DATA MINING
PROJEKT

STATLOG (GERMAN CREDIT DATA) DATA SET

OPRACOWAŁA:
ALEKSANDRA GRZESZCZUK
NUMER ALBUMU: 255707

SPIS TREŚCI

1 CZĘŚĆ II - 21.06.2023

1.1 Analiza skupień wraz z oceną jakości

Analiza skupień jest narzędziem do eksploracyjnej analizy danych, której celem jest ułożenie obiektów w grupy w taki sposób, aby stopień powiązania obiektów z obiektami należącymi do tej samej grupy był jak największy, a z obiektami z pozostałych grup jak najmniejszy. Analiza skupień może być wykorzystywana do wykrywania struktur w danych bez wyprowadzania interpretacji/wyjaśnienia. Mówiąc krótko: analiza skupień jedynie wykrywa struktury w danych bez wyjaśniania dlaczego one występują. Warto dodać, że analiza skupień jest przykładem uczenia nienadzorowanego [1].

Dzięki analizie skupień można:

- Wykryć czy otrzymane skupienia wskazują na jakąś prawidłowość (np. związek pomiędzy symptomami a faktycznym stanem chorobowym)
- Dokonać redukcji olbrzymiego zbioru danych do średnich poszczególnych grup
- Potraktować rozdzielenie na grupy jako wstęp do dalszych wielowymiarowych analiz

Analiza skupień typowo znajduje swoje zastosowanie w marketingu (wyszukiwanie grup podobnych klientów), biologii i medycynie (grupowanie pacjentów, wyodrębnienie grup genów pełniących podobne funkcje biologiczne) czy w segmentacji obrazów (podział obrazu na jednorodne zbiory pikseli).

Algorytmy analizy skupień można sklasyfikować na kilka sposobów, uwzględniając różne kryteria klasyfikacji. Poniżej prezentujemy niektóre z nich.

- Metody grupujące. Ich celem jest znalezienie podziału obiektów na K grup tak, aby optymalizować określone kryterium. Przykładowo
  - k-means
  - k-medians
  - k-medoids (PAM)
  - CLARA
- Metody hierarchiczne. Są to techniki grupowania danych, które tworzą hierarchiczną strukturę klastrów. Istnieją dwie główne kategorie metod hierarchicznych: aglomeracyjne (AGNES) oraz rozdzielające (DIANA)
- MCLUST
- Metody rozmyte
- Algorytmy grafowe

[1] Uczenie nienadzorowane polega na tym, że maszyna nie posiada „klucza odpowiedzi" i musi sama analizować dane, szukać wzorców i odnajdywać relacje. Ten rodzaj machine learning najbardziej przypomina sposób działania ludzkiego mózgu, który wyciąga wnioski na podstawie spontanicznej obserwacji i intuicji. Wraz ze wzrostem zbiorów danych prezentowane wnioski są coraz bardziej precyzyjne.

1.1.1 Metoda k - means

Metoda k - means to, jak już wspomnieliśmy wyżej, popularny algorytm analizy skupień, który polega na podziale zbioru danych na K klastrów, gdzie K jest ustaloną liczbą. Oto opis kroków, które są częścią metody k - means:

1. Wybór liczby klastrów (K). Pierwszym krokiem w metodzie k - means jest określenie liczby klastrów, na które chcemy podzielić zbiór danych. Może to być wartość ustalona na podstawie wcześniejszej wiedzy lub wynikająca z analizy danych

2. Inicjalizacja centrów skupień. Losowo wybieramy K punktów w przestrzeni danych jako początkowe centra skupień klastrów

3. Przypisanie obiektów do klastrów. Każdy obiekt z zbioru danych jest przypisywany do najbliższego centra skupień na podstawie odległości euklidesowej lub innej miary podobieństwa

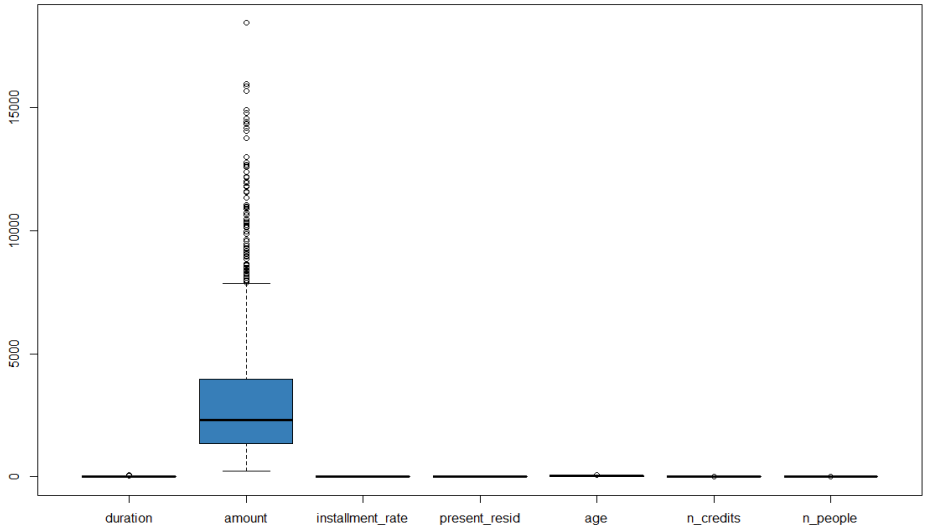
4. Aktualizacja centroidów. Po przypisaniu wszystkich obiektów do klastrów, centroidy są aktualizowane jako średnia wartość cech wszystkich obiektów przypisanych do danego klastra.

5. Powtarzamy kroki 3 - 4 do momentu spełnienia warunku zbieżności.

Warto dodać, że dla osób bardziej teoretycznych, algorytm k - means możemy interpretować jako iteracyjne rozwiązanie zadania minimalizacji kryterium

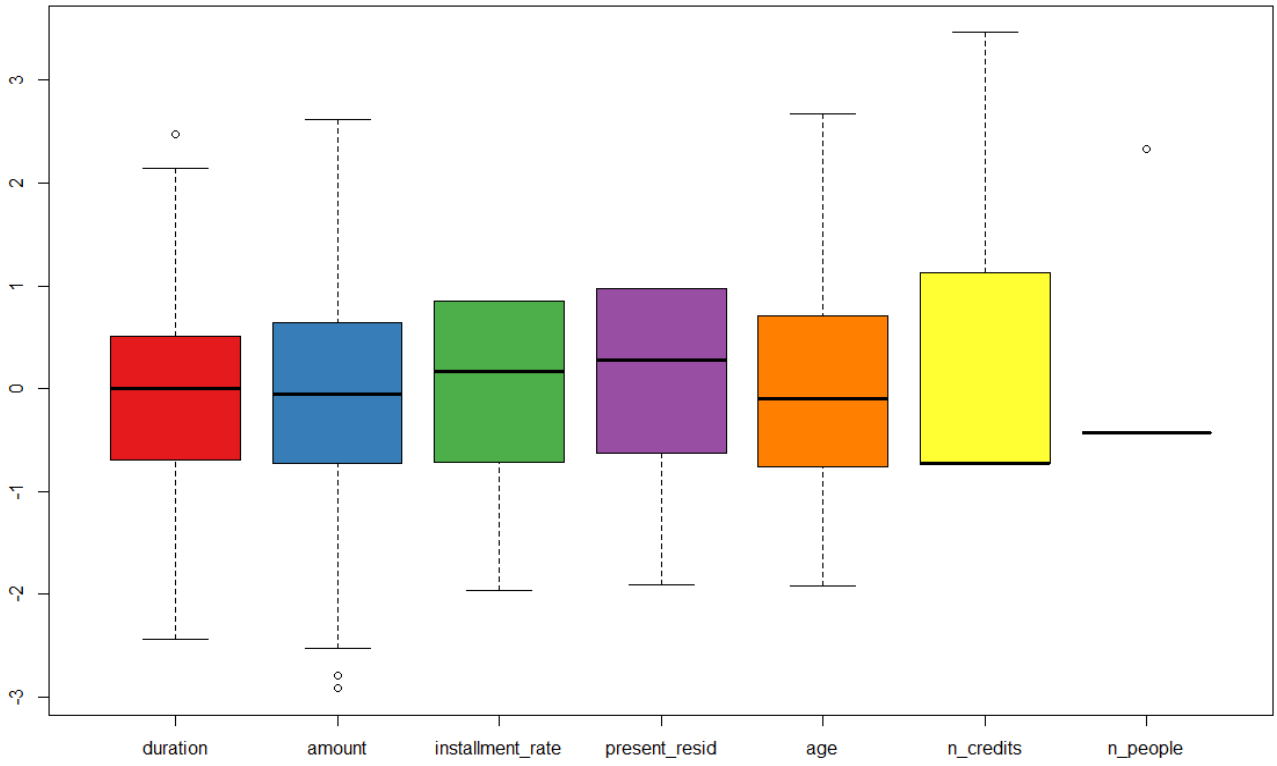
$$\tilde{W}(C) = \frac{1}{2}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{\{C(i)=k,C(j)=k\}}||x_i - x_j||^2 = \sum_{i=1}^{n}||x_i - m_{C(i)}||^2$$

Spośród całego zbioru danych *German Credit Data* wybieramy zmienne numeryczne, czyli *duration*, *amount*, *installment_rate*, *present_resid*, *age*, *n_credits*, *n_people*. Sprawdzamy, czy nasze dane są dość jednolite, spójne, czy jednak będziemy musieli dokonać ich standaryzacji. W tym celu wyznaczamy wykresy pudełkowe wybranych zmiennych.



Rysunek 1: Wykresy pudełkowe zmiennych niestandaryzowanych - opracowanie własne

Widzimy, że wykres *amount* znacznie różni się od pozostałych. Dlatego musimy dokonać standaryzacji zmiennych. Zrobimy to korzystając z funkcji `scale` z biblioteki `base` [2].



Rysunek 2: Wykresy pudełkowe zmiennych standaryzowanych - opracowanie własne

Znacznie lepiej teraz wyglądają nasze wykresy pudełkowe. Wyznaczymy teraz liczbę klastrów k za pomocą trzech metod.

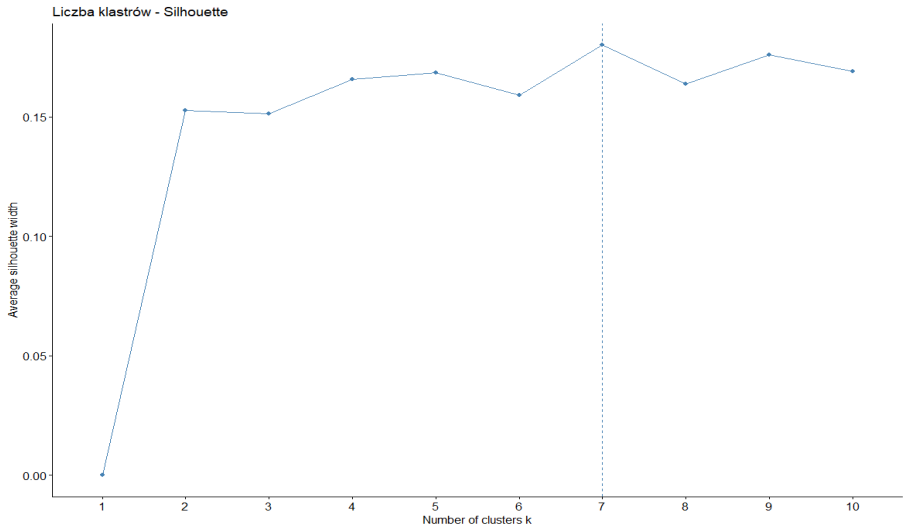
1. Indeks Silhouette. Dostarcza miary, która określa, jak dobrze obiekty w klastrze są podobne do siebie w porównaniu z innymi klastrami.

2. Metoda łokcia. Polega na analizie zmienności wyjaśnianej przez model w zależności od liczby klastrów. Należy dodać, że ocena na podstawie metody łokcia nie jest zawsze jednoznaczna. W praktyce może być trudno zidentyfikować wyraźny łokieć na wykresie.

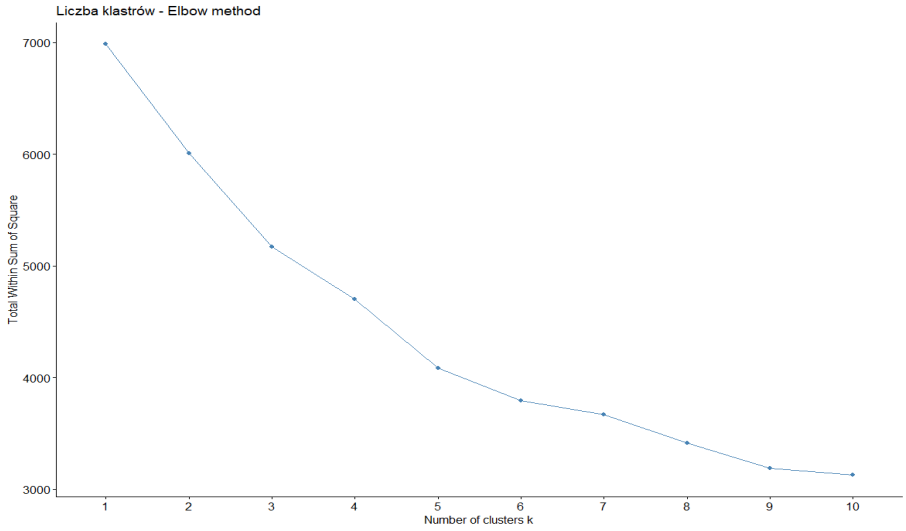
3. Metoda statystycznej luki. Metoda ta porównuje rzeczywiste dane z danymi wygenerowanymi losowo w celu określenia, czy struktura klastrów jest bardziej wyraźna niż można oczekiwać przez czysty przypadek.

[2]Informacje na temat funkcji dostępne pod adresem `https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/scale` zaś na temat biblioteki - `https://stat.ethz.ch/R-manual/R-devel/library/base/html/00Index.html`

W dalszej części sprawozdania będziemy korzystać z funkcji `fviz_nbclust` [3], która określa i wizualizuje optymalna liczbę klastrów na podstawie wybranych metod, oraz z funkcji `eclust` [4] zajmującej się wizualnym udoskonaleniem analizy skupień. Obydwie te funkcje dostępne są w bibliotecte `factoextra` [5].

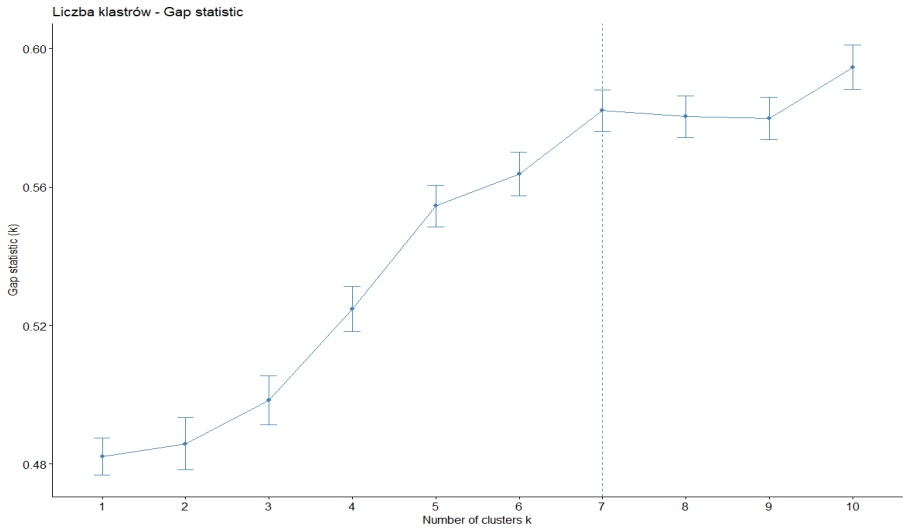


Rysunek 3: Liczba klastrów na podstawie indeksu Silhouette - opracowanie własne



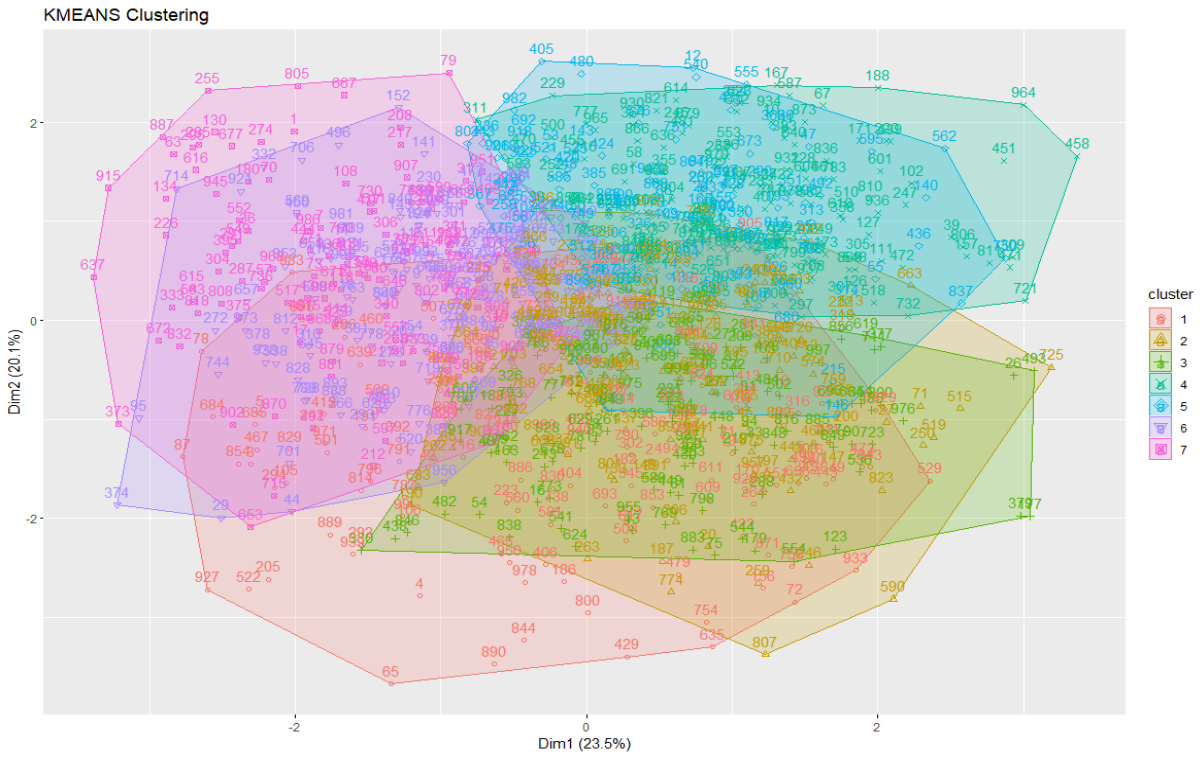
Rysunek 4: Liczba klastrów na podstawie metody łokcia - opracowanie własne

[3] `https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/fviz_nbclust`
[4] `https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/eclust`
[5] `https://cran.r-project.org/web/packages/factoextra/index.html`



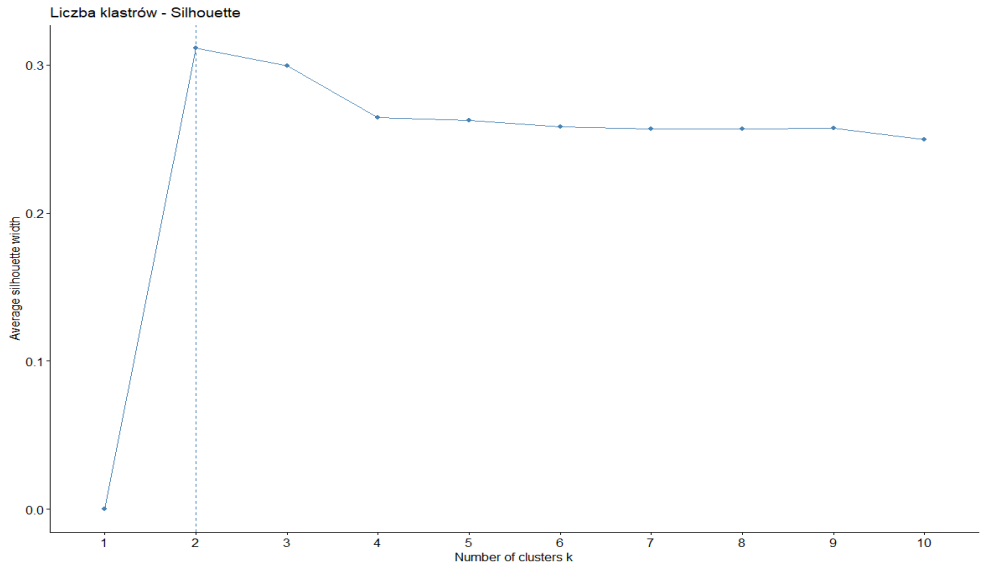
Rysunek 5: Liczba klastrów na podstawie metody statystycznej luki - opracowanie własne

Przyglądając się powyższym rysunkom (3), (4), (5) widzimy, że możemy za liczbę skupień k uznać wartość 7.

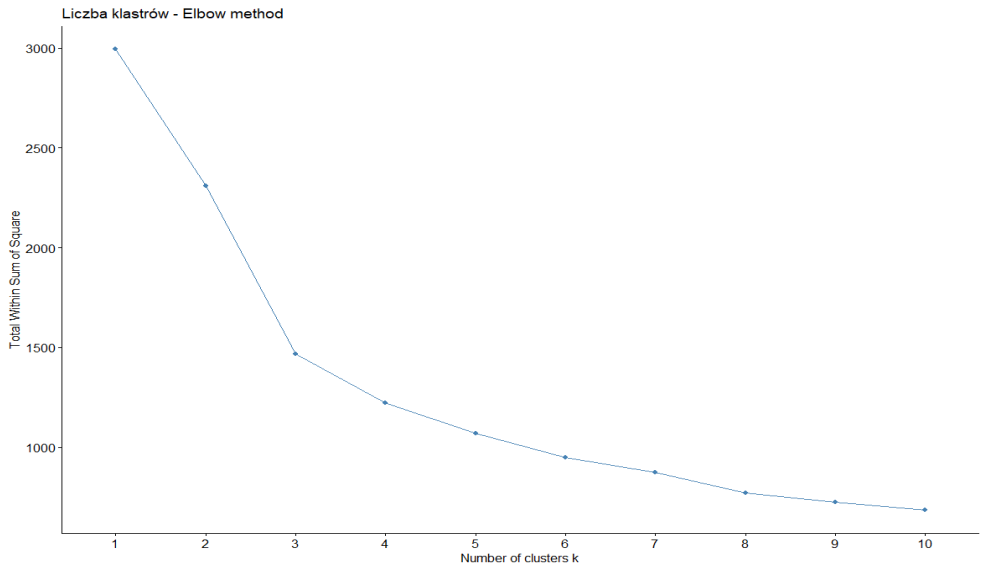


Rysunek 6: Wykres wyników analizy skupień - opracowanie własne

Widzimy, że dla takiej ilości zmiennych nie widać nic na wykresie analizy skupień. Dlatego teraz wybierzemy trzy zmienne numeryczne - *duration*, *age*, *amount*, dla których dokonamy tej samej analizy, mając nadzieję na lepsze efekty.



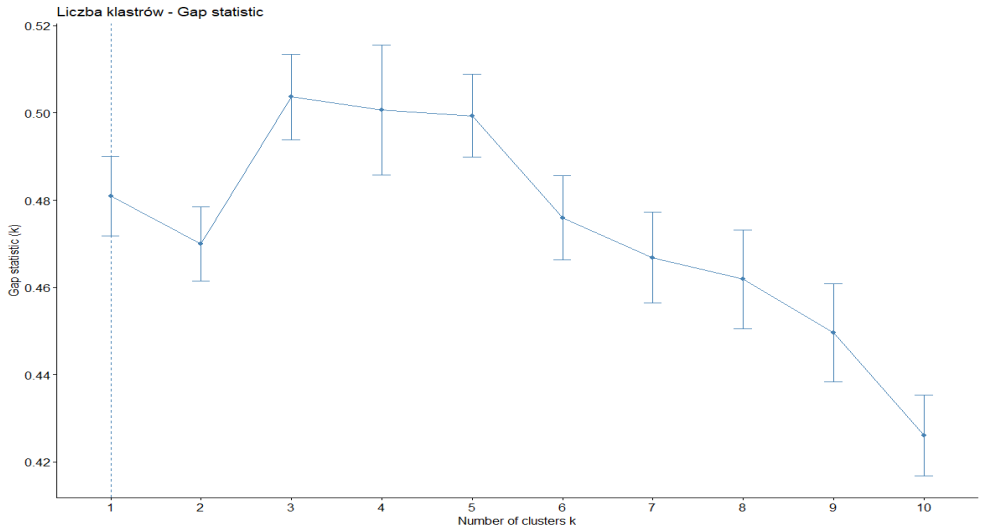
Rysunek 7: Liczba klastrów na podstawie indeksu Silhouette - opracowanie własne



Rysunek 8: Liczba klastrów na podstawie metody statystycznej luki - opracowanie własne

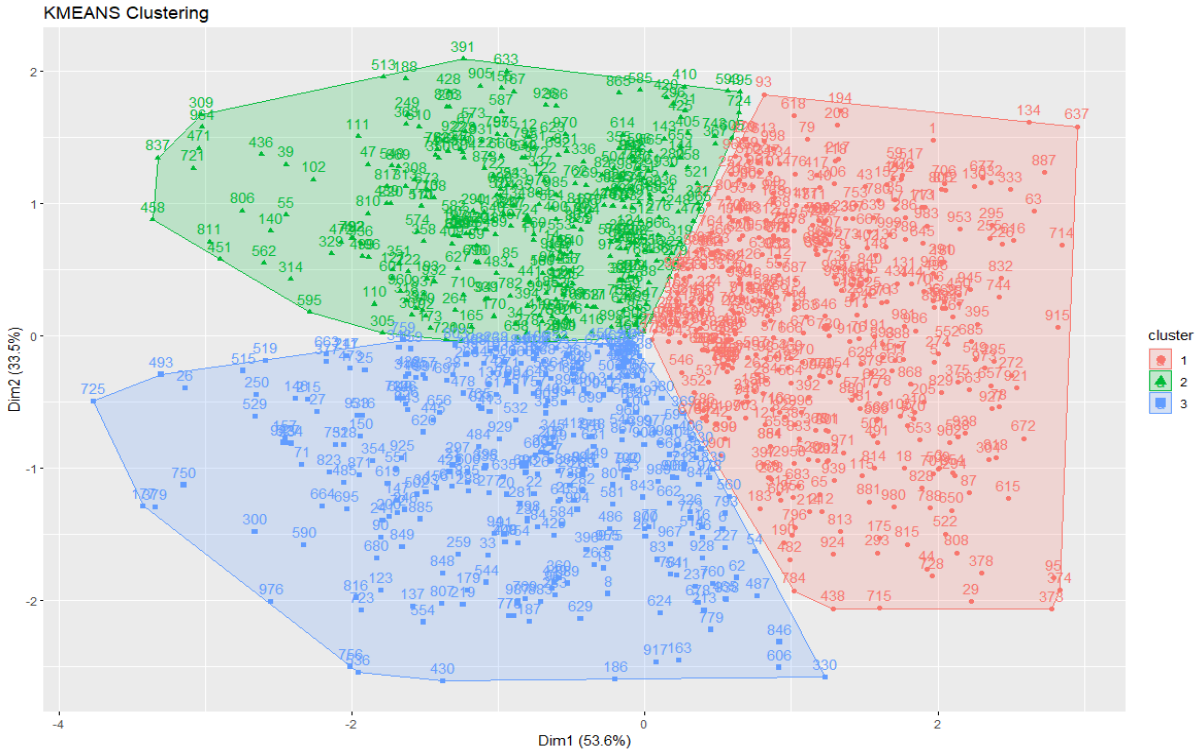
W tym przypadku otrzymaliśmy trzy różne wartości klastrów. Na podstawie rysunku (7) wybralibyśmy liczbę k = 2. Z rysunku (8) odczytujemy, że najlepszym będzie k = 3. Natomiast na poniższym (9) wyraźnie odznacza się liczba k = 1. Różne metody analizy skupień mogą prowadzić do różnych wartości liczby klastrów. Każda metoda ma swoje własne kryteria oceny i algorytmy, które wpływają na wybór optymalnej liczby klastrów. W związku z tym, różne metody mogą prowadzić do różnych rekomendacji co do optymalnej liczby klastrów. Sprawdzanie różnych metod pozwala na uzyskanie szerszej perspektywy i lepsze zrozumienie struktury danych. Ponadto nie ma jednego "rozmiaru" odpowiedniego dla wszystkich, jeśli chodzi o liczbę klastrów. Wybór optymalnej liczby klastrów powinien uwzględniać specyfikę danych. Przeanalizowanie różnych metod może dostarczyć bardziej kompleksowej oceny i dostosować się do różnych struktur danych.

Wnioskiem jest, że sprawdzanie liczby klastrów na podstawie różnych algorytmów jest ważne dla uzyskania bardziej wszechstronnego i informacyjnego spojrzenia na strukturę danych. Pomaga to uniknąć jednostronnej interpretacji i dostarcza bardziej pewnych i trafnych rekomendacji dotyczących optymalnej liczby klastrów w analizie skupień.



Rysunek 9: Liczba klastrów na podstawie metody statystycznej luki - opracowanie własne

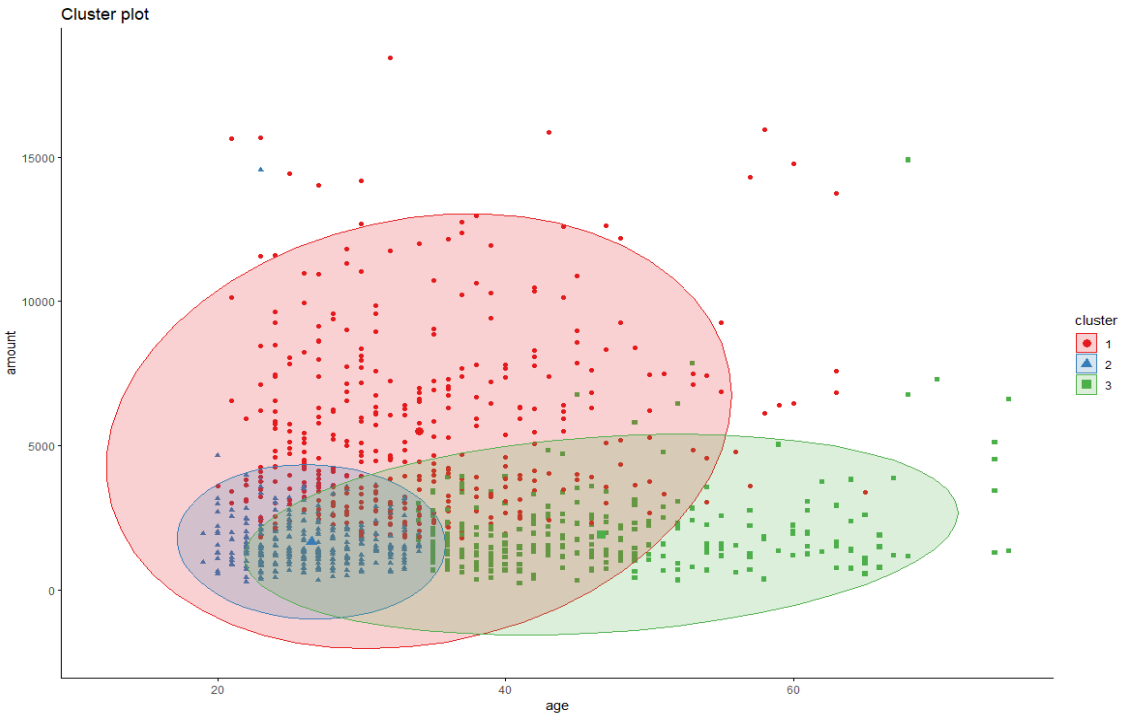
Wybraliśmy ostatecznie liczbę k = 3. Na jej podstawie uzyskaliśmy następujący wykres wyników analizy skupień.



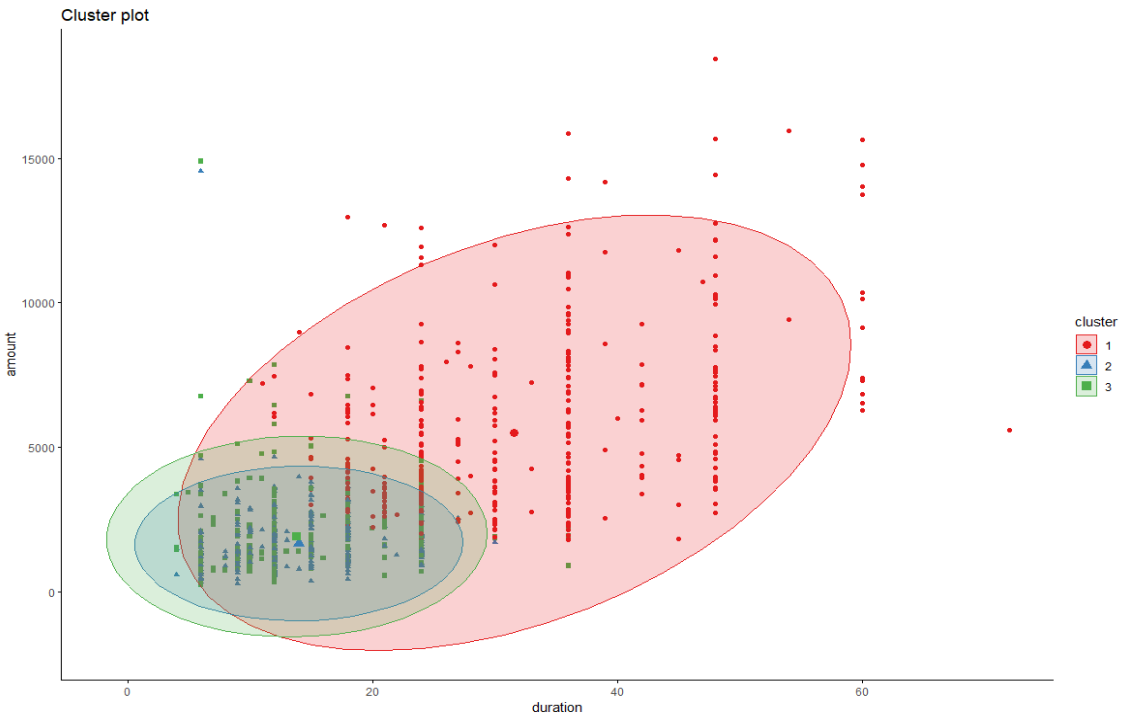
Rysunek 10: Wykres wyników analizy skupień - opracowanie własne

Widzimy, że powyższy rysunek (10) zdecydowanie lepiej rozdziela klastry niż wcześniejszy (18).

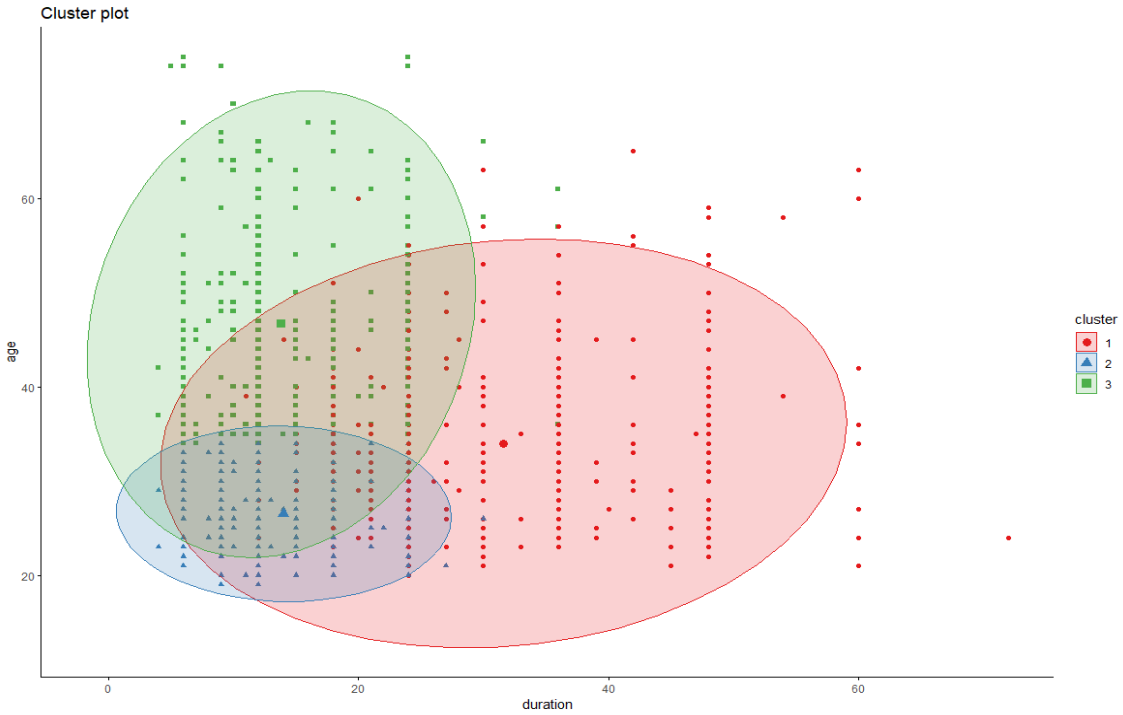
Teraz zaprezentujemy trzy wykresy trzech grup naszych klastrów.



Rysunek 11: Wykres klastrowy zmiennych *age* i *amount* - opracowanie własne



Rysunek 12: Wykres klastrowy zmiennych *duration* i *amount* - opracowanie własne



Rysunek 13: Wykres klastrowy zmiennych *duration* i *age* - opracowanie własne

Na podstawie powyższych trzech rysunków (11), (12), (13) możemy podzielić nasze dane na trzy zasadnicze grupy

- Pierwsza grupa to osoby w średnim wieku, które pożyczyły dość dużo pieniędzy na długi okres
- Druga grupa to osoby młode, które pożyczyły niewielkie kwoty pieniędzy na raczej krótki okres
- Trzecia grupa to osoby starsze z kredytem krótkoterminowym na niewielką kwotę

1.1.2 Metoda PAM

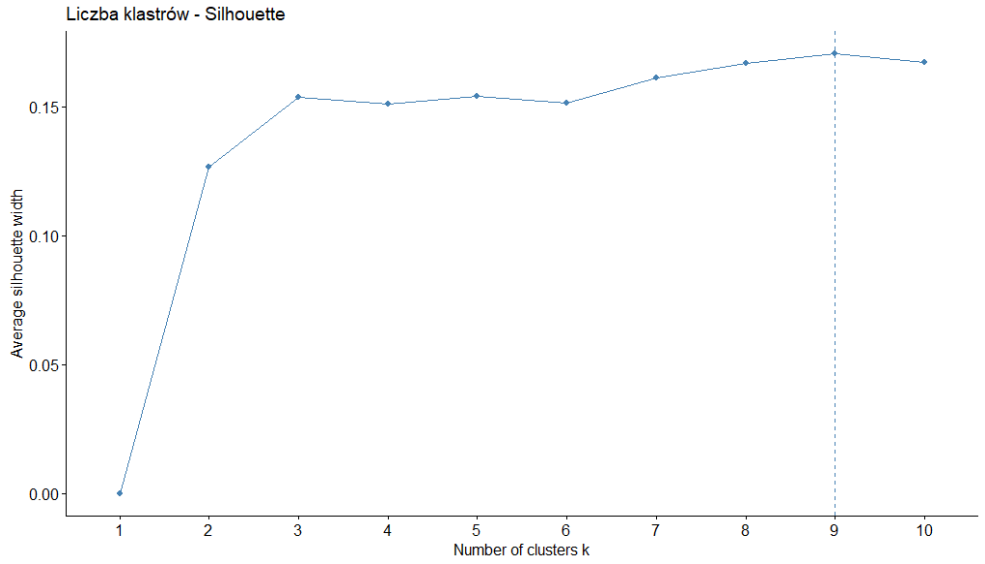
Metoda PAM jest jednym z algorytmów stosowanym w analizie skupień będącej elementem klasyfikacji bez nadzoru. PAM jest realizacją metody k- medoidowej, czyli takiej techniki grupowania, która dzieli zbiór danych zawierających n obiektów na k grup znanych a priori. Ma na celu minimalizację funkcji kosztu związanej z odległościami między obserwacjami a medoidami. Oto opis kroków będących częścią metody PAM:

1. Wybór początkowych medoidów. Na początku algorytmu konieczne jest wybranie początkowych medoidów, które będą reprezentować klastry
2. Przypisanie obserwacji do medoidów. W kolejnym kroku algorytmu każda obserwacja zostaje przypisana do najbliższego medoida na podstawie określonej miary odległości
3. Aktualizacja medoidów. Po przypisaniu obserwacji do medoidów następuje etap aktualizacji medoidów. Algorytm iteracyjne przegląda wszystkie kombinacje medoidów i ich nieprzypisanych obserwacji, sprawdzając, czy zamiana medoida na nową obserwację prowadzi do zmniejszenia funkcji kosztu

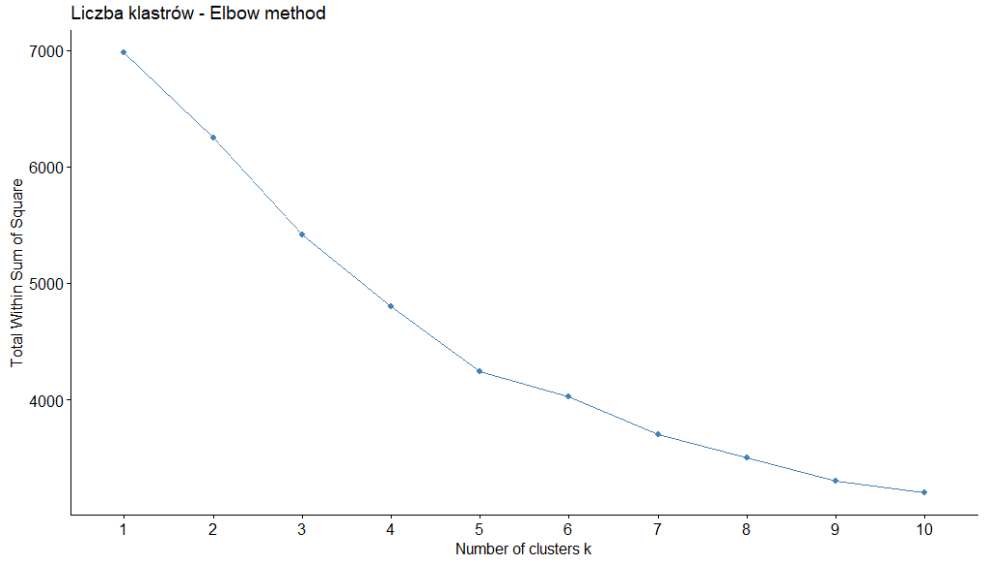
4. Powtarzanie kroków 2 i 3

Ostateczny wynik metody PAM to partycja danych na klastry, gdzie każdy klaster jest reprezentowany przez jeden medoid. Warto dodać, że algorytm PAM jest efektywny dla małych i średnich zbiorów danych, jednak dla dużych zbiorów może być czasochłonny obliczeniowo.

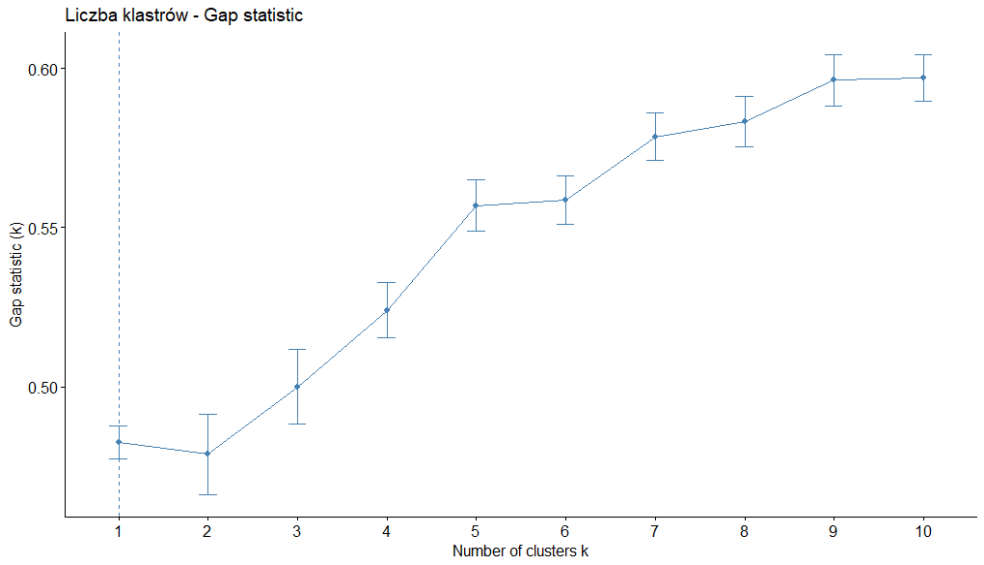
Korzystając z tych samych funkcji co wcześniej wyznaczamy liczbę klastrów korzystając z omówionych już wcześniej indeksu Silhouette, metody łokcia oraz metody statystycznej luki.



Rysunek 14: Liczba klastrów na podstawie indeksu Silhouette - opracowanie własne

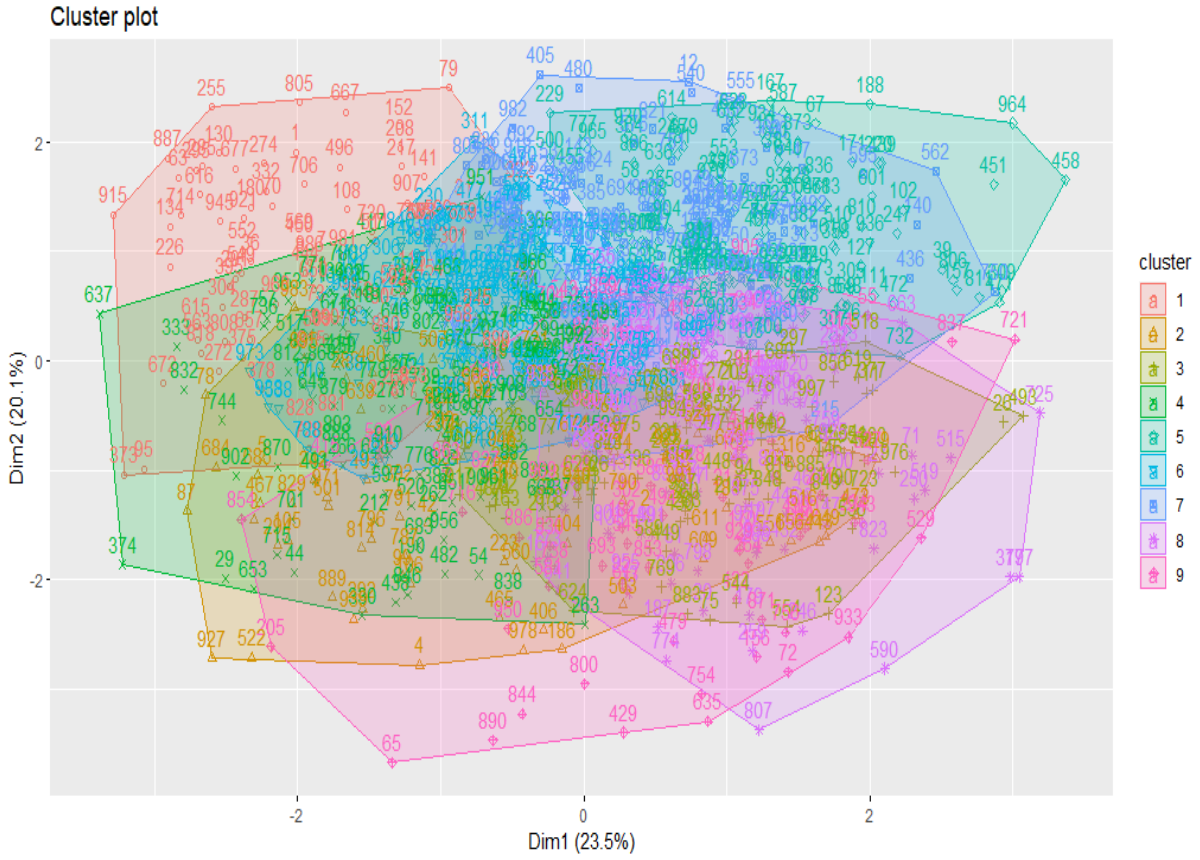


Rysunek 15: Liczba klastrów na podstawie metody łokcia - opracowanie własne

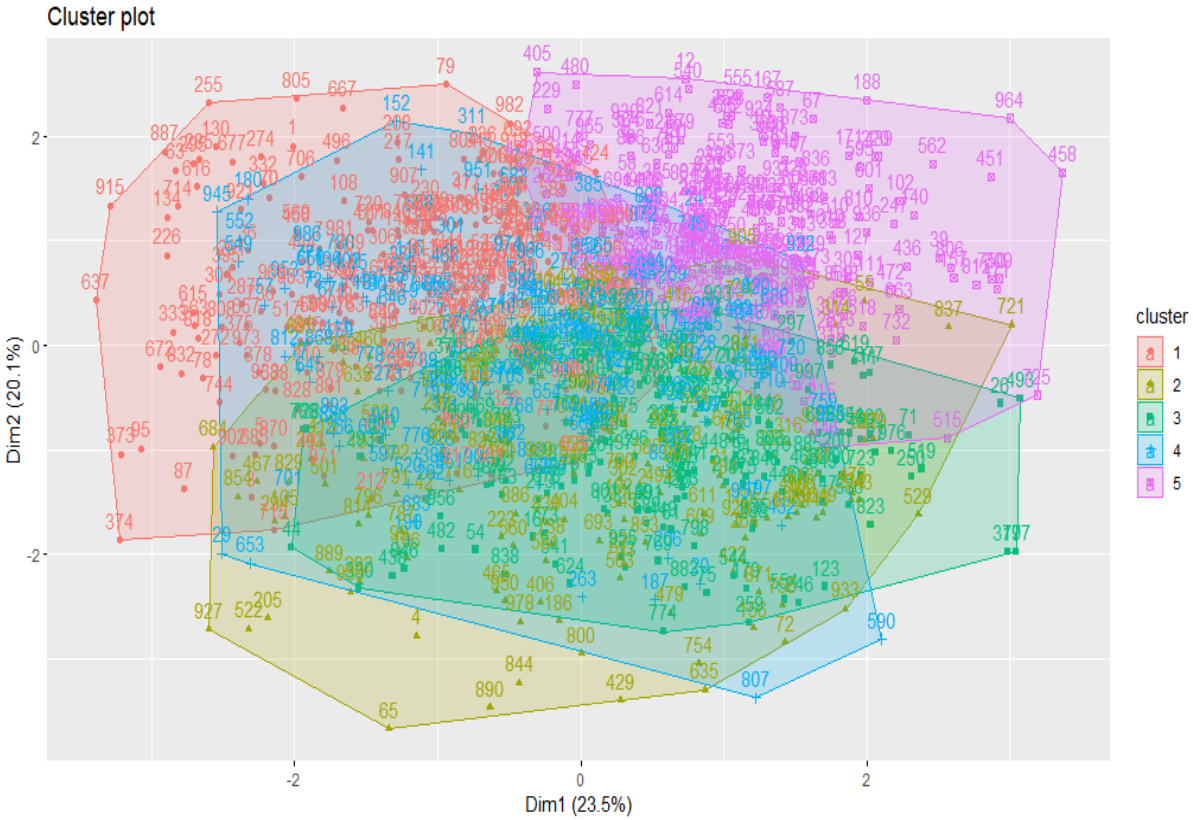


Rysunek 16: Liczba klastrów na podstawie metody luki statystycznej - opracowanie własne

Powyższe rysunki (14), (15), (16) wykazują się bardzo dużą skrajnością, niejednorodnością. Indeks Silhouette podał nam wartość 9, metoda łokcia 5 zaś metoda luki statystycznej jedynie 1. Wyznaczymy teraz wykresy analizy skupień dla k = 9,5.

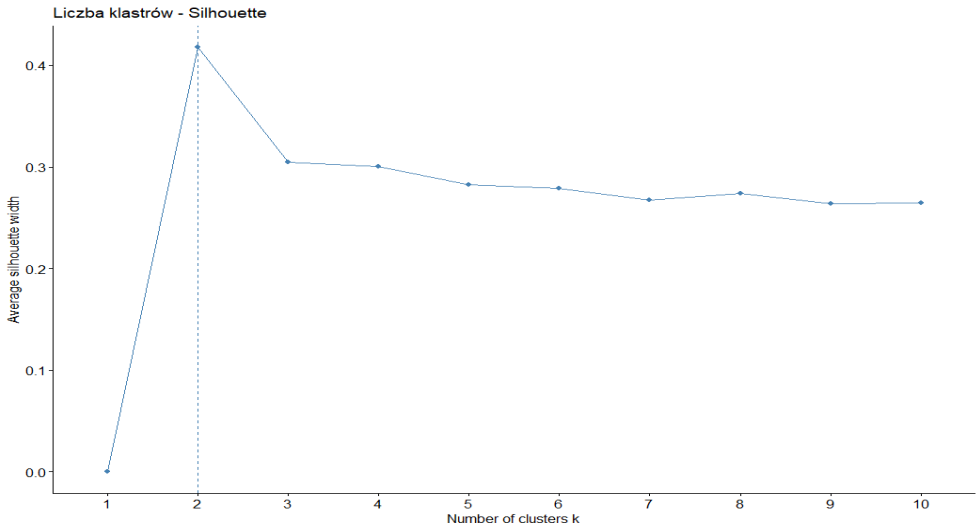


Rysunek 17: Wykres wyników analizy skupień dla k = 9 - opracowanie własne

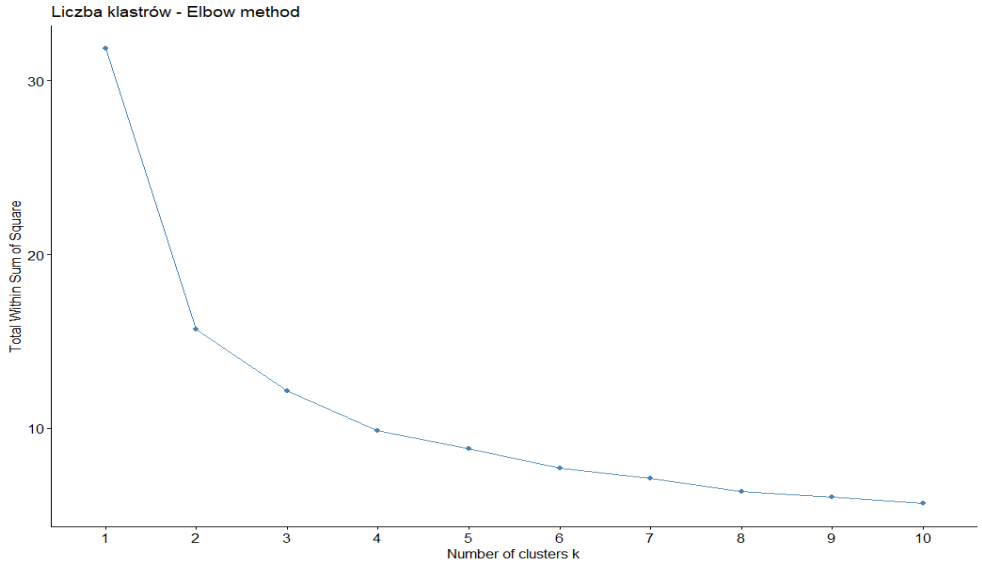


Rysunek 18: Wykres wyników analizy skupień dla k = 5 - opracowanie własne

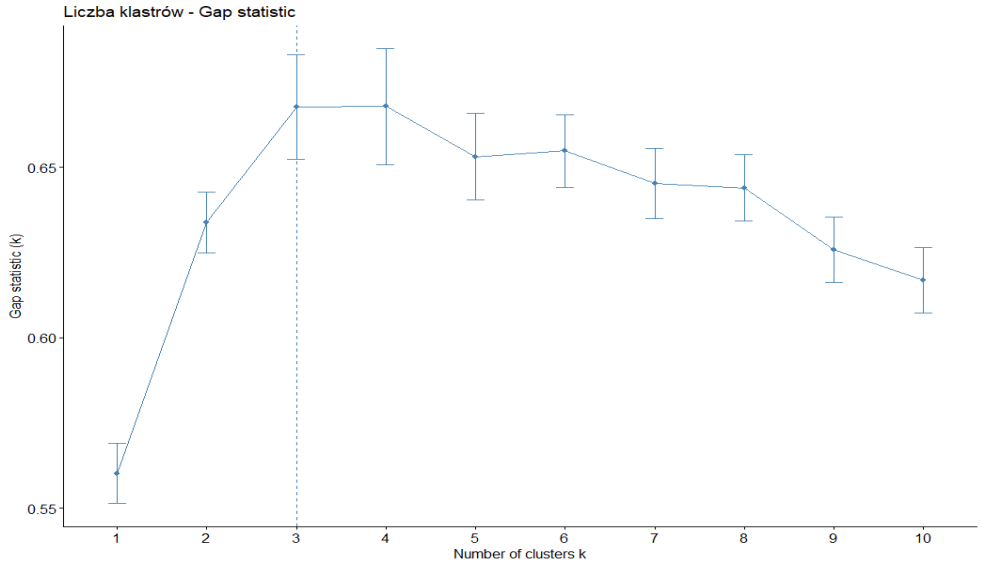
Możemy śmiało stwierdzić, iż wyniki analizy skupień dla wszystkich zmiennych numerycznych w naszym modelu, niezależnie od wartości parametru k, charakteryzują się nieczytelnością na wykresach. W związku z tym, postanowiliśmy ponownie wybrać trzy zmienne - *duration*, *age* oraz *amount* - i przeprowadzić na nich kolejną analizę, w nadziei na uzyskanie lepszych rezultatów. Pierwszym krokiem będzie jednak ponowne dobranie wartości parametru k.



Rysunek 19: Liczba klastrów na podstawie indeksu Silhouette - opracowanie własne

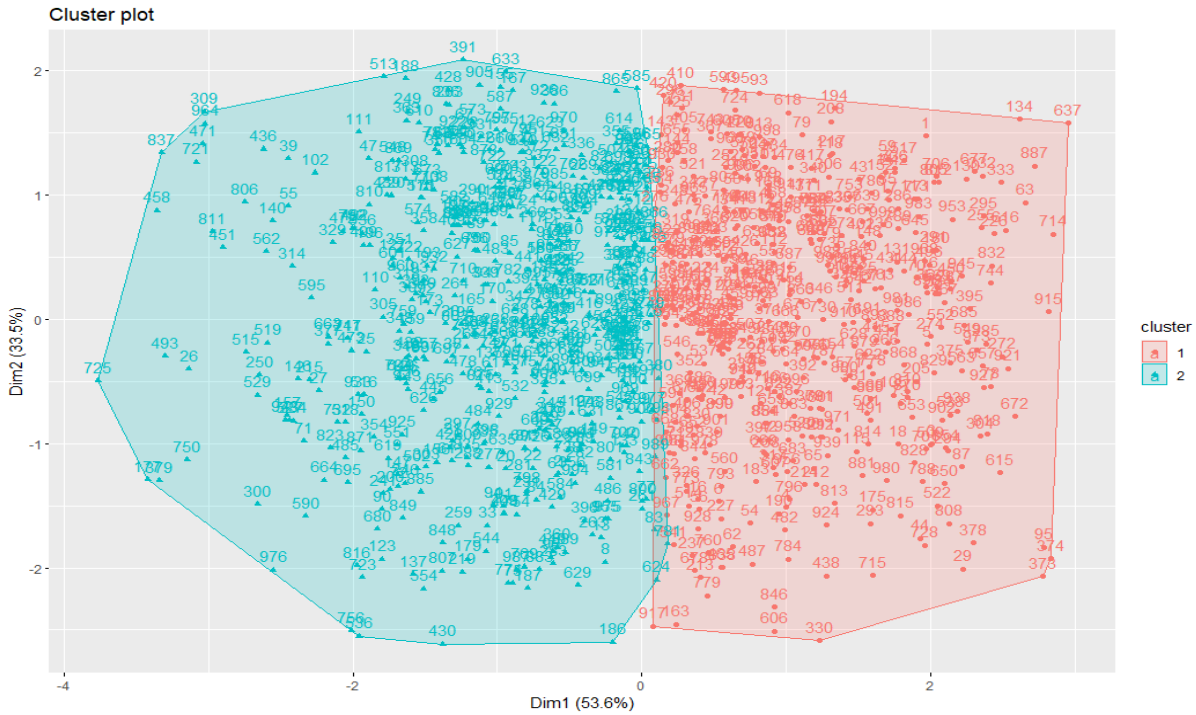


Rysunek 20: Liczba klastrów na podstawie metody łokcia - opracowanie własne

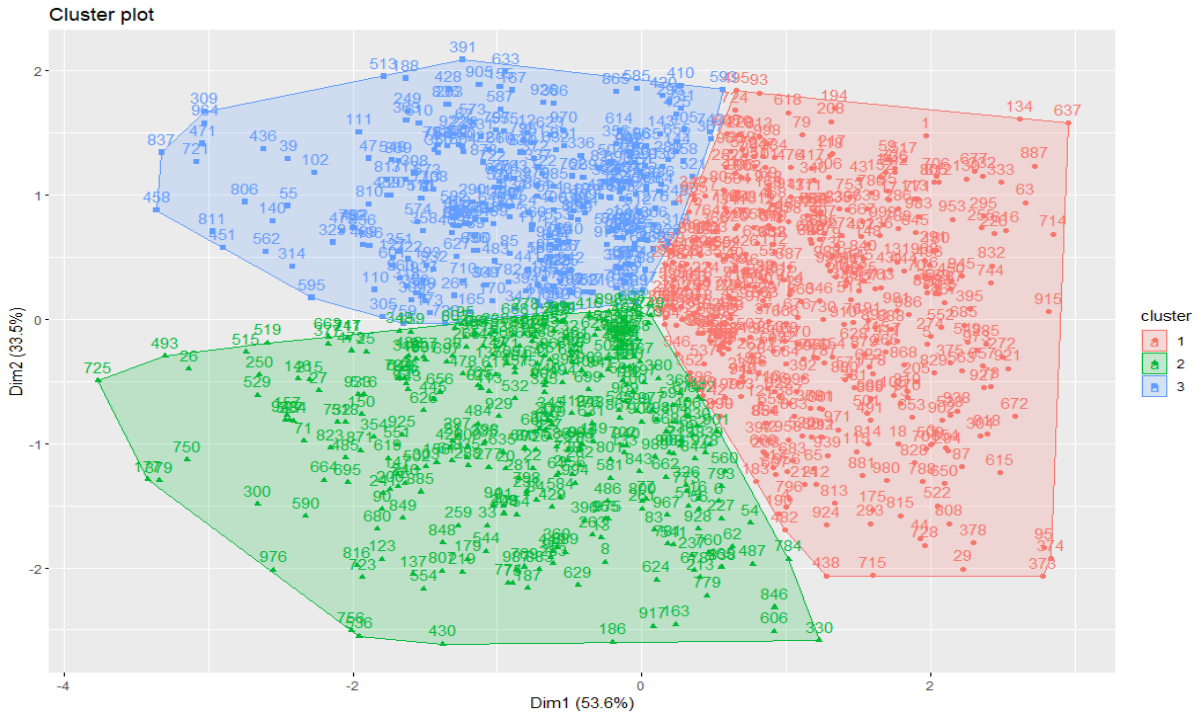


Rysunek 21: Liczba klastrów na podstawie metody luki statystycznej - opracowanie własne

Na podstawie powyższych trzech rysunków, (19), (20), (21) widzimy, że nie mamy już takich różnych wartości jak poprzednio. Indeks Silhouette oraz metoda łokcia wskazały, że liczba klastrów 2 będzie odpowiednia, zaś przy metodzie statycznej luki wyszło nam, że optymalne k = 3.



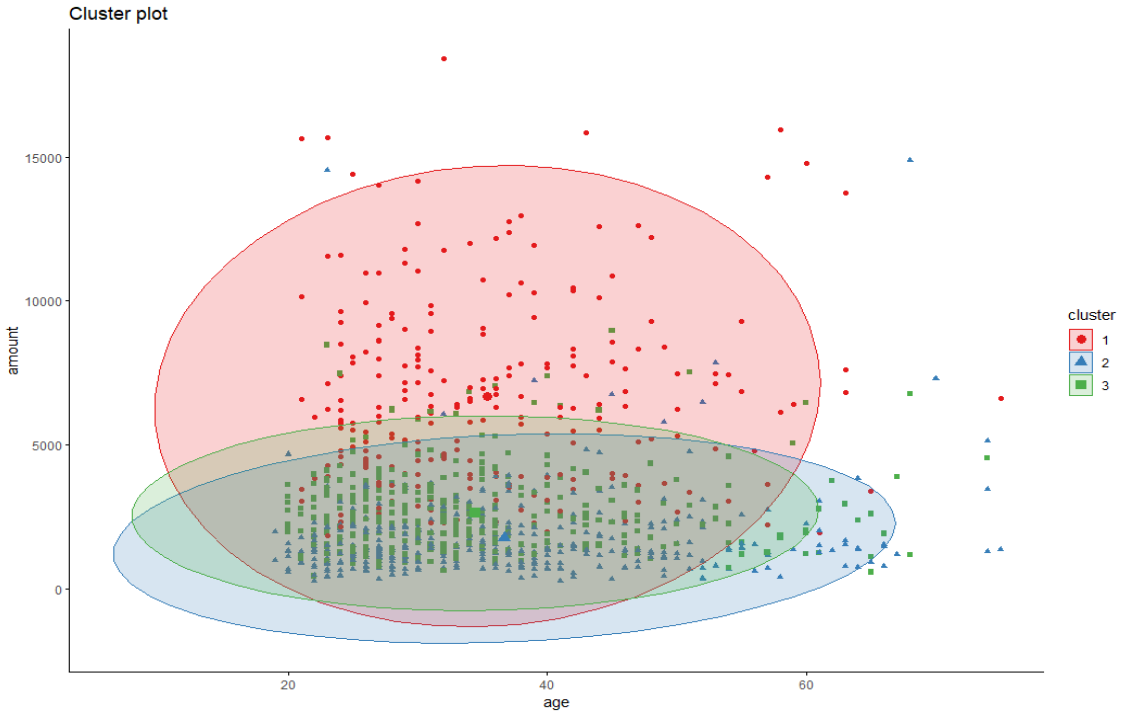
Rysunek 22: Wykres wyników analizy skupień dla k = 2 - opracowanie własne



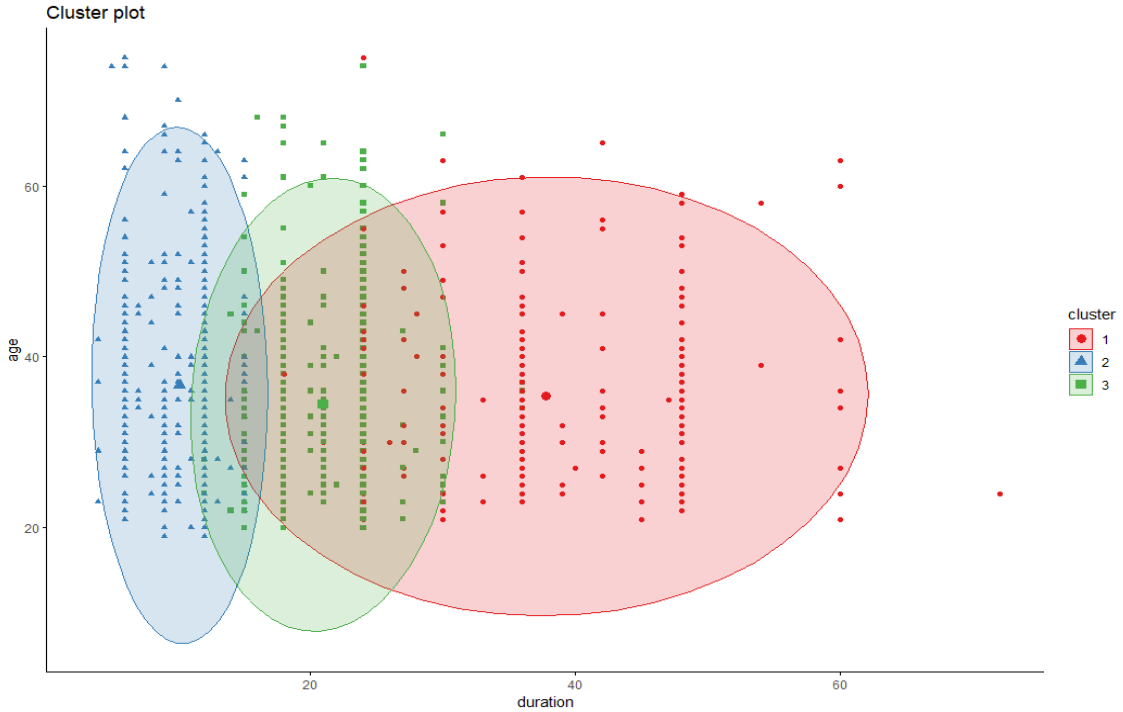
Rysunek 23: Wykres wyników analizy skupień dla k = 3 - opracowanie własne

Widzimy, że otrzymaliśmy zdecydowanie lepsze dopasowanie niż wcześniej. Możemy śmiało stwierdzić, że nie zawsze "więcej znaczy lepiej".

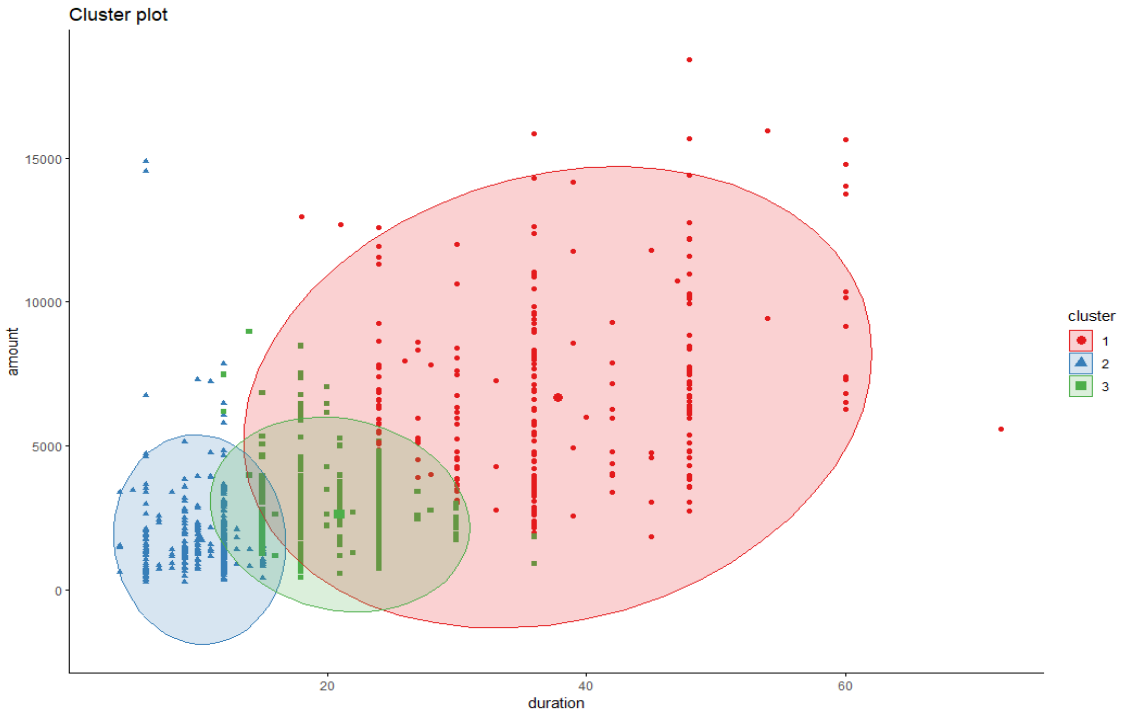
Dalej prezentujemy wykresy klastrowe w naszych grupach



Rysunek 24: Wykres klastrowy zmiennych *age* i *amount* - opracowanie własne



Rysunek 25: Wykres klastrowy zmiennych *duration* i *age* - opracowanie własne



Rysunek 26: Wykres klastrowy zmiennych *duration* i *amount* - opracowanie własne

Z rysunku (24) otrzymujemy właściwie tylko dwie grupy.

- I grupa - osoby, które wzięły kredyt duży (od 0 do 15 000)
- II grupa - osoby które wzięły mały kredyt (od 0 do 6000)

Z rysunku (25) otrzymujemy trzy wyraźne grupy. Warto podkreślić jednak, że nie otrzymaliśmy tutaj rozbieżności ze względu na wiek osób.

- I grupa - osoby z najdłuższym okresem kredytowania w miesiącach
- II grupa - osoby, które wzięły kredyt na najmniejszym okresem kredytowania w miesiącach
- III grupa - osoby ze średnim okresem kredytowania w miesiącach

Z rysunku (26) otrzymujemy dalej trzy wyraźne grupy.

- I grupa - osoby z dużym okresem kredytowania oraz znacznie większą wartością kredytu
- II grupa - osoby z małym okresem kredytowania oraz małą wartością kredytu
- III grupa - osoby ze średnim okresem kredytowania oraz małą wartością kredytu

Podsumowując, warto zauważyć, że największą ilość stanowi grupa I (czyli czerwona). Są to osoby w każdym wieku, z dużym kredytem na długi okres kredytowania. Drugą grupę stanowią osoby z małym kredytem na krótki okres kredytowania, zaś trzecią, ostatnią grupę, reprezentują osoby również z niskim kredytem ale na średni okres kredytowania.

1.1.3 Metoda AGNES

Metoda AGNES to technika analizy skupień, która polega na grupowaniu obiektów na podstawie ich podobieństwa. Wykorzystuje aglomeracyjne łączenie, łącząc pary grup o największym podobieństwie na podstawie obliczonej macierzy podobieństwa. Proces ten kontynuuje się aż do utworzenia jednej dużej grupy, umożliwiając również tworzenie hierarchicznej struktury skupień.

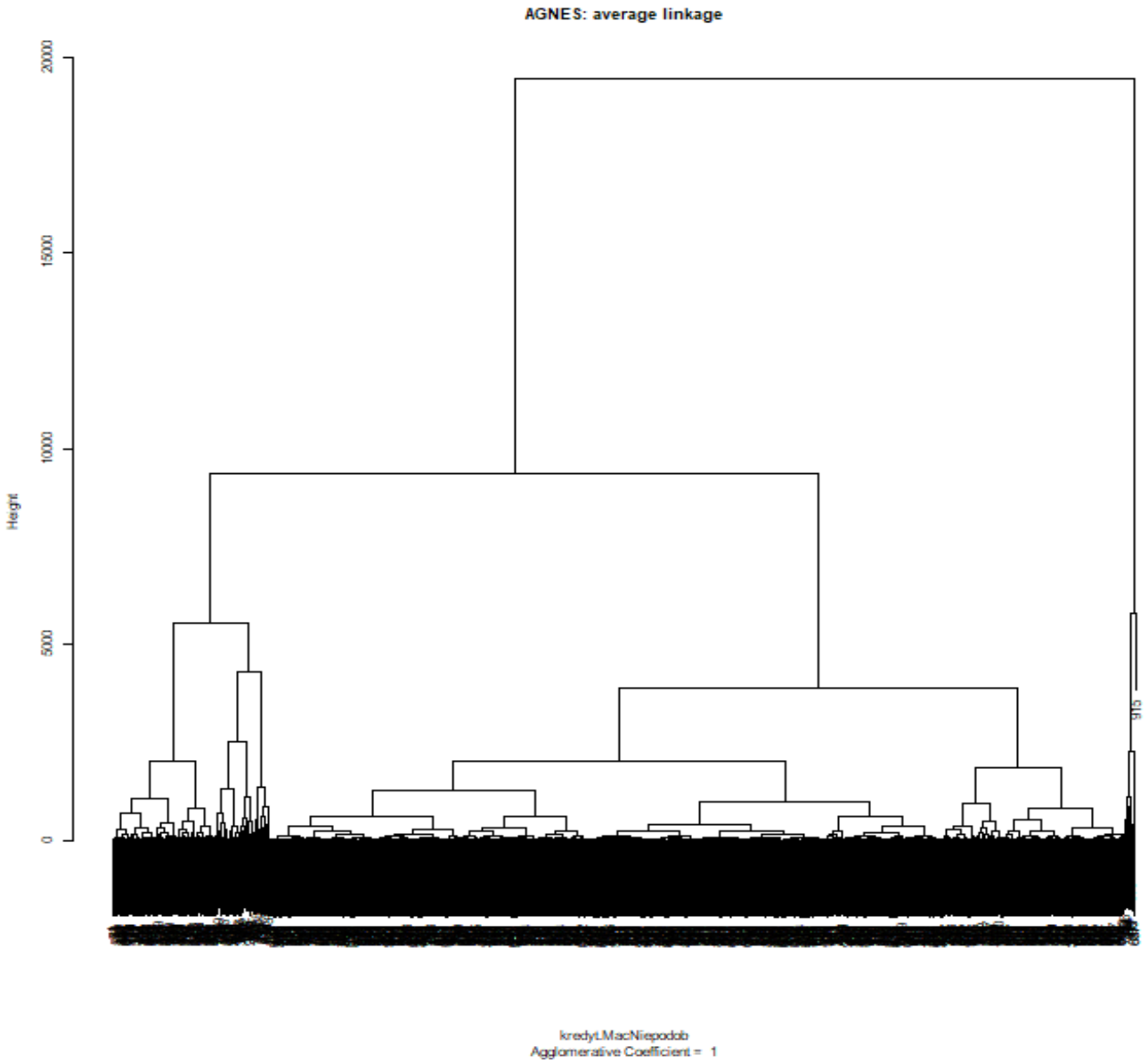
W tej części sprawozdania wykorzystamy głównie bibliotekę `cluster` [6] zawierającej różne metody analizy skupień.
W metodzie AGNES posiadamy trzy metody łączenia klastrów

- average - podczas tworzenia klastrów w tej metodzie, odległość między dwoma klastrami jest obliczana na podstawie średniej arytmetycznej odległości między wszystkimi parami obiektów z tych klastrów.
- complete - w tej metodzie odległość między dwoma klastrami jest obliczana na podstawie maksymalnej odległości między wszystkimi parami obiektów z tych klastrów.
- single - w tej metodzie odległość między dwoma klastrami jest obliczana na podstawie minimalnej odległości między wszystkimi parami obiektów z tych klastrów.

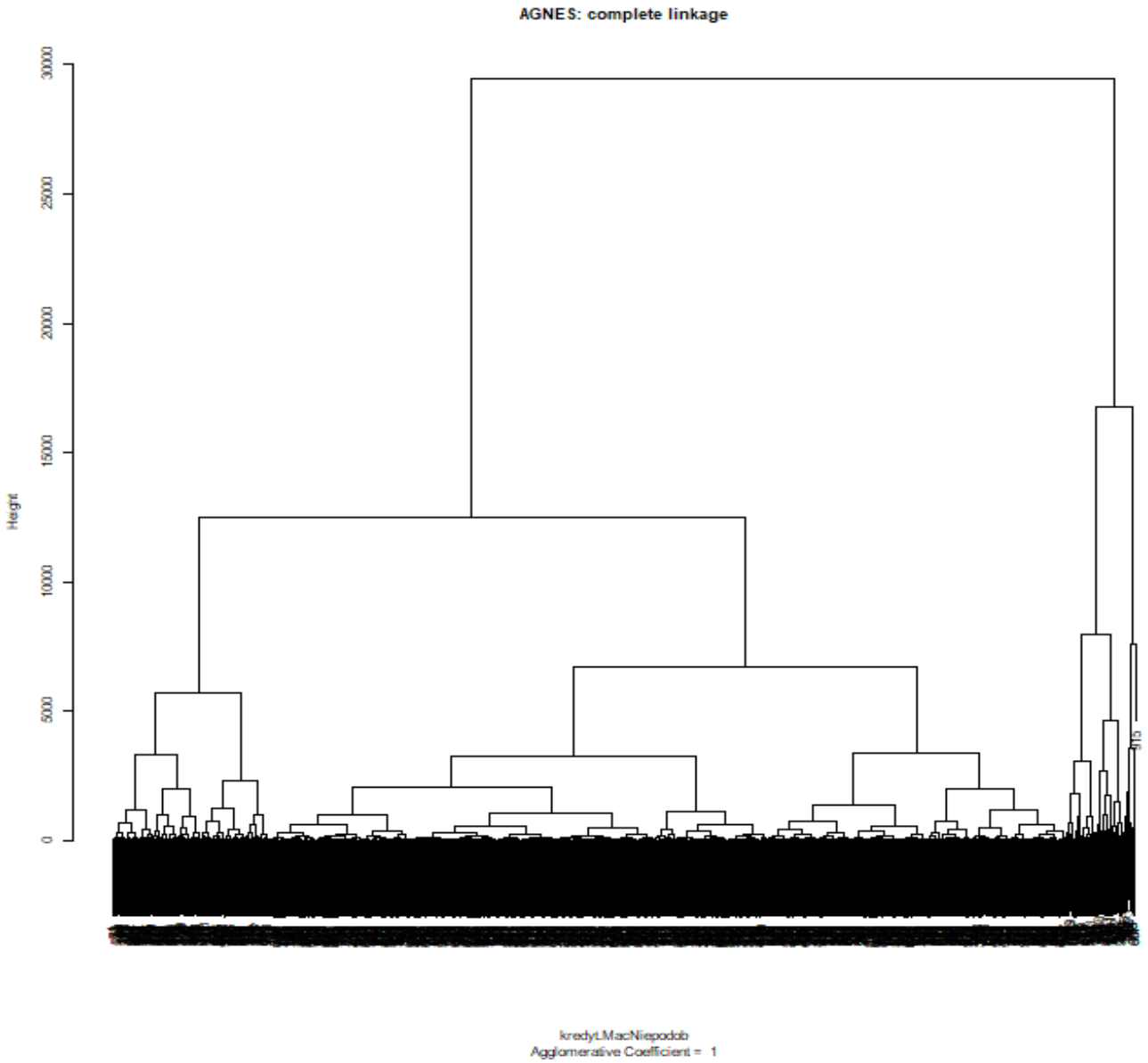
Pierwszy argument X w metodzie agnes określa macierz danych wejściowych, lub tak zwaną macierz odmienności, która zależy od argumentu diss. Jeśli diss = TRUE, przyjmujemy, że macierz X jest macierzą odmienności. W przeciwnym wypadku, macierz X jest traktowana jako macierz obserwacji. Każda zmienna (kolumna macierzy w danych) jest standaryzowana przez odjęcie średniej wartości zmiennej, a następnie podzielona przez średnie odchylenie bezwzględne zmiennej.

Przyglądając się trzem poniższym rysunkom (27), (28), (29) widzimy, że nie są one zbyt wyraźne. Podpisy konkretnych gałęzi zlewają się w jedno.

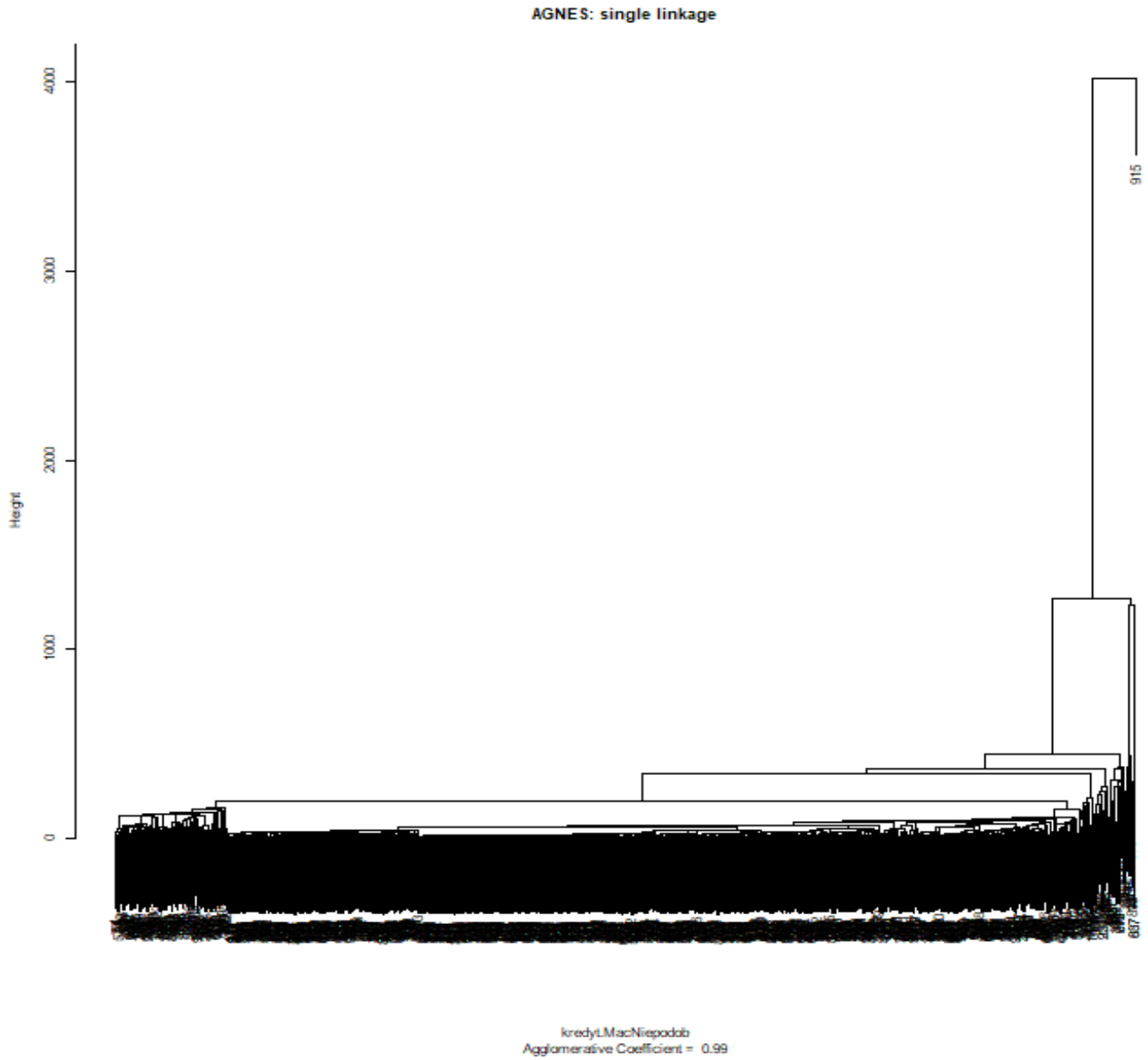
[6] Więcej informacji dostępnych pod linkiem `https://www.rdocumentation.org/packages/cluster/versions/2.1.4`



Rysunek 27: Wykres AGNES z metodą łączenia klastrów average - opracowanie własne

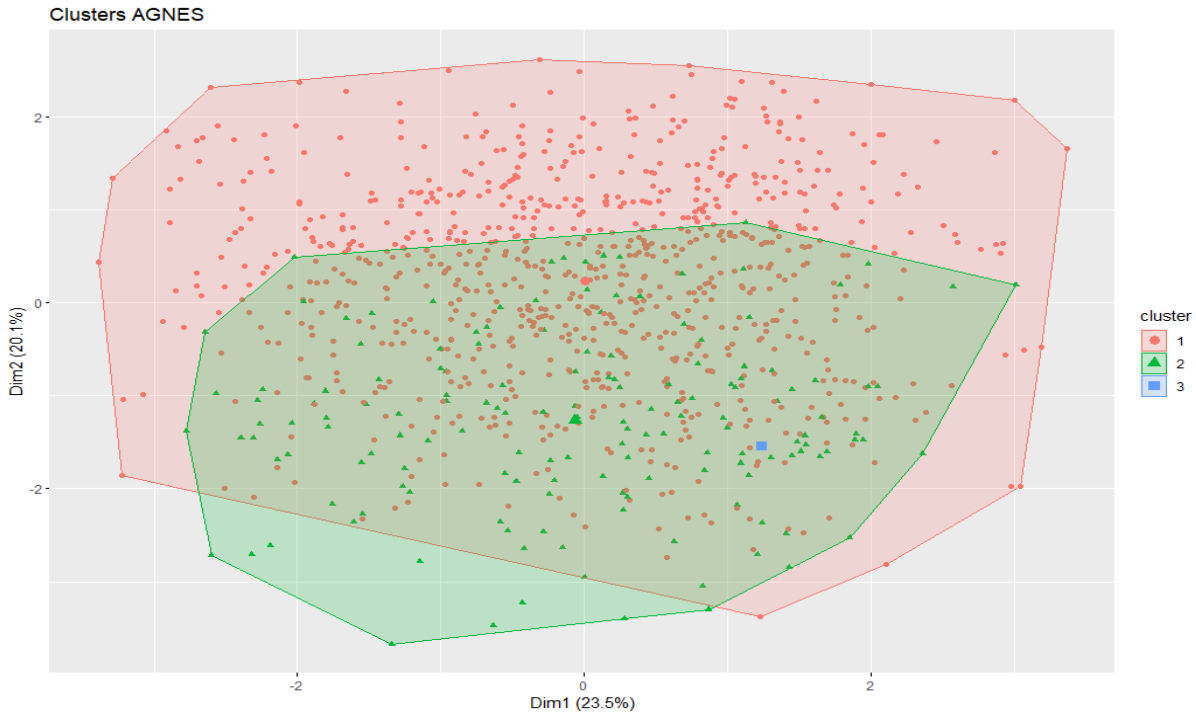


Rysunek 28: Wykres AGNES z metodą łączenia klastrów complete - opracowanie własne

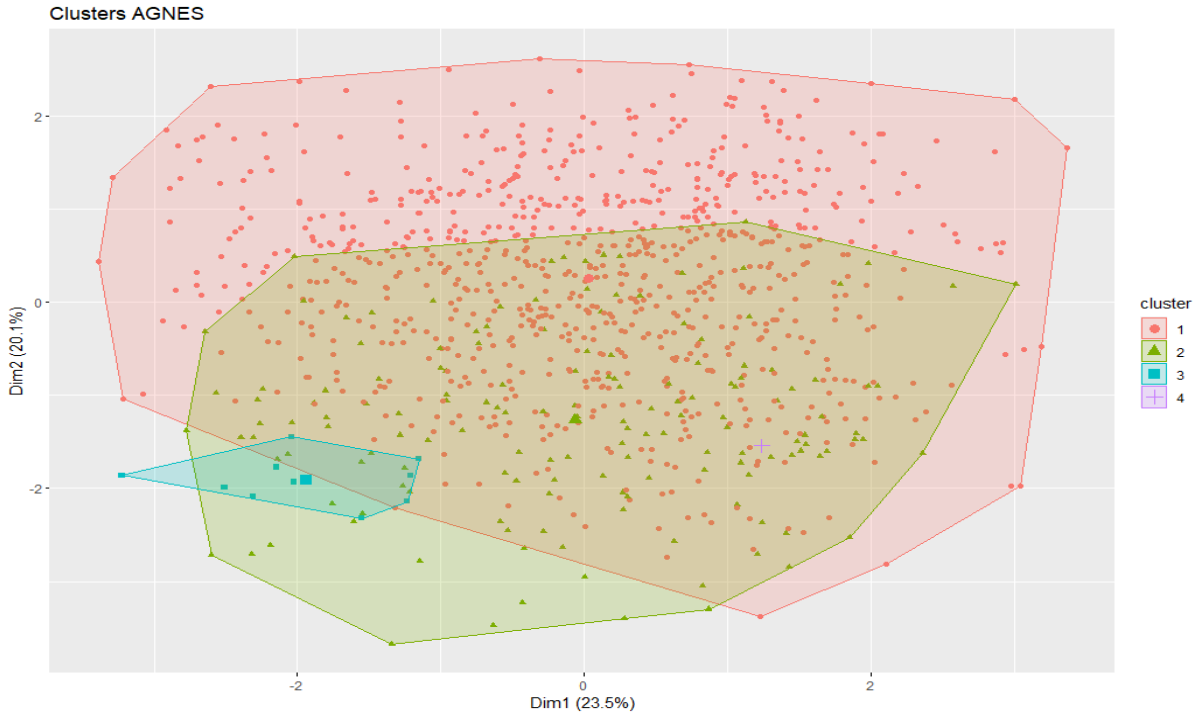


Rysunek 29: Wykres AGNES z metodą łączenia klastrów single - opracowanie własne

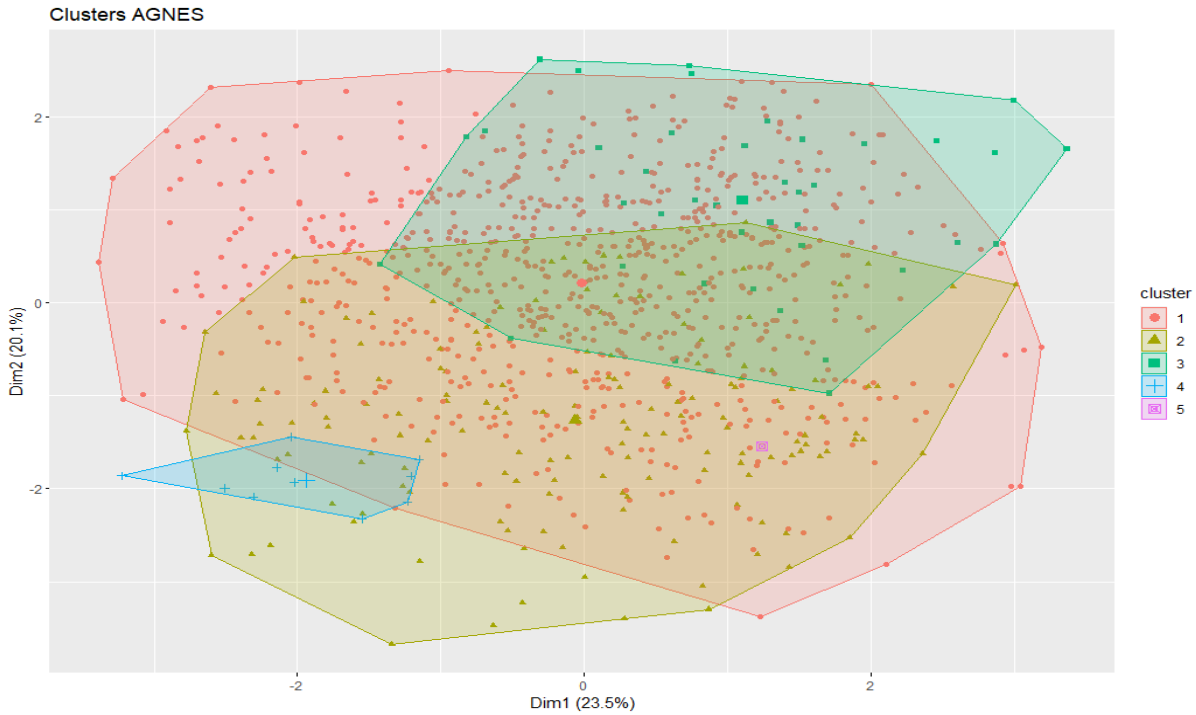
Poniżej przedstawiamy wynik zastosowania metody agnes dla różnych wartości klastrów k.



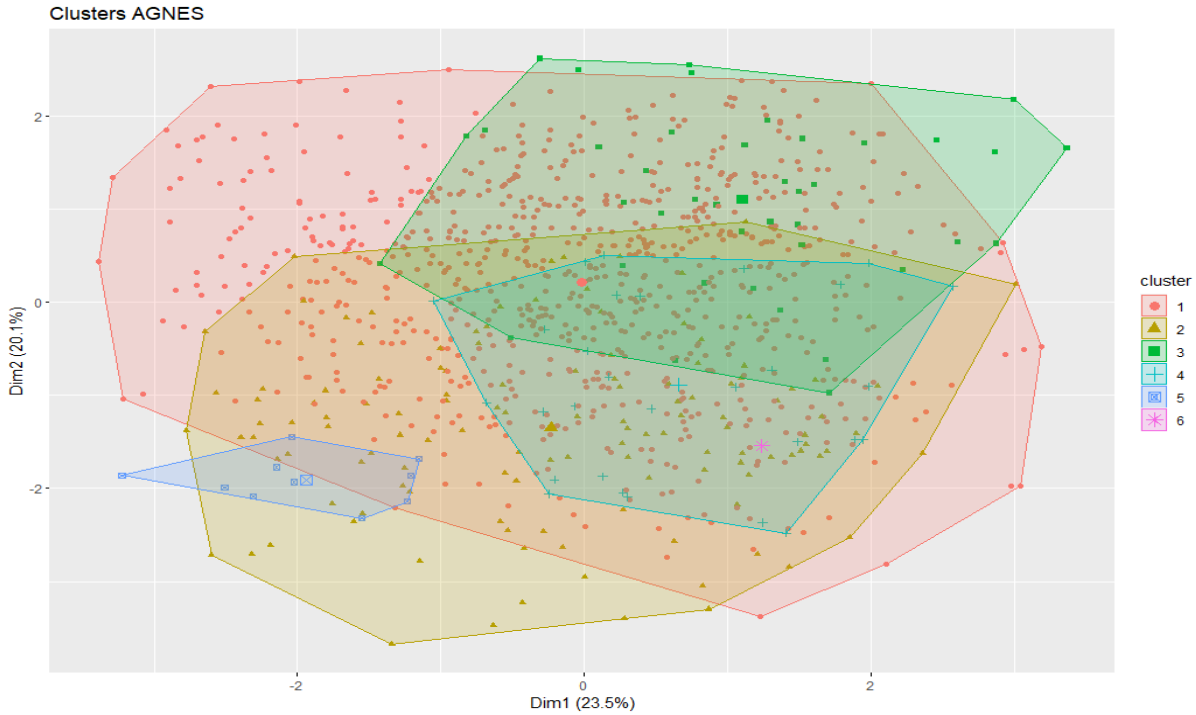
Rysunek 30: Wykres AGNES dla k = 3 - opracowanie własne



Rysunek 31: Wykres AGNES dla k = 4 - opracowanie własne



Rysunek 32: Wykres AGNES dla k = 5 - opracowanie własne



Rysunek 33: Wykres AGNES dla k = 6 - opracowanie własne

Widzimy, że w żadnym z zadanych przypadków nie otrzymaliśmy widocznego rozdzielenia klastrów, bez względu na ich liczbę.

1.2 Zastosowanie wybranej metody redukcji wymiaru w połączeniu z klasyfikacją i analizą skupień

Redukcja wymiaru w analizie skupień polega na zmniejszeniu liczby zmiennych (wymiarów) w zbiorze danych, zachowując jednocześnie jak najwięcej informacji. Jest to proces transformacji danych, który ma na celu zmniejszenie skomplikowanych struktur danych do bardziej przystępnej formy, jednocześnie minimalizując utratę istotnych informacji.
Najpopularniejsze metody redukcji wymiaru w analizie skupień to

- Analiza składowych głównych (PCA): PCA jest jedną z najpopularniejszych technik redukcji wymiaru. Polega na przekształceniu zbioru danych w nowe zestawy zmiennych (tzw. składowe główne), które są liniowymi kombinacjami oryginalnych zmiennych. Składowe główne są uporządkowane według stopnia wariancji, które wyjaśniają w danych. W praktyce można wybrać tylko kilka pierwszych składowych głównych, które wyjaśniają większość wariancji, a tym samym zmniejszyć wymiar danych.

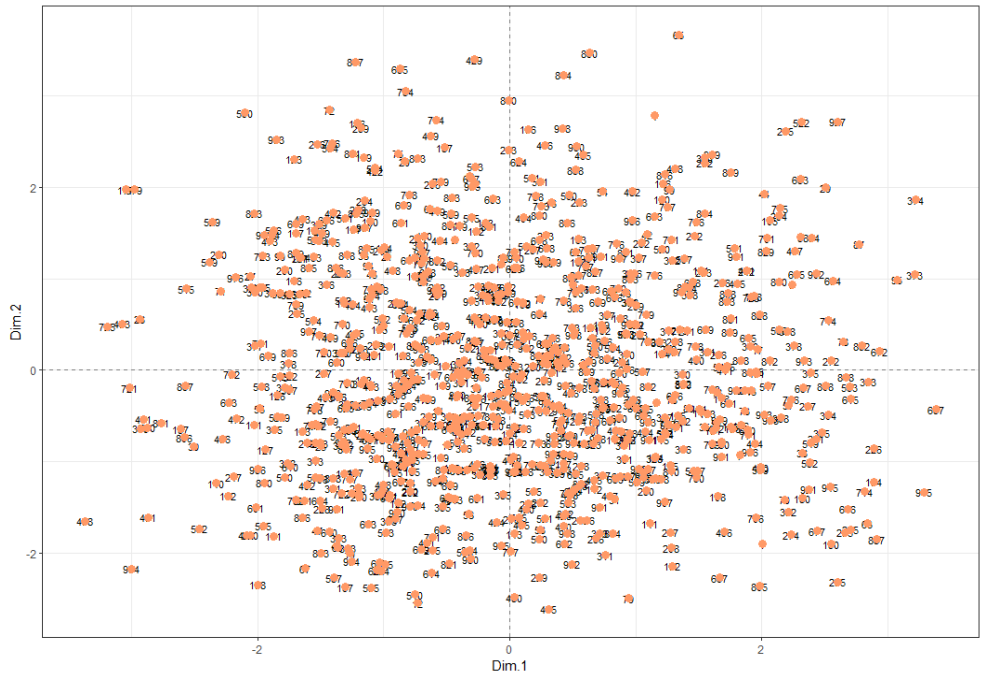
- Analiza czynnikowa (FA): FA jest podobna do PCA, ale zakłada, że zmienne obserwowane są wynikiem ukrytych czynników. FA identyfikuje te ukryte czynniki i próbuje wyjaśnić obserwowane zmienne jako kombinacje tych czynników. Może być stosowana jako technika redukcji wymiaru w analizie skupień.

- Uczenie nienadzorowane: Niektóre algorytmy uczenia maszynowego, takie jak autokodery, mogą być wykorzystywane do redukcji wymiaru danych. Autokoder to sieć neuronowa, która próbuje odtworzyć dane wejściowe na wyjściu. Jednak wewnętrzne warstwy autokodera reprezentują mniejszą liczbę wymiarów, co prowadzi do redukcji wymiaru danych.

- Selekcja cech: Ta technika polega na wyborze najbardziej informatywnych cech (zmiennych) na podstawie pewnych kryteriów, takich jak ważność cech, współczynniki korelacji, informacja wzajemna itp. Poprzez wyeliminowanie mniej ważnych cech, można zmniejszyć wymiar danych.

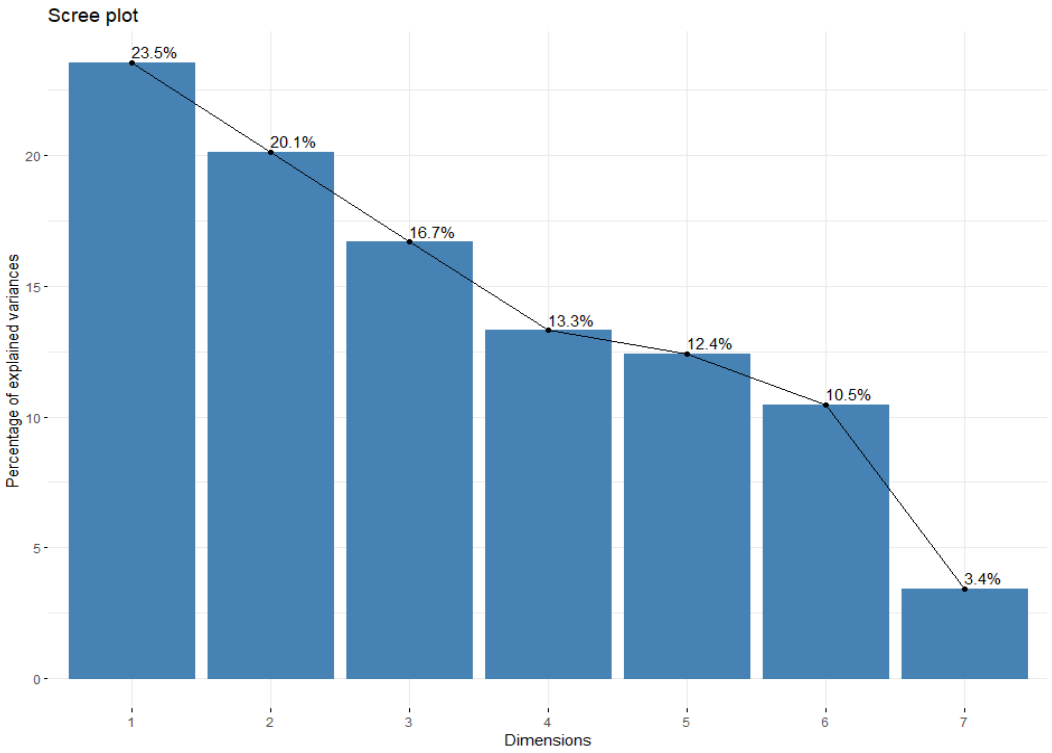
Warto dodać, że redukcja wymiaru w analizie skupień może pomóc w redukcji złożoności danych, eliminacji zbędnych szumów i nieistotnych informacji oraz ułatwieniu interpretacji wyników. Jednak warto pamiętać, że w niektórych przypadkach redukcja wymiaru może prowadzić do utraty istotnych szczegółów, dlatego ważne jest rozważenie tego aspektu i ocena wpływu redukcji wymiaru na analizę skupień.

Korzystając z metody PCA z pakietu `FactoMineR` zademonstrujemy teraz redukcję wymiaru. Wyniki PCA będą zawierały informacje o wartościach własnych, udziałach procentowych wariancji w poszczególnych składowych głównych, a także inne statystyki dotyczące poszczególnych składowych.



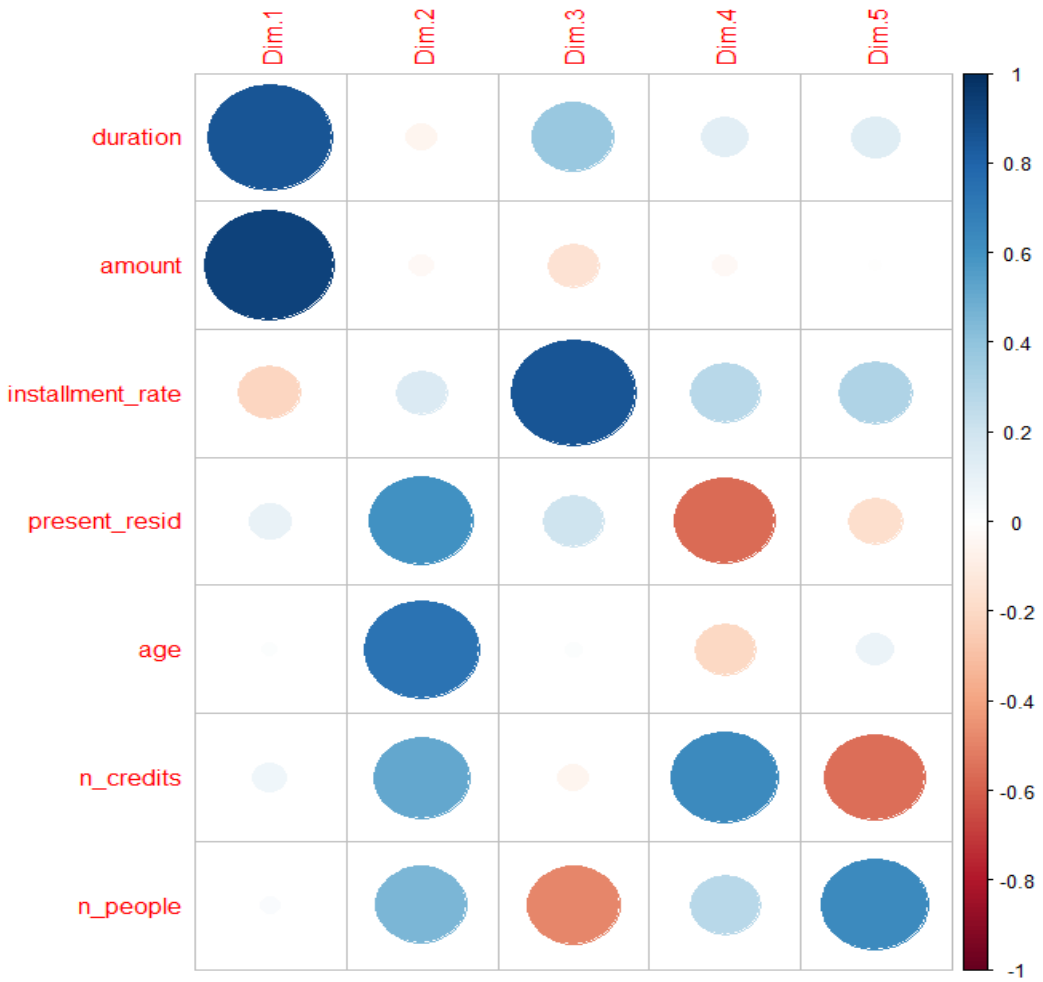
Rysunek 34: Wykres metody PCA - opracowanie własne

Dalej prezentujemy scree plot. Wykres ten przedstawia wartości własne (eigenvalues) dla kolejnych składowych głównych uzyskanych w wyniku analizy PCA (Principal Component Analysis). Składowe główne reprezentują nowe zmienne, które są liniowymi kombinacjami pierwotnych zmiennych. Na osi poziomej wykresu Scree plot przedstawiane są numery składowych głównych, natomiast na osi pionowej znajdują się odpowiadające im wartości własne. Wartość własna mierzy ilość wariancji w danych, która jest wyjaśniana przez daną składową główną. Im większa wartość własna, tym większy wkład w ogólną wariancję danych. Wykres Scree plot jest przydatny w analizie PCA, ponieważ pozwala ocenić, jak wiele składowych głównych jest potrzebnych do wyjaśnienia większości wariancji w danych. Na ogół, im większa wartość własna dla danej składowej głównej, tym większy jej wkład w wyjaśnienie wariancji. Punkt, w którym wartości własne zaczynają gwałtownie spadać, wskazuje na granicę, po której dalsze składowe główne mają znacząco mniejszy wkład i mogą być pomijane.



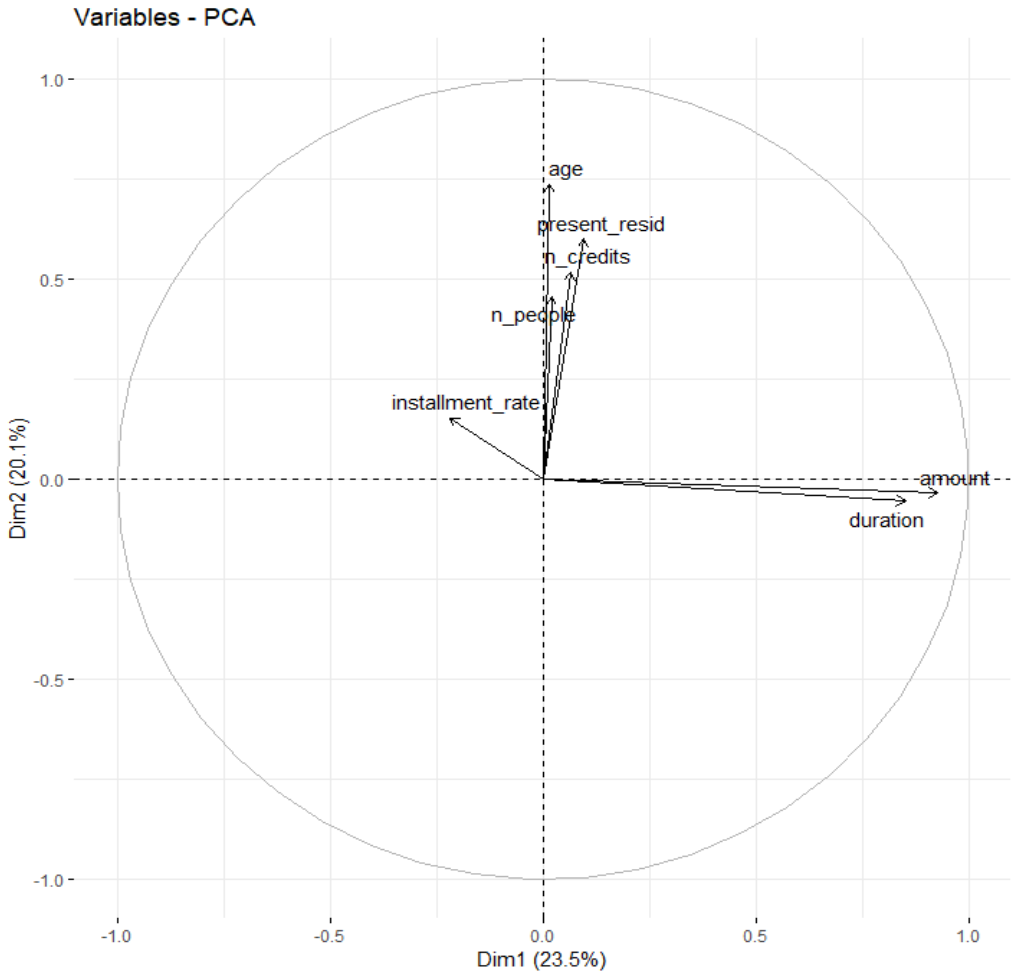
Rysunek 35: Scree Plot metody PCA - opracowanie własne

Dalej przedstawimy `corrplot`, czyli wykres korelacji między zmiennymi a składowymi głównymi uzyskanymi w wyniku analizy PCA. Na wykresie przedstawione są współczynniki korelacji między każdą zmienną a poszczególnymi składowymi głównymi. Wartości korelacji są reprezentowane przez kolory lub intensywność kolorów na wykresie. Typowy wykres corrplot przedstawia macierz korelacji, w której wartości dodatnie i silne korelacje są zazwyczaj reprezentowane przez jasne kolory, a wartości ujemne i słabe korelacje przez ciemne kolory. Ten wykres pozwala zobaczyć, jak poszczególne zmienne wpływają na składowe główne. Silna korelacja między zmienną a składową główną oznacza, że dana zmienna ma duży wpływ na tę składową główną. Można także zauważyć, czy istnieją grupy zmiennych, które są silnie skorelowane ze sobą i jednocześnie mają silne korelacje z tą samą składową główną. Jest przydatny w analizie PCA, ponieważ umożliwia wizualizację związków między zmiennymi a składowymi głównymi. Może pomóc w identyfikacji zmiennych, które mają największy wpływ na poszczególne składowe główne oraz zrozumieniu, jakie wzorce lub zależności między zmiennymi są reprezentowane przez te składowe główne.



Rysunek 36: Corrplot metody PCA - opracowanie własne

Ostatnim już wykresem jest wykresem prezentujący wpływ poszczególnych zmiennych na wartości dwóch pierwszych składowych głównych uzyskanych w wyniku analizy PCA (Principal Component Analysis). Na wykresie przedstawione są zmienne jako wektory w przestrzeni dwuwymiarowej, gdzie osie odpowiadają dwóm pierwszym składowym głównym. Długość i kierunek wektora reprezentuje wkład danej zmiennej w daną składową główną. Kolor punktów i strzałek wskazuje na wartość danej zmiennej w kontekście składowych głównych. Pozwala on zobaczyć, jak poszczególne zmienne wpływają na położenie obserwacji w przestrzeni dwuwymiarowej utworzonej przez dwie pierwsze składowe główne. Punkty i strzałki o większej długości wskazują na większy wkład danej zmiennej w daną składową główną. Można również zauważyć, które zmienne są podobne lub odległe od siebie w kontekście składowych głównych.



Rysunek 37: Wykres prezentujący wpływ poszczególnych zmiennych na wartości dwóch pierwszych składowych - opracowanie własne

Literatura

[1] Adam Zagdański *Eksploracyjna analiza danych (EDA) - wprowadzenie*, https://eportal.pwr.edu.pl/pluginfile.php/559567/mod_resource/content/0/EDA_wprowadzenie.pdf