Eksploracja Danych

Temat: Rozpoznawanie typu Pokemona po jego cechach.

Adam Kasperowicz - 279046 , Mateusz Sieczko - numerek

Spis treści

Problem

Celem projektu jest klasyfikacja typu pokemonów na podstawie ich atrybutów. Oznacza to, że powinniśmy mieć możliwość określenia typu pokemona wyłącznie po jego atrybutach.
Następujące prace zostaną wykonane:

1. Analiza i zrozumienie posiadanych danych
2. Obróbka danych w celu zmaksymalizowanie użyteczności posiadanych danych
3. Wybranie najlepszego klasyfikatora

Analiza danych

Struktura danych

Listing 1: Struktura danych

```
> str(pokemon)
'data.frame': 800 obs. of  13 variables:
 $ X.         : int  1 2 3 3 4 5 6 6 6 7 ...
 $ Name       : Factor w/ 800 levels "Abomasnow",
 "AbomasnowMega Abomasnow",..: 81 330 746 747 103 104 100 101 102 666 ...
 $ Type.1     : Factor w/ 18 levels "Bug","Dark",
 "Dragon",..: 10 10 10 10 7 7 7 7 7 18 ...
 $ Type.2     : Factor w/ 19 levels "","Bug","Dark",..: 15 15 15 15 1 1 9 4 9 1 ...
 $ Total      : int  318 405 525 625 309 405 534 634 634 314 ...
 $ HP         : int  45 60 80 80 39 58 78 78 78 44 ...
 $ Attack     : int  49 62 82 100 52 64 84 130 104 48 ...
 $ Defense    : int  49 63 83 123 43 58 78 111 78 65 ...
 $ Sp..Atk    : int  65 80 100 122 60 80 109 130 159 50 ...
 $ Sp..Def    : int  65 80 100 120 50 65 85 85 115 64 ...
 $ Speed      : int  45 60 80 80 65 80 100 100 100 43 ...
 $ Generation: int  1 1 1 1 1 1 1 1 1 1 ...
 $ Legendary : Factor w/ 2 levels "False","True": 1 1 1 1 1 1 1 1 1 1 ...
```

Plik “Pokemon.csv” zawiera 800 obiektów, z których każdy opisany jest przez 13 atrybutów. Atrybuty są zarówno typu numerycznego, binarnego jak i łańcuchowego.
Niektóre z atrybutów wymagają głębszego wyjaśnienia.

- **X** - Atrybut ID zliczający ile w zbiorze istnieje pokemonów o niepowtarzalnym atrybucie name.
- **Name, Type.1, Type.2** - Atrybuty łańcuchowe które razem specyfikują jednoznacznie każdego pokemona. Atrybut **Type.2** pojawia się tylko wyjątkowo dla pokemonów posiadających więcej niż jeden typ. Dlatego, często gdy nie ma potrzeby dokładniejszej specyfikacji ten atrybut jest pusty.
- **Total, HP, Attack, Defense, Sp. Atk,Sp. Def, Speed, Generation** - Atrybuty będące liczbami całkowitymi opisują statystki danego pokemona. Każdy z pokemonów ma ten zestaw zmiennych w pełni wypełniony.
- **Legendary** - Atrybut binarny przyjmujący wartość True dla pokemonów legendarnych. Widzimy, że ten atrybut chociaż w większości przypadków będzie miał wartość równą False niesie bardzo ważną informację.

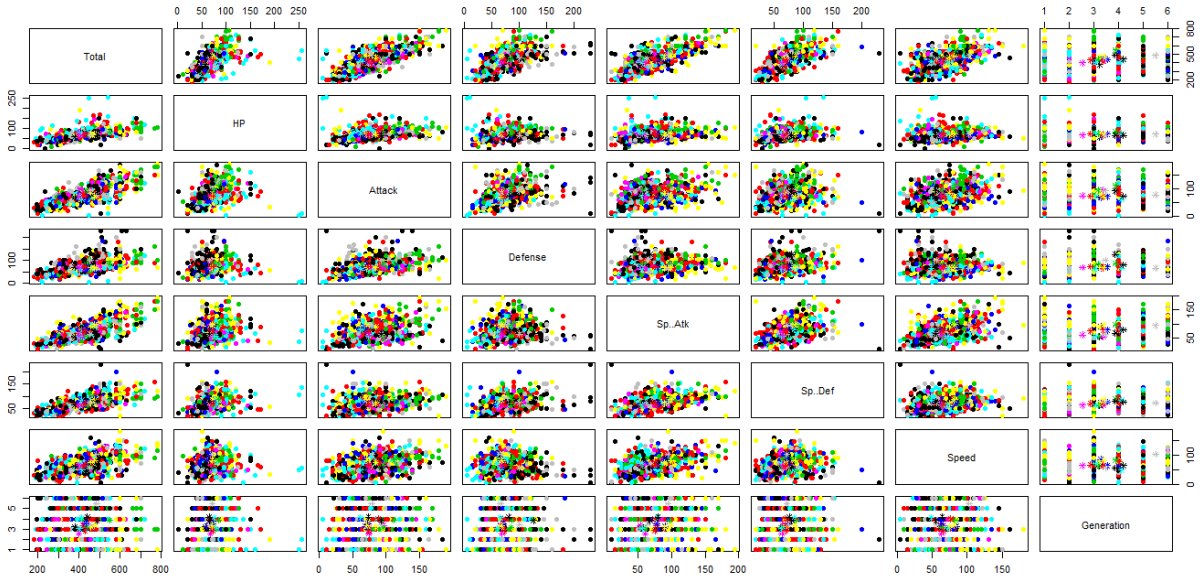
Po przebadaniu struktury danych możemy ustalić parę ważnych faktów.

- W zbiorze danych znajduje się 18 klas. Jest to ilość niepowtarzalnych elementów atrybutu **Type.1**
- Atrybut **X** musi zostać usunięty, gdyż był prawdopodobnie przydzielany arbitralnie i nie przekazuje żadnej informacji.
- Atrybut **Legendary** jest typu łańcuchowego i będzie musiał zostać zmieniony na typ liczb całkowitych o zakresie [0,1].
- Atrybut **Name** jest specyficzny pod tym względem, że prawie jednoznacznie wyznacza klasę obiektu. Atrybut ten sprawia, że problem klasyfikacji praktycznie znika, gdyż już po tylko tej zmiennej możemy określić typ pokemona. Na potrzeby projektu będziemy musieli wykluczyć ten atrybut z procesu klasyfikacji.
- Do klasyfikacji będziemy używać wszystkich atrybutów oprócz **X** i **Name**

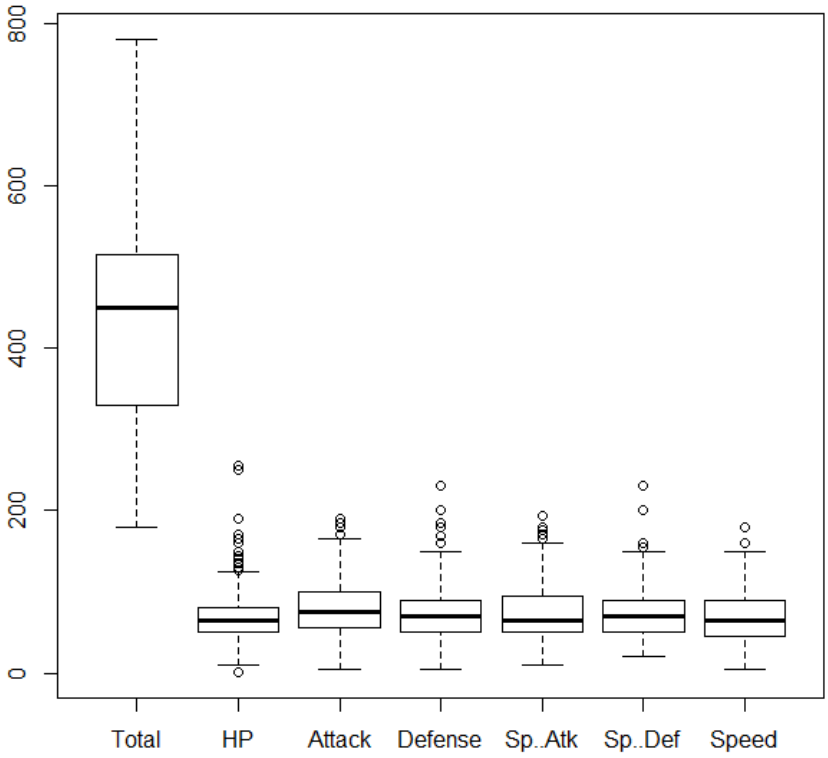
Analiza graficzna

Do przedstawienia danych w sposób graficzny posłużymy się dwuwymiarowymi wykresami oraz wykresami pudełkowymi. Z racji na swoją naturę graficznie przedstawione zostaną tylko atrybuty **Total, HP, Attack, Defense, Sp. Atk,Sp. Def, Speed, Generation**. Jako zmienna decyzyjna, jednocześnie koloryzująca punkty na wykresie, wykorzystana jest zmienna **Type.1**

Rysunek 1: Wykresy dwuwymiarowe



Rysunek 2: Wykresy pudełkowe



Z wykresów wyciągamy następujące wnioski.

1. Dane są generalnie równo rozłożone z paroma wartościami oddalonymi. Te nietypowe osobniki są to prawdopodobnie wyspecjalizowane lub ponadprzeciętnie silne pokemony.
2. Widzimy, że niektóre atrybuty są ze sobą silnie dodatnio skorelowane np. **Sp..Def i Total** a niektóre praktycznie w ogóle np. **HP i Attack**.
3. Wszystkie atrybuty przyjmują wartości z tego samego przedziału, wyjątkiem jest **Total** który będąc sumą pozostałych atrybutów musi się znacząco wywyższać. Z powodu tego jednego atrybutu będziemy musieli znormalizować dane. Inaczej wartości atrybutu **Total** dominowały by nad innymi atrybutami.
4. Nie widać żadnych szczególnych zgrupowań które pozwoliły by podzielić pokemony na jakieś większe podgrupy.

Analiza ilościowa

Kolejnym etapem jest próba skwantyfikowania obserwacji z poprzednich etapów.

Listing 2: Podsumowanie danych

```
> summary(pokemon)
       X.                                Name         Type.1         Type.2
 Min.   :  1.0   Abomasnow                :  1   Water  :112           :386
 1st Qu.:184.8   AbomasnowMega Abomasnow:  1   Normal : 98   Flying  : 97
 Median :364.5   Abra                     :  1   Grass  : 70   Ground  : 35
 Mean   :362.8   Absol                    :  1   Bug    : 69   Poison  : 34
 3rd Qu.:539.2   AbsolMega Absol          :  1   Psychic: 57   Psychic : 33
 Max.   :721.0   Accelgor                 :  1   Fire   : 52   Fighting: 26
                 (Other)                  :794   (Other):342   (Other) :189
     Total             HP              Attack       Defense          Sp..Atk
 Min.   :180.0   Min.   :  1.00   Min.   :  5   Min.   :  5.00   Min.   : 10.00
 1st Qu.:330.0   1st Qu.: 50.00   1st Qu.: 55   1st Qu.: 50.00   1st Qu.: 49.75
 Median :450.0   Median : 65.00   Median : 75   Median : 70.00   Median : 65.00
 Mean   :435.1   Mean   : 69.26   Mean   : 79   Mean   : 73.84   Mean   : 72.82
 3rd Qu.:515.0   3rd Qu.: 80.00   3rd Qu.:100   3rd Qu.: 90.00   3rd Qu.: 95.00
 Max.   :780.0   Max.   :255.00   Max.   :190   Max.   :230.00   Max.   :194.00

    Sp..Def          Speed          Generation    Legendary
 Min.   : 20.0   Min.   :  5.00   Min.   :1.000   False:735
 1st Qu.: 50.0   1st Qu.: 45.00   1st Qu.:2.000   True : 65
 Median : 70.0   Median : 65.00   Median :3.000
 Mean   : 71.9   Mean   : 68.28   Mean   :3.324
 3rd Qu.: 90.0   3rd Qu.: 90.00   3rd Qu.:5.000
 Max.   :230.0   Max.   :180.00   Max.   :6.000
```

Listing 3: Korelacje

```
> cor(pokemon[5:12])
                Total         HP     Attack    Defense    Sp..Atk    Sp..Def
Total      1.00000000 0.61874835 0.73621065 0.61278743 0.74724986 0.71760947
HP         0.61874835 1.00000000 0.42238603 0.23962232 0.36237986 0.37871807
Attack     0.73621065 0.42238603 1.00000000 0.43868706 0.39636176 0.26398955
Defense    0.61278743 0.23962232 0.43868706 1.00000000 0.22354861 0.51074659
Sp..Atk    0.74724986 0.36237986 0.39636176 0.22354861 1.00000000 0.50612142
Sp..Def    0.71760947 0.37871807 0.26398955 0.51074659 0.50612142 1.00000000
Speed      0.57594266 0.17595206 0.38123974 0.01522660 0.47301788 0.25913311
Generation 0.04838402 0.05868251 0.05145134 0.04241857 0.03643683 0.02848599
                Speed  Generation
Total       0.57594266  0.04838402
HP          0.17595206  0.05868251
Attack      0.38123974  0.05145134
Defense     0.01522660  0.04241857
Sp..Atk     0.47301788  0.03643683
Sp..Def     0.25913311  0.02848599
Speed       1.00000000 -0.02312106
Generation -0.02312106  1.00000000
```

Widzimy potwierdzenie naszych obserwacji w postaci numerycznej

- Atrybuty liczbowe przyjmują wartości które widzieliśmy na wykresie pudełkowym
- Istnieje jedna kolumna z pustymi polami i jest to **Type.2**. Zauważamy jednak, że jest to dla nas nadal informacja i dla tego nie musimy tych pustych pól wypełniać.
- W zbiorze danych istnieje 6 generacji pokemonów. Jest to ważna informacja która w sama sobie pozwala nam na pogrupowanie wszystkich pokemonów na 6 grup.
- Pokemonów legendarnych jest znacznie mniej niż pokemonów zwyczajnych
- Nasze obserwacje na temat korelacji atrybutów zostają potwierdzone przez liczby. Warta głębszej uwagi jest zmienna **Total** będąca sumą pozostałych atrybutów liczbowych niespecjalnych. Pomimo silnej korelacji z jej składowymi atrybut daje nam nową informację i pozwala na lepszą klasyfikację.

Przygotowanie danych

Obróbka danych

Po przebadaniu danych możemy przejść do procesu obróbki.

Listing 4: Obrabianie danych

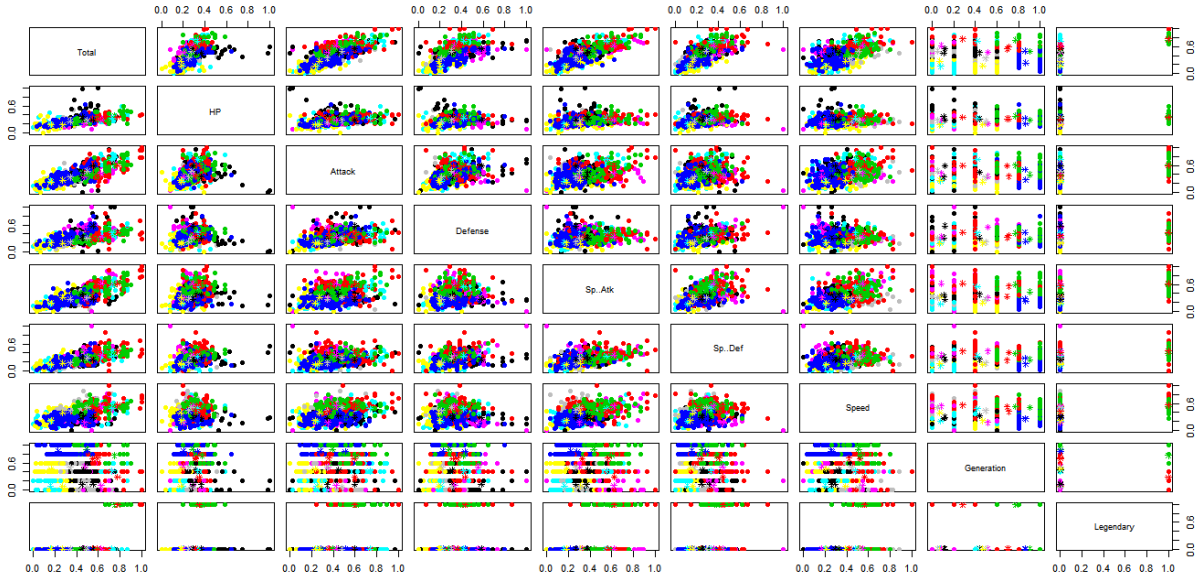
```
obrob <- function(m){

  # usuwamy kolumne X i Name
  m <- m[c(-1,-2)]

  # zamieniamy Legendary na boolean
  m$Legendary <- as.integer(as.logical(m$Legendary))

  # stosujemy normalizacje min-max na zbiorze
  m <- normalizuj(m, typ="norm", atryb=c(3,4,5,6,7,8,9,10))

  m
}
```

Nie istnieją braki w danych.

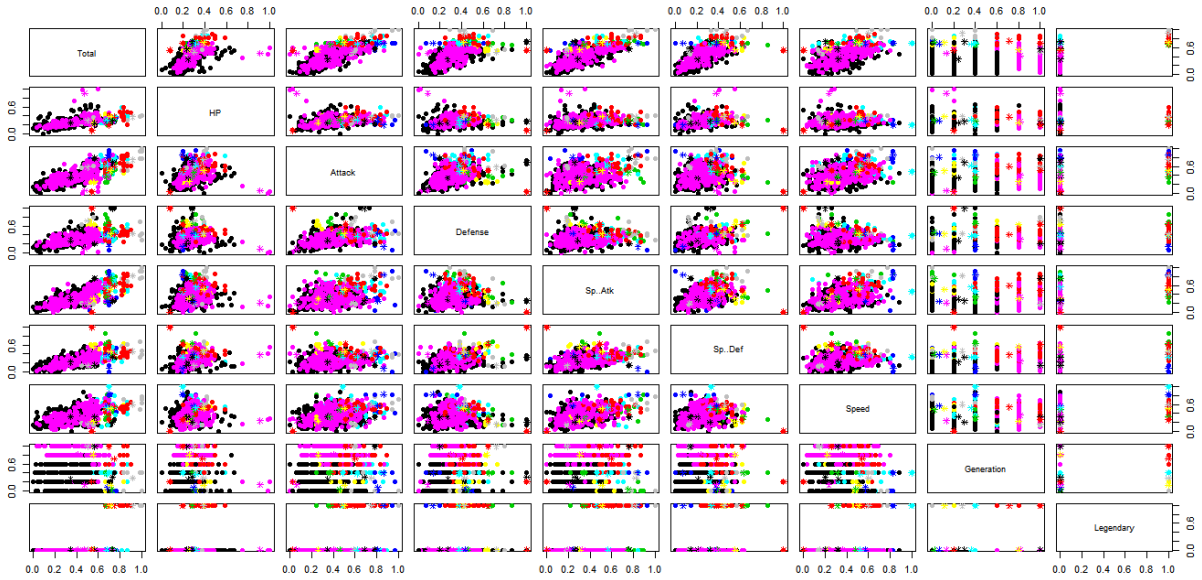
Grupowanie danych

Jesteśmy gotowi sprawdzić z jaką łatwością dane się grupują. Zaczynamy od próby zgrupowania danych względem atrybutu decyzyjnego, czyli **Type.1**

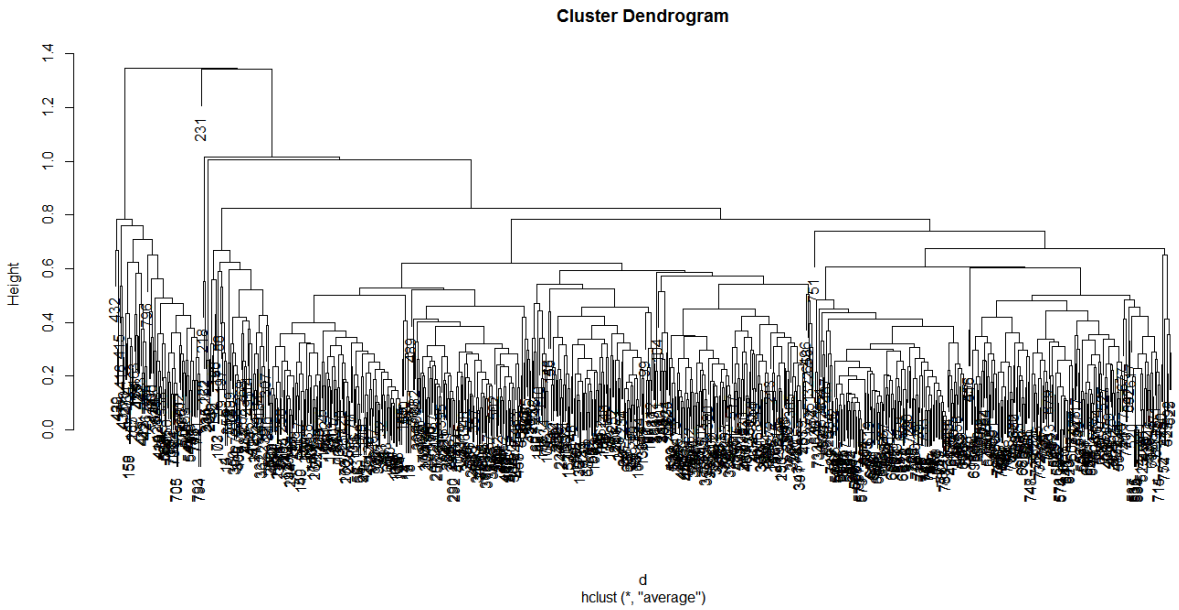
Rysunek 3: Grupowanie metodą kśrednich



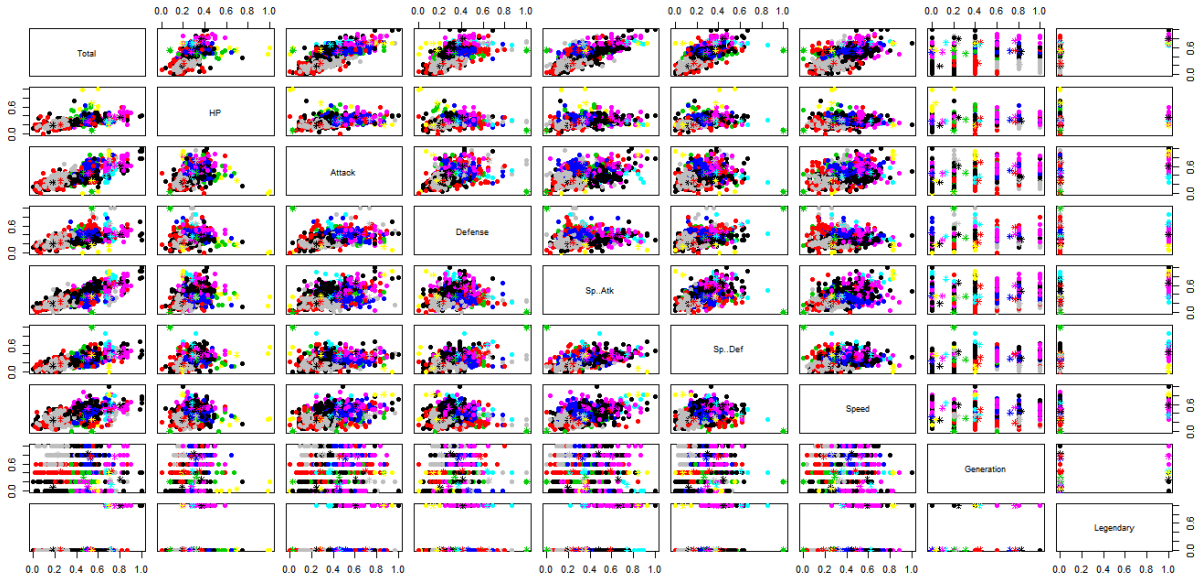
Rysunek 4: Grupowanie metodą hierarchiczną średniego połączenia



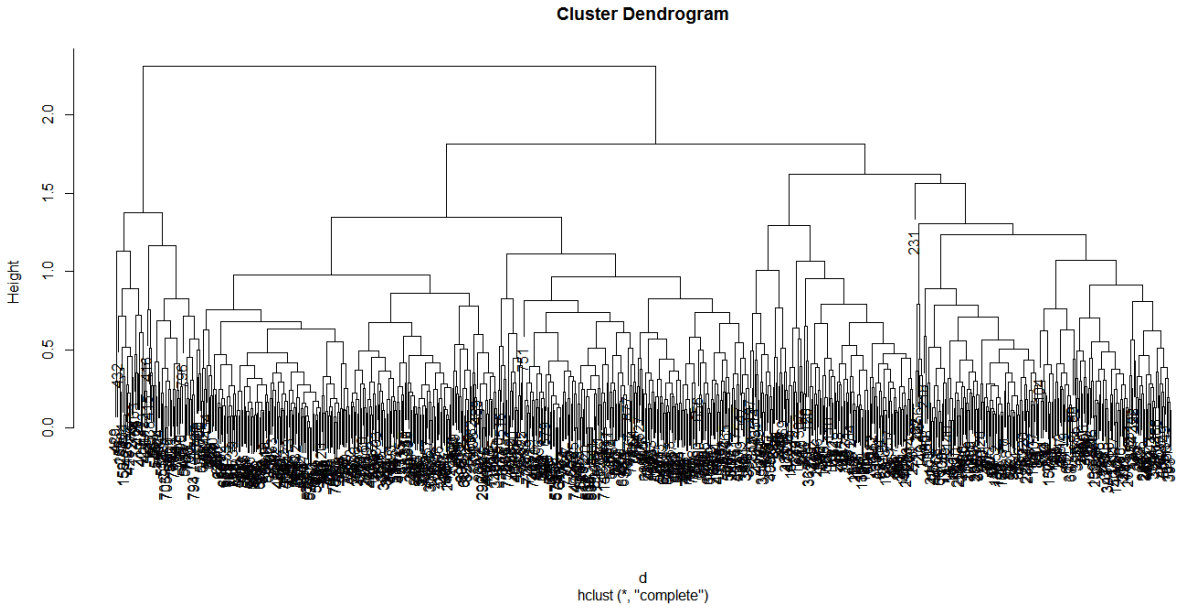
Rysunek 5: Grupowanie metodą hierarchiczną średniego połączenia - dendrogram



Rysunek 6: Grupowanie metodą hierarchiczną całkowitego połączenia



Rysunek 7: Grupowanie metodą hierarchiczną całkowitego połączenia - dendrogram



Nasuwają się następujące wnioski

- Zróżnicowanie danych jest na tyle skomplikowane, że grupowanie nie wystarczy do poprawnego zaklasyfikowania danych. Jak widzieliśmy na Rysunku 1. dane nie są pogrupowane w tak równomierny sposób jaki sugerują powyższe wykresy.
- Ważną informację przekazują nam dendrogramy. Widzimy na nich, że jeśli bierzemy pod uwagę wszystkie parametry to wyodrębnić można jedną małą grupę i jedną dużą. Jest to odzwierciedlone na wykresach, gdzie widzimy zawsze jedno duże skupisko punktów oraz jedno mniejsze poboczne.