Przedmiot: Eksploracja danych

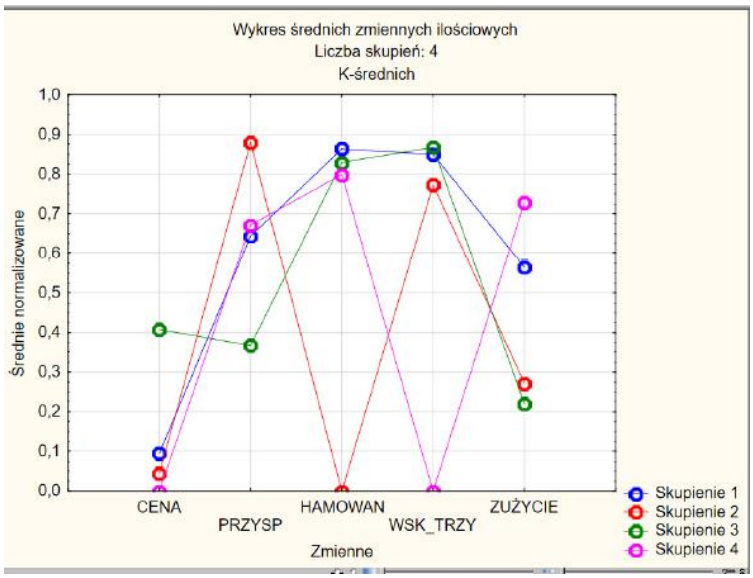
Kierunek: Informatyka – Data Science

Ćwiczenie: Analiza skupień - Cars

Autor: Bartłomiej Jamiołkowski

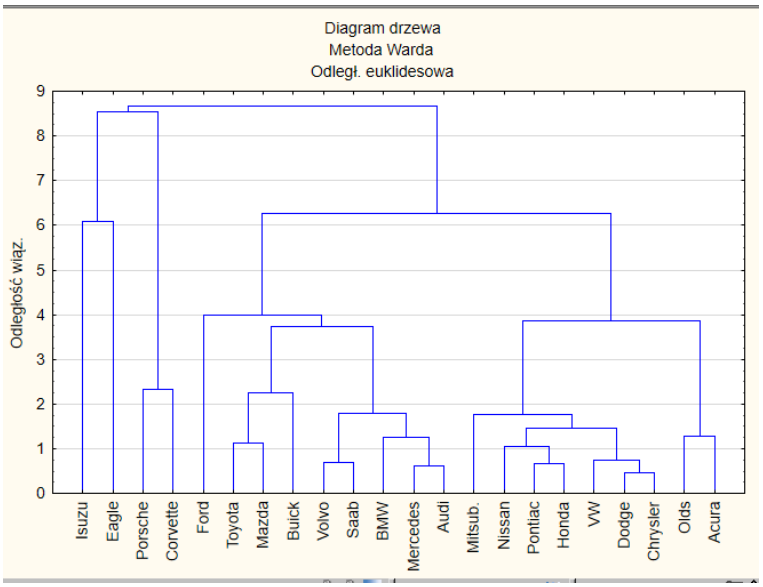
Ad 5)

Analiza skupień uogólnioną metodą k-średnich.



Przypadek	Elementy skupienia (Cars) Liczba skupień: 4 Całkowita liczba przypadków uczących: 22 Nr przypadku	Wynikowa klasyfikacja	CENA	PRZYSP	HAMOWAN	WSK_TRZY	ZUŻYCIE	Odległość od
Acura	1	1	-0,521072	0,47725	-0,00657	0,38162	2,07875	0
Audi	2	3	0,865652	0,20803	0,31870	-0,09137	-0,67706	0
BMW	3	3	0,495859	-0,80154	0,19220	-0,09137	-0,15381	0
Buick	4	1	-0,613521	1,68874	0,93309	-0,20962	-0,15381	0
Corvette	5	3	1,235446	-1,81111	-0,49447	0,97286	-0,67706	0
Chrysler	6	1	-0,613521	0,07342	0,42712	-0,20962	-0,15381	0
Dodge	7	1	-0,705969	-0,19580	0,48133	0,14512	-0,15381	0
Eagle	8	2	-0,613521	1,21761	-4,19889	-0,20962	-0,67706	0
Ford	9	3	-0,705969	-1,54189	0,98730	0,14512	-1,72357	0
Honda	10	1	-0,428624	0,40995	-0,00657	0,02687	0,36945	0
Isuzu	11	4	-0,798417	0,40995	-0,06078	-4,23006	1,06713	0
Mazda	12	3	0,126066	0,67917	-0,13306	0,49987	-1,72357	0
Mercedes	13	3	1,050549	0,00612	0,11992	-0,09137	-0,15381	0
Mitsub.	14	1	-0,613521	-1,00345	0,08378	0,38162	0,71829	0
Nissan	15	1	-0,428624	0,07342	-0,00657	0,26337	0,99736	0
Olds	16	1	-0,613521	-0,73423	0,40905	0,38162	2,11364	0
Pontiac	17	1	-0,613521	0,67917	0,53554	0,14512	0,19503	0
Porsche	18	3	3,454206	-2,21494	-0,29570	0,61812	-1,02590	0
Saab	19	1	0,588308	0,67917	0,24641	0,26337	0,02061	0
Toyota	20	1	-0,058831	1,21761	0,22834	0,73636	-0,85148	0
VW	21	1	-0,705969	-0,12849	0,10185	0,38162	0,19503	0
Volvo	22	1	0,218514	0,61186	0,13799	-0,20962	0,36945	0

Aglomeracja z wykorzystaniem metody Warda.



Porównanie metod:

Analiza skupień uogólniona metodą k-średnich wybiera losowo (k = 4) punkty jako początkowe centra klastrów w przeciwieństwie do Aglomeracji, gdzie każdy punkt jest traktowany jako oddzielny klaster. W pierwszej wymienionej metodzie w każdej iteracji punkty są przypisywane do najbliższych centrów k. Średnie wartości przypisanych punktów stanowią nowe centra klastrów. W ten sposób każda marka samochodu jest przypisana do jednej z 4 klas (kolumna wynikowa klasyfikacja). Oczywiście w tych klastrach występują różnice w parametrach co obrazuje wykres średnich zmiennych ilościowych.

W drugiej metodzie w każdej iteracji najbliższe klastry są łączone na podstawie funkcji kryterialnej (w tym wypadku metody Warda minimalizującej wariancję wewnątrz klastrów). Pokazuje to zamieszczony dendrogram. Widać na nim różnice w przyporządkowaniu marek samochodów do poszczególnych klastrów w porównaniu z Analizą skupień. Przede wszystkim Aglomeracja nie wymaga początkowego określenia liczby klastrów. Jest za to bardziej wrażliwa na obserwacje odstające.