



**Politechnika  
Śląska**

## **Algorytmy Eksploracji Danych**

Analiza zbioru sprzedaży detalicznej online

Członkowie zespołu:

*Jakub Darul, gr. 1/1*

*Mateusz Lamla, gr. 1/1*

Kierunek: Informatyka

Semestr 5, Stopień I

Specjalizacja: Inżynieria Analizy Danych

Politechnika Śląska

Wydział Matematyki Stosowanej

30 stycznia 2026

# **Spis treści**

<b>1 Wprowadzenie</b>	<b>2</b>
1.1 Cel projektu . . . . .	2
1.2 Opis projektu . . . . .	2
<b>2 Zbiór danych</b>	<b>2</b>
2.1 Atrybuty zbioru danych . . . . .	2
2.2 Charakterystyka danych . . . . .	3
<b>3 Zastosowane metody</b>	<b>4</b>
3.1 Redukcja wymiaru PCA . . . . .	4
3.1.1 Cel . . . . .	4
3.1.2 Opis metody . . . . .	4
3.2 Klasteryzacja K-means . . . . .	5
3.2.1 Cel . . . . .	5
3.2.2 Opis metody . . . . .	5
3.3 Reguły asocjacji . . . . .	6
3.3.1 Cel . . . . .	6
3.3.2 Opis metody . . . . .	6
<b>4 Opis implementacji</b>	<b>6</b>
4.1 Zastosowane biblioteki . . . . .	6
4.2 Zastosowane funkcje i algorytmy . . . . .	7
<b>5 Wyniki</b>	<b>8</b>
5.1 PCA . . . . .	8
5.2 K-means . . . . .	9
5.3 Reguły asocjacji . . . . .	11
5.4 Wnioski . . . . .	12
<b>6 Podsumowanie</b>	<b>12</b>

# 1 Wprowadzenie

## 1.1 Cel projektu

Głównym celem projektu jest przeprowadzenie procesu eksploracji danych na wybranym zbiorze danych, aby odkryć ukryte w nim struktury oraz zależności. Projekt zakłada zastosowanie metod odpowiadających za proces redukcji wymiarowości danych oraz ich segmentacji (klasteryzacji). Istotnym aspektem jest również analiza wpływu redukcji wymiaru na jakość i interpretowalność uzyskanych klastrów, a także próba wyekstrahowania charakterystycznych wzorców dla poszczególnych grup.

## 1.2 Opis projektu

Projekt składa się z kilku kluczowych etapów przetwarzania danych, zgodnych ze standardem procesu KDD. W pierwszej fazie następuje wstępne przetworzenie zbioru danych, obejmujące czyszczenie oraz standaryzację cech. Następnie, w celu wizualizacji oraz eliminacji szumu i redundancji, wykorzystana zostaje metoda Analizy Głównych Składowych (PCA). Na tak przygotowanych danych (zarówno w przestrzeni oryginalnej, jak i zredukowanej) stosowany jest algorytm K-means w celu wyodrębnienia jednorodnych podgrup obiektów. Końcowym etapem jest analiza uzyskanych wyników oraz interpretacja cech dominujących w wyznaczonych klastrach.

# 2 Zbiór danych

W projekcie wykorzystano zbiór danych *Online Retail II*, pochodzący z repreztorium UCI Machine Learning Repository i dostępny na platformie Kaggle. Zbiór zawiera dane transakcyjne zarejestrowane przez brytyjski sklep internetowy (e-commerce) w okresie od 01.12.2009 do 09.12.2011. Firma specjalizuje się w sprzedaży unikalnych upominków na różne okazje, a znaczną część jej klientów stanowią hurtownicy.

Zbiór danych ma charakter rzeczywisty i składa się z rekordów reprezentujących poszczególne pozycje na fakturach sprzedażowych.

## 2.1 Atrybuty zbioru danych

Oryginalny zbiór danych zawiera 8 atrybutów (cech). Poniżej przedstawiono ich szczegółowy opis:

- **Invoice** (Numer faktury): Typ nominalny. 6-cyfrowy numer całkowity unikalnie przypisany do każdej transakcji. Jeżeli kod ten zaczyna się od litery 'c', oznacza to anulowanie zamówienia (zwrot).
- **StockCode** (Kod produktu): Typ nominalny. 5-cyfrowy numer całkowity unikalnie przypisany do każdego odrębnego produktu w ofercie.
- **Description** (Opis produktu): Typ nominalny. Nazwa (opis) produktu.
- **Quantity** (Ilość): Typ numeryczny. Liczba sztuk danego produktu przypadająca na transakcję.
- **InvoiceDate** (Data faktury): Typ numeryczny (Data/Czas). Dzień i godzina wygenerowania każdej transakcji.
- **Price** (Cena jednostkowa): Typ numeryczny. Cena produktu za sztukę (wyrażona w funtach).
- **Customer ID** (Identyfikator klienta): Typ nominalny. 5-cyfrowy numer całkowity unikalnie przypisany do każdego klienta.
- **Country** (Kraj): Typ nominalny. Nazwa kraju, w którym znajduje się klient.

## 2.2 Charakterystyka danych

Zbiór danych charakteryzuje się kilkoma istotnymi właściwościami, które mają wpływ na dalsze etapy przetwarzania:

1. **Braki danych:** Atrybut *Customer ID* zawiera puste wartości dla części transakcji (np. zakupy dokonane przez gości bez rejestracji), co jest kluczowe w kontekście segmentacji klientów.
2. **Transakcje zwrotne:** Występowanie ujemnych wartości w kolumnie *Quantity* oraz prefiksu 'c' w *InvoiceNo* wskazuje na anulowane transakcje, które mogą wymagać odfiltrowania.
3. **Wymiarowość:** Dane mają charakter transakcyjny (wiersz = produkt w koszyku), co oznacza, że przed zastosowaniem algorytmów takich jak PCA czy K-means, konieczna będzie agregacja danych do poziomu klienta lub koszyka zakupowego (np. stworzenie macierzy RFM: Recency, Frequency, Monetary).

### 3 Zastosowane metody

#### 3.1 Redukcja wymiaru PCA

##### 3.1.1 Cel

Analiza Głównych Składowych (ang. *Principal Component Analysis*, PCA) została zastosowana w celu redukcji wymiarowości przestrzeni cech przy jednoczesnym zachowaniu jak największej ilości informacji (wariancji) zawartej w oryginalnym zbiorze danych. Metoda ta pozwala na:

- Wizualizację wielowymiarowych danych w przestrzeni 2D lub 3D.
- Usunięcie skorelowanych zmiennych.
- Zmniejszenie kosztu obliczeniowego dla kolejnych algorytmów (np. klaszteryzacji).

##### 3.1.2 Opis metody

Analiza Składowych Głównych (PCA) jest kluczową techniką statystyczną, której głównym celem jest redukcja liczby zmiennych opisujących badane zjawisko przy jednoczesnym zachowaniu maksymalnej ilości niesionej przez nie informacji (wariancji). Pozwala ona na przekształcenie rozbudowanego zestawu skorelowanych cech w mniejszy zbiór nowych zmiennych.

Charakterystyka składowych głównych:

- **Syntetyczność:** Są to zmienne „sztuczne” (nieobserwowalne bezpośrednio), będące liniowymi kombinacjami zmiennych pierwotnych.
- **Ortogonalność:** Nowe zmienne są wzajemnie nieskorelowane, co eliminuje problem nadmiarowości informacji.
- **Hierarchiczność:** Pierwsza składowa wyjaśnia największą część całkowitej wariancji układu. Każda kolejna jest konstruowana tak, aby wyjaśnić maksymalną część zmienności pozostalej po uwzględnieniu poprzednich składowych.

Istotnym elementem analizy jest interpretacja nowych zmiennych, dokonywana w oparciu o tzw. **ładunki czynnikowe**. Są to współczynniki korelacji pomiędzy zmiennymi wyjściowymi a poszczególnymi składowymi. Zmienne o wysokich ładunkach mają największy wkład w budowę danej składowej, co pozwala nadać jej merytoryczne znaczenie.

Kluczowym etapem PCA jest decyzja o liczbie pozostawionych składowych. W praktyce stosuje się trzy główne kryteria decyzyjne:

1. **Kryterium procentowe:** Zakłada uwzględnienie tylu początkowych składowych, aby suma wyjaśnianej przez nie wariancji przekroczyła określony próg (zazwyczaj 75% lub 80%).
2. **Kryterium Kaisera:** Rekomenduje pozostawienie wyłącznie tych składowych, którym odpowiadają wartości własne macierzy korelacji większe od 1. Oznacza to, że składowa musi wyjaśniać więcej zmienności niż pojedyncza zmienna oryginalna.
3. **Kryterium Cattella (wykres osypiska):** Metoda graficzna polegająca na analizie wykresu wartości własnych. Wybiera się składowe znajdujące się przed punktem, w którym wykres zaczyna łagodnie opadać (tworząc tzw. osypisko), co wskazuje na wygasanie istotnej zmienności.

## 3.2 Klasteryzacja K-means

### 3.2.1 Cel

Algorytm K-means (K-średnich) został użyty w celu pogrupowania obiektów w zbiorze danych na  $k$  rozłącznych klastrów. Celem jest taki podział danych, aby obiekty wewnątrz jednej grupy były do siebie jak najbardziej podobne, a obiekty z różnych grup jak najbardziej różne.

### 3.2.2 Opis metody

K-means jest algorytmem iteracyjnym, który dąży do zminimalizowania wariancji wewnętrzkklastrowej. Jako miarę niepodobieństwa najczęściej stosuje się kwadrat odległości euklidesowej.

Algorytm działa w następujących krokach:

1. **Inicjalizacja:** Losowy wybór  $k$  punktów jako początkowych środków klastrów (centroidów).
2. **Przypisanie:** Każdy punkt danych jest przypisywany do klastra, którego centroid znajduje się najbliżej:
3. **Aktualizacja:** Obliczane są nowe środki klastrów poprzez wyznaczenie średniej arytmetycznej punktów należących do danego klastra.
4. Kroki 2 i 3 są powtarzane aż do momentu osiągnięcia kryterium stopu (np. brak zmian w przypisaniach punktów).

### 3.3 Reguły asocjacji

#### 3.3.1 Cel

Celem zastosowania analizy asocjacji (reguł asocjacyjnych) jest wykrycie powiązań pomiędzy obiektami w dużych zbiorach danych. W kontekście sprzedaży detalicznej, technika ta, znana jako analiza koszykowa (Market Basket Analysis), pozwala odpowiedzieć na pytanie: „Które produkty są najczęściej kupowane razem?”. Wiedza ta jest kluczowa dla budowania systemów rekommendacyjnych oraz optymalizacji strategii cross-sellingu.

#### 3.3.2 Opis metody

Do wygenerowania reguł asocjacyjnych wykorzystuje się algorytmy takie jak Apriori lub FP-Growth. Reguła asocjacyjna przyjmuje postać  $X \rightarrow Y$ , gdzie  $X$  (poprzednik) i  $Y$  (następnik) to rozłączne zbiory produktów.

Siłę i użyteczność reguł ocenia się za pomocą trzech podstawowych miar:

- **Wsparcie (Support):** Określa, jak popularny jest dany zestaw produktów w całej bazie transakcji. Oblicza się je jako stosunek liczby transakcji zawierających zarówno  $X$ , jak i  $Y$ , do całkowitej liczby transakcji.
- **Ufność (Confidence):** Prawdopodobieństwo warunkowe zakupu produktu  $Y$ , pod warunkiem że w koszyku znajduje się już produkt  $X$ . Wysoka ufność oznacza silną zależność.
- **Lift:** Wskaźnik określający, ile razy częściej produkty  $X$  i  $Y$  występują razem, niż wynikłoby to z ich niezależności statystycznej.
  - $Lift > 1$ : Produkty są powiązane pozytywnie (występują razem częściej niż przypadkowo).
  - $Lift = 1$ : Produkty są niezależne.
  - $Lift < 1$ : Produkty są powiązane negatywnie (występują razem rzadziej niż przypadkowo).

## 4 Opis implementacji

### 4.1 Zastosowane biblioteki

Projekt został zrealizowany w języku Python, wykorzystując szereg bibliotek do analizy danych, obliczeń numerycznych oraz uczenia maszynowego.

Poniżej przedstawiono wykaz kluczowych pakietów użytych w procesie implementacji:

- **Pandas:** Podstawowa biblioteka służąca do manipulacji i analizy danych. W projekcie wykorzystana głównie do wczytania zbioru danych oraz przeprowadzania operacji czyszczenia, filtracji i agregacji.
- **NumPy:** Fundamentalny pakiet do obliczeń naukowych. Zapewnia obsługę wielowymiarowych tablic i macierzy oraz zestaw funkcji matematycznych niezbędnych do wykonywania operacji na wektorach.
- **Scikit-learn (sklearn):** Najpopularniejsza biblioteka do uczenia maszynowego, oferująca szeroki wachlarz efektywnych narzędzi do analizy danych. W ramach projektu wykorzystano moduły odpowiedzialne za:
  - *Preprocessing:* Standaryzacja i skalowanie danych.
  - *Decomposition:* Implementacja algorytmu redukcji wymiaru (PCA).
  - *Cluster:* Implementacja algorytmu grupowania (K-means).
  - *Metrics:* Ewaluacja jakości modeli.
- **Matplotlib:** Standardowa biblioteka do tworzenia wizualizacji. Użyta do generowania wykresów pomocniczych oraz wizualizacji 3D przestrzeni po redukcji wymiaru.
- **Seaborn:** Biblioteka bazująca na Matplotlib, służąca do wizualizacji danych statystycznych. Została wykorzystana do stworzenia bardziej czytelnych i estetycznych wykresów, w tym map ciepła (heatmap) obrazujących profile klastrów.
- **Factor Analyzer:** Specjalistyczna biblioteka umożliwiająca przeprowadzanie analizy czynnikowej oraz testów statystycznych. W projekcie posłużyła do weryfikacji zasadności zastosowania redukcji wymiaru (test sferyczności Bartletta oraz miara KMO).

## 4.2 Zastosowane funkcje i algorytmy

W procesie implementacji wykorzystano następujące kluczowe funkcje i metody:

- **Przetwarzanie danych:**
  - `pandas.groupby()` oraz `agg()`: Użyte do agregacji danych transakcyjnych do poziomu klienta w celu wyliczenia metryk RFM (Recency, Frequency, Monetary).

- `pandas.pivot_table()`: Wykorzystane do przekształcenia danych do formatu "koszykowego"(transakcja × produkt), niezbędnego do analizy reguł asocjacyjnych.

- **Standaryzacja i Redukcja wymiaru:**

- `StandardScaler.fit_transform()`: Metoda służąca do standaryzacji cech.
- `PCA(n_components=...)`: Inicjalizacja obiektu PCA.
- `pca.fit_transform()`: Wyznaczenie głównych składowych i rzu-towanie danych na nową przestrzeń.

- **Klasteryzacja i Ewaluacja:**

- `KMeans(n_clusters=...)`: Inicjalizacja modelu K-średnich.
- `kmeans.fit_predict()`: Dopasowanie modelu do danych i przy-pisanie etykiet klastrów.
- `silhouette_score()`, `davies_bouldin_score()`: Funkcje obli-czające metryki walidacji wewnętrznej klastrów.

## 5 Wyniki

### 5.1 PCA

W pierwszej fazie analizy sprawdzono zasadność zastosowania redukcji wy-miarowości. Przeprowadzono test sferyczności Bartletta, który dał wynik istotny statystycznie ( $p < 0.05$ ), co oznacza, że zmienne w zbiorze są ze sobą skorelowane. Dodatkowo, miara KMO (Kaiser-Meyer-Olkin) wyniosła **0.979**, co potwierdza wysoką przydatność danych do analizy czynnikowej i PCA.

Analiza wariancji (scree plot) pozwoliła na dobór optymalnej liczby składowych. Przyjęto próg wyjaśnionej wariancji na poziomie 90%. Pozwoliło to na zredukowanie przestrzeni cech do **28 głównych składowych**. Jest to znaczna redukcja w stosunku do pierwotnego wymiaru, przy zachowaniu kluczowych informacji o strukturze danych.



Rysunek 1: Wykres analizy wariancji (scree plot) dla wyboru liczby głównych składowych

## 5.2 K-means

Dla zredukowanego zbioru danych przeprowadzono klasteryzację metodą K-means. Aby wyznaczyć optymalną liczbę klastrów ( $k$ ), posłużyono się metodą łokcia (Elbow Method), analizując wykres inercji w funkcji  $k$ . Zauważalne załamanie krzywej nastąpiło dla  $k = 4$ , co przyjęto jako finalną liczbę grup.



Rysunek 2: Wykres metody łokcia dla wyboru liczby klastrów

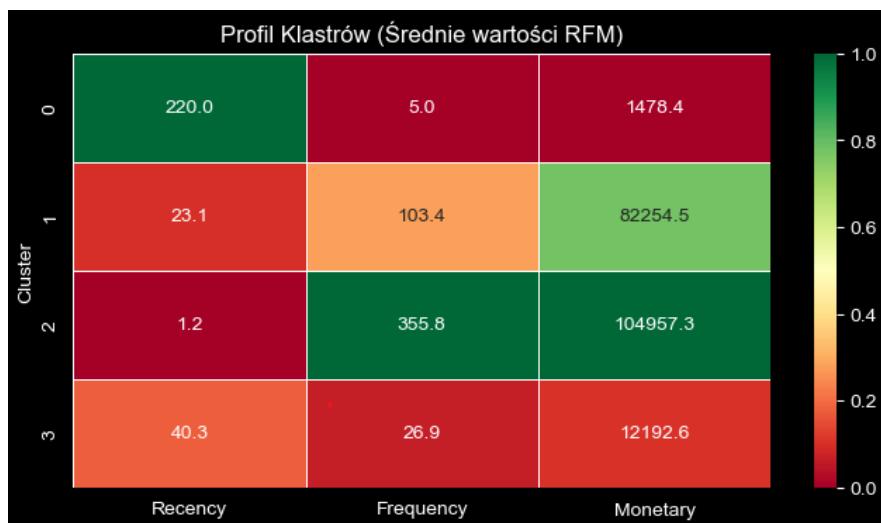
Jakość podziału zweryfikowano za pomocą:

- **Silhouette Score: 0.664** – Wynik ten świadczy o dobrej separacji klastrów i wysokiej spójności wewnętrz grup (wartości bliskie 1 oznaczają

bardzo dobry podział).

- **Davies-Bouldin Index:** **1.168** – Stosunkowo niska wartość wskaźnika potwierdza poprawność podziału (im niżej, tym lepiej).
- **Calinski-Harabasz Index:** **2476.73** – Wysoka wartość wskazuje na dobrze zdefiniowane, zwarte klastry.

W celu dokładniejszej interpretacji wyników, sporządzono mapę ciepła (heat-map), prezentującą średnie znormalizowane wartości atrybutów dla poszczególnych klastrów. Wizualizacja ta pozwala na identyfikację cech dominujących w każdej z grup.



Rysunek 3: Mapa ciepła (Heatmap) cech klientów w podziale na klastry

Analiza mapy ciepła prowadzi do następujących wniosków dotyczących profilu klientów w każdym z klastrów:

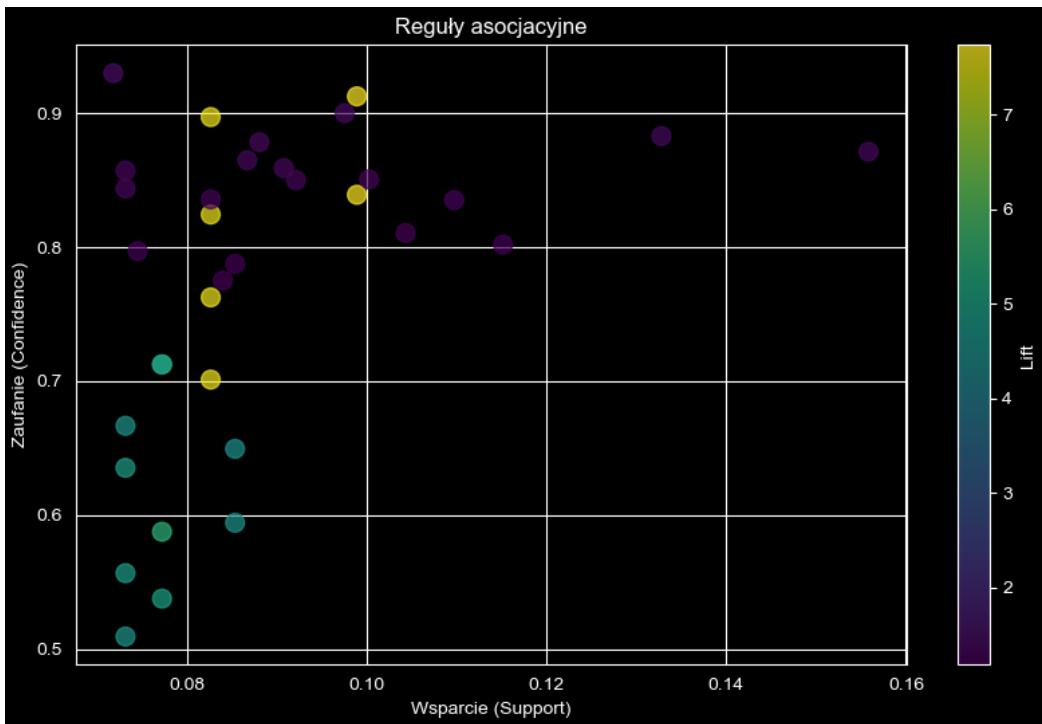
- **Klaster 0 (Okazjonalni):** Klienci o niskiej aktywności i wartości koszyka. Jest to grupa najliczniejsza, składająca się z kupujących sporadycznie.
- **Klaster 1 (Lojalni):** Klienci aktywni, o średniej wartości zakupów, ale wysokiej częstotliwości. Stanowią stabilną bazę lojalnych odbiorców.
- **Klaster 2 (VIP):** Grupa wyróżniająca się bardzo wysokimi wydatkami (Monetary) oraz częstotliwością. Mimo niewielkiej liczebności, generują znaczną część przychodu sklepu.

- **Klaster 3 (Uśpieni/Odchodzący):** Klienci, którzy w przeszłości dokonywali zakupów, ale obecnie zaprzestali aktywności (bardzo wysoki wskaźnik Recency).

### 5.3 Reguły asocjacji

Analiza reguł asocjacyjnych pozwoliła na zidentyfikowanie produktów, które klienci najczęściej kupują wspólnie. Wcielając się w rolę analityka doradzającego klientowi biznesowemu, możemy wskazać konkretne mechanizmy zakupowe:

- **Zjawisko komplementarności sprzętowej:** Zaobserwowano bardzo silną regułę ( $Lift > 5$ ) łączącą zakup monitorów lub komputerów z akcesoriami takimi jak kable HDMI czy uchwyty. *Interpretacja dla klienta:* Jeśli ktoś kupuje monitor, prawie na pewno potrzebuje kabla. To nie jest przypadek.
- **Zestawy dekoracyjne:** Klienci kupujący świece dekoracyjne w jednym kolorze, bardzo często dobierają do koszyka świeczniki lub serwetki z tej samej linii stylistycznej.
- **Rekomendacja:** Wykorzystując te dane, sklep może zwiększyć średnią wartość koszyka, proponując te produkty automatycznie („Klienci, którzy kupili ten produkt, wybrali również...”), zamiast liczyć na to, że klient sam je znajdzie.



Rysunek 4: Wykres najsiłniejszych reguł asocjacyjnych

## 5.4 Wnioski

Zastosowanie algorytmu PCA w połączeniu z K-means pozwoliło na skuteczną segmentację bazy klientów. Redukcja wymiarowości do 28 głównych składowych umożliwiła zachowanie istotnych informacji przy jednoczesnym uproszczeniu struktury danych. Klasteryzacja wykazała istnienie czterech wyraźnie zdefiniowanych grup klientów, co zostało potwierdzone przez wysokie wartości wskaźników walidacji wewnętrznej. Analiza charakterystyk poszczególnych klastrów pozwoliła na wyodrębnienie unikalnych wzorców zachowań zakupowych, co może być wykorzystane do celów marketingowych i personalizacji oferty.

## 6 Podsumowanie

Realizacja projektu pozwoliła na przejście przez pełny proces eksploracji danych (KDD). Wykazano, że surowe dane transakcyjne kryją w sobie cenną wiedzę o strukturze bazy klienckiej. Zastosowanie redukcji wymiarowości (PCA) znaczco usprawniło proces klasteryzacji, pozwalając na wyodrębnie-

nienie spójnych grup klientów.

Dzięki segmentacji sklep może przestać traktować wszystkich klientów tak samo, a dzięki regułom asocjacyjnym – lepiej odpowiadać na ich potrzeby, co bezpośrednio przekłada się na optymalizację zysków.