Analiza Preferencji Klientów

Weronika Kłujszo (223599), Michał Korzeniewski (223399),
Miłosz Malinowski (223391), Piotr Misiejuk (223302)

15 kwietnia 2025

Spis treści

1 Streszczenie **3**

2 Wprowadzenie **3**

3 Cel i zakres badania **4**

4 Przegląd literatury **4**

5 Słowa Kluczowe **5**

6 Zmienne wybrane do analizy oraz ich wizualizacja **5**

7 Wstępna analiza danych **9**
7.1 Statystyki opisowe 9
7.2 Brakujące oraz odstające dane 15
7.3 Skalowanie Danych 16
7.3.1 Standaryzacja 16
7.3.2 Normalizacja Min-Max 17
7.3.3 Skalowanie RobustScaler 18
7.3.4 One-hot encoding - dla zmiennych kategorycznych . . . 19
7.3.5 Macierze korelacji dla różnych metod normalizacji . . . 20

8 Omówienie metod klasteryzacji **24**
8.1 Klasteryzacja k-średnich 25
8.2 Klasteryzacja GMM (metoda mieszanki gaussowskiej) 26

9 Klasteryzacja dla danych po standaryzacji **28**
9.1 Klasteryzacja k-średnich 29
9.1.1 Wnioski 30
9.2 Klasteryzacja GMM 31
9.2.1 Wnioski 32

10 Klasteryzacja dla danych po normalizacji min-max **33**
10.1 Klasteryzacja k-średnich 35
10.1.1 Wnioski 37
10.2 Klasteryzacja GMM 38
10.2.1 Wnioski 39

11 Klasteryzacja dla danych po zastosowaniu normalizacji RobustScaler **41**
11.1 Klasteryzacja k-średnich 41
11.1.1 Wnioski 43
11.2 Klasteryzacja GMM 44
11.2.1 Wnioski 45

12 Rezultaty oraz omówienie wyników **46**

13 Podsumowanie **53**

14 Bibliografia **55**

1 Streszczenie

Projekt miał na celu segmentację klientów pewnego sklepu na podstawie danych demograficznych i zachowań zakupowych. Wykorzystano zbiór danych Customer Personality Analysis (2240 rekordów) z Kaggle, analizując zarówno zmienne ciągłe (np. dochód, wydatki), jak i kategoryczne (np. liczba dzieci, reakcje na kampanie). Dane zostały wstępnie przetworzone przy użyciu:

- standaryzacji,
- normalizacji min-max,
- skalowania RobustScaler (odpornego na wartości odstające),
- one-hot encoding dla zmiennych kategorycznych.

Do klasteryzacji zastosowano metody k-średnich i GMM (metodę mieszanki gaussowskiej), z optymalizacją liczby klastrów techniką łokcia. Analiza pozwoliła wyodrębnić wyraźnie różniące się segmenty klientów, co umożliwiło opracowanie spersonalizowanych rekomendacji marketingowych dostosowanych do charakterystyki każdej grupy.

2 Wprowadzenie

Współczesne przedsiębiorstwa gromadzą ogromne ilości danych o swoich klientach i stoją przed wyzwaniem ich efektywnego wykorzystania. Jednym z kluczowych zastosowań analizy danych w marketingu jest segmentacja klientów – podział bazy konsumentów na grupy o podobnych cechach. Segmentacja może opierać się na kryteriach demograficznych, behawioralnych czy psychograficznych. Odpowiednio przeprowadzona segmentacja umożliwia skierowanie właściwych ofert i komunikatów do odpowiednich odbiorców, zwiększając skuteczność kampanii marketingowych.

Analiza preferencji klientów stanowi szczegółowe badanie charakterystyki konsumentów. Pozwala ona lepiej zrozumieć zróżnicowanie bazy klientów oraz dostosować produkty i usługi do ich indywidualnych potrzeb. Firma może zidentyfikować segmenty najbardziej skłonne do zakupu danego produktu i skupić na nich swoje działania marketingowe. Tego rodzaju podejście przyczynia się do optymalizacji kosztów promocji oraz zwiększenia zadowolenia klientów poprzez bardziej spersonalizowaną ofertę.

W niniejszej pracy wykorzystano publicznie dostępny zbiór danych Customer Personality Analysis pochodzący z serwisu Kaggle. Dane te dotyczą kampanii marketingowej pewnej firmy i zawierają informacje o klientach, takie jak dane demograficzne oraz najchętniej kupowane produkty. Zbiór obejmuje 2240 klientów, których dane posłużyły do eksploracyjnej analizy danych oraz przeprowadzenia segmentacji z wykorzystaniem metod klasteryzacji. W dalszych rozdziałach przedstawiono proces przygotowania danych, zastosowane metody grupowania klientów oraz analizę otrzymanych wyników.

3 Cel i zakres badania

Celem projektu jest zastosowanie metod analizy skupień do segmentacji klientów na podstawie ich cech i zachowań zakupowych. Chcemy ustalić, czy możliwe jest wyodrębnienie sensownych grup konsumenckich oraz jakimi cechami się one charakteryzują. Projekt zakłada również porównanie kilku metod klasteryzacji (k-średnich oraz GMM) oraz ocenę ich skuteczności.

Zakres badania obejmuje:

- wczytanie i przygotowanie danych,
- skalowanie danych przy użyciu różnych metod (standaryzacja, min-max, RobustScaler, one-hot encoding dla zmiennych kategorycznych),
- zastosowanie dwóch metod klasteryzacji: k-średnich oraz GMM,
- ocenę i interpretację otrzymanych klastrów,
- podsumowanie wyników oraz możliwe rekomendacje marketingowe.

4 Przegląd literatury

Segmentacja klientów za pomocą analizy skupień jest szeroko opisywana w literaturze i stosowana w wielu sektorach. Wśród podobnych badań można wymienić:

W pracy [1] autor analizuje dane klientów firmy Amazon oraz stosuje metody nienadzorowane do ich segmentacji, z naciskiem na analizę skupień jako narzędzie wspierające decyzje biznesowe.

W badaniu [2] przedstawiono zastosowanie klasteryzacji k-średnich do identyfikacji cech charakterystycznych konsumentów, co miało na celu poprawę skuteczności działań marketingowych.

W artykule [3] poruszono temat segmentacji klientów w kontekście relacji, zaufania i zaangażowania. Autor sugeruje, że skuteczna segmentacja powinna uwzględniać psychograficzne aspekty zachowań konsumenckich.

Z kolei w publikacji [4] omówiono wpływ cech osobowości klientów na ich satysfakcję, postrzeganie marki i lojalność w kontekście usług mobilnych. Praca ta potwierdza znaczenie cech psychologicznych w segmentacji.

5 Słowa Kluczowe

POL: analiza osobowości klientów, strategia marketingowa, analiza danych, analiza skupień, klastrowanie
ENG: customer behaviour analysis, marketing strategy, data analysis, cluster analysis, clustering

6 Zmienne wybrane do analizy oraz ich wizualizacja

Zmienne podzielono następująco:

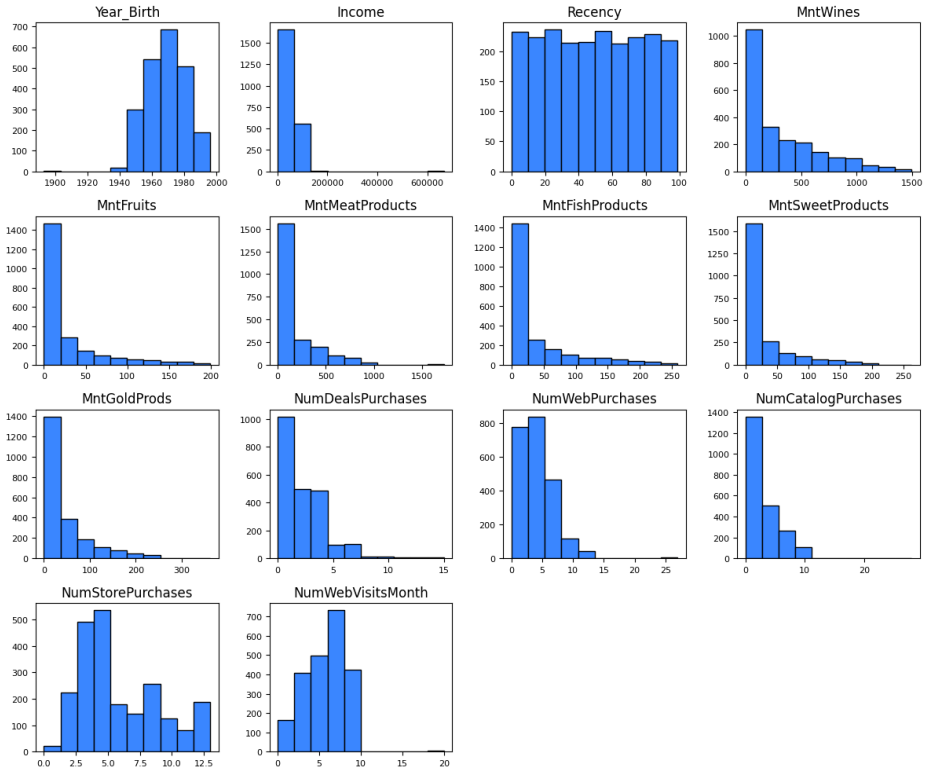
- **Ciągłe:**
  - Rok urodzenia klienta (Year_Birth),
  - Roczny dochód gospodarstwa domowego klienta (Income) - jednostka: dolary (USD),
  - Liczba dni od ostatniego zakupu klienta (Recency),
  - Kwota wydana na wino w ostatnich 2 latach (MntWines) - jednostka: dolary (USD),
  - Kwota wydana na owoce w ostatnich 2 latach (MntFruits) - jednostka: dolary (USD),
  - Kwota wydana na mięso w ostatnich 2 latach (MntMeatProducts) - jednostka: dolary (USD),

  - Kwota wydana na ryby w ostatnich 2 latach (MntFishProducts) - jednostka: dolary (USD),
  - Kwota wydana na słodycze w ostatnich 2 latach (MntSweetProducts) - jednostka: dolary (USD),
  - Kwota wydana na złoto w ostatnich 2 latach (MntGoldProds) - jednostka: dolary (USD),
  - Liczba zakupów dokonanych z rabatem (NumDealsPurchases),
  - Liczba zakupów dokonanych przez stronę internetową (NumWebPurchases),
  - Liczba zakupów dokonanych za pomocą katalogu (NumCatalogPurchases),
  - Liczba zakupów dokonanych w sklepie (NumStorePurchases),
  - Liczba wizyt na stronie internetowej w ostatnim miesiącu (NumWebVisitsMonth).

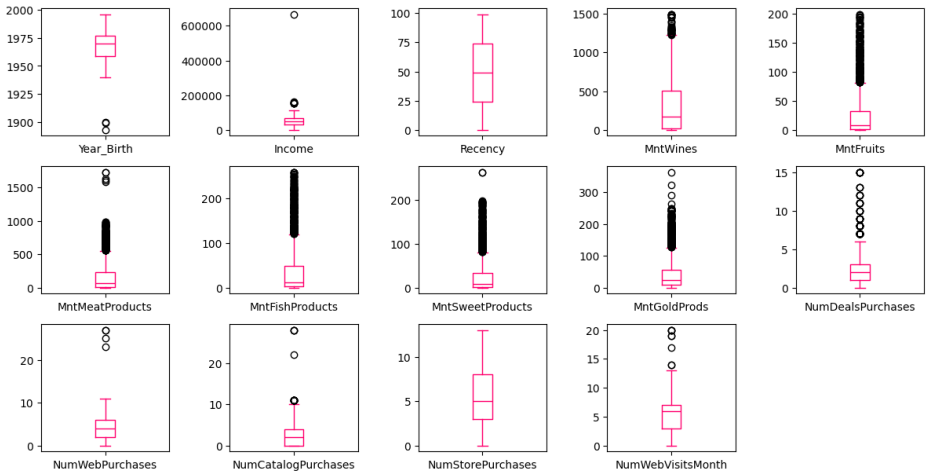
- **Kategoryczne:**
  - Liczba dzieci w gospodarstwie domowym klienta (Kidhome),
  - Liczba nastolatków w gospodarstwie domowym klienta (Teenhome),
  - Skarga (Complain): wartość = 1, jeśli klient składał skargę w ciągu ostatnich 2 lat, 0 w przeciwnym razie,
  - Akceptacja oferty w kampaniach (AcceptedCmp1 - AcceptedCmp5): wartość = 1, jeśli klient zaakceptował ofertę w kampaniach marketingowych,
  - Reakcja na ofertę w ostatniej kampanii (Response): wartość = 1, jeśli klient zaakceptował ofertę w ostatniej kampanii.

Stworzono wykresy rozkładu zmiennych. Dla zmiennych ciągłych powstały histogramy oraz wykresy pudełkowe.

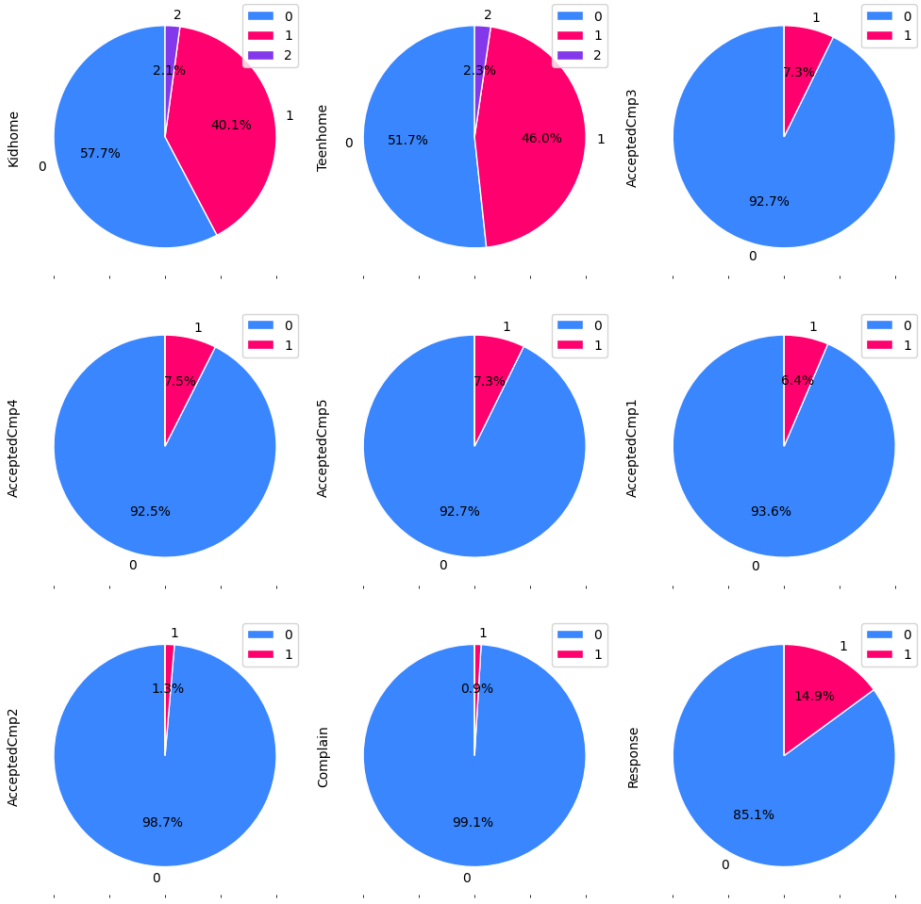
Dla zmiennych kategorycznych powstały wykresy kołowe.



Rys. 1. Histogramy rozkładu zmiennych ciągłych.



Rys. 2. Wykresy pudełkowe dla zmiennych ciągłych.



Rys. 3. Wykresy kołowe dla zmiennych kategorycznych.

7 Wstępna analiza danych

Przed przystąpieniem do klasteryzacji i analizy segmentacji klientów, dane zostały wstępnie przeanalizowane w celu zrozumienia ich struktury, identyfikacji brakujących wartości, wykrywania wartości odstających oraz zapoznania się z rozkładem zmiennych.

Zmienna `Year_Birth` została przekształcona w zmienną `Age`, reprezentującą wiek klienta. Dokonano tego, odejmując wartość `Year_Birth` od wartości `2021`, ponieważ to właśnie w tym roku ostatnio edytowano zbiór danych.

7.1 Statystyki opisowe

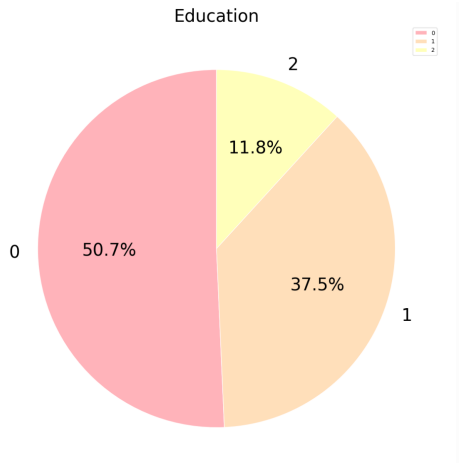
Aby uzyskać ogólny obraz danych, obliczono podstawowe statystyki opisowe dla zmiennych ciągłych. Dla każdej zmiennej obliczono:

- średnią,
- medianę,
- maksimum,
- minimum,
- odchylenie standardowe,
- skośność.

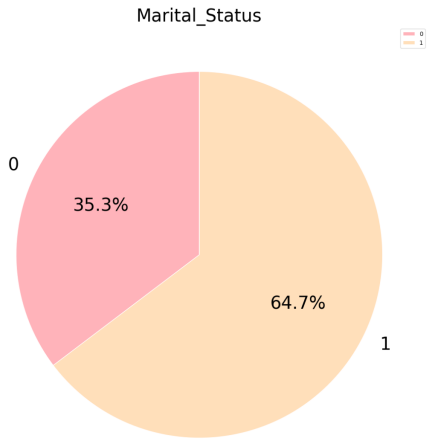
Zmienna	Średnia	Mediana	Min	Max	Odch. std.	Skośność
Income	50879.11	50008.00	5305.00	105471.00	20403.60	0.06
Recency	48.96	49.00	0.00	99.00	28.99	0.01
MntWines	279.29	158.00	0.00	1224.00	308.86	1.08
MntFruits	25.66	8.00	0.00	145.74	37.49	1.86
MntMeatProducts	157.35	60.00	0.00	839.85	208.74	1.70
MntFishProducts	37.22	12.00	0.00	201.89	53.33	1.75
MntSweetProducts	26.25	8.00	0.00	150.25	38.75	1.88
MntGoldProds	41.63	23.00	0.00	199.41	48.23	1.72
NumWebPurchases	3.96	3.00	0.00	11.00	2.57	0.74
NumCatalogPurchases	2.50	1.00	0.00	10.00	2.64	1.09
NumStorePurchases	5.74	5.00	0.00	13.00	3.23	0.79
NumWebVisitsMonth	5.28	6.00	0.00	13.00	2.28	-0.31
Age	51.92	51.00	25.00	81.00	11.65	0.10
Spent	570.71	322.00	8.00	2352.00	580.79	0.89

Tabela 1: Statystyki opisowe dla zmiennych ilościowych (średnia, mediana, minimum, maksimum, odchylenie standardowe, skośność)

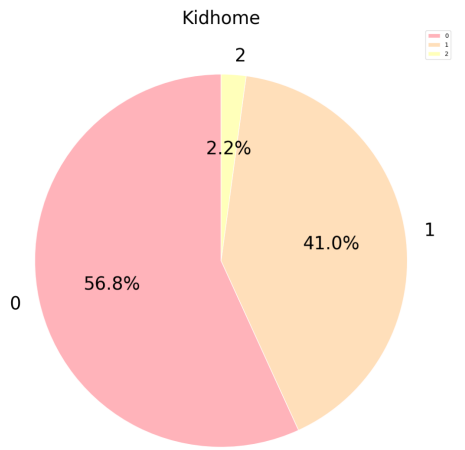
Statystyki te pozwalają na lepsze zrozumienie rozkładu danych oraz identyfikację zmiennych o dużym rozrzucie lub skrajnych wartościach.



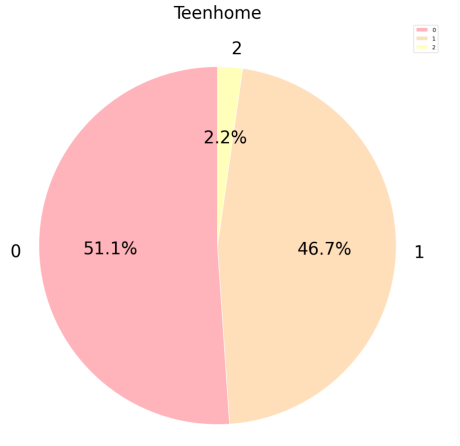
Rys. 4. Rozkład poziomu wykształcenia klientów: **50.7%** klientów ma wykształcenie średnie i podstawowe, **37.5%** wyższe pierwszego i drugiego stopnia, a **11.8%** posiada doktorat.



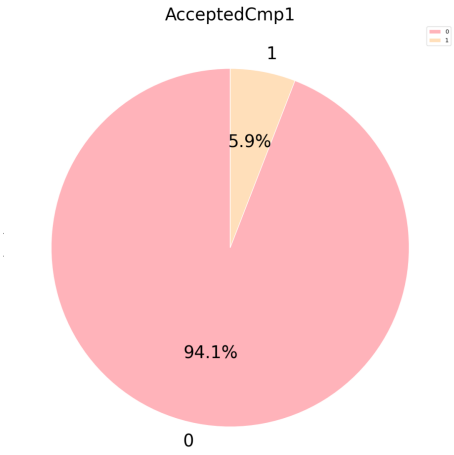
Rys. 5. Rozkład statusu związku klientów: **64.7%** klientów jest w związku, a **35.3%** to osoby bez partnera.



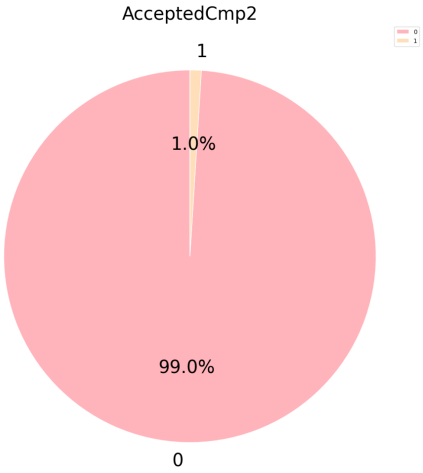
Rys. 6. Rozkład liczby dzieci w gospodarstwie domowym klientów: **56.8%** klientów nie mieszka z dziećmi, **41%** posiada jedno dziecko, a **2,2%** dwójkę dzieci.



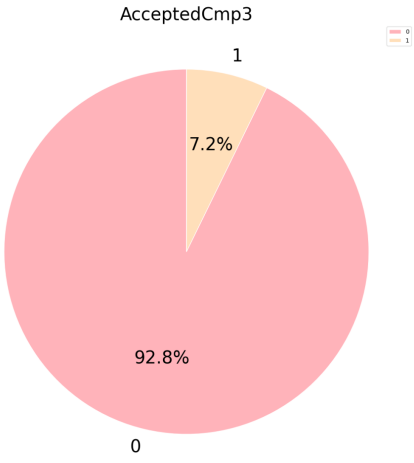
Rys. 7. Rozkład liczby nastolatków w gospodarstwie domowym klientów: **51.1%** klientów posiada dzieci w wieku nastoletnim, **46.7%** posiada jedno dziecko w wieku nastoletnim, a **2,2%** dwójkę.



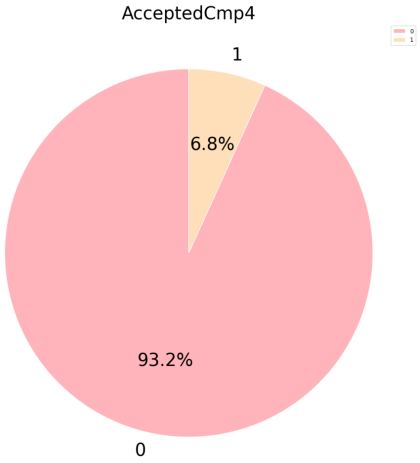
Rys. 8. Rozkład akceptacji oferty w kampanii marketingowej nr 1: **94.1%** klientów nie zaakceptowało oferty, a **5.9%** zaakceptowało.



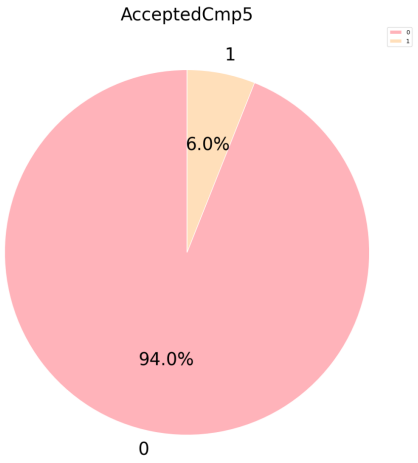
Rys. 9. Rozkład akceptacji oferty w kampanii marketingowej nr 2: **99%** klientów nie zaakceptowało oferty, a **1%** zaakceptował.



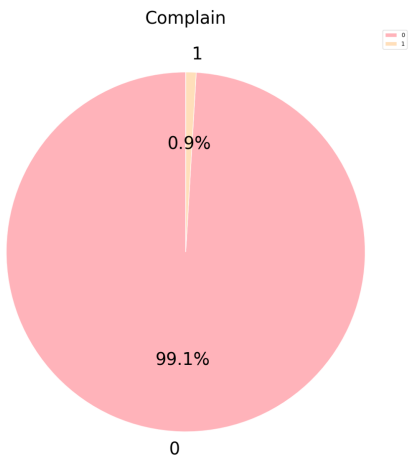
Rys. 10. Rozkład akceptacji oferty w kampanii marketingowej nr 3: **92.8%** klientów nie zaakceptowało oferty, a **7.2%** zaakceptowało.



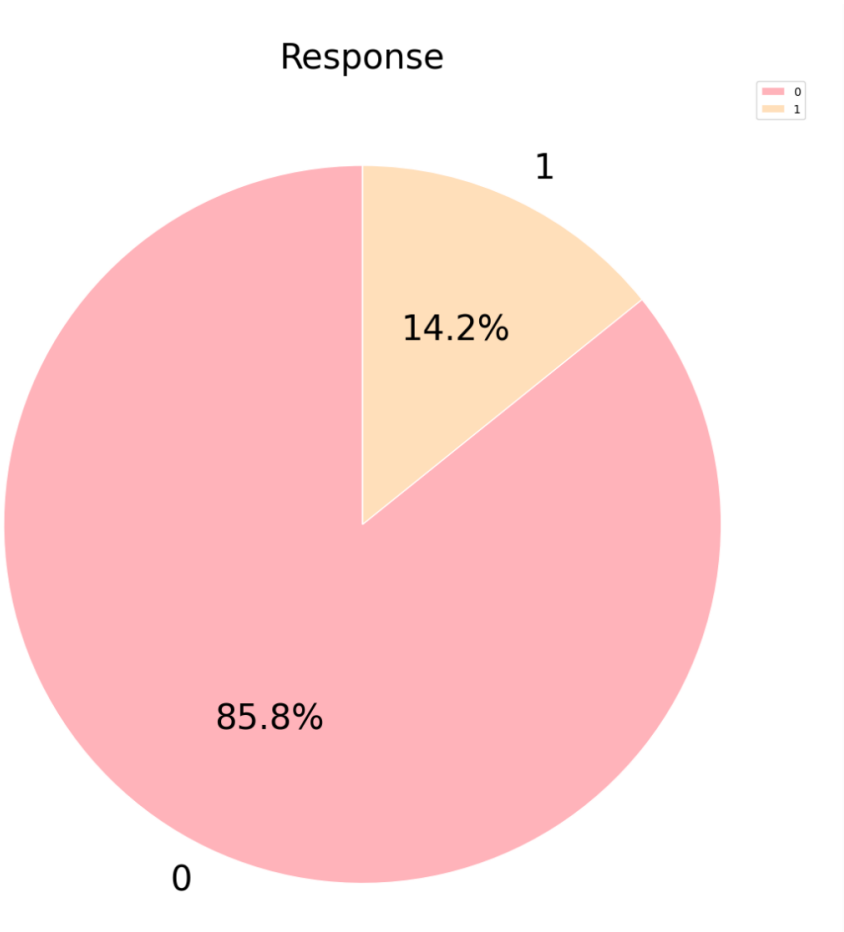
Rys. 11. Rozkład akceptacji oferty w kampanii marketingowej nr 4: **93.2%** klientów nie zaakceptowało oferty, a **6.8%** zaakceptowało.



Rys. 12. Rozkład akceptacji oferty w kampanii marketingowej nr 5: **94%** klientów nie zaakceptowało oferty, a **6%** zaakceptowało.



Rys. 13. Rozkład skarg wśród klientów: **99.1%** klientów nie złożyło skargi w przeciągu ostatnich 2 lat, a **0.9%** klientów złożyło.



Rys. 14. Rozkład odpowiedzi na ostatnią kampanię wśród klientów: **85.8%** klientów nie odpowiedziało na kampanię, a **14.2%** klientów odpowiedziało.

7.2 Brakujące oraz odstające dane

Brakujące wartości występują jedynie w kolumnie Income, a ich liczba jest niewielka (24, co stanowi 1.07% wszystkich wartości). Ze względu na ich małą ilość, rzędy zawierające brakujące wartości zostały usunięte i nie były brane pod uwagę w dalszej analizie.

Oznaczenia:

Q1 - pierwszy kwartyl,

Q3 - trzeci kwartyl,

IQR - odstęp ćwiartkowy.

Wartości odstające zostały wyznaczone zgodnie z regułą 1.5 IQR, gdzie wartości poniżej Q1-1.5*IQR bądź powyżej Q3+1.5*IQR zostały uznane za odstające.

W przypadku danych ilościowych, w kolumnach, gdzie wartości odstające stanowiły mniej niż 5% wszystkich danych, zostały one usunięte i nie były brane pod uwagę w dalszej analizie.

W pozostałych kolumnach danych ilościowych wartości odstające zostały ograniczone z góry przez Q3+1.5*IQR oraz z dołu przez Q1-1.5*IQR.

Kolumna	Liczba wartości odstających	Procent wartości odstających
Education	0	0.0
Marital_Status	0	0.0
Income	8	0.4
Kidhome	0	0.0
Teenhome	0	0.0
Recency	0	0.0
MntWines	35	1.6
MntFruits	246	11.1
MntMeatProducts	174	7.9
MntFishProducts	222	10.0
MntSweetProducts	246	11.1
MntGoldProds	205	9.3
NumDealsPurchases	84	3.8
NumWebPurchases	3	0.1
NumCatalogPurchases	23	1.0
NumStorePurchases	0	0.0
NumWebVisitsMonth	8	0.4
AcceptedCmp3	163	7.4
AcceptedCmp4	164	7.4
AcceptedCmp5	162	7.3
AcceptedCmp1	142	6.4
AcceptedCmp2	30	1.4
Complain	21	0.9
Response	333	15.0
Age	3	0.1
Spent	3	0.1

Tabela 2: Statystyki wartości odstających

7.3 Skalowanie Danych

Skalowanie danych jest jednym z kluczowych etapów przygotowania danych do analizy. Jego celem jest zapewnienie, że wszystkie cechy będą miały tę samą wagę i przyczynią się równomiernie do wyników modelu.

W tym projekcie zastosowano nastepujace metody skalowania: **standaryzację**, **skalowanie min-max**, **skalowanie RobustScaler** oraz - dla zmiennych kategorycznych - **one-hot encoding**.

7.3.1 Standaryzacja

Standaryzacja (inaczej normalizacja z wykorzystaniem średniej i odchylenia standardowego) polega na przekształceniu danych w taki sposób, że każda

cecha ma średnią wartość 0 oraz odchylenie standardowe 1. Dzięki temu zmienne z różnych zakresów wartości stają się porównywalne i nie mają wpływu na wyniki analiz statystycznych, takich jak klasteryzacja czy analiza głównych składowych (PCA).

Matematycznie, dla danej zmiennej x, przekształcenie za pomocą standaryzacji można zapisać w następujący sposób:

$$x' = \frac{x - \mu}{\sigma}$$

gdzie:

- x – wartość oryginalna,
- x' – wartość po standaryzacji,
- μ – średnia wartość zmiennej x,
- σ – odchylenie standardowe zmiennej x.

W wyniku tej operacji każda cecha w zbiorze danych ma średnią wartość równą 0, a odchylenie standardowe równe 1. Dzięki temu dane są "skalowane"w taki sposób, że zmienne o różnych jednostkach i zakresach wartości stają się porównywalne.

Standaryzacja jest szczególnie użyteczna, gdy dane mają różne jednostki miary lub różne zakresy wartości, a także w sytuacjach, gdy algorytmy takie jak k-średnie lub PCA zakładają, że dane są rozproszone wokół średniej z równym rozproszeniem (odchyleniem standardowym). Wadą może być wrażliwość na wartości odstające, które mogą zniekształcić obliczanie średniej i odchylenia standardowego.

7.3.2 Normalizacja Min-Max

Metoda min-max polega na przekształceniu danych w taki sposób, aby mieściły się one w określonym przedziale, zazwyczaj od 0 do 1. Jest użyteczna, gdy zależy nam na zachowaniu rozkładu wartości w tych samych proporcjach, szczególnie gdy chcemy zachować relacje między minimalną a maksymalną wartością zmiennej.

Matematycznie, dla danej zmiennej x, przekształcenie za pomocą normalizacji Min-Max można zapisać w następujący sposób:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

gdzie:

- x – wartość oryginalna,
- x' – wartość po normalizacji,
- $\min(x)$ – minimalna wartość w zbiorze danych,
- $\max(x)$ – maksymalna wartość w zbiorze danych.

W wyniku tego przekształcenia, minimalna wartość w zbiorze danych x zostaje przekształcona do 0, a maksymalna wartość do 1. Wartości pomiędzy nimi są skalowane proporcjonalnie.

Metoda min-max jest mniej oporna na wartości odstające, ponieważ po przekształceniu wartości nadal mieszczą się w określonym przedziale. Z tego względu, w przypadku dużej liczby wartości odstających, dane mogą zostać "ściśnięte"w wąskim przedziale, co może zniekształcić interpretację wyników. Jednakże, w sytuacji, gdy zależy nam na zachowaniu proporcji w rozkładzie, metoda ta jest bardzo skuteczna.

7.3.3 Skalowanie RobustScaler

RobustScaler jest techniką skalowania, która jest bardziej odporna na wartości odstające w porównaniu do tradycyjnych metod, takich jak standaryzacja czy normalizacja min-max. Zamiast używać średniej i odchylenia standardowego (jak w przypadku standaryzacji), RobustScaler opiera się na statystykach odpornościowych, takich jak mediana oraz IQR (odstęp ćwiartkowy).

Formuła skalowania w metodzie RobustScaler wygląda następująco:

$$X' = \frac{X - \text{Mediana}(X)}{\text{IQR}(X)}$$

gdzie:

- X – wartość oryginalna,
- $\text{Mediana}(X)$ – mediana danej cechy,
- $\text{IQR}(X)$ – odstęp ćwiartkowy, który oblicza się jako różnicę między trzecim i pierwszym kwartylem (Q3 - Q1).

RobustScaler skaluje dane poprzez odejmowanie mediany i dzielenie przez odstęp ćwiartkowy (IQR). Dzięki temu skala jest mniej wrażliwa na obecność skrajnych wartości w danych, co sprawia, że jest to technika bardziej odporna na wartości odstające.

Zaletą tej metody jest to, że nie jest ona wrażliwa na ekstremalne wartości w zbiorze danych, przez co nadaje się lepiej do pracy z danymi zawierającymi dużo wartości odstających, w porównaniu do tradycyjnej standaryzacji.

Zastosowanie:

- RobustScaler jest szczególnie przydatny w przypadku danych, które zawierają dużą liczbę wartości odstających.
- Skala danych po zastosowaniu tej metody będzie oparta na mediana i IQR, co pozwala na bardziej sprawiedliwą reprezentację danych w sytuacjach, gdzie standardowe metody mogą zawieść z powodu obecności odstających punktów.

7.3.4 One-hot encoding - dla zmiennych kategorycznych

Zmiennych kategorycznych nie można bezpośrednio wykorzystać w analizie klasteryzacyjnej, ponieważ algorytmy takie jak k-średnich czy GMM wymagają danych liczbowych. W tym celu do przekształcania danych kategorycznych na wartości liczbowe używamy techniki one-hot encoding, która działa w następujący sposób:

- **Tworzenie nowych kolumn:** Dla każdej unikalnej wartości w kolumnie kategorycznej tworzona jest nowa kolumna.
- **Zamiana wartości na 0 i 1:** W każdej z nowo utworzonych kolumn wstawiamy wartość 1 (gdy dany rekord ma tę kategorię) lub 0 (gdy nie ma tej kategorii).

Matematycznie proces ten można zapisać w następujący sposób. Załóżmy, że mamy kolumnę kategoryczną C z wartościami $\{c_1, c_2, \ldots, c_k\}$, gdzie k to liczba unikalnych kategorii.

Dla każdego wiersza i w tej kolumnie tworzymy nową kolumnę C_j, gdzie $j = 1, 2, \ldots, k$ i przypisujemy jej wartość:

$$C_j(i) = \begin{cases} 1 & \text{jeśli } C(i) = c_j \\ 0 & \text{w przeciwnym razie} \end{cases}$$

Gdzie:

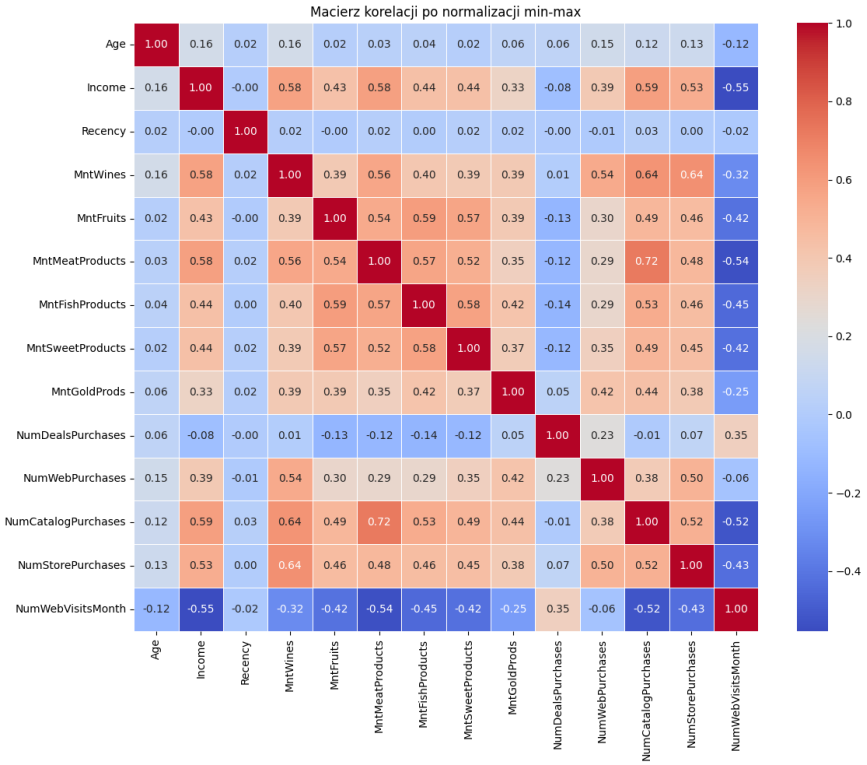
- $C(i)$ to wartość w kolumnie C dla wiersza i,
- $C_j(i)$ to wartość nowej kolumny (0 lub 1) w zależności od tego, czy wartość w $C(i)$ jest równa c_j.

W ten sposób każda kategoria z kolumny C zostaje reprezentowana jako zestaw binarnych kolumn (0 lub 1).

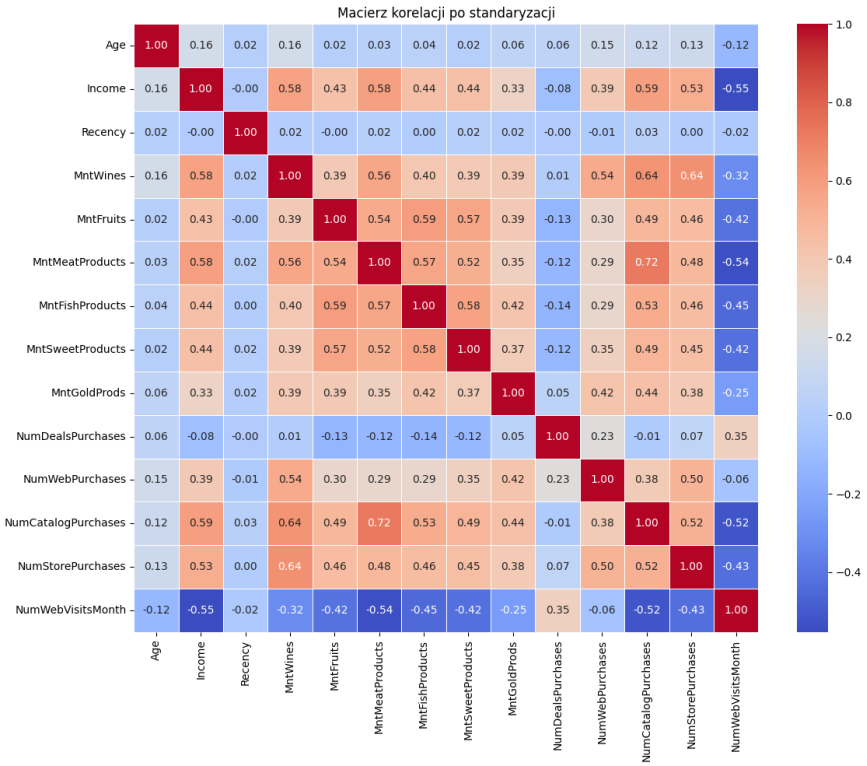
Wartość 1 w odpowiedniej kolumnie oznacza obecność danej kategorii, a 0 oznacza jej brak.

7.3.5 Macierze korelacji dla różnych metod normalizacji

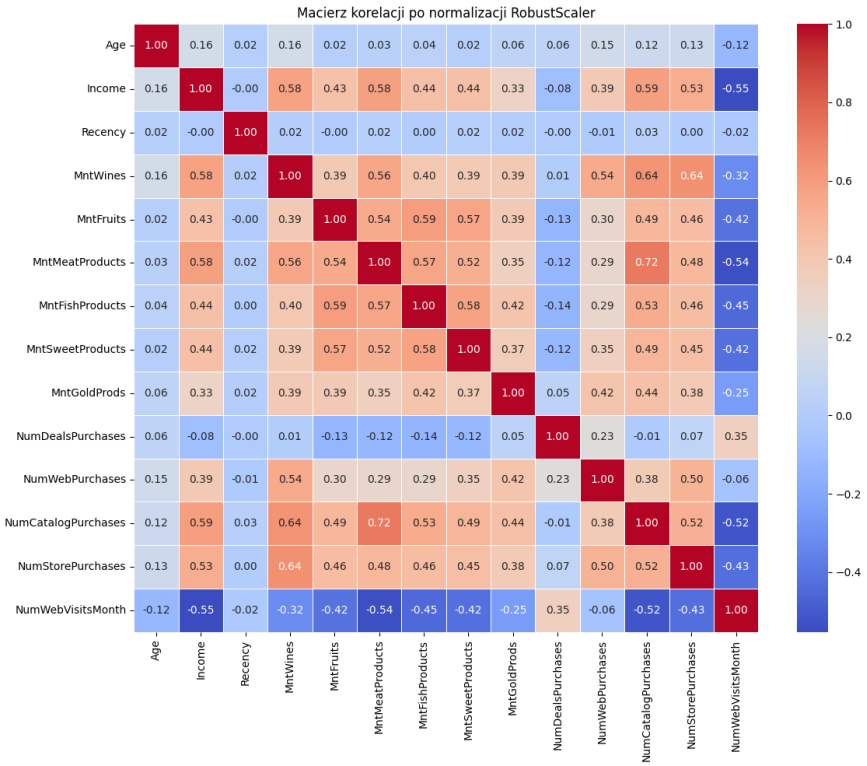
W tej sekcji przedstawiamy macierze korelacji dla danych po normalizacji min-max, standaryzacji, zastosowaniu RobustScaler oraz po przekształceniu zmiennych kategorycznych przy pomocy one-hot encodingu.



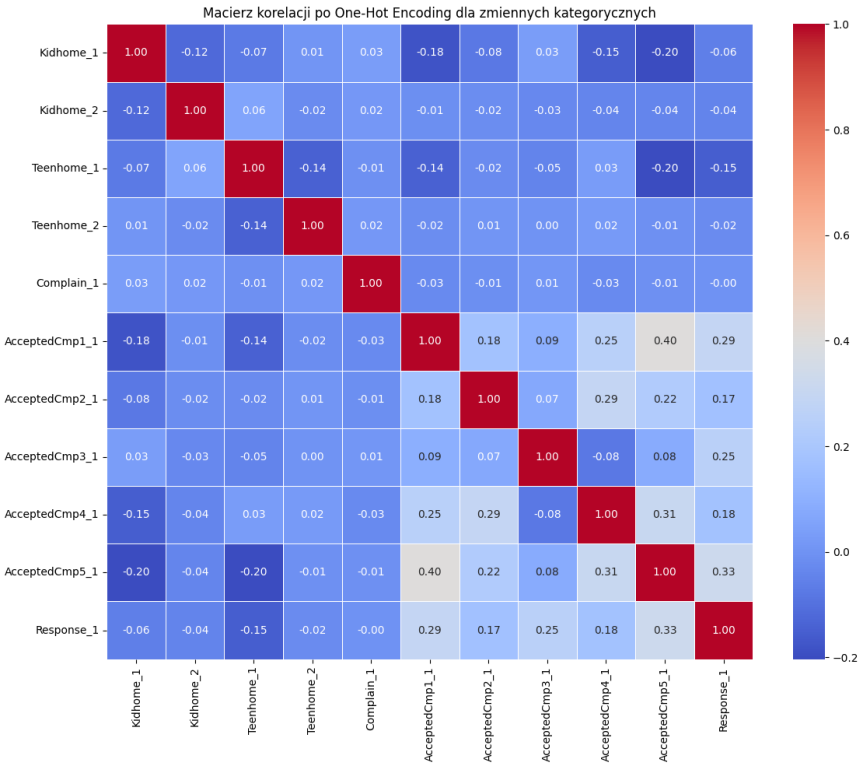
Rys. 15. Macierz korelacji dla danych po normalizacji Min-Max.



Rys. 16. Macierz korelacji dla danych po standaryzacji.



Rys. 17. Macierz korelacji dla danych ze zmiennych kategorycznych po zastosowaniu RobustScaler.



Rys. 18. Macierz korelacji dla danych po one-hot encoding.

1. Macierze korelacji po skalowaniu (RobustScaler, standaryzacja, min-max)

- Wszystkie trzy macierze korelacji (po zastosowaniu RobustScaler, standrayzacji oraz normalizacji min-max) są identyczne pod względem wartości współczynników korelacji Pearsona.
- Jest to zgodne z teorią, ponieważ korelacja Pearsona mierzy siłę i kierunek liniowej zależności między zmiennymi i jest niezależna od skali danych.
- Najsilniejsze dodatnie korelacje obserwuje się między:
  - `MntMeatProducts` a `NumStorePurchases` ($r = 0.72$),
  - `MntWines` a `NumCatalogPurchases` ($r = 0.64$),
  - `MntWines` a `NumStorePurchases` ($r = 0.64$).

- Najsilniejsze ujemne korelacje:
  - `Income` a `NumWebVisitsMonth` ($r = -0.55$),
  - `MntMeatProducts` a `NumWebVisitsMonth` ($r = -0.54$).
- Wysokie korelacje wewnątrz wydatków (np. `MntWines`, `MntMeatProducts`, `MntSweetProducts`) sugerują, że klienci wydający dużo w jednej kategorii, często wydają również w innych.

2. Macierz korelacji po One-Hot Encoding (zmienne kategoryczne)

- W analizie zmiennych binarnych zakodowanych za pomocą One-Hot Encodingu zaobserwowano ogólnie niskie wartości korelacji.
- Najwyższe dodatnie korelacje występują pomiędzy:
  - `AcceptedCmp1` a `AcceptedCmp5` ($r = 0.40$),
  - `AcceptedCmp5` a `Response` ($r = 0.33$),
  - `AcceptedCmp4` a `AcceptedCmp2` ($r = 0.29$).
- Te zależności mogą świadczyć o tym, że klienci reagujący pozytywnie na wcześniejsze kampanie są bardziej skłonni do odpowiedzi w przyszłości.
- Korelacje między zmiennymi takimi jak `Kidhome`, `Teenhome` czy `Complain` z innymi cechami są znikome, co może sugerować ich mniejszy wpływ na odpowiedzi klientów.

8 Omówienie metod klasteryzacji

W projekcie zastosowano dwie popularne metody klasteryzacji: k-średnich oraz GMM - metodę mieszanki gaussowskiej. Obie metody należą do klasy algorytmów nienadzorowanych, które służą do grupowania danych w oparciu o podobieństwa między obserwacjami. Wybór odpowiedniej metody klasteryzacji jest kluczowy, ponieważ może znacząco wpłynąć na jakość uzyskanych wyników oraz interpretację danych.

Dla obu rodzajów klasteryzacji (dla danych po standaryzacji, normalizacji min-max, a także normalizacji za pomocą RobustScaler) użyto danych należących do zmiennych kategorycznych po przejściu przez one-hot encoding.

8.1 Klasteryzacja k-średnich

Jako pierwszą metodę wybrano klasteryzację k-średnich. Jest to jeden z algorytmów najczęściej stosowanych w analizie segmentacji klientów, ponieważ:

- Ma niską złożoność obliczeniową, co sprawia, że działa efektywnie na dużych zbiorach danych, a z takim pracujemy przy projekcie.
- Algorytm jest szczególnie skuteczny w przypadku zmiennych ciągłych, jak w naszym przypadku, gdzie analizujemy dochody, wydatki oraz liczbę zakupów.

Celem algorytmu k-średnich jest wybranie centroidów, które minimalizują wariancję wewnątrzklastrową obliczaną wzorem:

$$\sum_{k=1}^{K} \sum_{x_l \in S_k} \|x_l - c_k\|^2$$

gdzie:

- K - liczba klastrów
- c_k - k-ty centroid
- S_k - zbiór punktów przypisanych do k-tego centroidu
- x_l - l-ty punkt zbioru S_k
- $\|x_i - c_k\|^2$ - kwadrat odległości euklidesowej między punktem x_l a centroidem c_k

Algorytm k-średnich składa się z trzech podstawowych kroków:

1. Wybór początkowych centroidów klastrów i powtarzanie dwóch następujących kroków, dopóki nie osiągnięta zostanie maksymalna liczba iteracji, bądź zmiana w centroidach będzie wystarczająco mała. Zmiana w centroidach jest określona przez odległość euklidesową między centroidami.

2. Przypisanie wszystkich punktów zbioru do najbliższego centroidu, odległość między punktem x a centroidem c jest obliczana wzorem:

$$\sqrt{\sum_{i=1}^{n}(x_i - c)^2}$$

gdzie:

- x_i - i-ta współrzędna punktu x
- c_i - i-ta współrzędna centroidu c
- n - liczba wymiarów punktów danych

3. Aktualizacja centroidów wzorem:

$$c_k = \frac{1}{|S_k|} \sum_{x_l \in S_k} x_l$$

gdzie:

- c_k - k-ty centroid
- S_k - zbiór punktów przypisanych do k-tego centroidu
- x_l - l-ty punkt zbioru S_k

8.2 Klasteryzacja GMM (metoda mieszanki gaussowskiej)

Metoda mieszanki gaussowskiej (GMM) jest probabilistycznym podejściem do klasteryzacji, które zakłada, że dane pochodzą z mieszanki rozkładów normalnych (Gaussa). Każdy klaster jest reprezentowany przez rozkład normalny, który ma określoną średnią oraz kowariancję.

Celem algorytmu GMM jest znalezienie najlepszych parametrów rozkładu (średnich, kowariancji) dla każdej z grup. GMM jest bardziej elastyczny niż k-średnie, ponieważ może wykrywać klastry o nieregularnych kształtach, takich jak elipsy.

GMM opiera się na algorytmie EM (Expectation Maximization), który składa się z dwóch etapów: estymacji oczekiwanych wartości i maksymalizacji parametrów.

Celem algorytmu jest maksymalizacja funkcji prawdopodobieństwa $p(x|\theta)$, gdzie x to wektory danych, a θ to zbiór parametrów (średnia, kowariancja, wagi). Funkcję prawdopodobieństwa dla modelu mieszanki gaussowskiej wyraża wzór:

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

gdzie:

- $p(x|\theta)$ – prawdopodobieństwo obserwacji x,
- K – liczba komponentów w mieszance,
- π_k – waga komponentu k (prawdopodobieństwo, że punkt x pochodzi z komponentu k),
- $\mathcal{N}(x|\mu_k, \Sigma_k)$ – rozkład normalny z średnią μ_k i macierzą kowariancji Σ_k.

Algorytm GMM składa się z dwóch głównych kroków: **Expectation** (E) oraz **Maximization** (M).

1. Estymacja prawdopodobieństw przynależności do klastra (expectation step):
$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}$$
gdzie γ_{ik} to prawdopodobieństwo, że punkt x_i pochodzi z klastra k.

2. Maksymalizacja parametrów (maximization step):
   - Waga komponentu π_k:
   $$\pi_k = \frac{N_k}{N}$$
   gdzie N_k to liczba punktów przypisanych do klastra k, a N to ogólna liczba punktów.
   - Średnia komponentu μ_k:
   $$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_i$$

gdzie μ_k to średnia komponentu k, a γ_{ik} to prawdopodobieństwo przynależności punktu x_i do komponentu k.

- Kowariancja komponentu Σ_k:

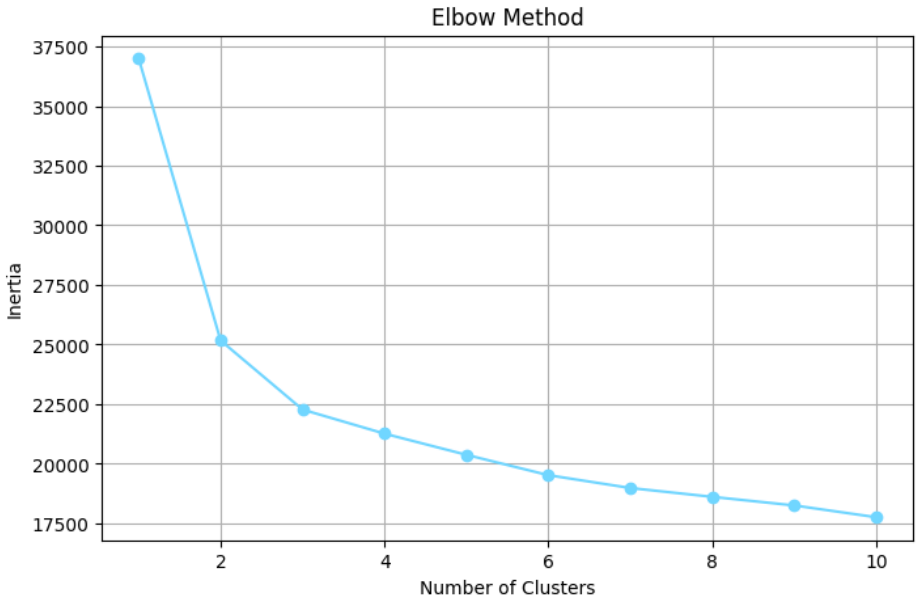
$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik}(x_i - \mu_k)(x_i - \mu_k)^T$$

gdzie Σ_k to macierz kowariancji komponentu k.

Algorytm powtarza kroki estymacji i maksymalizacji, aż wartości parametrów (średnich, kowariancji, wag) osiągną zbieżność, tj. zmiany w parametrach staną się minimalne lub nie będą już miały wpływu na wyniki.

9 Klasteryzacja dla danych po standaryzacji

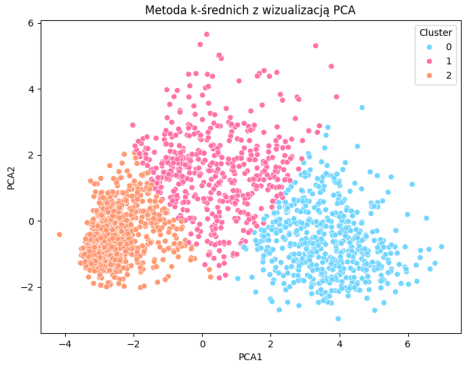
Zastosowano metodę łokcia do wyznaczenia optymalnej liczby klastrów.



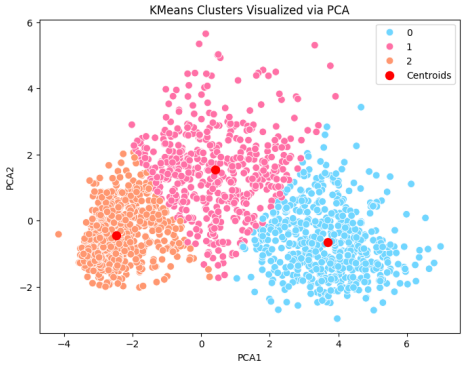
Rys. 19. Metoda łokcia w celu wyznaczenia optymalnej liczby klastrów.

Na podstawie otrzymanych wyników zdecydowano dokonać klasteryzacji dla 3 oraz 4 klastrów.

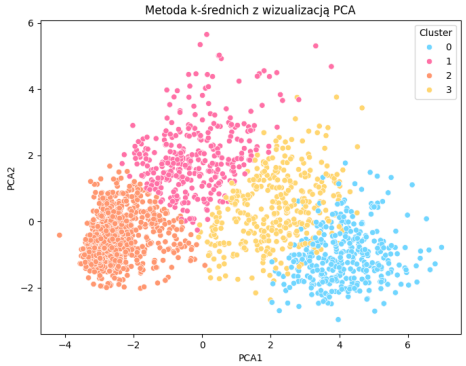
9.1 Klasteryzacja k-średnich



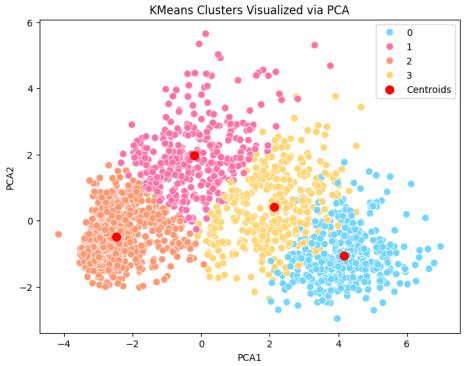
Rys. 20. Wyniki klasteryzacji k-średnich dla 3 klastrów.



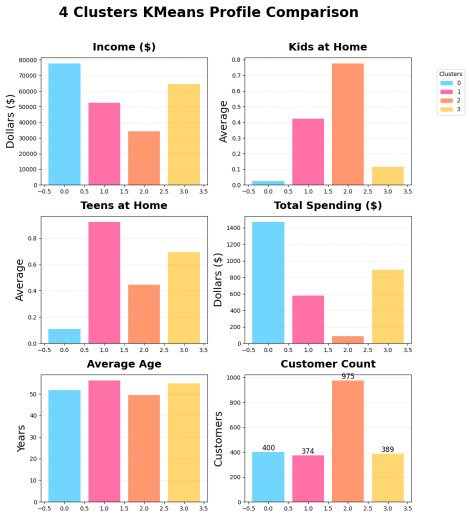
Rys. 21. Wizualizacja wyników klasteryzacji k-średnich dla 3 klastrów.



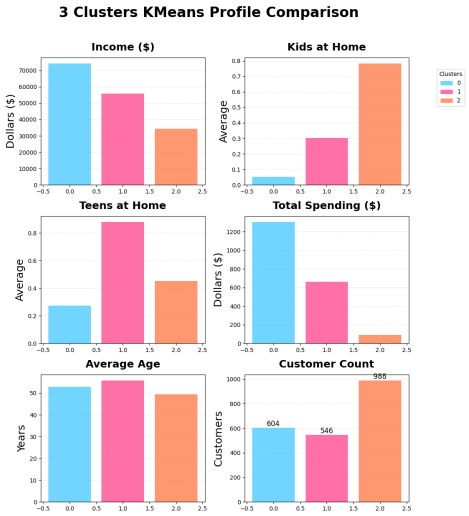
Rys. 22. Wyniki klasteryzacji k-średnich dla 4 klastrów.



Rys. 23. Wizualizacja wyników klasteryzacji k-średnich dla 4 klastrów.



Rys. 24. Wykres kolumnowy najważniejszych zmiennych metody k-średnich dla 4 klastrów.



Rys. 25. Wykres kolumnowy najważniejszych zmiennych metody k-średnich dla 3 klastrów.

Suma wyjaśnionej wariancji PCA1 oraz PCA2 wynosi 0.55.

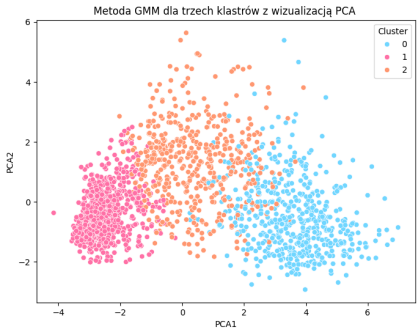
9.1.1 Wnioski

- **3 klastry:**
  - **Oszczędne rodziny z dziećmi** (1047 os.) - najniższe dochody, najwięcej dzieci, aktywni zakupowo (stacjonarnie/online), reagują na promocje, najwięcej reklamacji, wykształcenie licencjackie
  - **Bogate pary premium** (564 os.) - najwyższe dochody i wydatki (wino, mięso, ryby), głównie zakupy stacjonarne i katalogowe, najlepsza reakcja na kampanie, najmniej reklamacji
  - **Starsze małżeństwa** (605 os.) - średnie dochody, najwięcej nastolatków, wydatki na wino i złoto, wykształcenie magisterskie, aktywni zakupowo (online i stacjonarnie)
- **4 klastry:**
  - **Stabilne rodziny** (289 os.) - średnie dochody, najwięcej nastolatków, wrażliwi na promocje, zrównoważone zakupy

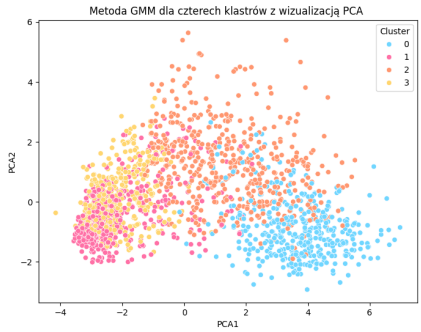
  - **Zamożni koneserzy** (473 os.) - najwyższe dochody i wydatki (produkty premium), lojalni wobec marek, najlepsza reakcja na kampanie
  - **Dojrzałe pary** (468 os.) - wysokie wydatki na wino, tradycyjne zakupy, najstarsza grupa, wykształcenie magisterskie
  - **Rodziny z małymi dziećmi** (986 os.) - najniższe dochody i wydatki, skupione na podstawowych potrzebach, brak reakcji na kampanie

- **Podsumowanie:** Podział na 3 klastry wydaje się bardziej optymalny (wg metody łokcia), z wyraźniejszym zróżnicowaniem grup. Kluczowe różnice to poziom dochodów, liczba dzieci i preferencje zakupowe.

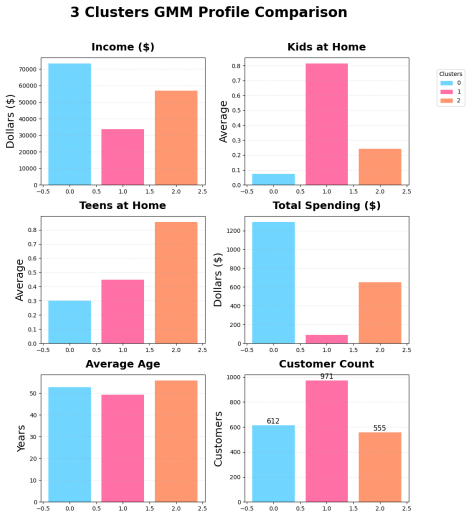
9.2 Klasteryzacja GMM



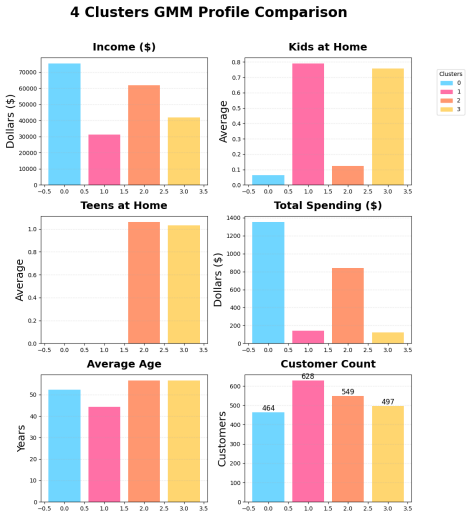
Rys. 26. Wizualizacja wyników klasteryzacji GMM dla 3 klastrów. Suma wyjaśnionej wariancji PCA1 oraz PCA2 wynosi 0.56



Rys. 27. Wizualizacja wyników klasteryzacji GMM dla 4 klastrów. Suma wyjaśnionej wariancji PCA1 oraz PCA2 wynosi 0.56



Rys. 28. Wykres kolumnowy najważniejszych zmiennych metody GMM dla 3 klastrów.



Rys. 29. Wykres kolumnowy najważniejszych zmiennych metody GMM dla 4 klastrów.

9.2.1 Wnioski

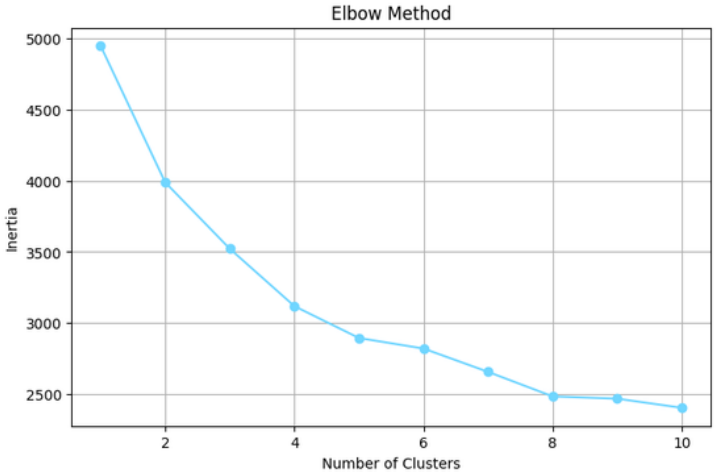
- **Dla 3 klastrów:**
  - **Klaster 0 (611 osób)** - Najbardziej wartościowi klienci: wysokie dochody (73k dolarów amerykańskich), duże wydatki (1290 dolarów amerykańskich), głównie na wino (555 dolarów amerykańskich) i mięso (375 dolarów amerykańskich). Preferują zakupy stacjonarne (8.66) i katalogowe (5.52). Najlepiej reagują na kampanie (19.5% Cmp5).
  - **Klaster 1 (556 osób)** - Rodziny z nastolatkami (0.85): średnie dochody (57k dolarów amerykańskich), wydatki (653 dolary amerykańskie). Aktywni online (5.93 zakupy, 5.66 wizyty). Chętnie korzystają z promocji (3.57 transakcje).
  - **Klaster 2 (971 osób)** - Najmniej atrakcyjni: niskie dochody (34k dolarów amerykańskich), minimalne wydatki (89 dolarów amerykańskich), dużo dzieci (0.81). Głównie zakupy podstawowe w sklepach (3.15).
- **Dla 4 klastrów:**

  - **Klaster 0 (339 osób)** - Elita: najwyższe dochody (77k dolarów amerykańskich) i wydatki (1487 dolarów amerykańskich), głównie wino (602 dolarów amerykańskich) i mięso (494 dolary amerykańskie). Najlepsza reakcja na kampanie (25.4% Cmp5).
  - **Klaster 1 (473 osoby)** - Rodziny z nastolatkami (0.92): średnie dochody (56k dolarów amerykańskich), wydatki (617 dolarów amerykańskich). Częste promocje (3.84 transakcje).
  - **Klaster 2 (968 osób)** - Najsłabsi: niskie dochody (34k dolarów amerykańskich), wydatki (89 dolarów amerykańskich), dużo dzieci (0.81). Brak reakcji na kampanie.
  - **Klaster 3 (358 osób)** - Zamożni (67k dochodu - w dolarach amerykańskich), wydatki 994 dolary amerykańskie (wino 487 dolarów amerykańskich, mięso 218 dolarów amerykańskich). Umiarkowana reakcja na promocje (10% Cmp5).

- **Podsumowanie:**
  - W obu wersjach wyraźnie widać segment najbogatszych klientów (Klastry 0) jako główny cel marketingowy
  - Rodziny z dziećmi stanowią osobną kategorię - te z nastolatkami są bardziej wartościowe
  - Najbiedniejsze rodziny mają marginalne znaczenie dla strategii sprzedażowej

10 Klasteryzacja dla danych po normalizacji min-max

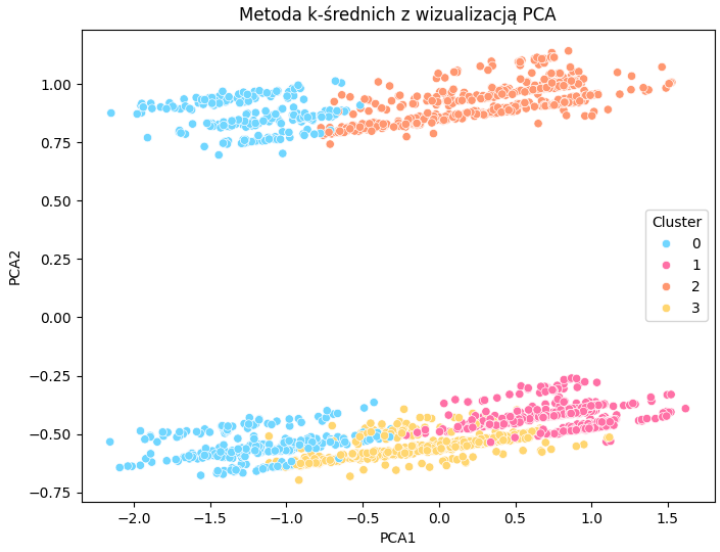
Zastosowano metodę łokcia do wyznaczenia optymalnej liczby klastrów.



Rys. 30. Metoda łokcia w celu wyznaczenia optymalnej liczby klastrów.

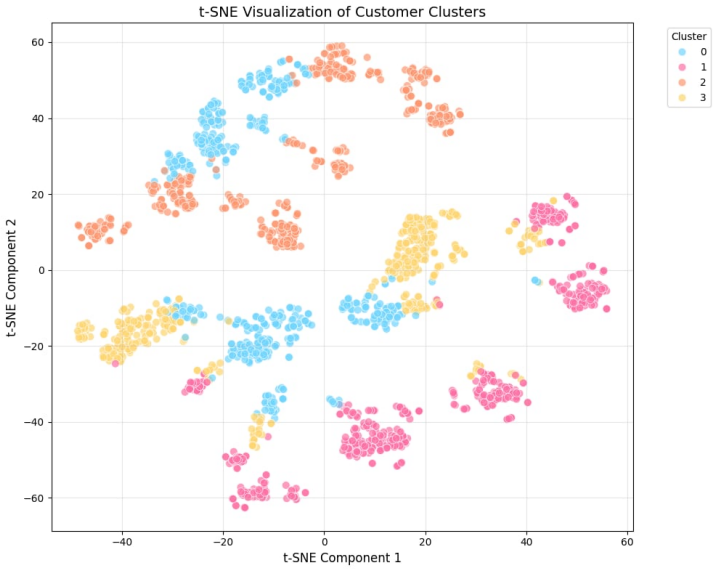
Na podstawie otrzymanych wyników zdecydowano się na wykonanie klasteryzacji dla 4 klastrów. W przypadku GMM zastosowano 3 oraz 4 klastry.

10.1 Klasteryzacja k-średnich

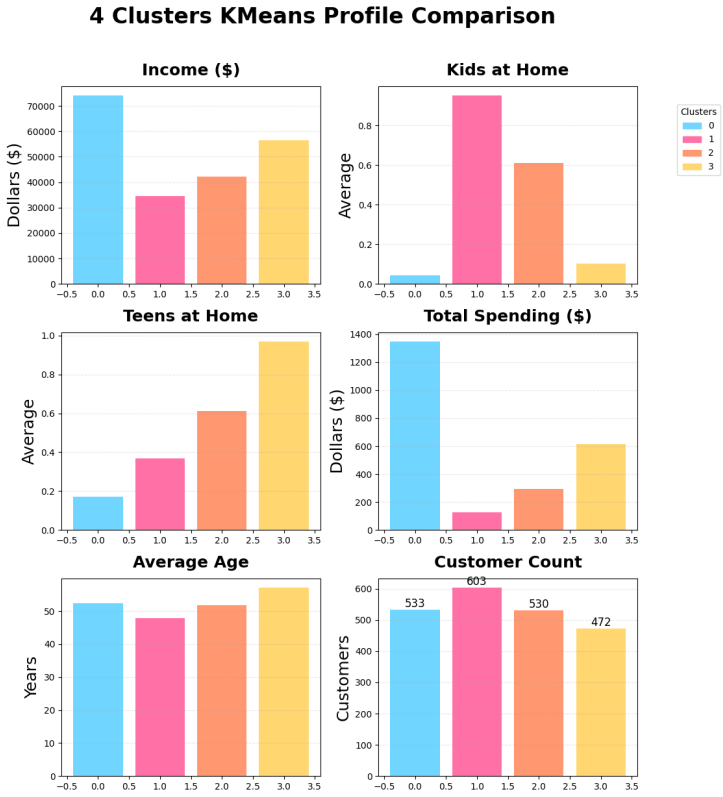


Rys. 31. Wyniki klasteryzacji k-średnich dla 4 klastrów z normalizacją min-max, z wizualizacją PCA.

Suma wyjaśnionej wariancji PCA1 oraz PCA2 wynosi 0.39.



Rys. 32. Wyniki klasteryzacji k-średnich dla 4 klastrów z normalizacją min-max, z wizualizacją t-SNE.



Rys. 33. Wykres kolumnowy najważniejszych zmiennych metody k-średnich z normalizacją min-max.

10.1.1 Wnioski

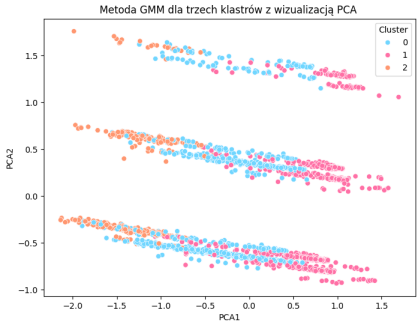
- **Dla 4 klastrów:**
  - **Klaster 0 (533 osoby)** - Najzamożniejsi klienci (74k dochodu) z niewielką liczbą dzieci. Rekordowe wydatki (1347), głównie na wino i mięso. Bardzo dobrze reagują na kampanie marketingowe, szczególnie na piątą i pierwszą kampanię. Preferują tradycyjne formy zakupów. Średni wiek 52 lata.
  - **Klaster 1 (603 osoby)** - Rodziny z małymi dziećmi: najniższe dochody i wydatki. Mało aktywni zakupowo, rzadko korzystają z kampanii. Najmłodsza grupa (średnio 48 lat).
  - **Klaster 2 (530 osób)** - Średniozamożne rodziny z nastolatkami.

Umiarkowane wydatki, głównie na podstawowe produkty. Średnia reakcja na promocje. Wiek około 52 lat.

- **Klaster 3 (472 osoby)** - Starsze osoby (średnio 57 lat) z dorastającymi dziećmi. Wyższe niż średnie dochody i wydatki. Rzadko korzystają z kampanii, preferują zakupy stacjonarne.

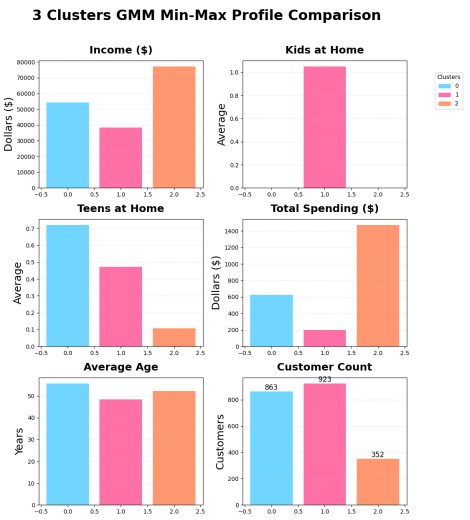
- **Podsumowanie:**
  - Najbardziej wartościową grupą są zamożni klienci bez dzieci (Klaster 0)
  - Rodziny z małymi dziećmi (Klaster 1) wymagają specjalnych programów lojalnościowych
  - Starsze osoby (Klaster 3) mogą być celem kampanii produktów premium
  - Średniozamożne rodziny (Klaster 2) stanowią stabilną bazę klientów

10.2 Klasteryzacja GMM

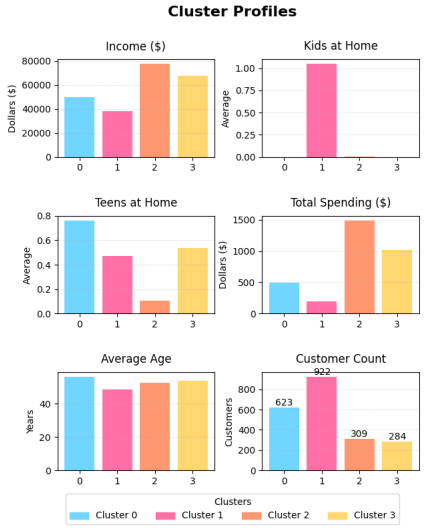


Rys. 34. Wizualizacja wyników klasteryzacji GMM dla 3 klastrów. Suma wyjaśnionej wariancji PCA1 oraz PCA2 wynosi 0.45

Rys. 35. Wizualizacja wyników klasteryzacji GMM dla 4 klastrów. Suma wyjaśnionej wariancji PCA1 oraz PCA2 wynosi 0.45



Rys. 36. Wykres kolumnowy najważniejszych zmiennych metody GMM dla 3 klastrów.



Rys. 37. Wykres kolumnowy najważniejszych zmiennych metody GMM dla 4 klastrów.

10.2.1 Wnioski

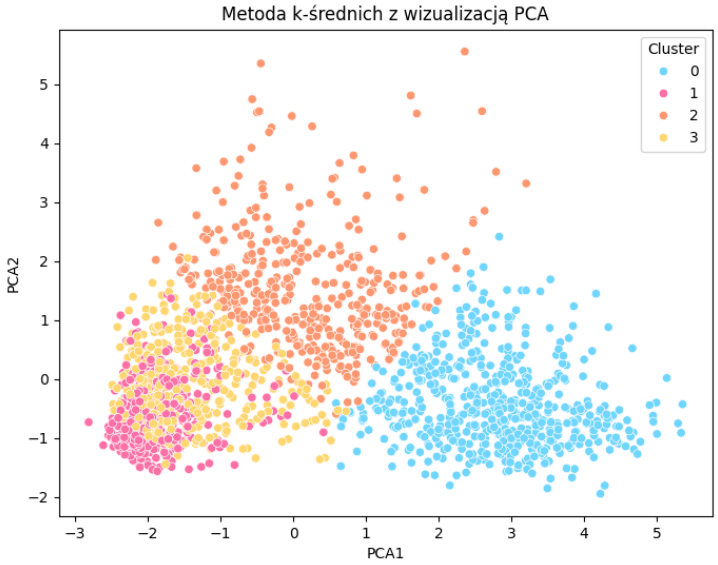
- **Dla 3 klastrów:**
  - **Klaster 0 (863 osoby)** - Średniozamożni klienci (54k dochodu - w dolarach amerykańskich) z nastolatkami. Umiarkowane wydatki (623 dolarów amerykańskich), głównie na wino i mięso. Częste zakupy online i promocyjne. Średnio reagują na kampanie marketingowe. Najstarsza grupa (56 lat).
  - **Klaster 1 (923 osoby)** - Rodziny z małymi dziećmi: najniższe dochody (38k dolarów amerykańskich) i wydatki (198 dolarów amerykańskich). Niska aktywność zakupowa, podstawowe produkty. Rzadko korzystają z promocji. Najmłodsza grupa (48 lat).
  - **Klaster 2 (352 osoby)** - Najzamożniejsi (77k dochodu - w dolarach amerykańskich), prawie bez dzieci. Rekordowe wydatki (1471 dolarów amerykańskich), szczególnie na wino i mięso. Najlepiej reagują na kampanie. Preferują zakupy stacjonarne i katalogowe.
- **Dla 4 klastrów:**

- **Klaster 0 (623 osoby)** - Starsze osoby (56 lat) z nastolatkami. Średnie dochody (50k dolarów amerykańskich), umiarkowane wydatki (493 dolary amerykańskie). Słabo reagują na kampanie. Częste wizyty online.
- **Klaster 1 (922 osoby)** - Podobny do 3-klastrowego: rodziny z małymi dziećmi, niskie dochody (38k dolarów amerykańskich) i wydatki (197 dolarów amerykańskich). Praktycznie brak reakcji na promocje.
- **Klaster 2 (309 osób)** - Najbogatsi (78k dochodu - w dolarach amerykańskich), bez dzieci. Najwyższe wydatki (1490 dolarów amerykańskich), doskonała reakcja na kampanie. Głównie zakupy premium.
- **Klaster 3 (284 osoby)** - Zamożni klienci (67k dochodu - w dolarach amerykańskich) z dorastającymi dziećmi. Wysokie wydatki (1017 dolarów amerykańskich), dobra reakcja na kampanie. Zrównoważone preferencje zakupowe.

- **Podsumowanie:**
  - Najbardziej wartościowi to zamożni klienci bez małych dzieci (Klastry 2)
  - Rodziny z małymi dziećmi (Klastry 1) wymagają specjalnego podejścia
  - Starsze osoby z nastolatkami (Klastry 0) stanowią średnio atrakcyjną grupę
  - Podział na 4 klastry lepiej wyodrębnia zamożnych klientów w średnim wieku

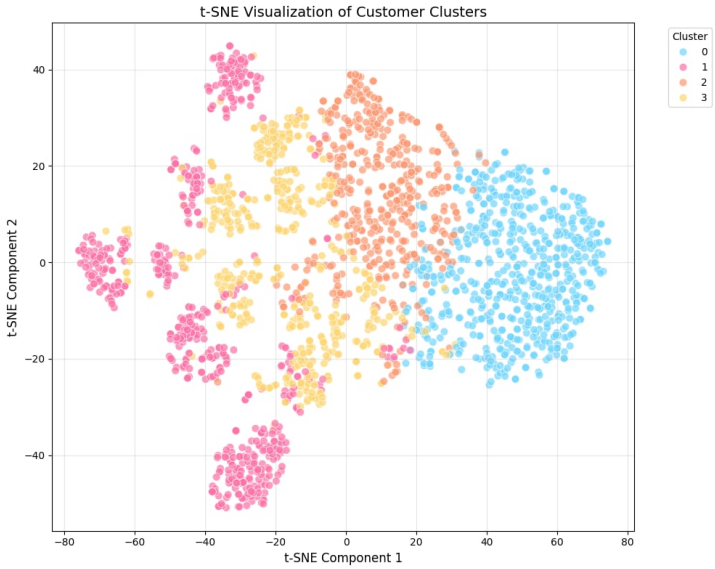
11 Klasteryzacja dla danych po zastosowaniu normalizacji RobustScaler

11.1 Klasteryzacja k-średnich

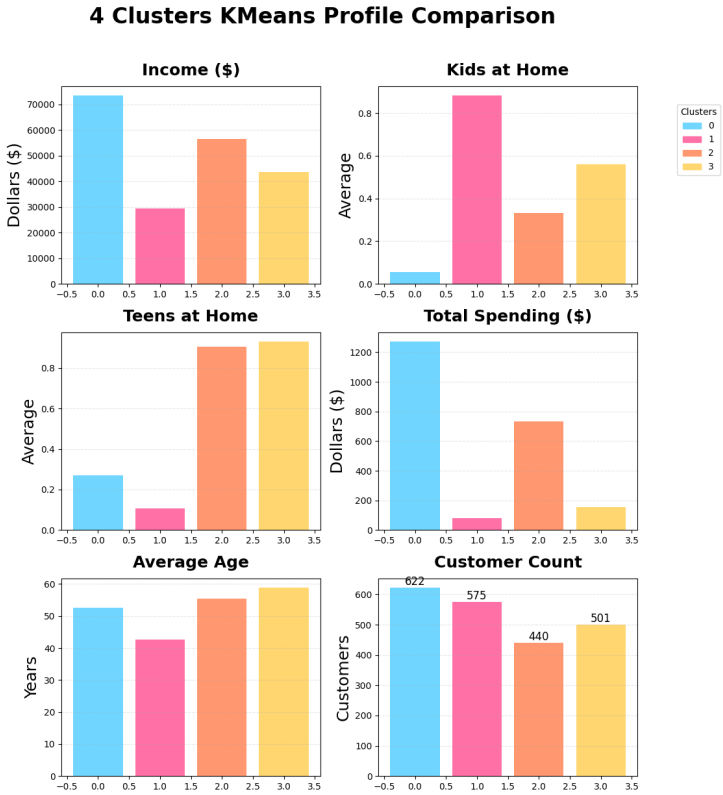


Rys. 38. Wyniki klasteryzacji k-średnich dla 4 klastrów z normalizacją min-max, z wizualizacją PCA.

Suma wyjaśnionej wariancji PCA1 oraz PCA2 wynosi 0.53.



Rys. 39. Wyniki klasteryzacji k-średnich dla 4 klastrów z normalizacją min-max, z wizualizacją t-SNE.



Rys. 40. Wykres kolumnowy najważniejszych zmiennych metody k-średnich z normalizacją RobustScaler.

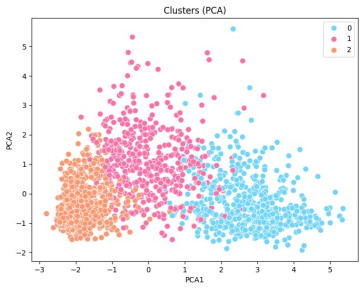
11.1.1 Wnioski

- **Dla 4 klastrów:**
  - **Klaster 0 (622 osoby)** - Najzamożniejsi (73k dochodu) z niewielką liczbą dzieci. Wysokie wydatki (1271), głównie na wino (550) i mięso (364). Aktywni w zakupach stacjonarnych (8.66) i katalogowych (5.42). Najlepiej reagują na kampanie marketingowe. Średni wiek 53 lata.
  - **Klaster 1 (575 osób)** - Rodziny z małymi dziećmi: najniższe dochody (29k) i wydatki (81). Minimalna aktywność zakupowa, podstawowe produkty. Rzadko korzystają z kampanii. Najmłodsza grupa (43 lata).

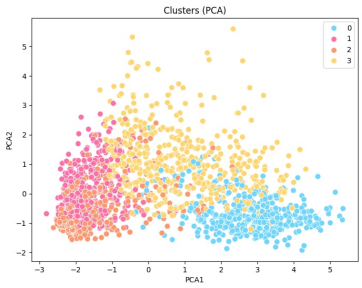
  - **Klaster 2 (440 osób)** - Średniozamożni (57k) z nastolatkami. Znaczne wydatki (734), szczególnie na wino (483). Częste zakupy online (6.59) i promocyjne (4.42). Średnia reakcja na kampanie. Starsza grupa (55 lat).
  - **Klaster 3 (501 osób)** - Starsze osoby (59 lat) z dorosłymi dziećmi. Niskie wydatki (155), głównie podstawowe produkty. Praktycznie brak reakcji na kampanie. Najstarsza grupa.

- **Podsumowanie:**
  - Najbardziej wartościowa grupa to zamożni klienci bez małych dzieci (Klaster 0)
  - Rodziny z małymi dziećmi (Klaster 1) wymagają specjalnych programów lojalnościowych
  - Rodziny z nastolatkami (Klaster 2) mogą być celem kampanii produktów średniej półki
  - Starsze osoby (Klaster 3) są mało atrakcyjne pod względem marketingowym

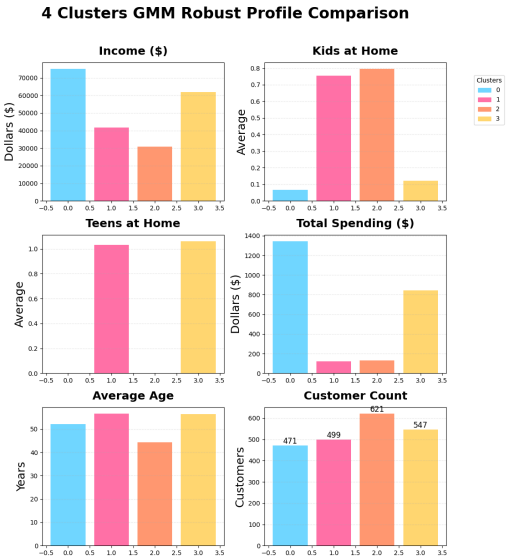
11.2 Klasteryzacja GMM



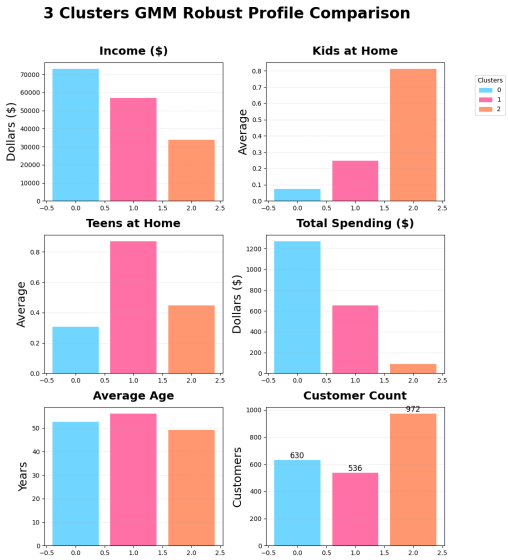
Rys. 41. Wizualizacja wyników klasteryzacji GMM dla 3 klastrów. Suma wyjaśnionej wariancji PCA1 oraz PCA2 wynosi 0.55



Rys. 42. Wizualizacja wyników klasteryzacji GMM dla 4 klastrów. Suma wyjaśnionej wariancji PCA1 oraz PCA2 wynosi 0.55



Rys. 43. Wykres kolumnowy najważniejszych zmiennych metody GMM dla 4 klastrów z normalizacją RobustScaler.



Rys. 44. Wykres kolumnowy najważniejszych zmiennych metody GMM dla 3 klastrów z normalizacją RobustScaler.

11.2.1 Wnioski

- **Dla 3 klastrów:**
  - **Klaster 0 (630 osób)** - Najzamożniejsi klienci (73k dochodu - w dolarach amerykańskich) z niewielką liczbą dzieci. Wysokie wydatki (1270 dolarów amerykańskich), głównie na wino i mięso. Aktywni w zakupach stacjonarnych i katalogowych. Najlepiej reagują na kampanie marketingowe. Średni wiek 53 lata.
  - **Klaster 1 (536 osób)** - Rodziny z nastolatkami: średnie dochody (57k dolarów amerykańskich), umiarkowane wydatki (655 dolarów amerykańskich). Częste zakupy promocyjne. Średnio reagują na kampanie. Starsza grupa (56 lat).
  - **Klaster 2 (972 osoby)** - Rodziny z małymi dziećmi: najniższe dochody (34k dolarów amerykańskich) i wydatki (89 dolarów amerykańskich). Podstawowe zakupy, głównie stacjonarne. Praktycznie brak reakcji na kampanie. Najmłodsza grupa (49 lat).

- **Dla 4 klastrów:**
  - **Klaster 0 (471 osób)** - Najbogatsi (75k dochodu - w dolarach amerykańskich), minimalna liczba dzieci. Rekordowe wydatki (1343 dolarów amerykańskich), szczególnie na wino i mięso. Doskonała reakcja na kampanie. Preferują zakupy katalogowe.
  - **Klaster 1 (499 osób)** - Rodziny z nastolatkami: niskie dochody (42k dolarów amerykańskich), minimalne wydatki (123 dolarów amerykańskich). Rzadko korzystają z promocji. Starsza grupa (57 lat).
  - **Klaster 2 (621 osób)** - Młodzi rodzice (44 lata) z małymi dziećmi: najniższe dochody (31k dolarów amerykańskich) i wydatki (133 dolarów amerykańskich). Podstawowe zakupy, słaba reakcja na kampanie.
  - **Klaster 3 (547 osób)** - Zamożni klienci (62k dochodu - w dolarach amerykańskich) z dorastającymi dziećmi. Wysokie wydatki (844 dolary amerykańskie), dobra reakcja na kampanie. Zrównoważone preferencje zakupowe.
- **Podsumowanie:**
  - Najbardziej wartościową grupę stanowią klienci z wysokimi dochodami i brakiem małych dzieci (klastry 0).
  - Rodziny z dziećmi dzielą się na dwie odrębne grupy: z małymi dziećmi (mniej atrakcyjne z punktu widzenia marketingowego) oraz z nastolatkami (o umiarkowanym potencjale zakupowym).
  - Podział klientów na cztery klastry pozwolił na lepsze zróżnicowanie rodzin z dziećmi – szczególnie pod względem wieku potomstwa.
  - Najstarsze grupy klientów (średnia wieku 56–57 lat) mogą wymagać specjalnie dostosowanych kampanii marketingowych, które uwzględniają ich potrzeby i preferencje zakupowe.

12 Rezultaty oraz omówienie wyników

Do porównania podobieństwa klastrów z różnych wyników użyty został indeks Jaccarda, którego wartość jest równa liczności częśc wspólnej obu zbiorów (klastrów) podzielonej przez liczność sumy obu zbiorów (klastrów). Im

bliższa wartość indeksu do 1, tym bardziej podobne są zbiory, a im bliższa do 0 tym mniej.

Podobieństwo pomiędzy porównywanymi zbriorami klastrów A i B opisuje wzór:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

	GMM Std	GMM MinMax	GMM Robust	K-means Std	K-means MinMax	K-means Robust
GMM Std	0: 0 1: 1 2: 2 3: 3	0: 2 1: 1 2: 0 3: 3	0: 0 1: 2 2: 3 3: 1	0: 0 1: 2 2: 3 3: 1	0: 0 1: 1 2: 3 3: 2	0: 0 1: 1 2: 2 3: 3
GMM MinMax	0: 2 1: 1 2: 0 3: 3	0: 0 1: 1 2: 2 3: 3	0: 3 1: 2 2: 0 3: 1	0: 1 1: 2 2: 0 3: 3	0: 3 1: 1 2: 0 3: 2	0: 2 1: 1 2: 0 3: 3
GMM Robust	0: 0 1: 3 2: 1 3: 2	0: 2 1: 3 2: 1 3: 0	0: 0 1: 1 2: 2 3: 3	0: 0 1: 1 2: 2 3: 3	0: 0 1: 2 2: 1 3: 3	0: 0 1: 3 2: 1 3: 2
K-means Std	0: 0 1: 3 2: 1 3: 2	0: 2 1: 0 2: 1 3: 3	0: 0 1: 1 2: 2 3: 3	0: 0 1: 1 2: 2 3: 3	0: 0 1: 2 2: 1 3: 3	0: 0 1: 2 2: 1 3: 3
K-means MinMax	0: 0 1: 1 2: 3 3: 2	0: 2 1: 1 2: 3 3: 0	0: 0 1: 2 2: 1 3: 3	0: 0 1: 2 2: 1 3: 3	0: 0 1: 1 2: 2 3: 3	0: 0 1: 1 2: 3 3: 2
K-means Robust	0: 0 1: 1 2: 2 3: 3	0: 2 1: 1 2: 0 3: 3	0: 0 1: 2 2: 3 3: 1	0: 0 1: 2 2: 1 3: 3	0: 0 1: 1 2: 3 3: 2	0: 0 1: 1 2: 2 3: 3

Tabela 3: Przedstawienie, które klastry z danej klasteryzacji odpowiadają którym klastrom z innych klasteryzacji

Z tabeli 3 wynika, że porównanie klastrów między różnymi metodami klasteryzacji wykazuje różnice, jednak zachowana jest relatywnie wysoka zgodność pomiędzy wynikami uzyskanymi dla różnych metod normalizacji.

Tabela ta przedstawia, które klastry z danej klasteryzacji odpowiadają którym klastrom z innych metod klasteryzacji. Wartości w tabeli są zapisane w formacie **w: k, w: k, w: k, w: k**, gdzie:

- **w** to numer klastra z metody, która jest wierszem,
- **k** to numer klastra z metody, która jest kolumną,
- każda komórka zawiera zestaw par (w: k), gdzie każda para wskazuje, który klaster z jednej metody odpowiada klastrowi z innej metody.

Na przykład w wierszu "GMM MinMax"i kolumnie "K-means Std", wartość **0: 1** oznacza, że klaster 0 w metodzie GMM MinMax odpowiada klastrowi 1 w metodzie K-means Std.

Ta tabela pozwala na łatwe porównanie przypisania klastrów między różnymi metodami klasteryzacji, pokazując, jak klastry są ze sobą powiązane w różnych podejściach.

	GMM Std	GMM MinMax	GMM Robust	K-means Std	K-means MinMax	K-means Robust
GMM Std	1.0	0.33	0.99	0.41	0.43	0.64
GMM MinMax	0.33	1.0	0.33	0.49	0.39	0.33
GMM Robust	0.99	0.33	1.0	0.41	0.43	0.65
K-means Std	0.41	0.49	0.41	1.0	0.39	0.43
K-means MinMax	0.43	0.39	0.43	0.39	1.0	0.43
K-means Robust	0.64	0.33	0.65	0.43	0.43	1.0

Tabela 4: Podobieństwo pomiędzy wynikami klasteryzacji

Podobieństwo pomiędzy wynikami klasteryzacji A i B opisuje wzór:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Jak widać, największe podobieństwo, bo wynoszące aż **0.99** obserwujemy przy porównaniu GMM Robust z GMM Std.

Za to najmniejsze, wynoszące jedynie **0.33** to podobieństwo pomiędzy GMM MinMax a:

- GMM Std,
- GMM Robust,
- K-means Robust.

Na podstawie wartości indeksu Jaccarda można wyciągnąć następujące wnioski dotyczące spójności klasteryzacji oraz wpływu metody normalizacji na wyniki:

- Niskie podobieństwo GMM MinMax do pozostałych metod (0,33) wskazuje, że MinMaxScaler znacząco zmienia rozkład danych, co prowadzi do zupełnie innych wyników segmentacji. Może to być spowodowane większą wrażliwością tej metody normalizacji na wartości odstające.

- Dobór metody normalizacji ma istotny wpływ na wyniki klasteryzacji. Dla bardziej spójnych i powtarzalnych wyników rekomendowane jest stosowanie RobustScaler lub StandardScaler zamiast MinMaxScaler.

Cecha	GMM Std	GMM MinMax	GMM Robust	K-means Std	K-means MinMax	K-means Robust
Income	75390.96	52635.05	29334.93	77654.62	55403.59	29334.93
Kidhome	0.06	0.0	0.88	0.03	0.0	0.88
Teenhome	0.0	0.91	0.11	0.11	0.47	0.11
Recency	49.4	48.48	48.38	51.09	43.47	48.38
MntWines	596.56	362.17	26.14	622.56	730.88	26.14
MntFruits	49.03	11.62	5.29	53.08	40.26	5.29
MntMeatProducts	398.07	84.47	21.59	446.37	208.24	21.59
MntFishProducts	77.67	15.57	7.82	87.09	66.62	7.82
MntSweetProducts	49.82	9.67	5.43	55.31	32.41	5.43
MntGoldProds	64.54	40.77	14.76	68.75	82.41	14.76
NumDealsPurchases	1.11	2.62	1.81	1.11	4.35	1.81
NumWebPurchases	4.99	5.02	2.07	4.88	8.0	2.07
NumCatalogPurchases	5.68	2.31	0.43	6.16	4.41	0.43
NumStorePurchases	8.55	6.45	2.99	8.41	8.94	2.99
NumWebVisitsMonth	2.77	5.54	6.93	2.54	6.82	6.93
AcceptedCmp3	0.08	0.06	0.09	0.09	0.24	0.09
AcceptedCmp4	0.12	0.13	0.0	0.11	0.06	0.0
AcceptedCmp5	0.25	0.02	0.0	0.27	0.12	0.0
AcceptedCmp1	0.22	0.03	0.0	0.23	0.06	0.0
AcceptedCmp2	0.03	0.02	0.0	0.02	0.12	0.0
Complain	0.01	0.01	0.01	0.0	0.0	0.01
Response	0.29	0.08	0.12	0.28	0.47	0.12
Age	52.22	58.44	42.73	51.67	56.94	42.73
Spent	1352.14	529.75	81.06	1468.62	1221.18	81.06
Education	Graduate	Postg.	Graduate	Graduate	Postg.	Graduate
Marital Status	Partner	Partner	Partner	Partner	Alone	Partner
Spending To Income Ratio	0.02	0.01	0.0	0.02	0.02	0.0
Count	464	507	575	400	17	575

Tabela 5: Porównanie klastra 0

Cecha	GMM Std	GMM MinMax	GMM Robust	K-means Std	K-means MinMax	K-means Robust
Income	31148.62	45609.33	43728.56	52437.86	42488.6	43728.56
Kidhome	0.79	1.06	0.56	0.42	1.0	0.56
Teenhome	0.0	0.94	0.93	0.92	0.79	0.93
Recency	49.05	48.99	49.57	46.63	52.62	49.57
MntWines	49.1	156.13	90.85	397.06	110.97	90.85
MntFruits	8.94	7.3	5.13	9.25	6.24	5.13
MntMeatProducts	35.48	58.12	30.58	94.88	43.6	30.58
MntFishProducts	13.26	10.46	7.26	14.65	8.22	7.26
MntSweetProducts	9.01	8.06	5.26	10.11	6.28	5.26
MntGoldProds	21.47	27.16	15.74	49.13	20.34	15.74
NumDealsPurchases	1.93	3.96	2.18	4.68	3.33	2.18
NumWebPurchases	2.6	3.52	2.48	6.25	3.05	2.48
NumCatalogPurchases	0.67	1.15	0.89	2.35	0.91	0.89
NumStorePurchases	3.51	4.52	3.99	6.64	4.17	3.99
NumWebVisitsMonth	6.72	6.32	5.44	6.7	6.15	5.44
AcceptedCmp3	0.09	0.06	0.04	0.08	0.06	0.04
AcceptedCmp4	0.01	0.04	0.04	0.16	0.04	0.04
AcceptedCmp5	0.0	0.0	0.0	0.01	0.0	0.0
AcceptedCmp1	0.0	0.01	0.0	0.03	0.02	0.0
AcceptedCmp2	0.0	0.0	0.0	0.02	0.01	0.0
Complain	0.01	0.01	0.01	0.01	0.01	0.01
Response	0.12	0.12	0.04	0.17	0.09	0.04
Age	44.28	55.55	58.85	56.17	55.84	58.85
Spent	139.32	269.43	154.95	579.63	197.14	154.95
Education	Graduate	Postg.	Postg.	Postg.	Graduate	Postg.
Marital Status	Partner	Partner	Partner	Partner	Partner	Partner
Spending To Income Ratio	0.0	0.0	0.0	0.01	0.0	0.0
Count	628	403	501	374	229	501

Tabela 6: Porównanie klastra 1

Cecha	GMM Std	GMM MinMax	GMM Robust	K-means Std	K-means MinMax	K-means Robust
Income	61862.95	66129.0	73425.08	34327.23	48896.44	73425.08
Kidhome	0.12	0.0	0.06	0.78	0.48	0.06
Teenhome	1.06	1.0	0.27	0.45	0.93	0.27
Recency	48.04	63.17	49.87	49.4	46.22	49.87
MntWines	497.73	501.17	549.96	37.24	254.21	549.96
MntFruits	28.01	11.33	52.52	4.75	9.76	52.52
MntMeatProducts	165.95	360.17	364.42	20.91	73.47	364.42
MntFishProducts	38.7	39.67	80.59	6.82	12.41	80.59
MntSweetProducts	28.49	17.0	53.89	4.88	8.95	53.89
MntGoldProds	60.07	19.17	67.64	14.04	35.09	67.64
NumDealsPurchases	3.56	3.33	1.44	1.9	3.11	1.44
NumWebPurchases	6.46	6.0	5.37	2.01	4.3	5.37
NumCatalogPurchases	3.63	6.5	5.42	0.51	1.86	5.42
NumStorePurchases	8.19	9.67	8.66	3.19	5.36	8.66
NumWebVisitsMonth	5.33	4.5	3.08	6.33	6.02	3.08
AcceptedCmp3	0.07	0.0	0.07	0.07	0.08	0.07
AcceptedCmp4	0.14	0.17	0.09	0.01	0.09	0.09
AcceptedCmp5	0.02	0.17	0.19	0.0	0.01	0.19
AcceptedCmp1	0.04	0.0	0.17	0.0	0.02	0.17
AcceptedCmp2	0.01	0.0	0.02	0.0	0.01	0.02
Complain	0.01	0.0	0.01	0.01	0.01	0.01
Response	0.1	0.33	0.23	0.09	0.14	0.23
Age	56.49	63.33	52.56	49.38	57.7	52.56
Spent	841.7	948.5	1271.35	88.72	397.13	1271.35
Education	Graduate	Postg.	Graduate	Graduate	Postg.	Graduate
Marital Status	Partner	Partner	Partner	Partner	Alone	Partner
Spending To Income Ratio	0.01	0.01	0.02	0.0	0.0	0.02
Count	549	6	622	975	316	622

Tabela 7: Porównanie klastra 2

Cecha	GMM Std	GMM MinMax	GMM Robust	K-means Std	K-means MinMax	K-means Robust
Income	41852.29	52191.8	56500.21	64687.47	54472.49	56500.21
Kidhome	0.76	0.0	0.33	0.12	0.13	0.33
Teenhome	1.03	0.8	0.9	0.69	1.0	0.9
Recency	49.75	51.4	48.07	48.29	49.06	48.07
MntWines	68.52	331.8	482.56	465.58	368.54	482.56
MntFruits	3.88	42.76	16.81	42.05	12.59	16.81
MntMeatProducts	26.09	142.6	128.76	197.02	92.97	128.76
MntFishProducts	5.41	52.92	22.81	54.84	17.77	22.81
MntSweetProducts	3.89	58.08	16.07	40.31	12.91	16.07
MntGoldProds	14.71	57.16	57.44	62.11	42.26	57.44
NumDealsPurchases	2.6	3.64	4.42	2.39	3.21	4.42
NumWebPurchases	2.36	5.68	6.59	6.22	5.14	6.59
NumCatalogPurchases	0.69	3.88	3.0	3.97	2.36	3.0
NumStorePurchases	3.56	7.44	7.58	8.94	6.68	7.58
NumWebVisitsMonth	6.01	5.44	6.35	4.42	5.52	6.35
AcceptedCmp3	0.05	0.12	0.09	0.06	0.05	0.09
AcceptedCmp4	0.02	0.0	0.15	0.07	0.13	0.15
AcceptedCmp5	0.0	0.0	0.02	0.05	0.01	0.02
AcceptedCmp1	0.01	0.0	0.04	0.05	0.02	0.04
AcceptedCmp2	0.0	0.0	0.02	0.01	0.01	0.02
Complain	0.01	0.0	0.01	0.02	0.01	0.01
Response	0.07	0.2	0.16	0.11	0.06	0.16
Age	56.53	59.44	55.45	54.83	57.77	55.45
Spent	122.51	705.12	733.75	892.39	551.67	733.75
Education	Graduate	Graduate	Postg.	Graduate	Postg.	Postg.
Marital Status	Partner	Partner	Partner	Partner	Partner	Partner
Spending To Income Ratio	0.0	0.01	0.01	0.01	0.0	0.01
Count	497	25	440	389	379	440

Tabela 8: Porównanie klastra 3

Analiza cech klientów w poszczególnych klastrach, przedstawiona w tabelach 5, 6, 7 i 8, ujawnia istotne różnice w takich zmiennych jak dochód, liczba dzieci, preferencje zakupowe (np. wydatki na wino, mięso czy owoce), oraz liczba dokonanych zakupów w różnych kanałach (strona, katalog, sklep). Z tych wyników można wyciągnąć wnioski o charakterystyce poszczególnych segmentów klientów, co pozwala na precyzyjniejsze dopasowanie oferty do ich potrzeb.

Cecha	Klaster 0	Klaster 1	Klaster 2	Klaster 3
Income	50199.47	42147.03	56511.54	54209.24
Kidhome	0.41	0.71	0.34	0.35
Teenhome	0.25	0.7	0.54	0.91
Recency	48.98	49.21	49.05	48.7
MntWines	296.27	136.08	348.56	363.04
MntFruits	22.31	7.05	28.64	17.93
MntMeatProducts	171.17	46.22	191.01	111.39
MntFishProducts	35.03	10.31	42.82	24.03
MntSweetProducts	22.44	7.36	29.24	17.56
MntGoldProds	38.02	23.99	45.96	45.84
NumDealsPurchases	1.75	2.85	2.14	3.41
NumWebPurchases	3.68	3.25	4.39	5.29
NumCatalogPurchases	2.69	1.09	3.19	2.55
NumStorePurchases	5.59	4.35	6.52	6.74
NumWebVisitsMonth	5.2	6.12	4.81	5.77
AcceptedCmp3	0.08	0.06	0.07	0.07
AcceptedCmp4	0.07	0.05	0.08	0.1
AcceptedCmp5	0.09	0.0	0.08	0.02
AcceptedCmp1	0.08	0.01	0.08	0.03
AcceptedCmp2	0.01	0.01	0.01	0.01
Complain	0.01	0.01	0.01	0.01
Response	0.17	0.1	0.15	0.11
Age	49.11	54.23	52.8	56.04
Spent	629.39	232.65	731.82	590.06
Spending To Income Ratio	0.01	0.0	0.01	0.01
Count	423.0	439.33	515.0	361.67

Tabela 9: Porównanie średnich wartości dla najbardziej podobnych do siebie klastrów

13 Podsumowanie

Przeprowadzona segmentacja klientów przy użyciu algorytmów k-średnich oraz GMM pozwoliła na wyodrębnienie czterech wyraźnych grup o odmiennych cechach demograficznych i zachowaniach zakupowych. Wyniki klasteryzacji stanowią cenną podstawę do budowy zróżnicowanych kampanii marketingowych.

Porównanie metod wskazuje, że:

- GMM w połączeniu z RobustScaler zapewnia najbardziej spójne i stabilne rezultaty.

- Metoda MinMaxScaler znacząco wpływa na strukturę klastrów i obniża ich zgodność (indeks Jaccarda na poziomie 0,33), co sugeruje jej ograniczoną przydatność w analizowanym kontekście.

Z tabel wynika, że poszczególne klastry można scharakteryzować następująco:

- **Klaster 0:** Najbardziej wartościowi pod względem potencjału zakupowego klienci – osoby o wysokich dochodach, niskiej liczbie dzieci, znaczących wydatkach (zwłaszcza na wino i mięso) oraz aktywności zakupowej. Jest to grupa najbardziej atrakcyjna z perspektywy działań marketingowych.

- **Klaster 1:** Rodziny z małymi dziećmi, o relatywnie niskich dochodach i niskim poziomie wydatków. Stanowią mniej dochodowy segment, potencjalnie bardziej wrażliwy cenowo.

- **Klaster 2:** Klienci o stabilnych dochodach, częściej z nastoletnimi dziećmi, zrównoważone wydatki i aktywność zakupowa zarówno online, jak i offline. Grupa średnio atrakcyjna, ale stabilna.

- **Klaster 3:** Starsze osoby (średnio 56 lat), z umiarkowanymi dochodami i większą aktywnością w kanałach online. Wymagają one odrębnego podejścia marketingowego, uwzględniającego ich wiek oraz cyfrowe nawyki.

Z analizy wynika, że podział na cztery klastry skutecznie różnicuje klientów nie tylko pod względem wydatków, ale również wieku, obecności dzieci oraz kanałów zakupowych. Co istotne, segmentacja pozwala lepiej zrozumieć różnice między rodzinami z małymi dziećmi a rodzinami z nastolatkami — co wcześniej mogło być trudne do uchwycenia.

Rekomendacje marketingowe:

- Skoncentrować kampanie premium na klastrze 0 (największy potencjał dochodowy).

- Opracować dedykowane oferty dla rodzin z małymi dziećmi (klaster 1), np. pakiety rodzinne lub zniżki ilościowe.

- W klastrze 2 warto testować strategie lojalnościowe i kampanie promujące zakupy online.

- W przypadku klastra 3 kluczowe jest dopasowanie komunikacji do preferencji starszych klientów i promowanie kanałów cyfrowych w sposób przyjazny i zrozumiały.

Dalsze działania powinny obejmować testy A/B oraz monitorowanie skuteczności kampanii dla każdego z segmentów.

14 Bibliografia

Literatura

[1] Ramireddy, K. (2019). *Customer Segmentation Analysis for Amazon Data*. California State University. Dostępny pod adresem: `https://scholarworks.calstate.edu/downloads/wm117x269`

[2] Acheme, D., & Enoyoze, E. (2023). Customer personality analysis and clustering for targeted marketing. *ResearchGate*. Dostępny pod adresem: `https://www.researchgate.net/publication/381943308`

[3] Hultén, B. (2007). Customer segmentation: The concepts of trust, commitment and relationships. *ResearchGate*. Dostępny pod adresem: `https://www.researchgate.net/publication/31963685`

[4] Smith, T. A. (2020). The role of customer personality in satisfaction, attitude-to-brand and loyalty in mobile services. *ResearchGate*. Dostępny pod adresem: `https://www.researchgate.net/publication/343285320`