

Analiza skupień zestawu parametrów samochodów amerykańskich, europejskich i japońskich wyprodukowanych w latach 1970-1982

Jerzy Marczewski Dawid Rozumkiewicz

Wojciech Pietraszuk Karol Hetmański

Przemysław Chachaj

1 Streszczenie

Przedmiotem badań są samochody pochodzenia amerykańskiego, europejskiego lub japońskiego wyprodukowane w latach 1970-1982. Dane pobraliśmy z repozytorium na GitHub o nazwie exploratory-data-analysis-dataset-cars należącego do użytkownika o pseudonimie RodolfoViana. Dane składają się z nazwy samochodu i kilku podstawowych informacji o nich. Omawiane dane wczytaliśmy do programu RStudio, a następnie dokonaliśmy obróbki danych. Na początku usunęliśmy co drugi wiersz, ponieważ dane były tak obszerne, że prezentacja graficzna była dosyć nieczytelna. Kolejnym krokiem było usunięcie wierszy, gdzie występowały znaki puste. Następnie usunęliśmy wiersze, gdzie powtarzały się nazwy samochodów, dzięki czemu te były unikalne, co sprawiło, że mogliśmy ich używać jako identyfikatorów. Obliczyliśmy statystyki opisowe. Wykorzystaliśmy metodę lokciową i silhouette w celu oszacowania optymalnej liczby klastrów, którą wykorzystaliśmy do metody k-srednich, Warda, complete, average i single. Potem użyliśmy techniki sieci Kohonen.

2 Słowa kluczowe

- klaster - zgrupowanie przestrzenne
- klasteryzacja - metoda klasyfikacji bez nadzoru (ang. unsupervised learning), która grupuje elementy na względzie jednorodne klasy
- mpg - (*ang. miles per gallon*) mile na galon

- objętość skokowa cylindra - różnica pomiędzy maksymalną a minimalną objętością cylindra
- objętość skokowa silnika - iloczyn objętości skokowej cylindra i liczby cylinderów
- hp - (*ang. Horsepower*) konie mechaniczne
- metoda k-średnich - jest jednym z algorytmów stosowanym w analizie skupień, wykorzystywanym m.in. w kwantyzacji wektorowej
- metoda Warda - jedna z aglomeracyjnych metod grupowania
- SOM - (*ang. Self-organizing map*) sieć Kohonena
- BMU - (*ang. best matching unit*) najlepiej dopasowana jednostka

3 Wprowadzenie

Okres 1970-1982 to ciekawy i ważny czas w motoryzacji. Na rynku występowała wtedy duża różnorodność, która niejako wymuszała podział tego rynku na segmenty. Gdyby chcieć przyporządkować jednoznacznie każdy model do danego segmentu mogłoby się okazać, że może być z tym problem. W rozwiązaniu tego problemu z pomocą przychodzą nam różne metody analizy skupień. Mają one zastosowanie w fazie eksploracyjnej badań, gdy nie dysponujemy żadnymi hipotezami. Celem analizy skupień jest ułożenie obiektów w grupy w taki sposób, by obiekty należące do tej samej grupy były ze sobą jak najbardziej powiązane, a jednocześnie były jak najmniej związane z obiektami z pozostałych grup. Tym sposobem możemy uzyskać bardzo dobry podział modeli samochodów na różne grupy.

4 Przedmiot badania

4.1 Cel i zakres badania

Badanie prowadzimy na podstawie danych różnych samochodów ze Stanów Zjednoczonych, Europy i Japonii. Dane składają się z nazwy samochodów i ich parametrów technicznych.

Naszym celem jest wykonanie analizy skupień samochodów. Badanie podzieli nam dane na klastry, na podstawie których będziemy mogli wywnioskować, które samochody są do siebie podobne, a które się między sobą różnią.

4.2 (min. 1 cytowanie powiązane tematycznie i krótki opis co było badane)

4.3 Zmienne wybrane do analizy

Do analizy wybrano następujące cechy diagnostyczne:

- X_1 - mpg
- X_2 - liczba cylindrów
- X_3 - objętość skokowa silnika
- X_4 - hp
- X_5 - waga (podana w funtach)
- X_6 - przyspieszenie
- X_7 - model (rok modelowy auta wyrażony w dwóch ostatnich cyfrach roku)
- X_8 - kraj pochodzenia (1 - Stany Zjednoczone, 2 - Europa, 3 - Japonia)

gdzie:

- X_1, X_2, X_3, X_4, X_6 to stymulanty
- X_5 to destymulanta
- X_7, X_8 to neutralna

4.4 Wstępna analiza danych

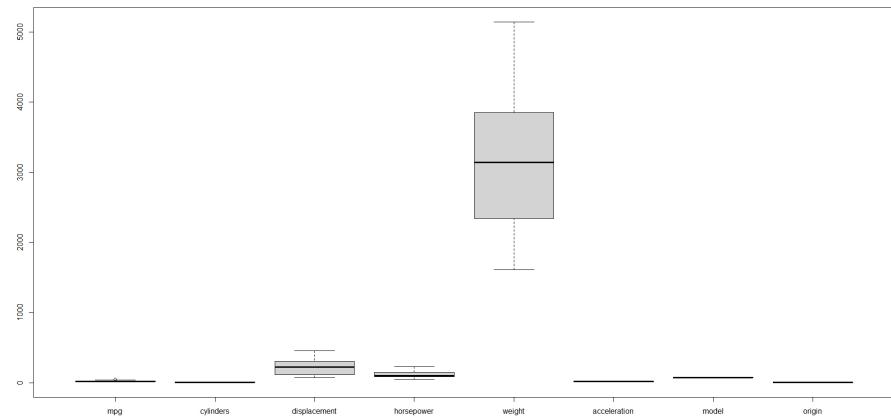
Statystyki opisowe

Wyniki średniej, mediany, minimum, maksimum, odchylenia standardowego, skośności dla każdej z cech zaokrąglone do dwóch miejsc po przecinku:

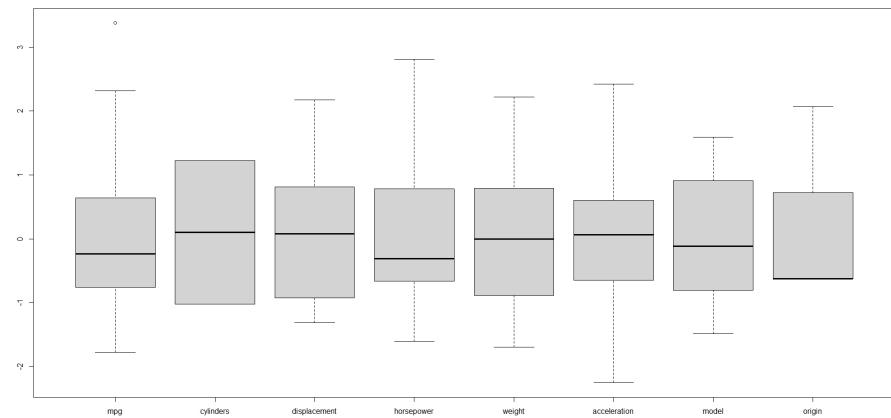
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
średnia	20.76	5.81	216.09	112.91	3139.32	15.31	74.35	1.46
mediana	19.2	6.0	225.0	100.0	3139.0	15.5	74.0	1.0
min	9	4	72	46	1613	9	70	1
max	43.1	8.0	455.0	230.0	5140.0	22.1	79.0	3.0
odchylenie standardowe	6.62	1.78	109.93	41.64	900.74	2.80	2.93	0.74
skośność	0.71	0.18	0.34	0.79	0.24	0.06	0.04	1.21

Podstawowa wizualizacja danych

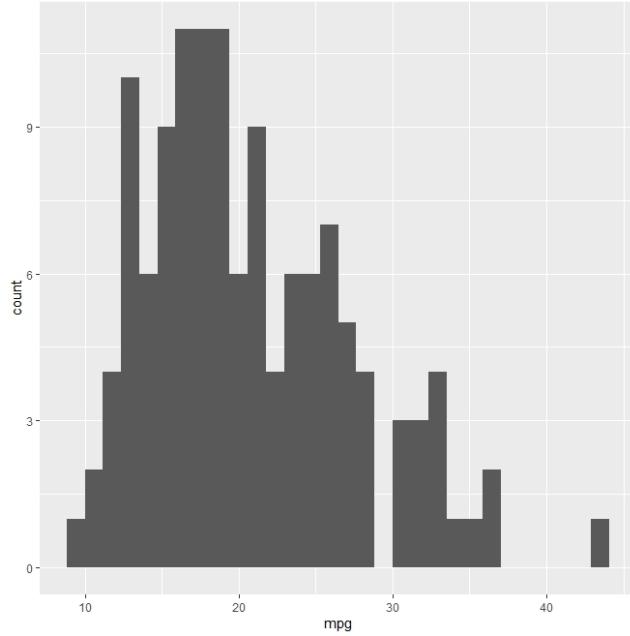
Rysunek 1: Wykres pudełkowy przed standaryzacją



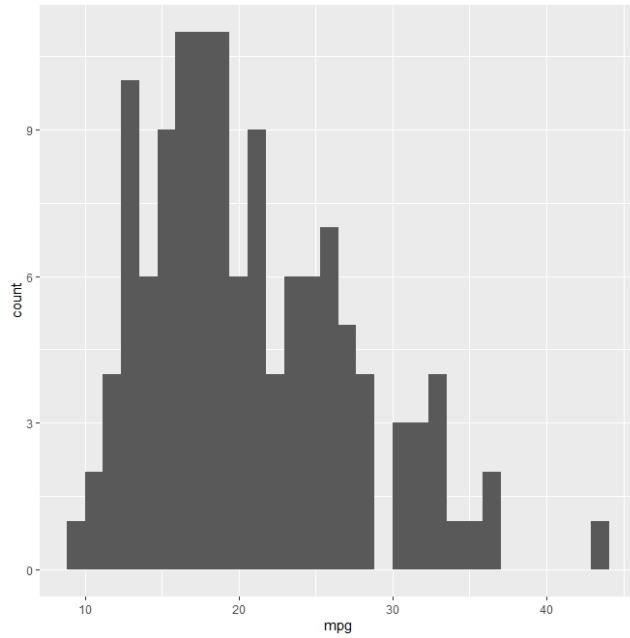
Rysunek 2: Wykres pudełkowy po standaryzacji



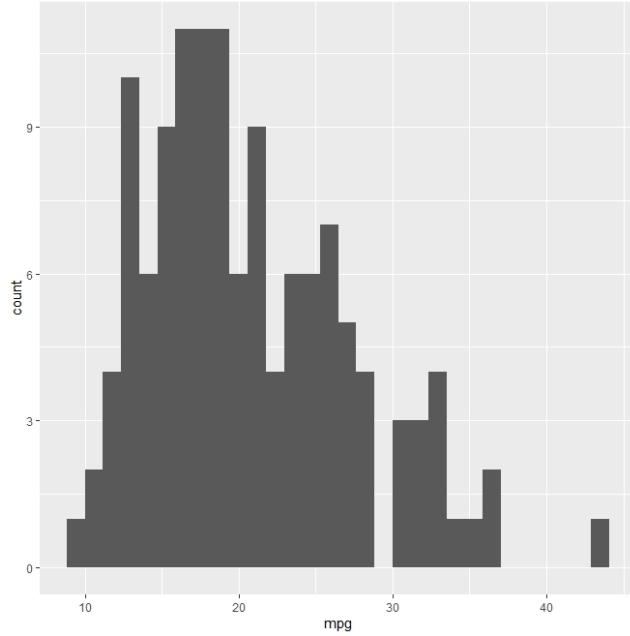
Rysunek 3: Histogram mpg



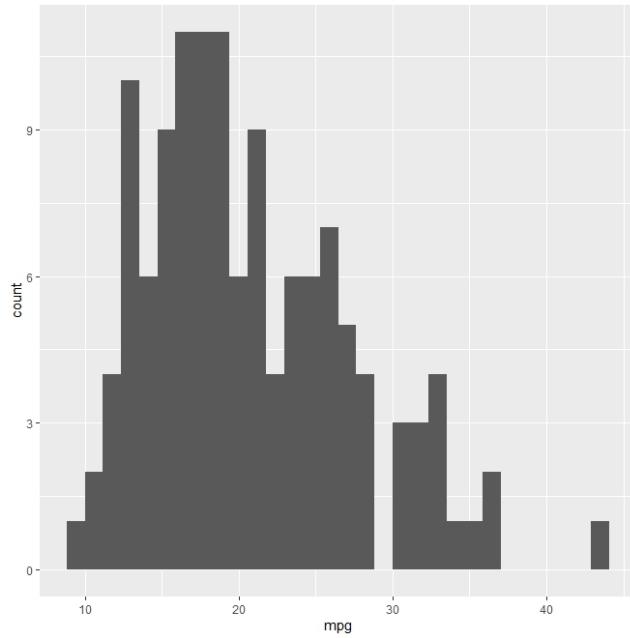
Rysunek 4: Histogram cylindrów



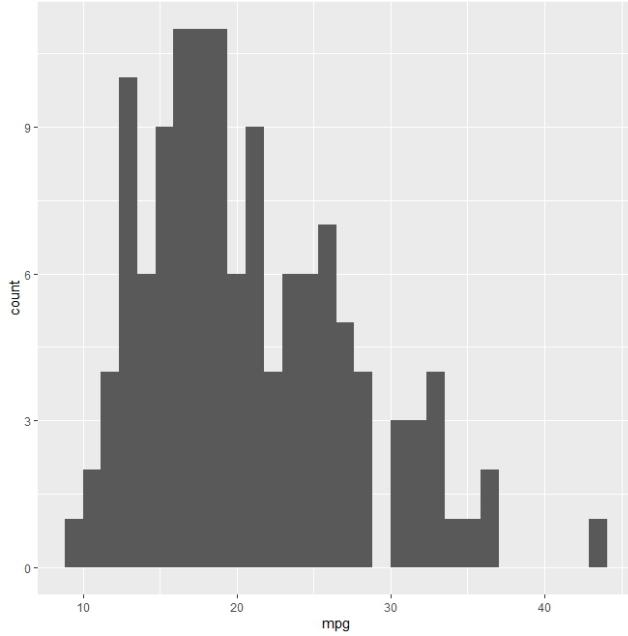
Rysunek 5: Histogram objętości skokowej silnika



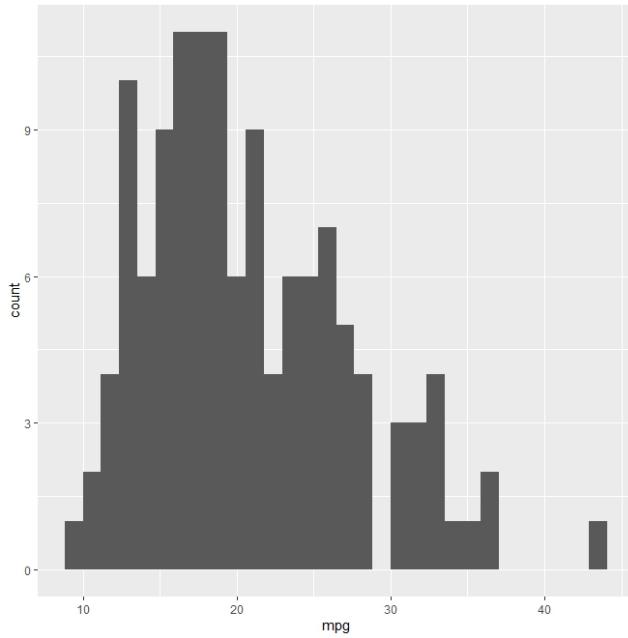
Rysunek 6: Histogram hp



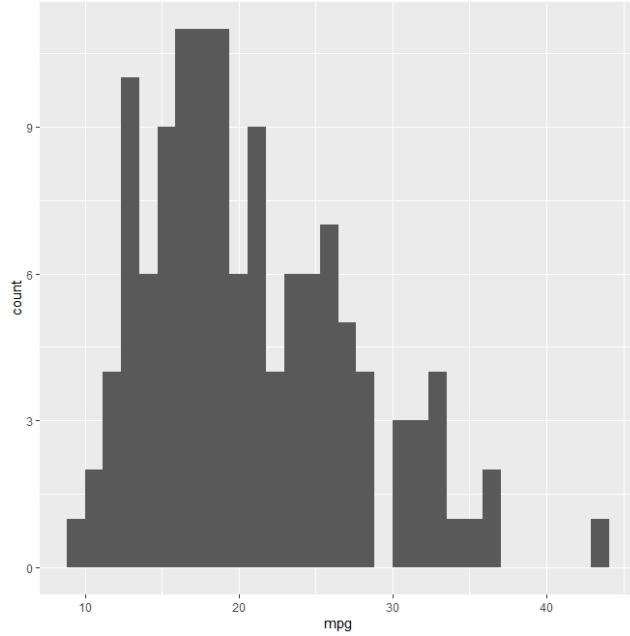
Rysunek 7: Histogram wagi



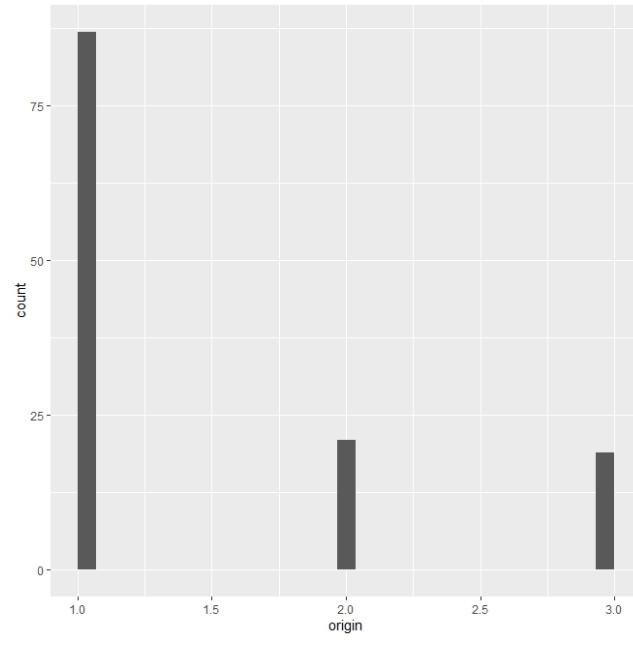
Rysunek 8: Histogram przyspieszenia



Rysunek 9: Histogram modelu



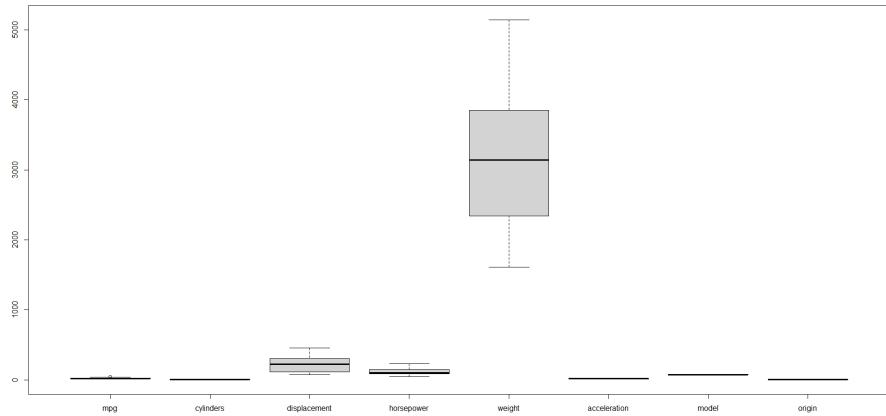
Rysunek 10: Histogram kraju pochodzenia



Braki danych i ich obsługa

Braki danych występują jedynie dla cechy hp. Wszystkie samochody z brakiem danych zostały usunięte z danych za pomocą skryptu w R.

Obserwacje odstające i ich obsługa



Obserwacja odstająca występuje jedynie dla cechy mpg. Różnica wartości jest na tyle mała, że praktycznie nie wpływa na przebieg badania.

5 Opis metod

5.1 wzory wraz z opisami oznaczeń

Metoda k-średnich

Celem metody jest przypisanie do wektorów r_i n wymiarowych wektorów danych, przy jak najmniejszym średnim błędzie kwantyzacji. Średni błąd kwantyzacji opisany jest wzorem:

$$D = \frac{1}{K} \sum_{i=1}^K d(x_i, r)$$

- K - liczba elementów x_i przypisanych do wektora r
- d - miara błędu kwantyzacji, najczęściej błąd kwadratowy opisany wzorem:

$$d(x, r) = \sum_{j=1}^n (x_j - r_j)^2$$

Metoda Warda

Odległość nowego skupienia od każdego pozostałego:

$$D_{pr} = a_1 \cdot d_{pr} + a_2 \cdot d_{qr} + b \cdot d_{pq}$$

- r - numery skupień różne od p i q
- D_{pr} - odległość nowego od skupienia r
- d_{pr} - odległość pierwotnego skupienia p od skupienia r
- d_{qr} - odległość pierwotnego skupienia q od skupienia r
- d_{pq} - wzajemna odległość pierwotnych skupień p i q
- $a_1 = \frac{n_p+n_r}{n_p+n_q+n_r}$, $a_2 = \frac{n_q+n_r}{n_p+n_q+n_r}$, $b = \frac{-n_r}{n_p+n_q+n_r}$
- n - liczliwość pojedyńczych obiektów w poszczególnych obiektach

SOM

Wzór aktualizowania neuronu v z wagą wektora $W_v(s)$:

$$W_v(s+1) = W_v(s) + \theta(u, v, s) \cdot \alpha(s) \cdot (D(t) - W_v(s))$$

- s - obecna iteracja
- t - indeks docelowego wektora danych wejściowych w zbiorze danych wejściowych D
- D(t) - docelowy wektor danych wejściowych
- v - indeks wektora w mapie
- W_v - aktualny wektor wagi węzła v
- u - indeks BMU na mapie
- $\theta(u, v, s)$ - jest ograniczeniem ze względu na odległość od BMU, zwykle nazywaną funkcją sąsiedztwa

5.2 cytowanie pracy w której zaproponowano metodę/ewentualnie pracę, w której użyto metodę

"Metoda Warda dąży do uzyskania raczej małych skupień i jest uznawana za bardzo efektywną. W wyniku analizy otrzymujemy dendrogram, będący graficzną interpretacją uzyskanych efektów. W zależności od przyjętych założeń badania, w tym zwłaszcza akceptowanej odległości taksonomicznej między obiekttami ze względu na zaproponowany zestaw cech, możemy wyróżniać większe lub mniejsze

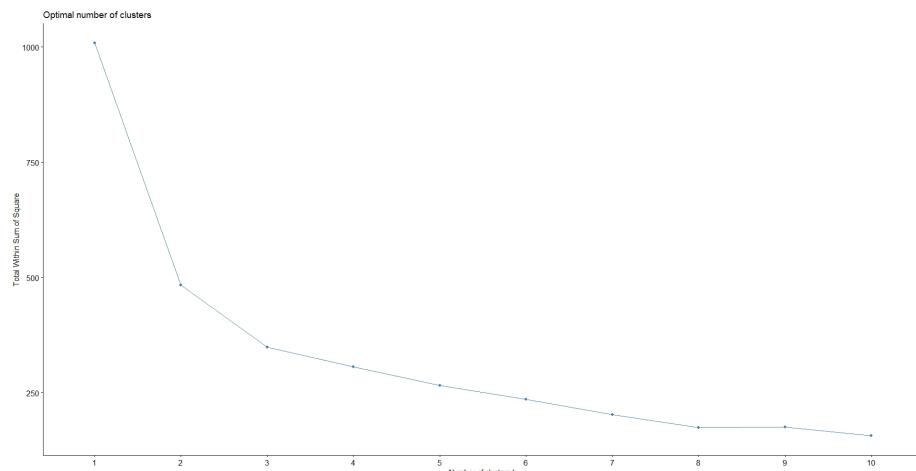
skupienia, a co za tym idzie – mniejszą lub większą ich liczbę." - Grupowanie państw Unii Europejskiej ze względu na zasoby kapitału ludzkiego i intelektualnego - Dr Małgorzata Stec, Mgr Agata Janas, Mgr Artur Kuliński - Uniwersytet Rzeszowski

"Dla potrzeb klasyfikacji bezwzorcowej skonstruowano sieć typu SOM. Jest to jeden z najbardziej zaawansowanych modeli sieci neuronowych, który dostarcza topologicznego odwzorowania przestrzeni wielowymiarowej na dwuwymiarową mapę neuronów. Może ona być zastosowana do wizualizacji skupisk w zbiorze danych, zachowując nieliniowe relacje między jednostkami i lokując bliskie jednostki bliżej siebie. Podczas trenowania sieci SOM wagie neuronów modelowane są w taki sposób, by bardzo zbliżone do siebie przypadki reprezentowały ten sam neuron, a podobne reprezentowane były przez neurony sąsiednie." - Grupowanie państw Unii Europejskiej ze względu na zasoby kapitału ludzkiego i intelektualnego - Dr Małgorzata Stec, Mgr Agata Janas, Mgr Artur Kuliński - Uniwersytet Rzeszowski

6 Rezultaty w postaci graficznej oraz omówienie wyników

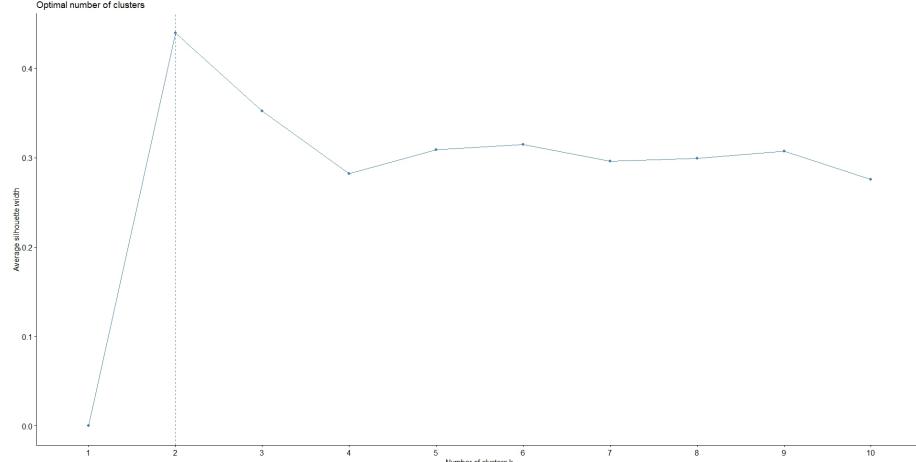
6.1 Metody wyznaczania optymalnej liczby zbiorów

6.1.1 Metoda lókciowa



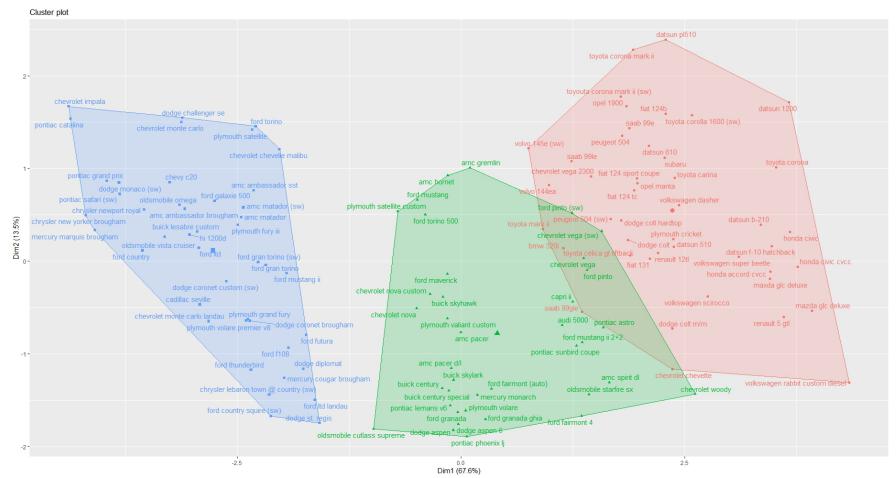
Z wykresu metody lókciowej można zaobserwować, że od wartości 3 na osi poziomej wykres maleje liniowo, więc ten punkt możemy przyjąć jako liczbę zbiorów do naszych danych.

6.1.2 Metoda silhouette



Kolejna metoda to ustalenia odpowiedniej liczby grupy, skupisk do naszych danych. Z tej metody wyszło, że nasze dane powinny być podzielone na 2 grupy. Ostatecznie wybraliśmy wybraliśmy 3 grupy zgodnie z metodą łokciową.

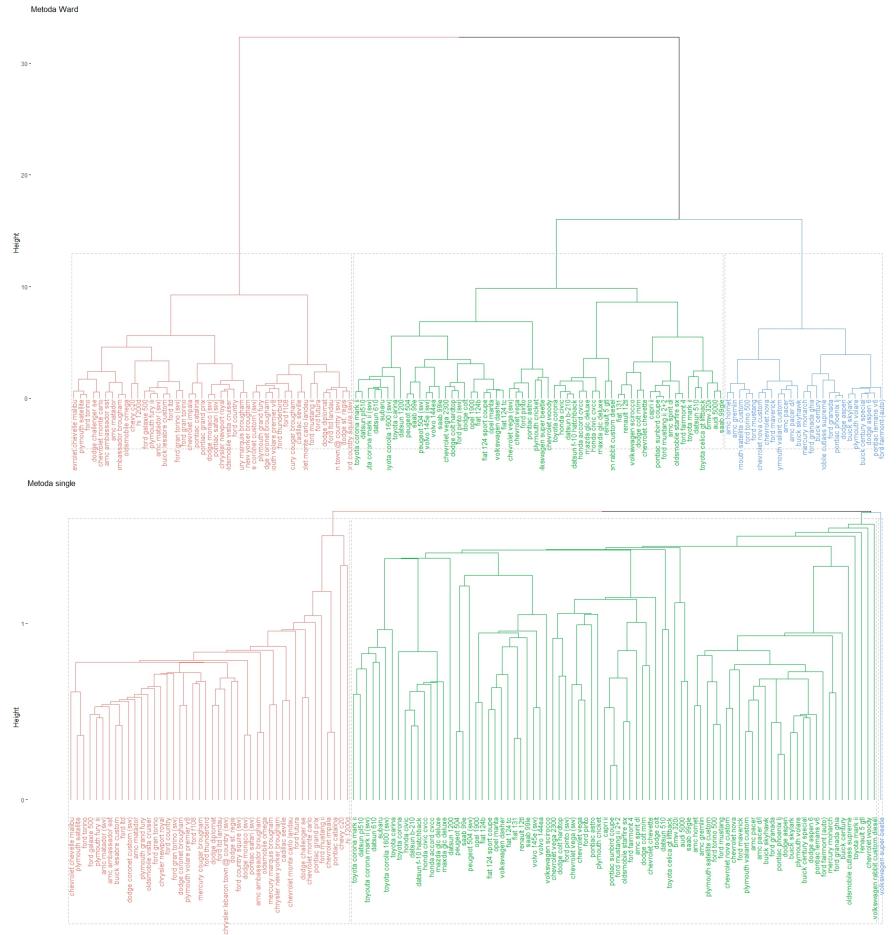
6.2 Metoda k-średnich

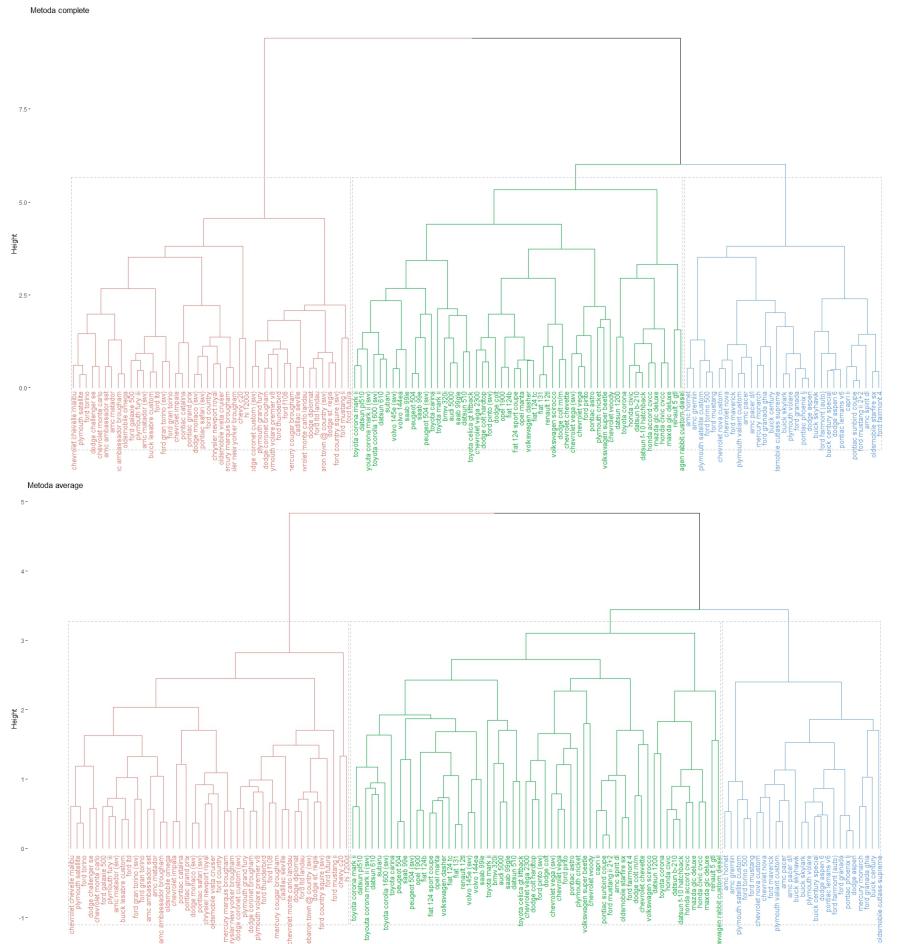


Dzięki metodom, które pomagają wybrać optymalną liczbę grup mogliśmy przeprowadzić analizę metodą k-średnich. Nasza optymalna liczba grup to 3. Z tego wykresu możemy odczytać zbiory o podobnych elementach oraz odległość między nimi. Co świadczy o wielkości podobieństwa. Na naszym wykresie można dostrzec, że dwa zbiory się pokrywają co świadczy o tym, że dane, które znajdują się w obu zbiorach kwalifikują się zarówno do zbioru zielonego i czerwonego, czyli ich parametry są do siebie zbliżone.

W zbiorze niebieskim są tylko samochody pochodzenia amerykańskiego, a w zbiorach zielonym i czerwonym są samochody pochodzenia europejskiego i japońskiego. Samochody amerykańskie charakteryzują dużą pojemność silnika, wysoką objętość skokowa silnika oraz duża waga.

6.3 Metoda Warda



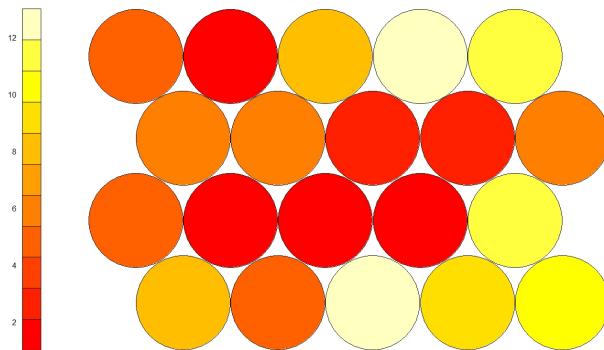


Jak widzimy wykresy wyglądają bardzo różnie. Grupy znalezione metodą Warda mają zbliżone wielkości. W przypadku metody single wykres jest najmniej czytelny oraz w klastrze trzecim, czyli niebieskim występuje tylko jeden obiekt. W tej metodzie grupa druga, czyli zielona przyjęła elementy z grupy niebieskiej porównując do reszty metod łączenia. Metoda Warda ma najbardziej czytelny wykres, łatwo z niego można odczytać otrzymane wyniki. Wykresy metod complete oraz average są dosyć podobne do siebie. Można zauważać również, że zbiory zielone i niebieski się zawsze łączą a dopiero później łączą się z ostatnim czerwonym. Wynika z tego więc, że zbiór zielony i niebieski mają więcej wspólnych cech niż zbiorem czerwonym. W przypadku wykresu complete wysokość osiąga wartość 8, w average około 5, w single około 1.5 i w metodzie Warda około 33. Im większa wysokość tym mniejsze powiązanie między aglomeracjami. Z tego wynika, że w metodzie single jest najmniejsza różnica powiązań, a w metodzie Warda największa.

6.4 SOM

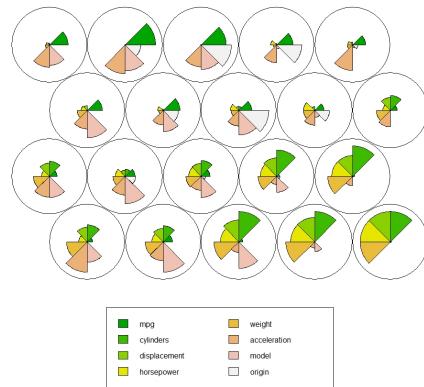
6.4.1 SOM counts

Rozmiar siatki to 5×4



Ustawiliśmy siatkę w wymiarach $x = 5$ i $y = 4$, ponieważ przy takich wymiarach nie pojawiają się szare okręgi, czyli takie, które są puste. Przy takich wymiarach siatki liczliwość zbiorów liczy od 1 do 14. Największe zbioru są w prawym dolnym rogu, na środku przy dolnej krawędzi wykresu oraz w prawym górnym rogu. Zbiory o najmniejszej liczbie elementów mieszczą się w środku. Po lewej stronie zbiory mają po około 6 elementów.

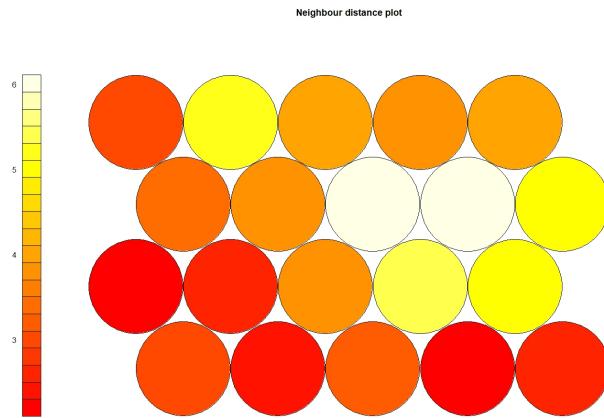
6.4.2 SOM codes



W prawym dolnym rogu są zbiory według liczby cylindrów, objętości skokowej silnika, liczby koni mechanicznych i wagi. Były to też najliczniejsze zbiory według wykresu som counts. Od środka do lewej strony widać, że zbiory są

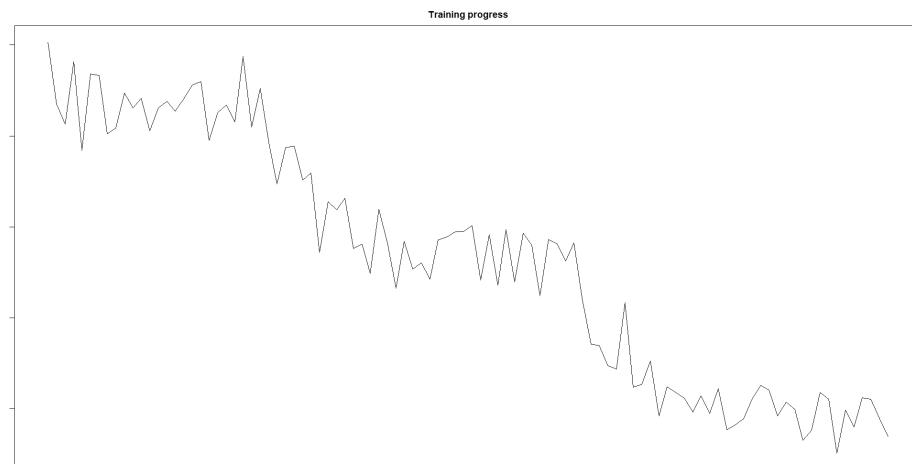
dobierane według modelu, czyli roku produkcji. W środku duże znaczenie ma pochodzenie samochodu.

6.4.3 SOM neighbour distance



Skala od 0 do 6. Najmniejsze odległości między zbiorami można zaobserwować w lewym dolnym roku oraz w prawym dolnym rogu. W drugim i trzecim wierszy od środka w prawo są największe odległości między zbiorami, co oznacza, że mają mniej wspólnych cech. Reszta zbiorów ma odległość na poziomie około 4, według skali przedstawionej na wykresie.

6.4.4 SOM training progress



Oś x to liczba iteracji a os y to średnia odległość do najbliższej jednostki W miarę postępu iteracji uczenia SOM zmniejsza się odległość od wag każdego węzła do próbek reprezentowanych przez ten węzeł. Idealna odległość powinna

osiągnąć minimalny poziom plateau. Ta opcja wykresu pokazuje postęp w czasie. Na naszym wykresie widać jak pod koniec funkcja coraz wolniej maleje, ale nadal występuje duża fluktuacja danych. Problem Plateau polega na wykazaniu istnienia minimalnej powierzchni przy danej granicy.

7 Podsumowanie (ocena realizacji celu, odniesienie do pozycji z przeglądu literatury)

8 Bibliografia

- Algorithm AS 136: A K-Means Clustering Algorithm - J. A. Hartigan and M. A. Wong
- www.statystyka.az.pl
- www.wikipedia.org
- https://github.com/RodolfoViana/exploratory-data-analysis-dataset-cars/blob/master/cars_multi.csv
- <https://mubi.pl/poradniki/najlepsze-samochody-lat-70/>
- https://pbiecek.github.io/NaPrzelajDataMiningR/part-3.html#part_35
- <https://nauka.metodolog.pl/metody-analizy-skupien-segmentacji-grupowanie/>
- https://www.statsoft.pl/textbook/stathome_stat.html?https\%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstcluan.html
- <https://www.r-bloggers.com/2014/02/self-organising-maps-for-customer-segmentation-using-kohonen-networks/>
- <https://pbiecek.gitbooks.io/przewodnik/content/Analiza/beznadzoru/agnes.html>