# Analiza Preferencji Klientów

Weronika Kłujszo, Michał Korzeniewski, Miłosz Malinowski, Piotr Misiejuk

7 kwietnia 2025

## Spis treści

# 1 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.

# 2 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),
- zastosowanie dwóch metod klasteryzacji: k-średnich oraz GMM,
- ocenę i interpretację otrzymanych klastrów,
- podsumowanie wyników oraz możliwe rekomendacje marketingowe.

# 3 Przegląd Podobnych Prac

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

- **Customer Segmentation Analysis for Amazon Data** – Karthik Ramireddy. Praca magisterska analizująca dane klientów Amazona i ich segmentację z wykorzystaniem metod nienadzorowanych. Link do pracy
- **Customer personality analysis and clustering for targeted marketing** – D. Acheme, E. Enoyoze. Badanie z wykorzystaniem klasteryzacji k-średnich w celu lepszego targetowania kampanii marketingowych. Link do pracy
- **Customer segmentation: The concepts of trust, commitment and relationships** – B. Hultén. Autor analizuje znaczenie relacji i zaufania jako elementów segmentacji. Link do pracy
- **The role of customer personality in satisfaction, attitude-to-brand & loyalty in mobile services** – T. A. Smith. Badanie powiązań między osobowością klientów a ich lojalnością wobec marki. Link do pracy

# 4 Słowa Kluczowe

**POL:** analiza osobowości klientów, strategia marketingowa, analiza danych, analiza skupień, klastrowanie

**ENG:** customer behaviour analysis, marketing strategy, data analysis, cluster analysis, clustering

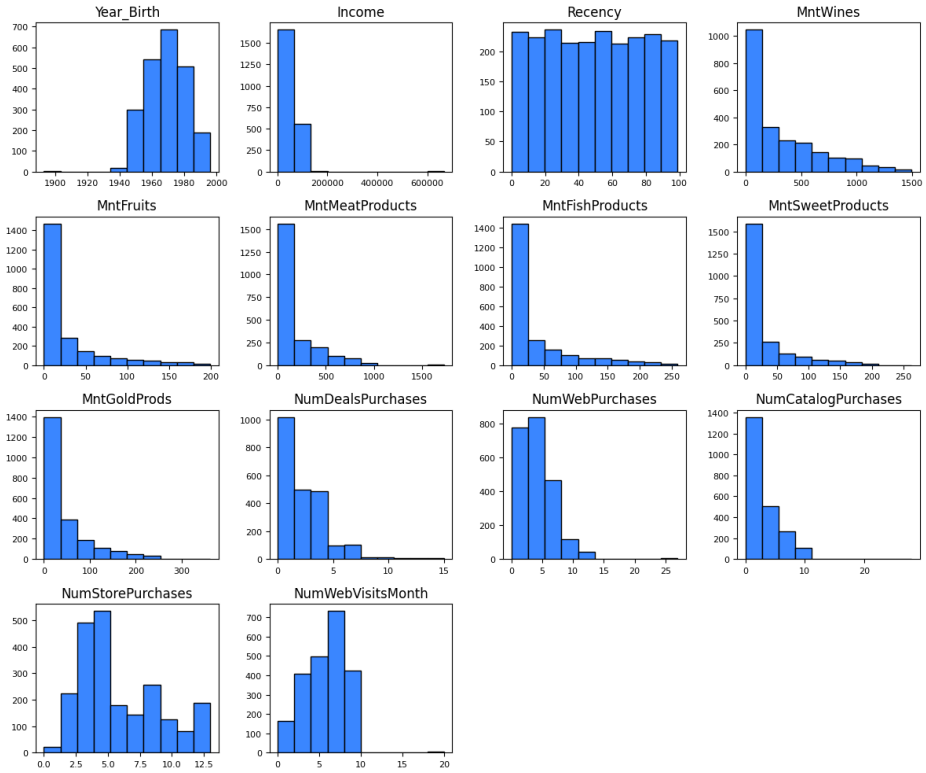
# 5 Wykresy dla zmiennych

Zmienne podzielono następująco:

- **Ciągłe:**
  - Rok urodzenia klienta (YearBirth),
  - Roczny dochód gospodarstwa domowego klienta (Income),
  - Liczba dni od ostatniego zakupu klienta (Recency),
  - Kwota wydana na wino w ostatnich 2 latach (MntWines),
  - wota wydana na owoce w ostatnich 2 latach (MntFruits,
  - Kwota wydana na mięso w ostatnich 2 latach (MntMeatProducts),
  - Kwota wydana na ryby w ostatnich 2 latach (MntFishProducts),
  - Kwota wydana na słodycze w ostatnich 2 latach (MntSweetProducts),
  - Kwota wydana na złoto w ostatnich 2 latach (MntGoldProds),
  - Liczba zakupów dokonanych z rabatem (NumDealsPurchases),
  - Liczba zakupów dokonanych przez stronę internetowąt (NumWebPurchases),
  - Liczba zakupów dokonanych za pomocą katalogu (NumCatalogPurchases),
  - Liczba zakupów dokonanych w sklepie (NumStorePurchases):,
  - Liczba wizyt na stronie internetowej w ostatnim miesiącu (NumWebVisitsMonth).
- **Jakościowe:**
  - 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.



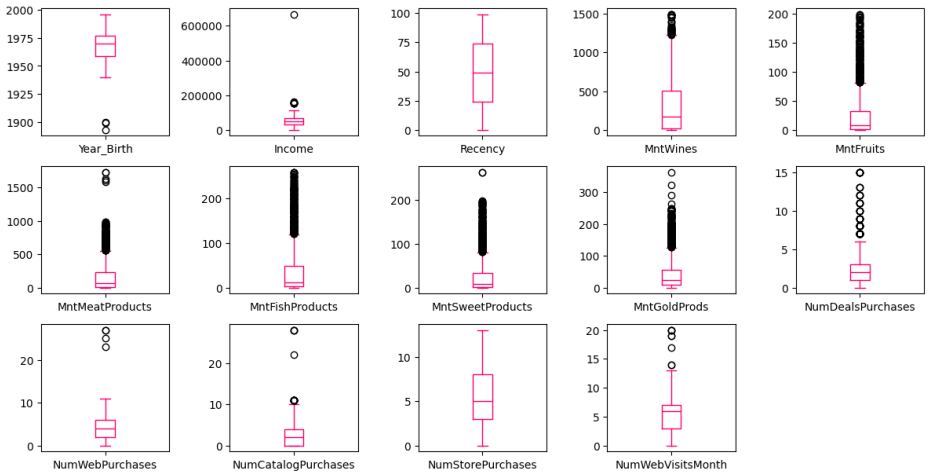

Rys. 1. Histogramy rozkładu zmiennych ciągłych.

Dla zmiennych jakościowych powstały wykresy kołowe.

# 6 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 dwie popularne metody skalowania: **standaryzację** oraz **skalowanie min-max**.




Rys. 2. Wykresy pudełkowe dla zmiennych ciągłych.

## 6.1 Standaryzacja

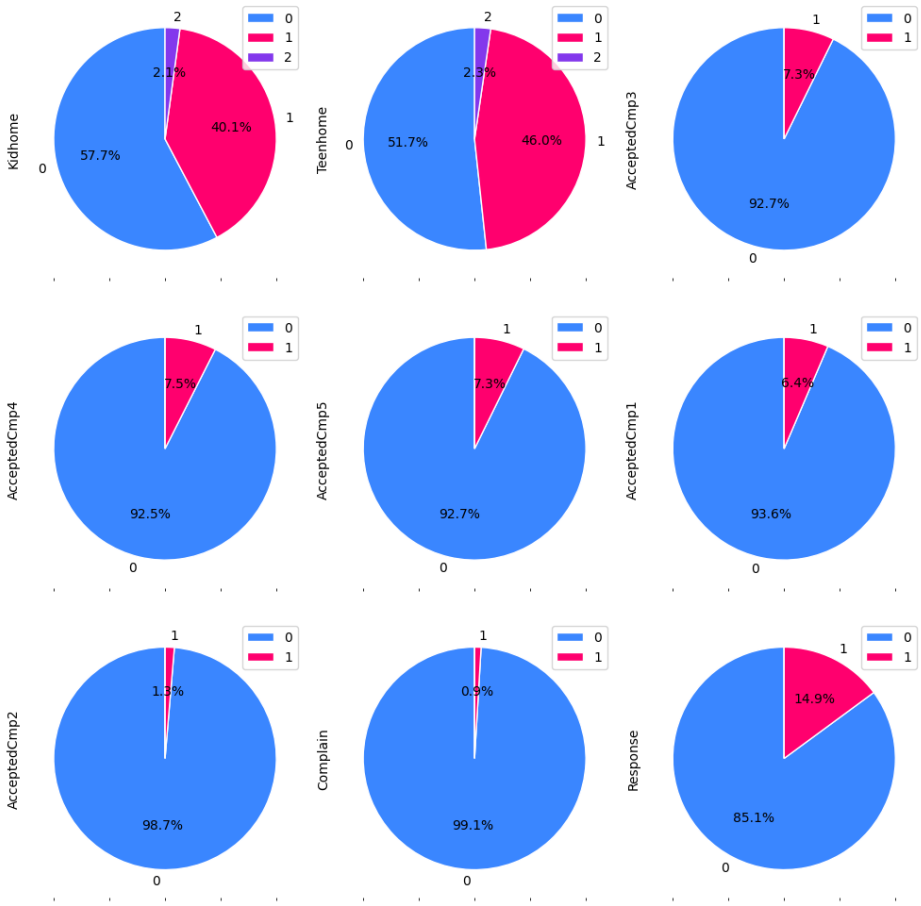
Standaryzacja (inaczej normalizacja z wykorzystaniem średniej i odchylenia standardowego) polega na przekształceniu danych w taki sposób, że każda cecha ma średnią wartość 0 oraz odchylenie standardowe 1. Dzięki temu zmienne z różnych zakresów wartości stają się porównywalne.

## 6.2 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. Jest mniej oporna na wartości odstające.

## 6.3 Porównanie metod

W projekcie zastosowano obie metody i porównano je ze sobą. Dane zostały wcześniej przygotowane, usunięto błędne wartości (na przykład wartość stanu cywilnego klienta wynoszącą "YOLO"lub rok urodzenia równy 1893). Z uwagi na wybór k-średnich jako jednej z metod klasteryzacji zdecydowano się na korzystanie ze znormalizowanych danych, ponieważ algorytm ten jest wrażliwy na wartości odstające.



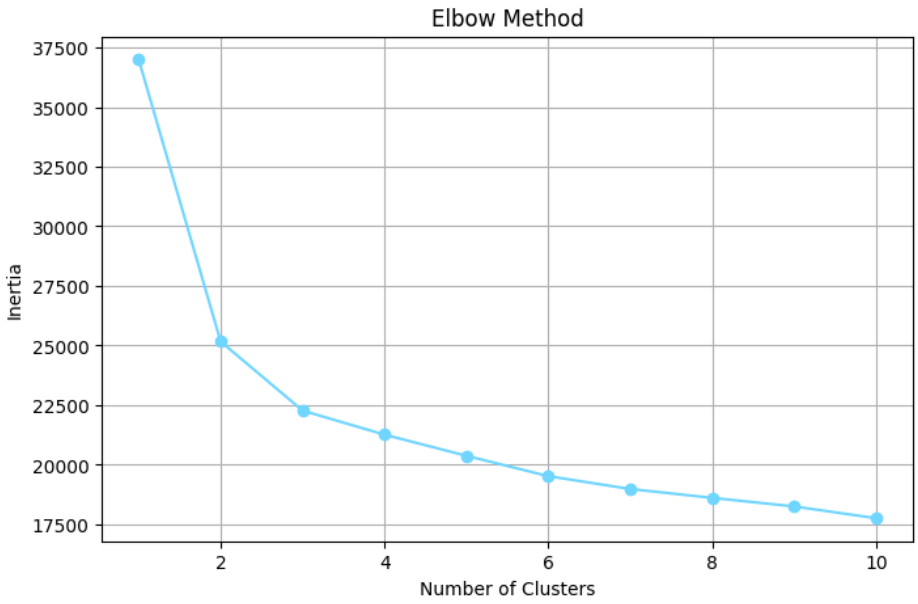

Rys. 3. Wykresy kołowe dla zmiennych jakościowych.

**tutaj to genuinely nie wiem jak stoimy, na razie napisałam takie cuś**

W tym projekcie, dla porównania, zastosowano obie metody skalowania: standaryzację oraz min-max skalowanie, aby sprawdzić, która metoda daje lepsze wyniki w kontekście segmentacji klientów za pomocą algorytmu k-średnich oraz GMM.

# 7 Klasteryzacja

Zastosowano metodę łokcia do wyznaczenia optymalnej liczby klastrów.



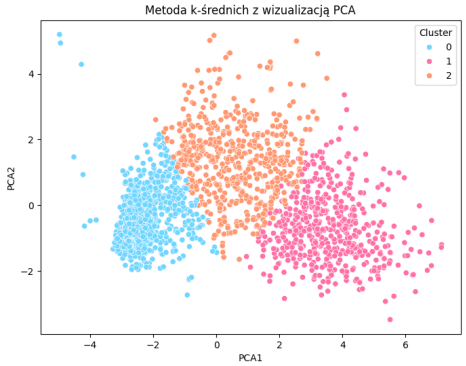

Rys. 4. Metoda łokcia w celu wyznaczenia optymalnej liczby klastrów.

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

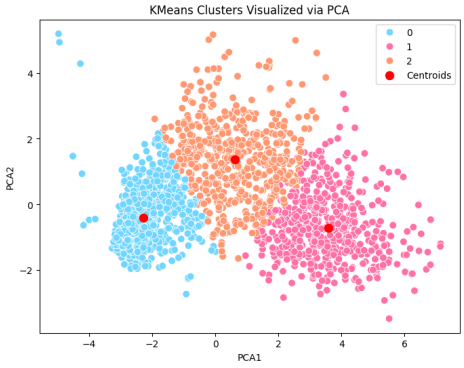
## 7.1 Klasteryzacja k-średnich

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

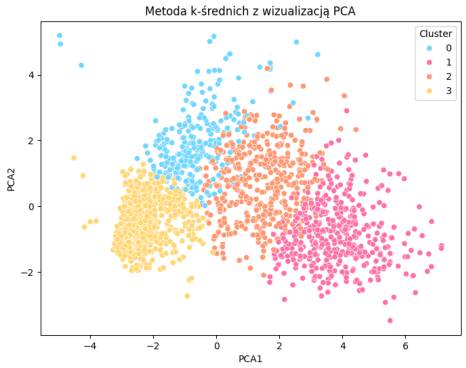
- Ma niską złożoność obliczeniową, co sprawia, że działa efektywnie nawet w przypadku dużych zbiorów danych, a z takim pracujemy przy projekcie.
- Algorytm jest szczególnie skuteczny w przypadku zmiennych ciągłych, jak w naszym przypadku, gdzie analizujemy dochody, wydatki oraz liczbę zakupów.



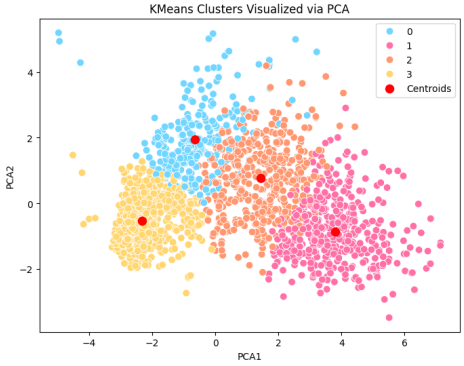

Rys. 5. Wyniki klasteryzacji k-średnich dla 3 klastrów.




Rys. 6. Wizualizacja wyników klasteryzacji k-średnich dla 3 klastrów.




Rys. 7. Wyniki klasteryzacji k-średnich dla 4 klastrów.




Rys. 8. Wizualizacja wyników klasteryzacji k-średnich dla 4 klastrów.

### 7.1.1 Wnioski

- **Dla 3 klastrów**
  - **Klaster 0: Oszczędne rodziny z dziećmi (1047 osób)**
    - Niski dochód, najniższy w porównaniu do innych klastrów
    - Duża liczba dzieci (średnio 1,23 dziecka)

    * Aktywność zakupowa: kupują zarówno stacjonarnie, jak i online
    * Reagują umiarkowanie na promocje i nie zależą od kampanii reklamowych
    * Edukacja wyższa (licencjat)
    * Składają najwięcej reklamacji

    Podsumowanie: Są to rodziny z dziećmi, które preferują oszczędne zakupy i są aktywne w poszukiwaniach promocji.

  - **Klaster 1: Bogate pary kupujące towary wyższej klasy (564 osoby)**
    * Najwyższy dochód w porównaniu do innych klastrów
    * Niska liczba dzieci i mała liczba nastolatków
    * Wydają najwięcej na wino, mięso i ryby
    * Najczęściej kupują stacjonarnie i przez katalog
    * Najlepiej reagują na kampanie marketingowe
    * Edukacja wyższa (licencjat)
    * Rzadko składają reklamacje

    Podsumowanie: Klienci premium, skłonni do zakupu drogich produktów, z lojalnością wobec marek wysokiej jakości.

  - **Klaster 2: Średniozamożne starsze małżeństwa (605 osób)**
    * Średni dochód i średnie wydatki
    * Najwięcej nastolatków, umiarkowana liczba dzieci
    * Wydają głównie na wino, mięso i złoto
    * Duża aktywność zakupowa, zarówno online, jak i stacjonarnie
    * Skłonność do składania umiarkowanej liczby reklamacji
    * Najwyższy poziom wykształcenia (magister)

    Podsumowanie: Starsze małżeństwa z preferencjami do bardziej premium produktów, z wykształceniem wyższym.

- **Dla 4 klastrów**
  - **Klaster 0: Stabilne rodziny z nastolatkami (289 osób)**
    * Dochód średni, około 49 261

    - Umiarkowana liczba dzieci i najwięcej nastolatków
    - Średnie wydatki, najwięcej na wino i mięso
    - Często korzystają z promocji
    - Równowaga między zakupami online i stacjonarnymi
    - Umiarkowana reakcja na kampanie
    - Średni wiek 55 lat
    - Większość w związku

  Podsumowanie: Stabilne rodziny z dziećmi, umiarkowane dochody, wrażliwe na promocje.

- **Klaster 1: Zamożni koneserzy (473 osoby)**
    - Najwyższy dochód ( 78 511)
    - Brak dzieci, bardzo mało nastolatków
    - Najwyższe wydatki, szczególnie na wino i produkty premium
    - Preferują zakupy katalogowe i stacjonarne
    - Najlepsza reakcja na kampanie
    - Średni wiek 52 lata
    - Większość w związku

  Podsumowanie: Zamożni klienci preferujący wysokiej jakości produkty, lojalni wobec marek.

- **Klaster 2: Dojrzałe pary lubiące wino (468 osób)**
    - Wyższy średni dochód ( 64 346)
    - Mało dzieci, sporo nastolatków
    - Wysokie wydatki na wino i mięso
    - Częste zakupy stacjonarne i online
    - Umiarkowana reakcja na kampanie
    - Najstarsza grupa (średni wiek 56)
    - Wykształcenie magisterskie

  Podsumowanie: Dojrzałe pary z wykształceniem wyższym, o ugruntowanych preferencjach zakupowych, lubiące wino i tradycyjne zakupy.

- **Klaster 3: Oszczędne rodziny z małymi dziećmi (986 osób)**
    - Najniższy dochód ( 34 781)

* Najwięcej dzieci, umiarkowana liczba nastolatków
* Najniższe wydatki, głównie zakupy stacjonarne
* Najwięcej wizyt na stronie, ale mało zakupów online
* Brak reakcji na kampanie
* Najmłodsza grupa (średni wiek 50)
* Wykształcenie wyższe

Podsumowanie: Rodziny z małymi dziećmi, skoncentrowane na podstawowych potrzebach, z niewielkim budżetem.

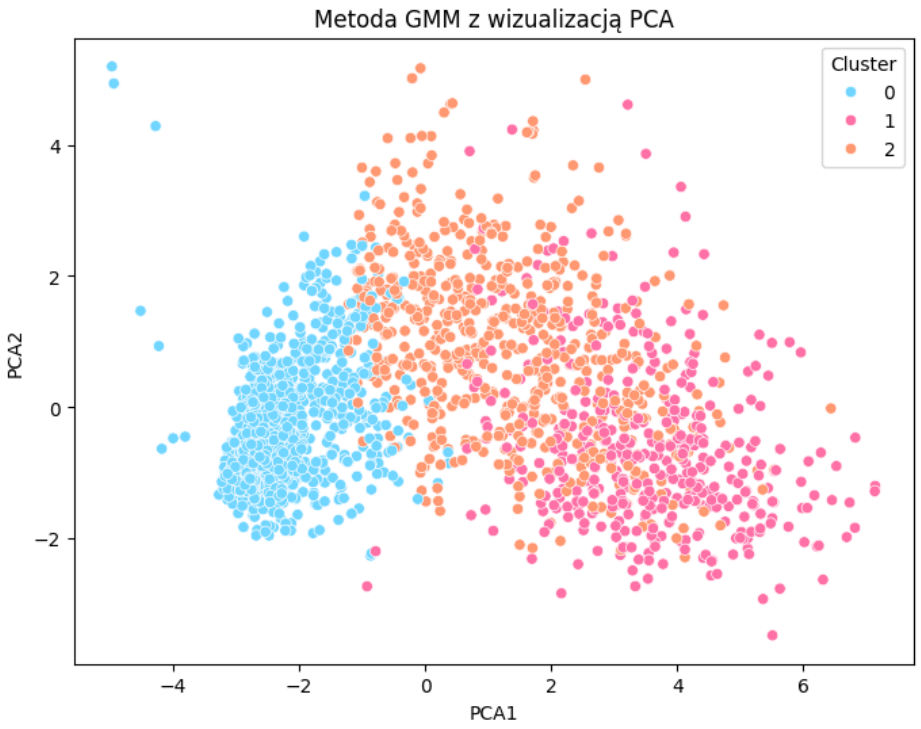
Według metody łokcia oraz otrzymanych wyników preferowane jest wykorzystanie 3 klastrów.

## 7.2 Klasteryzacja GMM

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. W przeciwieństwie do algorytmu k-średnich, który przypisuje punktom jednoznacznie przynależność do jednego klastra, GMM przypisuje prawdopodobieństwo przynależności punktu do każdego z klastrów.

- GMM jest bardziej elastyczne niż k-średnie, ponieważ może wykrywać klastry o nieregularnych kształtach, takich jak elipsy.
- Każdemu punktowi przypisywane jest prawdopodobieństwo przynależności do każdego klastra.
- Liczba klastrów (k) musi być określona przed rozpoczęciem procesu klasteryzacji.

Metodę mieszanki gaussowskiej wykonano tylko dla 3 klastrów.




Rys. 9. Metoda GMM dla 3 klastrów.

### 7.2.1 Wnioski

- **Klaster 0 (1096 osób)**
  - Największe zarobki ze wszystkich
  - Średnio najwięcej małych dzieci, i dzieci w ogóle
  - Mają kilka razy niższe zarobki ale nawet kilkanaście razy niższe wydatki
  - W miarę zainteresowani promocjami, ale nie najbardziej ze wszystkich grup.
  - Kupują głównie przez internet i stacjonarnie, praktycznie nie korzystają z katalogu, ale bardzo często odwiedzają stronę internetową
  - Co ciekawe, rzadko korzystają z kampanii promocyjnych, oraz rzadko narzekają

- Najmłodsza grupa ze wszystkich
- Tak jak pozostałe grupy, najczęściej mają za sobą edukację na poziomie licencjatu, oraz są w związku
- Składają średnio najwięćej skarg
- Wydają relatywnie najmniej w sklepie, średnio 0.3

Podsumowanie: Skrajnie oszczędni wcześni rodzice trudni do manipulacji poprzez akcje promocyjne oraz skorzy do skarżenia się na niedogodności

- **Klaster 1 (446 osób)**
  - Najrzadziej mają dzieci, a jeżeli już to starsze
  - Wydają najwięcej ze wszystkich grup na wszystkie produkty, ulubionym produktem jest wino, za nim ryby (w co może wliczać się kawior) i złoto
  - Kupują stacjonarnie, z taką samą częstotliwością korzystają z katalogu i sklepu internetowego.
  - Bardzo mało wchodzą na stronę internetową
  - Bardzo chętni do brania udziału w kampaniach promocyjnych
  - Wydają zdecydowanie najwięcej ze wszystkich grup, zarówno bewzględnie jak i procentowo (prawie 2
  - Tak jak pozostałe grupy, najczęściej mają za sobą edukację na poziomie licencjatu, oraz są w związku

Podsumowanie: bogaci i/lub rozrzutni zblazowani bezdzietni (lub z już dorosłymi dziećmi)łatwi do zmanipulowania przez kampanie promocyjne

- **Klaster 2**
  - Średnio najwięcej nastoletnich dzieci
  - Przychody, wydatki na poszczególne produty jak i chęć korzystania z promocji pomiędzy pozostałymi klastrami
  - Wydatki ogólne zarówno bezwzględnie jak i procentowo pomiędzy dwoma pozostałymi klastrami

- Wydają najwięcej na wino, mięso i złoto
- Średnio najstarsi
- Tak jak pozostałe grupy, najczęściej mają za sobą edukację na poziomie licencjatu, oraz są w związku

Podsumowanie: Nieco starsi, ale poza tym bardzo "średni"klienci. Brak znaczących znaków szczególnych

# 8 Wnioski

Analiza segmentacji klientów przeprowadzona za pomocą metod klasteryzacji (k-średnich i GMM) ujawnia wyraźne różnice między grupami klientów, które mogą być pomocne w tworzeniu spersonalizowanych strategii marketingowych. Na podstawie wyników klasteryzacji udało się wyróżnić kilka istotnych grup konsumentów, które charakteryzują się odmiennymi preferencjami i zachowaniami zakupowymi.

Wyróżniono:

- **Klaster 0: Oszczędne rodziny z dziećmi:** Grupa ta charakteryzuje się najniższym dochodem oraz najwyższą liczbą dzieci wśród wszystkich klastrów. Często korzystają z promocji, jednak ich zakupy są mniej zależne od kampanii reklamowych. Preferują zakupy stacjonarne, jednak wykazują także aktywność online, choć ich wydatki są stosunkowo niskie. Są to rodziny z dziećmi, które stawiają na oszczędność i planowanie wydatków.

- **Klaster 1: Bogate pary kupujące towary wyższej klasy:** Klienci w tej grupie charakteryzują się najwyższym dochodem i najniższą liczbą dzieci. Wydają dużo na wino, mięso i ryby, preferują zakupy stacjonarne oraz katalogowe. Są to osoby, które najlepiej reagują na kampanie marketingowe, co czyni ich idealną grupą do targetowania wysokiej jakości produktów. Z racji swojej lojalności wobec marek premium, ta grupa stanowi główną grupę docelową dla drobnych luksusowych produktów.

- **Klaster 2: Średniozamożne starsze małżeństwa:**

  Klientów w tej grupie cechuje średni dochód oraz wykształcenie magisterskie. Preferują zakupy zarówno online, jak i stacjonarnie, a ich

wydatki są średnie, przy czym spora część wydawana jest na wino, mięso i złoto. To starsze pary, które są mniej podatne na reklamy, ale ich preferencje zakupowe są jasno określone. Mogą być dobrym celem dla produktów bardziej luksusowych, ale z umiarem w promocjach.

# 9 Rekomendacje marketingowe

## 9.1 Dla Klastra 0 (Oszczędne rodziny z dziećmi)

- **Strategia marketingowa:** Należy skoncentrować się na oferowaniu promocji, rabatów i programów lojalnościowych, które mogłyby przyciągnąć tę grupę. Można rozważyć skierowanie ofert do rodzin z dziećmi, podkreślając oszczędności i atrakcyjne ceny produktów.

- **Kampanie online:** Skierowanie reklam do tej grupy online, z naciskiem na ekonomiczne podejście do zakupów (np. rabaty, promocje).

## 9.2 Dla Klastra 1 (Bogate pary kupujące towary wyższej klasy)

- **Strategia marketingowa:** Należy skierować ofertę produktów premium (np. luksusowe wina, artykuły spożywcze wysokiej jakości, biżuteria) do tej grupy. Można również zaoferować im ekskluzywne oferty i programy lojalnościowe.

- **Kampanie targetowane:** Zwiększyć inwestycje w reklamy targetowane, które zwiększą zaangażowanie tej grupy, zwłaszcza podczas kampanii marketingowych.

## 9.3 Dla Klastra 2 (Średniozamożne starsze małżeństwa)

- **Strategia marketingowa:** Proponować produkty klasy średniej, które łączą jakość z umiarkowaną ceną. Może to być idealna grupa docelowa dla produktów, które są w średniej cenie, ale wysokiej jakości.

- **Kampanie umiarkowane:** Zastosować umiarkowane kampanie promocyjne, które nie będą nachalne, ale będą docierały do tej grupy w sposób subtelny.

## 9.4 Podsumowanie

Analiza segmentacji klientów przeprowadzona za pomocą metod klasteryzacji (k-średnich i GMM) pozwoliła na wyodrębnienie kilku grup o odmiennych preferencjach i zachowaniach zakupowych. Na podstawie wyników klasteryzacji opracowane zostały rekomendacje marketingowe, które uwzględniają specyfikę każdej z grup. Kluczowym elementem w przyszłości będzie testowanie tych rekomendacji oraz monitorowanie efektywności działań marketingowych skierowanych do każdej z grup.