Raport 2

SD

23 maja 2023

Spis treści

1 Analiza składowych głównych

```
dane <- read.csv(file="uaScoresDataFrame.csv", stringsAsFactors = TRUE)

dane <- dane[2:21]#Usuwamy kolumnę z identyfikatorem kraju.

dim(dane) #liczba przypadków/cech.
```

[1] 266 20

```
sum(sapply(dane, is.factor)) #ile cech jakościowych.
```

[1] 3

```
sum(sapply(dane, is.numeric)) #ile cech ilościowych.
```

[1] 17

```
sum(sapply(dane, is.nan)) #Sprawdzamy czy jakaś kolumna jest pusta.
```

[1] 0

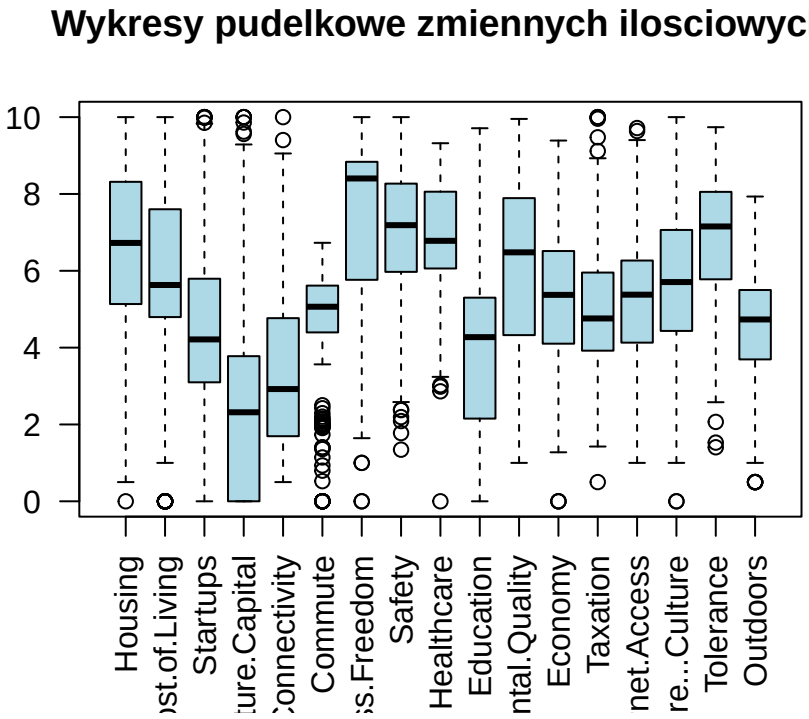
```
names(which(sapply(dane, is.numeric) == TRUE)) #Poszczgólne cechy ilościowe.
```

[1] "Housing" "Cost.of.Living" "Startups" [4] "Venture.Capital" "Travel.Connectivity" "Commute" [7] "Business.Freedom" "Safety" "Healthcare" [10] "Education" "Environmental.Quality" "Economy" [13] "Taxation" "Internet.Access" "Leisure...Culture" [16] "Tolerance" "Outdoors"

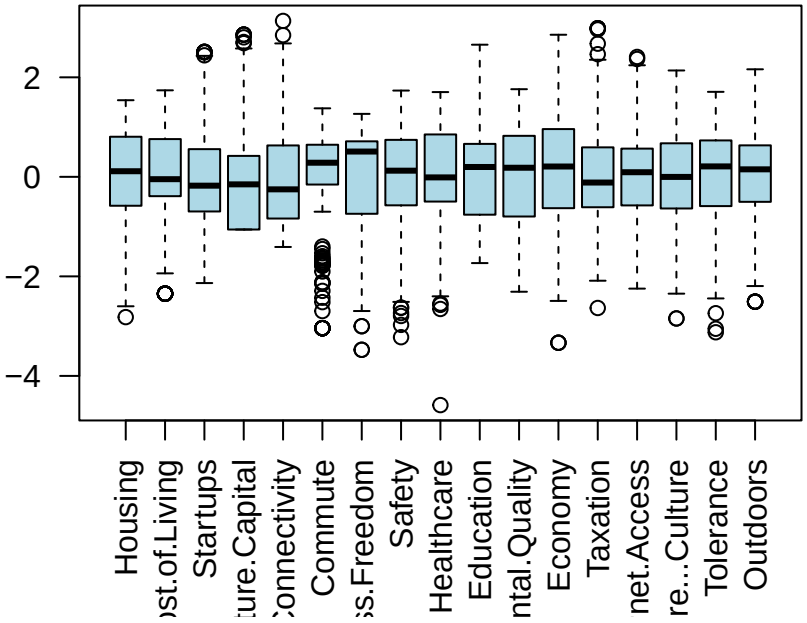
```
#Pozostałe dane mają typ jakościowy.

dane_ilosciowe <- dane[, 4:20]
dim(dane_ilosciowe)
```

[1] 266 17

```
# Rysowanie wykresów boxplot dla wszystkich kolumn jednocześnie
```



Zmienne ilosciowe po standaryzacji



Wykres pudelkowy skladowych głównych.

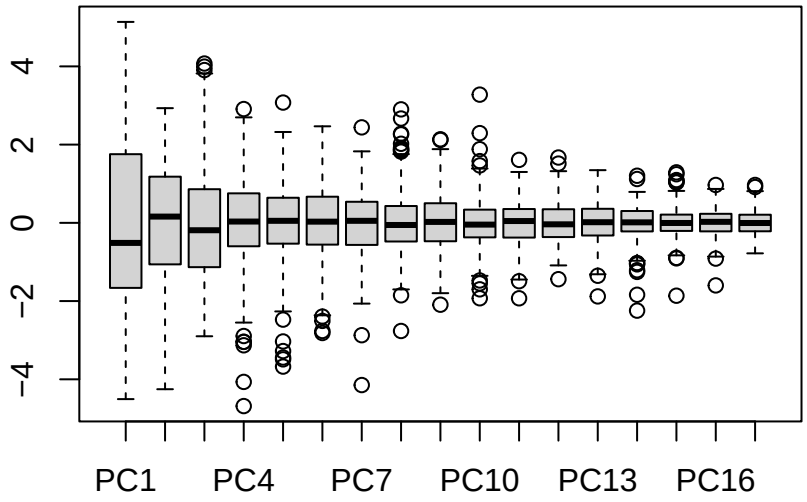
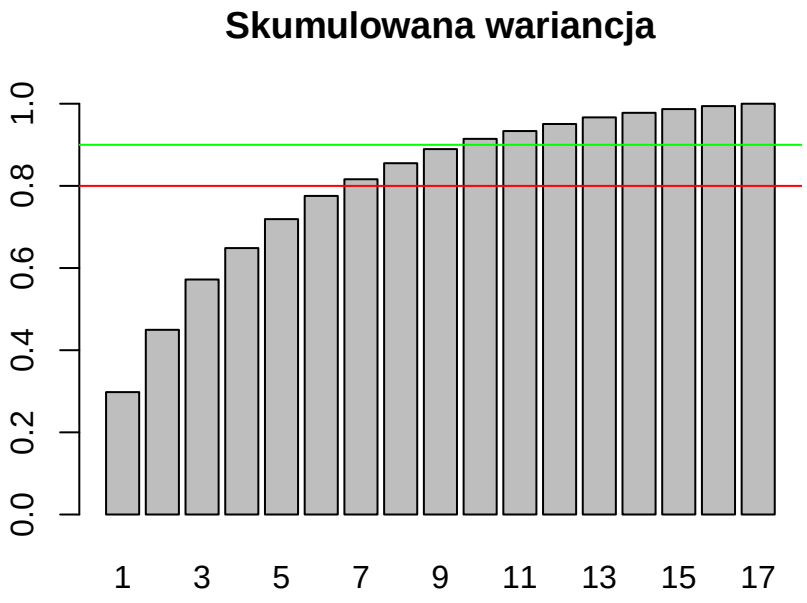


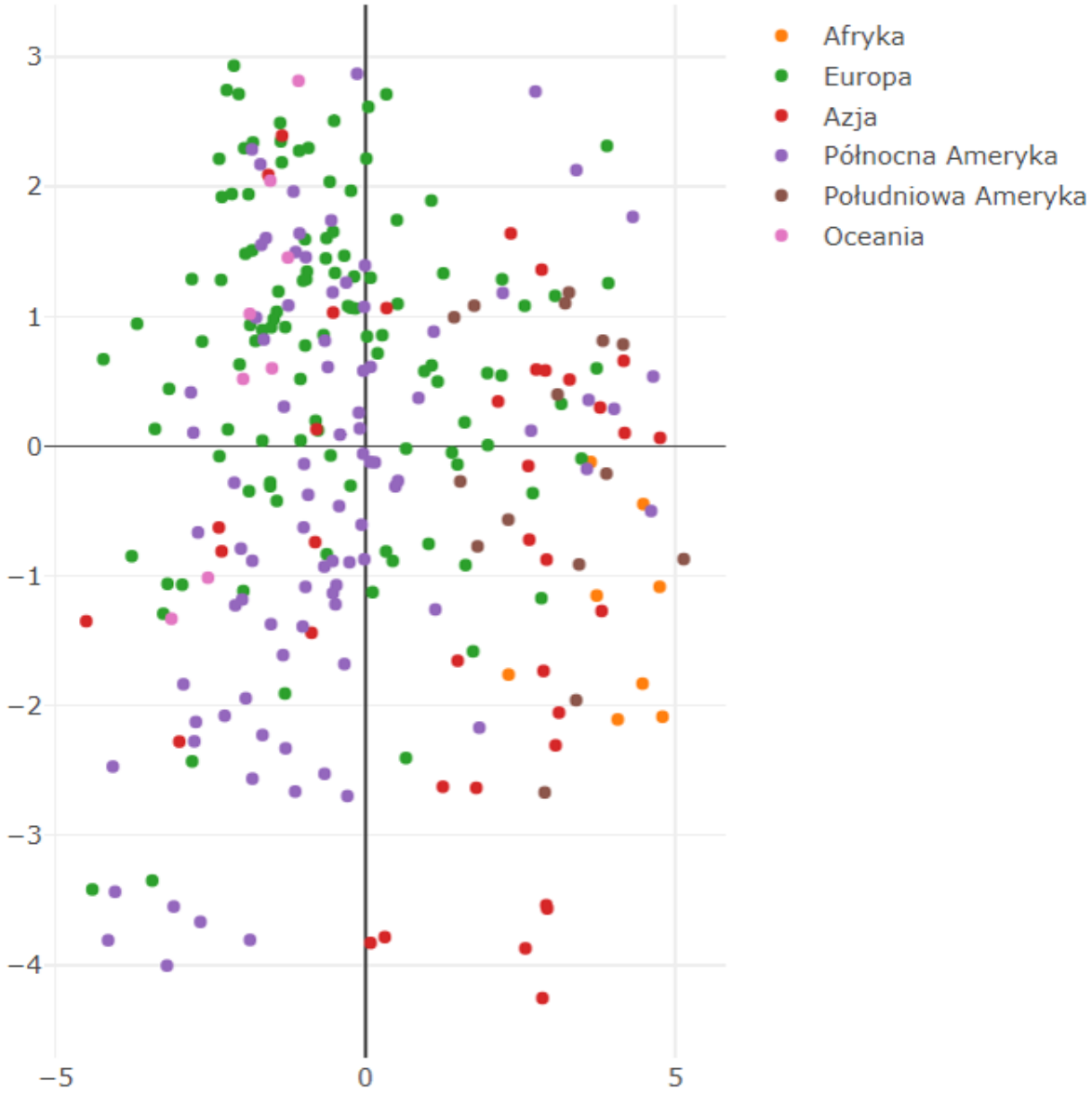
Tabela 1: Wektory ładunków dla PC1, PC2, i PC3.

	PC1	PC2	PC3
Housing	0.31	0.05	-0.31
Cost.of.Living	0.26	-0.18	-0.33
Startups	-0.18	-0.48	0.01
Venture.Capital	-0.24	-0.43	0.01
Travel.Connectivity	-0.21	-0.14	-0.34
Commute	-0.11	0.03	-0.51
Business.Freedom	-0.38	0.10	0.02
Safety	-0.04	0.29	-0.33
Healthcare	-0.28	0.24	-0.28
Education	-0.40	-0.05	-0.07
Environmental.Quality	-0.33	0.25	0.05
Economy	-0.27	-0.07	0.31
Taxation	0.03	0.11	-0.02
Internet.Access	-0.28	0.02	0.03
Leisure...Culture	-0.07	-0.36	-0.31
Tolerance	-0.19	0.36	-0.10
Outdoors	-0.09	-0.19	-0.15



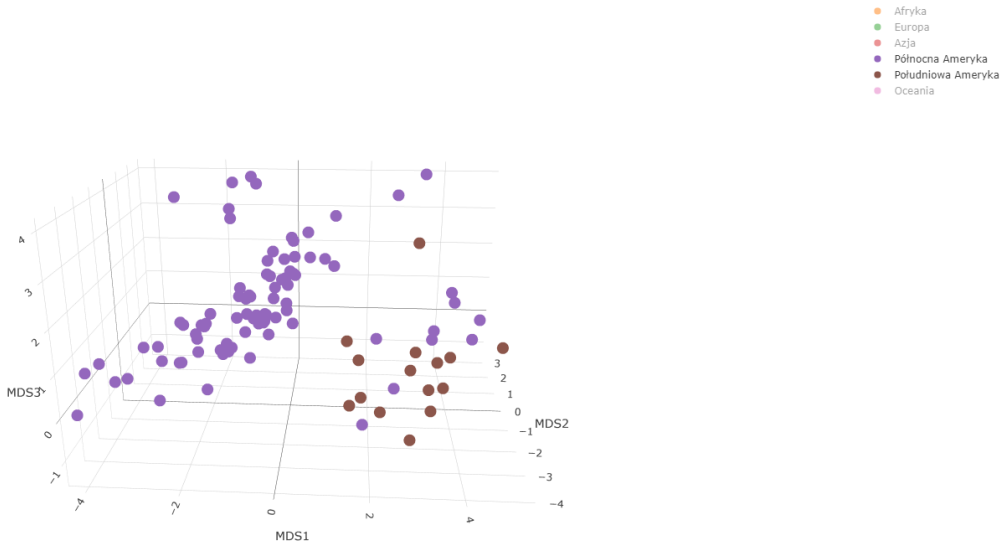
Potrzeba 7 składowych głównych aby wyjasnić 80% przypadków, oraz 10 skladowych aby wyjaśnić 90% przypakdów.

1.1 Wykresy rozrzutu PCA



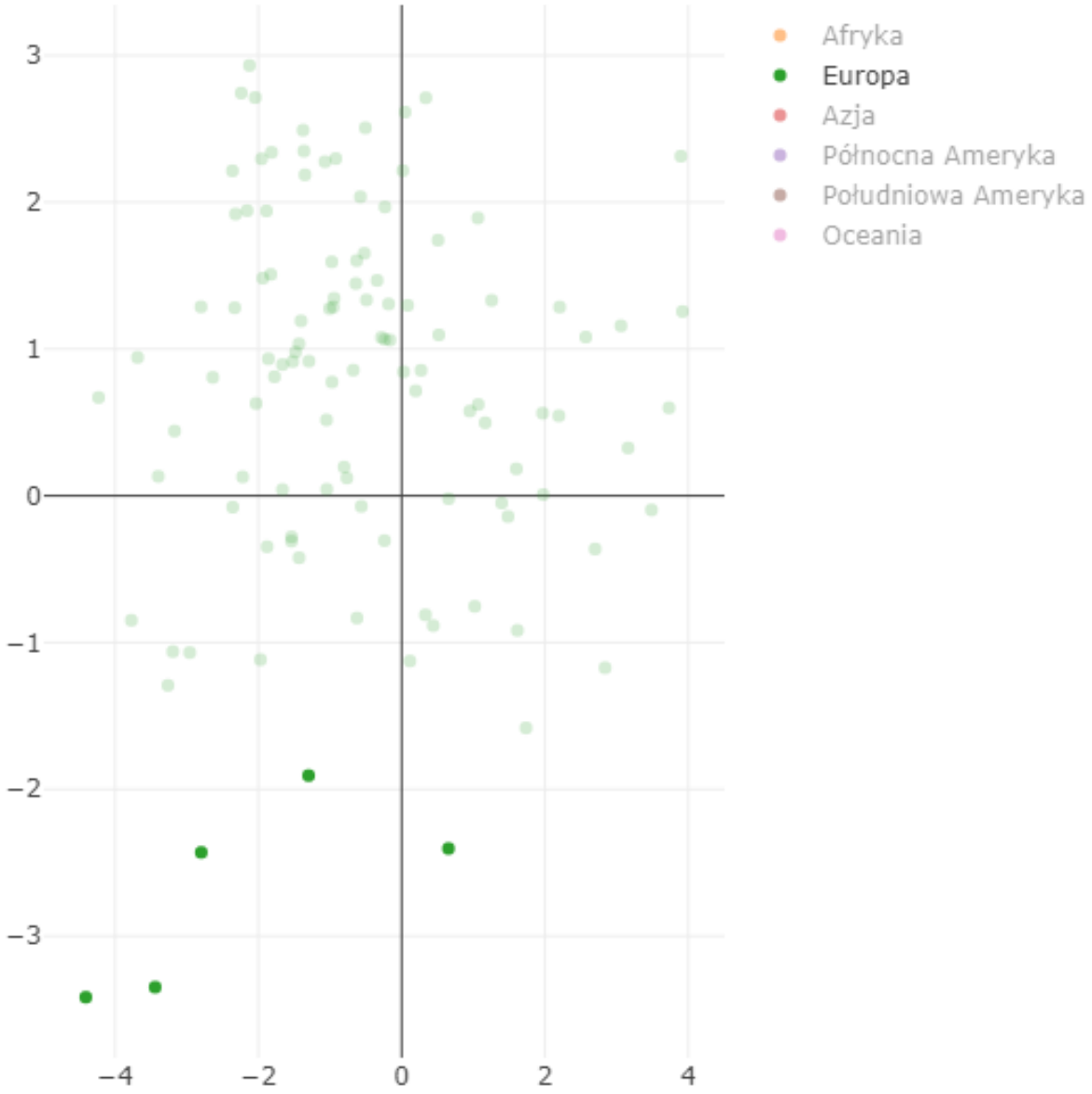
Rysunek 1: Wykres rozrzutu uzyskany za pomocą PCA

Obserwując wykres 2D wykorzystujący dwie składowe główne lub 3D wykorzystujący 3 składowe można dojść do wniosku, że miasta układają się w grupy zarówno względem kontynentu jak i państwa w którym się znajdują. Przypadki odstające opisze poniżej:



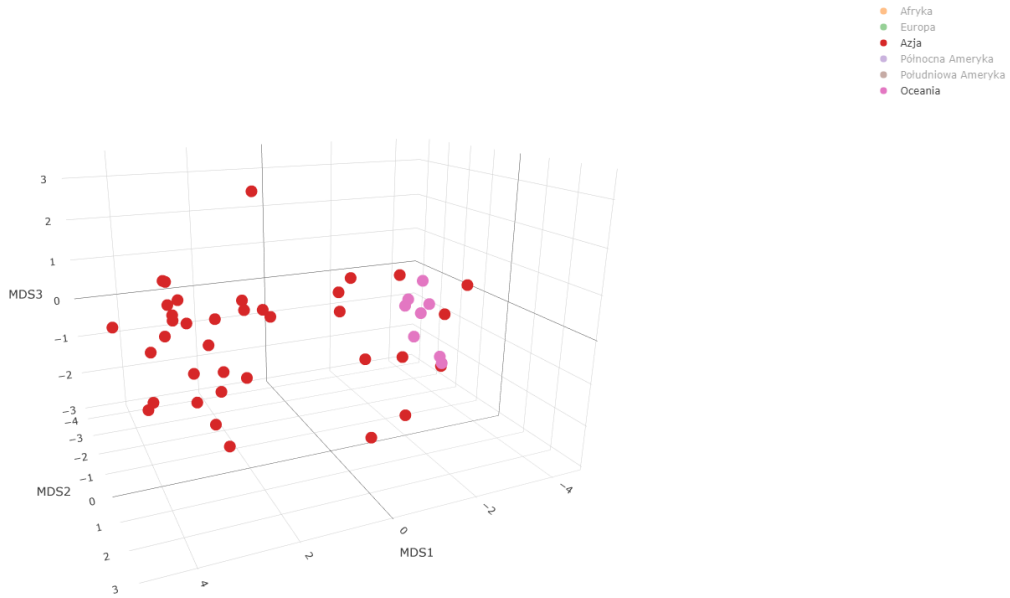
Rysunek 2: Wykres rozrzutu uzyskany za pomocą PCA

San Jose, Guatemala City, Havana, San Juan, San Salvador, Kingston, Santo Domingo, Guadalajara, Mexico City to obiekty odstające od grupy Północna Ameryka, jednocześnie obserwacje te znajdują się bardzo blisko grupy punktów z Południowej Ameryki, zbliżony rozrzut najpewniej jest spowodowany tym, że miasta te znajdują się w Ameryce Środkowej.



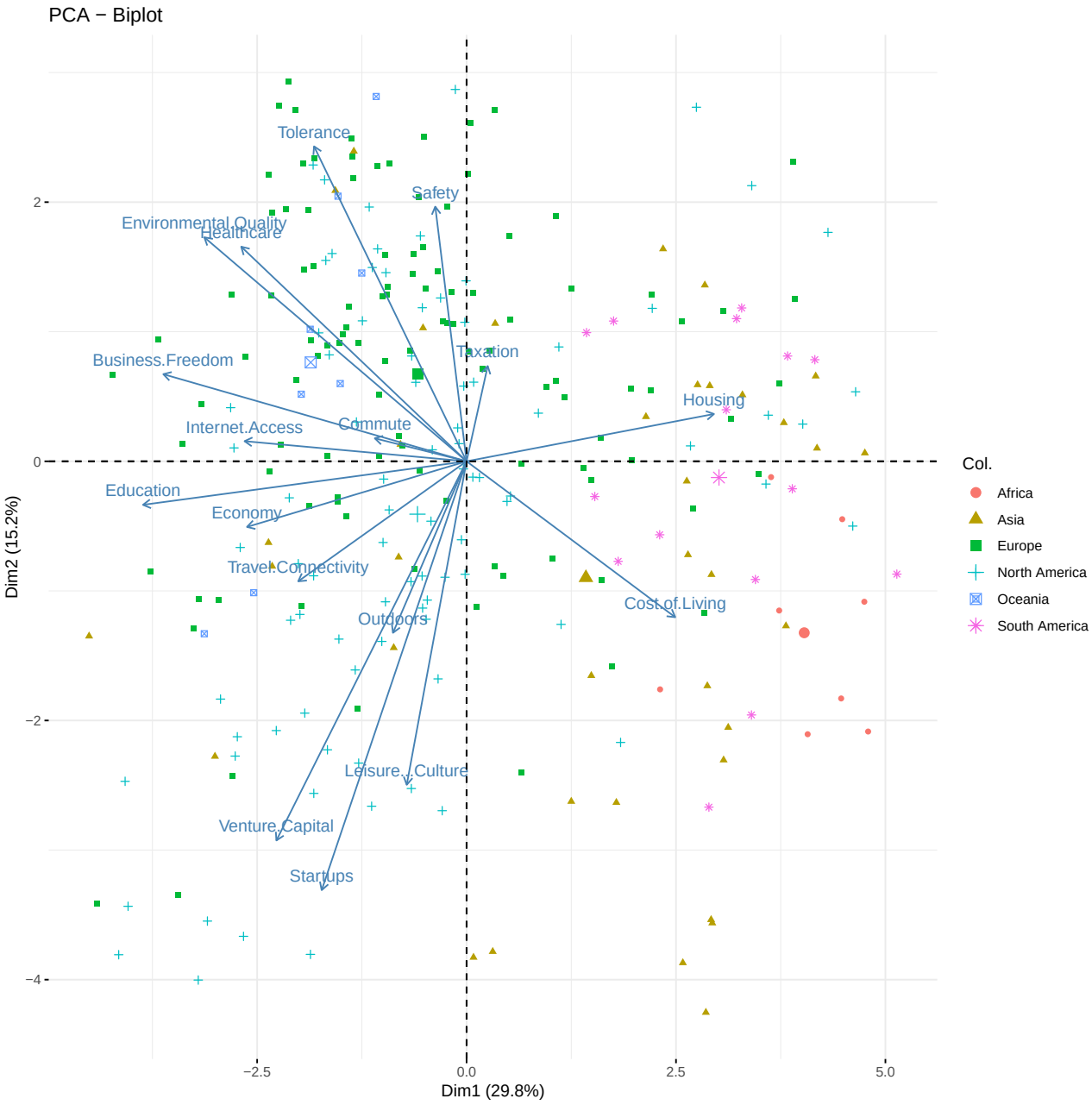
Rysunek 3: Wykres rozrzutu uzyskany za pomocą PCA

Obserwując rozrzut miast Europy w dla dwóch zmiennych składowych można zaobserwować odstające miasta takie jak: Londyn, Berlin, Paryż, Moskwa ich cecha wspólną jest fakt, że są to stolice. Na wykresie 3d mozna również zaobserować że miasta takie jak Giblartar, Andora, Valletta też odstają, cechą charakterystyczną tych miast jest to, że znajdują się w najmniejszych państwach Europy.

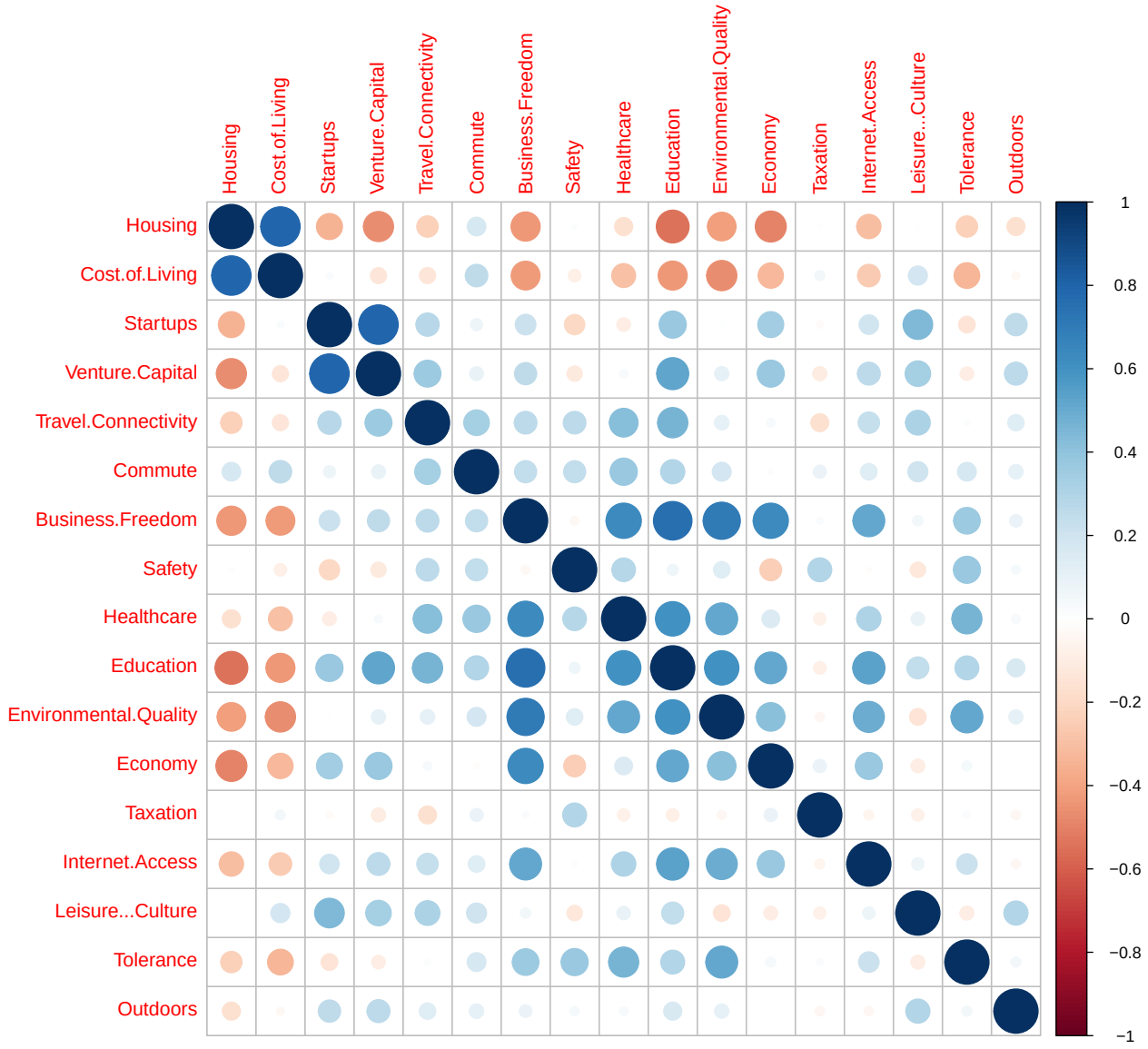


Rysunek 4: Wykres rozrzutu uzyskany za pomocą PCA

Hong Kong, Singapur, Tokyo, Seoul, Fukuoka, Kyoto, Osaka, Tajpej to miasta z Azji najbardziej odstające. Jednocześnie zbliżone są one do punktów oznaczających miasta Oceanii. Miasta te mają wiele podobieństw: są położone na wyspach lub na wybrzeżu, pod względem rzeczywistej odległości znajdują się blisko Oceanii co może tłumaczyć ich zbliżone wartości na wykresie. Odstającą wartością jest również Tashkent znajdujący się w Uzbekistanie. Można również zauważyć, że obserwacje miast które znajdują się w jednym państwie jak Japonia, Indie lub Chiny znajdują się również blisko siebie, więc tu również można zaobserować, że miasta grupują się względem kraju w którym się znajdują.



Obserwując biplot możemy dostrzec, że występuje bardzo duża dodatnia korelacja pomiędzy Environmental.Quality, a Healthcare natomiast te dwie zmienne są ujemnie skorelowane z Cost.of.Living. Kolejną dodatnią korelacje możemy zaobserować pomiędzy Venture capital i startups, można również zaobserować ujemną korelację tych dówch zmiennych z Taxation. Ujemna korelacja widocznia jest również między Economy a Housing.



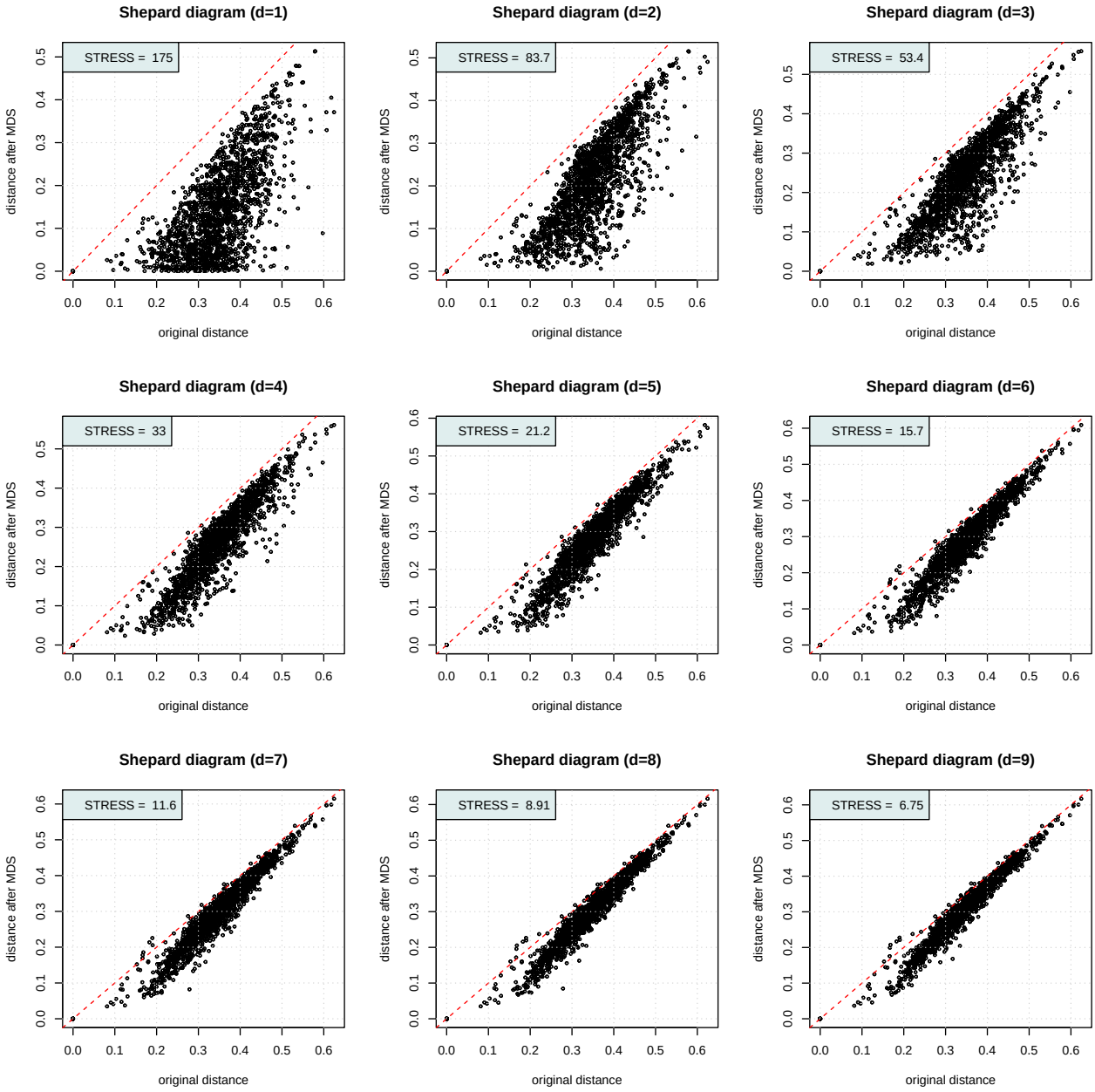
Porównując do macierzy korelacji wnioski które się powtrzają to: Venture.Capital dodatnio skorelowane z Startups, Housing ujemna korelacja z Economy.

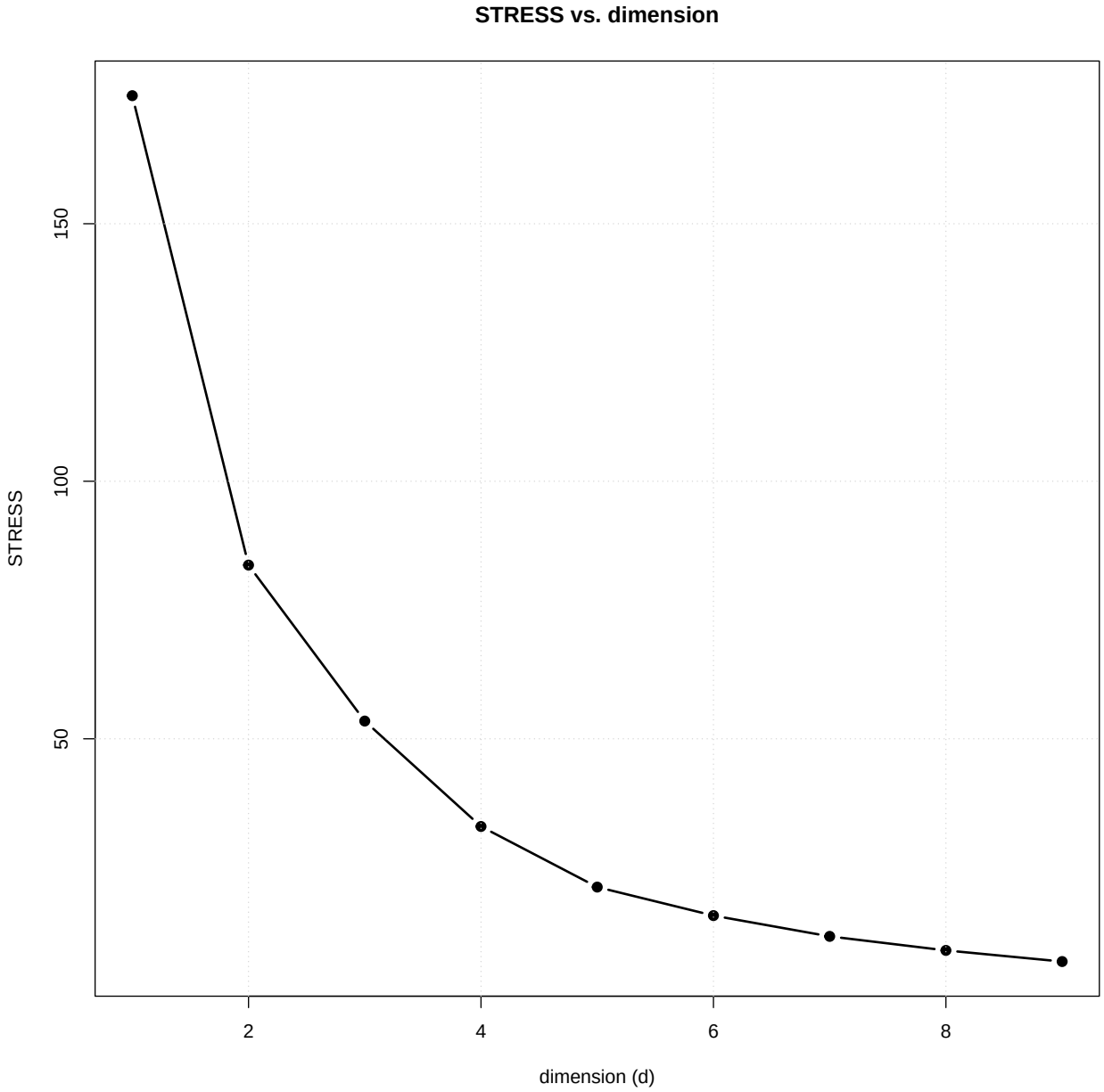
1.2 Podsumowanie PCA

Opisane analizy dotyczą wykorzystania dwóch metod wizualizacji wielowymiarowych danych - biplotów i wykresów 2D/3D. W przypadku wykresów dotyczących miast na różnych kontynentach, zauważono, że miasta grupują się względem kontynentu i państwa, a obserwacje odstające, takie jak stolice lub miasta w małych państwach, mogą być wyjaśnione tą przynależnością. W tym przypadku zastosowanie standaryzacji nie miało wpływu na otrzymane wyniki i wnioski, ponieważ wszystkie zmienne w danych były wyrażone w tej samej skali. Podsumowując, wykorzystanie biplotów i wizualizacji wielowymiarowych danych może pomóc w zrozumieniu wzajemnych relacji między zmiennymi oraz identyfikacji obserwacji odstających i grupowania się danych względem różnych kategorii.

2 MDS

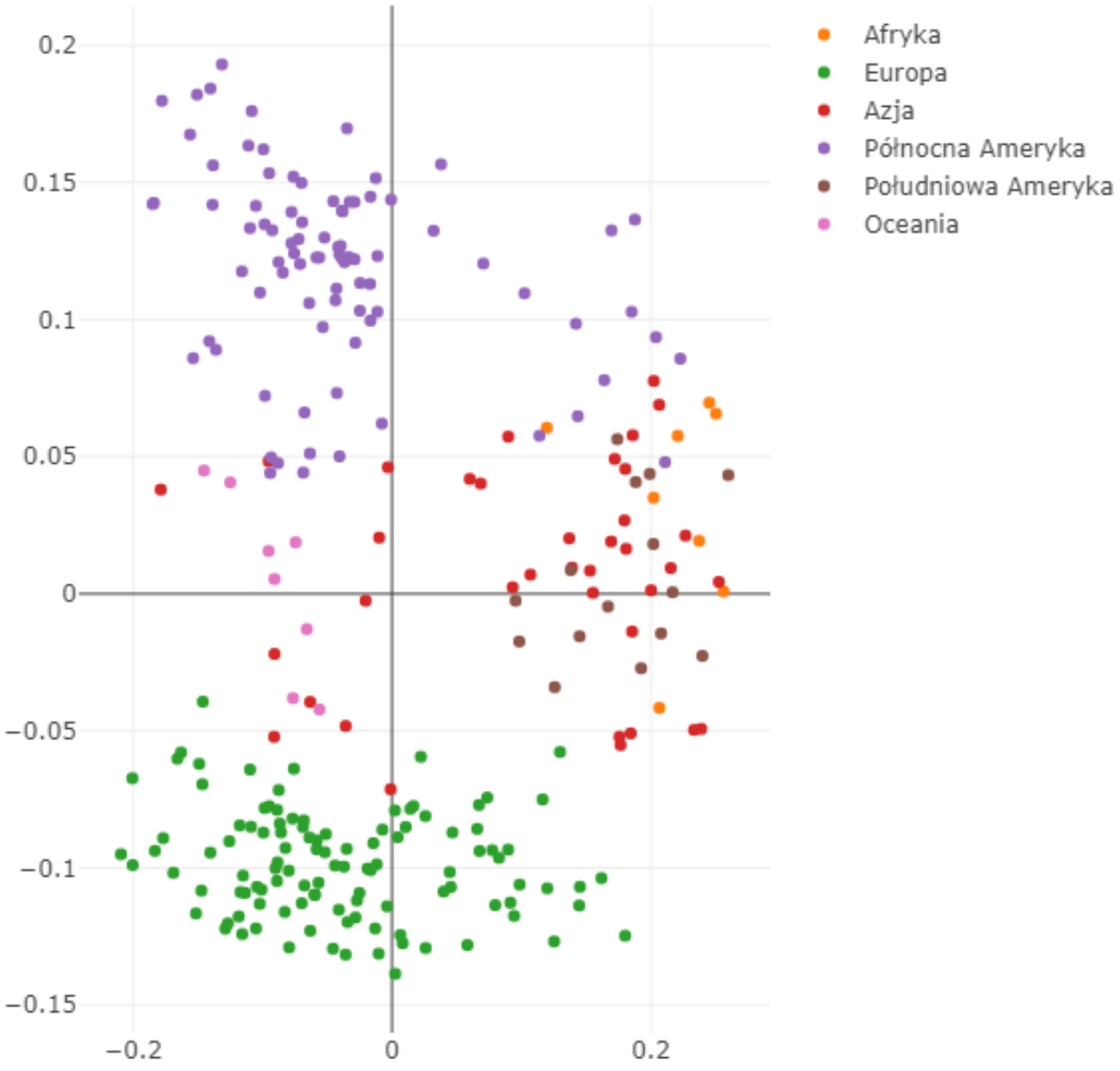
2.1 Diagramy Sheparda dla wymiarów 1-9



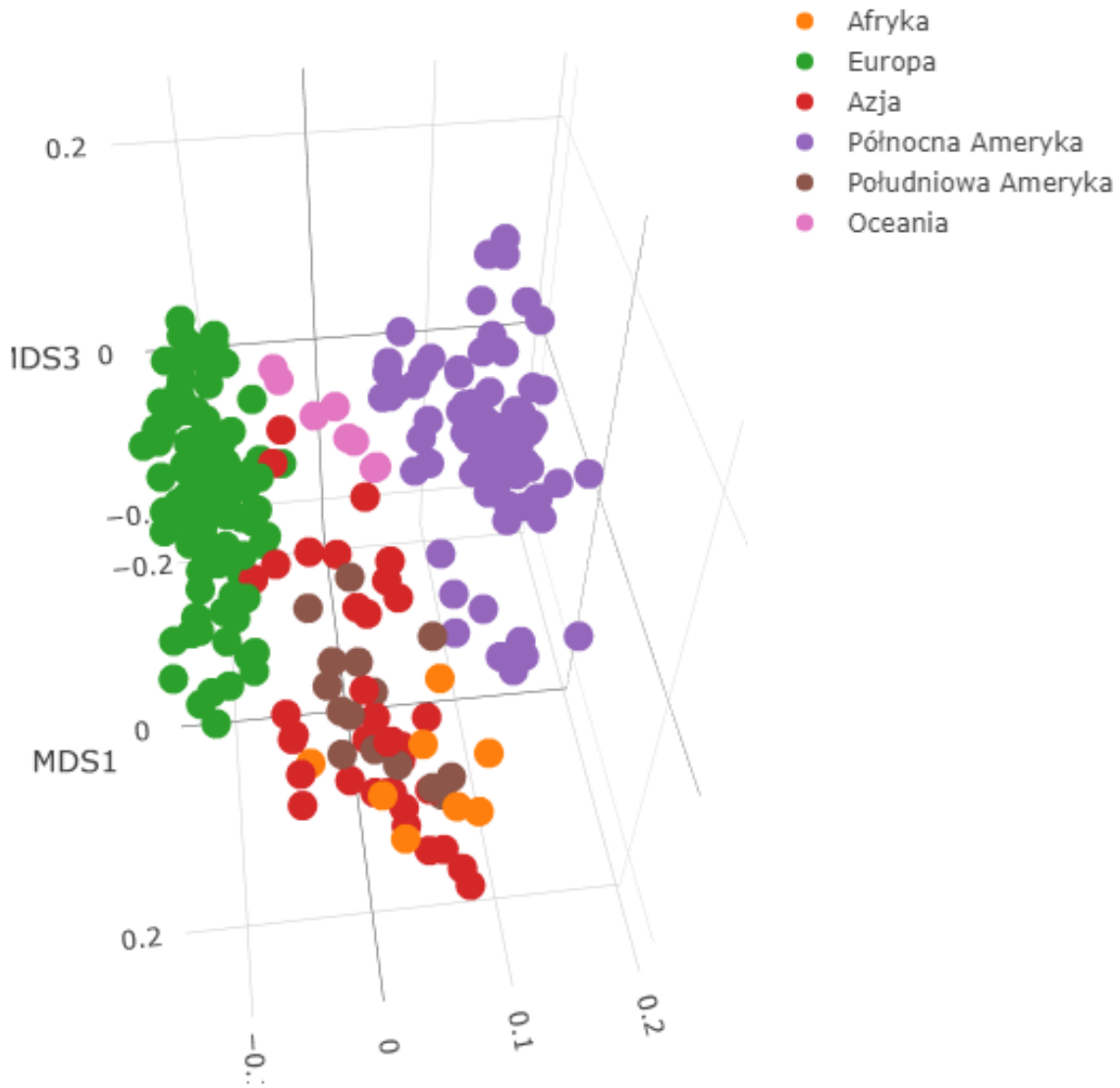


Interpretując diagramy Sheparda oraz kryterium STRESS wymiar d=3 wydaje się być optymalny.

2.2 Wykresy rozrzutu MDS

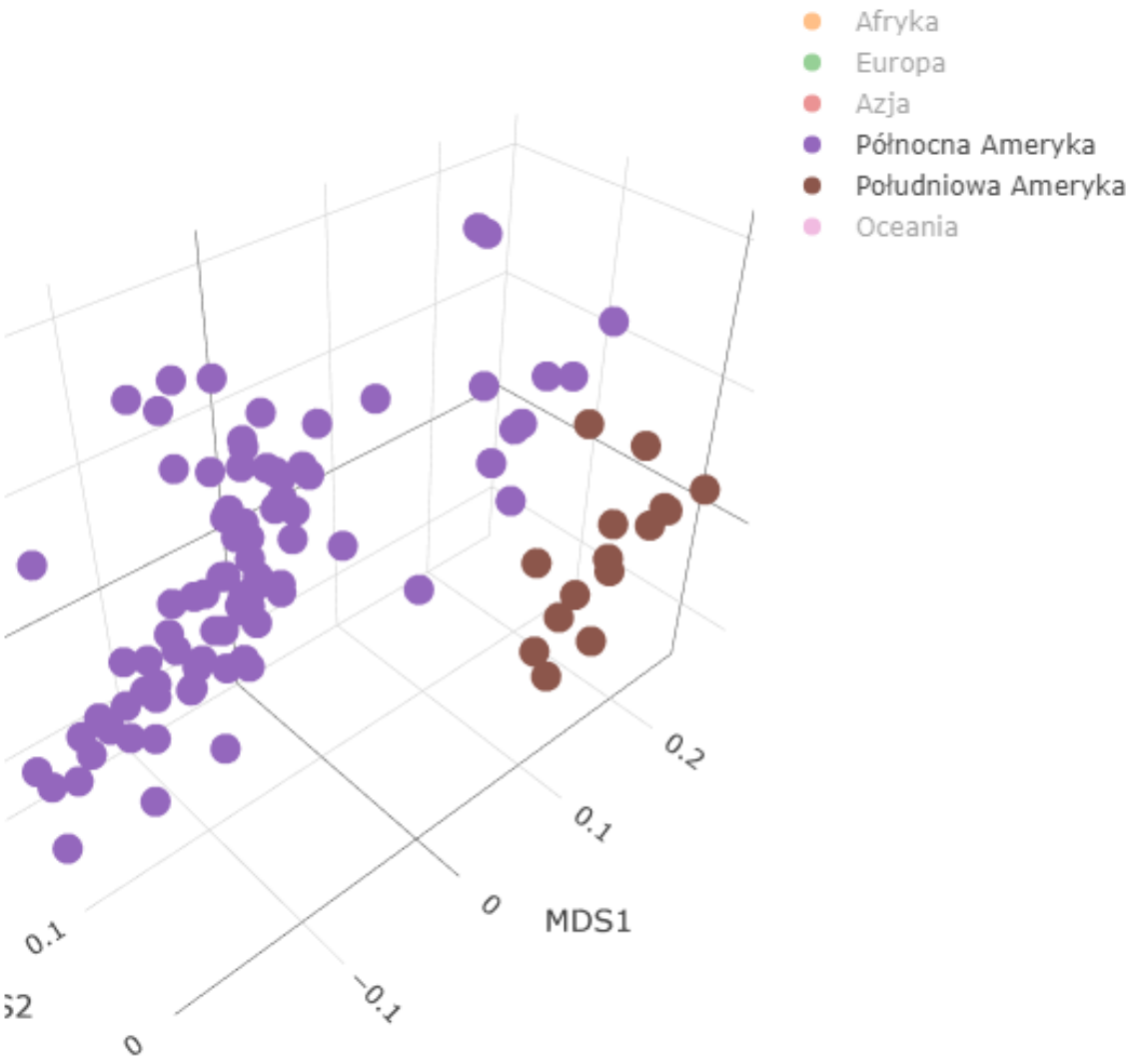


Rysunek 5: Wykres rozrzutu uzyskany za pomocą MDS

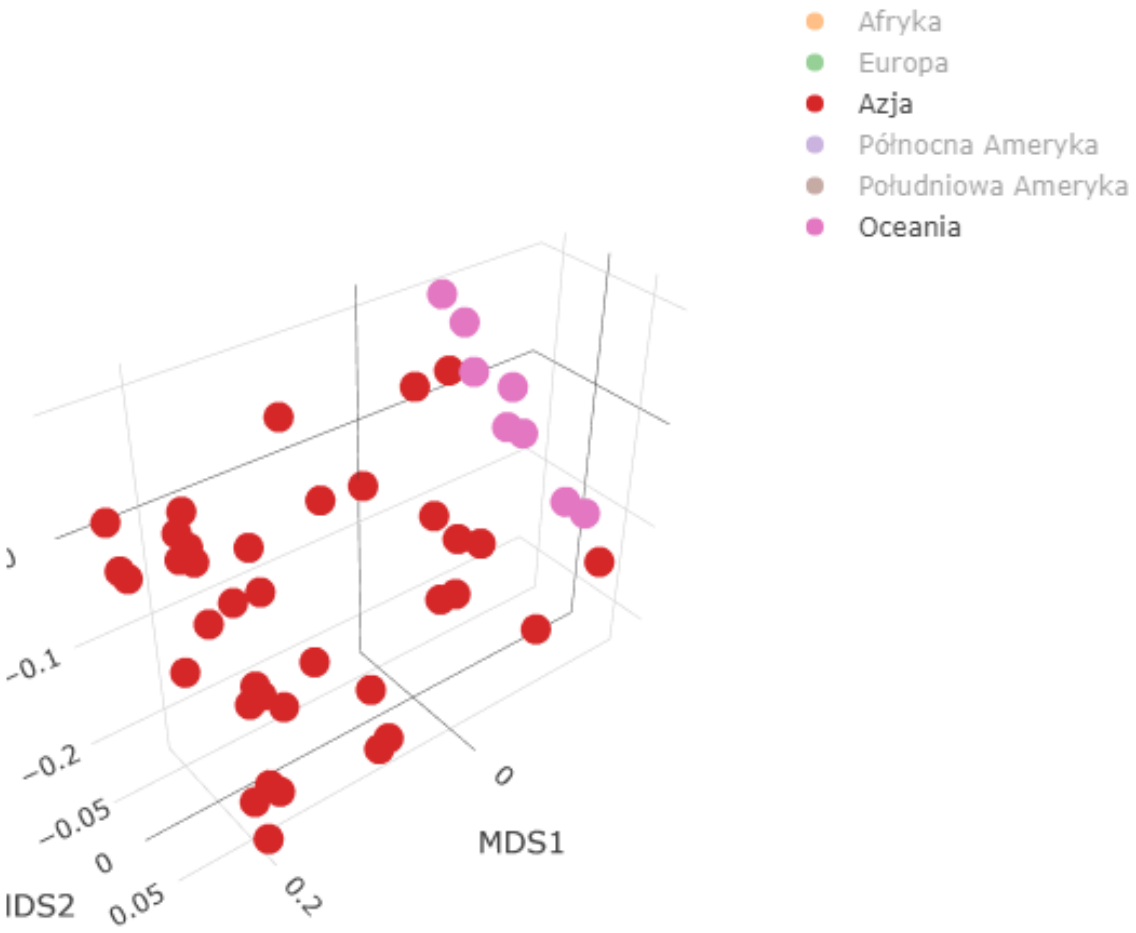


Rysunek 6: Wykres rozrzutu uzyskany za pomocą MDS

Podobnie jak w metodzie PCA widać, że miasta układają się w grupy zarówno względem kontynentu jak i państwa w którym się znajdują. Jednak w porównaniu do pca podział na grupy względem kontynentu czy państwa zdaje się być wyraźniejszy. Obserwacje odstające są takie same jak w przypadku PCA jednak nie odstają już tak wyraźnie jak w przypadku PCA.



Rysunek 7: Wykres rozrzutu uzyskany za pomocą MDS



Rysunek 8: Wykres rozrzutu uzyskany za pomocą MDS

3 Podsumowanie

Można zauważyć podobieństwo między metodą PCA a badaniami grupowymi miast, ponieważ obie metody pozwalają na zidentyfikowanie wzorców w danych i grupowanie ich na podstawie tych wzorców. Jednak w przypadku grupowania miast, podział na grupy względem kontynentów i państw wydaje się być bardziej wyraźny w metodzie MDS. Obserwacje odstające są podobne w obu metodach, ale w grupowaniu miast nie są one tak wyraźnie widoczne jak w PCA.