

Sprawozdanie 2

Eksploracja danych

Kacper Szmigielski, 282255 i Mateusz Wizner, 277508

2025-04-30

Spis treści

1 ZADANIE 1 (Dyskretyzacja(przedziałowanie) cech ciągłych)	2
1.1 a) Dane: iris (R-pakiet datasets).	2
1.2 b) Wybór cech	2
1.3 c) Porównanie nienadzorowanych metod dyskretyzacji	4
1.3.1 Metoda : Równe częstotliwości(Frequency)	4
1.3.2 Metoda : Równe szerokości (Interval)	7
1.3.3 Metoda : k najbliższych sąsiadów (K-means)	9
1.3.4 Dyskretyzacja z przedziałami zadanymi przez użytkownika (fixed) .	11
1.4 Wnioski :	14
2 ZADANIE 2 (Analizaskładowych głównych (Principal Component Analysis (PCA)))	15
2.1 a) Przygotowanie i opis danych	15
2.2 b) Wyznaczenie składowych głównych	22
2.3 c) Zmienna odpowiadająca poszczególnym składowym	26
2.4 d) Wizualizacja danych wielowymiarowych	28
2.5 e) Korelacja zmiennych	32
2.5.1 Wnioski z biplotu:	32
2.6 f) Końcowe wnioski	33
2.6.1 Wnioski ogólne:	34

3 ZADANIE 3 (Skalowanie wielowymiarowe (Multidimensional Scaling (MDS)))	34
3.1 a) Dane: titanic_train (R-pakiet titanic)	34
3.2 b) Przygotowanie danych	34
3.3 c) Redukcja wymiaru na bazie MDS	34
3.4 d) Wizualizacja danych	36

1 ZADANIE 1 (Dyskretyzacja(przedziałowanie) cech ciągłych)

1.1 a) Dane: iris (R-pakiet datasets).

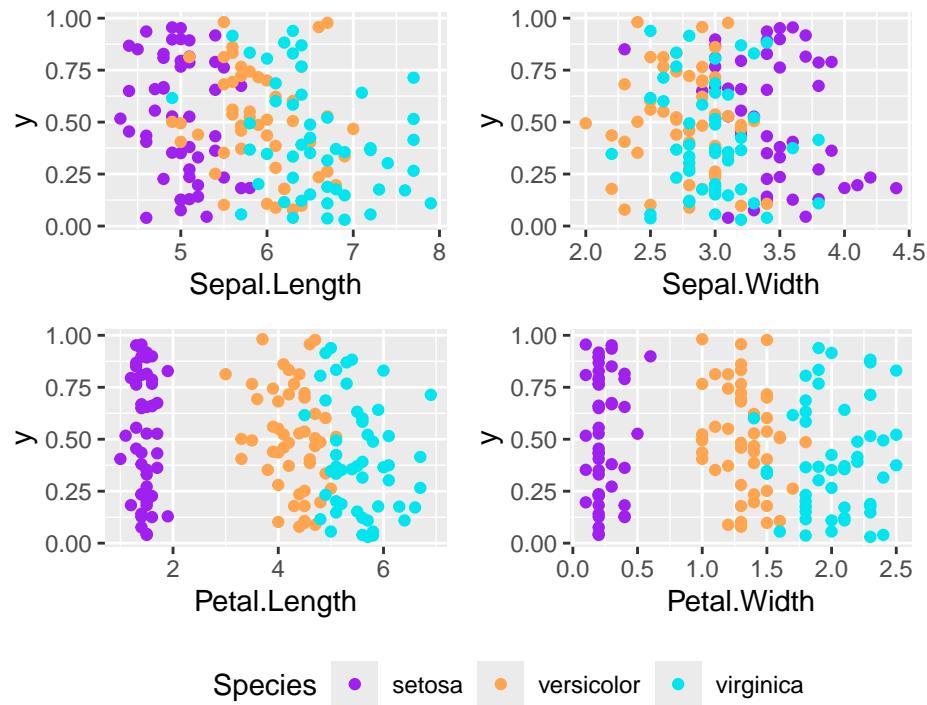
3 Pierwsze wiersze z pakietu iris

Zbiór danych zawiera wyniki pomiarów uzyskanych dla **trzech gatunków irysów** (tj. setosa, versicolor i virginica) i został **udostępniony przez Ronaldą Fishera w roku 1936**.

– Pomiary dotyczą **długości oraz szerokości** dwóch różnych części kwiatu– działki **kiełicha** (ang. sepal) oraz **płatka** (ang. petal).

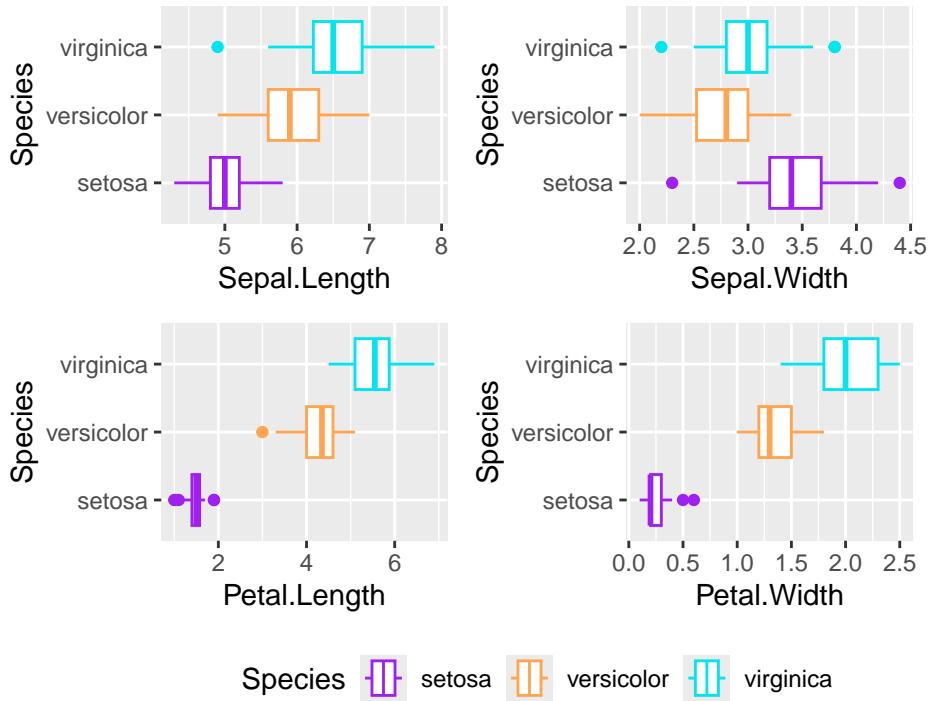
1.2 b) Wybór cech

Szukamy cech, których różnice są najbardziej spójne z różnicami pomiędzy gatunkami.



Po przeanalizowaniu scatter-plotów, widać, że podczas szukania cechy o najlepszej zdolności dyskretyzacyjnej warto zwrócić uwagę na **Petal.Length i Petal.Width**, natomiast jeśli poszukujemy kolumny o najgorszej zdolności dyskretyzacyjnej to wybór rozszerzy gamy spośród **Sepal.Length i Sepal.Width**

Musimy jednak wybrać **wartości najlepsze i najgorsze** do dyskretyzacji, aby to zrobić przeanalizujemy **box-ploty**.

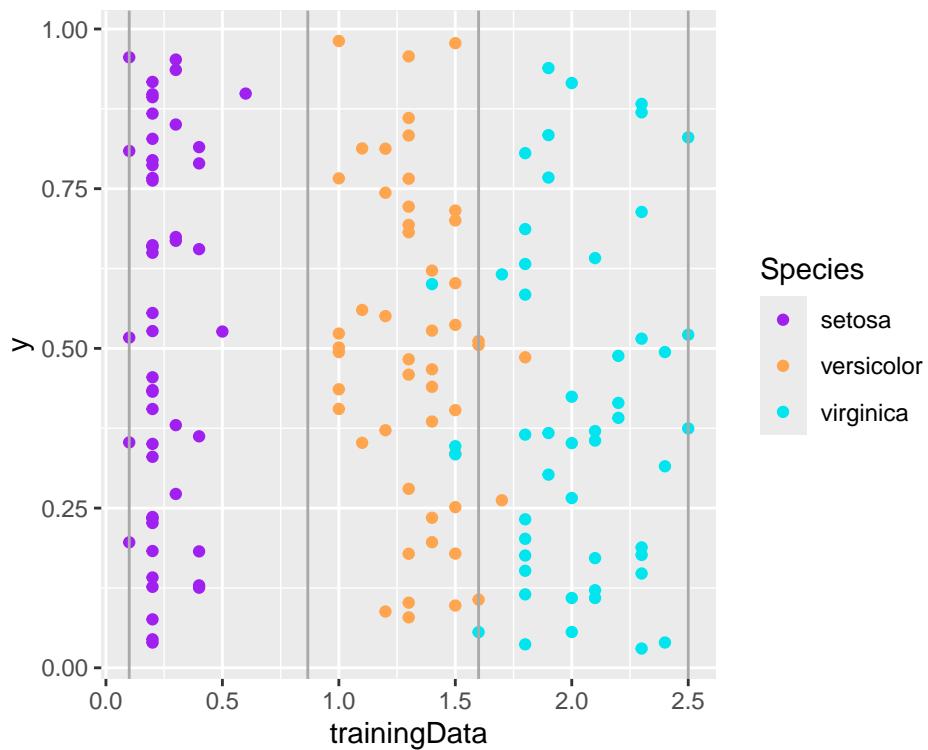


Na ich podstawie możemy uznać, że Petal.Width może stanowić najlepszy wyznacznik gatunku roślin. Najgorszym natomiast jest Sepal.Width ponieważ dla Petal.Width gatunki w najmniejszym stopniu się pokrywają ze względu na tą cechę , a w Sepal.Width w największym.

1.3 c) Porównanie nienadzorowanych metod dyskretyzacji

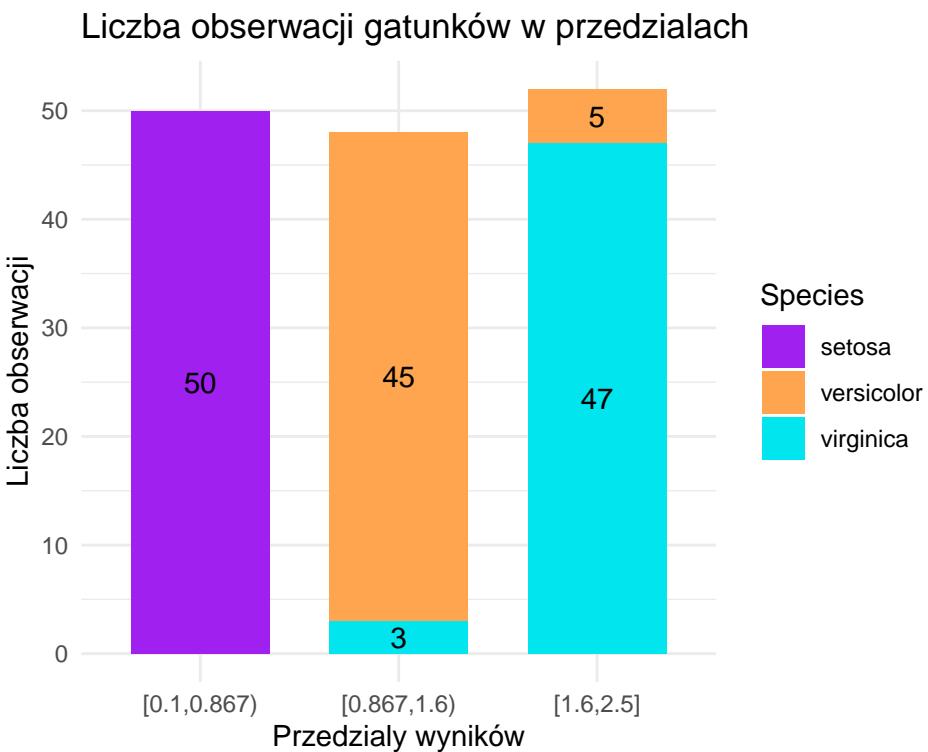
1.3.1 Metoda : Równe częstości(Frequency)

1.3.1.1 Dla najlepszej cechy : Petal.Length (Frequency)



Widać, że linie uzyskane za pomocą **Frequency** dość dobrze rozdzielają nasze dane .

Jeżeli chcemy dokładniej przeanalizować zależność podziału od gatunków, narysujemy specjalne bar-ploty

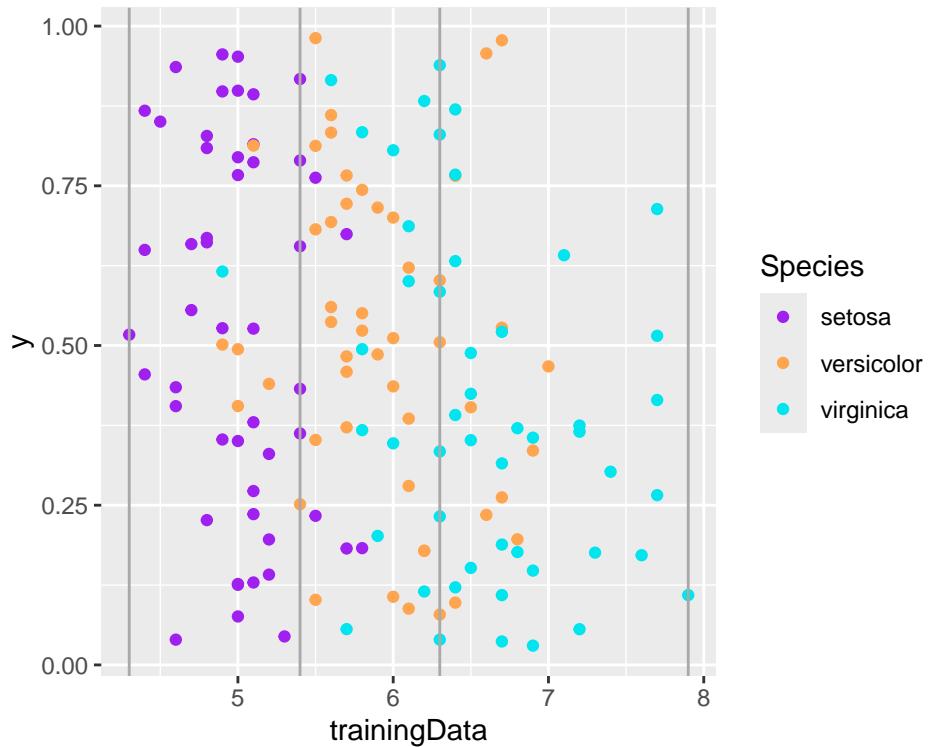


Świadczą one o tym, że metoda Frequency dla zmiennej **Petal.Width** bezproblemowo oddziela gatunek setosa, lecz wśród pozostałych występuje zjawisko mieszania się (3 virginica przyporządkowano do versicolor, a 5 versicolor do virginica)

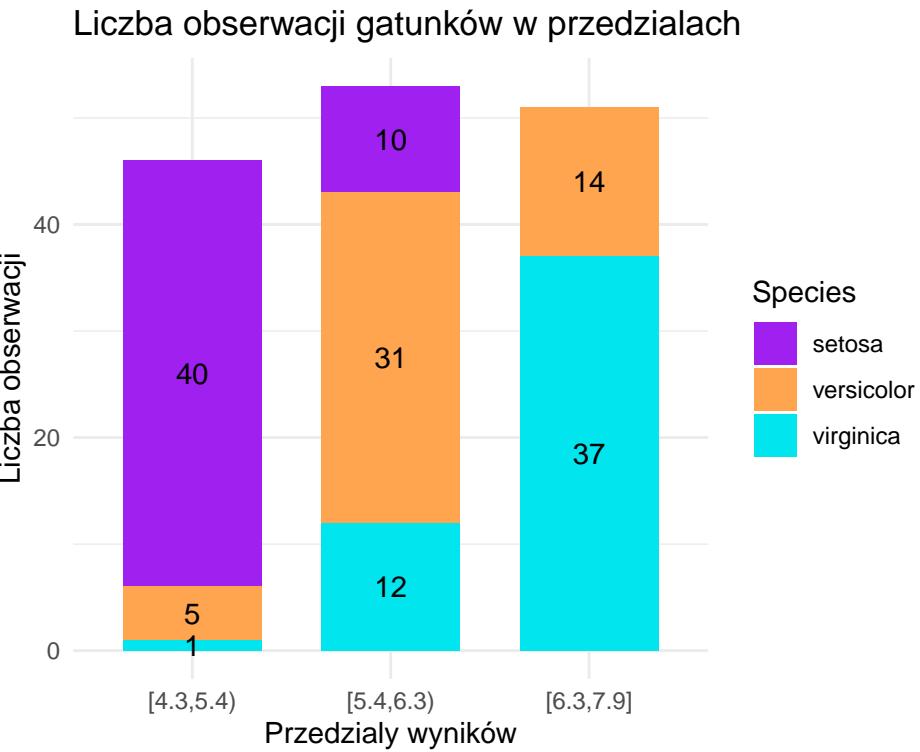
W przypadku tej metody **zgodność** uzyskanego grupowania z realnymi wartościami **wynosi** :

```
## [1] 0.9466667
```

1.3.1.2 Dla najgorszej cechy : Sepal.Length (Frequency)



Scatter-plot wskazuje, że dla Sepal.Length grupowanie może być dość problematyczne, widać, że obserwacje są dość wymieszane, i trudno będzie w prosty sposób oddzielić je tak, aby gatunki były prawidłowo rozłożozone, te sam problem pojawia się w pozostałych metodach grupowań, dlatego scatter-plot Sepal.Length analizujemy tylko tutaj.



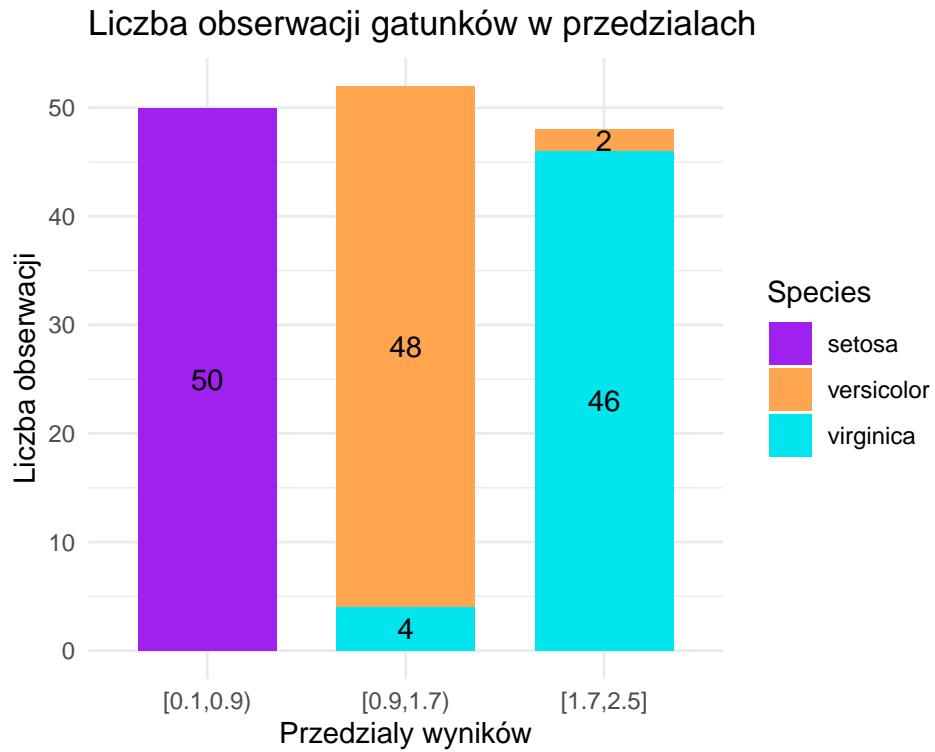
Na tabeli przyporządkowań widać, problemy metody Frequency, przy grupowaniu dla zmiennej Sepal.Length, gatunki są dość mocno przemieszane, brakuje jednolitego podziału.

```
## [1] 0.72
```

Zgodność dla nagjroszej cechy wynosi jedynie ok **72%**, co mówi o znacznym spadku wiarygodności (**o ok 23 %**) w porównaniu do Petal.Width

1.3.2 Metoda : Równe szerokości (Interval)

1.3.2.1 Dla najlepszej cechy : Petal.Width (Interval)



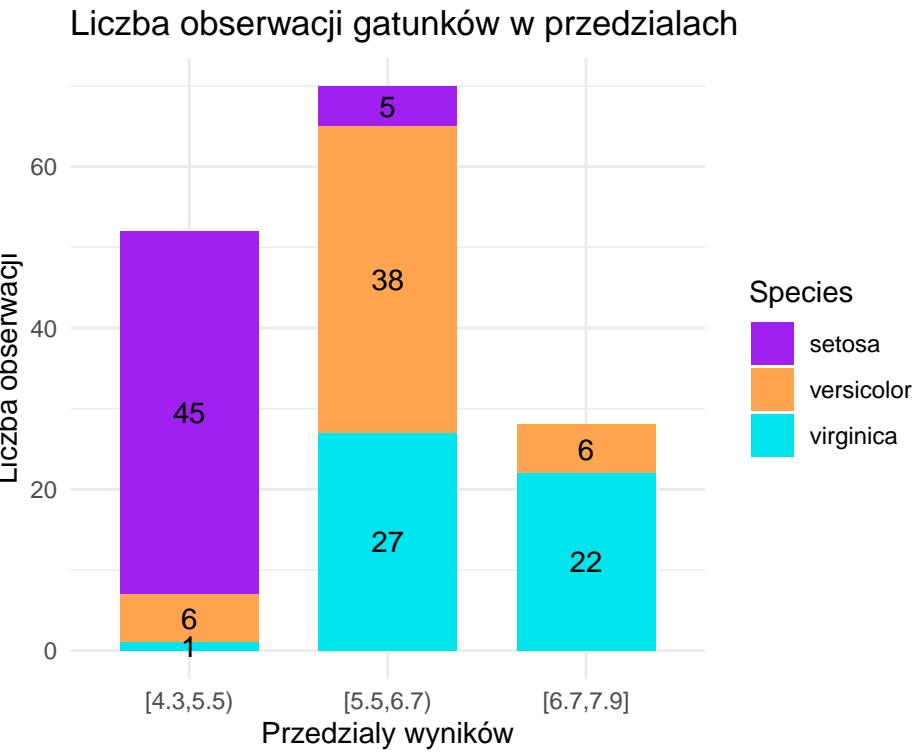
Po tabeli przyporządkowań widać, że mamy trochę lepsze odróżnienie versicolor od virginica

Dla tej metody również mamy **zgodność na poziomie** :

```
## [1] 0.96
```

Widac lekki wzrost zgodności w porównaniu do poprzedniej metody (**o ok 1%**)

1.3.2.2 Dla najgorszej cechy ; Sepal.Length (Interval)



Dla tabeli zgodności widać, że metoda w zły sposób rozdziela przypadki. Bardzo duża ich ilość znajduje się w środkowym przedziale, więc nie jest to dobry podział gatunkowy

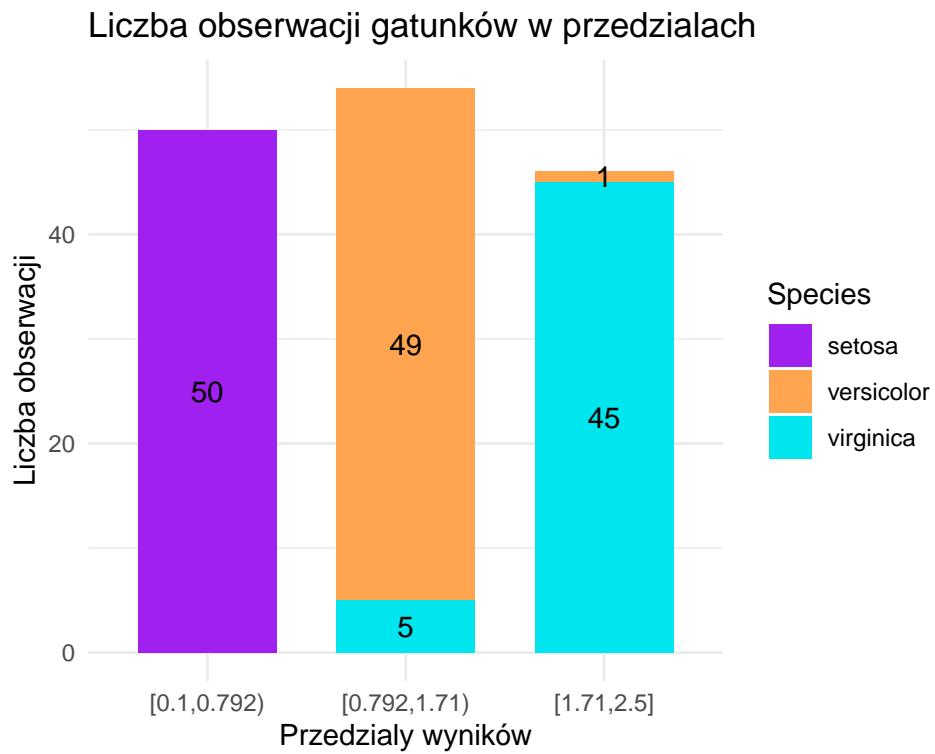
Metoda ta, dla najgorszej cechy dyskretyzuje ze zgodnością :

```
## [1] 0.5729167
```

Czyli w porównaniu do metody Frequency mamy **spadek aż o ok 16%**

1.3.3 Metoda : k najbliższych sąsiadów (K-means)

1.3.3.1 Dla najlepszej cechy : Petal.Width (K-means)



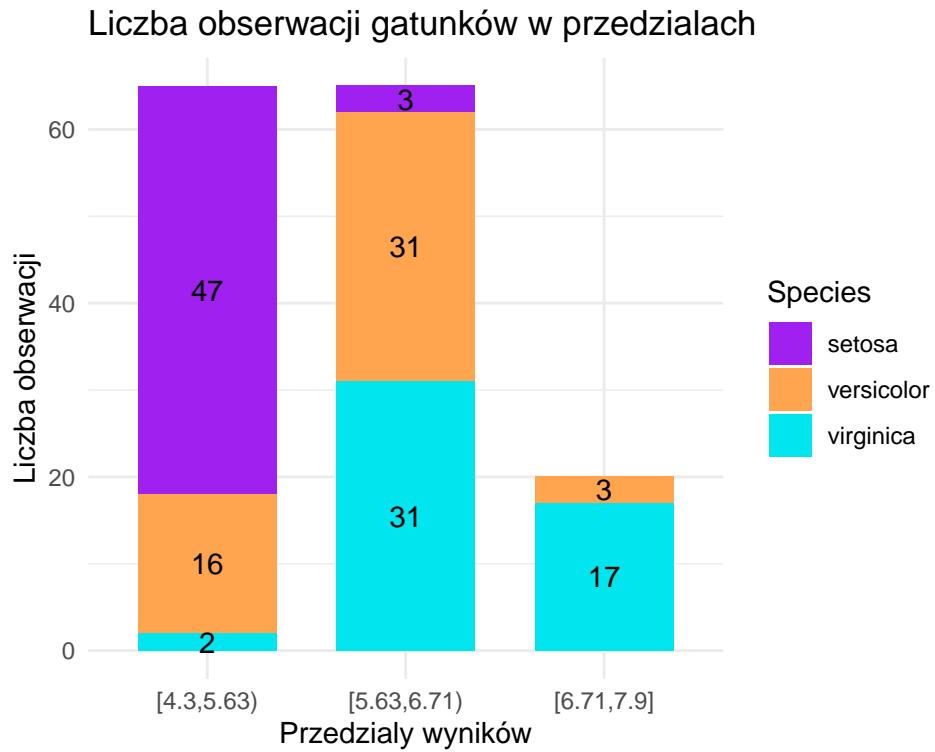
Analogiczny podział jak w poprzedniej metodzie

Zgodność na poziomie :

```
## [1] 0.96
```

Lepsza o ok 3% od ubiegłej metody

1.3.3.2 Dla najgorszej cechy : Sepal.Length (K-means)



Bardziej równomierne rozłożenie niż w metodzie poprzedniej, lecz nie jest wciąż dobre pod względem gatunkowym.

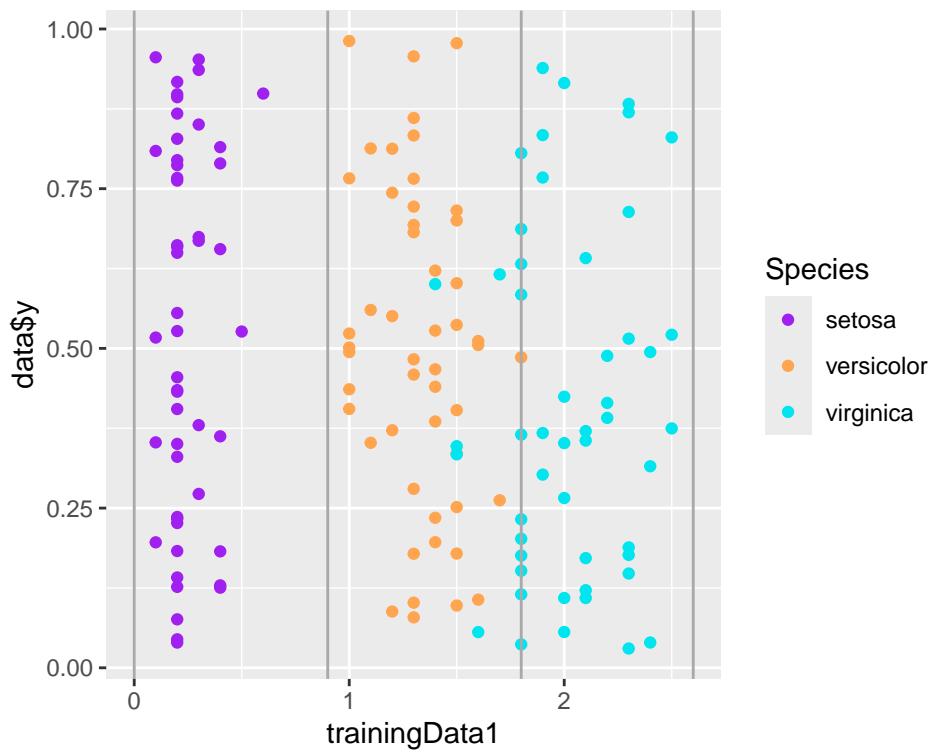
Dla najgorszej cechy mamy zgodność :

```
## [1] 0.7066667
```

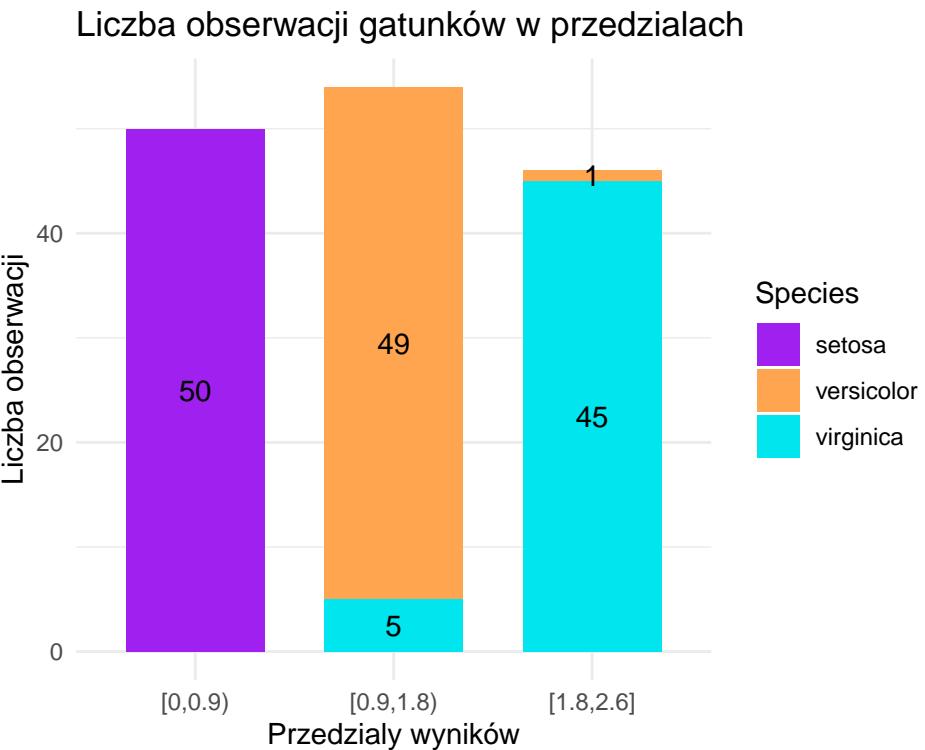
W tym przypadku jest ona **na poziomie metody Frequency (gorsza o 1)**

1.3.4 Dyskretyzacja z przedziałami zadanymi przez użytkownika (fixed)

1.3.4.1 Dla najlepszej cechy : Petal.Width (fixed)



Na wykresie mamy zaznaczone też końce przedziałów, co jest potrzebne podczas wizualizacji przedziałów zadanych przez użytkownika.

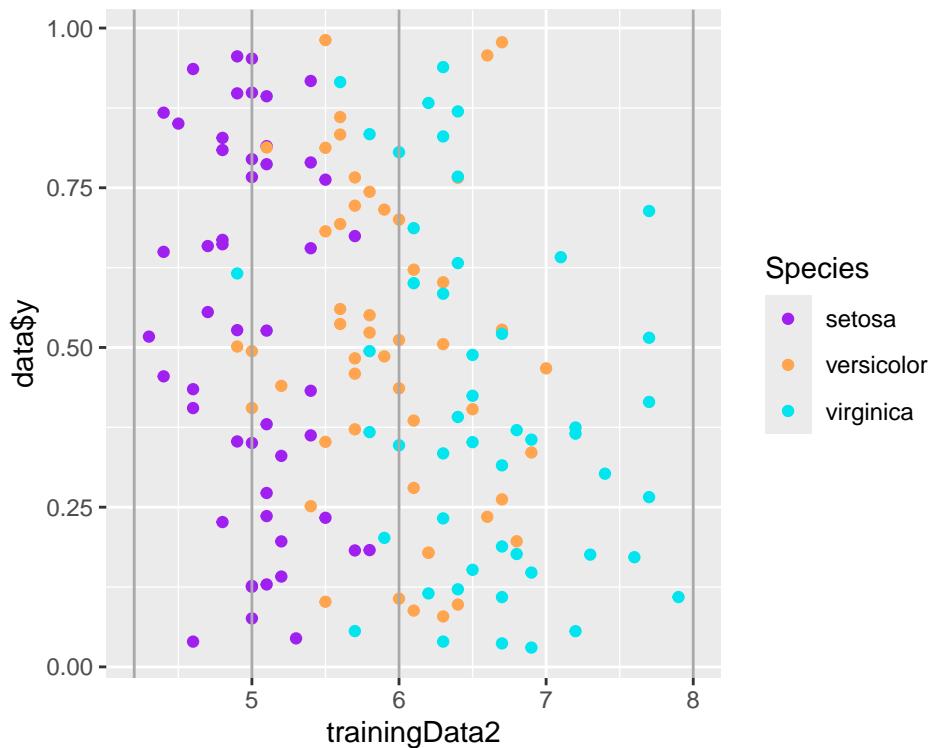


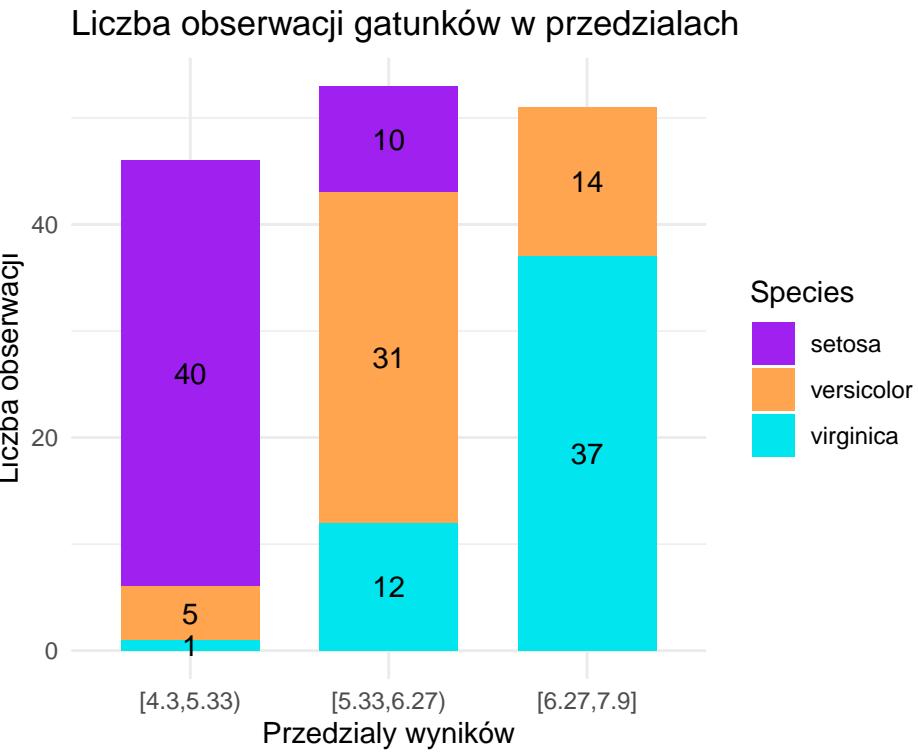
Mamy najmniejsze rozmieszanie virgnica i versicolor. Tylko 1 versicolor została źle przyporządkowana w porównaniu do aż 5 virginic.

Zgodność na poziome poprzednich dwóch metod, wynosi :

```
## [1] 0.96
```

1.3.4.2 Dla najgorszej cechy : Sepal.Length (fixed)





Równomierny rozkład między pierwszymi dwoma przedziałami ale dalej nieroóżnialne na postawie tej metody, więc nie powinniśmy używać jej do dyskretyzacji.

Dla cechy o najgorszej zdolności dyskretyzacyjnej:

```
## [1] 0.72
```

1.4 Wnioski :

Porównamy teraz zgodności procentowe wyników, dla poszczególnych algorytmów

	frequency	interval	cluster	fixed
Petal.Width	0.9466667	0.9600000	0.9600000	0.96
Sepal.Length	0.7200000	0.5729167	0.7066667	0.72

Na podstawie tabeli **przyporządkowań** dla cech najlepszych i najgorszych pod względem dyskretyzacji możemy wnioskować, że dla obecnych danych **najlepszym** algorytmem jest **frequency**(częstość). Odznacza się dobrym przyporządkowaniem dla **Petal.Width** i najlepszym dla **Sepal.Length**

2 ZADANIE 2 (Analizaskładowych głównych (Principal Component Analysis (PCA)))

2.1 a) Przygotowanie i opis danych

Podstawowe informacje nt. danych `uaScoresDataFrame`

rows	266
columns	21
discrete_columns	3
continuous_columns	18
all_missing_columns	0
total_missing_values	0
complete_rows	266
total_observations	5586
memory_usage	73496

Dane zawierają informacje o **266** miastach, obejmujące **21** cech, z których **18** to zmienne ciągłe, a **3** dyskretne. Zbiór jest **kompletny**, bez brakujących wartości, co oznacza pełne **5586** obserwacji.

Typy danych w zbiorze

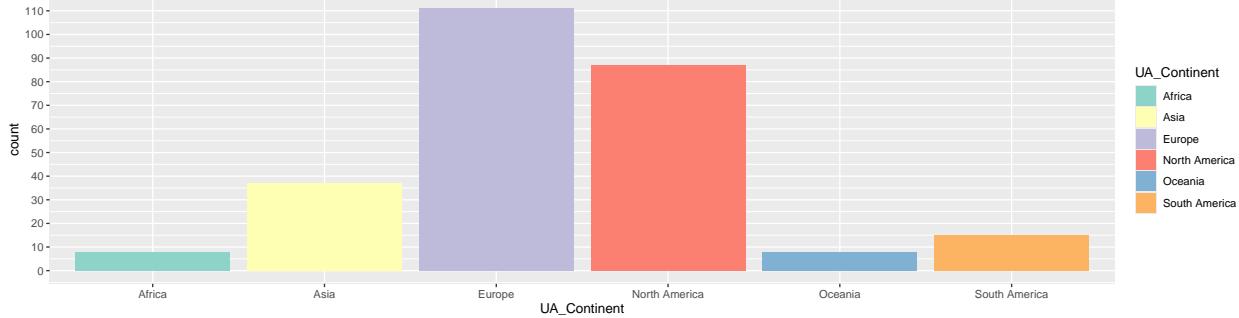


Tabela poniżej przedstawia pięć przykładowych wierszy danych.

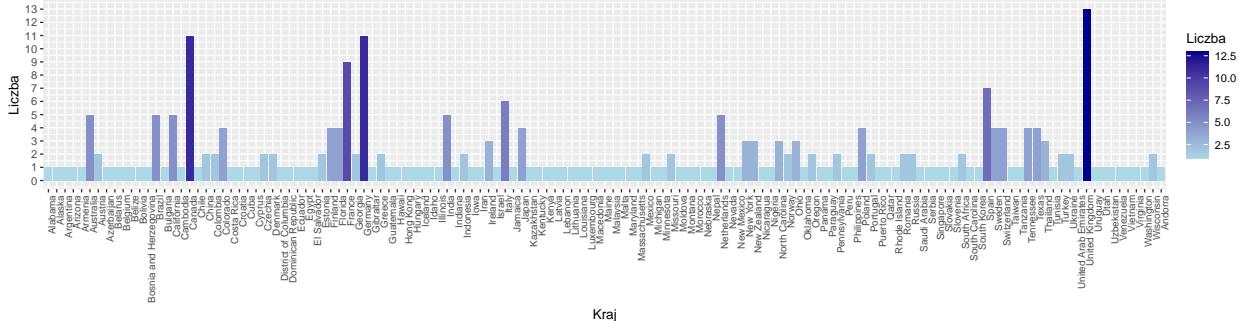
X	UA_Name	UA_Country	UA_Continent	Housing	Cost.of.Living	Startups	Venture.Capital	Travel.Connectivity	Commute
0	Aarhus	Denmark	Europe	6.132	4.015	2.827	2.512	3.536	6.312
1	Adelaide	Australia	Oceania	6.310	4.692	3.136	2.640	1.777	5.336
2	Albuquerque	New Mexico	North America	7.262	6.059	3.772	1.493	1.456	5.056
3	Almaty	Kazakhstan	Asia	9.282	9.333	2.458	0.000	4.592	5.871
4	Amsterdam	Netherlands	Europe	3.053	3.824	7.972	6.107	8.325	6.118
5	Anchorage	Alaska	North America	5.434	3.141	2.795	0.000	1.738	4.715

X	Business.Freedom	Safety	Healthcare	Education	Environmental.Quality	Economy	Taxation	Internet.Access	Leisure..Culture	Tolerance	Outdoors
0	9.940	9.617	8.704	5.367	7.633	4.887	5.068	8.373	3.187	9.739	4.130
1	9.400	7.926	7.937	5.142	8.331	6.070	4.588	4.341	4.328	7.822	5.531
2	8.671	1.343	6.430	4.152	7.319	6.514	4.346	5.396	4.890	7.028	3.515
3	5.568	7.309	4.546	2.283	3.857	5.269	8.522	2.886	2.937	6.540	5.500
4	8.837	8.504	7.907	6.180	7.597	5.053	4.955	4.523	8.874	8.368	5.307
5	8.671	3.470	6.060	3.624	9.272	6.514	4.772	4.964	3.266	7.093	5.358

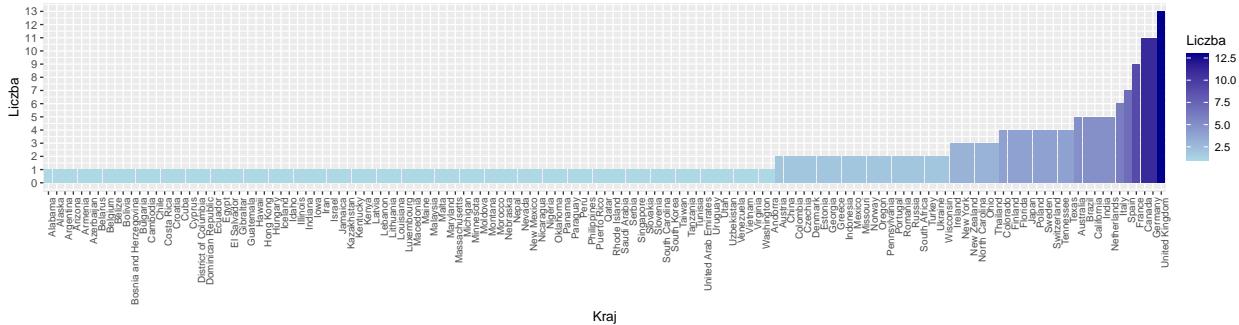
Wykres słupkowy pokazujący ilość rekordów dla każdego z kontynentów



Wykres słupkowy dla każdego kraju alfabetycznie

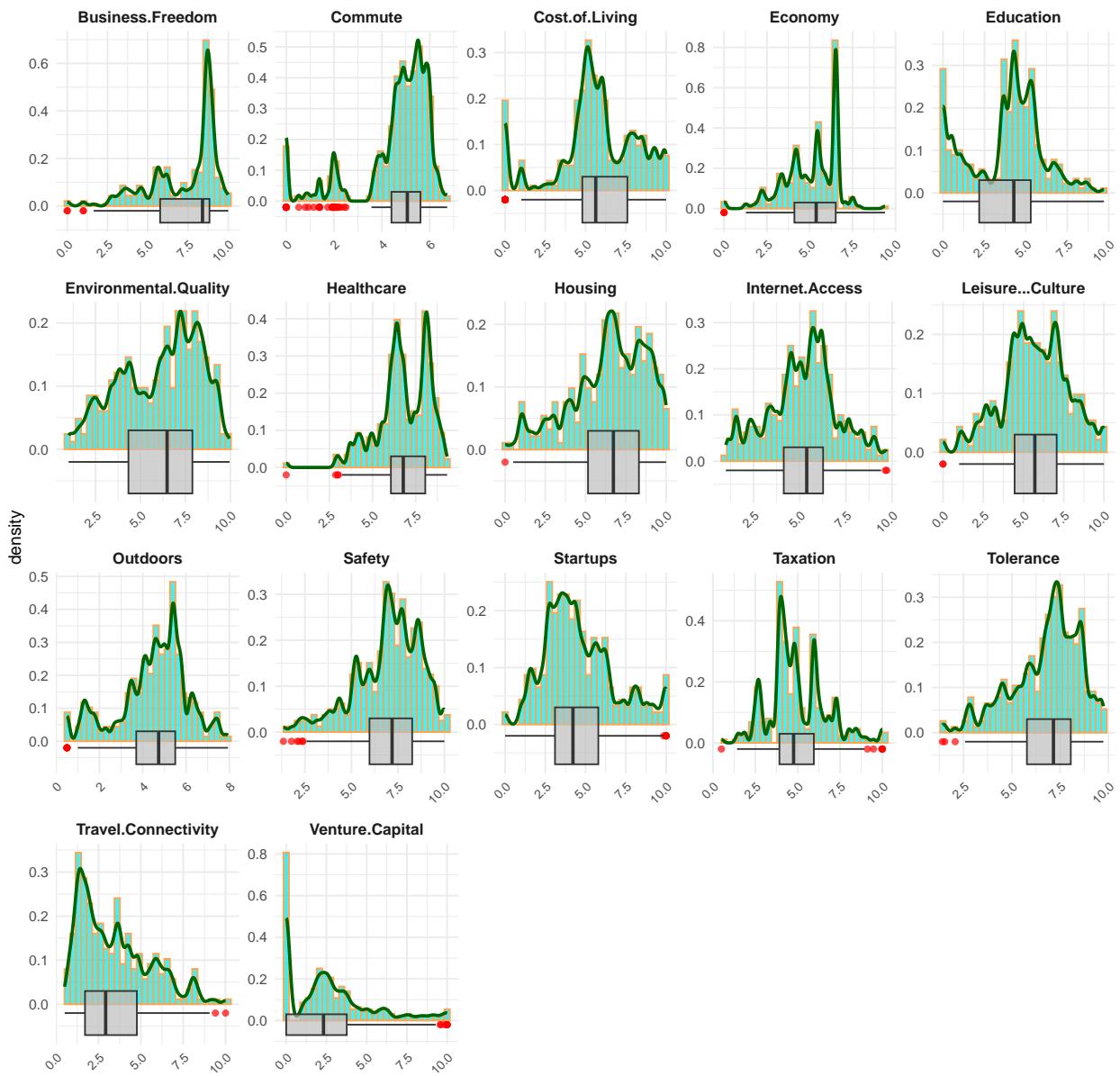


Wykres słupkowy dla kazdego kraju rosnaco



Analiza wykresów wskazuje, że większość rekordów pochodzi z krajów rozwiniętych, głównie z Europy i Ameryki Północnej.

Histogramy z estymatorami gestosci i boxplotami dla zmiennych



Wolność biznesowa (Business.Freedom):

Większość miast zapewnia dobre warunki dla biznesu (szczyt ok. 8.5), niewiele wypada słabo (<5).

Dojazdy (Commute):

Większość miast cechuje się przeciętnym/słabym poziomem skomunikowania (4–6).

Koszty życia (Cost.of.Living):

Podział na miasta średnio drogie (szczyt 5–6) i drogie (7–8).

Gospodarka (Economy):

Większość miast ma silną gospodarkę (szczyt 8–9), niewiele słabych (<5).

Edukacja (Education):

Wyraźny podział – bardzo niski (0–2) i przeciętny (4–6) poziom edukacji.

Jakość środowiska (Environmental.Quality):

Większość miast z dobrą jakością środowiska (6–8).

Opieka zdrowotna (Healthcare):

Podział na miasta z dobrą (8–9) i przeciętną (5–6) opieką, mało słabych

Mieszkalnictwo (Housing):

Dominują przeciętne warunki (5–6), część z bardzo dobrymi (8–9).

Dostęp do internetu (Internet.Access):

Powszechnie przeciętny dostęp (5–7)

Kultura i rozrywka (Leisure.&.Culture):

Przyzwoity poziom w większości miast (5–7).

Aktywności na świeżym powietrzu (Outdoors):

Główne średni poziom (4 - 6)

Bezpieczeństwo (Safety):

Większość miast jest bezpieczna (6-9)

Startupy (Startups):

Większość miast przeciętna (4–5), mniejsza grupa z bardzo dobrymi warunkami (9–10).

Podatki (Taxation):

Większość miast z umiarkowanymi lub wysokimi podatkami (4–5).

Tolerancja (Tolerance):

Dominują wysokie oceny (7–8), bardzo mało niskich (<4).

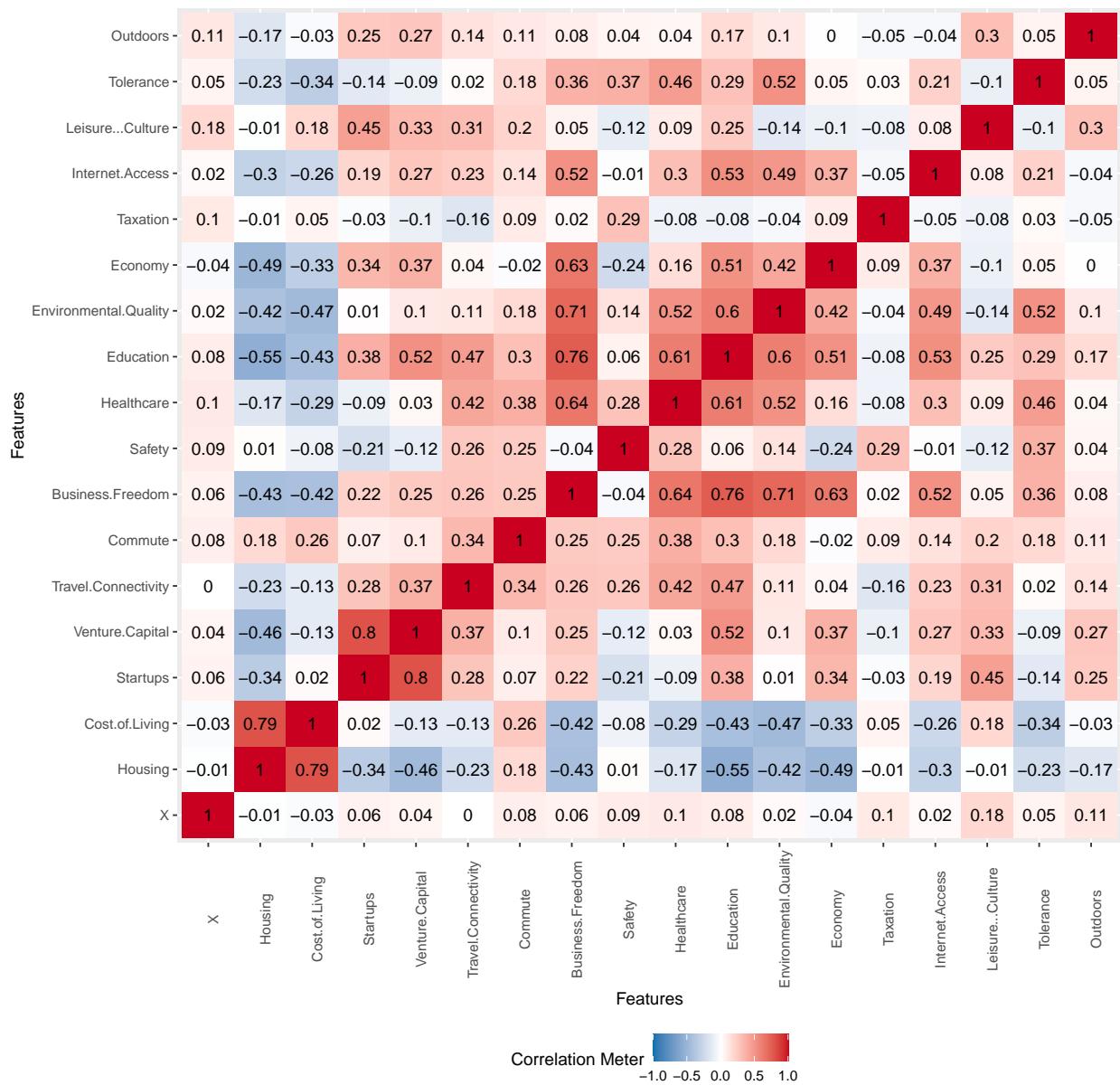
Połączenia komunikacyjne (Travel.Connectivity):

Większość miast ze słabymi połączeniami (2–3), tylko nieliczne dobre (6–7).

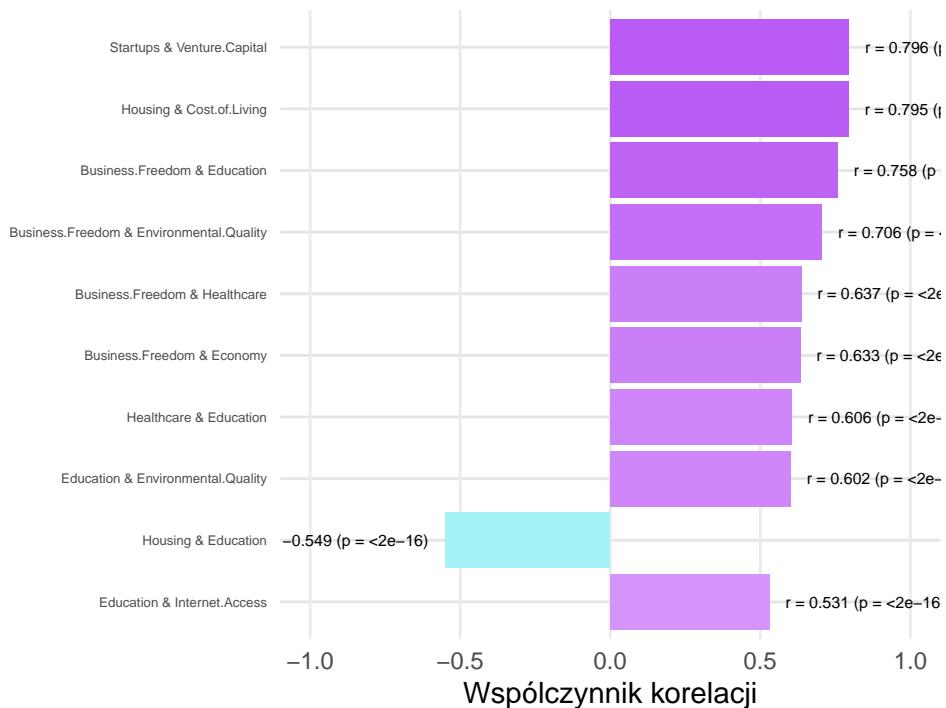
Kapitał venture (Venture.Capital):

Dostęp bardzo ograniczony – większość miast w przedziale 1–2, nieliczne wyjątki.

Macierz korelacji dla zmiennych



10 najsilniejszych istotnych korelacji



Z wykresu wynika, że **najsilniejsza korelacja** występuje między **Startups i Venture Capital**, co sugeruje, że dostęp do kapitału inwestycyjnego silnie wspiera rozwój środowiska startupowego.

Silna zależność widoczna jest również pomiędzy **Housing i Cost of Living**, co oznacza, że lepsze warunki mieszkaniowe często wiążą się z wyższymi kosztami życia.

Silne korelacje dotyczą także:

- **Business Freedom & Education,**
- **Business Freedom & Environmental Quality,**
- **Business Freedom & Healthcare,**
- **Business Freedom & Economy**

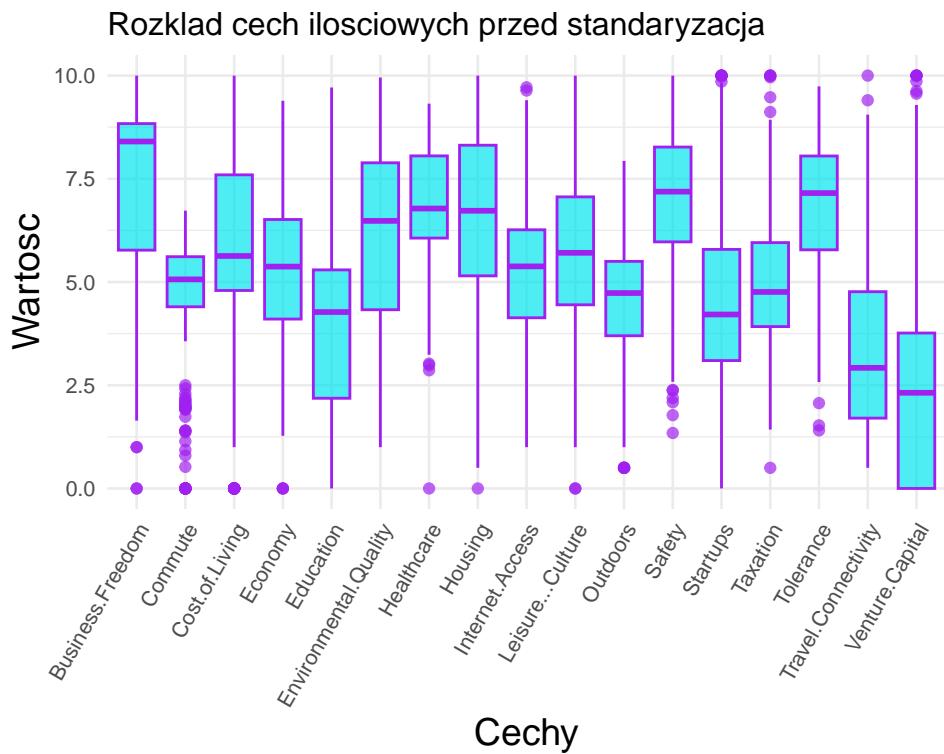
co sugeruje, że **większa swoboda gospodarcza** często idzie w parze z **lepszą edukacją, czystszym środowiskiem, lepszą opieką zdrowotną** i ogólnie **lepszą gospodarką**.

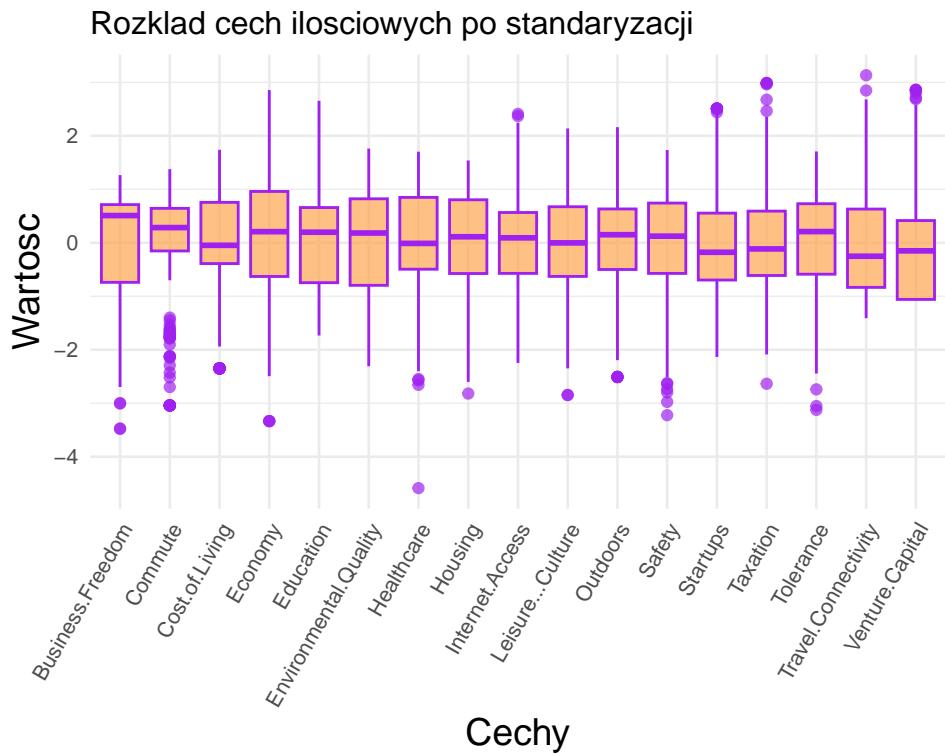
	Wariancja
Housing	5.265
Cost.of.Living	5.988
Startups	4.635
Venture.Capital	6.520
Travel.Connectivity	4.375
Commute	2.320
Business.Freedom	4.450

	Wariancja
Safety	3.051
Healthcare	2.196
Education	4.897
Environmental.Quality	4.840
Economy	2.302
Taxation	2.855
Internet.Access	3.505
Leisure... Culture	4.027
Tolerance	2.974
Outdoors	2.534

Dlaczego standaryzacja jest konieczna?

- Bez standaryzacji **PCA faworyzuje zmienne o większym zróżnicowaniu**, co może prowadzić do **błędnych wniosków**.
- **Standaryzacja** (średnia = 0, odchylenie = 1) zapewnia **równomierne traktowanie** wszystkich zmiennych, eliminując wpływ **skali**.



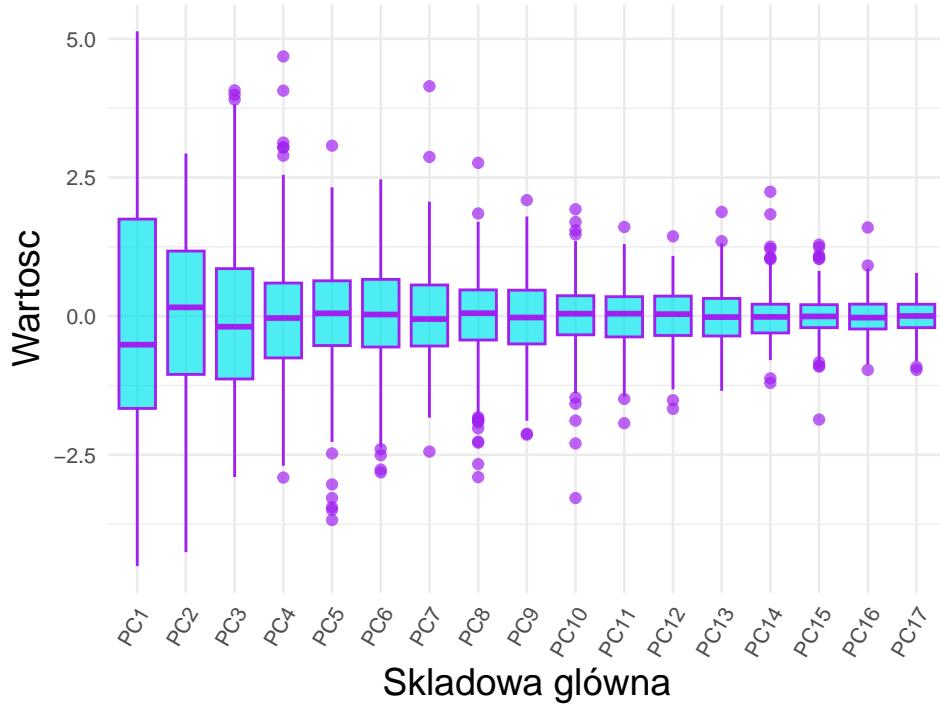


2.2 b) Wyznaczenie składowych głównych

Analiza PCA

Składowa	Odchylenie_standardowe	Procent_wariancji	Kumulatywna_wariancja
PC1	2.251	29.80	29.80
PC2	1.606	15.16	44.96
PC3	1.443	12.25	57.21
PC4	1.140	7.65	64.86
PC5	1.095	7.05	71.90
PC6	0.980	5.65	77.55
PC7	0.831	4.06	81.62
PC8	0.815	3.90	85.52
PC9	0.764	3.43	88.95
PC10	0.651	2.50	91.45
PC11	0.569	1.90	93.35
PC12	0.539	1.71	95.06
PC13	0.524	1.62	96.68
PC14	0.434	1.11	97.79
PC15	0.393	0.91	98.69
PC16	0.352	0.73	99.42
PC17	0.313	0.58	100.00

Rozkład wartości składowych głównych



PC1 wykazuje **największą zmienność**, tłumacząc największą część wariancji. Kolejne składowe mają coraz mniejszy wpływ na strukturę danych. Od PC7–PC8 zmienność jest **niewielka**, co sugeruje ograniczone znaczenie analityczne dalszych komponentów.

Macierz korelacji zmiennych głównych

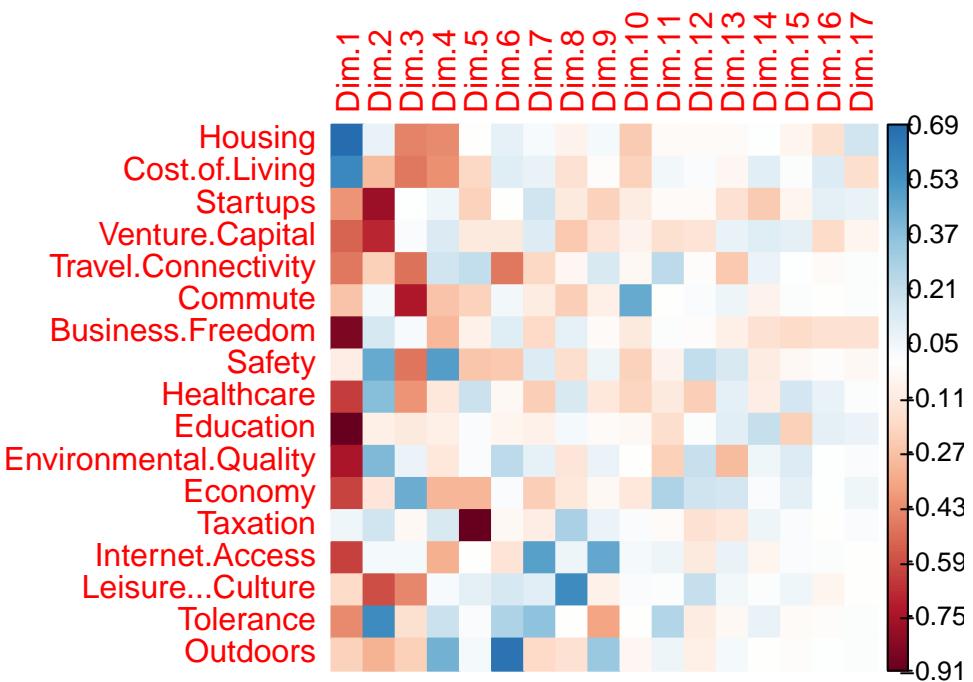


Tabela 5: Wektory ładunków dla PC1, PC2 i PC3

	PC1	PC2	PC3
Housing	0.3078251	0.0533534	-0.3135465
Cost.of.Living	0.2596091	-0.1757815	-0.3305352
Startups	-0.1802385	-0.4834415	0.0061000
Venture.Capital	-0.2365974	-0.4274509	0.0148768
Travel.Connectivity	-0.2094543	-0.1353067	-0.3397760
Commute	-0.1142045	0.0259310	-0.5057359
Business.Freedom	-0.3772809	0.0982196	0.0241046
Safety	-0.0389355	0.2871039	-0.3330100
Healthcare	-0.2803590	0.2419482	-0.2810248
Education	-0.4025620	-0.0490795	-0.0738645
Environmental.Quality	-0.3262220	0.2525355	0.0535717
Economy	-0.2731752	-0.0740033	0.3086705
Taxation	0.0262992	0.1074151	-0.0201849
Internet.Access	-0.2761922	0.0227056	0.0284416
Leisure...Culture	-0.0744466	-0.3647324	-0.3050545
Tolerance	-0.1897496	0.3550911	-0.1027251
Outdoors	-0.0915866	-0.1933825	-0.1485868

Składowa główna 1 (PC1): Jakość życia vs. dostępność ekonomiczna

PC1 kontrastuje miasta o wysokiej jakości usług z miastami ekonomicznie dostępnymi. Naj-silniejsze ładunki:

- **Edukacja** (-0.40),
- **Wolność biznesowa** (-0.38),
- **Jakość środowiska** (-0.33) - wszystkie **ujemne**
- **Mieszkalnictwo** (0.31),
- **Koszty życia** (0.26) - **dodatnie**

Wysokie wartości PC1 wskazują na miasta o niższych kosztach życia, ale słabszej infrastrukturze społecznej. Niskie wartości PC1 charakteryzują rozwiniętą infrastrukturę społeczną przy wyższych kosztach.

Składowa główna 2 (PC2): Środowisko startupowe vs. jakość społeczna

PC2 przeciwstawia ośrodkie technologiczne miastom o wysokich wskaźnikach społecznych:

- **Startupy** (-0.48),
- **Kapitał venture** (-0.43),
- **Kultura i rozrywka** (-0.36) - **ujemne**
- **Tolerancja** (0.36),
- **Jakość środowiska** (0.25),
- **Bezpieczeństwo** (0.29) - **dodatnie**

Wysokie wartości PC2 oznaczają miasta bardziej przyjazne społecznie, niskie wartości wskazują na dynamiczne centra technologiczne.

Składowa główna 3 (PC3): Komfort codziennego życia vs. gospodarka

PC3 zestawia komfort życia codziennego z rozwojem ekonomicznym:

- **Dojazdy** (-0.51),

- **Połączenia komunikacyjne** (-0.34),
- **Bezpieczeństwo** (-0.33) - ujemne
- **Gospodarka** (0.31) - dodatnie

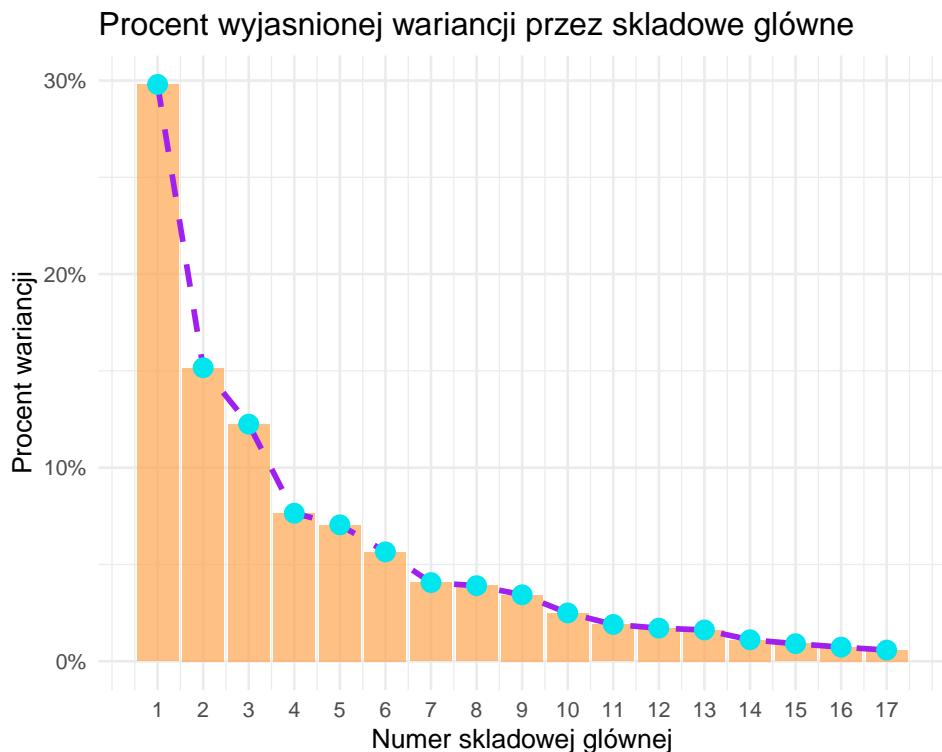
Miasta o wysokich wartościach PC3 mają silną gospodarkę kosztem wygody życia codziennego, podczas gdy niskie wartości PC3 wskazują na większy komfort przy mniej dynamicznej ekonomii.

Podsumowanie

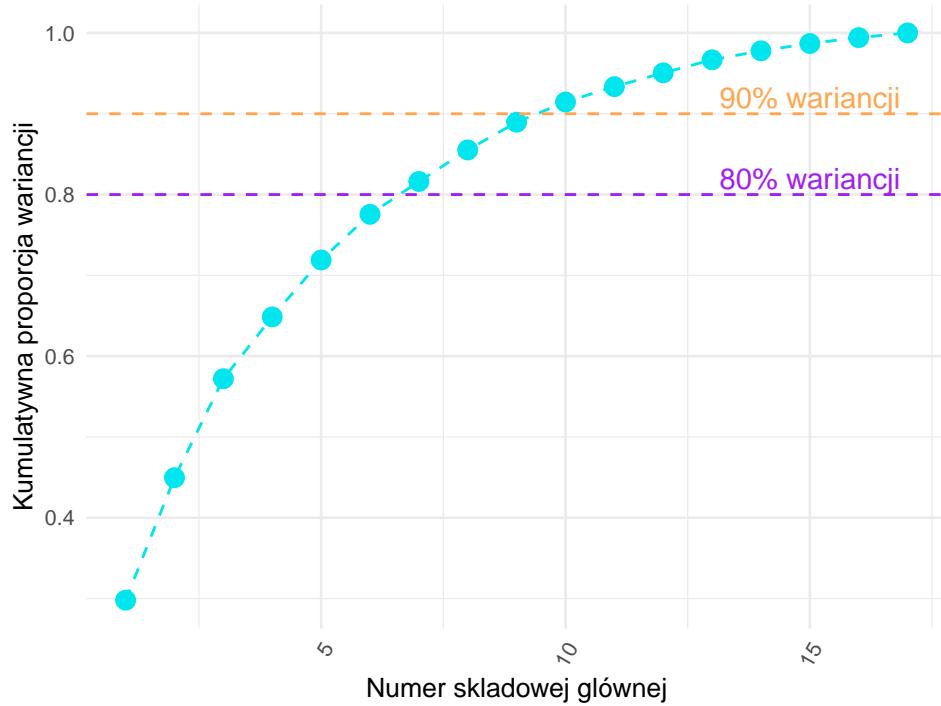
Te trzy wymiary tworzą kompleksowe ramy do klasyfikacji miast:

- **PC1:** Balans między rozwojem społecznym a dostępnością ekonomiczną
- **PC2:** Równowaga między ekosystemem startupowym a jakością życia społecznego
- **PC3:** Kompromis między codziennym komfortem a silną gospodarką

2.3 c) Zmienność odpowiadająca poszczególnym składowym



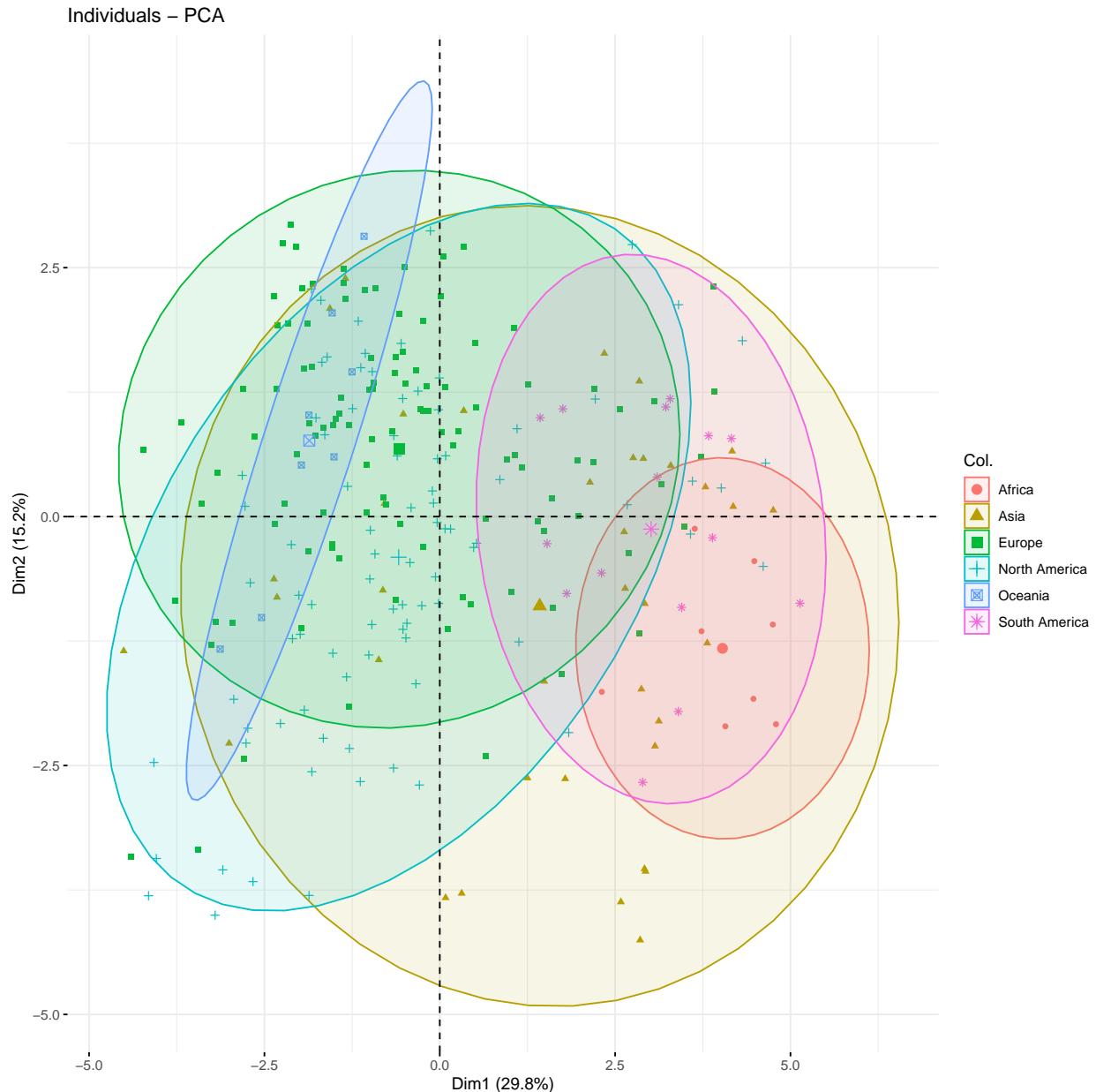
Kumulatywna wariancja wyjaśniona przez składowe główne



Na podstawie przedstawionych wykresów można zauważyć, że **pierwsze składowe główne** mają największy wpływ na wyjaśnienie wariacji danych. W szczególności **pierwsza składowa (PC1)**. Kolejne składowe, takie jak **PC2** i **PC3**, również wnoszą istotne informacje, ale ich udział w wyjaśnieniu wariacji jest stopniowo coraz mniejszy.

Z wykresu skumulowanej wariacji można wywnioskować, że pierwsze **7** składowych wyjaśnia około **80%** całkowitej zmienności, a pierwsze **10** składowych odpowiada za **90%** wariacji, co sugeruje, że można ograniczyć liczbę analizowanych cech bez znaczącej utraty informacji.

2.4 d) Wizualizacja danych wielowymiarowych



Europa (zielone kwadraty)

- Po lewej stronie PC1 → wysoka jakość życia, silna infrastruktura społeczna (edukacja, bezpieczeństwo, środowisko).
- Wyższe koszty życia, mniejsza dostępność ekonomiczna.
- Oś PC2 lekko dodatnia → zrównoważony rozwój społeczno-startupowy.

Ameryka Północna (turkusowe plusy)

- Również lewa strona PC1 → wysoka jakość życia.
- Niższe wartości PC2 → dynamiczne środowisko technologiczne, kosztem aspektów

społecznych.

- Duże **zróżnicowanie** – od wybitnie startupowych miast (np. USA) po umiarkowane.

Azja (*żółte trójkąty*)

- **Niskie koszty życia**, ale słabsza jakość usług społecznych.
- **Silna obecność centrów startupowych** (np. Singapur, Chiny).

Afryka (*czerwone koła*)

- **Wysokie PC1** → **niska jakość usług społecznych**, ale **dobra dostępność ekonomiczna**.
- **PC2 bliskie zeru** → przeciętna jakość społeczna, **słabe środowisko innowacyjne**.

Oceania (*niebieskie romby*)

- **Lewa górną ćwiartką** (PC1 ujemny, PC2 dodatni):
- **Bardzo wysoka jakość życia**, dobre wskaźniki społeczne (tolerancja, środowisko).
- Mało miast, ale **wyraźnie pozytywne wyniki**.

Ameryka Południowa (*fioletowe gwiazdki*)

- **Blisko środka** (lejko dodatni PC1):
- **Umiarkowana jakość życia**, rozsądne koszty.
- **Niska aktywność startupowa**.

Mapa miast względem ich PC1 i PC2

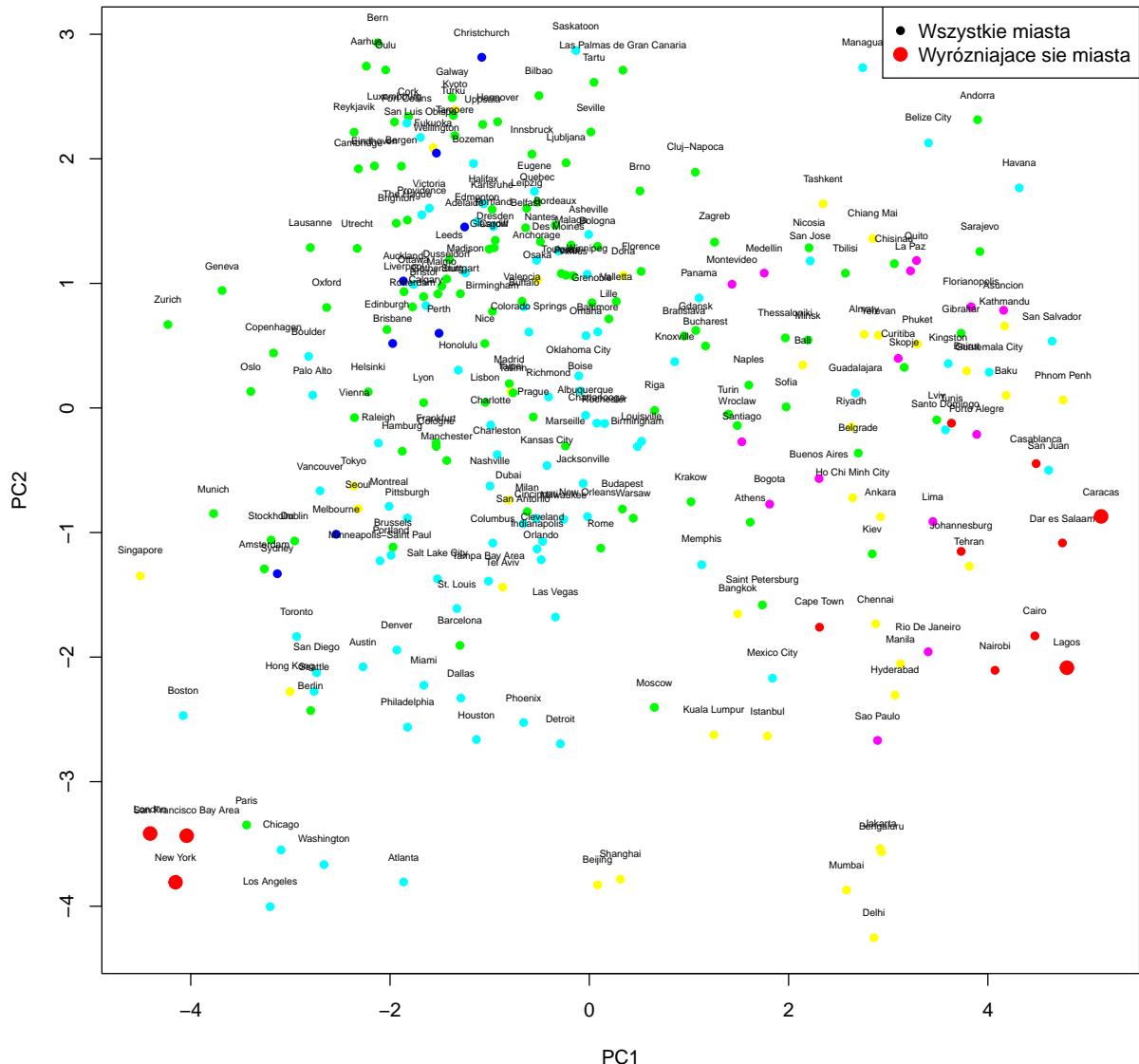


Tabela 6: Miasta najbardziej oddalone od środka układu współrzędnych (PCA)

	PC1	PC2	Miasto	Odległość od środka
172	-4.15	-3.81	New York	5.63
139	-4.41	-3.42	London	5.58
213	-4.04	-3.43	San Francisco Bay Area	5.30
127	4.79	-2.09	Lagos	5.23
53	5.14	-0.87	Caracas	5.21

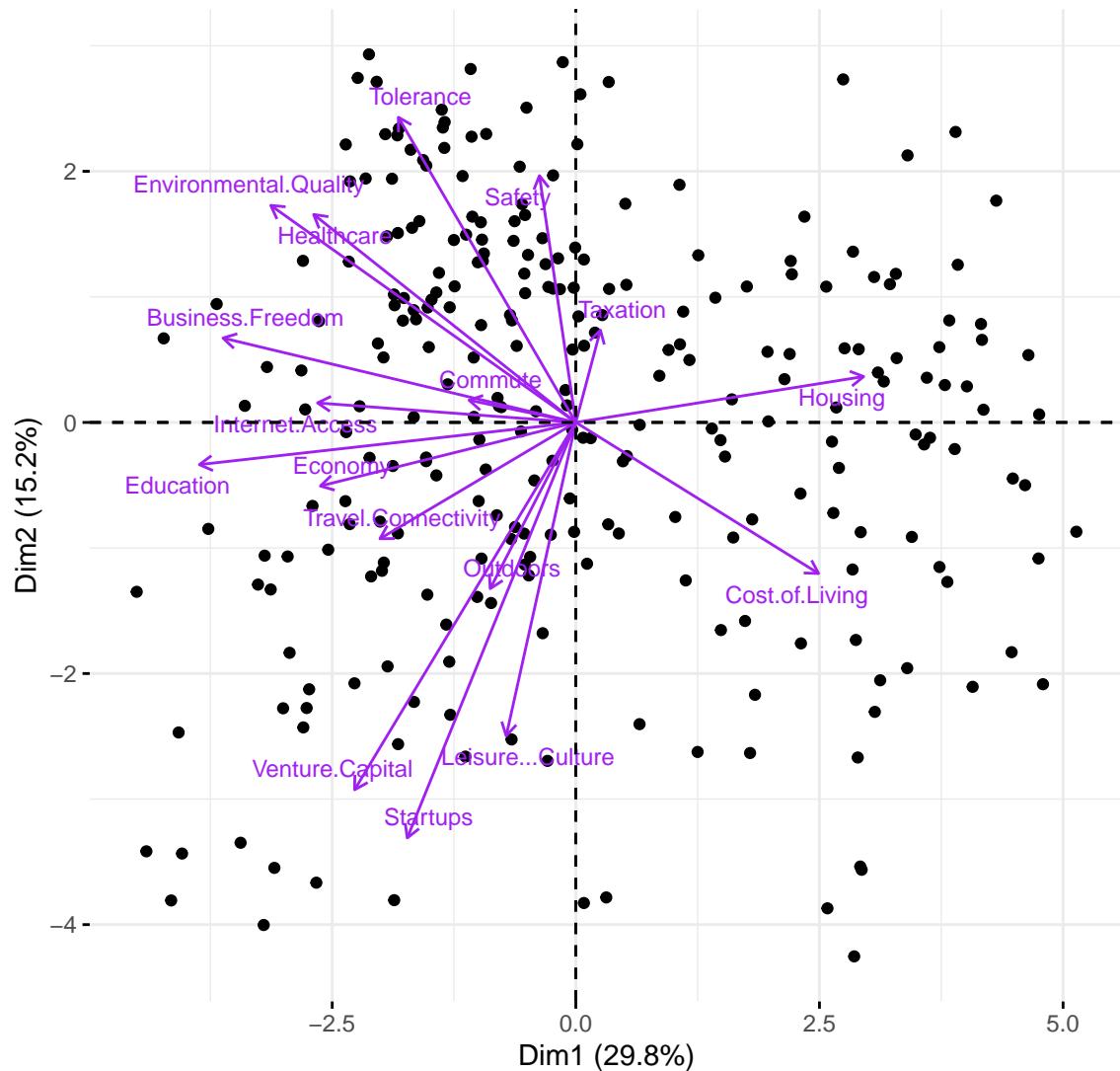
Miasta najbardziej oddalone od środka układu PCA charakteryzują się skrajnymi wartościami dla głównych składowych, co oznacza, że wyróżniają się na tle reszty pod względem profilu jakości życia, kosztów, rozwoju technologicznego i społecznego.

- Nowy Jork (**PC1 = -4.15, PC2 = -3.81**) – silnie ujemne wartości obu składowych wskazują na wysoką jakość infrastruktury społecznej (PC1) oraz intensywnie rozwinięte środowisko startupowe (PC2). To miasto dynamiczne, ale jednocześnie bardzo kosztowne.
 - Londyn (**PC1 = -4.41, PC2 = -3.42**) – podobnie jak Nowy Jork, łączy drogie życie z zaawansowaną infrastrukturą społeczną i wysokim rozwojem technologicznym.
 - San Francisco Bay Area (**PC1 = -4.04, PC2 = -3.43**) – silne centrum technologiczne o niskiej dostępności ekonomicznej, ale z najwyższą koncentracją startupów i kapitału venture.
 - Lagos (**PC1 = 4.79, PC2 = -2.09**) – wysoka wartość PC1 oznacza niski koszt życia i ograniczoną infrastrukturę społeczną, przy jednoczesnym udziale w środowisku startupowym (ujemne PC2). To przykład miasta rozwijającego się, ale jeszcze bez zaplecza społecznego.
 - Caracas (**PC1 = 5.14, PC2 = -0.87**) – miasto o niskiej jakości życia społecznego i dużej ekonomicznej dostępności, z niewielkim zaangażowaniem w nowoczesne sektory gospodarki.
-

Te wyniki pokazują, że największe odległości od środka PCA mają zarówno najbardziej rozwinięte miasta świata, jak i najbardziej marginalne – ale z różnych powodów: jedne z powodu nadmiaru infrastruktury i kosztów, inne z powodu braku zasobów społecznych i niskich kosztów życia.

2.5 e) Korelacja zmiennych

Biplot PCA – zmienne



2.5.1 Wnioski z biplotu:

- Długość strzałki oznacza wpływ zmiennej na PC1 i PC2
- Kierunek strzałek:
 - *Zbieżne* → dodatnia korelacja
 - *Przeciwnie* → ujemna korelacja
 - *Prostopadłe* → brak korelacji

2.5.1.1 Silne zależności: Dodatnie: - Startups – Venture Capital – Leisure & Culture

- Safety – Tolerance
- Business.Freedom - Education - Environmental.Quality

Ujemne: - Cost of Living vs Environmental Quality, Education, Business Freedom

- Housing vs Education

Brak istotnej korelacji: - Safety – Housing

- Commute – Venture Capital

2.5.1.2 Porównanie z macierzą korelacji: Wyniki biplotu są spójne z macierzą `cor()`:

- Najsilniejsza korelacja: Startups – Venture Capital
- Housing – Cost of Living
- Housing - Education
- Business Freedom silnie koreluje z:
 - Education
 - Environmental Quality
 - Healthcare
 - Economy

2.6 f) Końcowe wnioski

Na podstawie przeprowadzonych analiz i wyników biplotu, kilka istotnych wniosków:

1. Reprezentacja danych:

- **PC1** i **PC2** wyjaśniają główną część zmienności danych, szczególnie **PC1**, która tłumaczy różnice w jakości życia i dostępności ekonomicznej. **PC3** dostarcza dodatkowych informacji, ale ma mniejszy wpływ. Pierwsze 2–3 składowe wyjaśniają około **80% wariancji**.

2. Składowe główne:

- **PC1** (Jakość życia vs. dostępność ekonomiczna): Większość miast znajduje się na przeciwnych końcach tej osi, pokazując różnice w równowadze między wysokimi kosztami życia a rozwiniętą infrastrukturą społeczną.
- **PC2** (Środowisko startupowe vs. jakość społeczna): Składa się z zmiennych takich jak “Startups”, “Venture Capital” i “Leisure & Culture”, które opisują dynamiczne ośrodkie technologiczne.
- **PC3** (Komfort życia vs. gospodarka): Zestawia miasta o silnej gospodarce z tymi, które oferują wyższy komfort życia codziennego.

3. Znaczenie standaryzacji:

- **Standaryzacja** zmiennych miała kluczowy wpływ na wyniki PCA. Bez niej, zmienne o większym zróżnicowaniu (np. koszty życia) mogłyby dominować, prowadząc do błędnych wniosków. Po standaryzacji, każda zmienna ma równy wpływ, co zapewnia bardziej sprawiedliwą ocenę.

4. Geograficzne różnice:

- Z analizy biplotu wynika, że miasta rozmieszczone są zgodnie z globalnymi różnicami w jakości życia, dostępności ekonomicznej i rozwoju startupów. Duże zróżnicowanie występuje między miastami rozwiniętymi (np. Nowy Jork, Londyn) a tymi na początku drogi rozwoju (np. Lagos, Caracas). Wiele miast z krajów rozwiniętych znajduje się w lewym dolnym rogu biplotu, co wskazuje na **wysoką jakość usług społecznych i wyższe koszty życia**.

2.6.1 Wnioski ogólne:

- **Analiza PCA** dostarcza cennych informacji o relacjach między różnymi aspektami życia w miastach. Widać, że miasta o wyższych kosztach życia mają rozwiniętą infrastrukturę społeczną, podczas gdy te o niższych kosztach życia oferują mniejszy rozwój infrastruktury, ale większy dostęp do rozwoju gospodarczego.
- **Standaryzacja** jest kluczowa do uzyskania rzetelnych wyników PCA, eliminując wpływ różnic w skali danych.

3 ZADANIE 3 (Skalowanie wielowymiarowe (Multidimensional Scaling (MDS)))

3.1 a) Dane: titanic_train (R-pakiet titanic)

Zbiór danych zawiera wybrane charakterystyki opisujące pasażerów Titanica (w tym m.in. takie zmienne jak: wiek, płeć, miejsce rozpoczęcia podróży czy klasa pasażerska) wraz z informacją czy dana osoba przeżyła katastrofę (zmienna Survived).

3.2 b) Przygotowanie danych

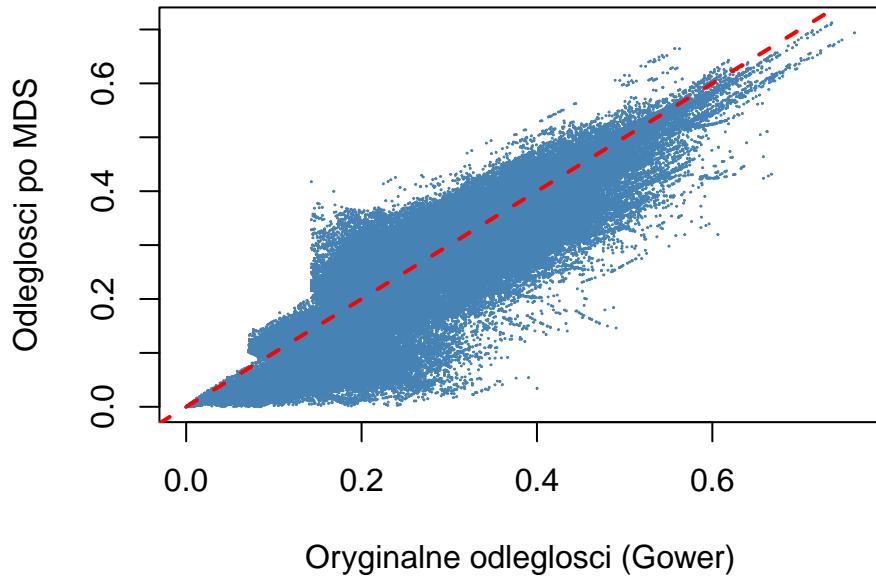
Wczytane dane, niepotrzebne kolumny zostały usunięte, oraz typy poszczególnych cech zostały zaaktualizowane na odpowiednie czyt. ordered, numeric

3.3 c) Redukcja wymiaru na bazie MDS

Redukuję wymiar danych korzystając z **metody metrycznej (Funkcja cmdscale)**

Kolejno tworzymy diagram Shephera

Diagram Shephera



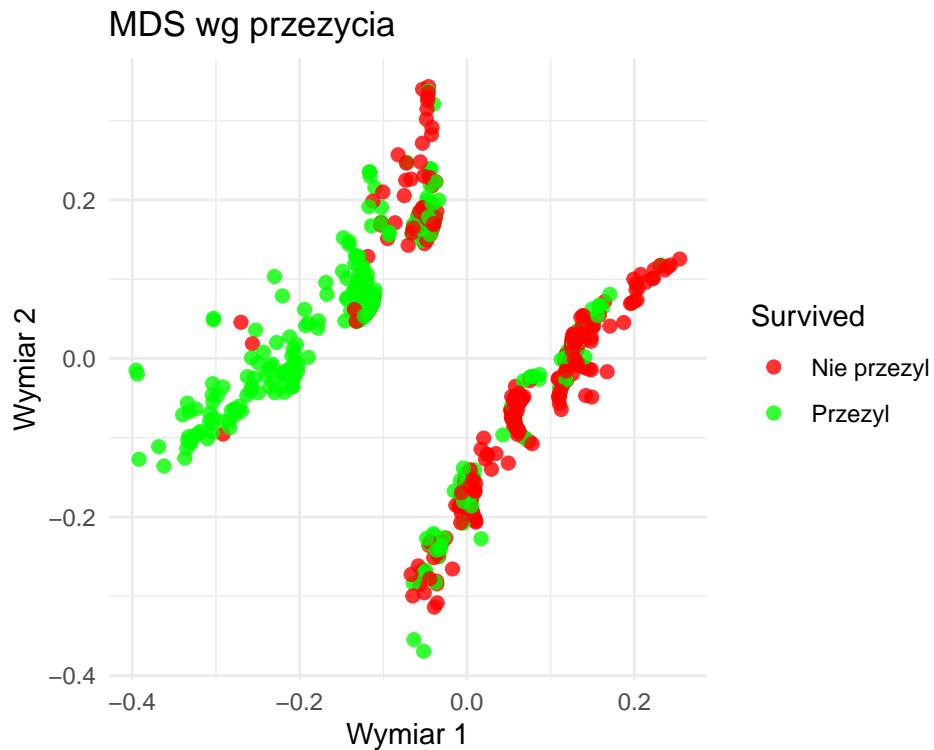
Ocena jakości odwzorowania – Diagram Shephera

Na powyższym wykresie porównano oryginalne odległości (metoda Gowera) z odległościami otrzymanymi po zastosowaniu MDS.

Większość punktów układa się wzdłuż linii $x = y$, co wskazuje na dobre odwzorowanie struktury danych. Choć niektóre punkty od niej odbiegają, to **różnice są stosunkowo niewielkie** – zwłaszcza w zakresie mniejszych odległości, które są najbardziej istotne dla zachowania lokalnej struktury danych.

Wniosek: Transformacja MDS zachowała strukturę danych na akceptowalnym poziomie. Diagram Shephera potwierdza, że skalowanie wielowymiarowe wiernie odtworzyło relacje między punktami, szczególnie w przypadku najbliższych sąsiadów.

3.4 d) Wizualizacja danych



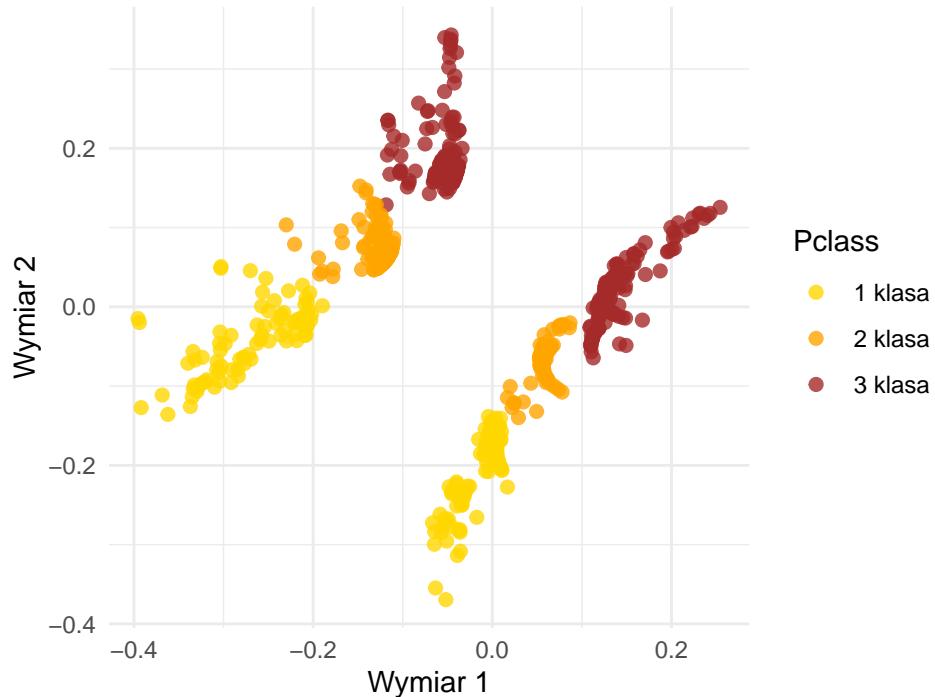
Na powyższym wykresie przedstawiono wynik analizy MDS, w której obiekty zostały rozróżnione ze względu na zmienną Survived.

Na wykresie wyraźnie widoczny jest **podział obiektów na dwa skupiska**. Pierwsze z nich (po lewej stronie wykresu) charakteryzuje się **dużym odsetkiem osób, które przeżyły**, natomiast drugie (po prawej) skupia głównie osoby, które **nie przeżyły**.

Nie zaobserwowano **obserwacji odstających** ani punktów jednoznacznie nietypowych — wszystkie dane mieszczą się w obrębie naturalnych skupisk.

Na podstawie wykresu można wnioskować, że **analiza MDS skutecznie wydobyła ukrytą strukturę danych**, związaną ze zmienną Survived.

MDS wg klasy



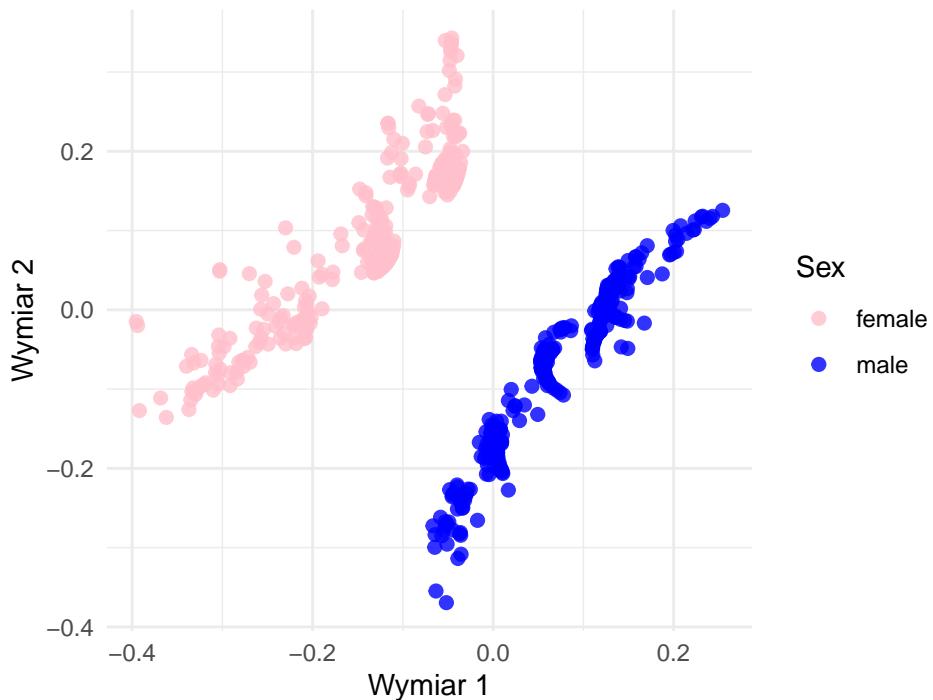
Na powyższym wykresie przedstawiono wynik analizy MDS z uwzględnieniem klasy, w której podróżowali pasażerowie (Pclass).

Rozkład punktów w przestrzeni odwzorowanej za pomocą MDS jest stosunkowo **równomierny**, a poszczególne klasy są **rozproszone w obrębie różnych skupisk**. Nie obserwuje się jednoznacznej separacji przestrzennej ze względu na klasę podróży. Wskazuje to na brak silnej zależności między zmienną Pclass a układem punktów w przestrzeni MDS.

W odróżnieniu od zmiennej Survived, która wykazywała wyraźną strukturę klasową, tutaj nie ma widocznych skupisk odpowiadających konkretnym klasom. W każdej z trzech głównych grup przestrzennych występują pasażerowie różnych klas, co sugeruje, że klasa podróży **nie była głównym czynnikiem różnicującym obserwacje** w tej analizie wielowymiarowej.

Nie zaobserwowano również punktów odstających ani obserwacji nietypowych — dane układają się w sposób naturalny i uporządkowany.

MDS wg płci



Na powyższym wykresie przedstawiono wynik analizy MDS z podziałem na płeć pasażerów (Sex). Widoczny jest **wyraźny podział** na dwie grupy, odpowiadające kobietom i mężczyznom.

Porównując ten wykres z wcześniejszą analizą przeżywalności (Survived), można zauważać, że grupa odpowiadająca kobietom **częściej pokrywa się** z obszarami o wyższej przeżywalności. Jest to zgodne zarówno z intuicją, jak i historycznymi danymi dotyczącymi katastrofy Titanica, gdzie kobiety miały znacznie większe szanse przeżycia niż mężczyźni.

Podobnie jak wcześniej, **nie zaobserwowano istotnych obserwacji odstających**, co świadczy o dobrej jakości danych i prawidłowym odwzorowaniu relacji między obserwacjami.