Raport 2
Dyskretyzacja, Analiza składowych głównych, Skalowanie wielowymiarowe

Romana Żmuda
249706
Adrian Kit
249746

8 listopada 2020

Spis treści

1 Wst<ea>p- wprowadzenie do podanych zagadnień

W tym sprawozdaniu mamy do zbadania 3 zbiory danych. Pierwszy z nich dotyczy kwiatków irysów, drugi różnorakich informacji o stanach w USA. Trzeci zbiór, wybrany przez nas dotyczy ..... Spróbujemy odpowiedzieć na pytania: -Czy obecność wartości odstających wpływa na efektywność metod dyskretyzacji? -Czy do pełnej analizy zawsze potrzebujemy wszystkich danych?

W tym celu posłużymy się m.in. dyskretyzacją danych, analizą składowych głównych oraz metodą skalowania wielowymiarowego.

1.1 Krótki opis działania podanych metod

- *Dyskretyzacja*
  Dyskretyzacjajest procesem w którym wartości dla atrybutów ciągłych są za-stępowane wartością dyskretną, odpowiadającą pewnemu przedziałowi ciągłychwartości oryginalnego atrybutu. Przedziały te są uporządkowane, co sprawia,że w wyniku dyskretyzacji otrzymujemy zamiast atrybutu ciągłego atrybutporządkowy o skończonej liczbie wartości.

- *Analiza składowych głównych (PCA)*
  Analiza składowych głównych (PCA) - służy m.in. do redukcji liczby zmiennych opisujących zjawiska, czy do odkrycia prawidłowości między zmiennymi. Polega ona na wyznaczeniu składowych będących kombinacją liniową badanych zmiennych. Dokładna analiza składowych głównych umożliwia wskazanie tych zmiennych początkowych, które mają duży wpływ na wygląd poszczególnych składowych głównych czyli tych, które tworzą grupę jednorodną. Składowa główna (u której wariancja jest zmaksymalizowana) jest wówczas reprezentantem tej grupy.

- *Skalowanie wielowymiarowe (MDS)*
  Skalowanie wielowymiarowe (MDS) może być rozważane jako alternatywa analizy czynnikowej. Ogólnie, celem tej analizy jest wykrycie sensownych ukrytych wymiarów, które pozwalają badaczowi wyjaśnić obserwowane podobieństwa lub odmienności (odległości) między badanymi obiektami. W analizie czynnikowej podobieństwa między obiektami (np. zmiennymi) są wyrażone w postaci macierzy korelacji. Przy pomocy MDS, oprócz macierzy korelacji, można analizować dowolny rodzaj macierzy podobieństwa lub odmienności.

2 Zadanie 1 - Dyskretyzacja (przedziałowanie) cech ciągłych na zbiorze o irysach

Podstawowe informacje o podanym pliku danych.

```
> library("datasets")
> data("iris")
> attach(iris)
> head(iris)

  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
```

```
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa

> ncol(iris) # ilośc kolumn

[1] 5

> nrow(iris) #ilość przypadków

[1] 150

> sapply(iris, class) # identyfikacja cech

Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species
   "numeric"    "numeric"    "numeric"    "numeric"     "factor"

> ilebrakujacych<-sum(is.na(iris))
> ilebrakujacych # liczba brakujących danych

[1] 0

> summary(iris)

  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50

>
```

2.1 Ocena zdolności dyskryminacyjnej

Histogramy poszczególnych cech ciągłych

Z histogramów możemy odczytać następujące zależności:

- Zmienna Sepal.length ma równomiernie rozmieszczone wartości, przez co ciężko będzie znaleźć jakiekolwiek zróżnicowanie bazując tylko na niej.
- Zmienna Sepal.Width przypomina rozkład normalny. Przez symetrię, zmienna ta również nie nadaje się do separacji gatunkóW.
- Zmienne Petal.Length oraz Petal.Width charakteryzują się dużym rozstrzałem wartości, widoczne są także wyraźne wartości odstające. Dzięki temu mogą dobrze różnicować poszczególne gatunki irysów.
- Warto też zauważyć, że w obu tych zmiennych jest luka w wartościach, co świadczy o zróżnicowaniu wartości.

W następnym kroku naszej analizy sprawdzimy, czy różnice w wartościach są powiązane z gatunkami naszych irysów. Mówiąc wprost, odpowiemy na pytanie: Czy badając poszczególne gatunki irysów możemy otrzymać znacząco zróżnicowane wartości zmiennych? Powtórzymy analizę zmiennych (ograniczając się jednak jedynie do zmiennych Petal.Length oraz Petal.Width) z podziałem na gatunki irysów.

Zmienna Petal Length

Wnioski:

- Wyraźne zróżnicowanie zakresów zmiennych
- Wartości mocno uśrednione, znikoma ilość wartości odstających (sztuczne wydłużenie zakresu).

Zmienna Petal Width

Wnioski:

- Gatunek setosa ma w zdecydowanej większości krótkie płatki długości maksymalnie 0,2. Są zdecydowanie krótsze od płatków innych gatunków (czy wypisujemy dokładnie?)
- Skrajne wartości z prawej strony są przyjmowane w bardzo małych ilościach, co sztucznie wydłuża zakres.
- Gatunki Versicolor i Virginica mają uśrednione wartości

Ogólne podsumowanie:

- Poszczególne gatunki irysów mają zdecydowanie zróżnicowane wartości zmiennych.
- Irysy gatunku setosa są zdecydowanie mniejsze od pozostałych, po środku uplasował się gatunek Versicolor, największe są irysy gatunku virginica
- Różnica między setosą a versicolorem jest zdecydowanie większa niż między versicolorem a virginicą.

2.2 Porównanie nienadzorowanych metod dyskretyzacji

2.2.1 Informacje wstępne do dyskretyzacji

Po analizie z podsekcji *Ocena zdolności dyskryminacyjnej* wybraliśmy 2 cechy: Petal.Length, Sepal.Width, które poddamy badaniu dyskretyzacji na najlepszą metodę nienadzorowaną, bedą to:

- Dyskretyzacja według równej częstości - taka sama liczba obiektów w przedziale
- Dyskretyzacja według równej szerokości - określona liczba przedziałów
- Dyskretyzacja oparta na algorytmie k-średnich - k różnych możliwie odmiennych skupień, u nas k = 3 odmiany irysów
- Metoda oparta na wizualizacji - ręczny podział ze względu na obserwacje

Krótka obróbka danych:

```
> # Wczytujemy pakiet pozwalający na konwersję zmiennej ciągłej (numeric)
> # na zmienną jakościową (factor)
> library(arules)
> x_len <- iris[,"Petal.Length"] # lista zmiennych Petal.Lenght
> x_wid <- iris[,"Sepal.Width"] # lista zmienych Sepal.Width
> y_len <- runif(length(x_len))
> # losowo wybierane wartości na osi Y dla zmiennych Petal.Lenght
> y_wid<-runif(length(x_wid))
> # losowo wybierane wartości na osi Y dla zmiennych Sepal.Width
>
```

Badanie zdolności dyskryminacyjnych :

```
> table(Species)

Species
    setosa versicolor  virginica
        50         50         50

> #zbadanie ilości klas, czyli u nas ilość gatunków - zmienna jakościowa
```

Naszą zmienną separacyjną będzie ilość klas gatunków. Mamy 3 różne gatunki, a ich zmienne noszą nazwy:

- setosa
- versicolor
- virginica

Sprawdzamy jej zdolności, czy dobrze separuje klasy za pomocą boxplotów: Zmienna Petal.Lenght, widzimy, iż taki separator doskonale ukazuje róznice dla obydwóch zmiennych.Zmienne w klasach Versicolor i Virginica są znacznie bliżej siebie, niż wartości Setosy i Versicolor, jak róWnież wartości z Ver i Virg skrajne mogą na wykresie zostać sklasyfikowane do nie swojego gatunku. Zmienna Sepal.Width ma zbiory, których wartości z odpowiednich klas nakładają się na siebie, możliwe że poszczególne elementy nie będą należeć do właściwych gatunków.
Teraz przyjrzymy się badanym zmiennym (bez podziału na klasy) w celu szybkiego zrozumienia i dokonaniu pierwszych spotrzeżeń w ich położeniach. Krótka obserwacja dla Petal.Length: znaczący podział na dwie grupy, co może potwierdzać wcześniejsze wnioski, iż setosa odstaje od pozostałych dwóch grup. Krótka obserwacja dla Sepal.Width: zbiór skoncentrowany w punkcie 3.

2.2.2 Dyskretyzacja według równej częstości

- Zmienna Petal.Length

```
[1] "Podzia■ly metodą równych części: "

x_len_cz
  [1,2.63) [2.63,4.9)  [4.9,6.9]
        50         49         51
```

Porównanie z podziałem na rzeczywiste klasy gatunków: Ostatni etap to wyznaczenie macierzy kontyngencji i sprawdzenie, w jakim stopniu obiekty należące do poszczególnych klas są przypisane do tej samej kategorii. Najpierw zobaczmy czy odpowiednie gatunki znalazły się w swoich podziałach:

```
[1] "Podzia■ly metodą równych części z uwzgędnieniem klas : "

            Species
x_len_cz     setosa versicolor virginica
  [1,2.63)       50          0         0
  [2.63,4.9)      0         46         3
  [4.9,6.9]       0          4        47
```

Oraz wyznaczony współczynnik zgodności:

```
Cases in matched pairs: 95.33 %
    [1,2.63)   [2.63,4.9)    [4.9,6.9]
    "setosa" "versicolor"  "virginica"
```

- Zmienna Sepal.Width

```
[1] "Podzia■ly metodą równych części: "

x_wid_cz
  [2,2.9) [2.9,3.2) [3.2,4.4]
       47        47        56
```

Porównanie z podziałem na rzeczywiste klasy gatunków: Ostatni etap to wyznaczenie macierzy kontyngencji i sprawdzenie, w jakim stopniu obiekty należące do poszczególnych klas są przypisane do tej samej kategorii. Najpierw zobaczmy czy odpowiednie gatunki znalazły się w swoich podziałach:

```
[1] "Podzia■ly metodą równych części z uwzgędnieniem klas : "
```

```
           Species
x_wid_cz    setosa versicolor virginica
  [2,2.9)        1         27        19
  [2.9,3.2)     11         18        18
  [3.2,4.4]     38          5        13
```

Oraz wyznaczony współczynnik zgodności:

```
Cases in matched pairs: 55.33 %
     [2,2.9)    [2.9,3.2)    [3.2,4.4]
"versicolor" "versicolor"     "setosa"
```

2.2.3 Dyskretyzacja według równej szerokości

- Zmienna Petal.Length

```
[1] "Podzia■ly metodą równej szerokości: "
```

```
x_len_sz
   [1,2.97) [2.97,4.93)  [4.93,6.9]
         50          54          46
```

Porównanie z podziałem na rzeczywiste klasy gatunków: Ostatni etap to wyznaczenie macierzy kontyngencji i sprawdzenie, w jakim stopniu obiekty należące do poszczególnych klas są przypisane do tej samej kategorii. Najpierw zobaczmy czy odpowiednie gatunki znalazły się w swoich podziałach:

```
[1] "Podzia■ly metodą równej szerokośći z uwzgędnieniem klas : "
```

```
             Species
x_len_sz      setosa versicolor virginica
  [1,2.97)        50          0         0
  [2.97,4.93)      0         48         6
  [4.93,6.9]       0          2        44
```

Oraz wyznaczony współczynnik zgodności:

```
Cases in matched pairs: 94.67 %
    [1,2.97)  [2.97,4.93)   [4.93,6.9]
    "setosa" "versicolor"  "virginica"
```

- Zmienna Sepal.Width

```
[1] "Podzia■ly metodą równej szerokości: "
```

```
x_wid_sz
  [2,2.8) [2.8,3.6) [3.6,4.4]
       47        88        15
```

Porównanie z podziałem na rzeczywiste klasy gatunków: Ostatni etap to wyznaczenie macierzy kontyngencji i sprawdzenie, w jakim stopniu obiekty należące do poszczególnych klas są przypisane do tej samej kategorii. Najpierw zobaczmy czy odpowiednie gatunki znalazły się w swoich podziałach:

```
[1] "Podzia■ly metodą równej szerokości z uwzgędnieniem klas : "

           Species
x_wid_sz    setosa versicolor virginica
  [2,2.8)        1         27        19
  [2.8,3.6)     36         23        29
  [3.6,4.4]     13          0         2
```

Oraz wyznaczony współczynnik zgodności:

```
Cases in matched pairs: 50.67 %
     [2,2.8)    [2.8,3.6)    [3.6,4.4]
"versicolor"     "setosa"     "setosa"
```

2.2.4 Dyskretyzacja oparta na algorytmie k-średnich

- Zmienna Petal.Length

```
[1] "Podzia■ly metodą opartą na algorytmie grupowania: "

x_len_k
   [1,2.85) [2.85,4.89)  [4.89,6.9]
         50          49          51
```

Porównanie z podziałem na rzeczywiste klasy gatunków: Ostatni etap to wyznaczenie macierzy kontyngencji i sprawdzenie, w jakim stopniu obiekty należące do poszczególnych klas są przypisane do tej samej kategorii. Najpierw zobaczmy czy odpowiednie gatunki znalazły się w swoich podziałach:

```
[1] "Podzia■ly metodą opartą na algorytmie grupowania z uwzgędnieniem klas : "

             Species
x_len_k       setosa versicolor virginica
  [1,2.85)        50          0         0
  [2.85,4.89)      0         46         3
  [4.89,6.9]       0          4        47
```

Oraz wyznaczony współczynnik zgodności:

```
Cases in matched pairs: 95.33 %
    [1,2.85)  [2.85,4.89)   [4.89,6.9]
    "setosa" "versicolor"  "virginica"
```

- Zmienna Sepal.Width

```
[1] "Podzia■ly metodą opartą na algorytmie grupowania: "

x_wid_k
   [2,2.69) [2.69,3.28)  [3.28,4.4]
         24          83          43
```

Porównanie z podziałem na rzeczywiste klasy gatunków: Ostatni etap to wyznaczenie macierzy kontyngencji i sprawdzenie, w jakim stopniu obiekty należące do poszczególnych klas są przypisane do tej samej kategorii. Najpierw zobaczmy czy odpowiednie gatunki znalazły się w swoich podziałach:

```
[1] "Podzia■ly metodą opartą na algorytmie grupowania z uwzgędnieniem klas : "

             Species
x_wid_k       setosa versicolor virginica
  [2,2.69)         1         16         7
  [2.69,3.28)     16         32        35
  [3.28,4.4]      33          2         8
```

Oraz wyznaczony współczynnik zgodności:

```
Cases in matched pairs: 56 %
    [2,2.69)  [2.69,3.28)   [3.28,4.4]
"versicolor"  "virginica"     "setosa"
```

2.2.5 Dyskretyzacja oparta na przedziałach zadanych przez użytkownika

- Zmienna Petal.Length

```
[1] "Podzia■ly metodą opartą na podzi■lach podanych przez u■ytkownika: "

x_len_u
 small medium  large
    50     54     46
```

Porównanie z podziałem na rzeczywiste klasy gatunków: Ostatni etap to wyznaczenie macierzy kontyngencji i sprawdzenie, w jakim stopniu obiekty należące do poszczególnych klas są przypisane do tej samej kategorii. Najpierw zobaczmy czy odpowiednie gatunki znalazły się w swoich podziałach:

```
[1] "Podzia■ly podane przez u■ytkownika z uwzgędnieniem klas : "

        Species
x_len_u  setosa versicolor virginica
  small      50          0         0
  medium      0         48         6
  large       0          2        44
```

Oraz wyznaczony współczynnik zgodności:

```
Cases in matched pairs: 94.67 %
       small       medium        large
    "setosa" "versicolor"  "virginica"
```

- Zmienna Sepal.Width

```
[1] "Podzia■ly metodą opartą na podzi■lach podanych przez u■ytkownika: "

x_wid_u
 small medium  large
    24     89     37
```

Porównanie z podziałem na rzeczywiste klasy gatunków: Ostatni etap to wyznaczenie macierzy kontyngencji i sprawdzenie, w jakim stopniu obiekty należące do poszczególnych klas są przypisane do tej samej kategorii. Najpierw zobaczmy czy odpowiednie gatunki znalazły się w swoich podziałach:

```
[1] "Podzia■ly podane przez u■ytkownika z uwzgędnieniem klas : "

        Species
x_wid_u  setosa versicolor virginica
  small       1         16         7
  medium     18         33        38
  large      31          1         5
```

Oraz wyznaczony współczynnik zgodności:

```
Cases in matched pairs: 56.67 %
       small       medium        large
"versicolor"  "virginica"     "setosa"
```

2.3 Wpływ obserwacji odstających przy badaniu metod dyskretyzacji

2.3.1 Dyskretyzacja według równej częstości

Nowa zmienna Petal.Length

```
[1] "Podzia■ly metodą równych części: "

x_len_cz
 [-6,2.63) [2.63,4.9) [4.9,13.9]
        50         49         51
```

Porównanie z podziałem na rzeczywiste klasy gatunków: Ostatni etap to wyznaczenie macierzy kontyngencji i sprawdzenie, w jakim stopniu obiekty należące do poszczególnych klas są przypisane do tej samej kategorii. Najpierw zobaczmy czy odpowiednie gatunki znalazły się w swoich podziałach:

```
[1] "Podzia■ly metodą równych części z uwzgędnieniem klas : "
```

```
            Species
x_len_cz     setosa versicolor virginica
  [-6,2.63)      50          0         0
  [2.63,4.9)      0         46         3
  [4.9,13.9]      0          4        47
```

Oraz wyznaczony współczynnik zgodności:

```
Cases in matched pairs: 95.33 %
   [-6,2.63)   [2.63,4.9)   [4.9,13.9]
    "setosa" "versicolor"  "virginica"
```

2.3.2 Dyskretyzacja według równej szerokości

Nowa zmienna Petal.Length

```
[1] "Podzia■ly metodą równej szerokości: "

x_len_sz
  [-6,0.633) [0.633,7.27)  [7.27,13.9]
           1          148            1
```

Porównanie z podziałem na rzeczywiste klasy gatunków: Ostatni etap to wyznaczenie macierzy kontyngencji i sprawdzenie, w jakim stopniu obiekty należące do poszczególnych klas są przypisane do tej samej kategorii. Najpierw zobaczmy czy odpowiednie gatunki znalazły się w swoich podziałach:

```
[1] "Podzia■ly metodą równej szerokośći z uwzgędnieniem klas : "

              Species
x_len_sz       setosa versicolor virginica
  [-6,0.633)        1          0         0
  [0.633,7.27)     49         50        49
  [7.27,13.9]       0          0         1
```

Oraz wyznaczony współczynnik zgodności:

```
Cases in matched pairs: 34.67 %
  [-6,0.633) [0.633,7.27)  [7.27,13.9]
    "setosa" "versicolor"  "virginica"
```

2.3.3 Dyskretyzacja oparta na algorytmie k-średnich

Nowa zmienna Petal.Length

```
[1] "Podzia■ly metodą opartą na algorytmie grupowania: "

x_len_k
  [-6,-2.25) [-2.25,3.25)  [3.25,13.9]
           1           50           99
```

Porównanie z podziałem na rzeczywiste klasy gatunków: Ostatni etap to wyznaczenie macierzy kontyngencji i sprawdzenie, w jakim stopniu obiekty należące do poszczególnych klas są przypisane do tej samej kategorii. Najpierw zobaczmy czy odpowiednie gatunki znalazły się w swoich podziałach:

```
[1] "Podzia■ly metodą opartą na algorytmie grupowania z uwzgędnieniem klas : "
```

```
              Species
x_len_k        setosa versicolor virginica
  [-6,-2.25)        1          0         0
  [-2.25,3.25)     49          1         0
  [3.25,13.9]       0         49        50
```

Oraz wyznaczony współczynnik zgodności:

```
Cases in matched pairs: 66.67 %
  [-6,-2.25) [-2.25,3.25)  [3.25,13.9]
    "setosa"     "setosa"  "virginica"
```

2.3.4 Dyskretyzacja oparta na przedziałach zadanych przez użytkownika

Nowa zmienna Petal.Length

```
[1] "Podzia■ly metodą opartą na podzi■lach podanych przez u■ytkownika: "
```

```
x_len_u
 small medium  large
    50     54     46
```

Porównanie z podziałem na rzeczywiste klasy gatunków: Ostatni etap to wyznaczenie macierzy kontyngencji i sprawdzenie, w jakim stopniu obiekty należące do poszczególnych klas są przypisane do tej samej kategorii. Najpierw zobaczmy czy odpowiednie gatunki znalazły się w swoich podziałach:

```
[1] "Podzia■ly podane przez u■ytkownika z uwzgędnieniem klas : "
```

```
        Species
x_len_u  setosa versicolor virginica
  small      50          0         0
  medium      0         48         6
  large       0          2        44
```

Oraz wyznaczony współczynnik zgodności:

```
Cases in matched pairs: 94.67 %
       small       medium        large
    "setosa" "versicolor"  "virginica"
```

Wnioski: .....

3 Zadanie 2. Analiza składowych głównych PCA

3.1 Wczytanie danych

Poprawne wczytanie danych

```
> data(state); dane <- as.data.frame(state.x77)
```

Podstawowe informacje o danych

```
> head(dane)
```

```
           Population Income Illiteracy Life Exp Murder HS Grad Frost   Area
Alabama          3615   3624        2.1    69.05   15.1    41.3    20  50708
Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
Arizona          2212   4530        1.8    70.55    7.8    58.1    15 113417
Arkansas         2110   3378        1.9    70.66   10.1    39.9    65  51945
California      21198   5114        1.1    71.71   10.3    62.6    20 156361
Colorado         2541   4884        0.7    72.06    6.8    63.9   166 103766
```

```
> ncol(dane) # ilośc kolumn
```

```
[1] 8
```

```
> nrow(dane) #ilość przypadków
```

```
[1] 50
```

```
> sapply(dane, class) # identyfikacja cech
```

```
Population     Income Illiteracy   Life Exp     Murder    HS Grad      Frost
 "numeric"  "numeric"  "numeric"  "numeric"  "numeric"  "numeric"  "numeric"
      Area
 "numeric"
```

```
> ilebrakujacych<-sum(is.na(dane))
> ilebrakujacych # liczba brakujących danych
```

```
[1] 0
```

W tym zadaniu przeprowadzimy analizę danych składowych PCA na zbiorze danych state.x77. Zaczniemy od sprawdzenia zmienności cech. Jeśli wariancje poszczególnych zmiennych będą zbyt zróżnicowane, wtedy konieczna będzie standaryzacja.

3.2 Przygotowanie danych

Badanie zmienności cech (wariancja):

```
  Population   Income Illiteracy Life Exp   Murder  HS Grad    Frost       Area
1   19931684 377573.3  0.3715306  1.80202 13.62747 65.23789 2702.009 7280748061
```

Obserwacje: Wariancje wyjściowych zmiennych są bardzo zróżnicowane. Celem uniknięcia dominacji zmiennej o bardzo dużej wariancji (zmienna Area) przeprowadzimy standaryzację danych przed zastosowaniem metody PCA. Wnioski: Przed standaryzacją danych największą wariancję miała zmienna Area, co mogło sugerować, że to ona będzie miała największy wpływ na nasze zmienne. Jednak po normalizacji zmienna ta okazuje się być jedną z najmniej istotnych, co potwierdza, że standaryzacja była absolutnie konieczna

3.3 Wyznaczenie składowych głównych oraz zbadanie zmienności

Wyznaczenie składowych głównych oraz wyświetlenie macierzy która wskazuje, jak każda ze zmiennych w wyściowym zbiorze wpływa na poszczególne główne składowe:

```
[1] "Nowe cechy:"

                  PC1         PC2         PC3         PC4          PC5
Population  0.12642809  0.41087417 -0.65632546 -0.40938555  0.405946365
Income     -0.29882991  0.51897884 -0.10035919 -0.08844658 -0.637586953
Illiteracy  0.46766917  0.05296872  0.07089849  0.35282802  0.003525994
Life.Exp   -0.41161037 -0.08165611 -0.35993297  0.44256334  0.326599685
Murder      0.44425672  0.30694934  0.10846751 -0.16560017 -0.128068739
HS.Grad    -0.42468442  0.29876662  0.04970850  0.23157412 -0.099264551
Frost      -0.35741244 -0.15358409  0.38711447 -0.61865119  0.217363791
Area       -0.03338461  0.58762446  0.51038499  0.20112550  0.498506338
                  PC6          PC7         PC8
Population -0.01065617 -0.062158658 -0.21924645
Income      0.46177023  0.009104712  0.06029200
Illiteracy  0.38741578 -0.619800310 -0.33868838
Life.Exp    0.21908161 -0.256213054  0.52743331
Murder     -0.32519611 -0.295043151  0.67825134
HS.Grad    -0.64464647 -0.393019181 -0.30724183
Frost       0.21268413 -0.472013140  0.02834442
Area        0.14836054  0.286260213  0.01320320
```

Zbadamy wektory ładunków, w celu odkycia, które z PC1,PC2,PC3, ... mają największy wkład redukcji danych i przenoszonej informacji o cechach: Wartości dla poszczególnych PCA:

```
> summary(dane_pca)

Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     1.8971 1.2775 1.0545 0.84113 0.62019 0.55449 0.38006
Proportion of Variance 0.4499 0.2040 0.1390 0.08844 0.04808 0.03843 0.01806
Cumulative Proportion  0.4499 0.6539 0.7928 0.88128 0.92936 0.96780 0.98585
                           PC8
Standard deviation     0.33643
Proportion of Variance 0.01415
Cumulative Proportion  1.00000

>
```

W interpretacji wyników najbardziej istotna jest część objaśnianej zmienności (Proportion of Variance) widoczna po zastosowaniu funkcji summary. Na jej podstawie możemy stwierdzić, że pierwsza składowa główna wyjaśnia 44,99% zmienności, a druga - 20,40%.
W celu lepszej analizy zastosujemy boxplot rozkładu wariancji oraz barplot procentowego udziału odpowiednich wartości PCA:

Kolejne słupki oznaczają zsumowane wartości PCA, zaczynając od PC1 kończąc na ich sumie: Widzimy, że zmienna PC1+PC2+PC3 stanowi około 80% informacji o danych, natomiast PC1+PC2+PC3+PC4+PC5 stanowi 90 %.

4 Zadanie 3 - Skalowanie wielowymiarowe (MDS)

Celem tej metody jest wyznaczenie współrzędnych w nowym układzie współrzędnych, w taki sposób by odległości pomiędzy obiektami w nowym układzie współrzędnych były podobne do oryginalnych odległości pomiędzy obiektami, w tej metodzie szczególną uwagę przykładamy do zmiennych jakościowych. W naszym raporcie użyjemy metody skalowania metrycznego przy pomocy funkcji *cmdscale* . Jest to metoda ekstrakcji cech, na podstawie macierzy odległości lub macierzy niepodobieństwa pomiędzy obiektami.

4.1 Wprowadzenie danych

Wybraliśmy dane z pakietu *lattice* o nazwie "US Regional Mortality", poniżej krótka charakteryzacja pliku:

```
> library(lattice)
> mortal<-as.data.frame(USRegionalMortality)
> attach(mortal)
> head(mortal)

          Region Status    Sex         Cause  Rate  SE
5  HHS Region 01  Urban   Male Heart disease 188.2 1.0
6  HHS Region 01  Rural   Male Heart disease 199.1 2.6
7  HHS Region 01  Urban Female Heart disease 115.1 0.6
8  HHS Region 01  Rural Female Heart disease 124.5 1.7
9  HHS Region 02  Urban   Male Heart disease 226.8 0.8
10 HHS Region 02  Rural   Male Heart disease 248.8 3.3

> ncol(mortal) # ilośc kolumn

[1] 6

> nrow(mortal) #ilość przypadków

[1] 400

> sapply(mortal, class) # identyfikacja cech

   Region    Status       Sex     Cause      Rate        SE
 "factor"  "factor"  "factor"  "factor" "numeric" "numeric"

> ilebrakujacych<-sum(is.na(mortal))
> ilebrakujacych # liczba brakujących danych

[1] 0

>
```

Plik zawiera 6 cech i 400 przypadków, w których kolumny odpowiednio nazywają się:

- Region - podzielony na 10 różnych klas odpowiadającym odpowiednim organą w Stanach Zjednoczonych

- Status - miejsce zamieszkania (wiejski lub miejski)
- Sex - płeć
- Cause - przyczyna śmierci
- Rate - wskaźnik sposobu śmierci na 100 000 osób w danym regionie
- SE - standardowy błąd dla wskażnika

4.2 Redukcja wymiaru na bazie MDS - skalowanie Kruskala

W każdej metodzie skalowania musimy zacząć od tworzenia macierzy odmienności:

```
> n <- dim(mortal)[1]
> n.subset <- 30
> subset.index <- sample(1:n, n.subset) # losujemy 30 samochodów
> mortal_subset <- mortal[subset.index,]
> # Przypisujemy nazwy regionów (pomo■e nam to zidentyfikować odpowiednie regiony)
> regiony <- paste(mortal_subset$Region, sep=" ")
> # Usuwamy niepotrzebne zmienne
> mortal_mds <- subset(mortal_subset, col=-c("Region"))
> niepodobienstwa = daisy(mortal_mds,stand=T)
> niepodobienstwa <- as.matrix(niepodobienstwa) #macierz odmienności
```

Przeprowadzamy skalowanie metryczne do przestrzeni o dwóch wymiarach:

```
> mds_k2<-cmdscale(niepodobienstwa, k=2)
> #skalowanie wielowymiarowe
> #nasze wyniki zawierą współ■rzędne obserwacji w nowym uk■ladzie współ■rzędnych
> str(mds_k2)

 num [1:30, 1:2] -0.509 -0.2 -0.19 -0.111 -0.159 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:30] "42" "178" "318" "236" ...
  ..$ : NULL
```

4.2.1 Jakość odwzorowania MDS dla różnych wymiarów przestrzeni

W sekcji powyżej obliczyliśmy naturalne odległości występujące w zmiennych i przedstawiliśmy w postaci macierzowej, naszym zadaniem będzie zbadanie różnicy w podobieństwach po zastosowaniu skalowania MDS i porównanie o ile zmieniły się dane w stosunku do stanu przed zastosowaniem skalowania. Najpowszechniejszą miarą stosowaną do szacowania, na ile dobrze (lub źle) dana konfiguracja odtwarza obserwowaną macierz odległości jest kryterium Stress. Kryterium Stress:

```
> # obliczamy odleg■łość w nowej przestrzeni metodą euklidesową
> dist_mds_k2 <- dist(mds_k2, method="euclidean")
> dist_mds_k2 <- as.matrix(dist_mds_k2) #nowa odleg■łość dla przestrzeni d =2
> dis_original <- niepodobienstwa #odleg■łość oryginalna
> STRESS <- sum((dis_original-dist_mds_k2)^2) #kryterium stress
> print("Wartość odmienności :")
```

```
[1] "Wartość odmienności :"
```

```
> STRESS
```

```
[1] 23.7933
```