

Uniwersytet Przyrodniczy we Wrocławiu
Wydział Gospodarki Przestrzennej i Architektury Krajobrazu



Gospodarka Przestrzenna

Aleksandra Dybka

Nr indeksu 121062

**Wykorzystanie grafowych algorytmów uczenia maszynowego w analityce
społeczno-gospodarczej**

Praca magisterska

Opiekun pracy
dr inż. Grzegorz Chrobak

Wrocław, wrzesień 2025

*Szczególne podziękowania składam sobie: za pisanie, poprawianie, usuwanie, ponowne pisanie... i za to, że nie wybrałam łatwiejszego tematu.
A przede wszystkim, że dotrwałam do końca z reszkami godności i poczucia humoru.*

*Dziękuję także mojemu Promotorowi, za wymyślenie tej pracy oraz
subtelną presję intelektualną.*

*Z wyrazami szacunku
Aleksandra Dybka*

„Wykorzystanie grafowych algorytmów uczenia maszynowego w analityce społeczno-gospodarczej”

Streszczenie pracy

Celem pracy było zbadanie możliwości wykorzystania grafowych algorytmów uczenia maszynowego w analizie danych społeczno-gospodarczych, ze szczególnym uwzględnieniem modelowania osobowości klienta. Praca łączy podejścia teoretyczne z praktycznym zastosowaniem narzędzi eksploracji danych, bazując na zbiorze „Customer Personality Analysis” zawierającym informacje o cechach demograficznych, behawioralnych oraz decyzjach zakupowych klientów.

W pierwszej części pracy przedstawiono podstawy teorii grafów oraz ich przewagę nad tradycyjnymi bazami relacyjnymi w kontekście reprezentowania relacji między danymi. W środowisku Neo4j zbudowano grafową bazę danych, w której węzły reprezentują klientów, produkty, kanały sprzedaży i kampanie marketingowe, a relacje między nimi odzwierciedlają rzeczywiste interakcje, takie jak zakup, akceptacja kampanii czy wybór kanału sprzedaży.

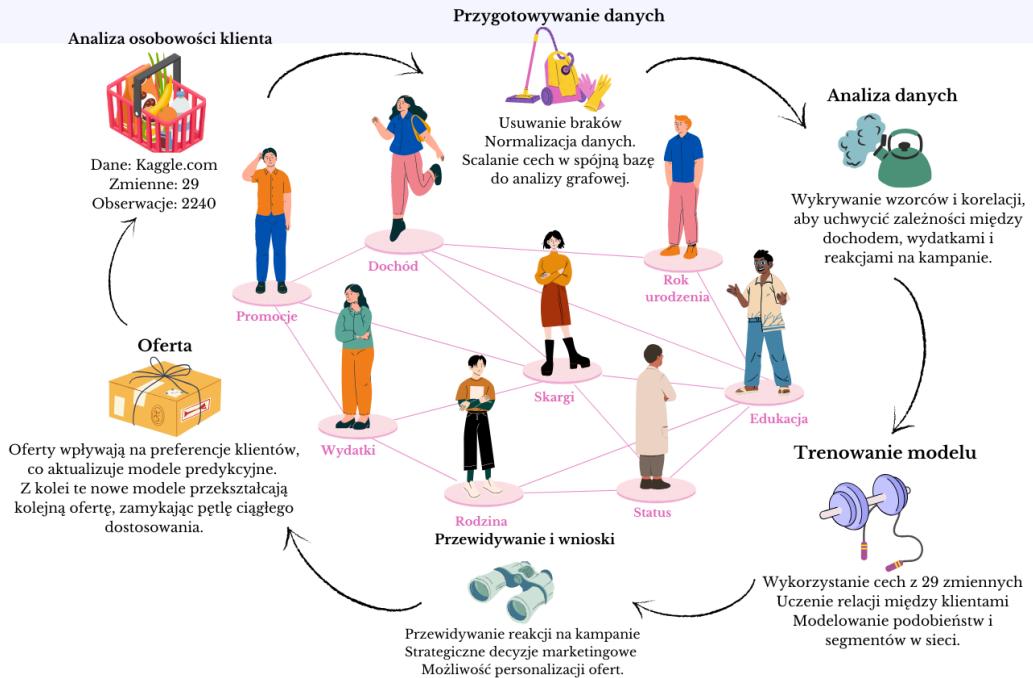
Następnie przeprowadzono eksploracyjną analizę grafową. Wykorzystano m.in. analizę głównych składowych (PCA), obliczenie macierzy odległości i budowę grafów sąsiedztwa, a także zastosowano klasteryzację i segmentację klientów.

Wyniki przeprowadzonych analiz wskazują, że zastosowanie podejścia grafowego pozwalać na efektywniejsze odwzorowanie relacji w zbiorze danych, trafniejsze profilowanie klientów oraz skuteczniejszą segmentację. Praca dowodzi, że połączenie metod uczenia maszynowego z analizą sieciową może być efektywnym narzędziem w analityce danych społeczno-gospodarczych.

Słowa kluczowe: grafowe uczenie maszynowe, grafowe bazy danych, analiza osobowości klienta.

Ryc. 1 Perspektywa 360 stopni klienta.

Grafowe uczenie maszynowe w analizie kampanii marketingowych



Źródło: Opracowanie własne.

“Using Graph Machine Learning Algorithms in Socio-Economic Analytics”

Summary

The objective of this thesis was to investigate the applicability of graph-based machine learning algorithms in the analysis of socio-economic data, with particular emphasis on customer personality modeling. The thesis combines theoretical considerations with the practical implementation of data mining tools, utilizing the “Customer Personality Analysis” dataset, which contains information on demographic, behavioral, and purchasing characteristics of customers.

The initial part of the thesis presents the theoretical foundations of graph theory and its advantages over traditional relational databases, particularly in terms of representing complex relationships between data. A graph database was constructed in the Neo4j environment, where nodes represent customers, products, sales channels, and marketing campaigns, while the edges capture real interactions, such as purchases, campaign acceptance, or sales channel selection.

Subsequently, an exploratory graph analysis was conducted. The applied methods included principal component analysis (PCA), distance matrix calculation, construction of neighborhood graphs, as well as clustering and segmentation of customers.

The results of the conducted analyses demonstrate that the graph-based approach enables a more accurate representation of relationships within the dataset, more precise customer profiling, and more effective segmentation. The thesis confirms that the integration of machine learning methods with network analysis may constitute an efficient tool in socio-economic data analytics.

Keywords: graph-based machine learning, graph databases, customer personality analysis.

Słownik pojęć:

Graf – struktura danych składająca się z węzłów i krawędzi, służąca do reprezentowania relacji między obiektami.

Węzeł – jednostka w grafie reprezentująca obiekt (np. klienta, produkt, kanał sprzedaży).

Krawędź – połączenie pomiędzy dwoma węzłami w grafie, reprezentujące relację lub interakcję między nimi.

Klasteryzacja (*clustering*) – proces grupowania obiektów w taki sposób, aby obiekty w tej samej grupie były bardziej podobne do siebie niż do obiektów z innych grup.

Klasteryzacja hierarchiczna (hierarchical clustering) – metoda grupowania, w której tworzona jest hierarchia klastrów, przedstawiana często w postaci dendrogramu.

Analiza głównych składowych (*PCA – Principal Component Analysis*) – metoda redukcji wymiarowości danych, która przekształca oryginalne zmienne w nowy zestaw zmiennych (składowych głównych) zachowując jak najwięcej informacji o wariancji danych.

Metryka euklidesowa (*Euclidean distance*) – klasyczna miara odległości między dwoma punktami w przestrzeni, obliczana jako pierwiastek kwadratowy z sumy kwadratów różnic ich współrzędnych; odzwierciedla rzeczywistą, geometryczną odległość w przestrzeni wielowymiarowej.

Neo4j – grafowa baza danych przeznaczona do przechowywania i analizy danych o strukturze grafowej.

Cypher – język zapytań używany w bazie Neo4j do wyszukiwania i manipulowania danymi w grafach.

Orange Data Mining – narzędzie do analizy danych i uczenia maszynowego, umożliwiające wizualne tworzenie przepływów pracy.

1. Wstęp.....	7
2. Cel i zakres opracowania.....	10
3. Metodyka opracowania.....	11
3.1 Założenia bazy danych.....	11
3.2 Implementacja bazy danych w środowisku Neo4j.....	14
3.3 Struktura bazy danych.....	20
4. Modelowanie klienta.....	22
4.1 Segmentacja klientów metodą RFM.....	26
5. Funkcjonalność bazy danych - przykłady.....	29
6. Analiza danych o klientach.....	32
6.1 Charakterystyka sieci.....	36
7. Podsumowanie.....	39
8. Bibliografia.....	40
8.1 Artykuły naukowe.....	40
8.2 Źródła internetowe.....	41
9. Spis tabel.....	42
10. Spis rycin.....	42

1. Wstęp

W dobie nieustannego dostępu do internetu, intensywnego przeglądania stron internetowych, portali społecznościowych, publikowania komentarzy czy wystawiania recenzji, każdego dnia generowane są ogromne ilości danych często nawet przez jednego użytkownika. Wydaje się, że na podstawie tych danych można coraz trafniej scharakteryzować człowieka lub przewidzieć jego kolejne działania, nawet bez bezpośredniej znajomości. Przykłady pozyskiwania danych są dziś wszechobecne, od zaznaczania odpowiednich kafelków ze znakami drogowymi w ramach weryfikacji Google ReCaptcha, po automatyczne logowanie dzięki plikom cookies.

Dane te wykorzystywane są na wiele sposobów, a jednym z głównych obszarów ich zastosowania jest sztuczna inteligencja traktowana jako osobna dziedzina wiedzy, obok takich nauk jak biologia czy chemia. Ważnym pod obszarem sztucznej inteligencji jest **uczenie maszynowe**, którego głównym celem jest przewidywanie wyników lub podejmowanie decyzji na podstawie dostępnych danych wejściowych. Proces ten opiera się na trzech fundamentalnych komponentach: danych i wspomnianych wcześniej przykładowych metod ich pozyskiwania, cechach zwanymi również zmiennymi lub parametrami oraz algorytmach, które zamieniają dane wejściowe w przewidywania.

Możliwości tego narzędzia można przedstawić na prostym przykładzie:

Uczenie maszynowe

Ola pracuje w urzędzie miejskim i otrzymała zadanie zaplanowania lokalizacji nowych ławek w centrum miasta. Chodzi o to, żeby zachęcić ludzi do odpoczynku, poprawić komfort przestrzeni i wspierać pieszych.

Pozornie zadanie wydaje się proste. Przejrzeć mapę, dodać ławki przy skwerach i przystankach, gotowe. Ale Ola wie, że nie każda ławka będzie używana, niektóre stoją puste a inne są przepalone.



Ryc. 2 Interpretacja uczenia maszynowego,
opracowanie własne.

Postanawia zebrać dane dotyczące m.in. natężenia ruchu pieszego, miejsc zatrzymywania się przechodniów, występowania zacienienia, lokalizacji punktów handlowo-usługowych, obecności osób starszych oraz odległości od ścian budynków i krawędzi jezdni. Po chwili uzyskuje setki punktów danych, ale nie widzi prostego wzoru i rozwiązania.

Wówczas decyduje się na wykorzystanie metod uczenia maszynowego. Wprowadza do modelu zebrane dane o przestrzeni, przepływach pieszych, punktach usługowych, nasłonecznieniu, a nawet porze dnia. Model „uczy się”, gdzie ludzie rzeczywiście chcą siadać i pokazuje, że kluczowe są nie odległości od budynków, ale widoczność i sąsiedztwo lokali gastronomicznych. W rezultacie Ola wskazuje lokalizację ławek, które rzeczywiście będą wykorzystywane. Miasto staje się przyjaźniejsze, bo decyzje planistyczne opierają się nie na intuicji, ale na analizie danych.

Dotychczasowe potrzeby używania uczenia maszynowego oraz jego dokonania można zauważać w wielu innych sferach, takich jak: rozpoznawanie głosu i twarzy w telefonie, rekomendacje na Netflixie lub wykrywanie spamu na poczcie. Wiele z tych sfer można by przedstawić za pomocą sieci i ich zbliżonych powiązań, takich jak sieć klientów w sklepie internetowym, sieć rozkładu miast i odległości między nimi, a nawet sieć rodziny, przyjaciół i ich przyjaciół.¹ Sieci nie tylko nas otaczają ale są także częścią nas samych. Dzięki złożonym układom powiązań między genami, komórki potrafią się różnicować i przyjmować różne funkcje, tworząc m.in. kości, mięśnie czy układ krwionośny. Nie sposób też pominąć najbardziej złożonej sieci w naszym organizmie: mózgu, zbudowanego z około 100 miliardów neuronów tworzących sieć powiązań o niespotykanej złożoności (Fronczak, 2021, s.12).

Dalsza część historii Oli ukazuje, jak istotną rolę odgrywają sieci na przykładzie planowania przestrzeni miejskiej.

¹ N.A.Christakis, J.H. Fowler, W sieci, Wydawnictwo Smak Słowa, Sopot 2011.

Sieci neuronowe

Doceniona za trafność swojego projektu, Ola otrzymuje nowe zadanie, tym razem znacznie bardziej złożone.

Jej zadaniem jest zaprojektowanie całej przestrzeni, z uwzględnieniem optymalnej lokalizacji przystanków, placów zabaw oraz stref relaksu, których usytuowanie sprzyjałoby realnemu wykorzystaniu przez mieszkańców. Patrząc na mapę, Ola zaczyna dostrzegać coś nowego, nie ma tu przypadkowych punktów, lecz sieć wzajemnych powiązań między miejscami i funkcjami przestrzeni.

Ludzie przemieszczają się między domem, szkołą i sklepem, dzieci podążają z placu zabaw do biblioteki, a seniorzy udają się ze skweru do przychodni, tworząc codzienną sieć powiązań i przepływów w przestrzeni miejskiej. Miasto działa jak żywy organizm. Zaczyna patrzeć na miasto jak na graf: ulice to krawędzie, ważne miejsca to węzły natomiast relacje między nimi to przepływy informacji, ludzi, energii.

Z pomocą przychodzi teoria grafów, dzięki której Ola dostrzega znacznie więcej, nie tylko to, gdzie coś się znajduje, ale przede wszystkim jak jest połączone z resztą. Myśląc sieciowo, nie tworzy już pojedynczych punktów na mapie, lecz projektuje relacje, przepływy i powiązania między elementami miejskiej przestrzeni.

W toku historii pojęcie sieci ewoluowało w zależności od dziedzin nauki, inaczej definiowano je w biologii, inaczej w matematyce, socjologii, ekonomii czy informatyce. Nauka o sieciach rozwijała się niezależnie w ramach każdej z tych dziedzin. Zainteresowanie naukowców fizyką w ubiegłym wieku, zapoczątkowało ogólną naukę o sieciach, czyli fizykę statyczną sieci złożonych.² Ich istotą jest formalne odwzorowanie relacji pomiędzy elementami za pomocą zbioru wierzchołków (nazywanych również węzłami) oraz zbioru krawędzi, które te wierzchołki łączą. W najprostszym ujęciu graf można przedstawić jako strukturę:

$$G = (V, E)$$



Ryc. 3 Interpretacja sieci neuronowych,
opracowanie własne.

² A.Fronczak, P.Fronczak, Świat sieci złożonych Od fizyki do Internetu, Wydawnictwo Naukowe PWN SA, Warszawa 2021, s. 17.

gdzie V oznacza zbiór wierzchołków, a E - zbiór krawędzi będących parami elementów ze zbioru V .

Wierzchołki reprezentują pewne obiekty, a połączenie między nimi, definiuje relacje. Obiektami mogą być na przykład dwa komputery, a krawędzią między nimi - przewód światłowodowy.³

Teoria grafów umożliwia modelowanie zarówno prostych, jak i złożonych systemów, w których istotne są nie tylko same elementy, lecz przede wszystkim relacje pomiędzy nimi. W zależności od przyjętego formalizmu, grafy mogą być skierowane lub nieskierowane, ważone lub nieważone, a także dynamiczne. Umożliwiają one analizę pojęć takich jak ścieżki, cykle, spójność, stopień wierzchołka czy centralność, które znajdują szerokie zastosowanie m.in. w analizie sieci komputerowych, struktur społecznych, systemów biologicznych czy procesów optymalizacyjnych.⁴

W dalszych częściach pracy przedstawiono, w jaki sposób narzędzia teorii grafów zostały wykorzystane w analizie zjawisk społeczno-gospodarczych.

2. Cel i zakres opracowania

Niniejsza praca ma na celu przedstawienie możliwości zastosowania narzędzi teorii grafów oraz uczenia maszynowego w analizie danych społeczno-gospodarczych. Zakres pracy obejmuje zarówno część teoretyczną, dotyczącą definicji i klasyfikacji podstawowych pojęć z zakresu grafów, jak i część praktyczną, w której zaprezentowano przykłady wykorzystania wybranych metod analitycznych na zbiorze bazy danych charakterystyki klienta.

Stawiana w pracy hipoteza zakłada, że grafowe algorytmy uczenia maszynowego umożliwiają skuteczne odwzorowanie relacji społeczno-gospodarczych oraz identyfikację kluczowych zależności między jednostkami, co przekłada się na lepsze zrozumienie ich zachowań zakupowych i wyższą trafność segmentacji w porównaniu z podejściem tradycyjnym.

³ A.Fronczak, P.Fronczak, Świat sieci złożonych Od fizyki do Internetu, Wydawnictwo Naukowe PWN SA, Warszawa 2021, s. 18.

⁴ Ł. Kowalik, Algorytmiczne problemy ścieżkowe w grafach planarnych, rozprawa doktorska, Uniwersytet Warszawski, Wydział Matematyki, Informatyki i Mechaniki, styczeń 2005, s. 17.

3. Metodyka opracowania

3.1 Założenia bazy danych

Podstawą omawianych baz danych są wcześniej wspomniane grafy, które w analizie danych klientów pozwalają uchwycić nie tylko ich indywidualne cechy, lecz także relacje zachodzące pomiędzy nimi a innymi elementami, takimi jak kanały sprzedaży, produkty, kampanie marketingowe czy decyzje zakupowe. Tradycyjne bazy danych o strukturze relacyjnej są zorientowane na przechowywanie danych tabelarycznych, gdzie powiązania między obiektami odwzorowuje się za pomocą kluczy obcych i złożonych zapytań. Choć są one efektywne w przetwarzaniu prostych struktur, ich użyteczność znaczco spada w przypadku analizy złożonych, wieloelementowych relacji, obszaru w których technologie grafowe mają przewagę.⁵ Grafowe bazy danych mogą efektywnie obsługiwać duże zbiory danych i rosnącą ilość połączonych ze sobą informacji.

Technologie grafowe są niezwykle elastyczne, co oznacza, że różnorodne, heterogeniczne informacje, takie jak adresy IP, geolokalizacja bankomatów, numery kart i identyfikatory kont, mogą być modelowane jako wierzchołki i krawędzie.⁶ Potrafią również wzbogacać dane dla modeli uczenia maszynowego, tworząc bogatsze, głębsze i bardziej kompletnie cechy. Na przykład, mogą uwzględniać relacje, takie jak znajomości wśród klientów, co prowadzi do późniejszych i dokładniejszych modeli.

Zastosowania grafów można zaobserwować w wielu branżach, takich jak usługi finansowe, gdzie umożliwiają szybsze wykrywanie oszustw; w przemyśle, gdzie przyspieszają śledzenie produktów; oraz w bezpieczeństwie publicznym, gdzie wspierają identyfikację zagrożeń i analizę powiązań między zdarzeniami.

W obszarze marketingu grafy znajdują zastosowanie w analizie 360 stopni klienta, umożliwiając integrację różnorodnych danych i dostarczając kompleksowego obrazu klienta oraz relacji między nim a produktami. Dzięki temu możliwe jest generowanie szybkich, realizowanych w czasie rzeczywistym rekomendacji, gdyż relacje są już odwzorowane w strukturach baz danych grafowych. Z tego względu podejście grafowe uznano za szczególnie adekwatne w kontekście analizy osobowości klienta oraz jego interakcji z elementami otoczenia marketingowego.

⁵ Oracle. (2024). *17 Use Cases for Graph Databases and Graph Analytics*. Pobrane z Oracle: Graph Database Use Cases Ebook.

⁶ Oracle. (2024). *17 Use Cases for Graph Databases and Graph Analytics*. Pobrane z Oracle: Graph Database Use Cases Ebook.

Spośród wielu typów grafów, w dalszej części pracy przedmiotem rozważań będą:

Graf skierowany - wierzchołki w tym grafie są połączone krawędziami skierowanymi, oznaczonymi strzałkami wskazującymi konkretny kierunek, w jakim można poruszać się po grafie.

Graf k-dzielny - w tym przypadku 4-dzielny, graf, którego zbiór wierzchołków można podzielić na k rozłącznych podzbiorów (partyj), w taki sposób, że żadne dwa wierzchołki należące do tego samego podzbioru nie są ze sobą połączone krawędzią.

Przykładem w analizowanej bazie danych będą węzły: Product, Customer, Campaign, Channel, gdzie klient jest powiązany z produktem, kampanią i kanałem sprzedaży ale nie ma połączeń między nimi wewnątrz jednej grupy.

Wykorzystany do analizy zbiór danych „Customer Personality Analysis” (Analiza profilu klienta) pochodzi z serwisu Kaggle (Romero-Hernandez, 2020) i zawiera informacje dotyczące 2240 klientów. Celem udostępnienia tej bazy jest umożliwienie prowadzenia analiz związanych z segmentacją klientów oraz modelowaniem ich zachowań rynkowych. Zbiór danych został opublikowany na platformie Kaggle przez dr. Omara Romero-Hernandezę i jest publicznie dostępny online.⁷ W niniejszym projekcie dane te wykorzystano do eksploracji oraz modelowania relacji społeczno-gospodarczych z zastosowaniem algorytmów uczenia maszynowego opartych na analizie grafów.

Zmienne występujące w zbiorze można sklasyfikować w następujący sposób:

Zmienna identyfikacyjna

- ID – unikalny identyfikator klienta; nie posiada wartości analitycznej i nie jest wykorzystywana w modelowaniu.

Zmienna daty

- Dt_Customer – data rejestracji klienta.

Zmienna pochodna

- Age – wiek klienta, wyliczony na podstawie roku urodzenia (Year_Birth).

⁷ Zbiór danych, <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>

Zmienna ciągła (ilościowa)

- Income – roczny dochód klienta (w walucie nieokreślonej w źródle danych).
- Recency – liczba dni od ostatniego zakupu (im wyższa wartość, tym mniej zaangażowany klient).
- MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds – wydatki (w jednostkach monetarnych) na poszczególne kategorie produktów w ciągu ostatnich dwóch lat.

Zmienna dyskretna (ilościowa całkowita)

- Kidhome, Teenhome – liczba dzieci i nastolatków mieszkających z klientem.
- NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth – liczba zakupów w różnych kanałach lub liczba wizyt na stronie w ostatnim miesiącu.

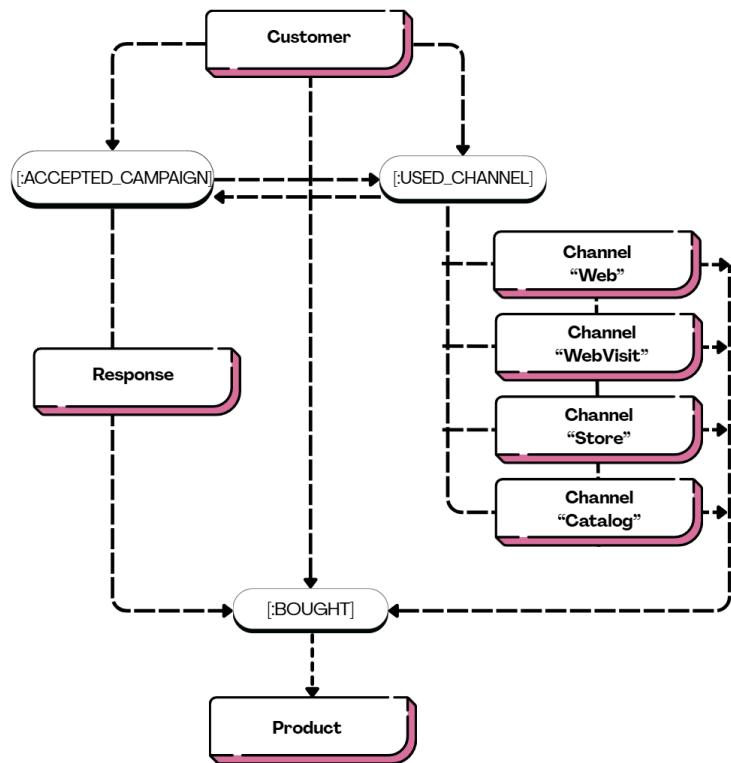
Zmienna kategoryczna (jakościowa)

- Education – poziom wykształcenia klienta (Basic, Graduation, Master, PhD, 2n Cycle).
- Marital_Status – stan cywilny klienta (Single, Married, Together, Divorced, Widow, Alone, Absurd, YOLO).

Zmienna binarna (0/1)

- AcceptedCmp1 – AcceptedCmp5 – zmienne wskazujące, czy klient zaakceptował daną kampanię marketingową.
- Complain – informacja, czy klient złożył reklamację.
- Response – ogólna reakcja na ostatnią kampanię marketingową (0 – brak odpowiedzi, 1 – pozytywna odpowiedź).

Ryc. 4 Schemat ideowy grafowej bazy danych.



Źródło: Opracowanie własne.

3.2 Implementacja bazy danych w środowisku Neo4j

Do utworzenia grafowej bazy danych posłużyono się deklaratywnym językiem zapytań grafowych jakim jest *Cypher* w programie *Neo4j Aura*. Poniższy kod przedstawia sposób uzyskania z bazy danych relacji i powiązań pomiędzy klientami (Customer), Produktami (Product), Kampaniami (Campaign) oraz Kanałami (Channel). W celu szczegółowego objaśnienia kodu, przedstawiono opis działania poszczególnych bloków kodu i ich funkcji w procesie budowy grafu.

```

None
LOAD CSV WITH HEADERS
FROM 'https://drive.google.com/uc?export=download&id=1ZP1cNVWz5mbJXJ2Nqu3CLrWRolqrLH1G' AS row
FIELDTERMINATOR '\t'
MERGE (c:Customer {id: toInteger(row.ID)})
```

```

SET
c.year_of_birth = toInteger(row.Year_Birth),
c.education = row.Education,
c.marital_status = row.Marital_Status,
c.income = toInteger(row.Income),
c.kid_home = toInteger(row.Kidhome),
c.teen_home = toInteger(row.Teenhome),
c.dt_customer = row.Dt_Customer,
c.recency = toInteger(row.Recency),
c.complain = toInteger(row.Complain)

/// Produkty
WITH c, row
FOREACH (ignoreMe IN CASE WHEN toInteger(row.MntWines) > 0 THEN [1] ELSE [] END |
MERGE (p:Product {name: "Wines"})
MERGE (c)-[r:BOUGHT]->(p)
SET r.amount = toInteger(row.MntWines)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.MntFruits) > 0 THEN [1] ELSE [] END |
MERGE (p:Product {name: "Fruits"})
MERGE (c)-[r:BOUGHT]->(p)
SET r.amount = toInteger(row.MntFruits)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.MntMeatProducts) > 0 THEN [1] ELSE [] END |
MERGE (p:Product {name: "Meat"})
MERGE (c)-[r:BOUGHT]->(p)
SET r.amount = toInteger(row.MntMeatProducts)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.MntFishProducts) > 0 THEN [1] ELSE [] END |
MERGE (p:Product {name: "Fish"})
MERGE (c)-[r:BOUGHT]->(p)
SET r.amount = toInteger(row.MntFishProducts)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.MntSweetProducts) > 0 THEN [1] ELSE [] END |
MERGE (p:Product {name: "Sweets"})
MERGE (c)-[r:BOUGHT]->(p)
SET r.amount = toInteger(row.MntSweetProducts)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.MntGoldProds) > 0 THEN [1] ELSE [] END |
MERGE (p:Product {name: "Gold"})
MERGE (c)-[r:BOUGHT]->(p)

```

```

        SET r.amount = toInteger(row.MntGoldProds)
    )
///Kampanie
FOREACH (ignoreMe IN CASE WHEN toInteger(row.NumDealsPurchases) > 0 THEN [1] ELSE [] END |
    MERGE (cam:Campaign {type: "Deals"})
    MERGE (c)-[r:ACCEPTED]->(cam)
    SET r.count = toInteger(row.NumDealsPurchases)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.AcceptedCmp1) > 0 THEN [1] ELSE [] END |
    MERGE (cam:Campaign {type: "Accepted1"})
    MERGE (c)-[r:ACCEPTED]->(cam)
    SET r.count = toInteger(row.AcceptedCmp1)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.AcceptedCmp2) > 0 THEN [1] ELSE [] END |
    MERGE (cam:Campaign {type: "Accepted2"})
    MERGE (c)-[r:ACCEPTED]->(cam)
    SET r.count = toInteger(row.AcceptedCmp2)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.AcceptedCmp3) > 0 THEN [1] ELSE [] END |
    MERGE (cam:Campaign {type: "Accepted3"})
    MERGE (c)-[r:ACCEPTED]->(cam)
    SET r.count = toInteger(row.AcceptedCmp3)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.AcceptedCmp4) > 0 THEN [1] ELSE [] END |
    MERGE (cam:Campaign {type: "Accepted4"})
    MERGE (c)-[r:ACCEPTED]->(cam)
    SET r.count = toInteger(row.AcceptedCmp4)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.AcceptedCmp5) > 0 THEN [1] ELSE [] END |
    MERGE (cam:Campaign {type: "Accepted5"})
    MERGE (c)-[r:ACCEPTED]->(cam)
    SET r.count = toInteger(row.AcceptedCmp5)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.Response) > 0 THEN [1] ELSE [] END |
    MERGE (cam:Campaign {type: "Response"})
    MERGE (c)-[r:ACCEPTED]->(cam)
    SET r.count = toInteger(row.Response)
)

/// Kanały zakupowe
FOREACH (ignoreMe IN CASE WHEN toInteger(row.NumWebPurchases) > 0 THEN [1] ELSE [] END |
    MERGE (ch:Channel {type: "Web"})

```

```

MERGE (c)-[r:USED_CHANNEL]->(ch)
SET r.count = toInteger(row.NumWebPurchases)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.NumCatalogPurchases) > 0 THEN [1] ELSE [] END |
MERGE (ch:Channel {type: "Catalog"})
MERGE (c)-[r:USED_CHANNEL]->(ch)
SET r.count = toInteger(row.NumCatalogPurchases)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.NumStorePurchases) > 0 THEN [1] ELSE [] END |
MERGE (ch:Channel {type: "Store"})
MERGE (c)-[r:USED_CHANNEL]->(ch)
SET r.count = toInteger(row.NumStorePurchases)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.NumWebVisitsMonth) > 0 THEN [1] ELSE [] END |
MERGE (ch:Channel {type: "WebVisit"})
MERGE (c)-[r:USED_CHANNEL]->(ch)
SET r.count = toInteger(row.NumWebVisitsMonth)
)

```

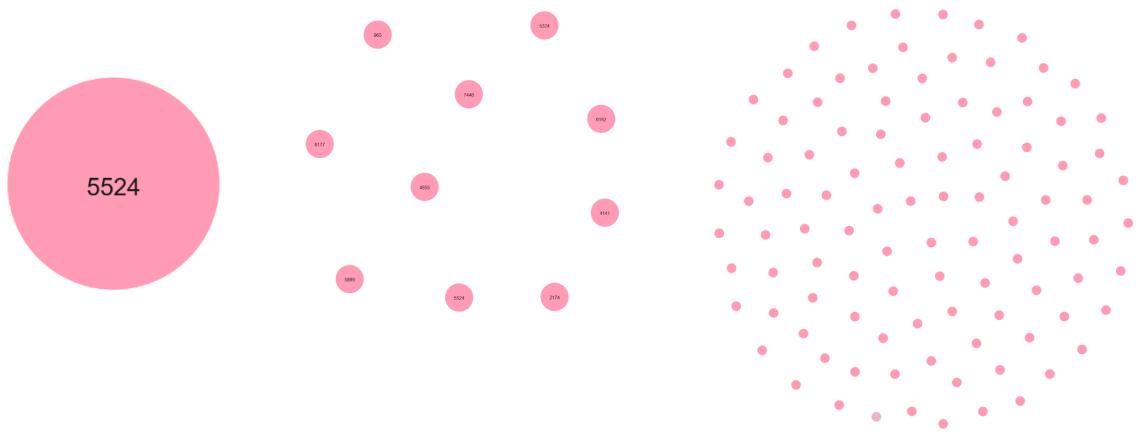
Zainportowanie danych do systemu zarządzania bazą danych *Neo4j*, a następnie utworzenie węzłów reprezentujących klientów wraz z przypisanymi im właściwościami personalnymi.

```

None
LOAD CSV WITH HEADERS
FROM 'https://drive.google.com/uc?export=download&id=1ZP1cNVWz5mbJXJ2Nqu3CLrWRolqrLH1G' AS row
FIELDTERMINATOR '\t'
MERGE (c:Customer {id: toInteger(row.ID)})
SET
  c.year_of_birth = toInteger(row.Year_Birth),
  c.education = row.Education,
  c.marital_status = row.Marital_Status,
  c.income = toInteger(row.Income),
  c.kid_home = toInteger(row.Kidhome),
  c.teen_home = toInteger(row.Teenhome),
  c.dt_customer = row.Dt_Customer,
  c.recency = toInteger(row.Recency),
  c.complain = toInteger(row.Complain)

```

Ryc. 5 Wizualizacja węzłów klientów w bazie *Neo4j* przy różnych limitach zapytania.



```
MATCH (c:Customer)
RETURN c LIMIT 1
```

```
MATCH (c:Customer)
RETURN c LIMIT 10
```

```
MATCH (c:Customer)
RETURN c LIMIT 100
```

Źródło: Opracowanie własne.

Poniższy fragment kodu tworzy węzeł reprezentujący kanał zakupowy. Jeśli klient skorzystał z danego kanału (warunek > 0), zostaje utworzona relacja między klientem a tym kanałem.

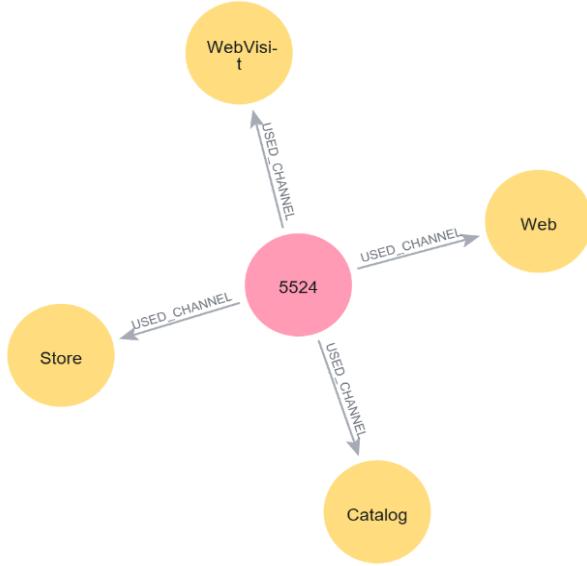
```
None
/// Kanały zakupowe
FOREACH (ignoreMe IN CASE WHEN toInteger(row.NumWebPurchases) > 0 THEN [1] ELSE [] END |
    MERGE (ch:Channel {type: "Web"})
    MERGE (c)-[r:USED_CHANNEL]->(ch)
    SET r.count = toInteger(row.NumWebPurchases)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.NumCatalogPurchases) > 0 THEN [1] ELSE [] END |
    MERGE (ch:Channel {type: "Catalog"})
    MERGE (c)-[r:USED_CHANNEL]->(ch)
    SET r.count = toInteger(row.NumCatalogPurchases)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.NumStorePurchases) > 0 THEN [1] ELSE [] END |
    MERGE (ch:Channel {type: "Store"})
    MERGE (c)-[r:USED_CHANNEL]->(ch)
    SET r.count = toInteger(row.NumStorePurchases)
)
FOREACH (ignoreMe IN CASE WHEN toInteger(row.NumWebVisitsMonth) > 0 THEN [1] ELSE [] END |
    MERGE (ch:Channel {type: "WebVisit"})
```

```

MERGE (c)-[r:USED_CHANNEL]->(ch)
SET r.count = toInteger(row.NumWebVisitsMonth)
)
MATCH (c)-[r:USED_CHANNEL]->(ch)
RETURN c, r , ch LIMIT 4

```

Ryc. 6 Węzeł relacyjny pomiędzy klientem a kanałem zakupowym.



Źródło: Opracowanie własne.

Węzeł reprezentujący klienta o identyfikatorze 5524 został połączony skierowaną krawędzią z węzłem kanału sprzedawczego, który został przez niego zaakceptowany.

Analogicznie do wcześniejszego przykładu, kod tworzy węzeł produktu, a w przypadku zakupu (warunek > 0) dodawana jest relacja łącząca klienta z tym produktem.

```

None
WITH c, row
FOREACH (ignoreMe IN CASE WHEN toInteger(row.MntWines) > 0 THEN [1] ELSE [] END |
  MERGE (p:Product {name: "Wines"})
  MERGE (c)-[r:BOUGHT]->(p)
  SET r.amount = toInteger(row.MntWines)
)
MATCH (c)-[r:BOUGHT]->(p)
RETURN c, r , p LIMIT 1

```

Ryc. 7 Węzeł relacyjny pomiędzy klientem a produktem.



Źródło: Opracowanie własne.

Klient 5524 zakupił wina.

Kod tworzy węzeł kampanii, a w przypadku akceptacji (warunek > 0) dodawana jest relacja łącząca klienta z tą kampanią.

```

None
MATCH (c)-[r:ACCEPTED]->(cam)
RETURN c, r, cam LIMIT 1
  
```

Ryc. 8 Węzeł relacyjny pomiędzy klientem a zaakceptowaną kampanią.



Źródło: Opracowanie własne.

Klient 2174 zaakceptował kampanię promocyjną.

3.3 Struktura bazy danych

Celem skryptu jest jednorazowe zimportowanie pełnego zestawu danych do bazy grafowej Neo4j oraz zbudowanie modelu połączeń między klientami, produktami, kampaniami i kanałami zakupowymi, przedstawionej w tabeli 1. Import obejmuje:

Tab. 1 Struktura bazy danych w systemie Neo4j.

Encja	Liczba właściwości	Typ relacji	Semantyka relacji
Customer	9	—	Dane socjodemograficzne i behawioralne
Product	1 (name)	(:Customer)-[:BOUGHT]-> (:Product)	Łączna wartość wydatków na kategorię
Campaign	1 (type)	(:Customer)-[:ACCEPTED]-> (:Campaign)	Liczba akceptacji/ofert
Channel	1 (type)	(:Customer)-[:USED_CHANNEL]-> (:Channel)	Liczba interakcji/purchases

Źródło: Opracowanie własne.

Przebieg importu

```
None
MERGE (c:Customer {id: toInteger(row.ID)})
SET c.year_of_birth = ... , c.complain = ...
```

- MERGE gwarantuje idempotencję – ponowne uruchomienie skryptu nie zduplikując węzłów.
- Wszystkie pola liczbowo-całkowite rzutowane są funkcją toInteger, co zapobiega zapisywaniu liczb jako danych o typie “tekst”, zachowując spójność typów danych.

Relacje zakupu produktów

Każda kategoria produktów jest tworzona lub wyszukiwana przez MERGE (p:Product {name: ...}). Sekcja FOREACH + CASE WHEN ... THEN [1] ELSE [] END pełni rolę warunkowego tworzenia relacji:

- Jeśli klient wydał > 0 w danej kategorii, powstaje krawędź (:Customer)-[:BOUGHT {amount}]->(:Product).

Dzięki temu nie występują puste krawędzie – relacje pojawiają się tylko, gdy istnieje realne zdarzenie (warunek > 0). Wykres pozostaje rzadki, a zapytania szybsze.

- Dzięki podejściu strumieniowemu (FOREACH) działa w tym samym cyklu transakcyjnym co wczytanie wiersza.

Relacje akceptacji kampanii

Analogiczny schemat jak wcześniej, ale z relacją [:ACCEPTED {count}] do węzłów Campaign. Każdy atrybut AcceptedCmp, NumDealsPurchases i Response przekłada się na inny typ kampanii (type), co ułatwia filtrowanie.

Relacje użycia kanałów

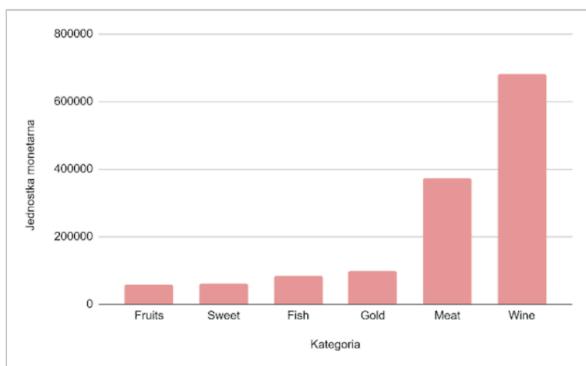
Sekcja tworzona w identyczny sposób, relacja [:USED_CHANNEL {count}] wskazuje częstotliwość korzystania z kanału lub liczbę zakupów.

4. Modelowanie klienta

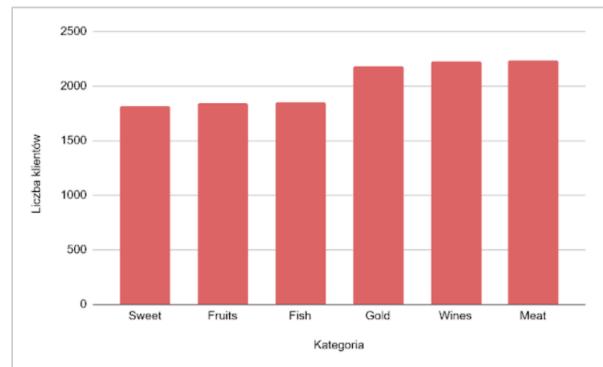
W warunkach gospodarki wolnorynkowej, gdzie rynek cechuje się trwałą nierównowagą sprzyjającą stronie popytu, producenci muszą aktywnie zabiegać o uwagę i wybory konsumentów. W takiej sytuacji kluczowe znaczenie zyskuje znajomość mechanizmów rządzących zachowaniami nabywców oraz ich reakcjami na różnorodne bodźce i narzędzia marketingowe.⁸ Podobnie jak rozwój technologii cyfrowych i internetu zrewolucjonizował sposób komunikowania się, tak również marketing uległ istotnym przeobrażeniom. Współczesne strategie marketingowe, coraz skuteczniej wpływają na decyzje konsumenckie. Dzięki precyzyjnemu profilowaniu, możliwe stało się nie tylko zrozumienie potrzeb odbiorców, ale także ich kształtowanie, niejednokrotnie prowadząc do podejmowania decyzji zakupowych w sposób bardziej impulsywny niż racjonalny. Poznanie klienta sprzyja precyzyjnemu kierowaniu kampanii marketingowych, a analiza historii jego zakupów i wcześniejszych wyborów umożliwia prognozowanie przyszłych decyzji konsumenckich.

⁸ A. Wiśniewska, „Znaczenie wiedzy o zachowaniach konsumentów dla tworzenia przekazu reklamowego”, w: A. Wiśniewska, A. Kozłowska (red.), *Reklama i PR z perspektywy współczesnych problemów komunikacji marketingowej*, Wyższa Szkoła Promocji, Mediów i Show Businessu, Warszawa 2016.

Ryc.9 Suma wydatków na kategorię.



Ryc.10 Liczba klientów kupujących z danej kategorii.



Źródło: Opracowanie własne.

Do analizy zachowań klienta, wybrano analizę marketingową, metodę behawioralną RFM. Metoda ta opiera się na trzech wymiarach: aktualności ostatniego zakupu (Recency), częstotliwości zakupów (Frequency) oraz wartości generowanych przychodów (Monetary), co pozwala na segmentację klientów według ich znaczenia i aktywności.

Recency (R) odnosi się do czasu, jaki upłynął od ostatniej transakcji klienta. Wysoka wartość wskaźnika R oznacza, że klient dokonał zakupu stosunkowo niedawno, co wskazuje na jego aktualne zainteresowanie ofertą i większe prawdopodobieństwo ponownej interakcji z marką. Niska wartość R jest natomiast sygnałem, że klient od dłuższego czasu pozostaje nieaktywny, co może wiązać się z ryzykiem jego utraty.

Frequency (F), określa częstotliwość dokonywania zakupów w analizowanym okresie. Klienci o wysokim poziomie F to osoby regularnie dokonujące transakcji, stanowiące często trzon lojalnej bazy odbiorców. Niższe wartości F wskazują na klientów okazjonalnych, których potencjał zakupowy może wymagać dodatkowej stymulacji.

Monetary (M), odnosi się do wartości pieniężnej zakupów dokonywanych przez klienta w określonym czasie. Wysoki wskaźnik M charakteryzuje tzw. klientów premium, którzy

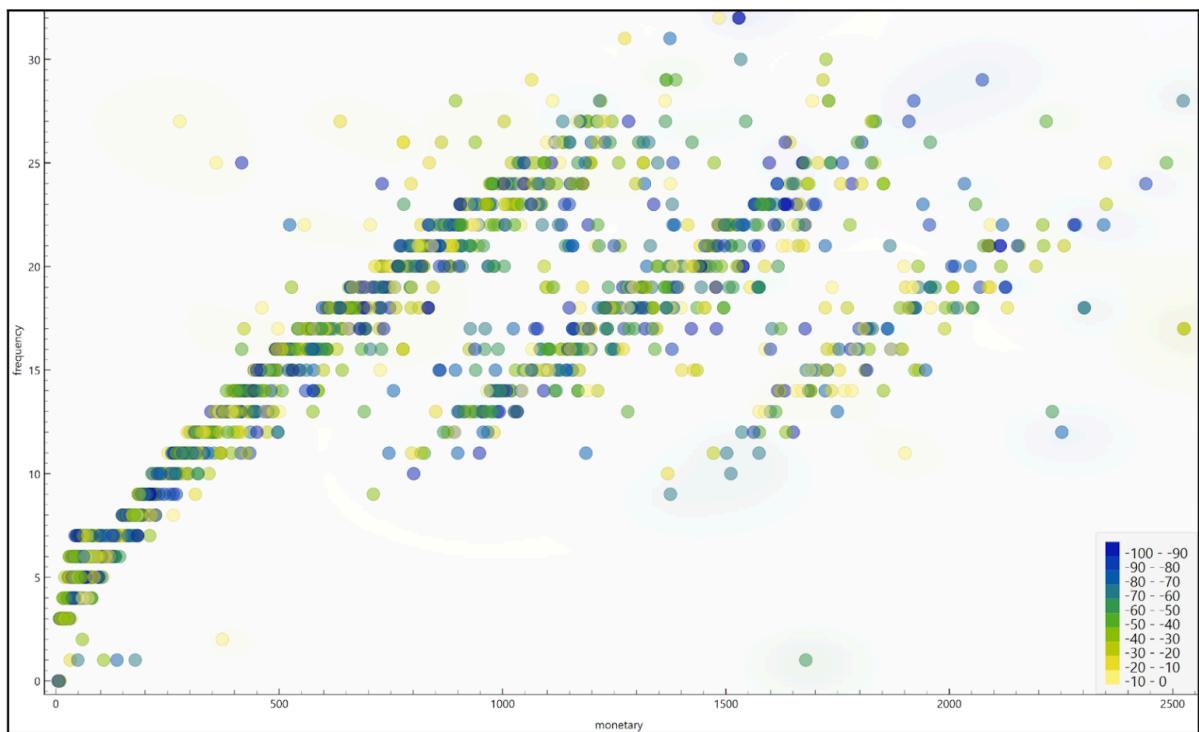
generują znaczące przychody, natomiast niskie wartości M są typowe dla klientów oszczędnych lub dokonujących niewielkich transakcji.

Metoda umożliwia podejmowanie działań oddziałujących na zachowanie klienta wyłącznie w kolejnym okresie, nie dostarczając jednak informacji o jego wartości w całym cyklu zakupowym ani w całym okresie współpracy z przedsiębiorstwem.⁹

Ze względu na obecność pierwszego wskaźnika *Recency* w opracowywanej bazie danych, obliczono pozostałe dwa wymiary modelu RFM: *Frequency*, obejmujący sumę transakcji dokonanych we wszystkich kanałach sprzedaży, oraz *Monetary*, określający łączną wartość wszystkich artykułów zakupionych przez klienta w analizowanym okresie.

⁹ M. Pawłowski, J. Banaś, Z. Pastuszak, *Wykorzystanie metody RFM do segmentacji klientów w celach marketingowych. Badanie na podstawie danych z firmy handlowej*, „Annales Universitatis Mariae Curie-Skłodowska. Sectio H” 2016, vol. L, nr 2, s. 55.

Ryc.11 Wykres rozrzutu wskaźników: Monetary, Frequency i Recency.



Źródło: Opracowanie własne.

Na wykresie rozrzutu przedstawiono zależność pomiędzy wskaźnikami *Monetary* (oś X) i *Frequency* (oś Y), natomiast kolor punktów odpowiada wartości wskaźnika *Recency*. W analizie zastosowano odwrócony wskaźnik *Recency*, aby zachować spójność interpretacyjną z pozostałymi wymiarami, w każdym przypadku wyższa wartość oznacza bardziej pożądane zachowanie klienta. Dzięki temu osoby dokonujące zakupów w krótszym czasie od momentu analizy oznaczone są jaśniejszym kolorem, natomiast klienci mniej aktywni ciemniejszym. Wykres obrazuje zależności między częstotliwością i wartością zakupów a ostatnią transakcją, wskazując na istnienie wyraźnych grup klientów o zróżnicowanych profilach zakupowych.

W celu nadania wartości punktowych każdemu klientowi, zmienne *recency*, *frequency* oraz *monetary* poddano procesowi dyskretyzacji metodą podziału na przedziały o równej liczności, dzieląc dane na sześć przedziałów, co zapewniło równomierny rozkład liczby klientów w poszczególnych kategoriach. Następnie przekształcono je do postaci zmiennych porządkowych, aby zachować hierarchię wartości w dalszych etapach analizy. W odniesieniu do każdego z wymiarów zastosowano pięciostopniową skalę ocen, w której wyższe wartości odpowiadały bardziej pożdanym zachowaniom zakupowym krótszemu czasowi od ostatniego zakupu, większej częstotliwości transakcji oraz wyższej wartości wydatków.

Tak przygotowane oceny posłużyły do wygenerowania kodów RFM oraz przypisania klientów do odpowiednich segmentów.

Ryc.12 Fragment przypisywania kodów RFM do klientów wraz z segmentacją.

	Kolumna: AE	Kolumna: AF	Kolumna: AG		
ID klienta	Recency	Frequency	Monetary	RFM kod	Segment
0	1	3	4	134	At Risk
1	5	4	3	543	Loyal Customers
9	0	2	2	22	Hibernating
13	2	0	0	200	About to Sleep
17	1	5	4	154	At Risk
20	0	2	2	22	Hibernating
22	0	3	2	32	Hibernating
24	0	2	0	20	Hibernating

```
=JEŽELI(ORAZ(AE2=5;AF2<=2;AG2<=2);"New Customers";
JEŽELI(ORAZ(AE2=4;AF2<=2;AG2<=2),"Promising";
JEŽELI(ORAZ(AE2=2;LUB(AF2>=3;AG2>=3));"Need
Attention";
JEŽELI(ORAZ(AE2=2;AF2<=2;AG2<=2); "About to Sleep";
JEŽELI(ORAZ(AE2=1;LUB(AF2>=2;AG2>=2));"At Risk";
JEŽELI(SUMA(AE2;AF2;AG2)>=14;"Champions";
JEŽELI(SUMA(AE2;AF2;AG2)>=11;"Loyal Customers";
JEŽELI(SUMA(AE2;AF2;AG2)>=8;"Potential Loyalist";
"hibernating"))))))
```

Źródło: Opracowanie własne.

4.1 Segmentacja klientów metodą RFM.

Metoda marketingowa pozwoliła na wyodrębnienie dziewięciu grup klientów, różniących się poziomem aktywności zakupowej, częstotliwością zakupów oraz wartością generowanych transakcji. Poniżej przedstawiono charakterystykę poszczególnych segmentów wraz z ich udziałem procentowym.

1. Hibernating (25,18%)

Segment obejmuje klientów, którzy w przeszłości dokonywali zakupów, jednak od dłuższego czasu pozostają nieaktywni. Wskaźniki recency oraz frequency wskazują na znaczny upływ czasu od ostatniej transakcji oraz niską częstotliwość zakupów. Grupa ta charakteryzuje się ograniczonym potencjałem generowania przychodów w obecnym stanie, jednak stanowi istotny cel działań reaktywacyjnych, takich jak kampanie przypominające czy oferty specjalne.

2. Loyal Customers (15,27%)

Do tej grupy należą klienci utrzymujący stabilny poziom zakupów w dłuższym okresie. Charakteryzują się oni wysoką częstotliwością zakupów przy umiarkowanym poziomie wartości pojedynczych transakcji. Segment ten stanowi trwałe źródło przychodów, a działania

marketingowe powinny koncentrować się na utrzymaniu ich zaangażowania poprzez programy lojalnościowe, spersonalizowane komunikaty oraz oferty specjalne.

3. Potential Loyalist (11,47%)

Segment obejmuje klientów, którzy niedawno rozpoczęli zakupy w analizowanym podmiocie i wykazują rosnącą częstotliwość transakcji. Potencjalnie mogą przejść do grupy klientów lojalnych lub nawet „Champions”. Istotne jest zastosowanie działań wzmacniających ich zaangażowanie, takich jak rabaty na kolejne zakupy czy rekommendacje produktowe dostosowane do ich wcześniejszych wyborów.

4. At Risk (11,38%)

Grupa ta obejmuje klientów, którzy w przeszłości generowali wysoki poziom przychodów, lecz obecnie wykazują oznaki spadku aktywności. Czas od ostatniej transakcji jest wydłużony, co wiąże się z ryzykiem całkowitej utraty klienta. Rekomendowane jest wdrożenie działań ukierunkowanych na przywrócenie ich aktywności, np. poprzez spersonalizowane kampanie promocyjne.

5. Need Attention (9,33%)

Klienci w tej grupie wykazują średni poziom zaangażowania, ich aktywność jest niższa niż w segmentach lojalnych, lecz wyższa niż w grupach nieaktywnych. Potencjał tej grupy polega na możliwości przesunięcia jej członków do segmentów bardziej wartościowych przy odpowiednim wsparciu marketingowym.

6. New Customers (7,81%)

Segment obejmuje osoby, które niedawno dokonały pierwszych zakupów. Etap ten ma kluczowe znaczenie dla dalszej relacji z klientem, odpowiednie działania mogą skutkować zwiększeniem częstotliwości i wartości transakcji. Wskazane jest wprowadzenie mechanizmów powitalnych, mających na celu utrzymanie zainteresowania ofertą.

7. Promising (7,72%)

Do grupy tej należą klienci, którzy dokonali zakupu w niedawnym czasie i charakteryzują się umiarkowanym potencjałem zakupowym. Ich dalsza aktywność wymaga wzmacnienia poprzez działania zachęcające do kolejnych transakcji, np. poprzez oferty powiązane z dotychczasowymi zakupami.

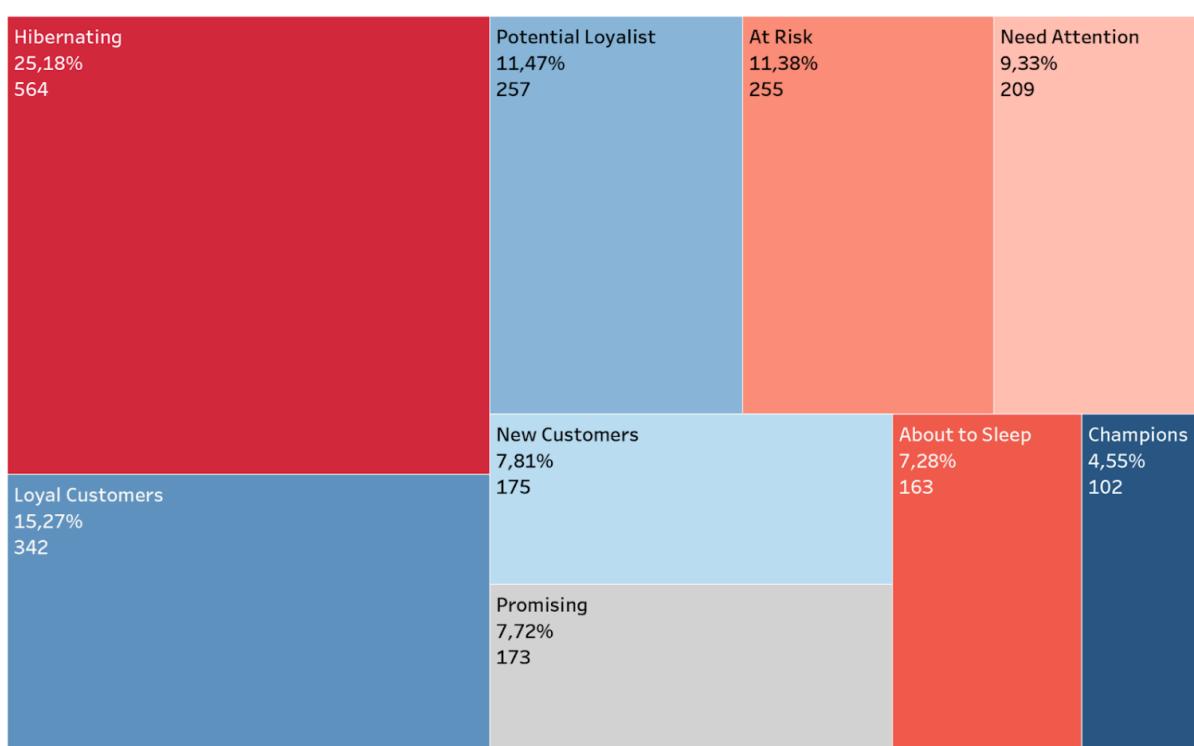
8. About to Sleep (7,28%)

Segment obejmuje klientów, którzy wykazywali aktywność w przeszłości, jednak w ostatnim okresie nastąpił spadek częstotliwości zakupów. Istnieje ryzyko, że bez podjęcia odpowiednich działań przejdą do segmentów nieaktywnych.

9. Champions (4,55%)

To najbardziej wartościowy segment klientów. Osoby dokonujące zakupów często, generujące wysokie przychody i pozostają aktywne w krótkim okresie od ostatniej transakcji. Segment ten stanowi kluczowy element bazy klientów i może pełnić funkcję „ambasadorów marki”. Rekomenduje się oferowanie im benefitów o charakterze ekskluzywnym, np. pierwszeństwa w dostępie do nowych produktów.

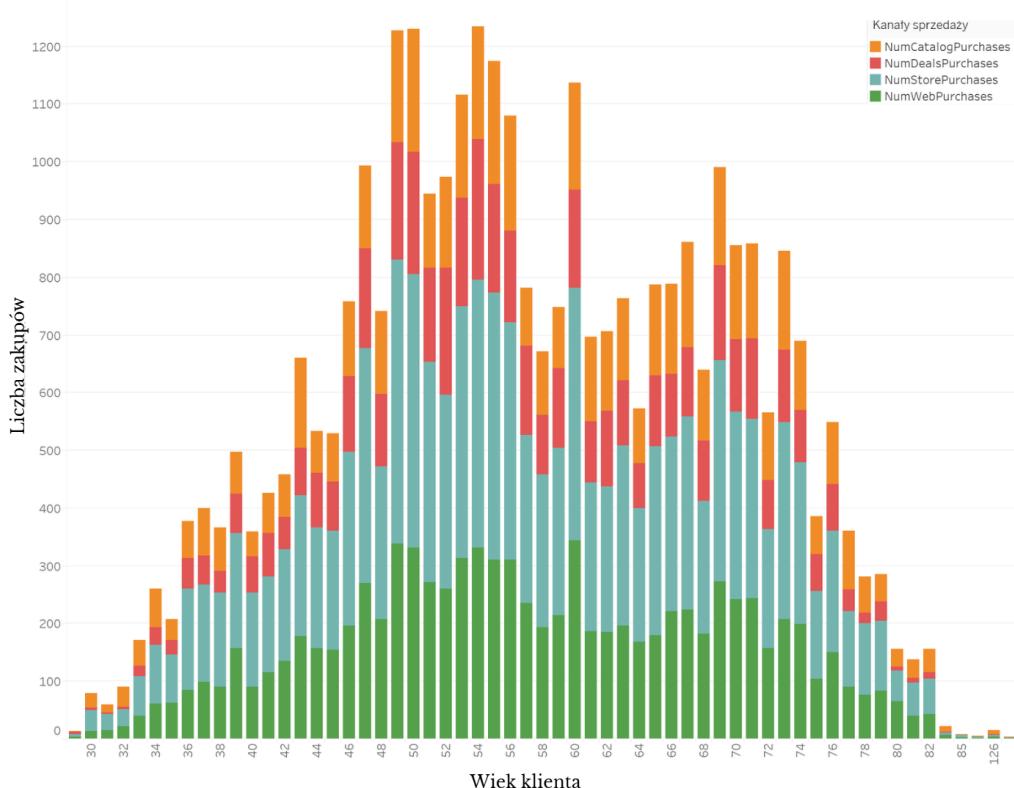
Ryc. 13 Analiza RFM.



Źródło: Opracowanie własne.

Dla osiągnięcia lepszego zrozumienia preferencji zakupowych klientów w różnych grupach wiekowych przeanalizowano ich aktywność w obrębie czterech kanałów sprzedaży: online (Web), katalogowego (Catalog), stacjonarnego (Store) oraz promocyjnego (Deals). Wizualizacja przedstawia sumaryczną liczbę dokonanych zakupów z podziałem na wiek konsumentów oraz sposób realizacji transakcji.

Ryc. 14 Liczba zakupów dokonanych w wybranych kanałach sprzedaży według wieku klienta.



Źródło: Opracowanie własne.

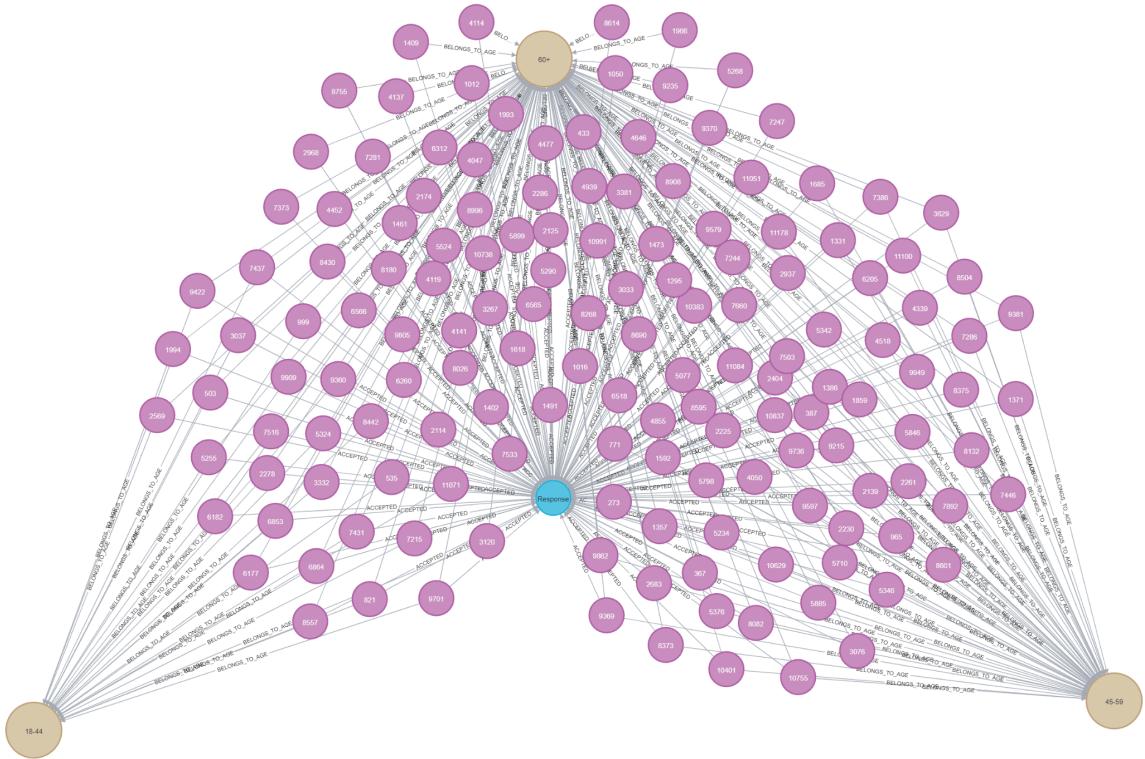
Dane wskazują, że największą aktywność zakupową wykazują klienci w przedziale wiekowym od około 45 do 65 lat. To właśnie ta grupa najczęściej korzysta z różnych kanałów sprzedaży, przy czym szczególnie wyróżnia się liczba zakupów stacjonarnych oraz przez katalog.

5. Funkcjonalność bazy danych - przykłady

W tym rozdziale przedstawiono przykłady zapytań grafowych, które ilustrują możliwości analityczne bazy danych zbudowanej w oparciu o strukturę grafową. Celem zaprezentowanych zapytań jest nie tylko pozyskiwanie konkretnych danych, ale przede wszystkim odkrywanie relacji między klientami, produktami, kanałami zakupu oraz reakcją na działania marketingowe (kampanie).

Przy pomocy zastosowania grafu możliwe jest intuicyjne analizowanie zachowań klientów w kontekście ich cech demograficznych, preferencji zakupowych oraz lojalności. Poniżej kilka z wielu możliwych zapytań grafowych analizy osobowości klienta:

Ryc. 15 Podział na grupy wiekowe, które najczęściej reagują na kampanię oraz która najbardziej dominuje.



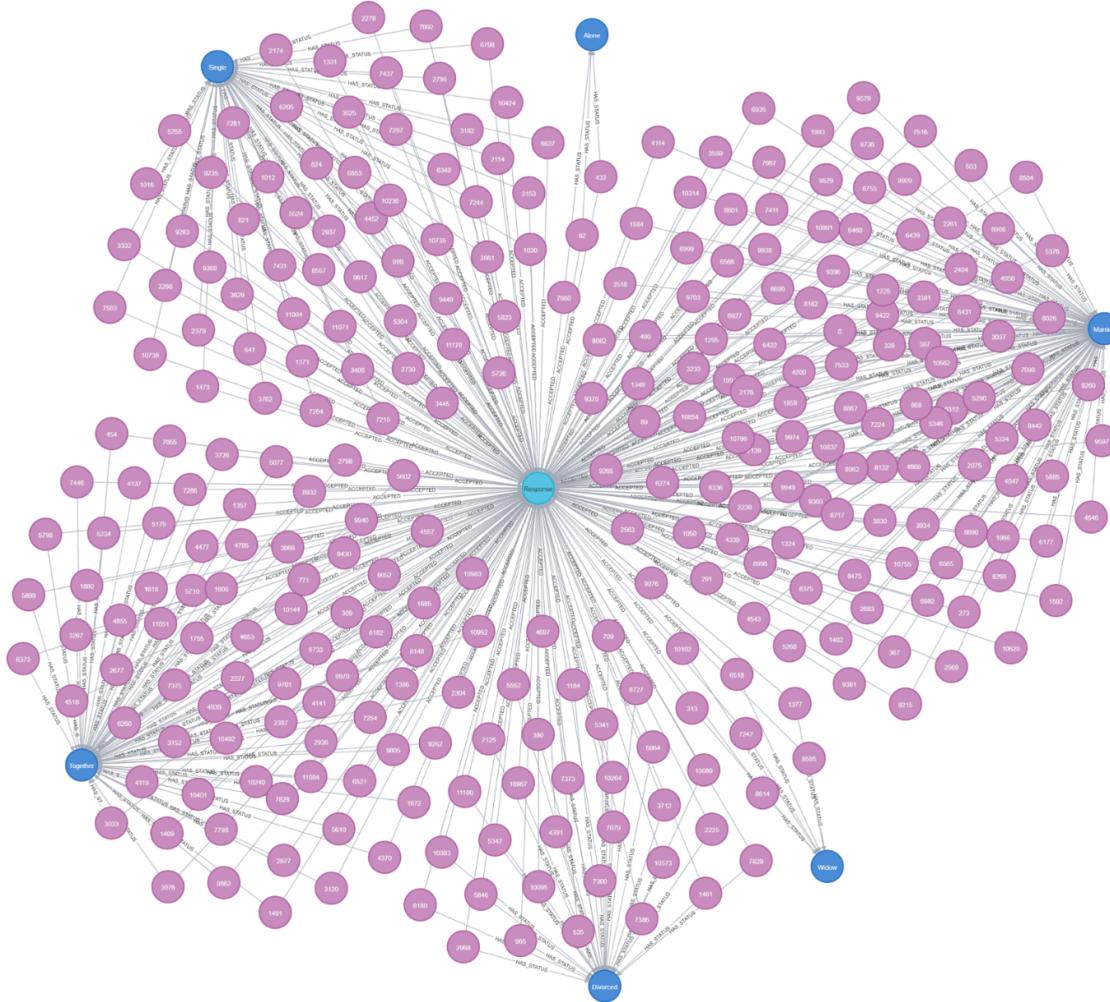
Źródło: Opracowanie własne.

```

MATCH (c:Customer)-[:ACCEPTED]->(cam:Campaign {type: "Response"})
WITH c, cam, CASE
    WHEN (2025 - c.year_of_birth) < 45 THEN '18-44'
    WHEN (2025 - c.year_of_birth) < 60 THEN '45-59'
    ELSE '60+'
END AS age_group
MERGE (g:AgeGroup {range: age_group})
MERGE (c)-[:BELONGS_TO_AGE]->(g)
RETURN g, c, cam
LIMIT 150

```

Ryc. 16 Klienci według statusu małżeńskiego, którzy odpowiedzieli na ostatnią kampanię (*response*).

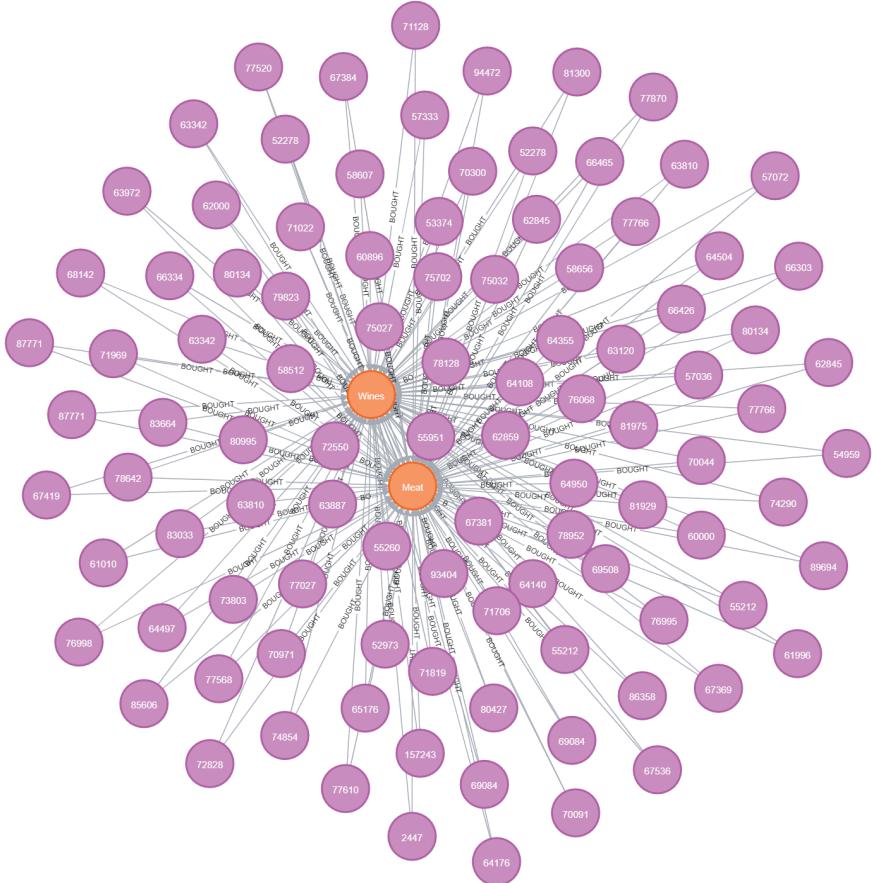


Źródło: Opracowanie własne.

```

MATCH (c:Customer)-[:ACCEPTED]->(cam:Campaign {type: "Response"})
WHERE c.marital_status IS NOT NULL
MERGE (s:MaritalStatus {status: c.marital_status})
MERGE (c)-[:HAS_STATUS]->(s)
RETURN s, c, cam
    
```

Ryc. 17 Pierwszych 100 klientów posiadających dzieci lub nastolatków, którzy dokonali największych zakupów produktów (łącznie według wartości zakupu).



Źródło: Opracowanie własne.

```

MATCH (c:Customer)-[r:BOUGHT]->(p:Product)
WHERE c.kid_home > 0 or c.teen_home > 0
RETURN c, r, p
ORDER BY r.amount DESC LIMIT 100

```

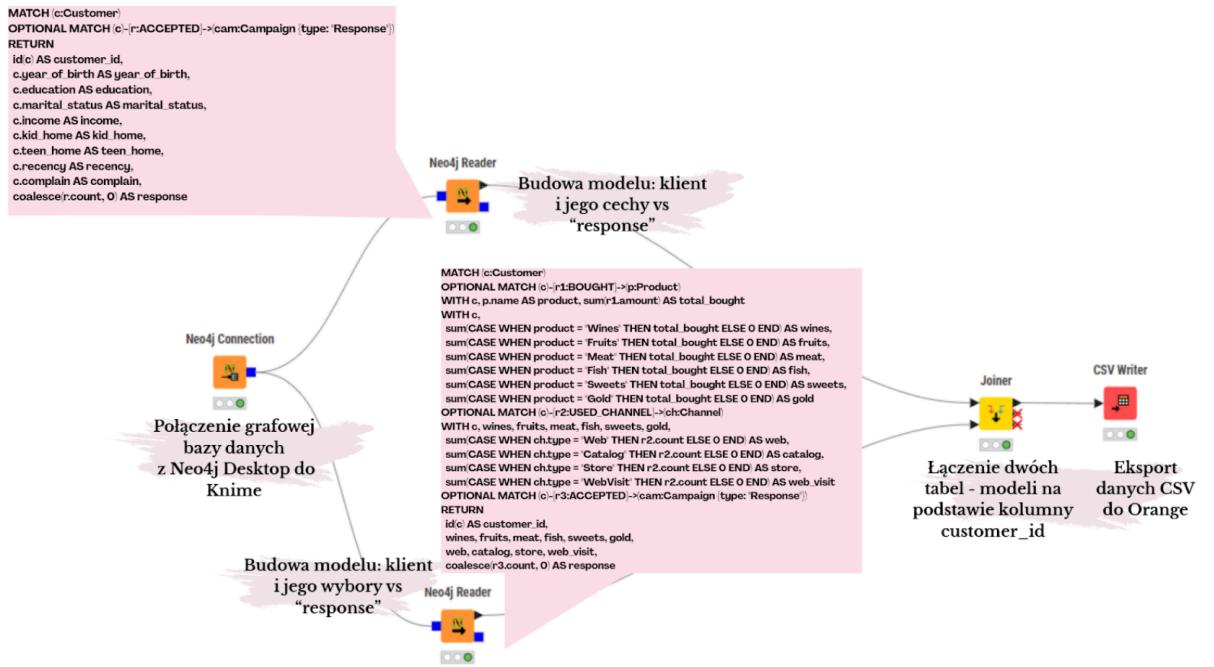
6. Analiza danych o klientach

Kluczowym elementem analizy osobowości klienta jest identyfikacja cech oraz decyzji zakupowych, które mają wpływ na reakcję klienta na kampanie marketingowe.

W tym celu rozpatrywane są dwa modele predykcyjne:

1. Model oparty na cechach klienta – analizuje, jakie właściwości (np. wiek, wykształcenie, dochód) współwystępują z akceptacją kampanii (*response*).
2. Model oparty na wyborach klienta – koncentruje się na wcześniejszych decyzjach zakupowych lub kanałach komunikacji i ich związku z reakcją na kampanię.

Ryc. 18 Pozyskiwanie i łączenie danych z bazy Neo4j w środowisku KNIME, z wykorzystaniem zapytań języka Cypher.



Źródło: Opracowanie własne.

Schemat przedstawia proces budowy dwóch modeli analitycznych na podstawie grafowej bazy danych w środowisku *Neo4j*, z wykorzystaniem programu *KNIME*. Dane zostały pobrane z lokalnej bazy *Neo4j Desktop* za pomocą węzła połączniowego (*Neo4j Connection*), a następnie przetworzone za pomocą zapytań *Cypher*.

Pierwszy model opiera się na analizie zależności pomiędzy cechami klienta (takimi jak wiek, wykształcenie, dochód czy status cywilny), a jego reakcją na kampanię marketingową (*response*). Drugi model uwzględnia zachowania zakupowe klientów i ich preferencje produktowe oraz wykorzystywane kanały sprzedaży, również w kontekście odpowiedzi na działania marketingowe.

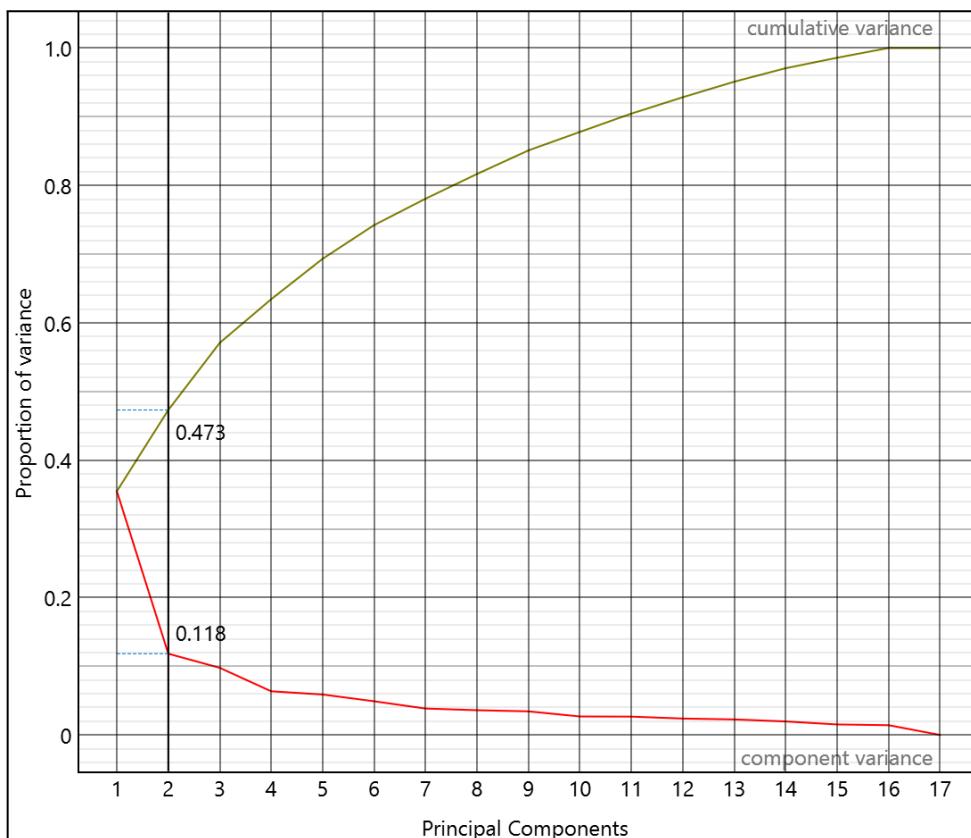
Oba modele zostały odczytane przez węzły *Neo4j Reader*, a następnie połączone przy użyciu węzła *Joiner*, bazując na wspólnej kolumnie identyfikującej klienta (*customer_id*).

Końcowym etapem było wyeksportowanie przetworzonych danych do formatu CSV za pomocą węzła *CSV Writer*, w celu dalszej analizy w środowisku Orange, aby tam następnie w poszczególnych etapach, przeprowadzić strukturę sieci oraz klasteryzację.

W pierwszym etapie analizy grafowej zastosowano analizę głównych składowych (PCA), której celem była redukcja wymiarowości oraz eliminacja współliniowości między zmiennymi. Technika ta umożliwia przekształcenie pierwotnych cech w nowy zestaw komponentów, które zachowują istotną część informacji zawartej w danych, przy jednoczesnym zmniejszeniu liczby wymiarów analizy.

Przed przystąpieniem do analizy wszystkie zmienne zostały poddane normalizacji, co umożliwiło uwzględnienie ich wpływu niezależnie od różnic w skali wartości. Na podstawie wykresu wariancji skumulowanej zdecydowano o zachowaniu 2 pierwszych składowych głównych, które łącznie wyjaśniają około 47% zmienności danych. Wstępnie rozważano zachowanie większej liczby komponentów (np. ośmiu), które łącznie wyjaśniałyby około 81% zmienności danych. Jednak ze względu na większą przejrzystość i interpretowalność wizualizacji w przestrzeni dwuwymiarowej, umożliwiając czytelniejsze zobrazowanie rozkładu klientów oraz relacji pomiędzy nimi, zdecydowano na pozostanie przy pierwszym założeniu.

Ryc. 19 Analiza głównych składowych (PCA).

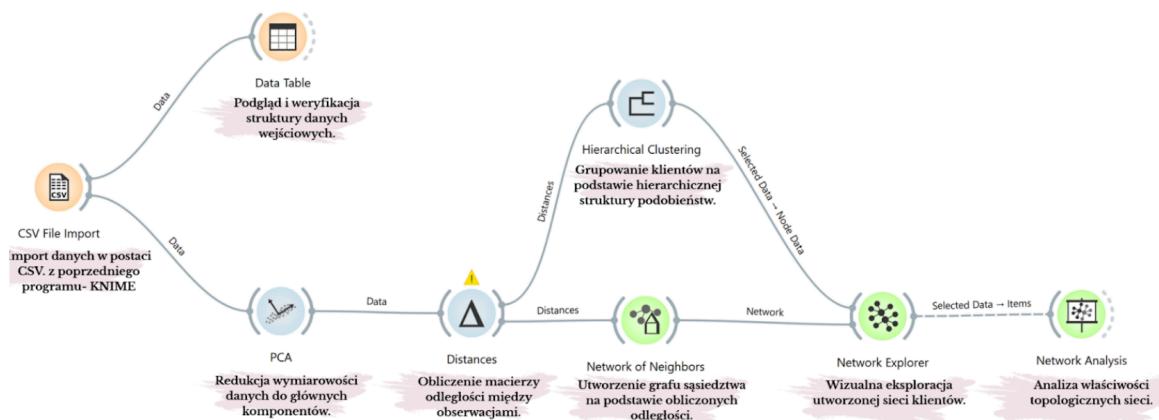


Źródło: Opracowanie własne.

Po zredukowaniu wymiarowości danych kolejnym etapem było obliczenie odległości między obserwacjami, które następnie posłużyły do utworzenia grafu sąsiedztwa. W narzędziu *Distances* zastosowano miarę odległości euklidesowej (*Euclidean distance*), będącą jedną z najczęściej stosowanych metryk w analizie danych. Odległość euklidesowa mierzy różnicę geometryczną w przestrzeni wielowymiarowej i pozwala na intuicyjną interpretację podobieństwa między obserwacjami po transformacji PCA.

Baza klientów zawiera zróżnicowane cechy, które mają różne skale. W związku z tym zdecydowano się na zastosowanie miary odległości kosinusowej, odpornej na różnice w skali zmiennych. Dodatkowo, dzięki wcześniejszej redukcji wymiarowości (PCA), dane zostały przekształcone w sposób umożliwiający porównywanie kierunku wektorów cech, czyli tzw. profilu klienta.

Ryc. 20 Etapy eksploracyjnej analizy grafowej w środowisku Orange.



Źródło: Opracowanie własne.

W tym celu wykorzystano narzędzie *Network of Neighbors*, umożliwiające modelowanie relacji pomiędzy obserwacjami (klientami) na podstawie wcześniej wyliczonej macierzy odległości dla danych zredukowanych do 16 komponentów głównych. Graf został zbudowany w oparciu o zasadę łączenia każdego wierzchołka z jego najbliższymi sąsiadami, w analizie przyjęto przedział najbliższych sąsiadów ($k=3, k=4, k=5, k=6$).

Następnie przeprowadzono klasteryzację z wykorzystaniem narzędzia *Hierarchical Clustering*, w wyniku której wyodrębniono grupy klientów o zbliżonych cechach.

Do wizualizacji sieci z uwzględnieniem przypisanych klastrów posłużyono się komponentem *Network Explorer*, umożliwiającym graficzne odwzorowanie relacji między klientami.

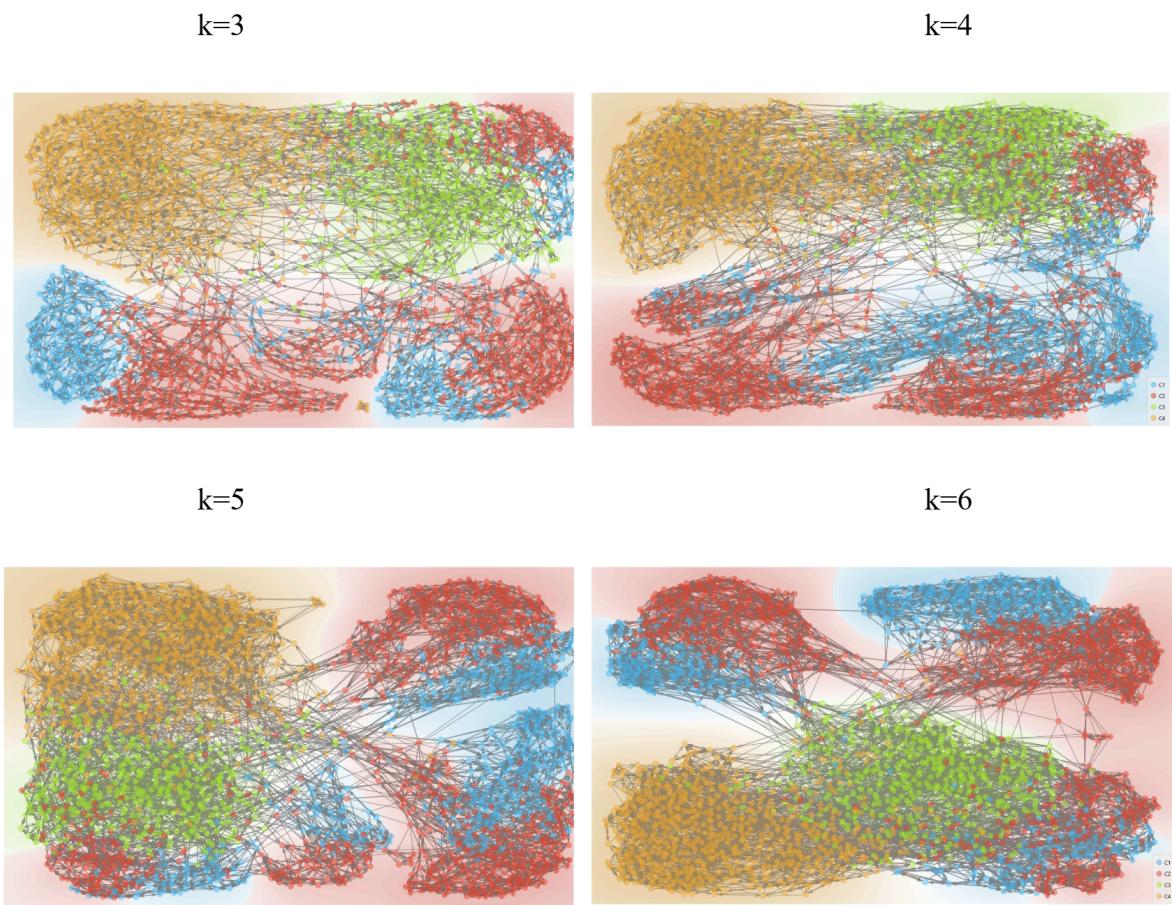
6.1 Charakterystyka sieci.

Na rycinach przedstawiono sieci klientów z podziałem na cztery klastry (C1–C4), uzyskane w module *Network Explorer* programu *Orange*. Sieci te zostały zbudowane w oparciu o algorytm k-NN, przy różnych wartościach parametru k , oznaczającego minimalną liczbę sąsiadów przypisanych do każdego węzła. Wzrost wartości parametru prowadzi do wyraźnych zmian w strukturze i spójności całej sieci. W analizie zastosowano niższą wartość parametru liczby sąsiadów k (3-6), ponieważ przy wyższych wartościach sieć stawała się zbyt gęsta, co utrudniało interpretację struktury i wyodrębnienie klastrów. Wybór ten pozwolił na uzyskanie bardziej czytelnego obrazu relacji pomiędzy klientami.

Sieć obejmuje 2240 węzłów (klientów) oraz 6720 krawędzi (relacji), przy średnim stopniu równym 3,0 i gęstości 0,00134. Występują w niej dwa komponenty spójne, co wskazuje na pewien stopień fragmentyzacji struktury. Parametry te pozostają niezmienne przy

wszystkich wartościach k (3–6), ponieważ odnoszą się do ogólnej wielkości i charakterystyki całej sieci. Zmienia się natomiast topologia grafu, czyli rozkład powiązań między klientami, co wpływa na modularność i kohezję.

Ryc. 21 Wizualizacja sieci klientów w module *Network Explorer* programu *Orange* z wyróżnieniem klastrów, przy różnych wartościach parametru liczby sąsiadów.



Źródło: Opracowanie własne.

Sieć o wartościach parametru ($k=3$), charakteryzuje się stosunkowo wysoką gęstością wewnętrz poszczególnych klastrów, przy jednocześnie czytelnych granicach pomiędzy grupami klientów. Struktura jest wyraźnie modularna, widać silne powiązania wewnętrz klastrów oraz ograniczoną liczbę połączeń między nimi. Wraz ze wzrostem liczby sąsiadów ($k=4$), następuje częściowe zatarcie granic między klastrami. W sieci pojawia się więcej relacji między klastrowymi, co skutkuje zwiększeniem spójności całej struktury. Klienci zaczynają tworzyć mosty komunikacyjne pomiędzy grupami. Na kolejnej sieci ($k=5$), widoczna jest dalsza intensyfikacja powiązań między poszczególnymi klastrami. Chociaż klastry zachowują swoje

główne obszary skupienia, to relacje między nimi stają się coraz silniejsze. Struktura sieci przybiera charakter bardziej zintegrowany, sprzyjający przepływowi informacji pomiędzy klientami z różnych grup.

Ostatnia sieć ($k=6$) swoją strukturą osiąga wysoki poziom kohezji, rozumianej jako spójność sieciowa wyrażająca stopień powiązań między węzłami¹⁰. Węzły (klienci) są powiązane wieloma ścieżkami, co wskazuje na silne przenikanie się grup i utrudnia jednoznaczne wyodrębnienie ich granic. Struktura staje się wysoce spójna i skomunikowana, co z jednej strony podnosi integrację sieci, a z drugiej zmniejsza jej modularność.

Wraz ze wzrostem wartości k sieć przechodzi od bardziej modularnej i wyraźnie podzielonej ($k=3$) do struktury zintegrowanej i silnie spójnej ($k=6$). Zjawisko to obrazuje kompromis pomiędzy wyraźnym podziałem na segmenty klientów a całościową spójnością sieci.

Na podstawie klasteryzacji wyodrębniono cztery klastry klientów (C1–C4), które obejmują osoby reagujące na kampanie marketingowe. Liczebności te odpowiadają więc liczbie pozytywnych odpowiedzi w poszczególnych segmentach. Tabela 2 przedstawia ich szczegółową charakterystykę.

Tab. 2 Charakterystyka klastrów.

Klaster	Liczba klientów	Dominujący wiek	Dochód	Preferowany kanał	Rodzina
C1	464	54	36166	Web	0.80
C2	748	54	37712	Web	0.74
C3	396	62	60673	Store	0.08
C4	632	56	73993	Store	0.06

Klaster	Wina	Owoce	Mięso	Ryby	Słodycze	Złoto
C1	52	4.8	24.4	7.8	5.4	15.2

¹⁰ Moody, J. i White, D.R. (2003) ‘Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups’, *American Sociological Review*, 68(1), ss. 103–127.

C2	94	7.0	41.1	10.1	6.6	22.6
C3	615	21.5	153	28.2	21.2	60.9
C4	542	68.0	429	97.7	70.8	80.0

Źródło: Opracowanie własne.

Na podstawie porównania struktury koszyka zakupowego można zauważyć, że klastry C1 i C2 charakteryzują się niskim poziomem konsumpcji i ograniczają się głównie do produktów podstawowych. Odmienny profil reprezentuje klaster C3, w którym obserwuje się bardzo wysokie wydatki na wina oraz znaczące zakupy mięsa i dóbr luksusowych, takich jak złoto. Klaster C4 stanowi natomiast grupę najbardziej konsumpcyjną, klienci w tym segmencie wydają zdecydowanie najczęściej we wszystkich kategoriach, a szczególnie na wina i mięso, co czyni tę grupę liderem pod względem wartości koszyka zakupowego.

7. Podsumowanie

Współczesna gospodarka opiera się w coraz większym stopniu na danych, a zdolność ich gromadzenia, przetwarzania i interpretacji staje się kluczowym zasobem zarówno w nauce, jak i w praktyce biznesowej. Złożoność relacji społecznych i ekonomicznych wymaga nowoczesnych metod analitycznych, które pozwolą nie tylko uchwycić pojedyncze zdarzenia, ale także zrozumieć sieci powiązań i mechanizmy stojące za zachowaniami jednostek oraz całych grup. W tym kontekście narzędzia oparte na teorii grafów i uczeniu maszynowym stają się istotnym kierunkiem rozwoju badań i zastosowań praktycznych.

Niniejsza praca stanowi próbę weryfikacji potencjału grafowych algorytmów uczenia maszynowego w analizie danych społeczno-gospodarczych, ze szczególnym uwzględnieniem modelowania osobowości klienta. Opracowana w środowisku *Neo4j* baza danych pozwoliła odwzorować sieci relacji między klientami, produktami, kanałami sprzedaży i kampaniami marketingowymi, co stworzyło solidny fundament do dalszych analiz. Wykorzystanie metod takich jak segmentacja RFM, analiza głównych składowych (PCA), macierz odległości czy klasteryzacja hierarchiczna umożliwiło wyodrębnienie różnorodnych segmentów klientów oraz zidentyfikowanie istotnych wzorców konsumpcyjnych.

Przeprowadzone badania dowiodły, że podejście grafowe pozwala nie tylko na trafniejsze profilowanie klientów, lecz także na pełniejsze uchwycenie dynamiki ich zachowań i reakcji na działania marketingowe. Kluczowym wynikiem było wskazanie

odmiennych grup konsumenckich, od klientów nieaktywnych i niskokosztowych, po segmenty o wysokiej wartości, które mogą stanowić fundament dla dalszych działań strategicznych. Ponadto analiza sieciowa ujawniła kompromis pomiędzy modularnością a integracją relacji w obrębie sieci klientów, pokazując złożoność procesów zachodzących w środowisku rynkowym.

Wnioski z pracy mają znaczenie zarówno dla rozwoju badań nad zastosowaniami grafowych algorytmów uczenia maszynowego, jak i dla praktyki biznesowej. Wskazują one na możliwość budowania bardziej spersonalizowanych, elastycznych i efektywnych strategii marketingowych, które odpowiadają na rosnące wymagania współczesnego rynku.

8. Bibliografia

8.1 Artykuły naukowe

1. N.A.Christakis, J.H. Fowler, W sieci, Wydawnictwo Smak Słowa, Sopot 2011.
2. A.Fronczak, P.Fronczak, Świat sieci złożonych Od fizyki do Internetu, Wydawnictwo Naukowe PWN SA, Warszawa 2021, s. 17.
3. Kowalik, Ł. (2005) Algorytmiczne problemy ścieżkowe w grafach planarnych. Rozprawa doktorska, Uniwersytet Warszawski, Wydział Matematyki, Informatyki i Mechaniki.
4. Wiśniewska, A. (2016) „Znaczenie wiedzy o zachowaniach konsumentów dla tworzenia przekazu reklamowego”, w: A. Wiśniewska, A. Kozłowska (red.), Reklama i PR z perspektywy współczesnych problemów komunikacji marketingowej, Wyższa Szkoła Promocji, Mediów i Show Businessu, Warszawa, s. 19–29.
5. Oracle. (2024). *17 Use Cases for Graph Databases and Graph Analytics*. Pobrane z Oracle: Graph Database Use Cases Ebook.
6. Oracle. (2024). *17 Use Cases for Graph Databases and Graph Analytics*. Pobrane z Oracle: Graph Database Use Cases Ebook.
7. R. Parameswari, G. Raghavendra, „Graph Databases and Applications: A Survey”, *International Journal of Innovative Science and Research Technology*, vol. 9, no. 3, 2024, s. 568–571.
8. Zbiór danych,
<https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis> - Źródło internetowe

9. A. Wiśniewska, „Znaczenie wiedzy o zachowaniach konsumentów dla tworzenia przekazu reklamowego”, w: A. Wiśniewska, A. Kozłowska (red.), *Reklama i PR z perspektywy współczesnych problemów komunikacji marketingowej*, Wyższa Szkoła Promocji, Mediów i Show Businessu, Warszawa 2016.
10. M. Pawłowski, J. Banaś, Z. Pastuszak, *Wykorzystanie metody RFM do segmentacji klientów w celach marketingowych. Badanie na podstawie danych z firmy handlowej*, „Annales Universitatis Mariae Curie-Skłodowska. Sectio H” 2016, vol. L, nr 2, s. 55.
11. Moody, J. i White, D.R. (2003) ‘Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups’, *American Sociological Review*, 68(1), ss. 103–127.
12. Fortunato, S. (2009) ‘Community detection in graphs’, *Physics Reports*, 486(3–5), pp. 75–174.
13. Schaeffer, S.E. (2007) ‘Graph clustering’, *Computer Science Review*, 1(1), pp. 27–64.
14. Mondal, R. (2024) ‘Clustering graph data: the roadmap to spectral techniques’, *Machine Learning with Knowledge Extraction*, available online.
15. Dong, X., Thanou, D., Toni, L., Bronstein, M. and Frossard, P. (2020) ‘Graph Signal Processing for Machine Learning: A Review and New Perspectives’, *IEEE Journal on Selected Topics in Signal Processing*, 14(3), pp. 589–604.
16. Kipf, T.N. and Welling, M. (2017) ‘Semi-Supervised Classification with Graph Convolutional Networks’, in *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France.
17. Hamilton, W.L., Ying, R. and Leskovec, J. (2017) ‘Inductive Representation Learning on Large Graphs’, in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, pp. 1024–1034.

8.2 Źródła internetowe

- Zbiór danych,
<https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>
- https://vas3k.com/blog/machine_learning/
- <https://mirosławmamczur.pl/czym-jest-uczenie-maszynowe-i-jakie-są-rodzaje/>
- <https://www.datasciencerobie.pl/rodzaje-ml-i-najpopularniejsze-algorytmy/>
- https://www.researchgate.net/publication/332523909_Modelowanie_strukturalne_w_analizie_zachowan_konsumentow_porownanie_metod_opartych_na_analizie_kowariancji_CB-SEM_i_czesciowych_najmniejszych_kwadratow_PLS-SEM

- <https://open.icm.edu.pl/server/api/core/bitstreams/315b5a02-0c8e-492d-87a9-62adcdca3f22/content>
- <https://pracenaukowe.wwszip.pl/prace/prace-naukowe-40.pdf#page=87>

9. Spis tabel

Tab. 1 Struktura bazy danych w systemie Neo4j.....	21
Tab. 2 Charakterystyka klastrów.....	39

10. Spis rycin

Ryc. 1. Analiza 360 stopni klienta.....	3
Ryc. 2. Interpretacja uczenia maszynowego.....	7
Ryc. 3. Interpretacja sieci neuronowych.....	9
Ryc. 4. Schemat ideowy grafowej bazy danych.....	14
Ryc. 5. Wizualizacja węzłów klientów w bazie Neo4j przy różnych limitach zapytania.....	18
Ryc. 6. Węzeł relacyjny pomiędzy klientem a kanałem zakupowym.....	19
Ryc. 7. Węzeł relacyjny pomiędzy klientem a produktem.....	20
Ryc. 8. Węzeł relacyjny pomiędzy klientem a zaakceptowaną kampanią.....	20
Ryc. 9. Suma wydatków na kategorię.....	23
Ryc. 10. Liczba klientów kupujących z danej kategorii.....	23
Ryc. 11. Wykres rozrzutu wskaźników: Monetary, Frequency i Recency.....	25
Ryc. 12. Fragment przypisywania kodów RFM do klientów wraz z segmentacją.....	26
Ryc. 13. Analiza RFM – wizualizacja segmentów klientów.....	28
Ryc. 14. Liczba zakupów dokonanych w wybranych kanałach sprzedaży według wieku klienta.....	29
Ryc. 15. Podział na grupy wiekowe, które najczęściej reagują na kampanię.....	30
Ryc. 16. Klienci według statusu małżeńskiego odpowiadający na ostatnią kampanię.....	31
Ryc. 17. Pierwszych 100 klientów posiadających dzieci lub nastolatków z największą wartością zakupów.....	32
Ryc. 18. Pozyskiwanie i łączenie danych z bazy <i>Neo4j</i> w środowisku <i>KNIME</i>	33
Ryc. 19. Analiza głównych składowych (PCA).....	34
Ryc. 20. Etapy eksploracyjnej analizy grafowej w środowisku <i>Orange</i>	35
Ryc. 21. Wizualizacja sieci klientów w module <i>Network Explorer</i> programu <i>Orange</i> z wyróżnieniem klastrów, przy różnych wartościach parametru liczby sąsiadów.....	37

