SPRAWOZDANIE

Zajęcia: Zbiory Big Data i Eksploracja Danych

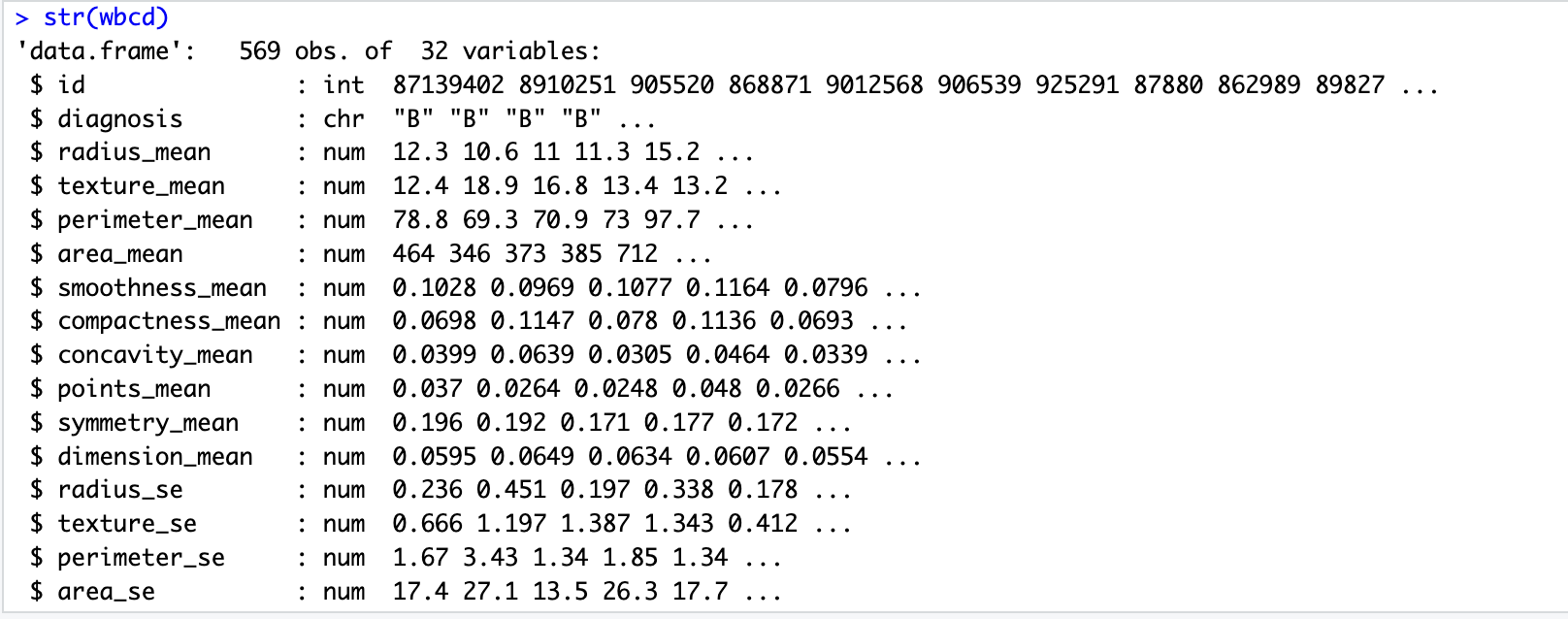
Prowadząca: dr inż. Ruslana Ziubina

|  |  |
| --- | --- |
| Laboratorium nr 6  Data rozpoczęcia: 12.01.2024  Temat: Klasyfikacja i grupowanie | Rafał Klinowski  Informatyka  II stopień, stacjonarne,  Semestr 2, gr. a |

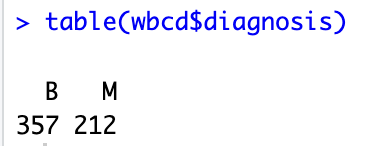
Poszczególne ćwiczenia będą wykonywane w pliku źródłowym edytowanym przy pomocy środowiska RStudio, opisanego w poprzednich laboratoriach.

# Ćw. 1.

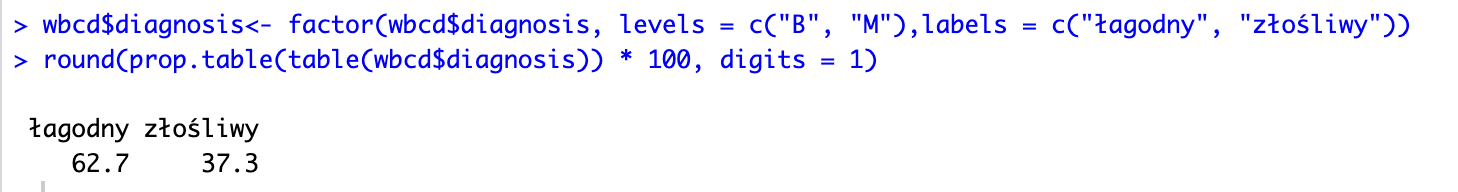
Celem było utworzenie klasyfikatora opartego o algorytm kNN dla danych dotyczących raka.



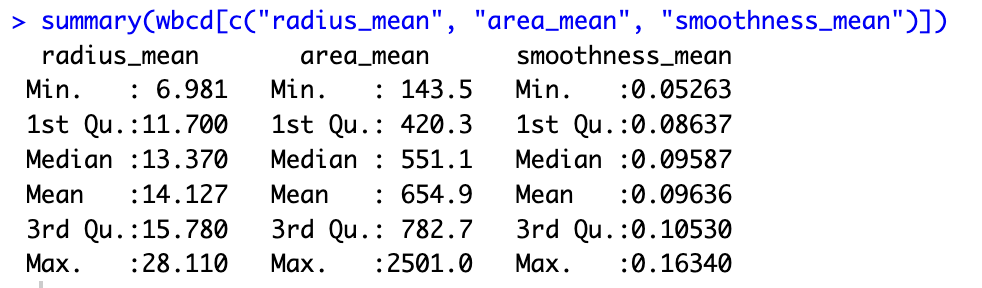
Rysunek . Fragment podsumowania zbioru danych.



Rysunek . Rozkład zmiennej – podział na dwie klasy występujące w zbiorze.

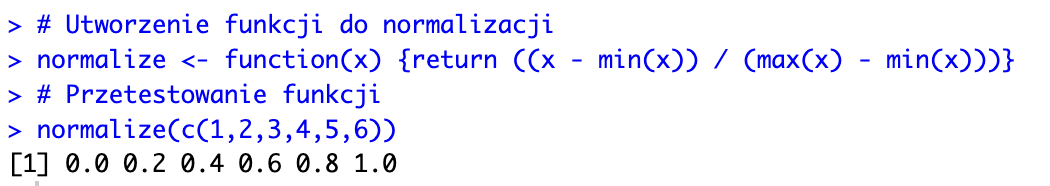


Rysunek . Zmiana oznaczeń klas i podgląd rozkładu procentowego podziału na klasy.



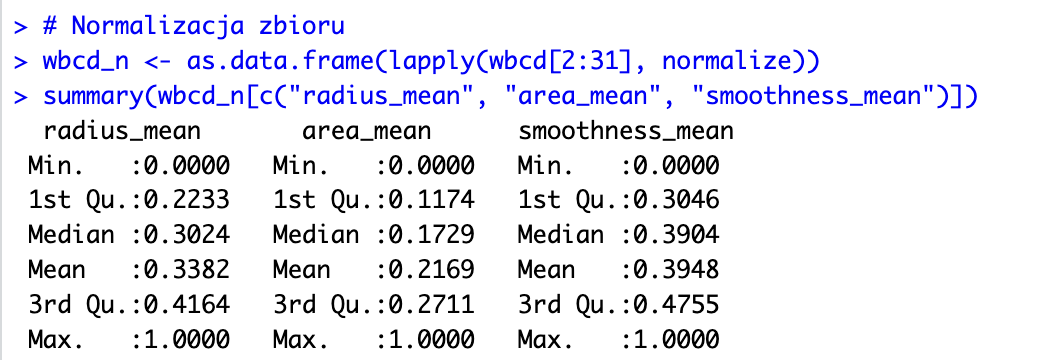
Rysunek . Wyświetlenie wartości dla pierwszych trzech zmiennych.

Ponieważ zmienne mają zupełne różne zakresy wartości, wymagana jest normalizacja danych.



Rysunek . Utworzenie funkcji przeprowadzającej normalizację i przetestowanie jej dla przykładowych danych.

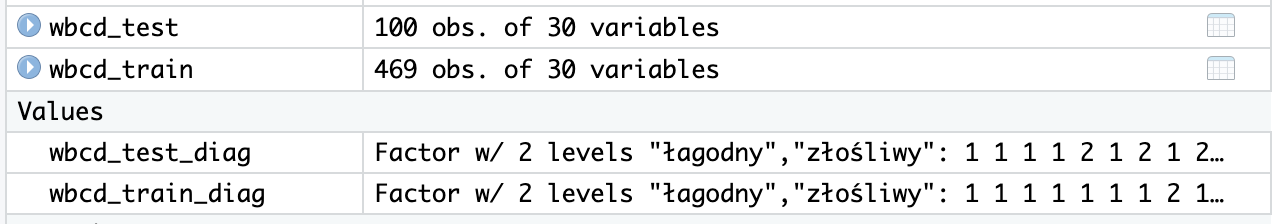
Następnie przeprowadzono normalizację wszystkich kolumn liczbowych ze zbioru wejściowego.



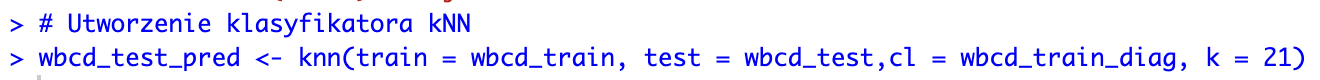
Rysunek . Normalizacja i wyświetlenie nowych zakresów danych.

Do normalizacji zbioru danych wykorzystano funkcję lapply, która wywołuje podaną funkcję normalize dla wszystkich określonych kolumn danych.

Kolejnym krokiem było zbudowanie zbioru treningowego i testowego.

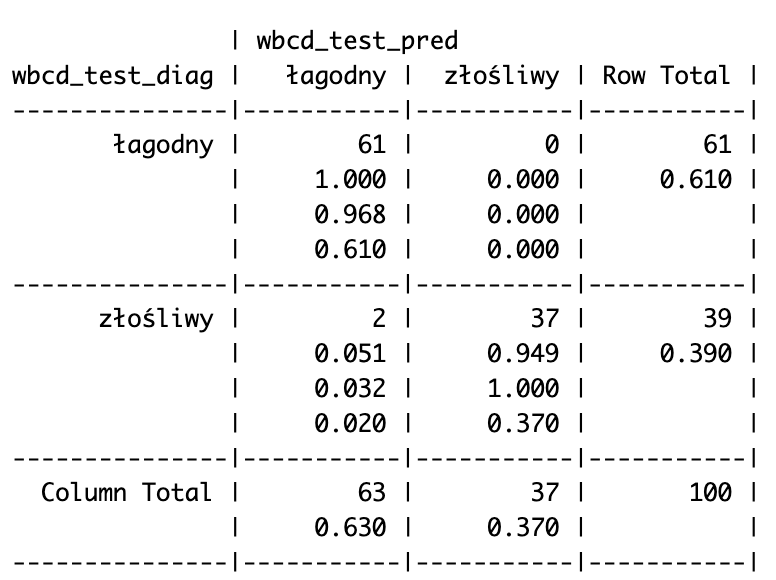


Rysunek . Utworzone zbiory. Zbiór treningowy i testowy zawierają atrybuty, a zmienne „diag” zawierają tylko klasy, do których należy dany rekord ze zbioru danych.



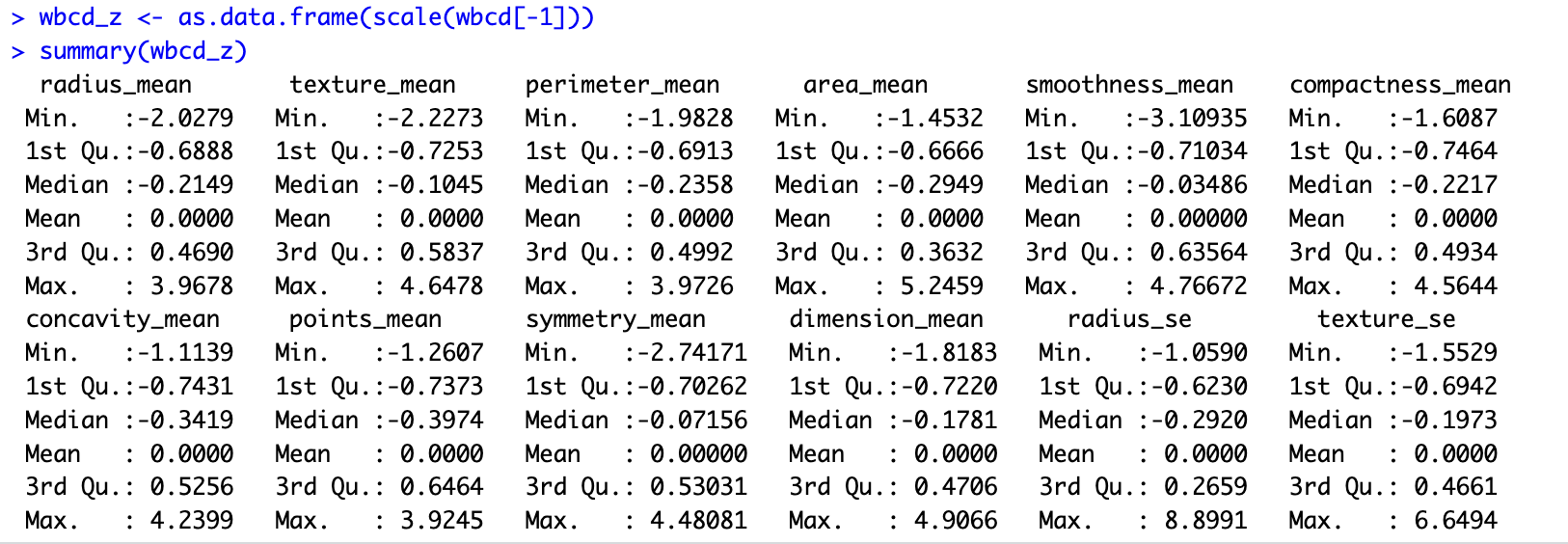
Rysunek . Utworzenie klasyfikatora kNN dla powyższych zbiorów oraz wartości domyślnej k=21.

Uzyskano wyniki klasyfikacji, które należy teraz porównać do wyników oczekiwanych („wbcd\_train\_diag”).

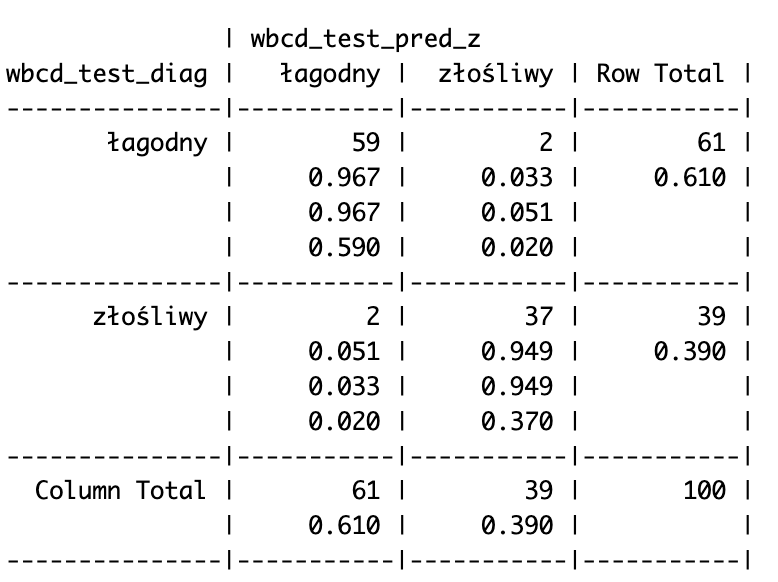


Rysunek . Uzyskana tabela z informacjami o poprawnych i niepoprawnych klasyfikacjach.

Powtórzymy teraz proces tworzenia zbioru danych i wykorzystania kNN po standaryzacji danych.



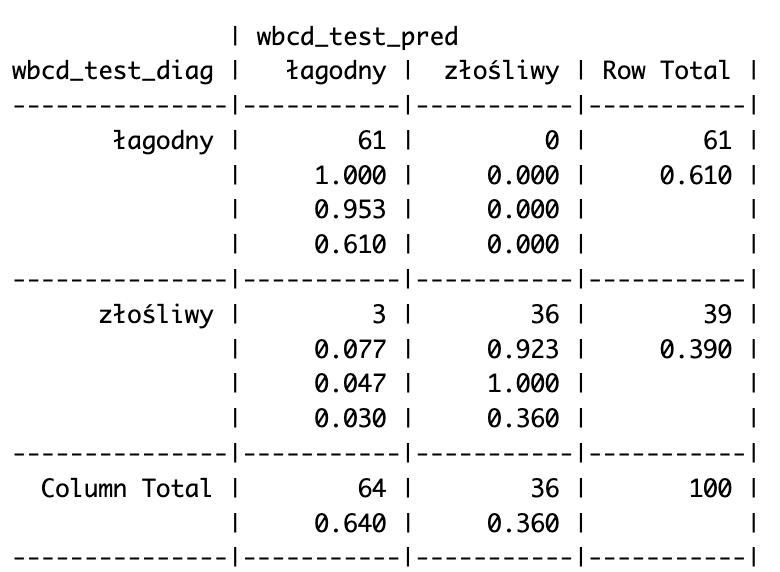
Rysunek . Zbiór danych po standaryzacji. W oczy rzucają się wartości ujemne kolumn.



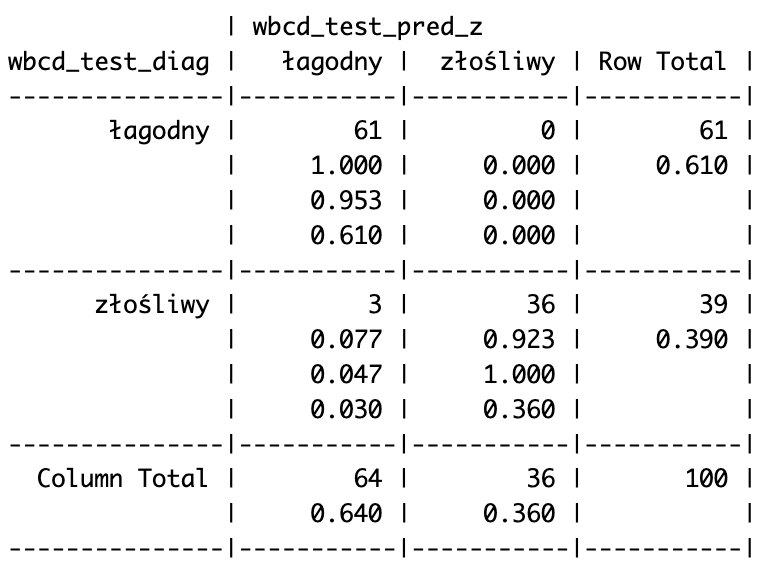
Rysunek . Tabela uzyskana dla drugiego przypadku.

Dla tych danych i parametrów uzyskano lepszą dokładność w przypadku normalizacji danych, niż w przypadku standaryzacji.

Przetestujmy teraz oba przypadki dla innych wartości parametru k.

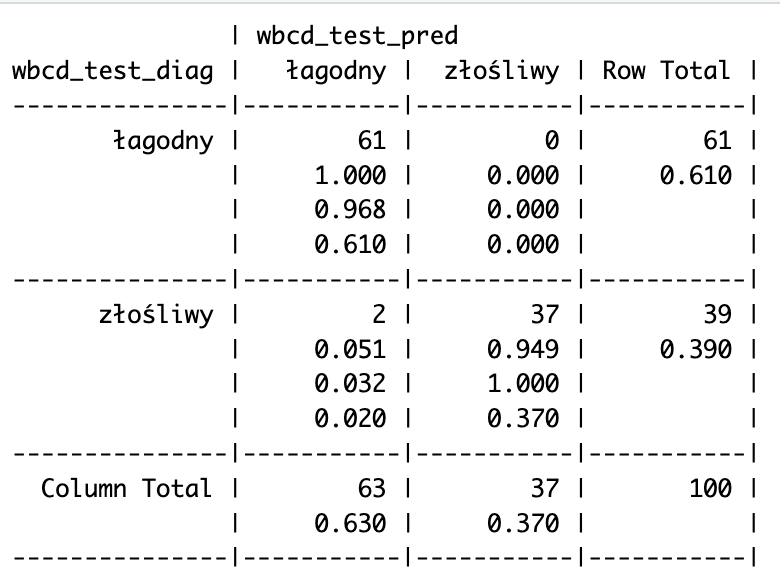


Rysunek . Klasyfikator kNN, k=15, dane normalizowane.

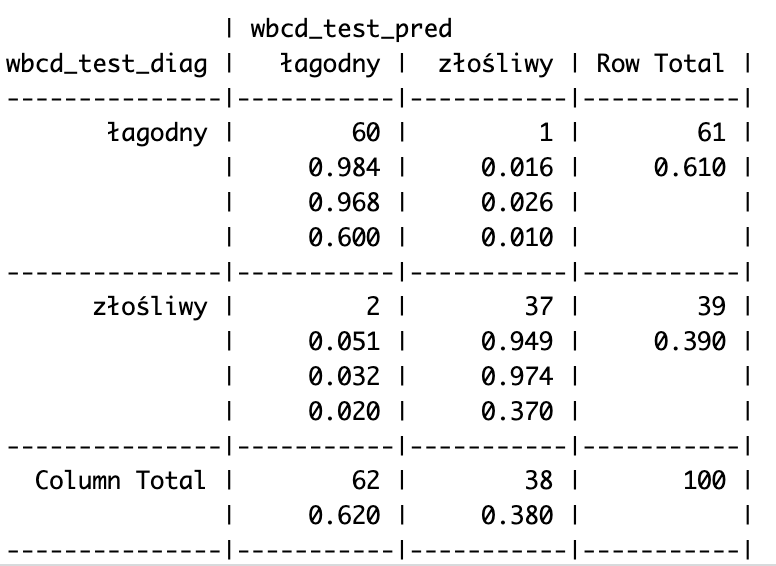


Rysunek . Klasyfikator kNN, k=15, dane standaryzowane.

Dla k=15 uzyskano identyczne wyniki, które nieznacznie różnią się od uzyskanych dla k=21.



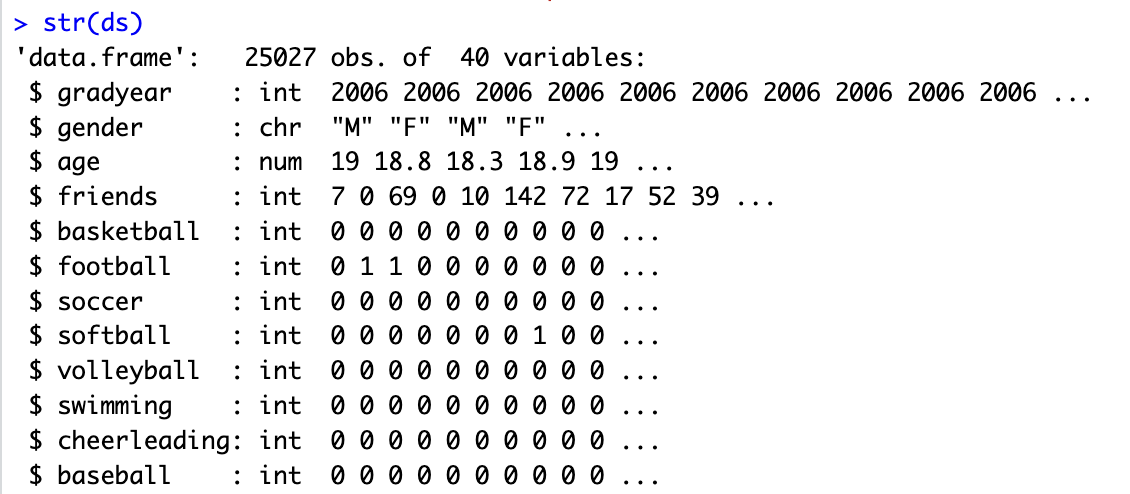
Rysunek . Klasyfikator kNN, k=5, dane normalizowane.



Rysunek . Klasyfikator kNN, k=3, dane normalizowane.

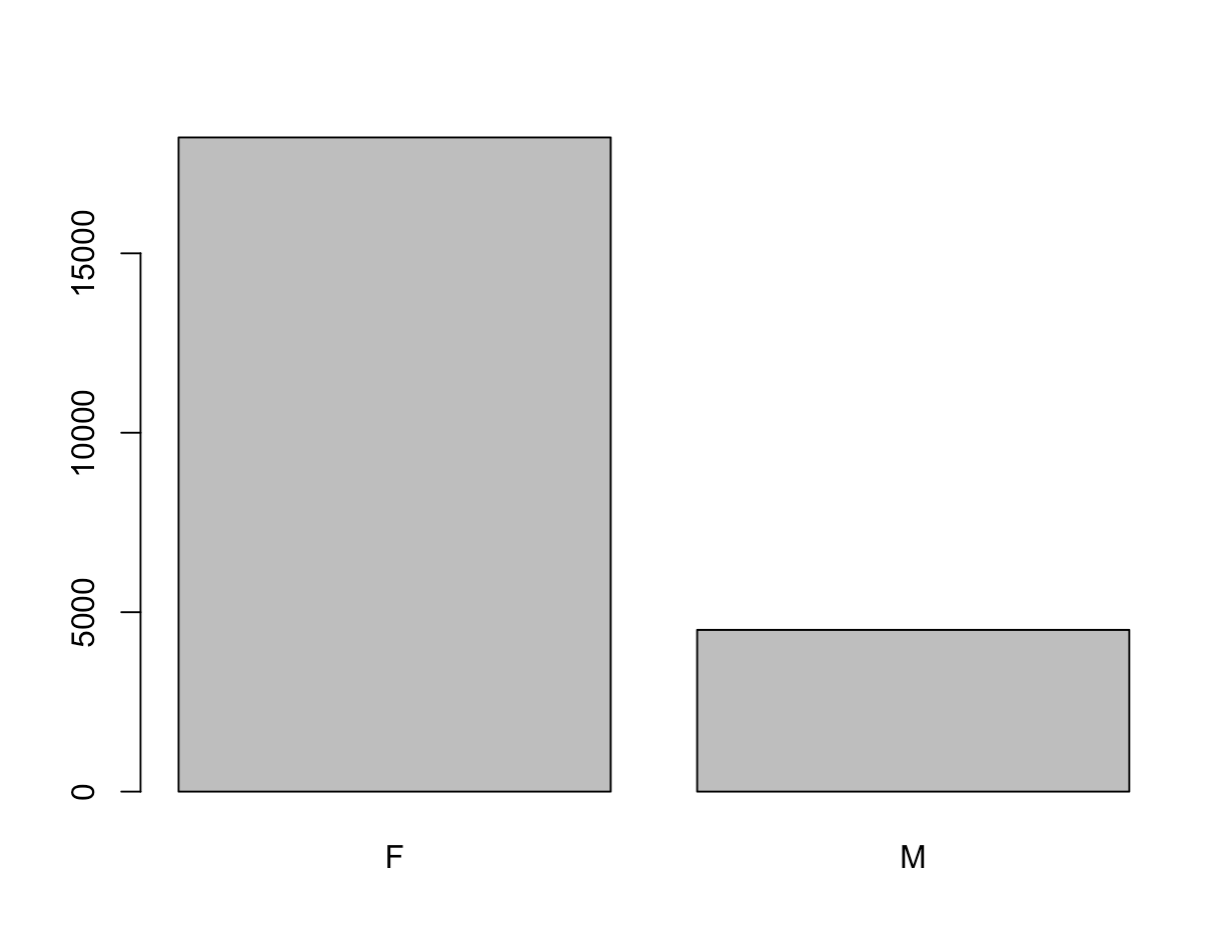
# Ćw. 2.

W tej części utworzono, w oparciu o klasyfikator „k średnich”, model grupowania danych dotyczących aktywności w portalach społecznościowych studentów szkół średnich.



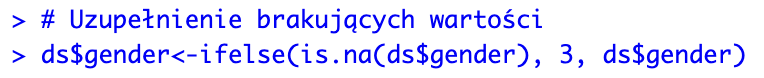
Rysunek . Fragment wczytanego zbioru danych.

Eksploracja danych – utworzono histogram płci występujących w zbiorze danych.

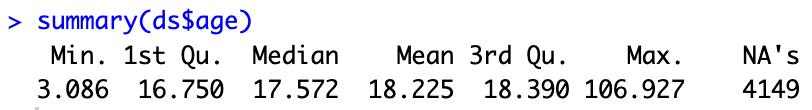


Rysunek . Histogram płci w zbiorze danych.

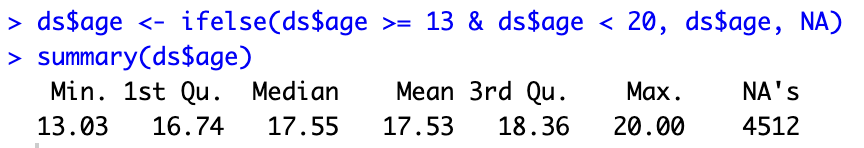
Należało rozwiązać problem brakujących danych w kolumnie gender i age. W przypadku gender, brakujące dane ustawiamy na wartość 3. Dla age, odrzucone zostały wartości niepoprawne i oddalone (np. osób, które są zbyt młode lub zbyt stare, by uczęszczać do szkoły średniej).



Rysunek . Zamiana brakujących wartości zmiennej „gender” na 3.

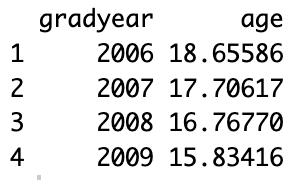


Rysunek . Dane dotyczące zmiennej age.

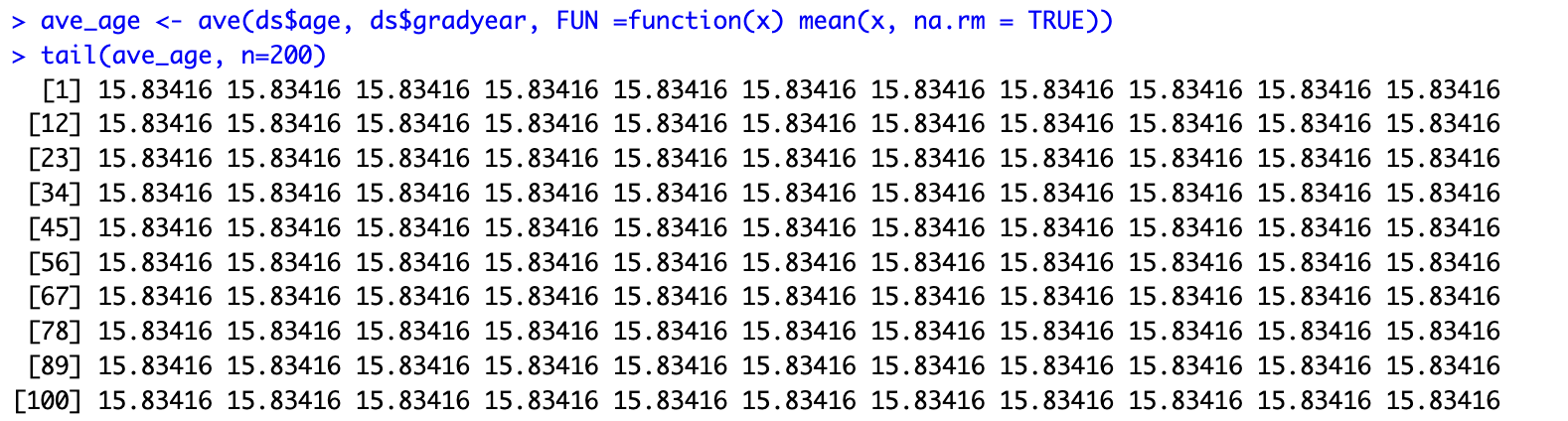


Rysunek . Zamiana wartości niepoprawnych i ponowne wyświetlenie danych dla tej zmiennej.

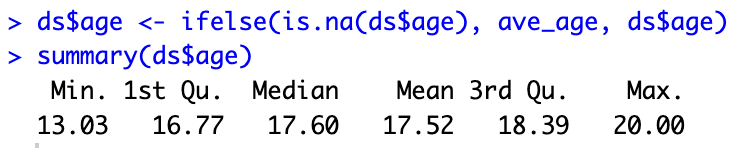
Teraz brakujące dane dotyczące wieku zostały zastąpione średnią posiadanych w zbiorze, w zależności od roku ukończenia szkoły.



Rysunek . Obliczony średni wiek ucznia w zależności od roku, w którym skończył szkołę.



Rysunek . Utworzenie wektora zawierającego odpowiedni wiek dla każdego ucznia po kolei, na podstawie wartości średniej zależnej od roku ukończenia szkoły.



Rysunek . Dane dotyczące wieku po zamianie brakujących wartości. Wartości średnie i kwartyli tylko nieznacznie się zmieniły.

Teraz konieczna była normalizacja danych. Przeprowadzono ją dla wszystkich kolumn „zainteresowań”, które będą istotne w procesie grupowania. Ponownie wykorzystano w tym celu funkcję lapply.



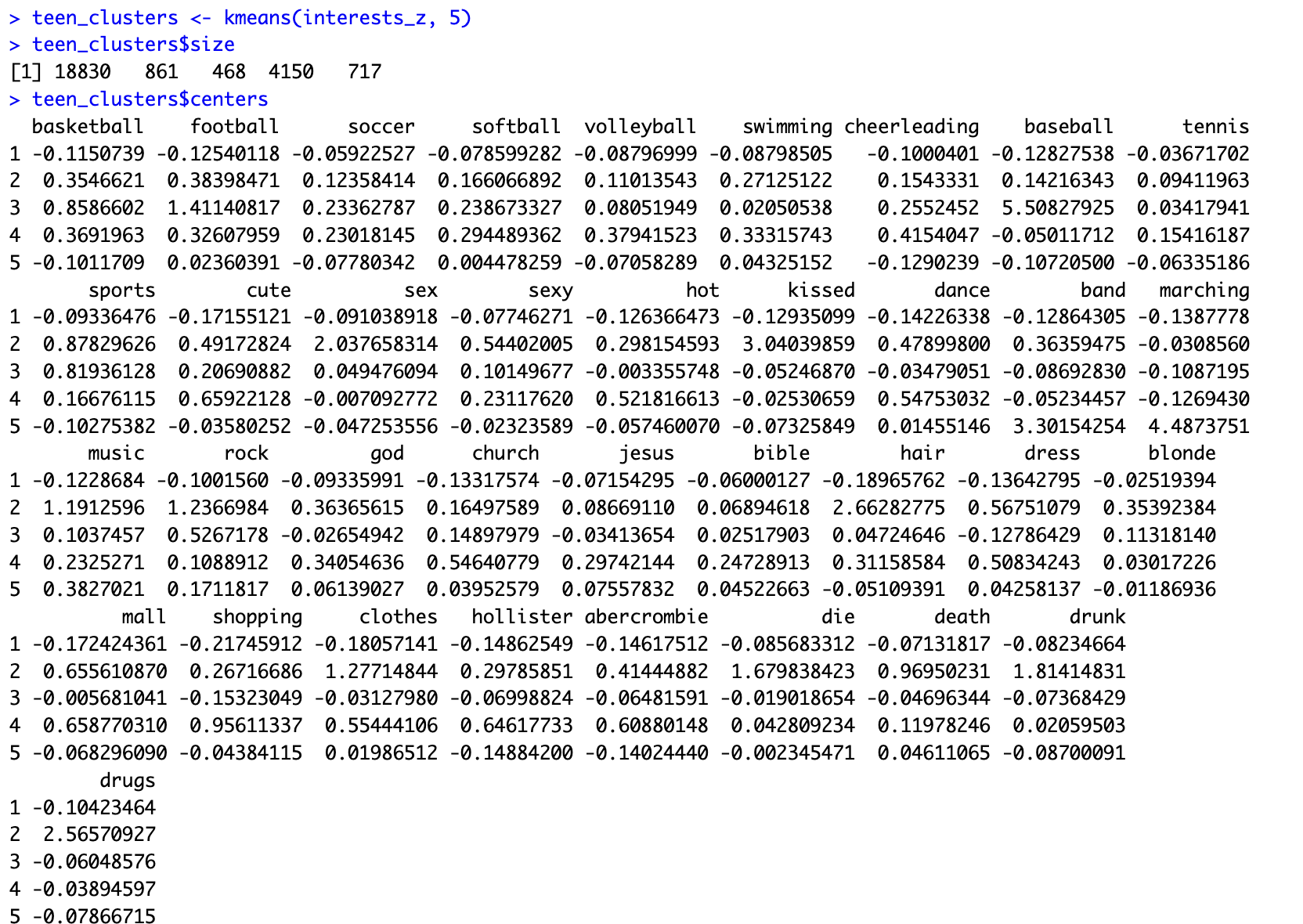
Rysunek . Przeprowadzenie normalizacji danych dotyczących zainteresowań.

Podczas normalizacji uzyskano informację o brakujących danych w wybranych kolumnach. Konieczna jest więc ich eliminacja. W tym celu wykorzystano funkcję na.omit().



Rysunek . Usunięcie brakujących danych przy pomocy funkcji na.omit(). Usunięty został 1 rekord.

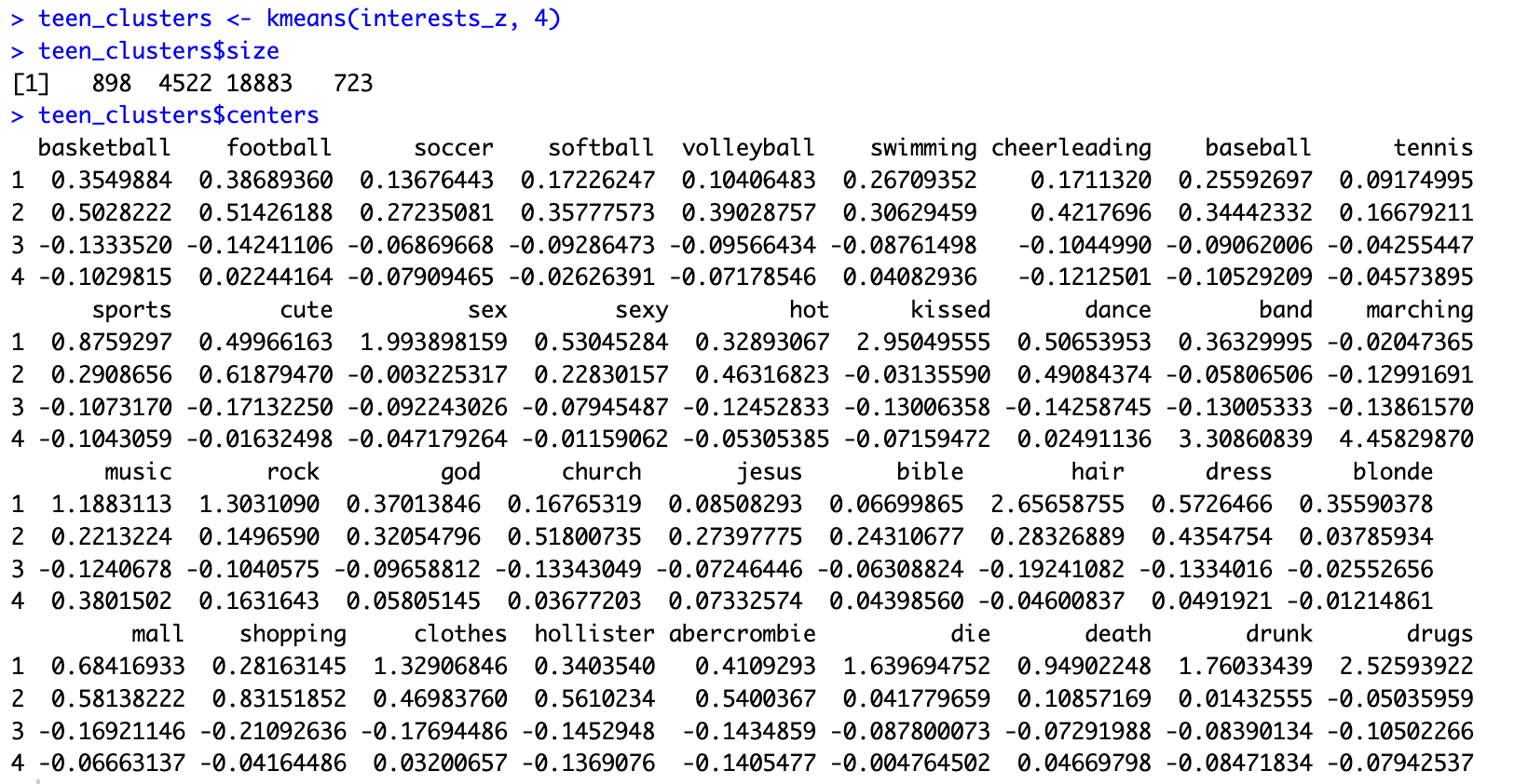
Teraz możliwe było grupowanie danych w oparciu o algorytm „k-Means”.



Rysunek . Przeprowadzenie grupowania danych z podziałem na 5 klas. Wyświetlenie liczności tych klas oraz ich środków dla poszczególnych atrybutów.

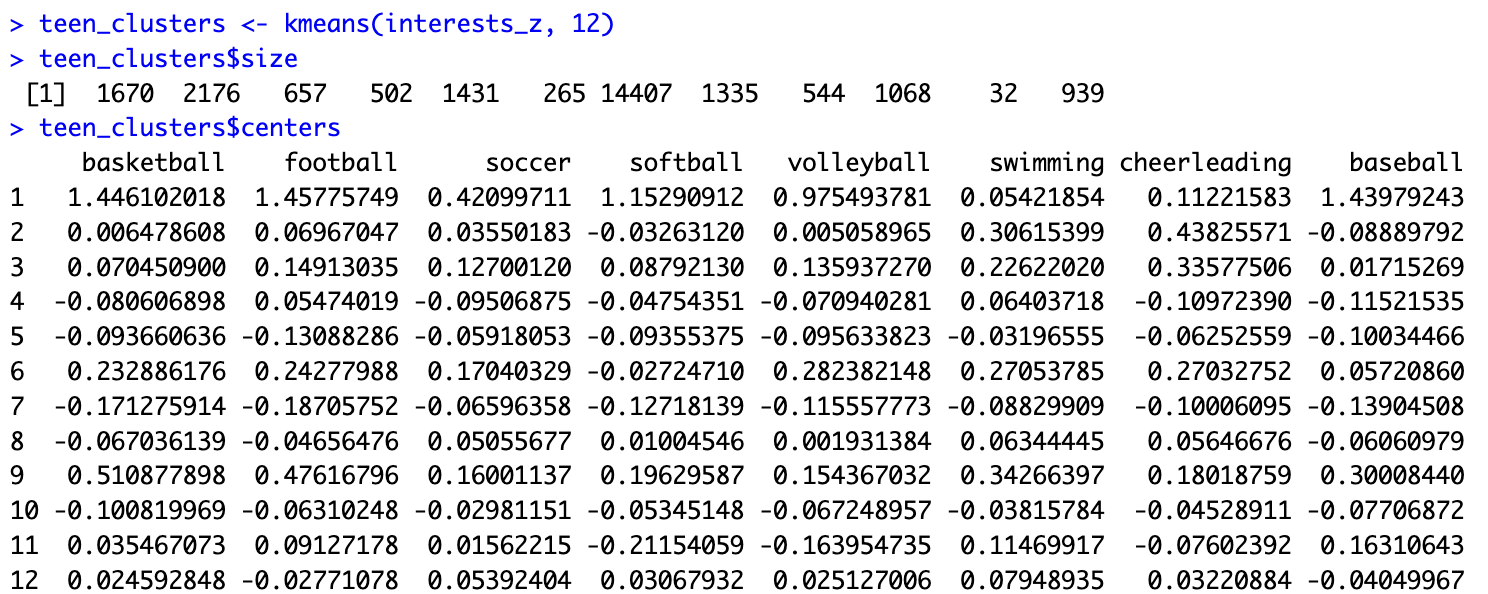
Przekazywany parametr 5 określa ilość grup, jakiej oczekujemy. W przypadku takich danych nie wiemy, ile grup w nich występuje, w związku z tym sprawdzimy wyniki algorytmu dla kilku parametrów.

Powtórzmy powyższe operacje dla ilości grup równej 4.



Rysunek . Przeprowadzenie grupowania z podziałem na 4 klasy.

Sprawdźmy, co się stanie dla większej ilości grup (k=12).



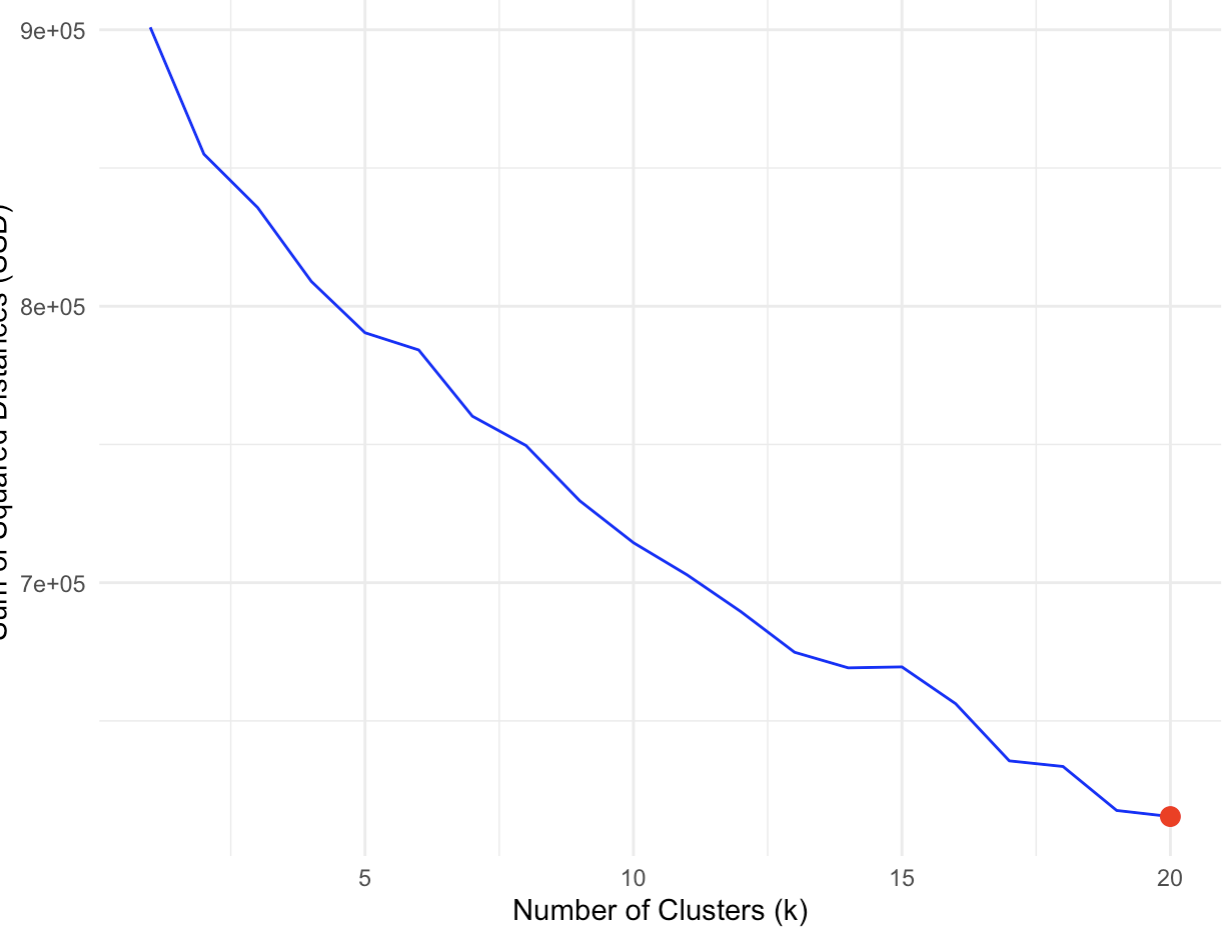
Rysunek . Przeprowadzenie grupowania z podziałem na 12 klas.

# Wnioski.

W przypadku klasyfikatora kNN, dla danych dotyczących raka nie uzyskano większych różnic w jakości uzyskanych wyników po zmianie wartości k. Najlepsze wyniki uzyskano dla k=5, jednak nawet dla większych wartości (15, 21) utworzono klasyfikator, który z dobrą skutecznością rozpoznawał klasę, do której należy każdy przypadek na podstawie innych atrybutów. Warto jednak zauważyć, że zbiór danych wejściowych, a w szczególności danych testowych, był dość mały, w związku z czym klasyfikator może nie radzić sobie tak dobrze dla przypadków, których nie było w oryginalnym zbiorze. Niemniej jednak przykład ten pokazuje, że nawet prosty klasyfikator oparty o algorytm „k najbliższych sąsiadów” może być skutecznie wykorzystywany w celach diagnozy.

W przypadku grupowania przy pomocy algorytmu k-Means widzimy, że – zarówno dla k=5, k=4 jak i k=12 – jedna z uzyskanych grup jest znacznie większa od pozostałych. Na tej podstawie można wyciągnąć wniosek, że większość uczniów szkoły średniej ma zbliżone zainteresowania. Nawet w przypadku podziału na 12 grup uzyskujemy taką, która jest kilkukrotnie większa od pozostałych, a więc mimo że uzyskujemy więcej szczegółowych grup, i tym samym możliwe jest znalezienie dodatkowych zależności w danych, algorytm k-Means nadal przyporządkowuje większość uczniów do tej samej grupy.

W celu znalezienia optymalnej ilości grup można przeprowadzić analizę na kilka sposobów. Jednym z nich jest metoda polegająca na obliczeniu, dla różnych wartości k, sumy kwadratów odległości punktów w grupie od wyznaczonego środka tej grupy. Następnie możliwe jest określenie k jako wartości, dla której suma ta przestaje rosnąć w dużym tempie („zakrzywia się”).



Rysunek . Przeprowadzenie opisanej wyżej metody wyznaczenia optymalnej wartości k. Widzimy, że dla tego zbioru danych, wartość reprezentująca sumę kwadratów odległości od środków przestaje szybko spadać dopiero dla dużych wartości k (w okolicy 18).