Raport 2
Dyskretyzacja cech ciągłych. Redukcja wymiaru (PCA i MDS).

Jan Solarz 243889
Szymon Suszek 237288

27 kwietnia 2020

Spis treści

1 Krótki opis przeprowadzanych analiz

- Zadanie 1. Analiza 4 cech 150 kwiatow irysa o 3 gatunkach. Na poczatku badamy zdolnosc do separacji klas ¡gatunkow¿. Nastepnym krokiem jest zastosowanie poznanych algortymow: qual width, equal frequency, k-means clustering oraz dyskretyzacji przedziałów zadanych przez użytkownika, ktore pomogą nam przyporzadkowac dane obiekty do konkretnej klasy. Przyjrzymy sie skutecznosci dyskretyzcji algorytmow.

- Zadanie 2. PCA. W tym zadaniu skupimy sie na skladowych glownych cech i przeanalizujemy najwiekszy wklad wektorow ladunkowych.
- Zadanie 3. Pierwszy kontakt z MDS, metodą skalowania wielowymiarowego, mającego na celu wykrycie zmiennych ukrytych, które choć nie obserwowane bezpośrednio, wyjaśniają podobieństwa i różnice pomiędzy badanymi obiektami.

2 Zadanie 1. Dyskretyzacja (przedzialowanie) cech ciągłych

2.1 Wprowadzenie do zadania i przygotowanie danych

```
## Error in install.packages(arules): nie znaleziono obiektu 'arules'
## Loading required package: Matrix
##
## Attaching package: 'arules'
## The following objects are masked from 'package:base':
##
##    abbreviate, write
```

2.2 Analiza właściwa danych iris

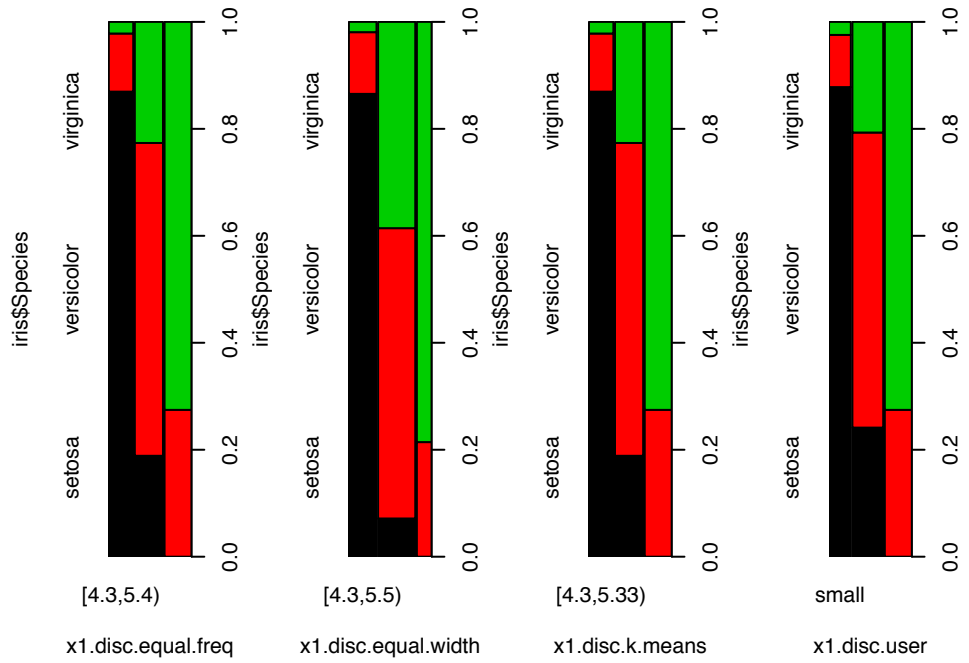
2.2.1 Porównanie nienadzorowanych metod dyskrytyzacji

- W tej sekcji przeprowadzimy kompletną analizę. Badać będziemy kolejno każdą cechę przy użyciu wszystkich czterech algorytmow.
- W raporcie ukazane są jedynie porównawcze wykresy przy każdej zmiennej

Przekształcenie zmiennej ciągłej na zmienną jakościową

1. Sepal.Length

```
##porównanie wszystkich metod
par(mfrow=c(1,4))
plot(iris$Species~x1.disc.equal.freq, col=1:3) #metoda1
plot(iris$Species~x1.disc.equal.width, col=1:3) #metoda2
plot(iris$Species~x1.disc.k.means, col=1:3) #metoda3
plot(iris$Species~x1.disc.user, col=1:3) #metoda4
```



Rysunek 1: Analiza Sepal.Length

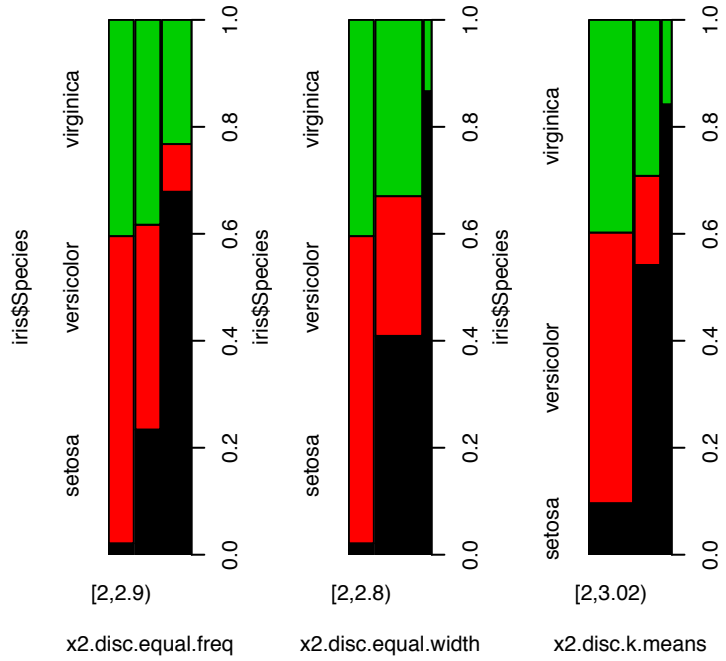
```
matchClasses(tab.equal.freq1) #metoda1

## Cases in matched pairs: 72 %
##    [4.3,5.4)    [5.4,6.3)    [6.3,7.9]
##     "setosa" "versicolor"  "virginica"

matchClasses(tab.equal.width1) #metoda2

## Cases in matched pairs: 70 %
##    [4.3,5.5)    [5.5,6.7)    [6.7,7.9]
##     "setosa" "versicolor"  "virginica"

matchClasses(tab.k.means1) #metoda3

## Cases in matched pairs: 72 %
##   [4.3,5.33)  [5.33,6.27)   [6.27,7.9]
##     "setosa" "versicolor"  "virginica"

matchClasses(tab.user1) #metoda4

## Cases in matched pairs: 70 %
##        small       medium        large
##     "setosa" "versicolor"  "virginica"
```

- Zaczynamy od histogramu zmiennej, aby zobaczyc rozkład długości i obrać przedziały do algorytmu "user".
- Skuteczność wszystkich algorytmów w okolicach 70 procent
- Skuteczność bardzo podobna, najlepiej equal.freq i k.means , ropzietosc 2 procent

2. Sepal.Width

```
## Error in discretize(x2, method = "fixed", breaks = przedzial2, labels =
c("small", : The calculated breaks are: -Inf, 2.8, 3, 3, Inf
##  Some breaks are not unique. Look at the distribution of the data (e.g.,
histogram) to determine appropriate breaks and use the discretization method
'fixed'.
## Error in table(x2.disc.user, iris$Species): nie znaleziono obiektu 'x2.disc.user
```

```
##porównanie wszystkich metod
par(mfrow=c(1,4))
plot(iris$Species~x2.disc.equal.freq, col=1:3) #metoda1
plot(iris$Species~x2.disc.equal.width, col=1:3) #metoda2
plot(iris$Species~x2.disc.k.means, col=1:3) #metoda3
plot(iris$Species~x2.disc.user, col=1:3) #metoda4

## Error in eval(predvars, data, env): nie znaleziono obiektu 'x2.disc.user'

matchClasses(tab.equal.freq2) #metoda1

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

matchClasses(tab.equal.width2) #metoda2

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

matchClasses(tab.k.means2) #metoda3

## Cases in matched pairs: 56 %
##     [2,3.02)  [3.02,3.55)   [3.55,4.4]
## "versicolor"     "setosa"     "setosa"

matchClasses(tab.user2) #metoda4

## Error in apply(tab, 1, which.max): nie znaleziono obiektu 'tab.user2'
```

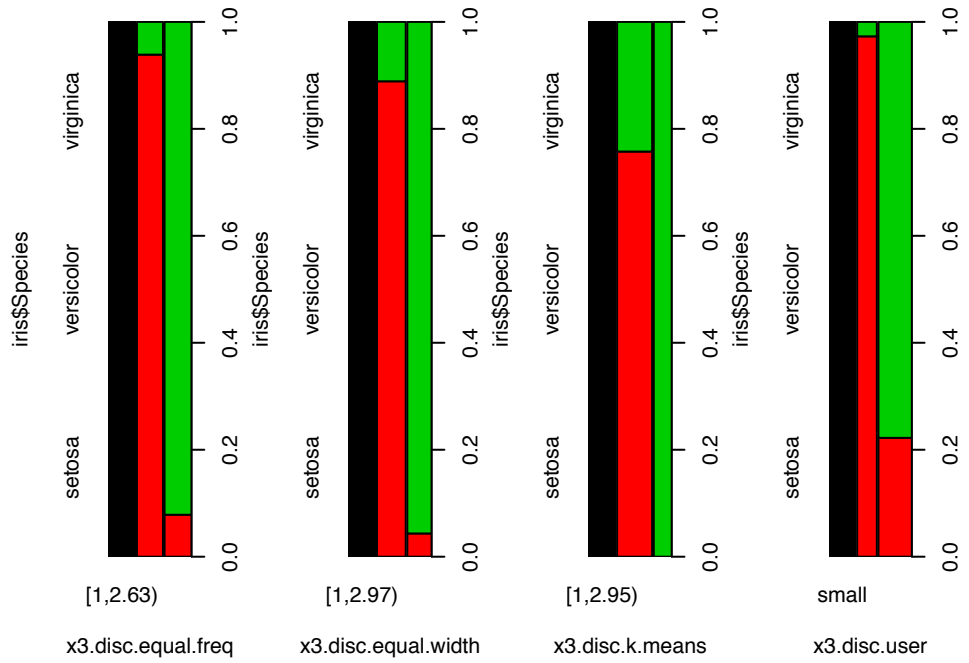


Rysunek 2: Analiza Sepal.Width

- Skuteczność w okolicach 50 procent
- Najlepiej equal.freq i k.means, rozpietosc 6 procent

3. Petal.Length

```
##porównanie wszystkich metod
par(mfrow=c(1,4))
plot(iris$Species~x3.disc.equal.freq, col=1:3) #metoda1
plot(iris$Species~x3.disc.equal.width, col=1:3) #metoda2
plot(iris$Species~x3.disc.k.means, col=1:3) #metoda3
plot(iris$Species~x3.disc.user, col=1:3) #metoda4
```



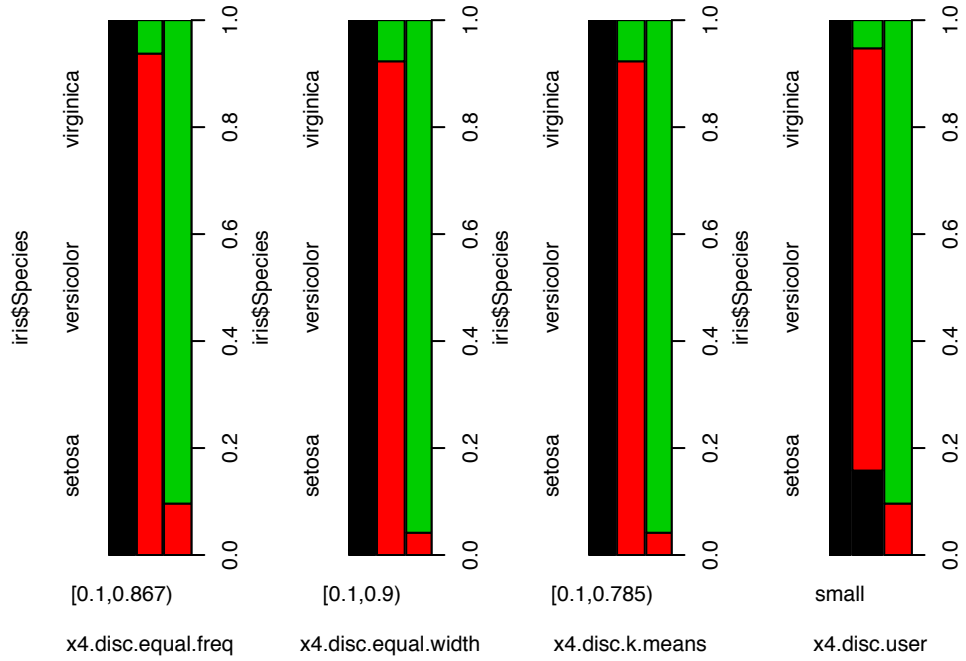
Rysunek 3: Analiza Petal.Length

```
matchClasses(tab.equal.freq3) #metoda1

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

matchClasses(tab.equal.width3) #metoda2

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

matchClasses(tab.k.means3) #metoda3

## Cases in matched pairs: 89.33 %
##     [1,2.95)  [2.95,5.13)   [5.13,6.9]
##     "setosa" "versicolor"  "virginica"

matchClasses(tab.user3) #metoda4

## Cases in matched pairs: 90 %
##        small       medium        large
##     "setosa" "versicolor"  "virginica"
```

- Skuteczność w okolicach 93 procent
- Bardzo wysoka skutecznosc, bliskie siebie wyniki, nejlepiej equal.freq i k.means, rozpietosc 5.33 procent.

4. Petal.Width

```
##porównanie wszystkich metod
par(mfrow=c(1,4))
plot(iris$Species~x4.disc.equal.freq, col=1:3) #metoda1
plot(iris$Species~x4.disc.equal.width, col=1:3) #metoda2
plot(iris$Species~x4.disc.k.means, col=1:3) #metoda3
plot(iris$Species~x4.disc.user, col=1:3) #metoda4
```



Rysunek 4: Analiza Petal.Width

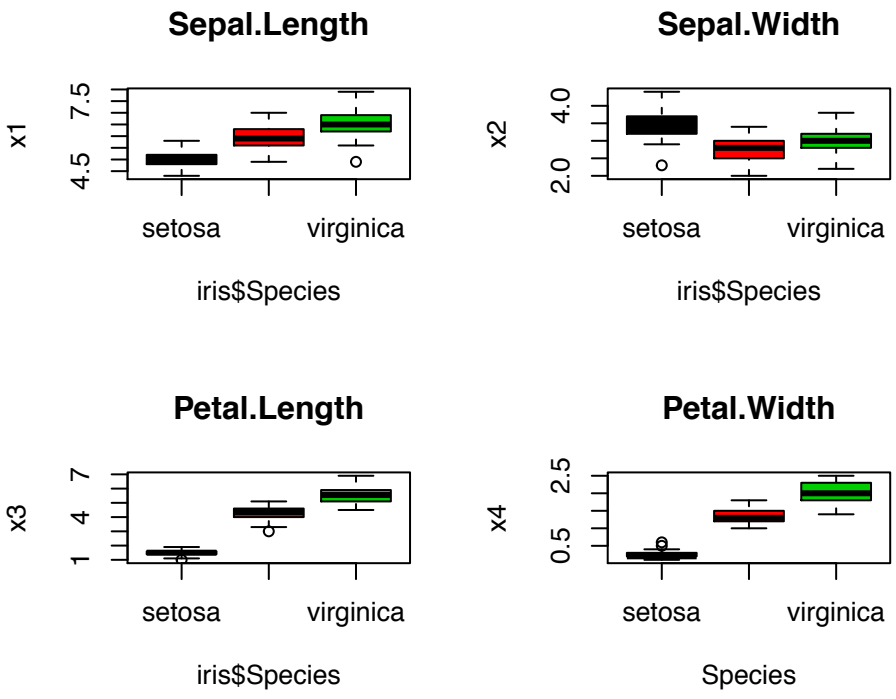
```
matchClasses(tab.equal.freq4) #metoda1

## Cases in matched pairs: 94.67 %
##  [0.1,0.867)  [0.867,1.6)    [1.6,2.5]
##     "setosa" "versicolor"  "virginica"

matchClasses(tab.equal.width4) #metoda2

## Cases in matched pairs: 96 %
##    [0.1,0.9)    [0.9,1.7)    [1.7,2.5]
##     "setosa" "versicolor"  "virginica"
```

```
matchClasses(tab.k.means4) #metoda3

## Cases in matched pairs: 96 %
##  [0.1,0.785) [0.785,1.69)   [1.69,2.5]
##     "setosa" "versicolor"  "virginica"

matchClasses(tab.user4) #metoda4

## Cases in matched pairs: 88.67 %
##        small       medium        large
##     "setosa" "versicolor"  "virginica"
```

- Skuteczność w okolicach 95 procent
- Bardzo wysoka skutecznosc, nejlepiej k.means 96 procent, rozpietosc 7.33 procent, najgorzej metoda user(reczna).

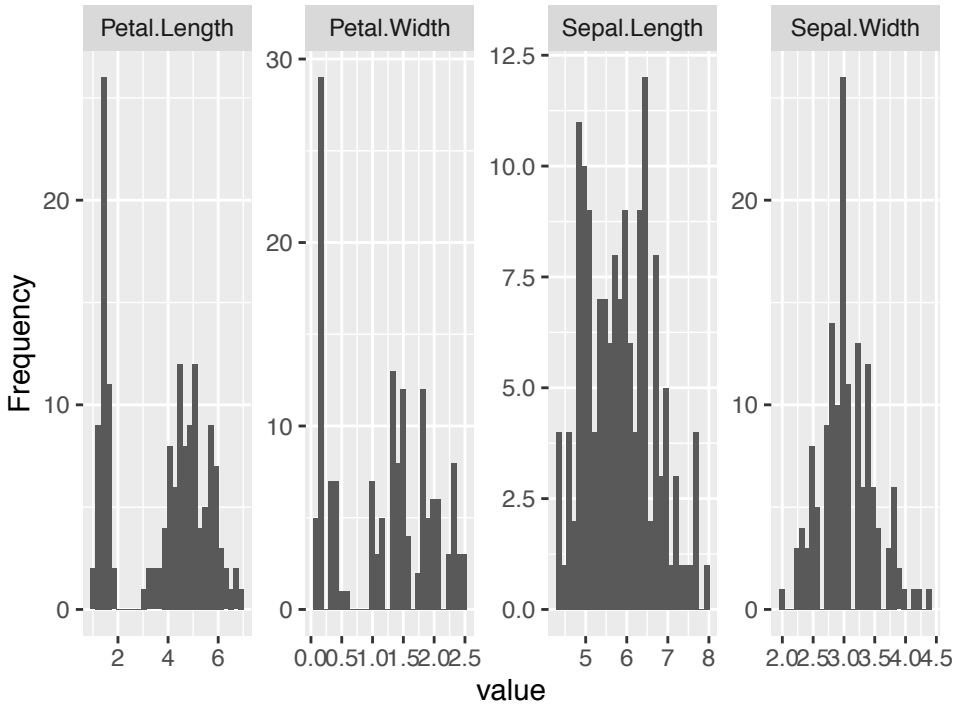
2.2.2 Wstępna analiza klas ze względu na cechy

```
par(mfrow=c(2,2))
boxplot(x1~iris$Species, col=1:3, main="Sepal.Length")
boxplot(x2~iris$Species, col=1:3, main="Sepal.Width")
boxplot(x3~iris$Species, col=1:3, main="Petal.Length")
boxplot(x4~Species, col=1:3, main="Petal.Width")
```

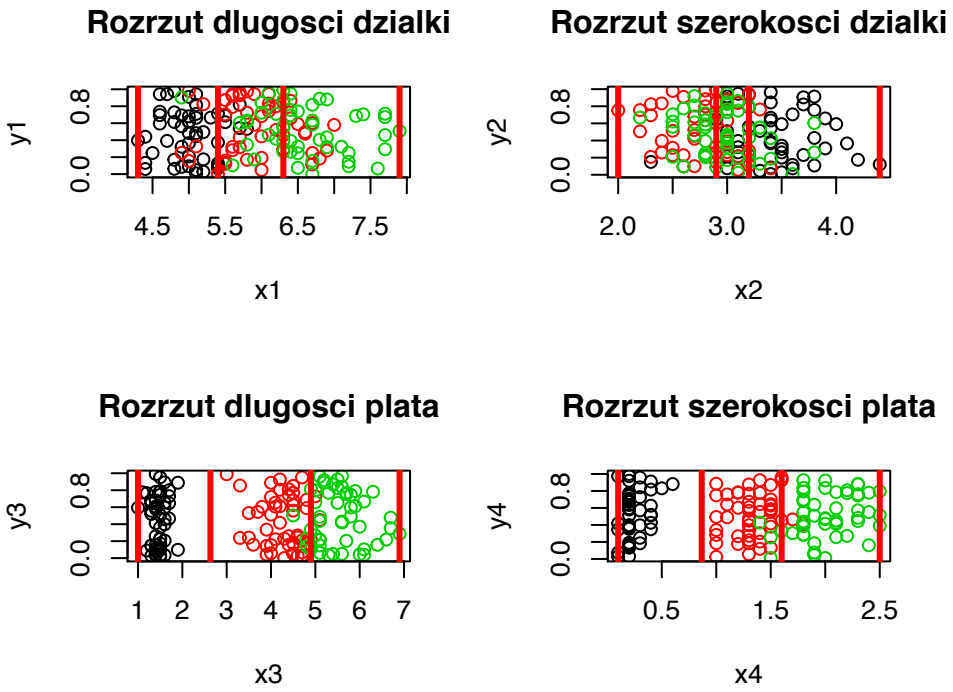


Rysunek 5: Wstępne wykresy pudełkowe

```
plot_histogram(iris)
```




```
par(mfrow=c(2,2))
plot(x1, y1, col=iris$Species, main = "Rozrzut dlugosci dzialki")
abline(v = breaks.equal.frequency1, col = "red", lwd=3)

plot(x2, y2, col=iris$Species, main = "Rozrzut szerokosci dzialki")
abline(v = breaks.equal.frequency2, col = "red", lwd=3)

plot(x3, y3, col=iris$Species, main = "Rozrzut dlugosci plata")
abline(v = breaks.equal.frequency3, col = "red", lwd=3)

plot(x4, y4, col=iris$Species, main = "Rozrzut szerokosci plata")
abline(v = breaks.equal.frequency4, col = "red", lwd=3)
```



- 4 cechy ilościowe, 1 jakościowa (klasa-gatunek), brak brakujacych danych
- Po wstepnych rysunkach widzimy lepsza separacje cech dotyczacych Płata irysa, a więc Petal.Length i Petal.Width niz w pzypadku dzialki kielicha.

2.3 Wpływ obserwacji odstających

```
x1[which.min(x1)] <- min(x1) - 2*IQR(x1)
x1[which.max(x1)] <- max(x1) + 2*IQR(x1)

x2[which.min(x2)] <- min(x2) - 2*IQR(x2)
x2[which.max(x2)] <- max(x2) + 2*IQR(x2)

x3[which.min(x3)] <- min(x3) - 2*IQR(x3)
x3[which.max(x3)] <- max(x3) + 2*IQR(x3)

x4[which.min(x4)] <- min(x4) - 2*IQR(x4)
x4[which.max(x4)] <- max(x4) + 2*IQR(x4)
```

- Equal.freq - nie wplynela, equal.width duzo mniejsza, k.means - troche mneijsza, user nie wplynela.
- Equal.freq - nie wplynela, equal.width duzo mniejsza, k.means - taka sama, user nie wplynela.
- Equal.freq - nie wplynela, equal.width z ponad 90 procent na 35 procent !!, k.means - z 95 procent na 67 procent, user nie wplynela.

- Equal.freq - nie wplynela, equal.width z 96 procent na 35 procent !!, k.means - z 95 procent na 86 procent, user nie wplynela.

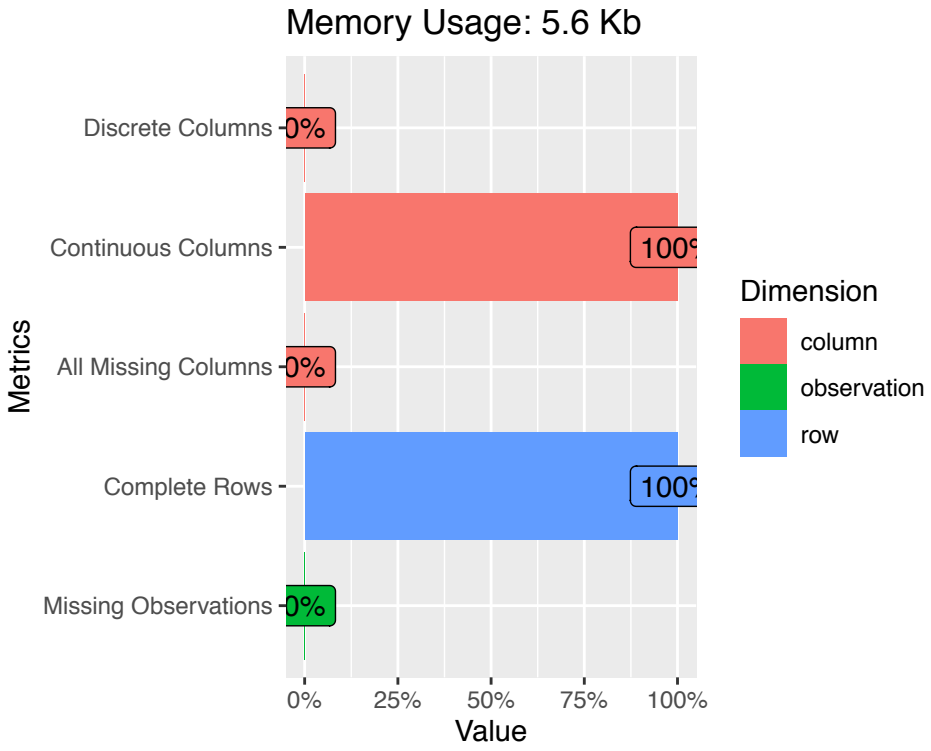
2.4 Wnioski dyskretyzacji

- Najlepszą zdolność dyskriminacyjną ma cecha Szerokosci Plata i Dlugosci Plata
- Najgorsza separacje na klasy ma szerokosc dzialki kielicha.
- Najbardziej skutecznym algorytmem jest k.means, zaraz po nim equal.freq. Rozpietossci sa wieksze przy cechach o lepszej separacji. Wplyw na to tez moze miec metoda wpisywania przedzialow przez uzytkownika.
- Przy algorytmie wpisywania przedziałów przez użytkownika działamy intuicyjnie patrząc na histogramy. Jest on jednak bardzo mylny i najmniej skuteczny, należy uważać na zmiane etykietyzacji.

3 Zadanie 2. Analiza składowych głównych PCA

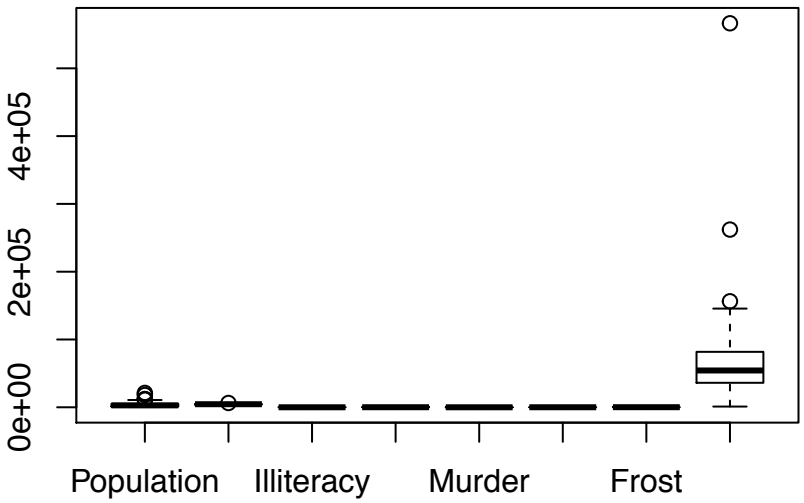
3.1 Wprowadzenie do zadania i przygotowanie danych

```
plot_intro(dane)
```



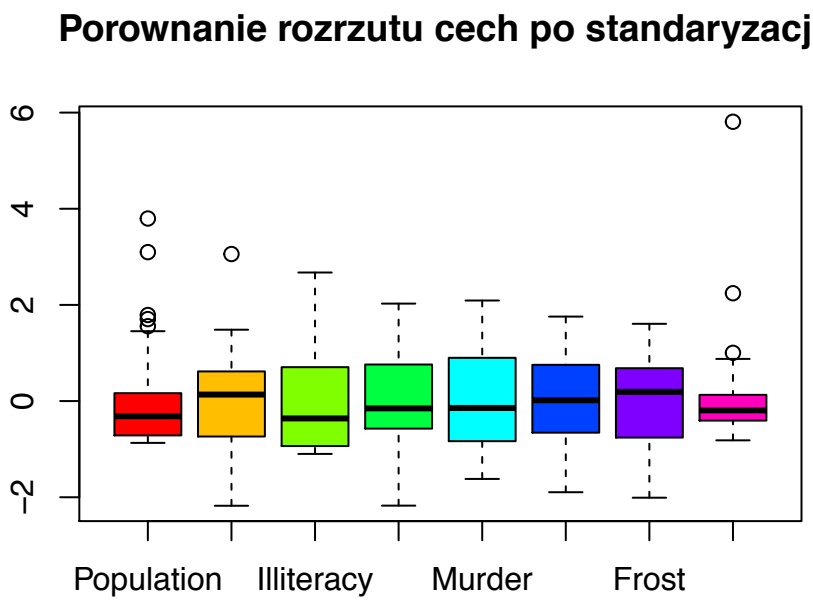
Rysunek 6: Wstępne wykresy

```
boxplot(dane) # dane zdominowane przez zmienna area
```



Rysunek 7: Wstępne wykresy

```
boxplot(scale(dane),col=rainbow(8),main="Porownanie rozrzutu cech po standaryzacji")
```



Rysunek 8: Wstępne wykresy

- Zaczynamy od wczytania danych, zajmować się będziemy analizą na 50 stanach USA.
- Zbiór charakteryzuje 8 zmiennych ilościowych, brak brakujących obserwacji/danych.
- Poza Area zmienna Population wyraża się także dużymi wartościami ¡ich wariancje odbiegaja od reszty¿, aby móc porównać wariancje cech należy dokonać stadaryzacji

3.2 Analiza właściwa

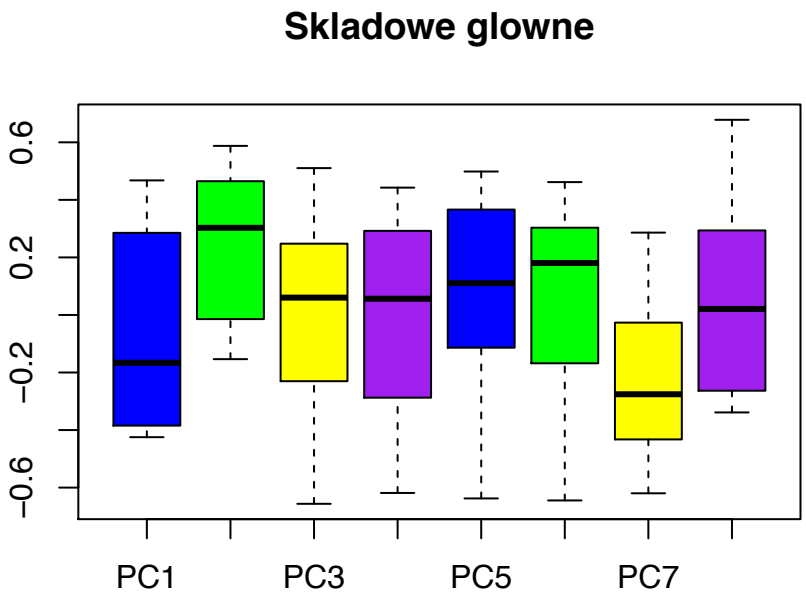
3.2.1 Wyznaczenie składowych głównych.

```
dane.po.pca <- prcomp(dane, scale.=TRUE)
dane.po.pca$rotation
```

```
##                    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
```

```
boxplot(dane.po.pca$rotation[,1:8], main="Skladowe glowne", col=c("blue","green","yellow
```

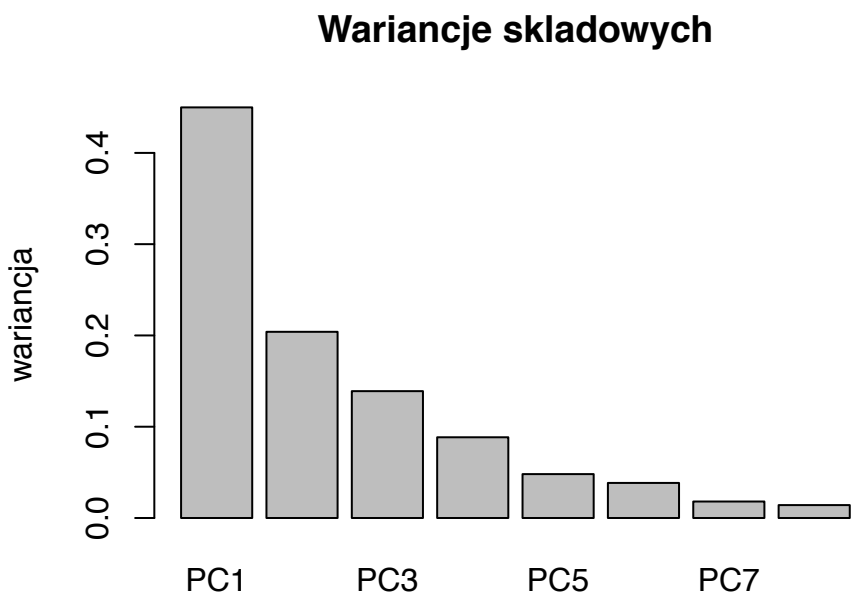


Rysunek 9: Pierwsze składowe główne

- Największą różnorodnością wartości charakteryzują się wektory ładunkowe PCA3-PCA6.
- Widzimy że PCA1 przypisuje się największą wagę cechom Illiteracy, LifeExp, Murder oraz HS Grad, są one na podobym poziomie około 0.43. PC1 możemy interpretować jako ogólny wskaźnik wykształcenia i długości życia, a więc wskaźznik związany z człowiekiem.
- Do PCA2 natomiast przypisane są wysokie wartości Area, Income oraz Population na poziomie 0.41-0.58, przypisać go możemy do charakterystyki danego Stanu związanego z zarobkami.

3.2.2 Zmienności odpowiadające poszczególnym składowym głównym.

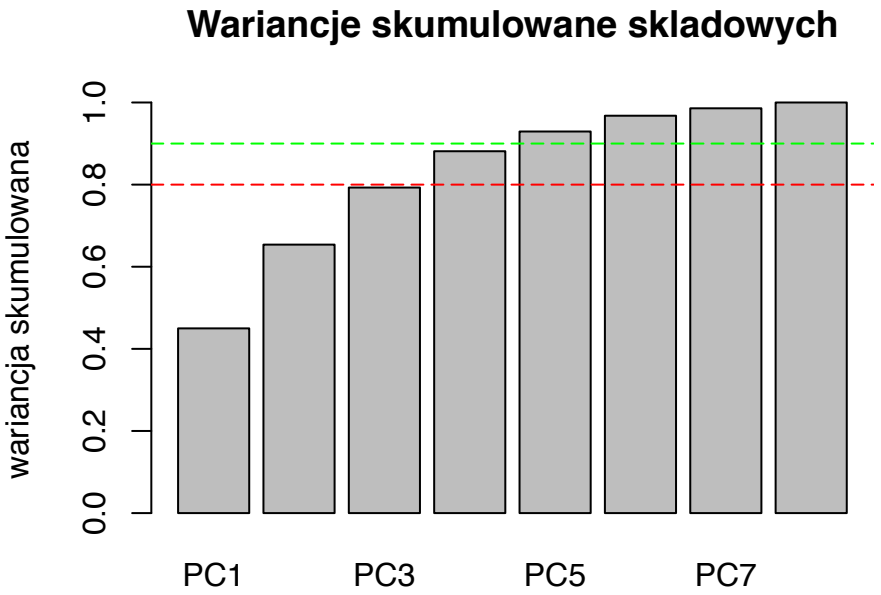
```
podsumowanie <- summary(dane.po.pca)
podsumowanie

## 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

barplot(podsumowanie$importance[2,], ylab = "wariancja")
title("Wariancje skladowych")
```



Rysunek 10: Wkład wyjaśnionej zmienności

- Widać zdecydowaną przewagę zmienności składowej PC1- aż 45 procent.
- Zmienność maleje z każdą kolejną składową.
- Standard deviation - odchylenie standardowe dla każdej składowej głównej PC1, ... (liczone dla elementów każdej kolumny)
- Cumulative Proportion - kumulacyjna proporcja zmienności np. w kolumnie PC2 jest to zmienność wszystkich danych opisana przez dwie pierwsze składowe główne 65,39 procent

```
barplot(podsumowanie$importance[3,], ylab = "wariancja skumulowana")
abline(h=0.8, col = "Red", lty = 5, )
abline(h=0.9, col = "Green", lty = 5)
title("Wariancje skumulowane skladowych")
```

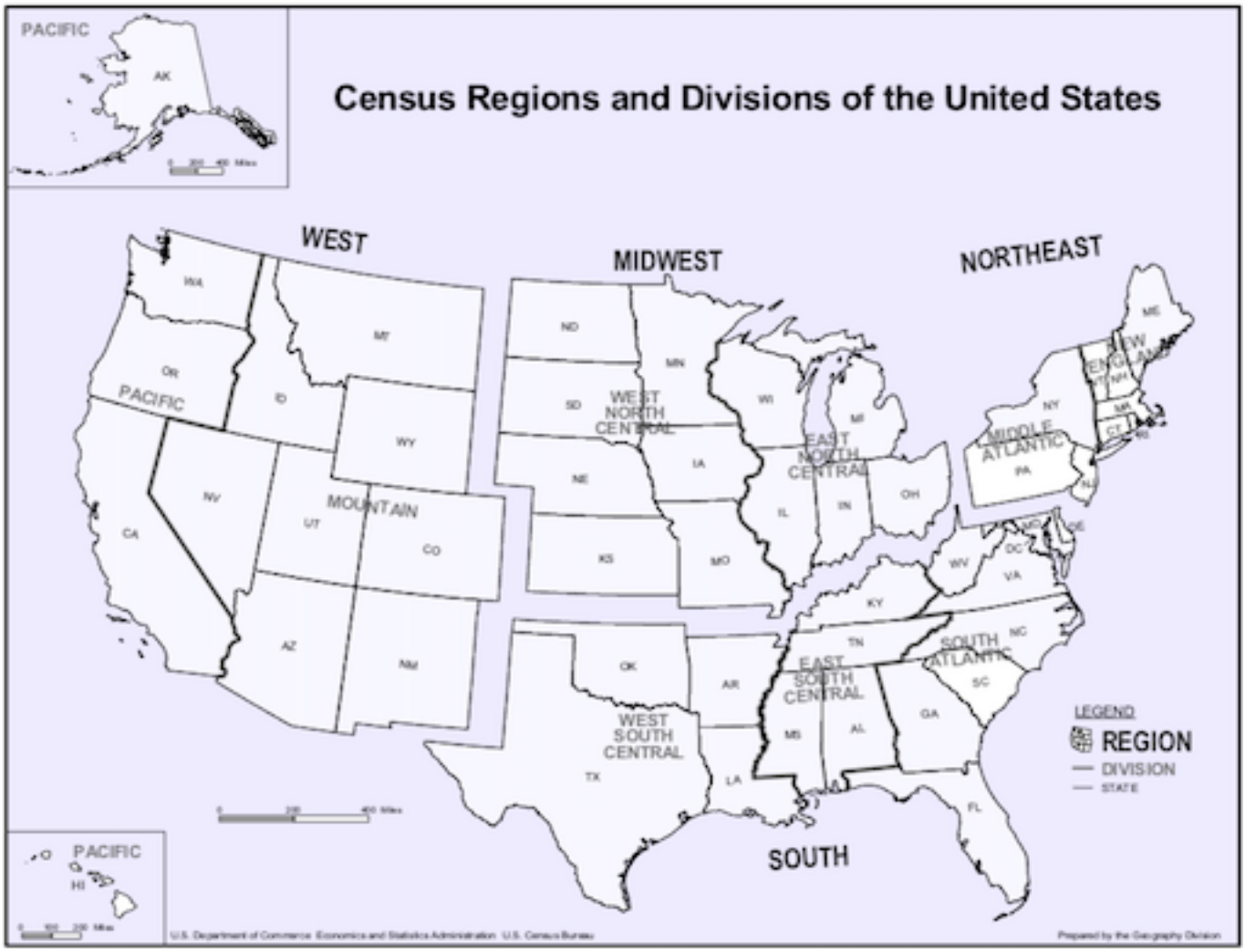


Rysunek 11: Ile składowych jest nam potrzebnych?

- Aby pokryć 80 procent zmienności wystarczyłyby prawie 3 składowe 79,28 procent, ale aby w ppełni pokryć 80 procent potrzebujemy jednak 4 składowych 88.5 procent (mozna łatwo odczytać z Cumulative proportion).
- do wyjaśnienia 90 procent zmienności potrzeba 5 składowych 92,9 procent

3.2.3 Wizualizacja danych wielowymiarowych

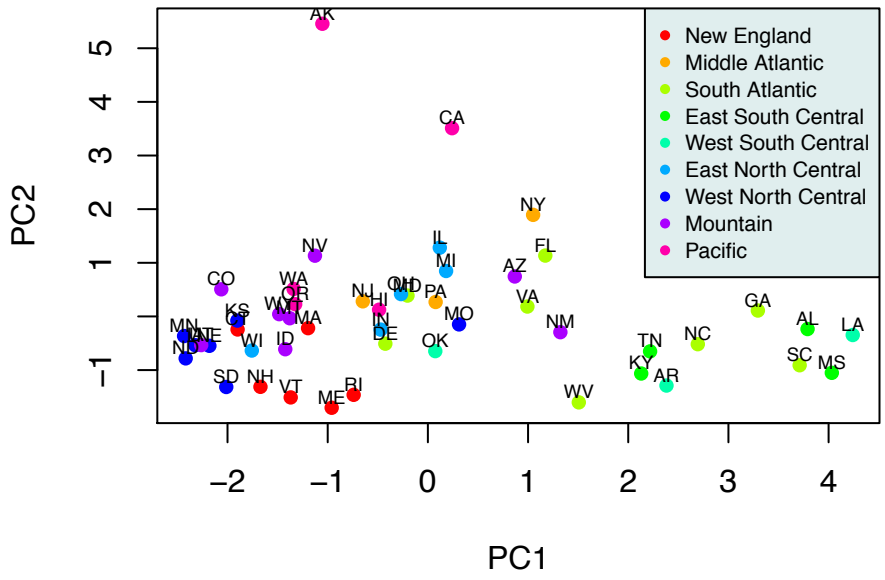
Badanie podobieństw poszczególnych stanów na podstawie charakterystycznych cech z głównych składowych



Mapa dywizji użytych w zadaniu

```
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
```

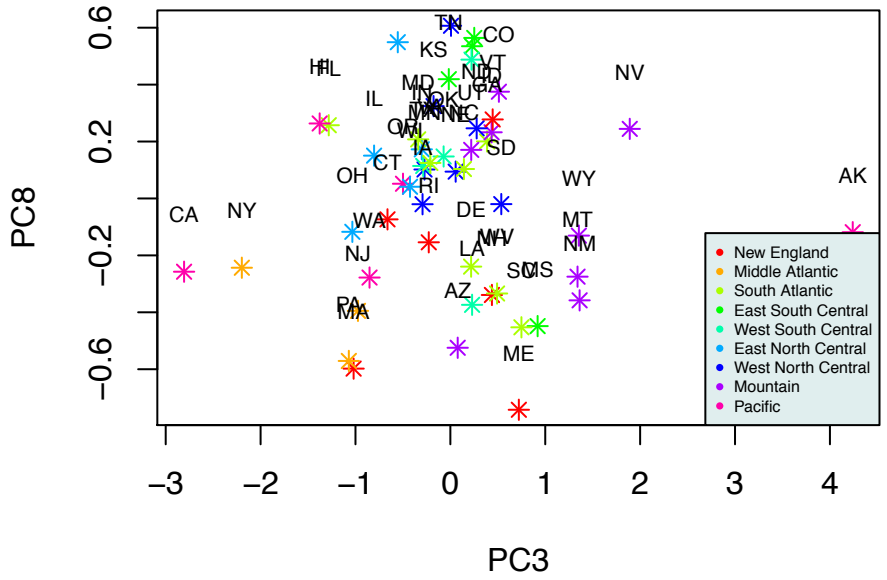
```
plot(dane.po.pca$x[,1], dane.po.pca$x[,2], col=kolory[as.numeric(state.division)], pch=1
text(dane.po.pca$x[,1], dane.po.pca$x[,2]+0.2, labels=state.abb, cex=0.6)
legend("topright",legend=levels(state.division), col=kolory, pch=16, cex=0.7, bg="azure2
```



Rysunek 12: Wykres rozrzutu podobieństwa stanów 2D PC1-PC2

- Przy zestawieniu PC1 i PC2 zwracmy uwage na cechy Area, Illeteracy i LifeExp
- Wyizolowany Stan Alaski, zdecydowanie najwieksza powierzchnia, ktora wyróżnia PC1. Drugim takim Stanem jest California
- Grupa stanów z dywizji South Atlantic i East South Central charakteryzuje się wysokim wskaźnikiem PC1 (położenie po dolnej prawej stronie)- wysoki analfabetyzm.

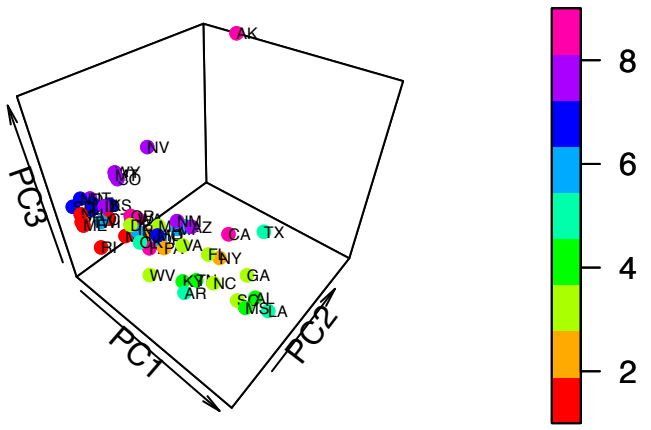
```
plot(dane.po.pca$x[,3], dane.po.pca$x[,8], col=kolory[as.numeric(state.division)], pch=8
text(dane.po.pca$x[,3], dane.po.pca$x[,8]+0.2, labels=state.abb, cex=0.7)
legend("bottomright",legend=levels(state.division), col=kolory, pch=16, cex=0.5, bg="azu
```



Rysunek 13: Wykres rozrzutu podobieństwa stanów 2D PC3-PC8

- Przy zestawieniu PC3 i PC8 zwracmy uwage na cechy Frost i Murder
- Wyizolowany Stan Alaski, zdecydowanie ma najwiecej dni poznizej zera stpopni, pozniej taka ceche wykazuja stany Wyoming, Minesota i Montana- wszystkie na polnocy USA
- Grupa stanów z dywizji South Atlantic i East South Central charakteryzuje się wysokim wskaźnikiem PC1 (położenie po dolnej prawej stronie)- wysoki analfabetyzm.
- Stany z północy charaktryzuja sie mniejszym wskaznikiem morderstw.

```
scatter3D(dane.po.pca$x[,1], dane.po.pca$x[,2], dane.po.pca$x[,3],  colvar=as.numeric(st
text3D(dane.po.pca$x[,1], dane.po.pca$x[,2], dane.po.pca$x[,3],  labels = state.abb, add
```

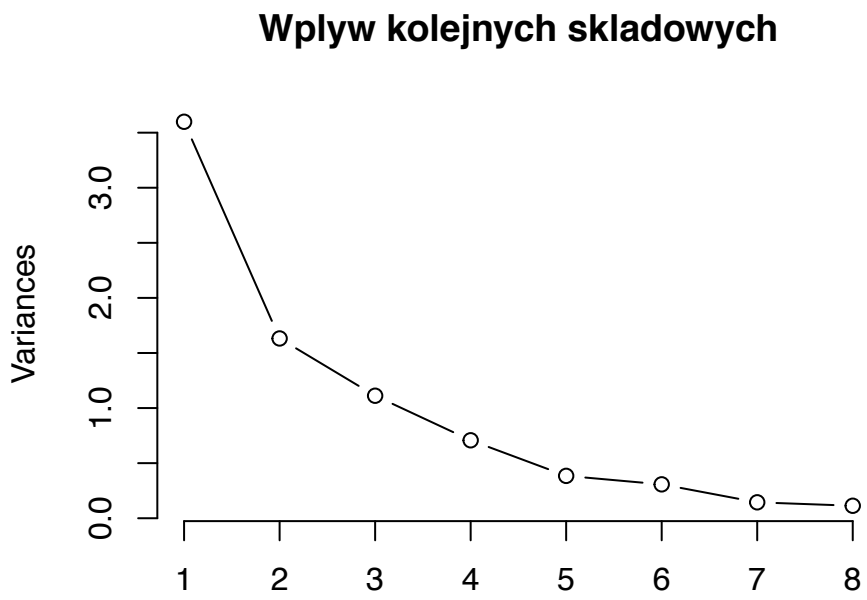


Rysunek 14: Wykres rozrzutu podobieństwa stanów 3D PC1-PC2-PC3

- Widzimy zdecydowane grupowanie sie stanów ze wzledu na dywziej(ten sam kolor), izolacja Alaski.
- Dywizja niebieska/zielona(PCA1 i PCA2), dywizja fioletowa (PC3 i PCA2)
- Duże skuepienie głównie kolorów czerwonych ale również ciemnoniebieskich w splocie osi PC1, PC2 i PC3- niskie wskazniki

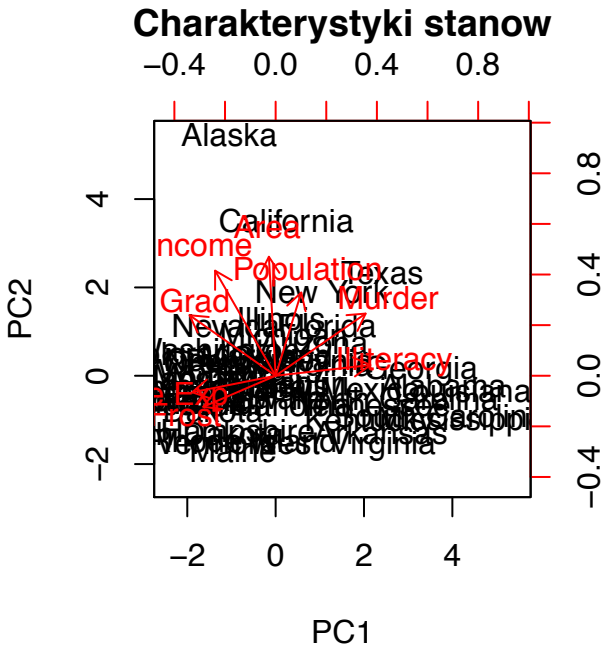
3.2.4 Korelacja zmiennych

```
plot(dane.po.pca, type='l', main="Wplyw kolejnych skladowych")
```



Rysunek 15: Wykresy korelacji

```
biplot(dane.po.pca, scale=0, main="Charakterystyki stanow")
```



Rysunek 16: Wykresy korelacji

- Czerwone strzałki (osie) wyznaczone są przez wektory własne dla każdej zmiennej.
- Wykres wizualizuje relację pomiędzy zmiennymi.
- Wyodrębnienie (wykres charakterystyki stanów) Alaski i Californi ze wzgledu na Powierzchnie, Texas, New York, również California i Florida na populacje.

```
##                    PC1         PC2         PC3         PC4          PC5
## Population  0.23984363  0.52487776 -0.69208615 -0.34434757  0.251765858
## Income     -0.56690291  0.66297778 -0.10582738 -0.07439531 -0.395428165
## Illiteracy  0.88720374  0.06766573  0.07476148  0.29677518  0.002186803
## Life Exp   -0.78085597 -0.10431289 -0.37954435  0.37225450  0.202555453
## Murder      0.84278855  0.39211733  0.11437750 -0.13929172 -0.079427577
## HS Grad    -0.80565843  0.38166418  0.05241692  0.19478457 -0.061563366
## Frost      -0.67803840 -0.19619845  0.40820686 -0.52036774  0.134807911
## Area       -0.06333314  0.75067024  0.53819393  0.16917324  0.309171079
##                    PC6          PC7          PC8
## Population -0.005908762 -0.023624278 -0.073761915
## Income      0.256048018  0.003460375  0.020284268
## Illiteracy  0.214819050 -0.235563887 -0.113946217
## Life Exp    0.121479057 -0.097377400  0.177446387
## Murder     -0.180318726 -0.112135329  0.228186671
## HS Grad    -0.357451480 -0.149372507 -0.103366533
## Frost       0.117931704 -0.179395279  0.009536022
## Area        0.082264769  0.108797248  0.004442002
```

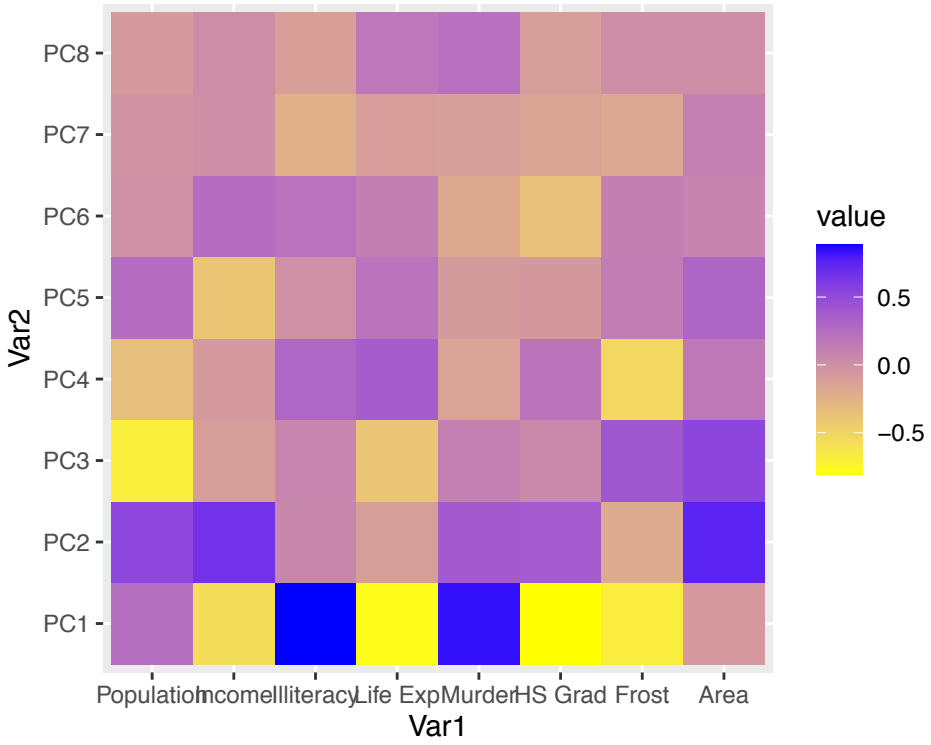
```
library(reshape2)
library(ggplot2)
K <- melt(korelacja)
head(K)
```

```
##         Var1 Var2      value
## 1 Population  PC1  0.2398436
## 2     Income  PC1 -0.5669029
## 3 Illiteracy  PC1  0.8872037
## 4   Life Exp  PC1 -0.7808560
## 5     Murder  PC1  0.8427885
## 6    HS Grad  PC1 -0.8056584
```

```
library(hrbrthemes)
```

```
## Error: package or namespace load failed for 'hrbrthemes' in dyn.load(file, DLLpath
= DLLpath, ...):
##  nie można załadować współdzielonego obiektu  '/Library/Frameworks/R.framework/Versio
##  dlopen(/Library/Frameworks/R.framework/Versions/3.6/Resources/library/systemfonts/li
6): Library not loaded: /opt/X11/lib/libfreetype.6.dylib
##  Referenced from: /Library/Frameworks/R.framework/Versions/3.6/Resources/library/syst
##  Reason: image not found
```

```
ggplot(data = K, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()+
  scale_fill_gradient(low="yellow", high="blue")
```



Rysunek 17: Mapa ciepła

- Widzimy wizaulizacje na mapie ciepła korealcji pomiędzy wartościami obserwacji dla poszczególnych zmiennych a wartościami po wykoaniu PCA
- Widzimy wyraziste kolory niebieskie i żółte przy PC1, które ma największy wpływ

3.3 Wnioski do PCA

- Zdecydowanie największy wpływ na zmienność ma główna składowa PC1- ponad 40 procent. Po korelacji widzimy, że ma ona wpływ głownie na zmienne Income, Illiteracy, LiveExp, Murder, Hs Grad i Frost.
- Każdy następny posiada mniejszy wpływ- wprowadza mniejszą zmienność
- Już 3 składowe zawierają niemal 80 procent całej zmienności.
- Zdecydowanie widać podobieństwo w danych zmiennych ze względu na położenie danego stanu (dystrykty). Główne składowe w większości przypadków poprawnie wskazują różnice stanów w poszczególnych zmiennych. Na wykresach widać jednak duże skupiska.

4 Zadanie 3. Skalowanie wielowymiarowe MDS

4.1 Wprowadzenie do zadania i przygotowanie danych

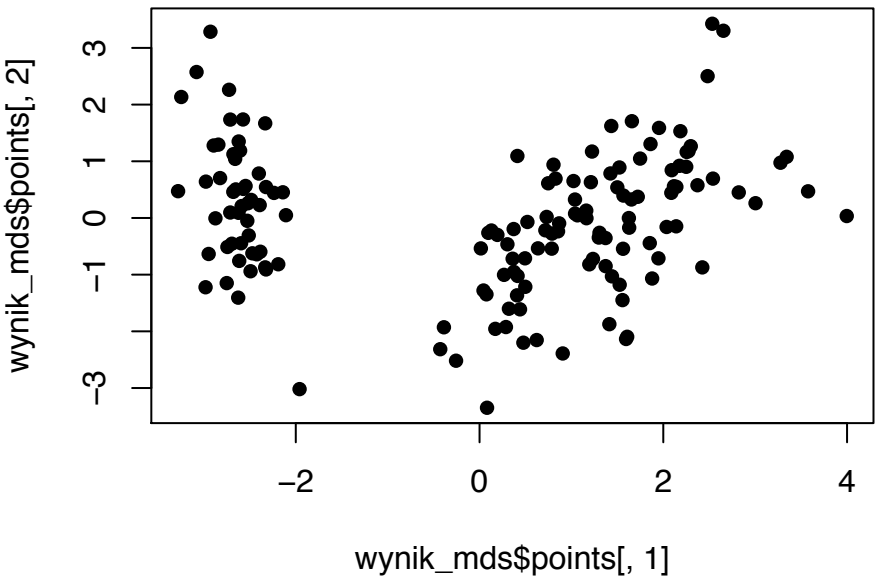
- Do analizy wybrano zbiór danych iris, który został opisany podczas omawiania zadania 1. Zmienna grupująca Species nie została wykorzystana do redukcji wymiaru, a jedynie podczas wizualizacji.

```
dane_iris <- iris[-102,]
podobienstwa_iris <- daisy(dane_iris[-5], stand=T)
# wykorzystanie metody skalowania niemetrycznego
wynik_mds <- isoMDS(podobienstwa_iris, k=2)

## initial  value 4.671392
## iter   5 value 4.072982
## iter   5 value 4.071831
## iter   5 value 4.071830
## final  value 4.071830
## converged
```

4.2 Analiza właściwa z wizualizacją daych

- Analiza skalowania wielowymiarowego NIEMETRYCZNEGO
- Usunięcie wiersza nr 102 ponieważ kolejna metoda nie przyjmuje jako argumentu wartości zerowych oraz ujemnych.
- Ponownie usuwamy zmienna gatunek - uzyta tylko podczas wizualizacji

```
plot(wynik_mds$points[,1], wynik_mds$points[,2], pch=16)
```

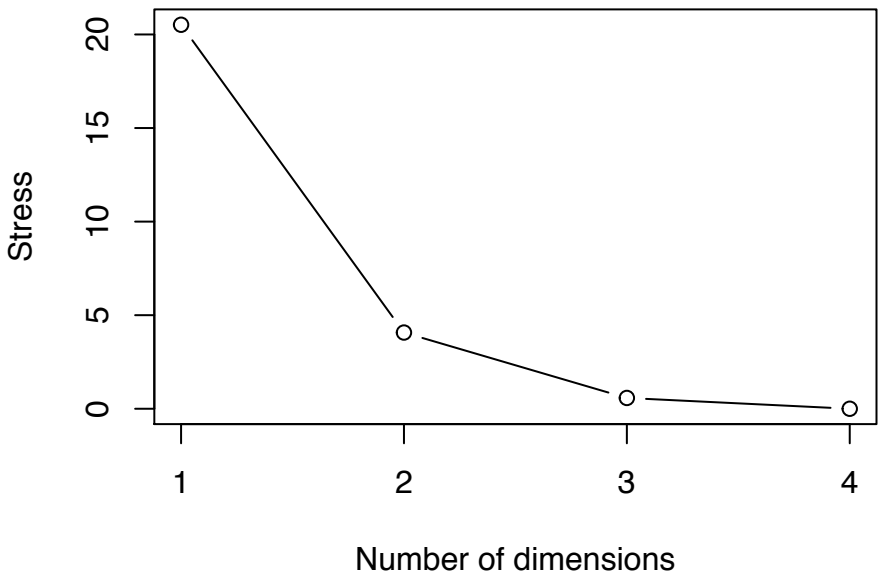


Rysunek 18: Wykres w 2 wymiarach

4.2.1 Funkcja do wyznaczania wykresu wartości funkcji kryterialnej STRESS

```
scree.plot = function(d, k) {
  stresses=isoMDS(d, k=k)$stress
  for(i in rev(seq(k-1)))
    stresses=append(stresses,isoMDS(d, k=i)$stress)
  plot(seq(k),rev(stresses), type="b", xaxp=c(1,k, k-1), ylab="Stress", xlab="Number of
}

#Wykres funkcji kryterialnej STRESS
scree.plot(podobienstwa_iris, k =4)

## initial  value 0.000000
## final  value 0.000000
## converged
## initial  value 0.687495
## iter   5 value 0.582213
## final  value 0.576638
## converged
## initial  value 4.671392
## iter   5 value 4.072982
## iter   5 value 4.071831
## iter   5 value 4.071830
## final  value 4.071830
```

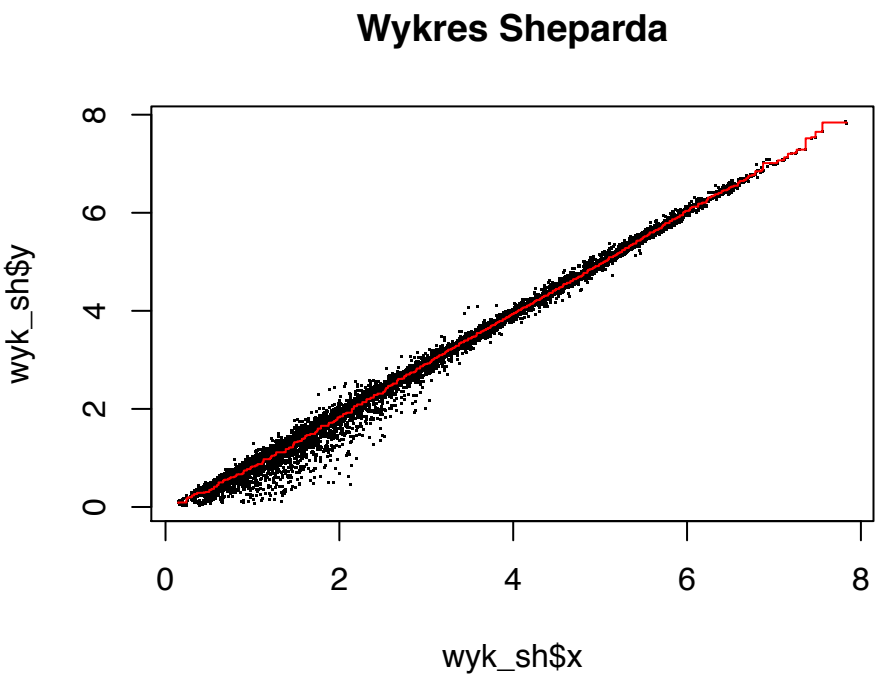
```
## converged
## initial  value 25.520935
## iter   5 value 20.532876
## iter   5 value 20.518557
## iter   5 value 20.518557
## final  value 20.518557
## converged
```



Rysunek 19: jakis tytul

Funkcja wyznacza dla każdego z wymiarów funkcję stresu a następnie tworzy wykres

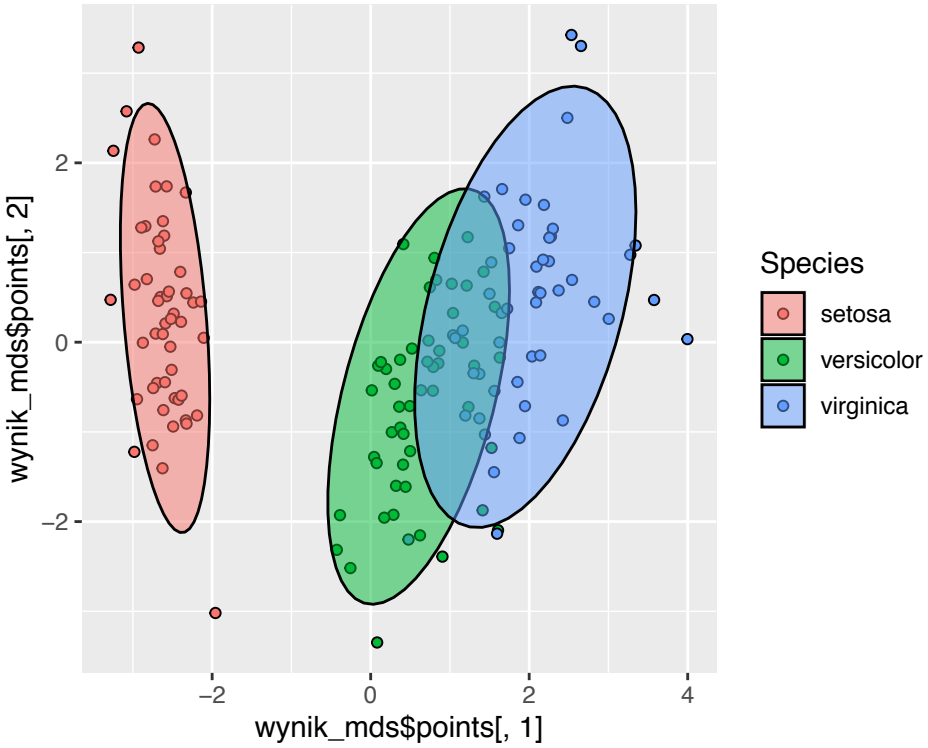
```
wyk_sh <- Shepard(podobienstwa_iris, wynik_mds$points)
plot(wyk_sh, pch = ".")
lines(wyk_sh$x, wyk_sh$yf, type = "S", col = "red")
title("Wykres Sheparda")
```



Rysunek 20: Wykres Sheparda

- Wykres ten przedstawia odtworzone odległości wykreślone na osi pionowej względem pierwotnych niepodobieństw wykreślonych na osi poziomej
- Pokazuje także funkcję krokową. Linia ta przedstawia wartości odległości, które są rzutami oryginalnych składowych (wynik transformacji monotonicznej danych wejściowych)
- Jeśli wszystkie odtworzone odległości znajdowałyby się na lini krokowej, to porządek rangowy odległości (lub niepodobieństw) byłby dokładnie odtworzony
- Odchylenia od lini krokowej wskazują na brak dopasowania
- UWAGA jak uwzględnimy zmienna gatunki podczas analiz czyli zamiast usuniete[-5] damy usuniete to otrzymamy na wykresach lepszą separację grup - obiekty z tych samych grup będą blisko siebie
- Dla usuniete[-5] mamy dobrą duże lepsze dopasowanie krzywej krokowej (czerwony) a punktów dla usuniete punkty są w mniejszym stopniu skupione wokół krzywej krokowej (czerwony)

```
iris3 <- cbind(dane_iris, wynik_mds)

ggplot(iris3, aes(wynik_mds$points[,1], wynik_mds$points[,2], col = Species, fill = Spec
geom_point(shape = 21, col = "black")+
stat_ellipse(geom = "polygon",alpha = 0.5, col = "black")
```



Rysunek 21: jakis tytul

- Dobra separacja klasy setosa - jest wyraźnie oddalona od innych obserwacji
- Nieco mniejsza separacja jest dla klas versicolor i virignica

Literatura

[1] dr inż. Adam Zagdański, http://prac.im.pwr.wroc.pl/~zagdan/polish_ver/ED2020/index.html, 2020.