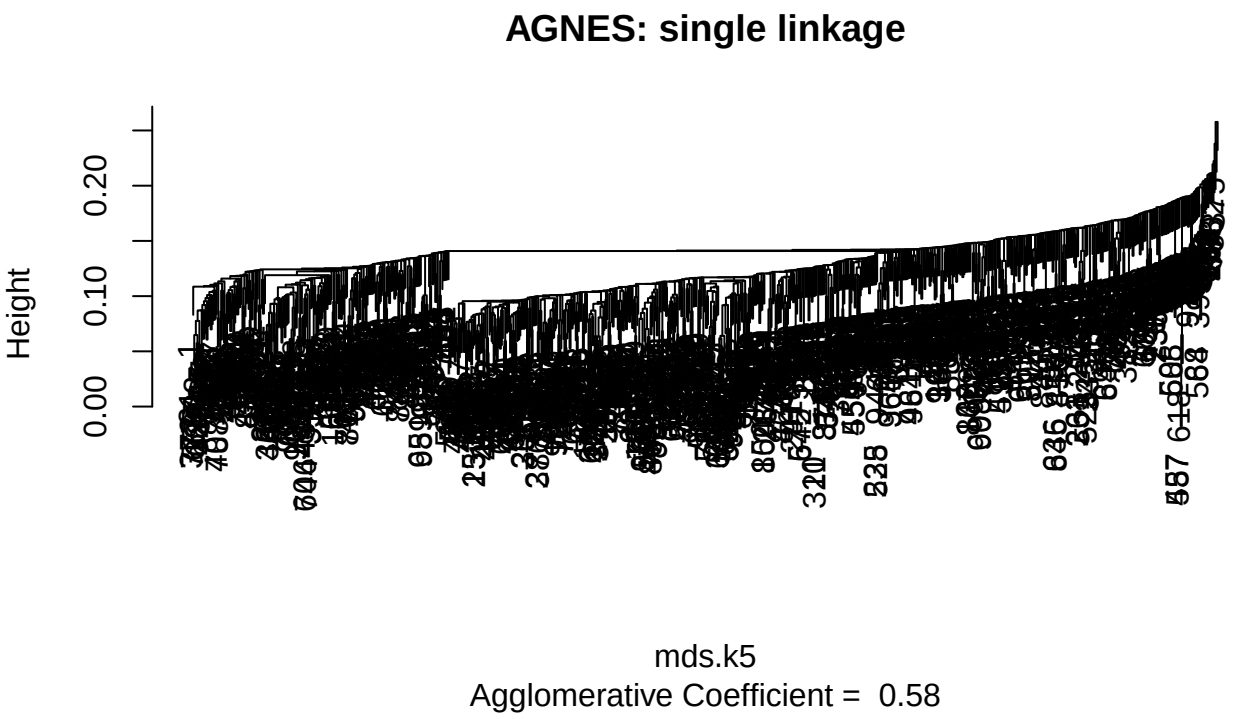
Tabela 44: Macierz kontyngencji - AGNES - average linkage

```
## Direct agreement: 1 of 2 pairs
## Iterations for permutation matching: 1
## Cases in matched pairs: 69.9 %
##   1   2
## "0" "1"
```

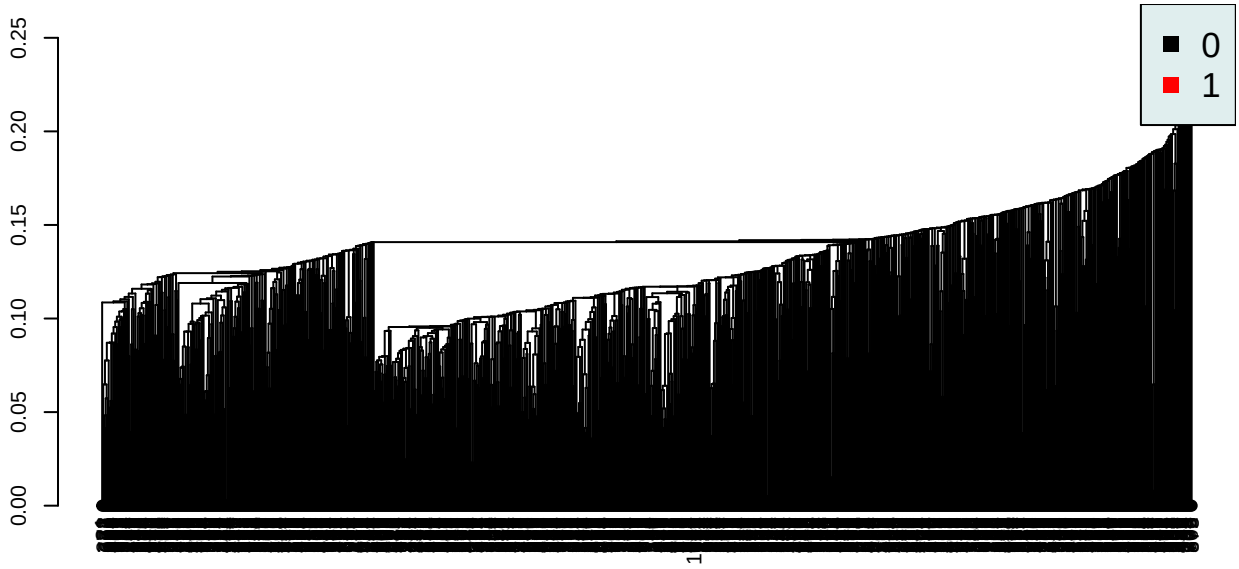
Podobnie jak przypadku klasteryzacji przed zastosowaniem MDS, jeden z klastrów jest mało liczny. Tym razem jest w nim zaledwie 13 obserwacji.

5.3.2 single linkage

Podobnie jak poprzednio, rysujemy dendrogram. Znajduje się on na Rysunku 44. Z kolei Rysunek 45 przedstawia dendrogram z zaznaczonymi klasami rzeczywistymi.



Rysunek 44: Dendrogram - AGNES - single linkage



Rysunek 45: Dendrogram - AGNES - single linkage z etykietkami rzeczywistymi

Poniżej przedstawiamy macierz kontyngencji dla $k = 2$.

	0	1
1	700	299
2	0	1

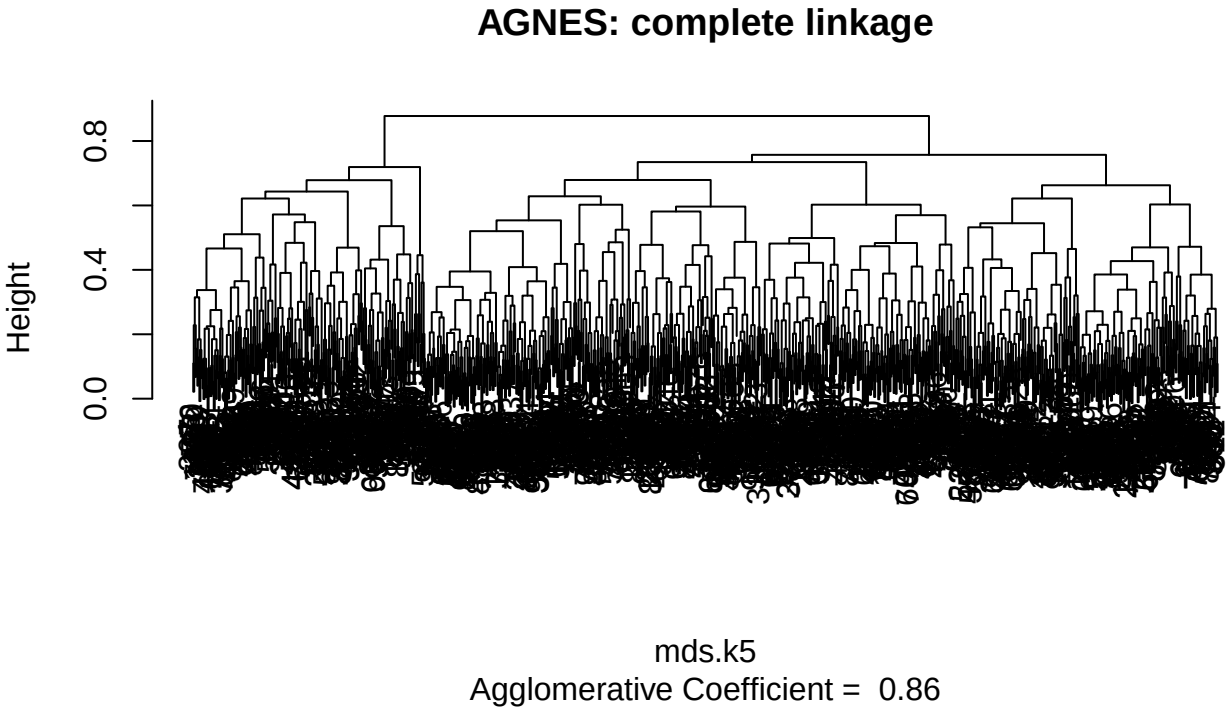
Tabela 45: Macierz kontyngencji - AGNES - single linkage

```
## Direct agreement: 1 of 2 pairs
## Iterations for permutation matching: 1
## Cases in matched pairs: 70.1 %
##   1   2
## "0" "1"
```

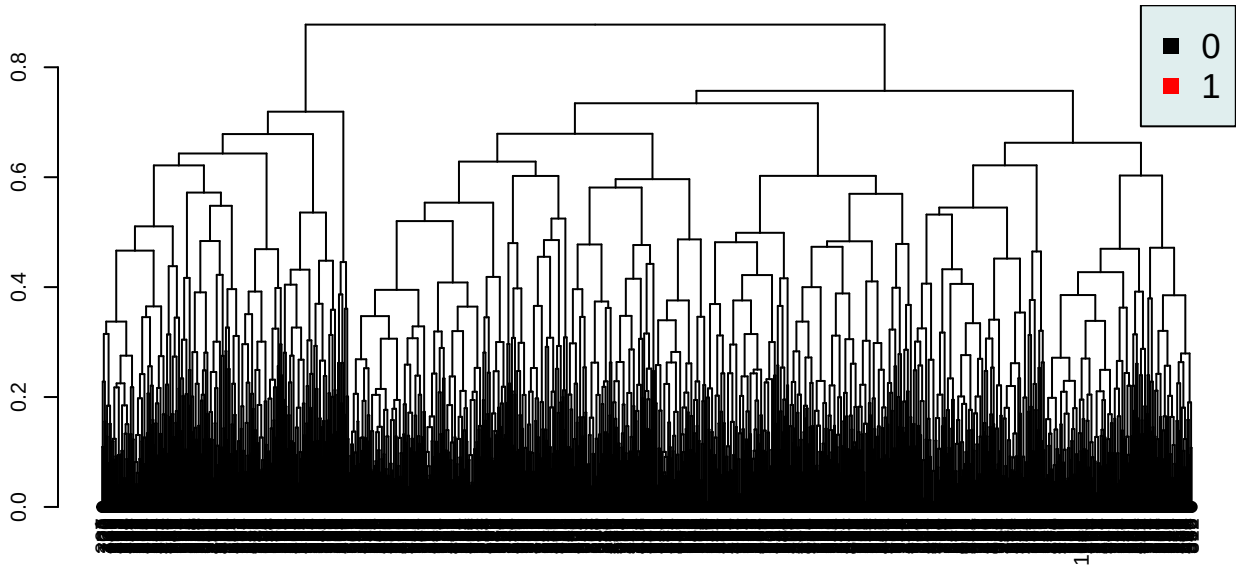
Podobnie jak przy poprzedniej metodzie łączenia, jeden z klastrów jest bardzo mało liczny - jest tam zaledwie jedna obserwacja.

5.3.3 complete linkage

Rysunek 46 przedstawia dendrogram. Na Rysunku 47 zaznaczyliśmy na dendrogramie klasy rzeczywiste.



Rysunek 46: Dendrogram - AGNES - complete linkage



Rysunek 47: Dendrogram - AGNES - complete linkage z etykietkami rzeczywistymi

Tabela 46 przedstawia macierz kontyngencji dla $k = 2$.

	0	1
1	151	76
2	549	224

Tabela 46: Macierz kontyngencji - AGNES - complete linkage

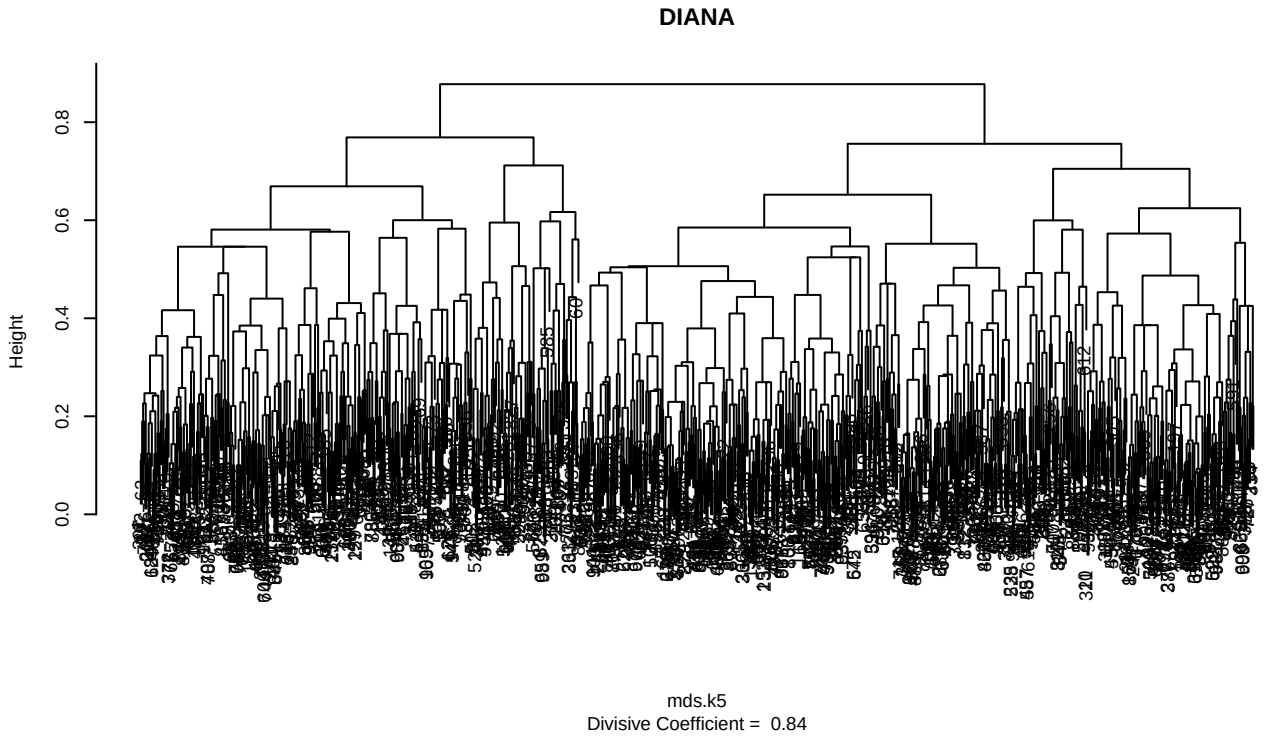
Z macierzy kontyngencji, możemy wyliczyć że do klastra pierwszego przypisano 227 obserwacji, a do drugiego 773. Mamy tu więc bardziej sensowny podział niż dla pozostałych metod łączenia .

```
## Direct agreement: 1 of 2 pairs
## Iterations for permutation matching: 1
## Cases in matched pairs: 62.5 %
##   1   2
## "1" "0"
```

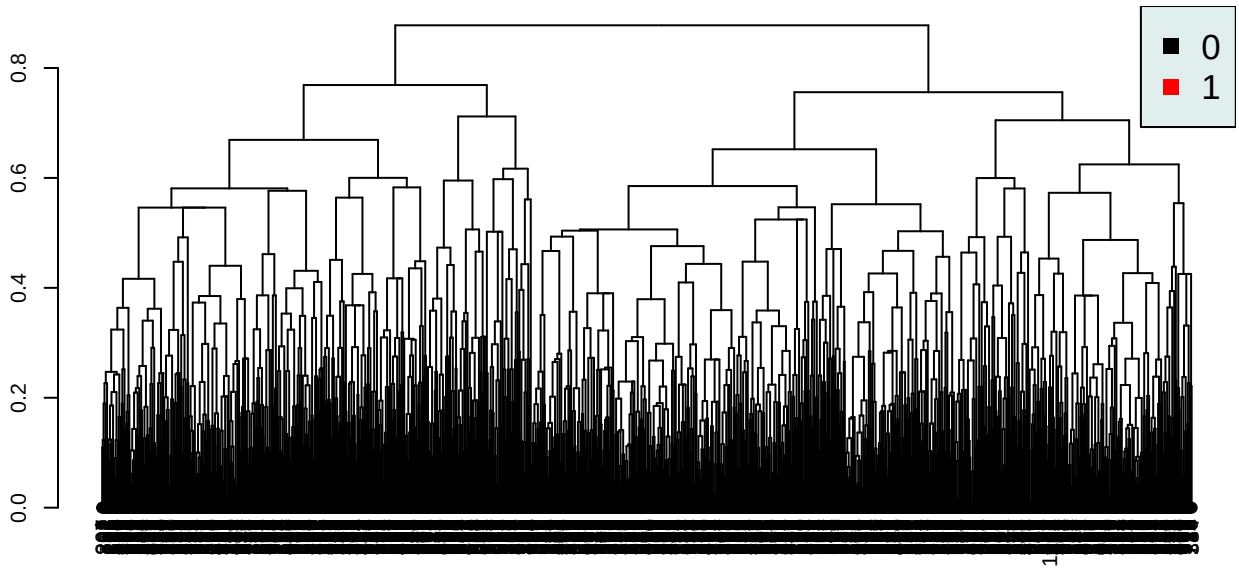
Procent zgodnych klas wynosi 62.5%.

5.4 DIANA

Na Rysunkach 48 i 49 znajdują się odpowiednio: dendrogram i dendrogram z klasami rzeczywistymi.



Rysunek 48: Dendrogram - DIANA



Rysunek 49: Dendrogram - DIANA z etykietkami rzeczywistymi

	0	1
1	281	113
2	419	187

Tabela 47: Macierz kontyngencji - DIANA

```
## Direct agreement: 0 of 2 pairs
## Iterations for permutation matching: 2
## Cases in matched pairs: 53.2 %
##   1   2
## "1" "0"
```

Po zastosowaniu algorytmu DIANA otrzymujemy dopasowanie na poziomie 53.2%. Jest to słabszy wynik niż np. ten uzyskany metodą AGNES z complete linkage.

5.5 Ocena

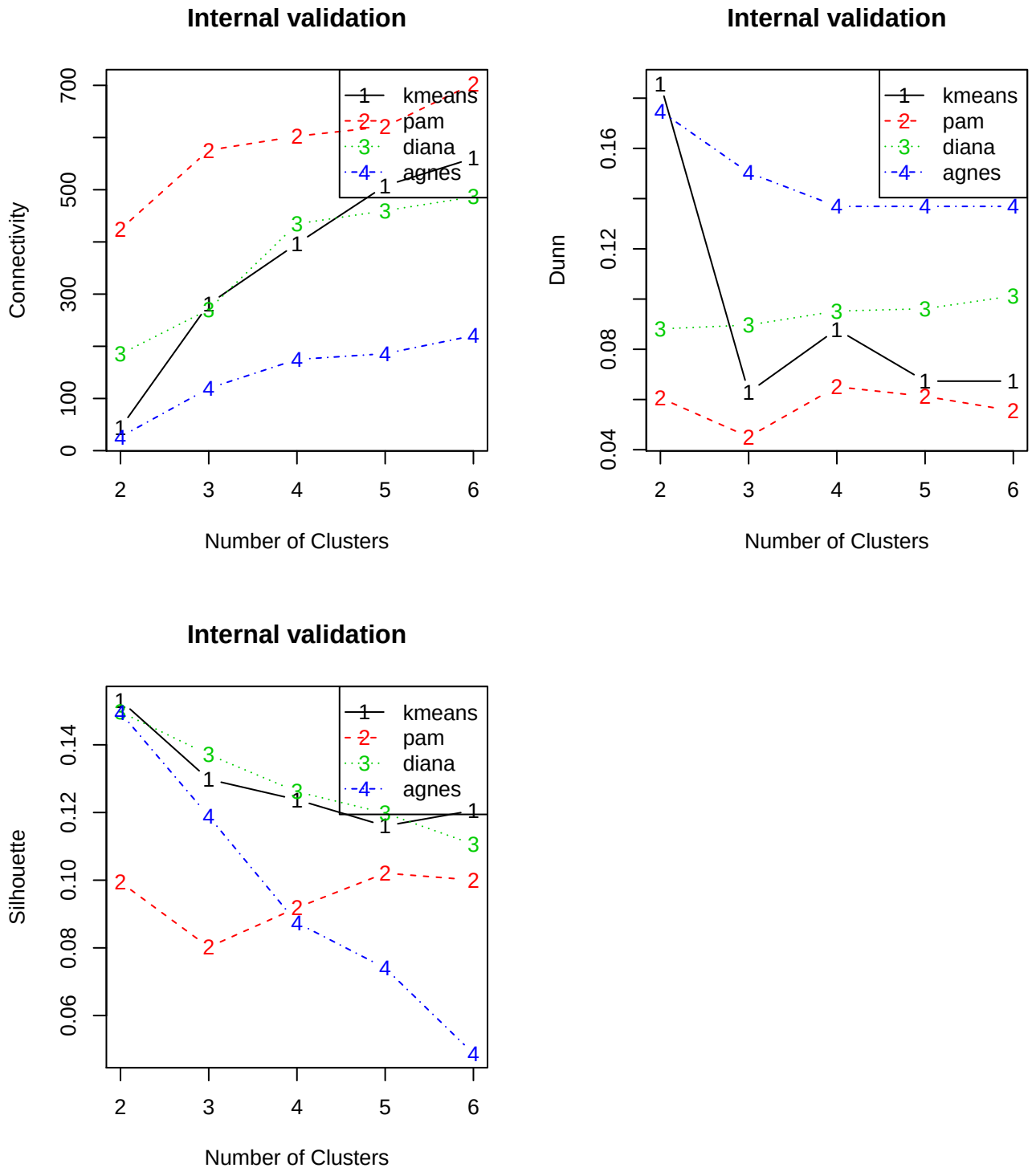
Przedstawimy teraz uzyskane za pomocą funkcji clValid wartości wybranych kryteriów dla metod k-means, PAM, DIANA, AGNES (average linkage) dla wartości $k = 2, \ldots, 6$. W Tabeli 48 znajdują się wskaźniki Connectivity, Dunn i Silhouette. Tabela 49 przedstawia optymalną według danego kryterium liczbę klastrów i metodę. Ilustracja otrzymanych wyników znajduje się na Rysunku 50.

method	measure	k=2	k=3	k=4	k=5	k=6
kmeans	Connectivity	43.692	280.874	396.458	507.610	560.499
	Dunn	0.186	0.063	0.088	0.067	0.067
	Silhouette	0.153	0.130	0.124	0.116	0.121
pam	Connectivity	424.279	575.714	602.525	621.119	703.004
	Dunn	0.061	0.045	0.065	0.061	0.055
	Silhouette	0.099	0.080	0.092	0.102	0.100
diana	Connectivity	185.009	270.452	434.645	459.835	486.373
	Dunn	0.088	0.090	0.095	0.096	0.101
	Silhouette	0.150	0.137	0.126	0.120	0.111
agnes	Connectivity	26.525	119.408	174.801	185.848	220.475
	Dunn	0.175	0.150	0.137	0.137	0.137
	Silhouette	0.149	0.119	0.087	0.074	0.049

Tabela 48: Wskaźniki wewnętrzne

	Score	Method	Clusters
Connectivity	26.525	agnes	2
Dunn	0.186	kmeans	2
Silhouette	0.153	kmeans	2

Tabela 49: Optymalne liczby klastrów i metoda w zależności od kryterium



Rysunek 50: Ocena jakości grupowania

Według wskaźników Dunn i Silhouette optymalną metodą i liczbą klastrów jest k-means z $k = 2$. Kryterium Connectivity wskazuje zaś na metodę AGNES z $k = 2$.

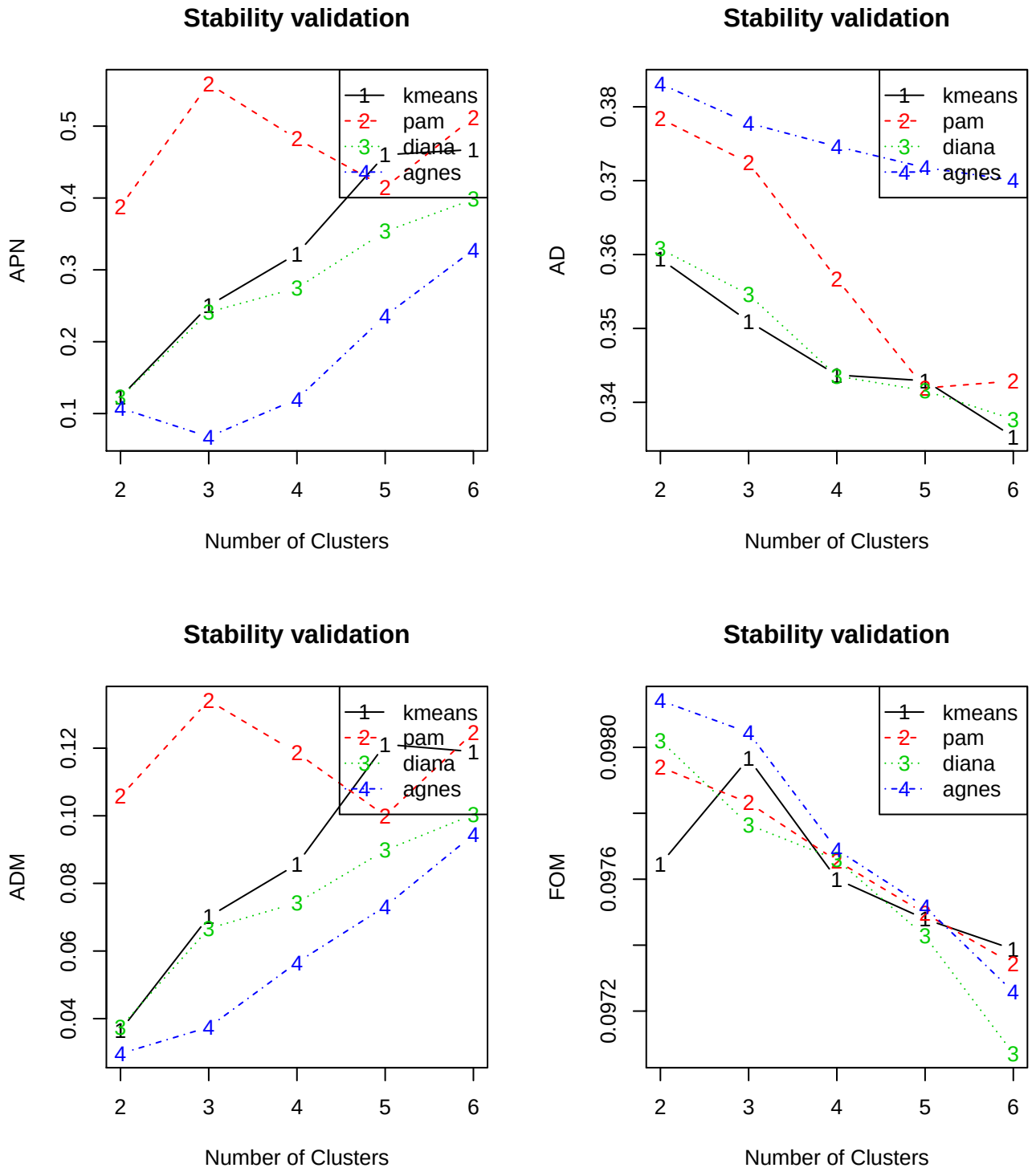
Sprawdzimy teraz wartości wskaźników stabilności: APN, AD, ADM i FOM. Zostały one przedstawione w Tabeli 50. Optymalna metoda i liczba klastrów według danego kryterium znajdują się w Tabeli 51. Rysunek 51 ilustruje otrzymane wyniki.

method	measure	k=2	k=3	k=4	k=5	k=6
kmeans	APN	0.123	0.250	0.322	0.460	0.467
	AD	0.359	0.351	0.344	0.343	0.335
	ADM	0.036	0.070	0.086	0.121	0.119
	FOM	0.098	0.098	0.098	0.097	0.097
pam	APN	0.388	0.559	0.482	0.415	0.512
	AD	0.378	0.372	0.357	0.342	0.343
	ADM	0.106	0.134	0.119	0.100	0.125
	FOM	0.098	0.098	0.098	0.097	0.097
diana	APN	0.123	0.241	0.275	0.353	0.398
	AD	0.361	0.355	0.344	0.342	0.338
	ADM	0.037	0.067	0.074	0.090	0.100
	FOM	0.098	0.098	0.098	0.097	0.097
agnes	APN	0.108	0.068	0.120	0.236	0.328
	AD	0.383	0.378	0.375	0.372	0.370
	ADM	0.030	0.037	0.057	0.073	0.094
	FOM	0.098	0.098	0.098	0.098	0.097

Tabela 50: Wskaźniki stabilności

	Score	Method	Clusters
APN	0.068	agnes	3
AD	0.335	kmeans	6
ADM	0.030	agnes	2
FOM	0.097	diana	6

Tabela 51: Optymalne liczby klastrów i metoda w zależności od kryterium



Rysunek 51: Ocena jakości grupowania

Każdy z rozważanych wskaźników stabilności wybiera inną metodę z liczbą klastrów jako optymalną. Zauważmy jednak, że wartości FOM są bardzo zbliżone dla wszystkich metod. Z kolei rozważane wcześniej wskaźniki Dunn i Silhouette wskazują, że najlepszą metodą jest k-średnich. Także jeden ze wskaźników stabilności - AD wskazuje na ten algorytm. Warto zauważyć, że także procent zgodności klas rzeczywistych i przypisanych dla tej metody był wysoki. Była to jedyna metoda, przy której łatwo udało nam się znaleźć cechy charakteryzujące dane klastry. Dwa ostatnie stwierdzenia były prawdziwe również dla danych przed zastosowaniem MDS.

6 Podsumowanie

Analiza skupień jest bardzo wymagającym tematem, wymaga dużego nakładu pracy i doświadczenia. Analizy przeprowadzane przez różne osoby mogą znacząco się różnić. Co więcej, trudno wybrać optymalną liczbę klastrów, ponieważ istnieje bardzo dużo różnych sposobów oceny grupowania. Wybór metody grupowania także nie jest prosty. Warto jednak rozważać stosowanie redakcji wymiaru, gdyż w wielu przypadkach może polepszyć uzyskane rezultaty.

W przypadku analizowanych przez nas danych *German Credit*, redukcja wymiaru poprzez zastosowanie skalowania wielowymiarowego sprawiła, że nastąpiła poprawa oceny dopasowywanych modeli klasyfikacji, jednak nie była ona znaczna. Należy również wziąć pod uwagę fakt, że stosując redukcję wymiaru utrudniamy interpretację otrzymywanych modeli. Ponadto w przypadku naszych danych, redukcja wymiaru poprzez MDS zmniejszyła wymiar o około połowę, co nie jest oszałamiającym rezultatem. Prawdopodobnie redukcja wymiaru ma większy sens, a być może nawet jej konieczna w przypadku, gdy dane mają bardzo dużą liczbę zmiennych, tak jak np. dane mikromacierzowe. Brak redukcji wymiaru może wówczas skutkować niemożnością dopasowania modelu klasyfikacji, na pewno zaś znacznie wydłuży czas obliczeń. U nas zastosowanie MDS skutkowało zazwyczaj pogorszeniem liczby zgodnych klas rzeczywistych i przypisanych. Może być to spowodowane utratą pewnej ilości informacji, z którą łączy się redukcja wymiaru. Wydaje się, że najlepsze rezultaty otrzymywaliśmy po zastosowaniu algorytmu k-means, zarówno przed jak i po zastosowaniu MDS. Mieliśmy wówczas najwyższą procentową zgodność klas rzeczywistych i przypisanych, a także łatwo udało nam się znaleźć cechy charakteryzujące dane klastry.