

Sprawozdanie 2

Eksploracja danych

Kacper Szmigielski, 282255 i Mateusz Wizner, 277508

2025-04-29

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	3
1.3.1 Metoda : Równe częstotliwości(Frequency)	3
1.3.2 Metoda : Równe szerokości (Interval)	6
1.3.3 Metoda : k najbliższych sąsiadów (K-means)	8
1.3.4 Dyskretyzacja z przedziałami zadanyimi przez użytkownika (fixed) .	11
1.4 Wnioski :	13
2 ZADANIE 2 (Analizaskładowych głównych (Principal Component Analysis (PCA)))	13
2.1 a) Przygotowanie danych	13
2.2 b) Wyznaczenie składowych głównych	20
2.3 c) Zmienność odpowiadająca poszczególnym składowym	21
2.4 d) Wizualizacja danych wielowymiarowych	23
2.5 e) Korelacja zmiennych	23
2.6 d) Wizualizacja danych wielowymiarowych	23
2.7 e) Korelacja zmiennych	23
2.8 f) Końcowe wnioski	23
2.9 f) Końcowe wnioski	23
3 ZADANIE 3 (Skalowaniewielowymiarowe (Multidimensional Scaling (MDS)))	23
3.1 a) Dane: titanic_train (R-pakiet titanic)	23
3.2 b) Przygotowanie danych	24
3.3 c) Redukcja wymiaru na bazie MDS	24
3.4 d) Wizualizacja danych	24

1 ZADANIE 1 (Dyskretyzacja(przedziałowanie) cech ciągłych)

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

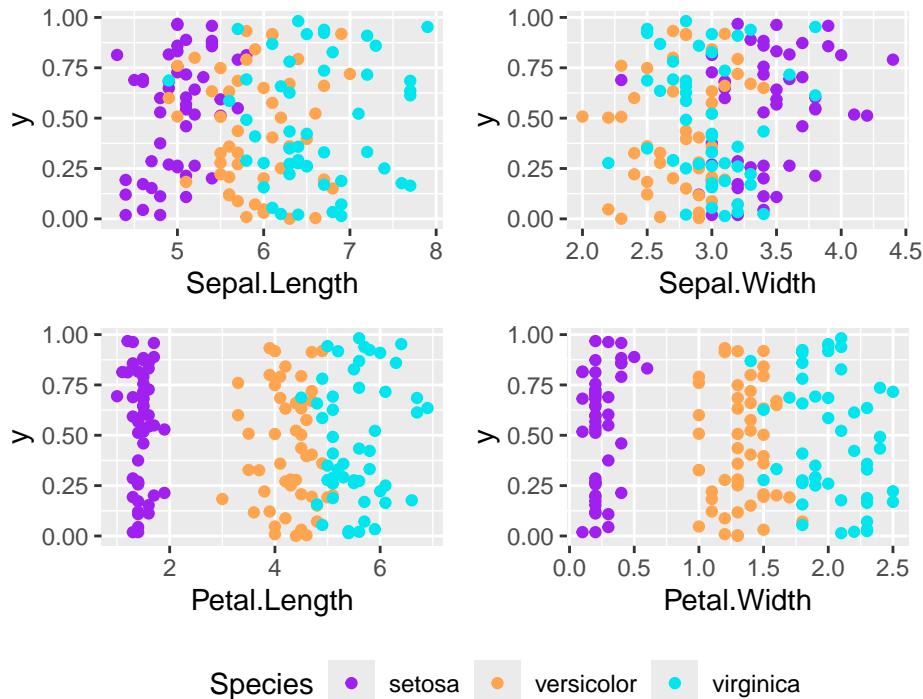
3 Pierwsze wiersze z pakietu iris

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

- Pomiary dotyczą **długości oraz szerokości** dwóch różnych części kwiatu – działa kielicha (ang. sepal) oraz płatka (ang. petal).

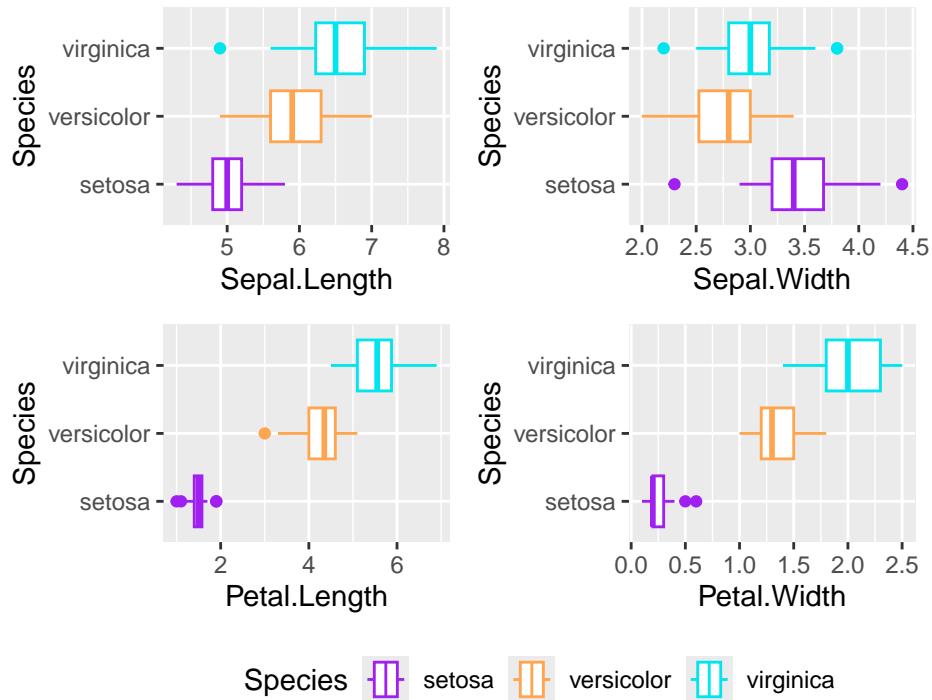
1.2 b) Wybór cech

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



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

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

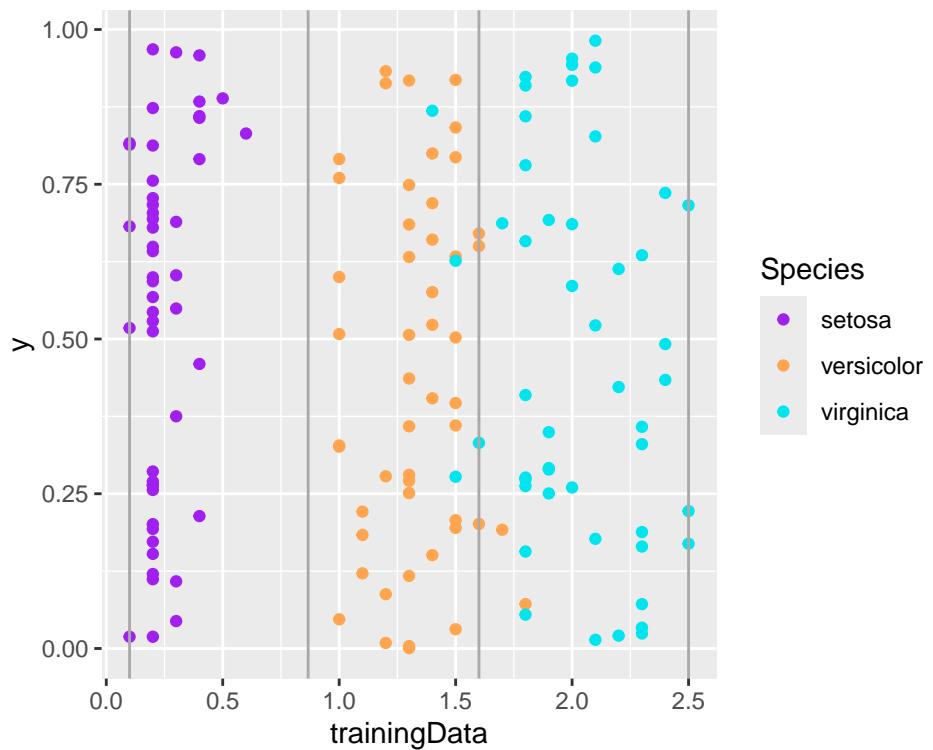


Na ich podstawie możemy uznać, że **Petal.Width** może stanowić najlepszy wyznacznik **gatunku** roślin. Najgorszym natomiast jest **Sepal.Width**. Ponieważ dla **Petal.Width** gatunki w najmniejszym stopniu się pokrywają ze względu na tą cechę, a w **Sepal.Width** w największym.

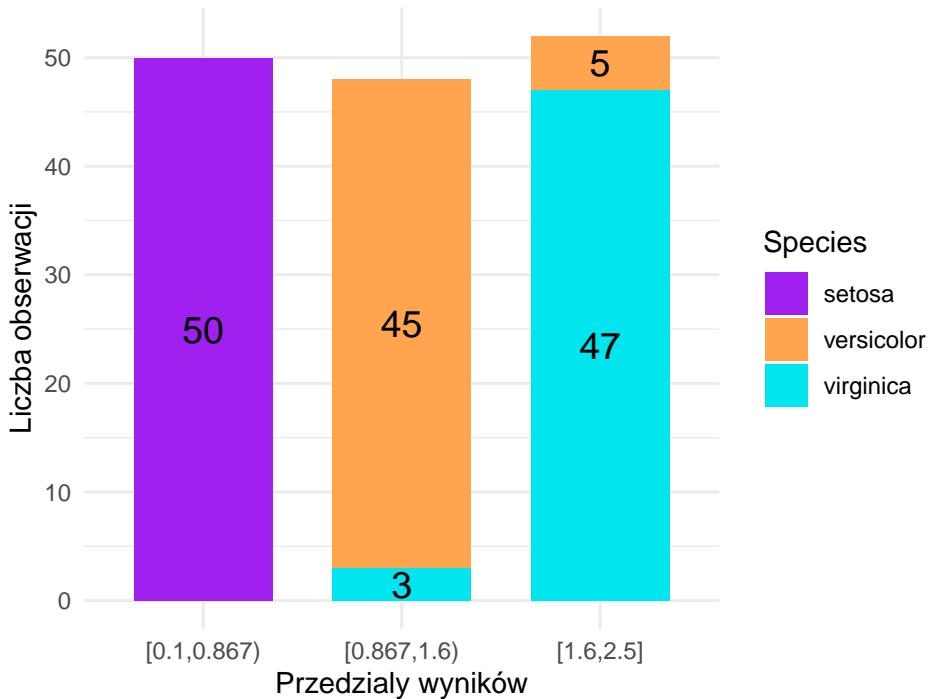
1.3 c) Porównanie nienadzorowanych metod dyskretyzacji

1.3.1 Metoda : Równe częstości(Frequency)

1.3.1.1 Dla najlepszej cechy : Petal.Length (Frequency)

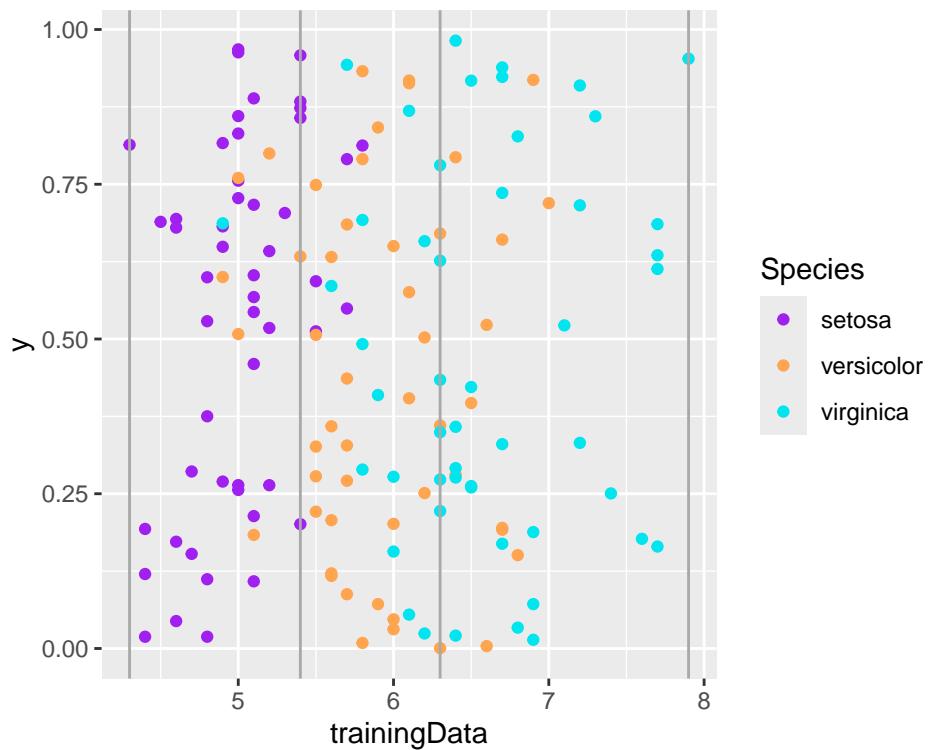


Liczba obserwacji gatunków w przedziałach

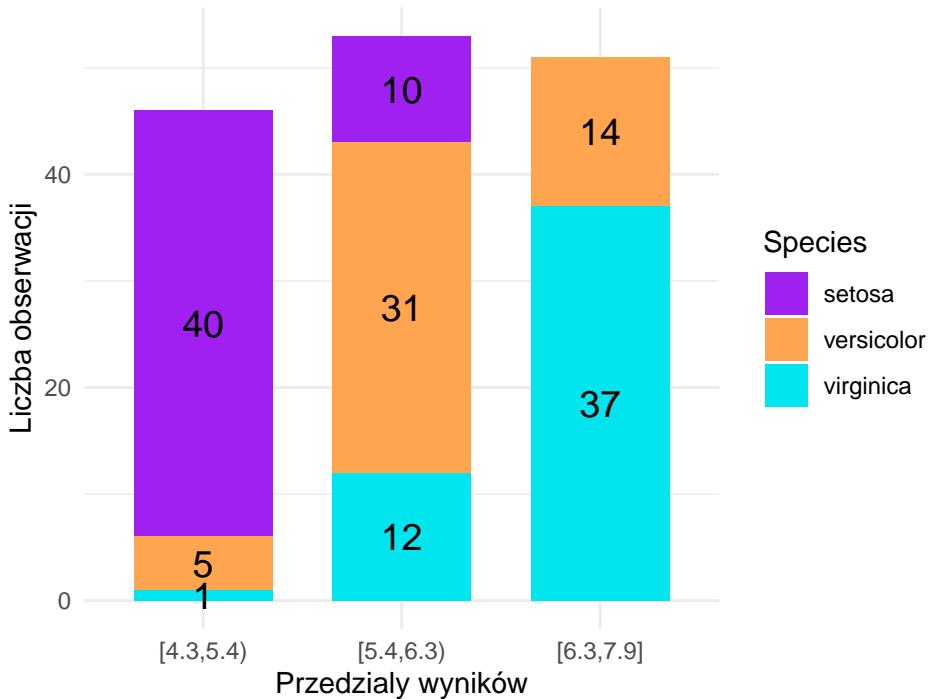


W przypadku tej metody **zgodność** uzyskanego grupowania z realnymi wartościami **wynosi** :

1.3.1.2 Dla najgorszej cechy : Sepal.Length (Frequency)



Liczba obserwacji gatunków w przedziałach

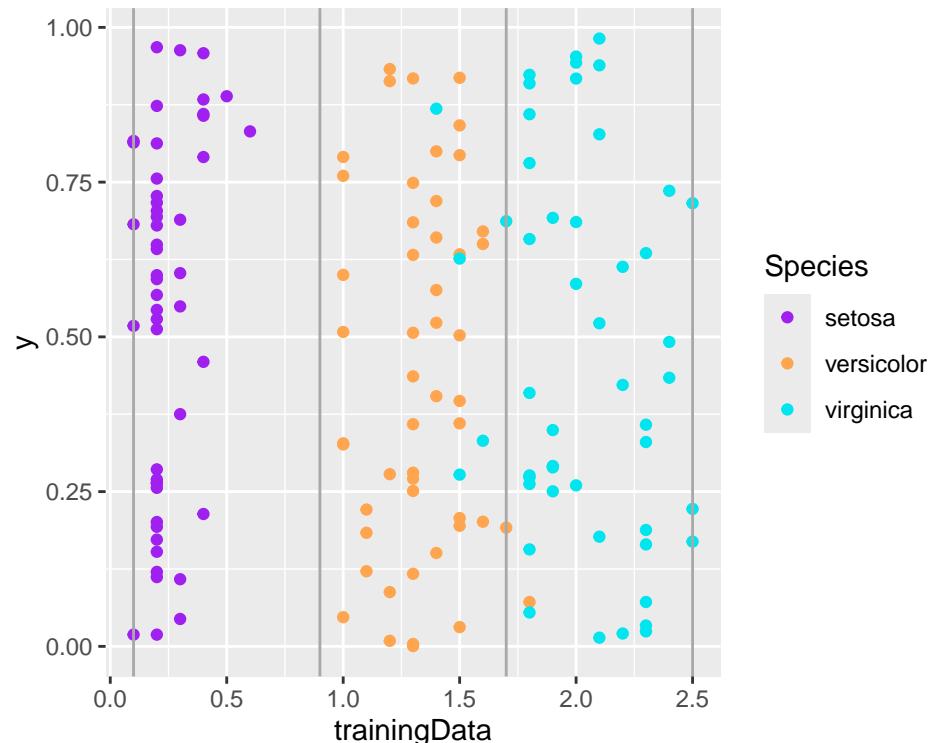


Zgodność dla nagjroszej cechy wynosi jedynie ok 72%, co mówi o znacznym spadku wiarygodności (**o ok 23 %**)

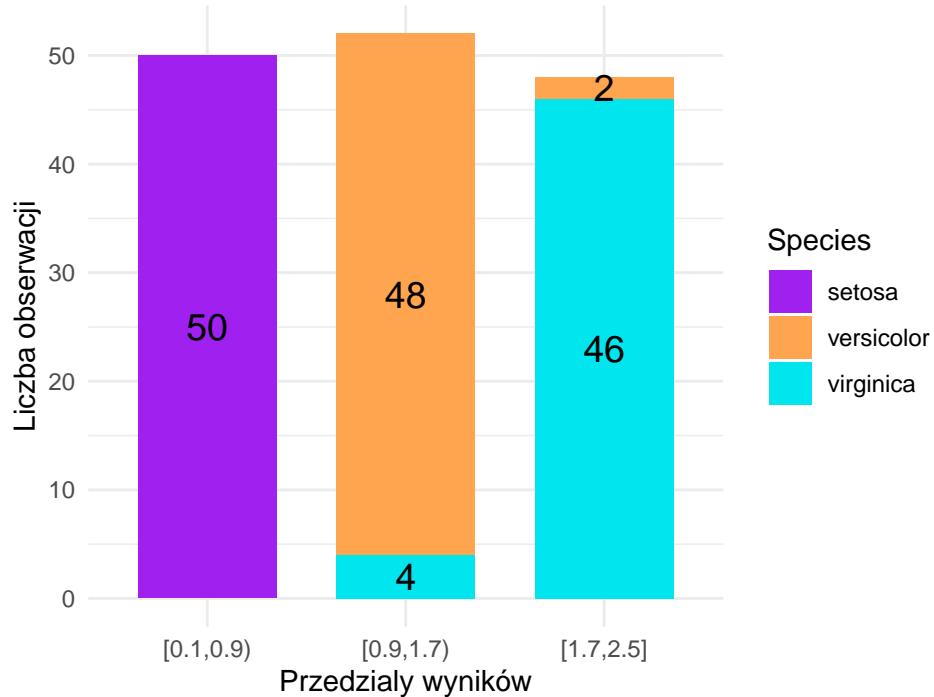
```
## [1] 0.72
```

1.3.2 Metoda : Równe szerokości (Interval)

1.3.2.1 Dla najlepszej cechy : Petal.Width (Interval)



Liczba obserwacji gatunków w przedziałach

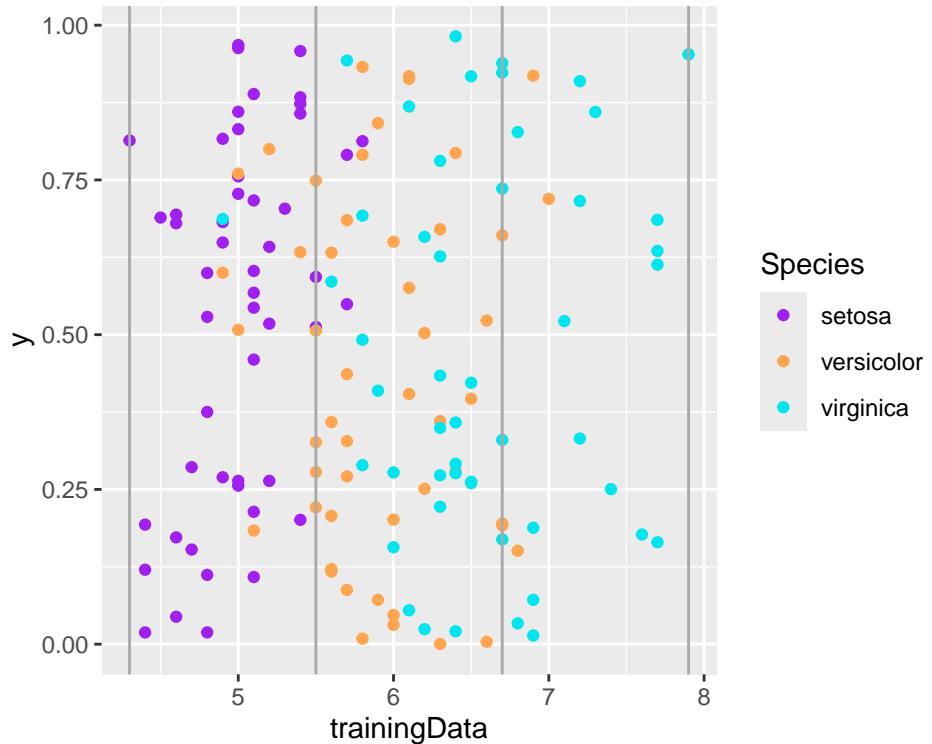


Dla tej metody również mamy zgodność na poziomie :

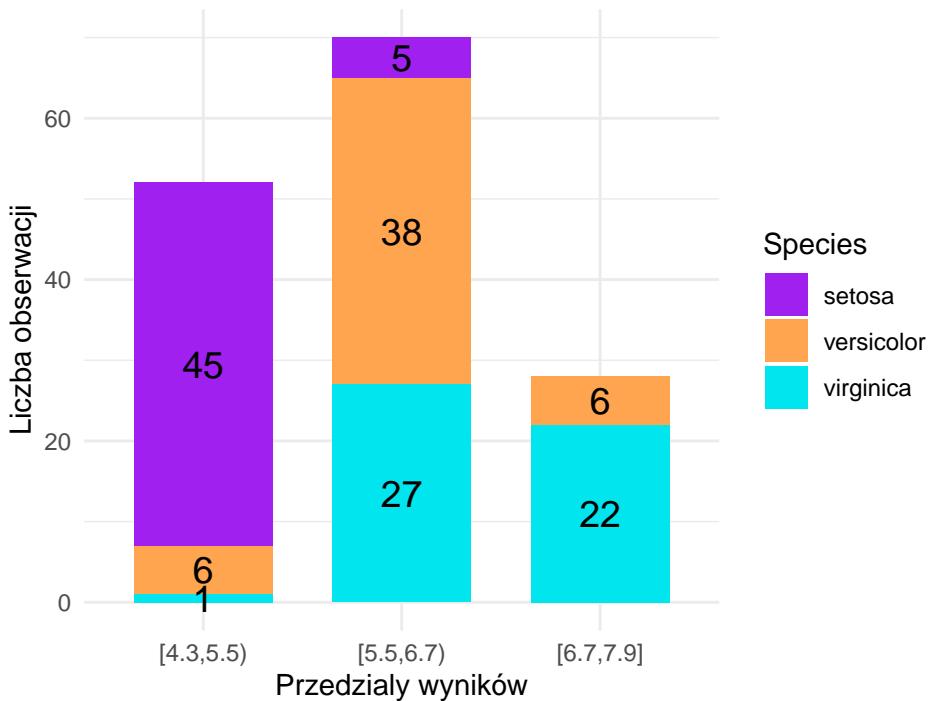
```
## [1] 0.96
```

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

1.3.2.2 Dla najgorszej cechy ; Sepal.Length (Interval)



Liczba obserwacji gatunków w przedziałach



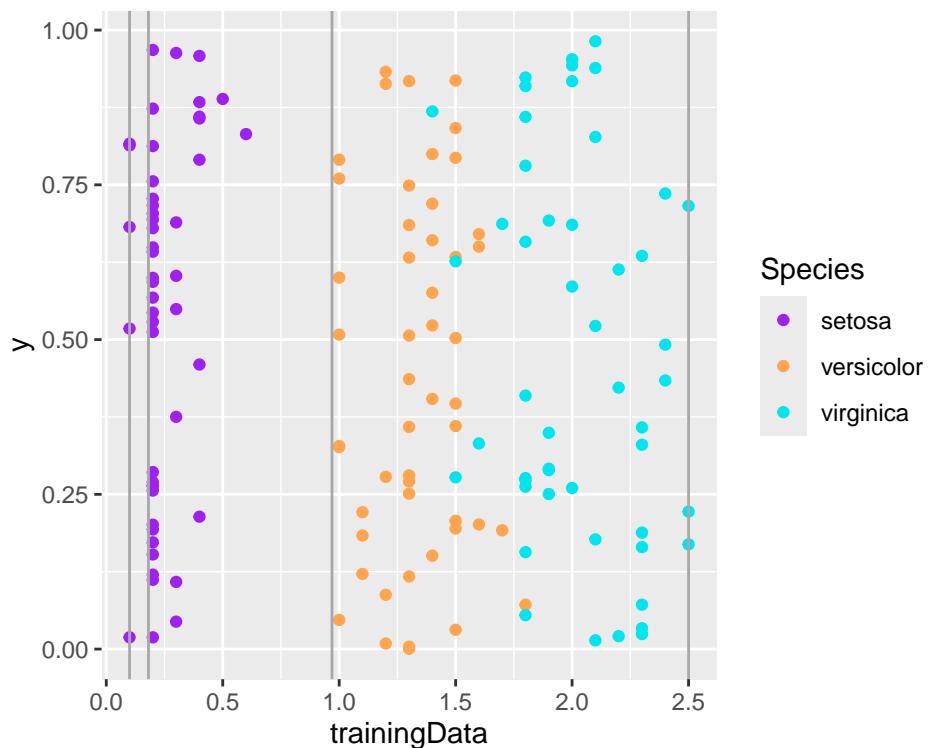
Metoda ta, dla najgorszej cechy dysryminuje ze zgodnością :

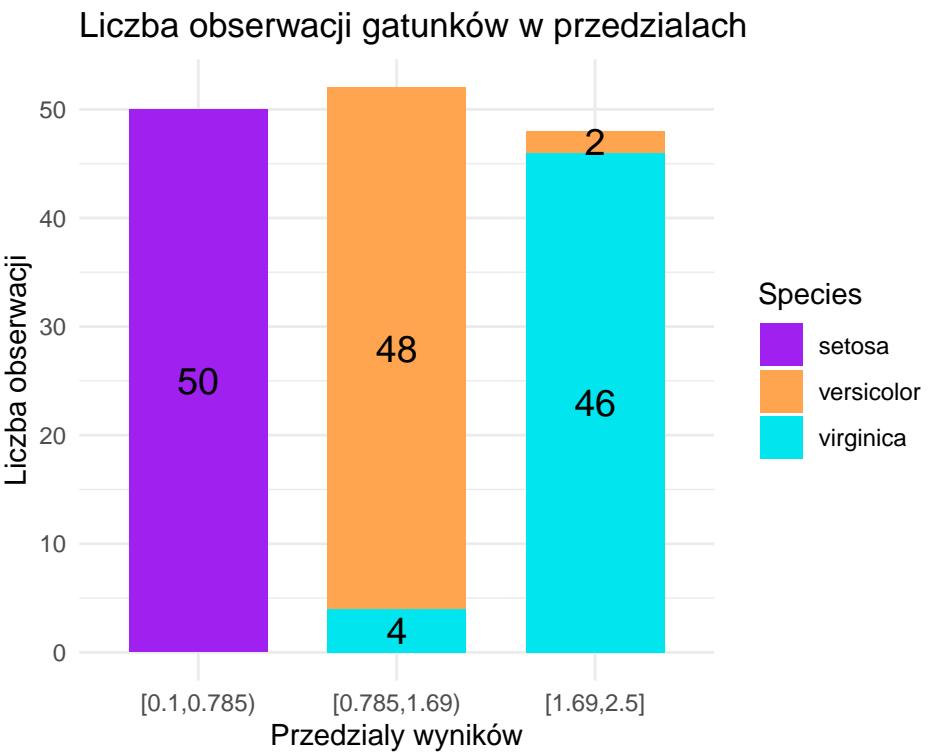
```
## [1] 0.5729167
```

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)



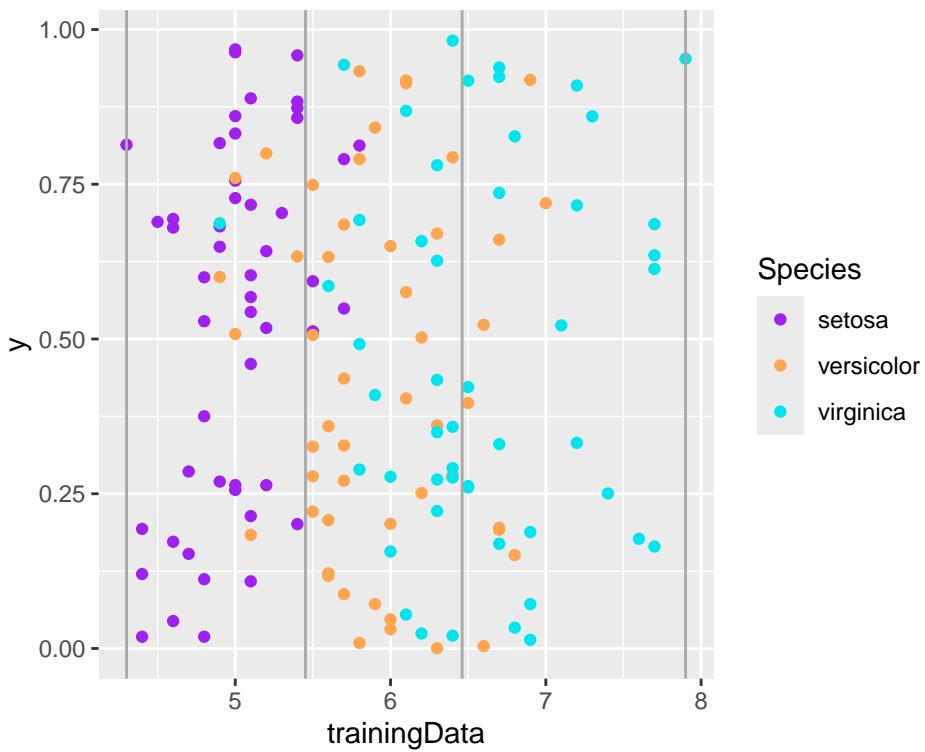


Zgodność na poziomie :

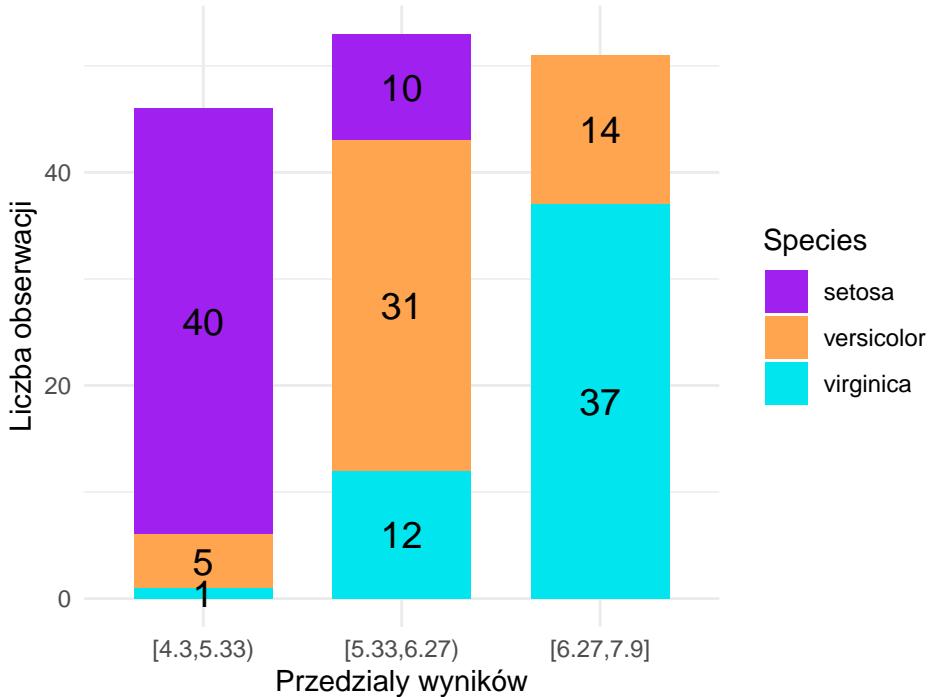
```
## [1] 0.96
```

Lepsza o ok 3% od ubiegłej metody

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



Liczba obserwacji gatunków w przedziałach



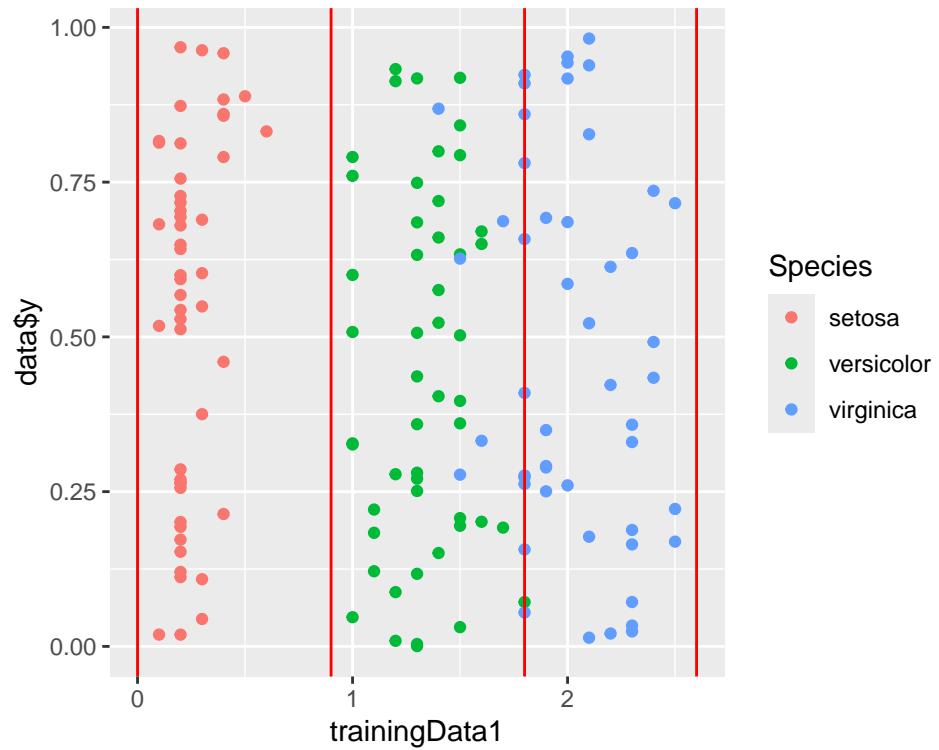
Dla najgorszej cechy mamy zgodność :

```
## [1] 0.5589744
```

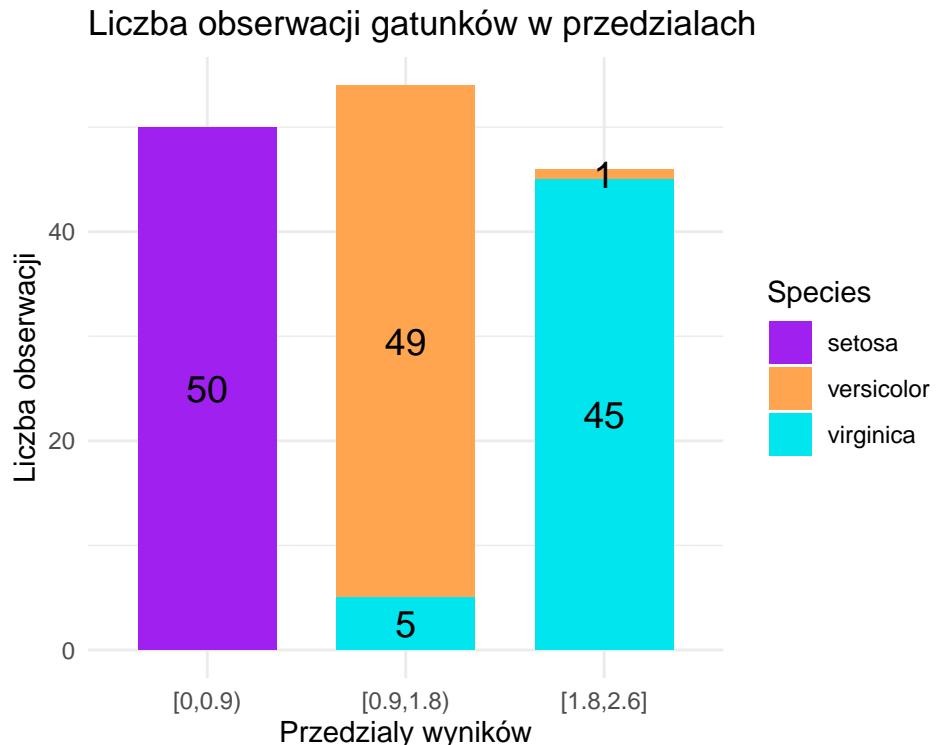
W tym przypadku jest ona na poziomie metody Frequency (gorsza o 1)

1.3.4 Dyskretyzacja z przedziałami zadanymi przez użytkownika (fixed)

1.3.4.1 Dla najlepszej cechy : Petal.Width (fixed)



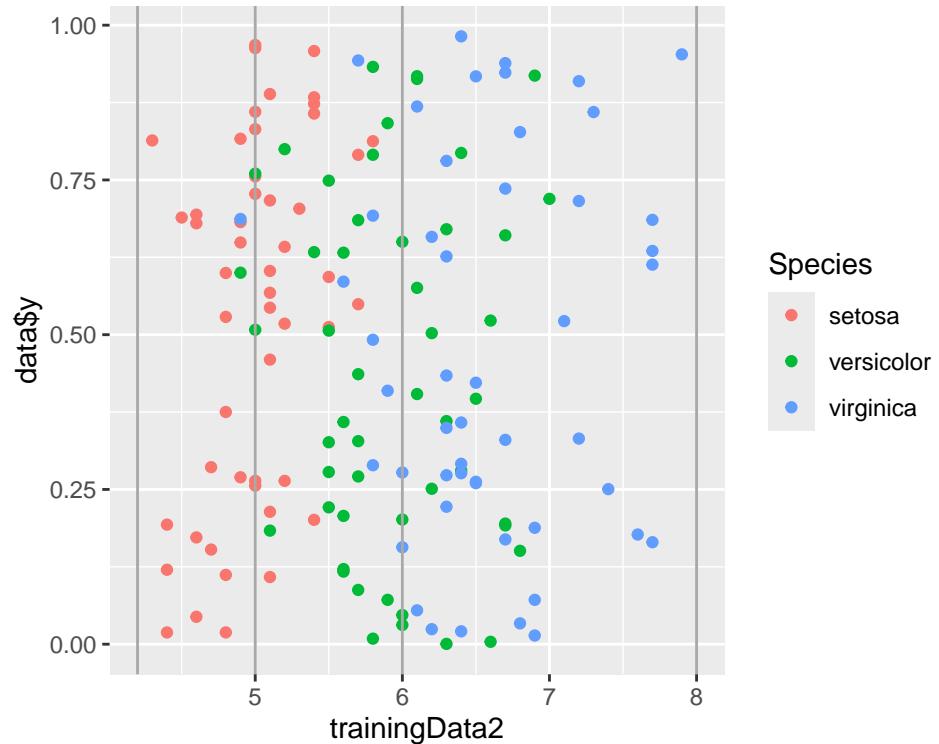
Na wykresie mamy zaznaczone też końce przedziałów, to jest potrzebne, co jest potrzebne podczas rysowania kolejnego wykresu



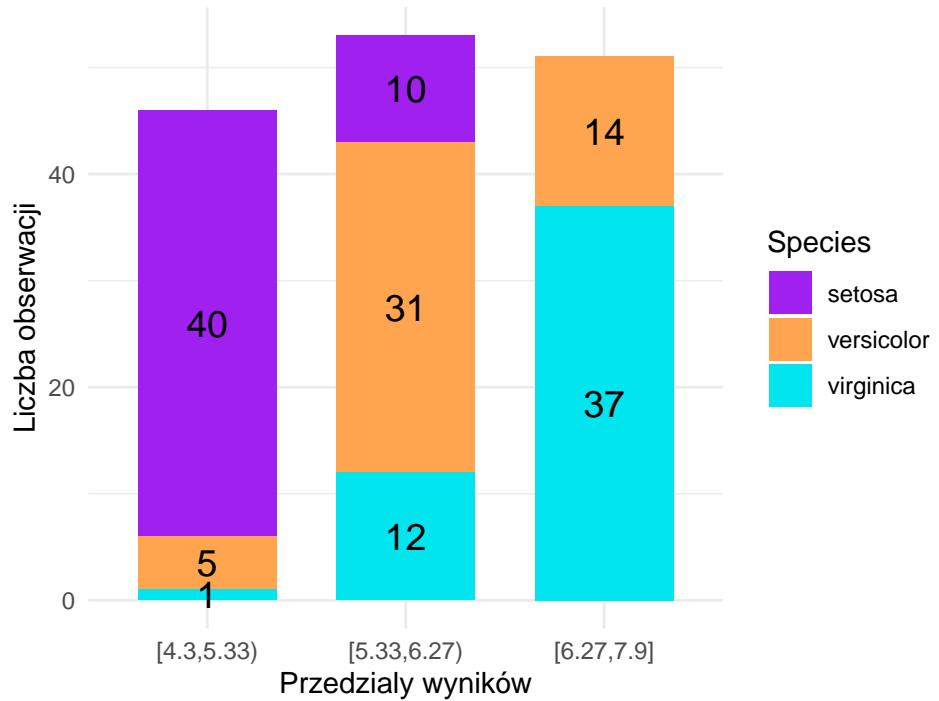
Zgodność na poziome poprzednich dwóch metod, wynosi :

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

1.3.4.2 Dla najgorszej cechy : Sepal.Length (fixed)



Liczba obserwacji gatunków w przedziałach



Dla cechy o najgorszej zdolności dyskryminacyjnej :

```
## [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	0.5589744	0.72

Na podstawie tabeli, dokładniej Porównania przyporządkować dla cech najgorszych i najlepszych pod względem dyskryminacji. Możemy wnioskować, że dla obecnych danych najlepszym algorytmem jest frequency(częstość) odznacza się najlepszym przyporządkowaniem zarówno dla Sepal.Length jak i Petal.Width

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

2.1 a) Przygotowanie danych

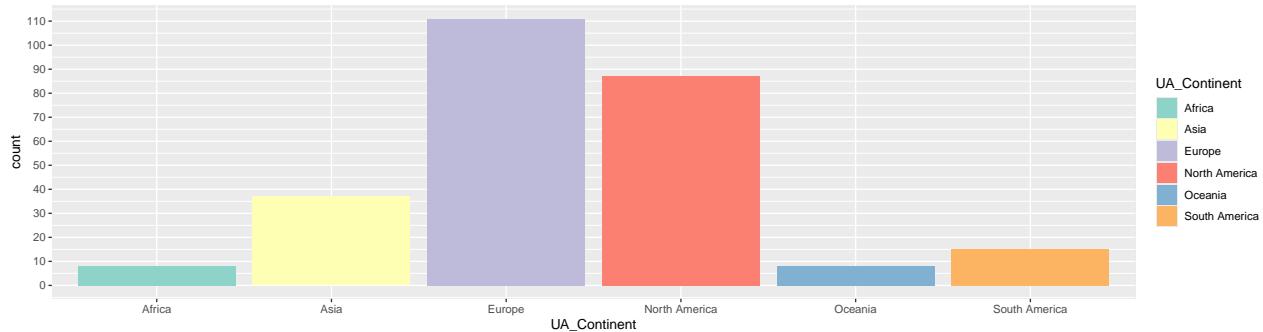
Tabela 2: Podstawowe informacje nt. danych
uaScoresDataFrame

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

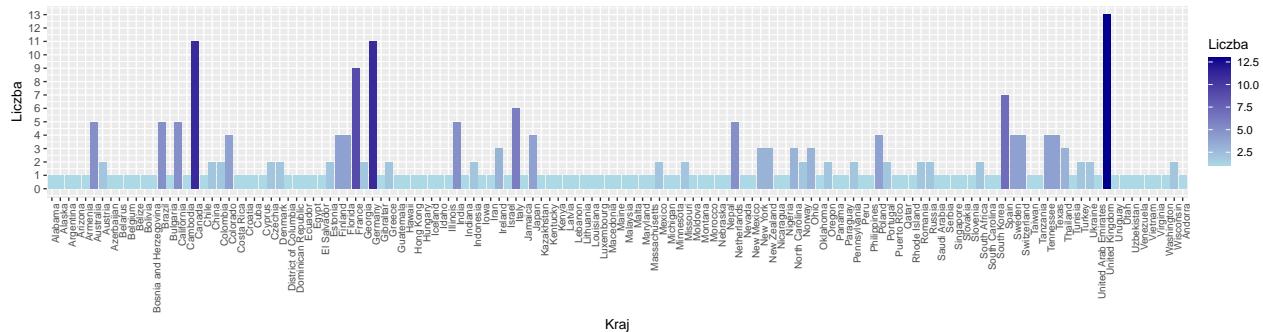
Typy danych w zbiorze



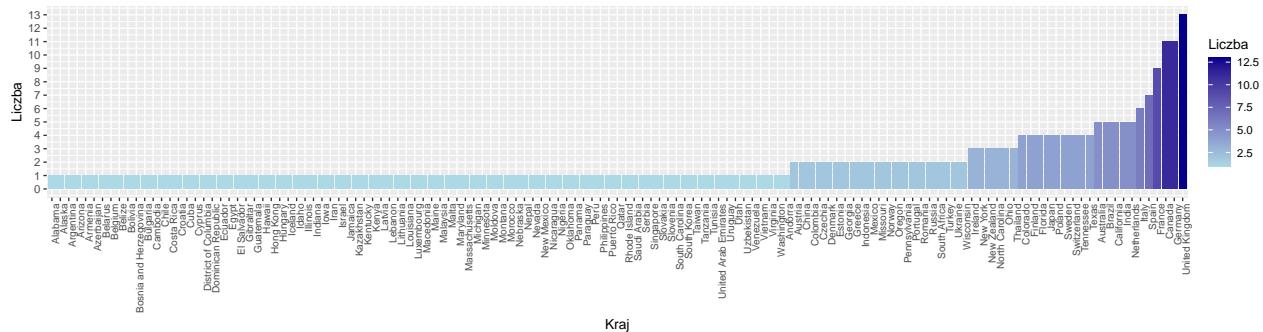
Wykres słupkowy dla UA_Continent



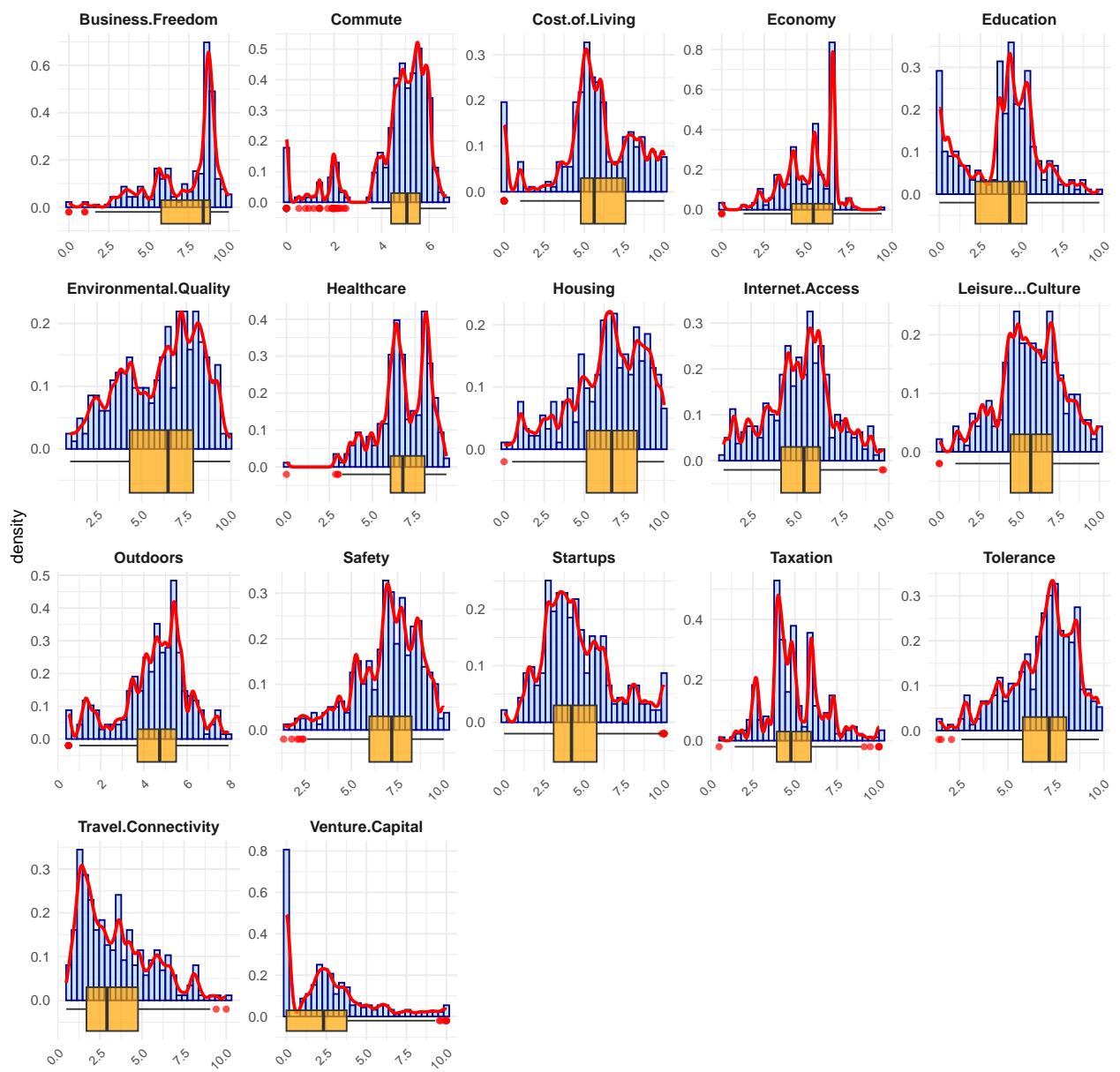
Wykres słupkowy dla UA_Country alfabetycznie



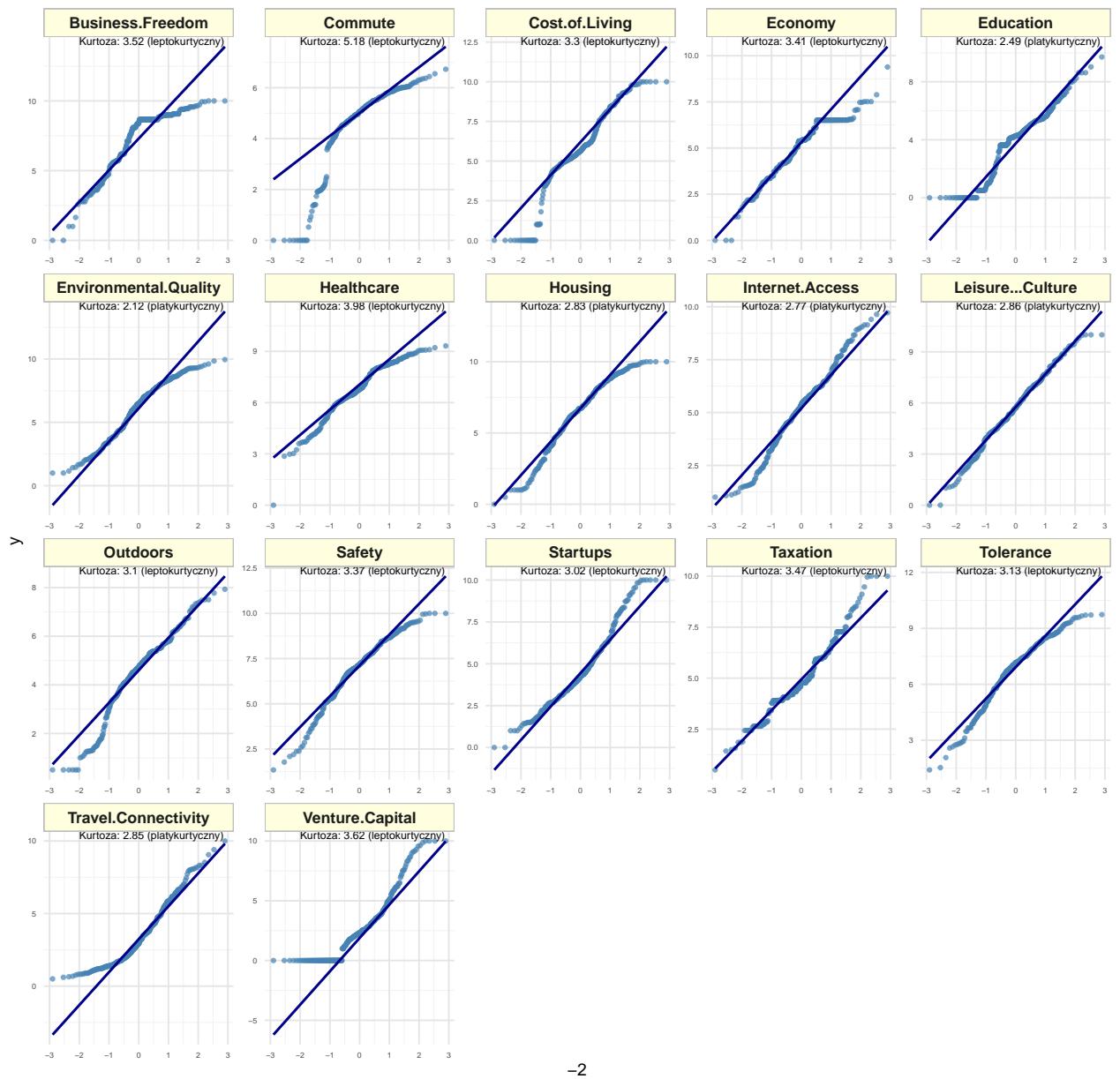
Wykres słupkowy dla UA_Country rosnaco



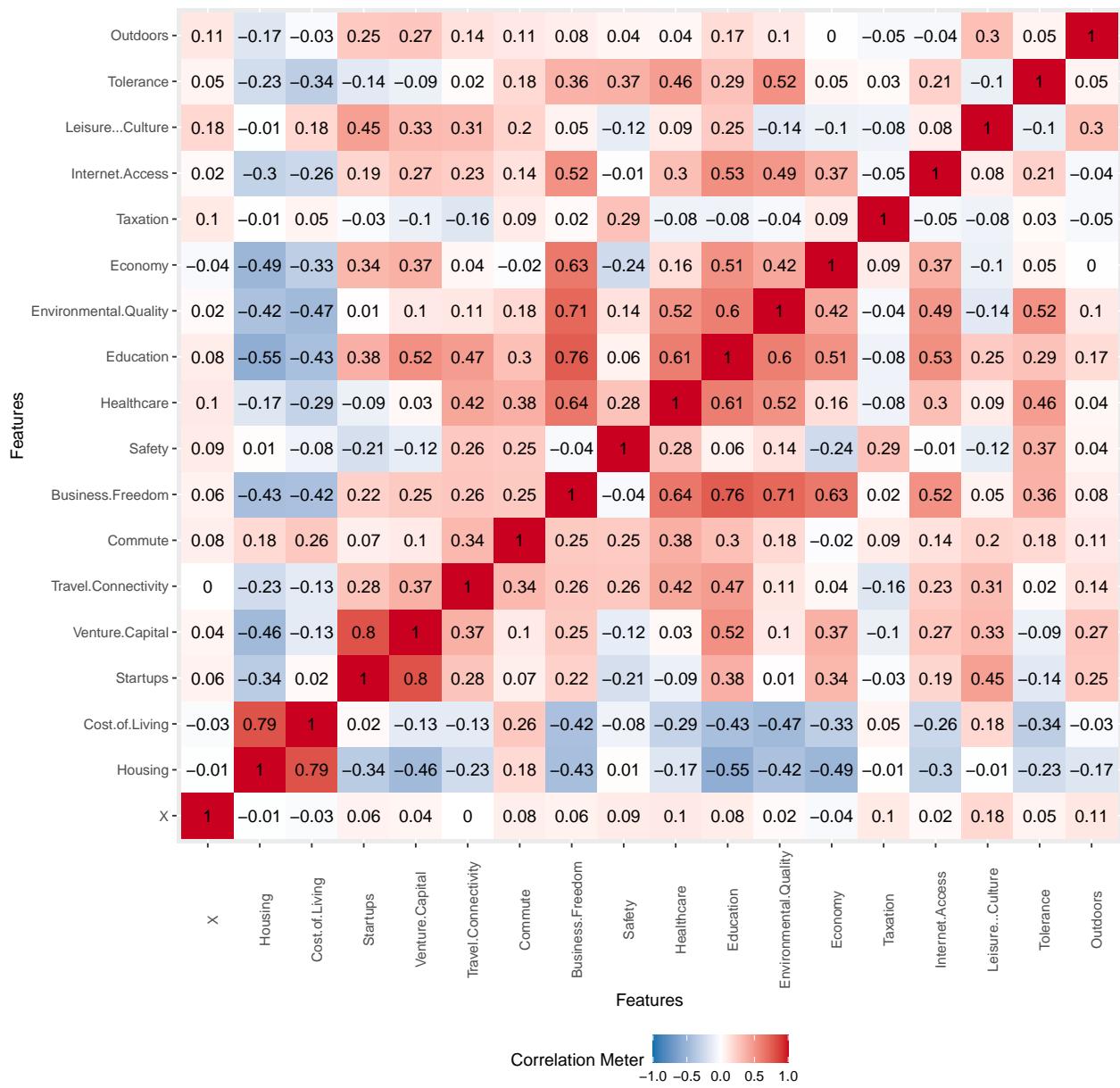
Histogramy z estymatorami gestosci i boxplotami dla zmiennych ilosciowych



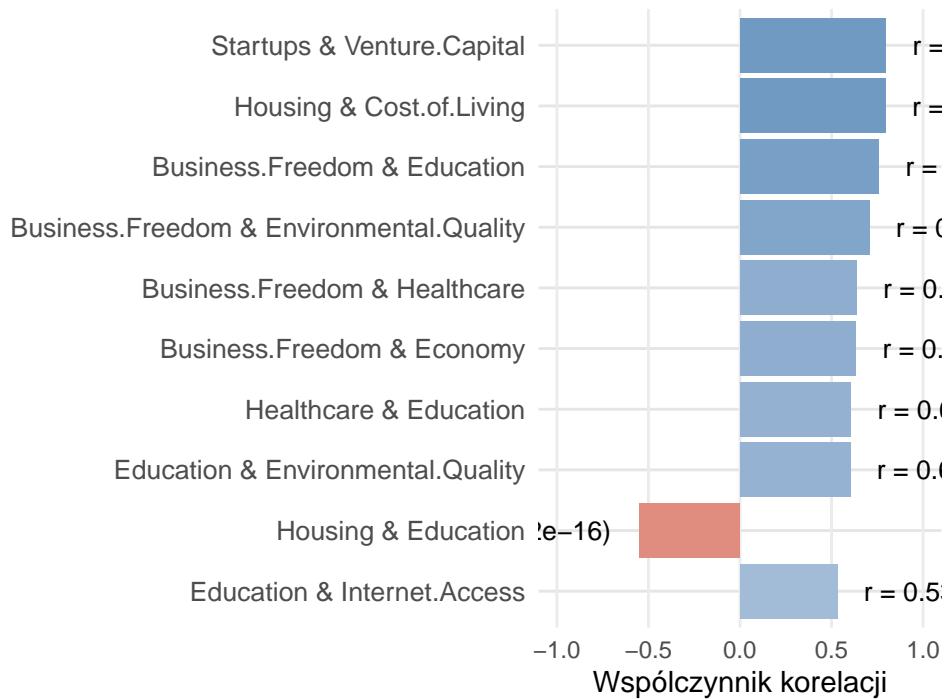
Wykresy Q-Q dla zmiennych ilościowych



Macierz korelacji dla zmiennych ciągkich



Najśilniejsze istotne korelacje (1)



X	UA_Name	UA_Country	UA_Continent	Housing	Cost.of.Living
0	Aarhus	Denmark	Europe	6.132	4.015
1	Adelaide	Australia	Oceania	6.310	4.692
2	Albuquerque	New Mexico	North America	7.262	6.059
3	Almaty	Kazakhstan	Asia	9.282	9.333
4	Amsterdam	Netherlands	Europe	3.053	3.824
5	Anchorage	Alaska	North America	5.434	3.141

X	Startups	Venture.Capital	Travel.Connectivity	Commute	Business.Freedom
0	2.827	2.512	3.536	6.312	9.940
1	3.136	2.640	1.777	5.336	9.400
2	3.772	1.493	1.456	5.056	8.671
3	2.458	0.000	4.592	5.871	5.568
4	7.972	6.107	8.325	6.118	8.837
5	2.795	0.000	1.738	4.715	8.671

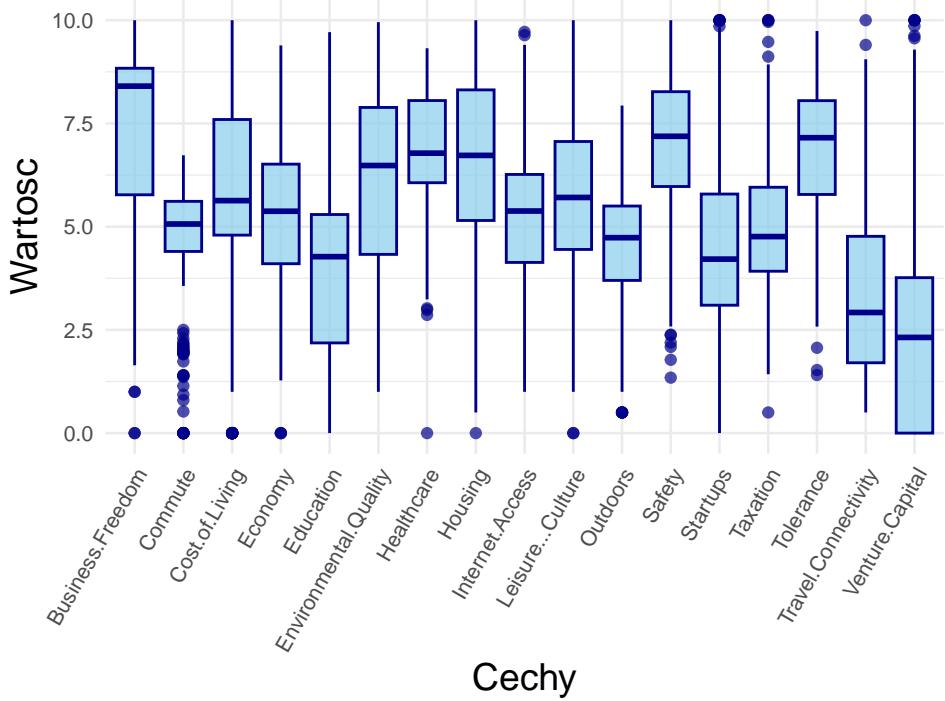
X	Safety	Healthcare	Education	Environmental.Quality	Economy
0	9.617	8.704	5.367	7.633	4.887
1	7.926	7.937	5.142	8.331	6.070
2	1.343	6.430	4.152	7.319	6.514
3	7.309	4.546	2.283	3.857	5.269

X	Safety	Healthcare	Education	Environmental.Quality	Economy
4	8.504	7.907	6.180	7.597	5.053
5	3.470	6.060	3.624	9.272	6.514

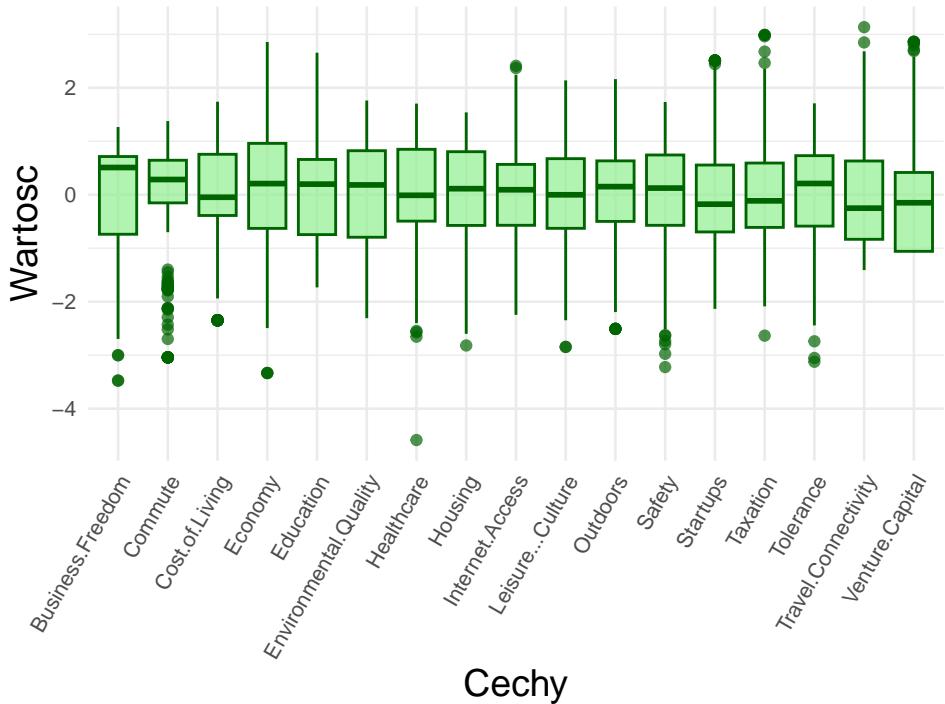
X	Taxation	Internet.Access	Leisure... Culture	Tolerance	Outdoors
0	5.068	8.373	3.187	9.739	4.130
1	4.588	4.341	4.328	7.822	5.531
2	4.346	5.396	4.890	7.028	3.515
3	8.522	2.886	2.937	6.540	5.500
4	4.955	4.523	8.874	8.368	5.307
5	4.772	4.964	3.266	7.093	5.358

	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
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

Rozkład cech ilościowych przed standaryzacją



Rozkład cech ilościowych po standaryzacji



2.2 b) Wyznaczenie składowych głównych

Tabela 8: Podsumowanie analizy 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

2.3 c) Zmienna odpowiadająca poszczególnym składowym

Rozkład wartości składowych głównych

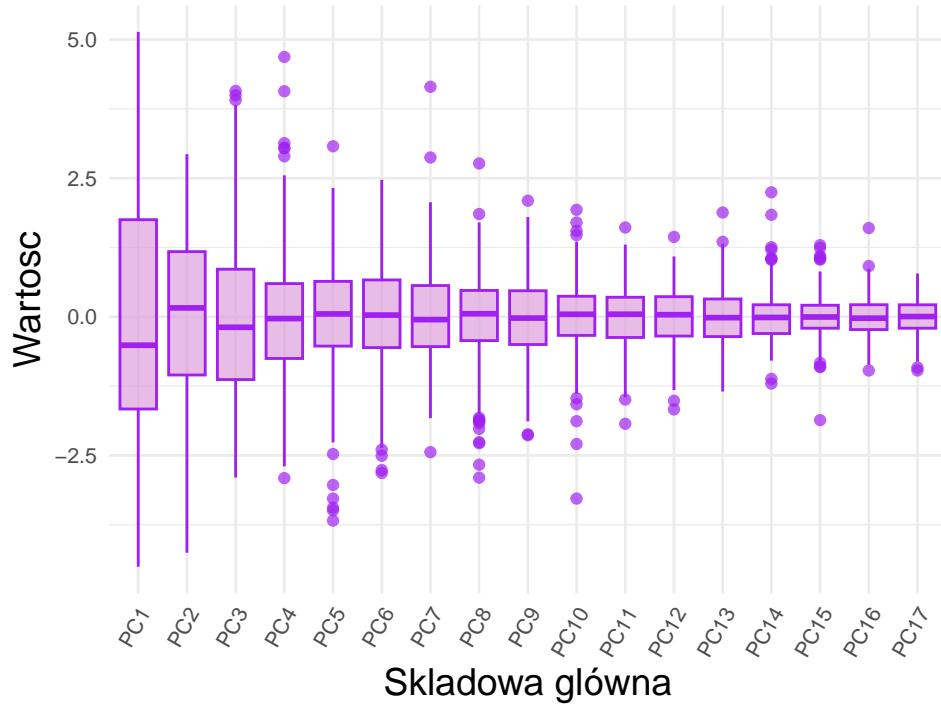
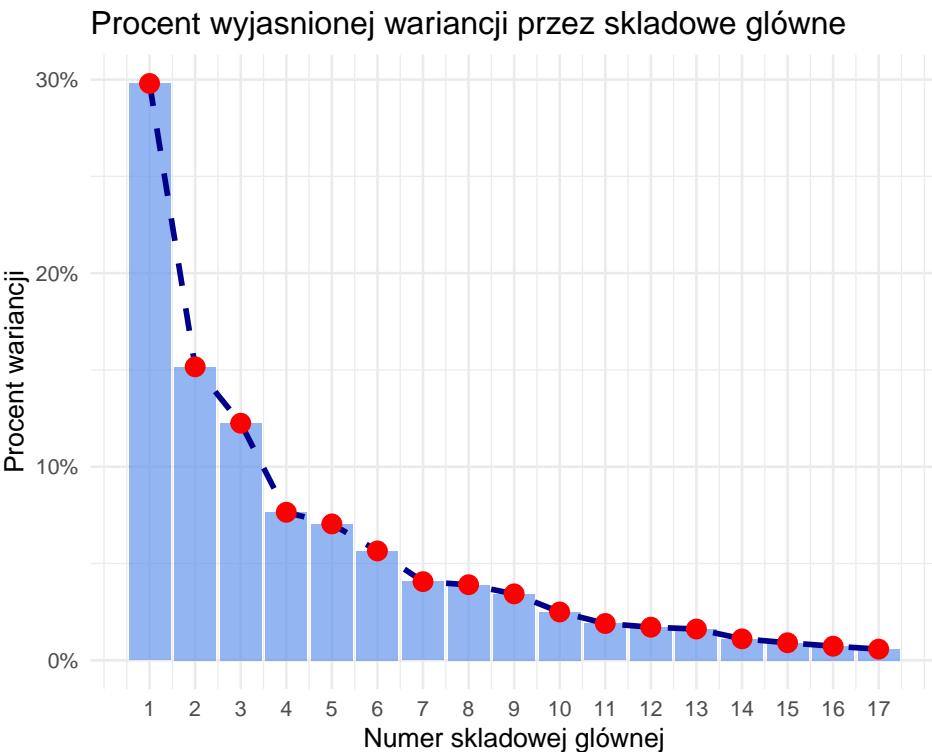
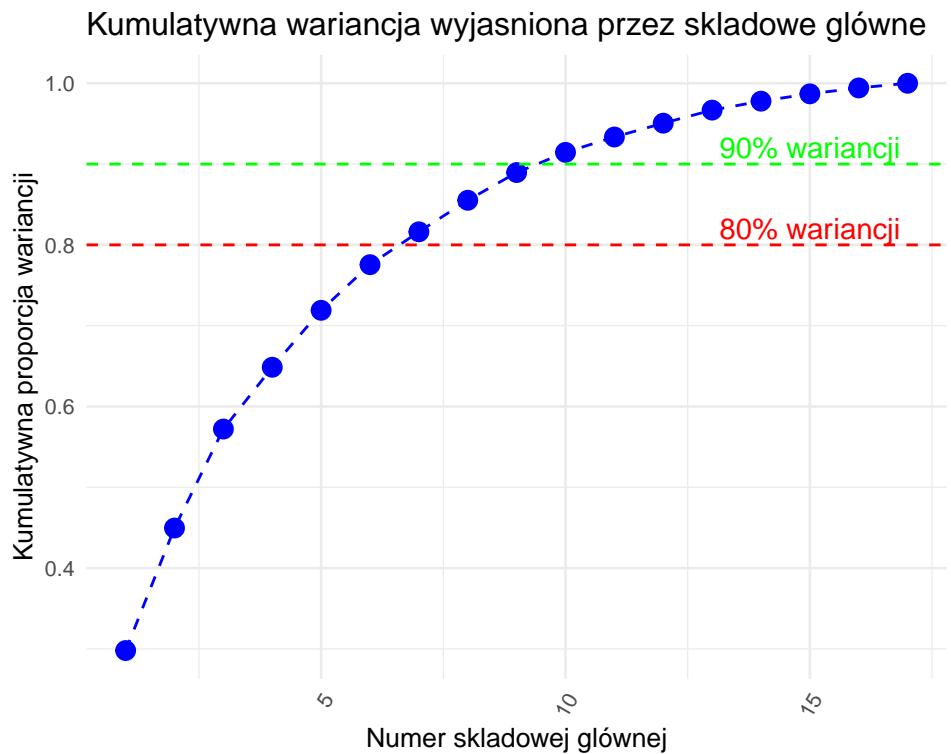


Tabela 9: Wektory ładunków dla PC1, PC2 i PC3

	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





Liczba składowych głównych wyjaśniających **80%** wariacji: **7**

Liczba składowych głównych wyjaśniających **90%** wariacji: **10**

2.4 d) Wizualizacja danych wielowymiarowych

2.5 e) Korelacja zmiennych

2.6 d) Wizualizacja danych wielowymiarowych

2.7 e) Korelacja zmiennych

2.8 f) Końcowe wnioski

2.9 f) Końcowe wnioski

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).

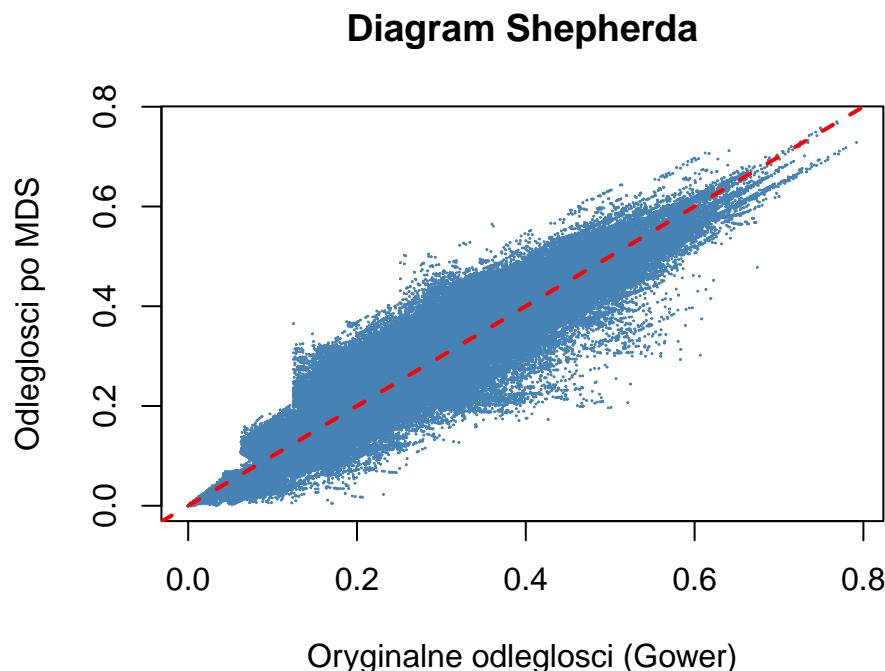
3.2 b) Przygotowanie danych

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



Widać, że odległości w nowej przestrzeni danych zmieniły się, nadal mają podobny charakter, największe skupisko danych znajduje się przy linii $x = y$ (w gdyby odległości zostały zachwoane, czyli w idealnym scenariuszu, to nasze nowe odległości przechodziły by właśnie przez tą linię).

Pomimo, że duża liczba punktów odbiega od linii $x = y$, to jednak różnica między ich nową odlegością a starą nie są duże (nie mamy znaczących rozrzutów na osi OY), więc możemy uznać skalowanie za dość dobre, chyba, że chcemy przeprowadzać bardzo dokładną analizę

3.4 d) Wizualizacja danych