Grupowanie gatunków nasion z wykorzystaniem metody k-means

Autor:
Radosław Zduńczyk

Zbiór danych:
Seed_Data.csv (UCI dataset)

Typ danych:
Numeryczne, wymiary geometryczne nasion pszenicy

Metoda ewaluacji:
Procentowy udział poszczególnych gatunków w klastrach

	A	P	C	LK	WK	A_Coef	LKG	target
0	15.26	14.84	0.8710	5.763	3.312	2.221	5.220	0
1	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	0
2	14.29	14.09	0.9050	5.291	3.337	2.699	4.825	0
3	13.84	13.94	0.8955	5.324	3.379	2.259	4.805	0
4	16.14	14.99	0.9034	5.658	3.562	1.355	5.175	0

210 x 8

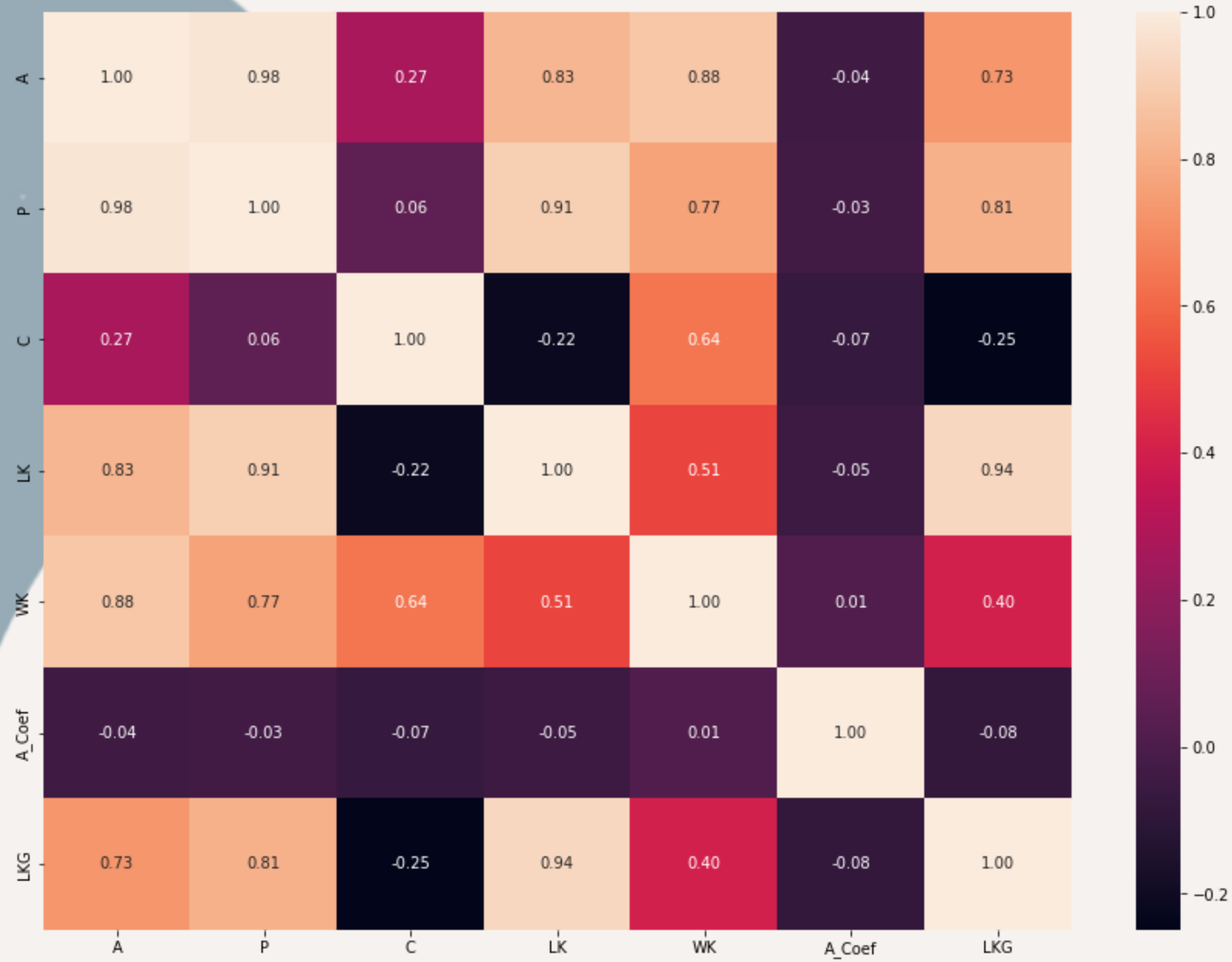
Faza 3 i 4

Ocena danych:

- **Dane są kompletne**
- **Nie ma nietypowych wartości**
- **Możemy użyć danych do rozwiązania zagadnienia**

Przetwarzanie danych:

- **Usunięcie kolumn 'A_coef' i 'C'**
- **Normalizacja danych metodą min-max**

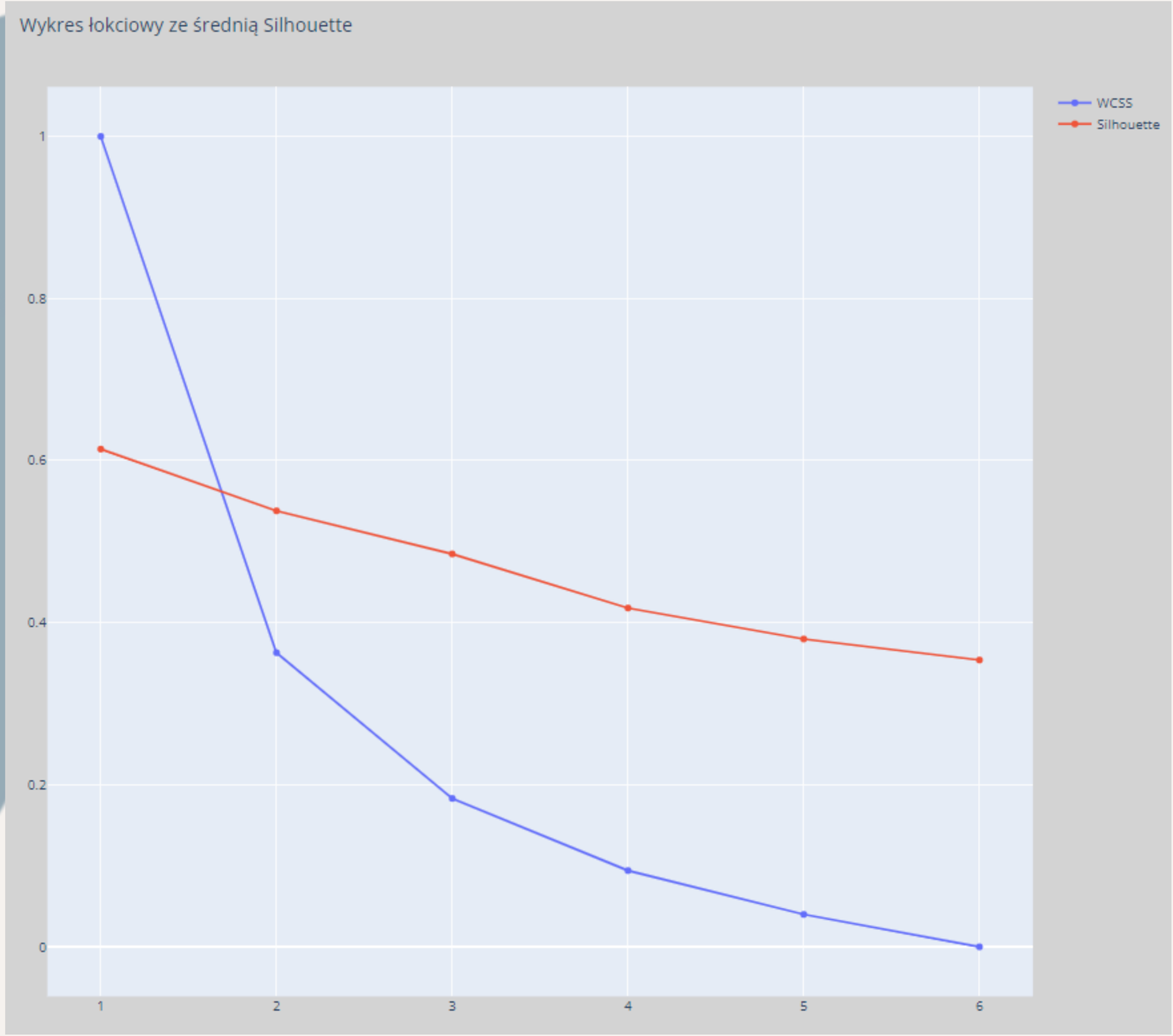


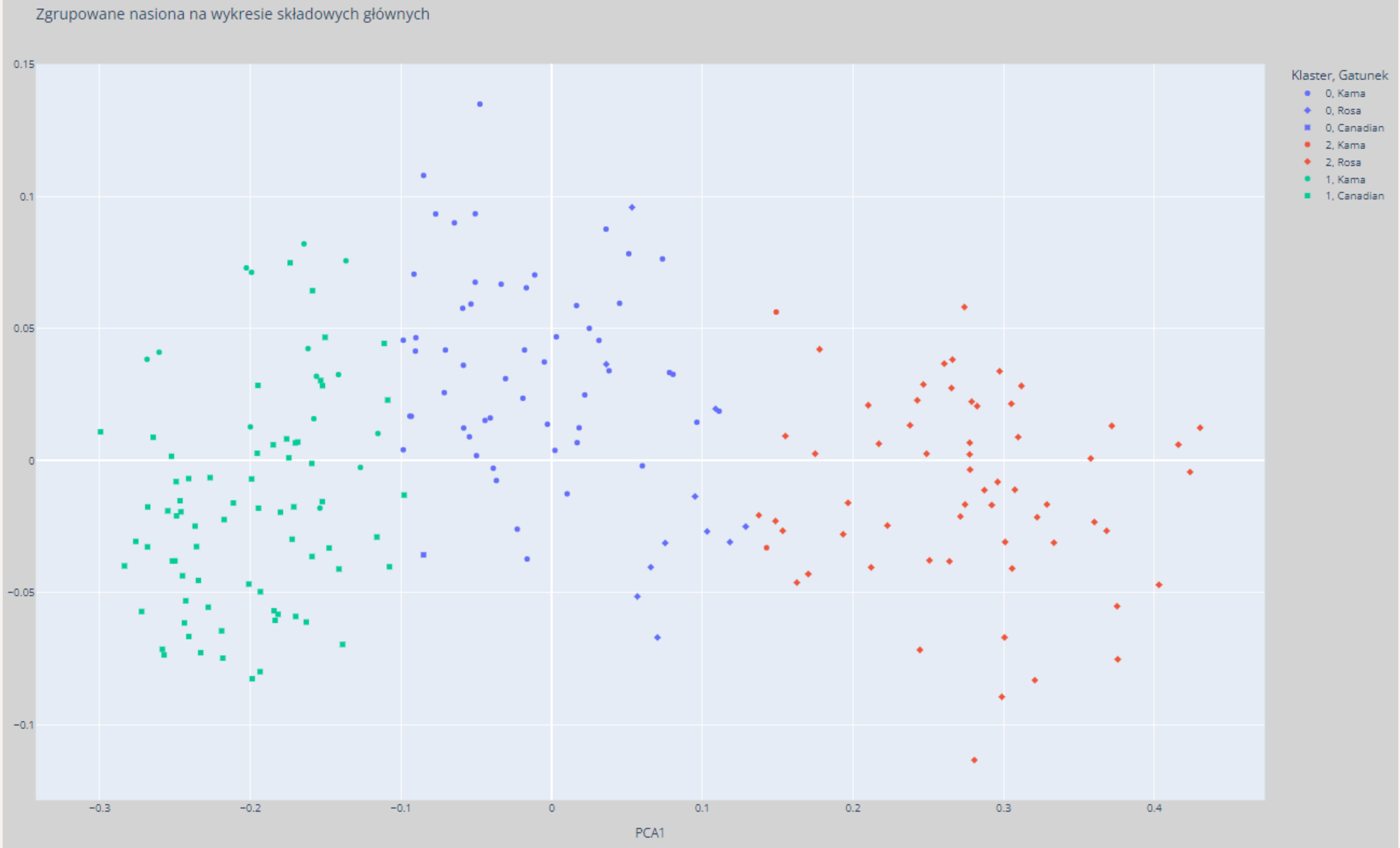
Macierz korelacji zmiennych

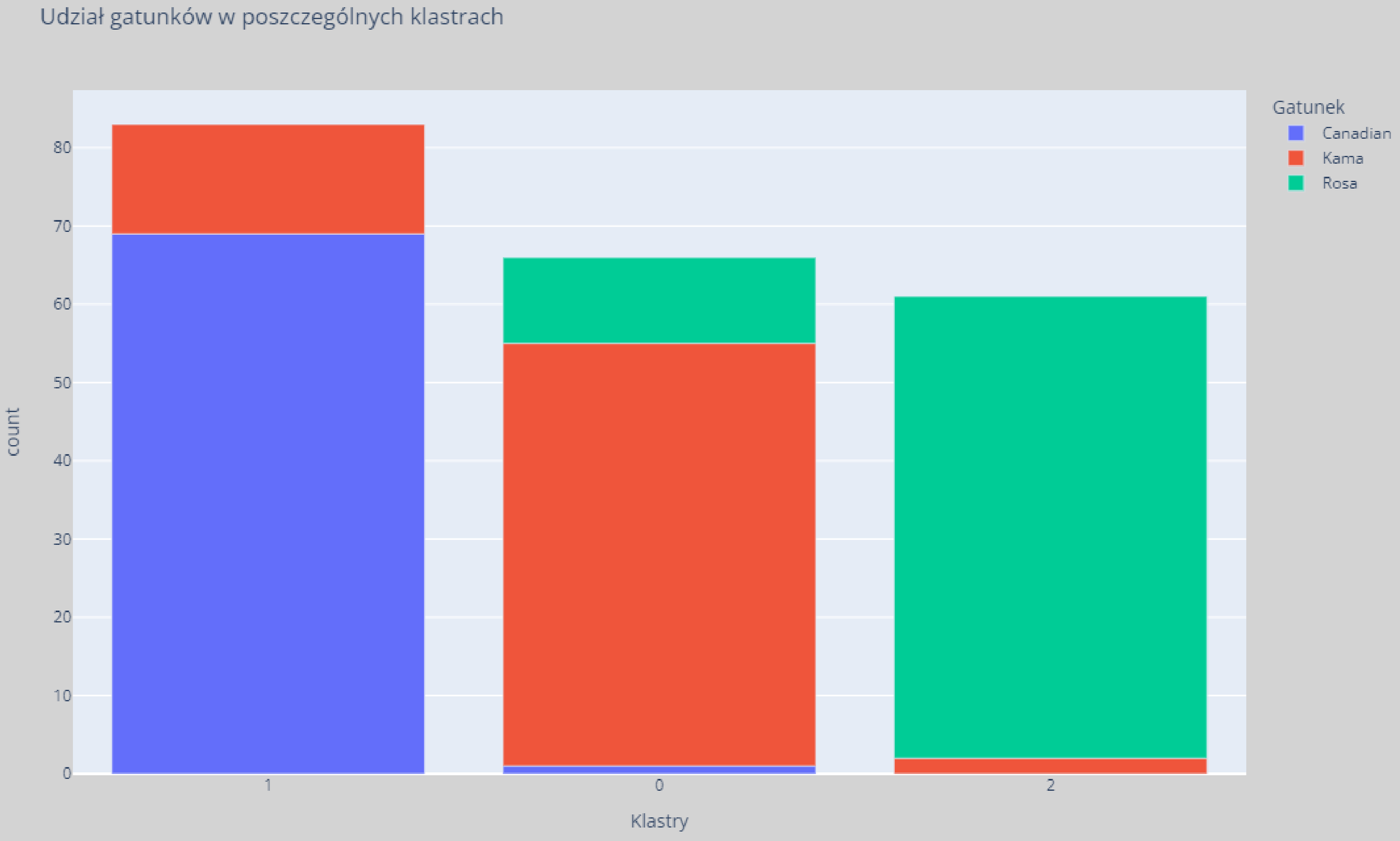
Faza 5

Dobór liczby klastrów:

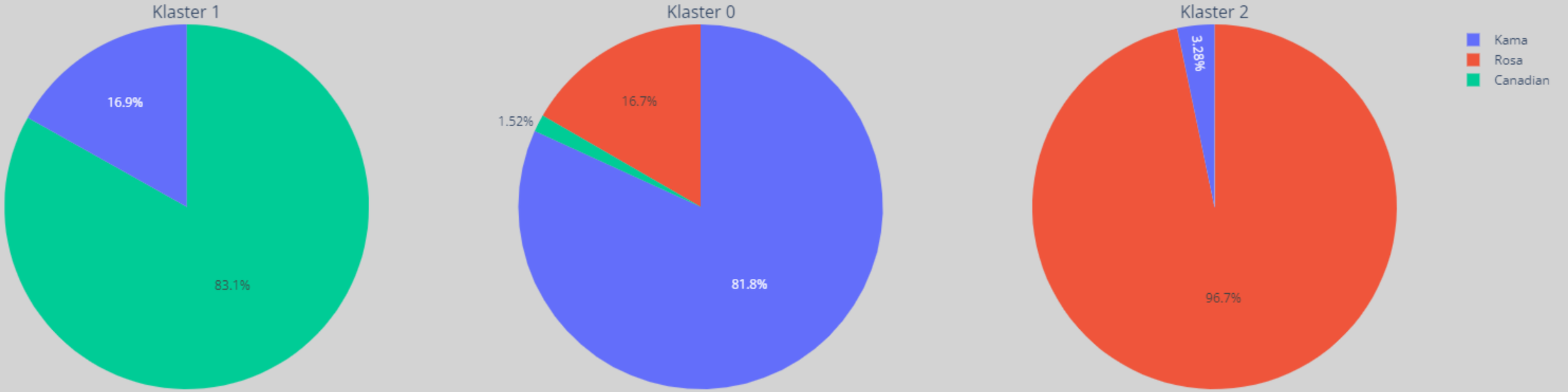
Na podstawie wykresu łokciowego oraz informacji o trzech różnych gatunkach pszenicy w zbiorze danych, przyjęto liczbę klastrów k=3




Zgrupowane nasiona na wykresie składowych głównych
Klaster, Gatunek
0, Kama
0, Rosa
0, Canadian
2, Kama
2, Rosa
1, Kama
1, Canadian
0.15
0.1
0.05
0
−0.05
−0.1
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
PCA1


Udział gatunków w poszczególnych klastrach
Gatunek
Canadian
Kama
Rosa
count
80
70
60
50
40
30
20
10
0
1
0
2
Klastry

Procentowy udział gatunków w poszczególnych klastrach



Obserwacje i wnioski

1. Nasiona gatunku Rosa są najmniej podobne geometrycznie do innych nasion.
2. Najwyższą czystość klastra osiągnięto w przypadku gdzie dominował gatunek Rosa.
3. Klaster z przewagą nasion gatunku Kama jako jedyny zanieczyszczony był dwoma innymi gatunkami.
4. Grupowanie metodą k-means dobrze sprawdziło się do realizacji celu eksperymentu.

Dziękuję za uwagę!