Przedmiot: Eksploracja danych

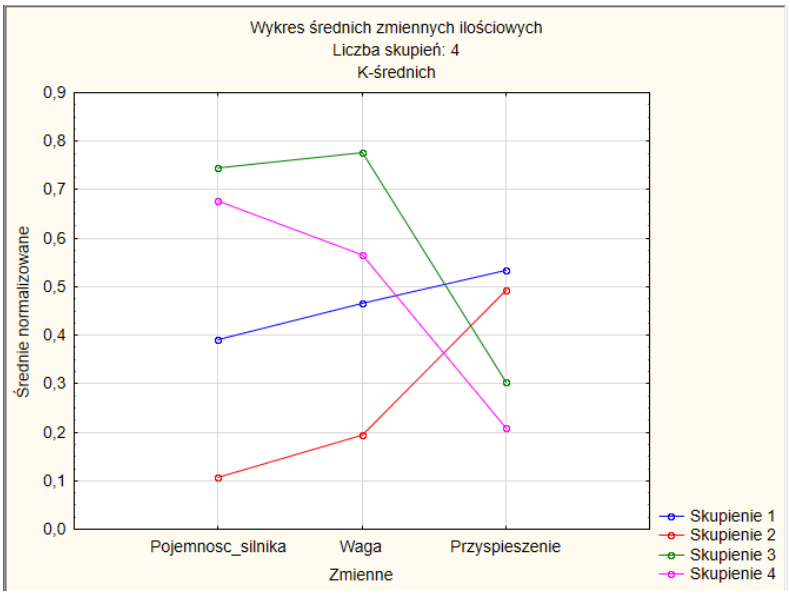
Kierunek: Informatyka – Data Science

Ćwiczenie: Analiza skupień – Auta all

Autor: Bartłomiej Jamiołkowski

Ad 5)

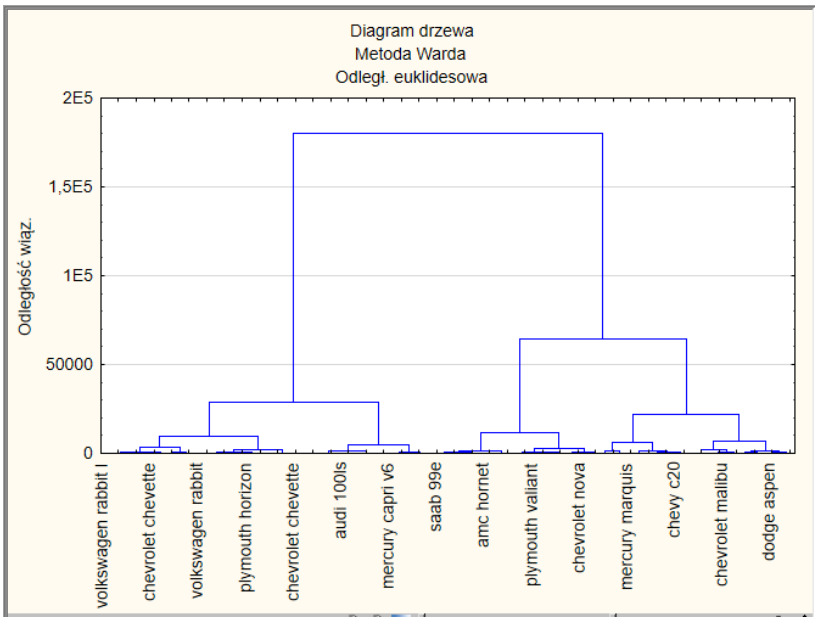
Analiza skupień uogólnioną metodą k-średnich.



Elementy skupienia (Auta_all)
Liczba skupień: 4
Całkowita liczba przypadków uczących: 398

Przypadek	Nr przypadku	Wynikowa klasyfikacja	Pojemnosc_silnika	Waga	Przyspie
chevrolet chevelle malibu	1	4	307,0000	3504,000	
buick skylark 320	2	4	350,0000	3693,000	
plymouth satellite	3	4	318,0000	3436,000	
amc rebel sst	4	4	304,0000	3433,000	
ford torino	5	4	302,0000	3449,000	
ford galaxie 500	6	3	429,0000	4341,000	
chevrolet impala	7	3	454,0000	4354,000	
plymouth fury iii	8	3	440,0000	4312,000	
pontiac catalina	9	3	455,0000	4425,000	
amc ambassador dpl	10	4	390,0000	3850,000	
dodge challenger se	11	4	383,0000	3563,000	
plymouth 'cuda 340	12	4	340,0000	3609,000	
chevrolet monte carlo	13	4	400,0000	3761,000	
buick estate wagon (sw)	14	4	455,0000	3086,000	
toyota corona mark ii	15	2	113,0000	2372,000	
plymouth duster	16	1	198,0000	2833,000	
amc hornet	17	1	199,0000	2774,000	
ford maverick	18	1	200,0000	2587,000	
datsun pl510	19	2	97,0000	2130,000	
volkswagen 1131 deluxe sedan	20	2	97,0000	1835,000	
peugeot 504	21	2	110,0000	2672,000	
audi 100 ls	22	2	107,0000	2430,000	
saab 99e	23	2	104,0000	2375,000	
bmw 2002	24	2	121,0000	2234,000	
amc gremlin	25	1	199,0000	2648,000	

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 każdy model samochodu jest przypisany do jednej z 4 klas (kolumna wynikowa klasyfikacja). W tych klastrach występują różnice w średnich wartościach parametrów 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 modeli 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.