DEBIUTY STUDENCKIE

2023

ZASTOSOWANIE METOD ILOŚCIOWYCH W EKONOMII I FINANSACH

pod redakcją
Alicji Grześkowiak
i Piotra Peterneka

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2023

Recenzja
Katarzyna Ostasiewicz

Redakcja wydawnicza
Elżbieta Żurawska-Łuczyńska

Korekta
Katarzyna Gwizda

Skład i łamanie
Beata Mazur

Projekt okładki
Beata Dębska

Na okładce wykorzystano zdjęcia z zasobów 123 Royalty Free

ISBN 978-83-67899-08-6 (wersja papierowa)
ISBN 978-83-67899-09-3 (wersja elektroniczna)

DOI: 10.15611/2023.09.3

Druk i oprawa: TOTEM

Kamil Sokołowski

e-mail: 180312@student.ue.wroc.pl

ORCID: 0009-0001-9814-404X

Uniwersytet Ekonomiczny we Wrocławiu

Zastosowanie wybranych metod *data mining* w segmentacji klientów

DOI: 10.15611/2023.09.3.06
JEL Classification: C40, M3

Streszczenie: Niniejszy artykuł przedstawia wykorzystanie wybranych metod eksploracji danych w segmentacji klientów. W artykule poruszono na wstępie problematykę segmentacji oraz reklamy, gdzie jest ona obecnie bardzo często wykorzystywana. Reklama jest obecnie bardzo ważnym czynnikiem w budowaniu i promocji marki. Następnie, po dokonaniu przeglądu literatury na tematy związane z algorytmami eksploracji danych, przeprowadzono analizy bazodanowe. Podzieliły one konsumentów na grupy różniące się między sobą zarówno cechami, jak i zachowaniem. Analizy te wskazały również, które klastry klientów mogą być ważne dla przyszłych kampanii marketingowych.

Słowa kluczowe: analiza danych, RFM, *k*-średnie, segmentacja klientów, reklama, targetowanie, klastry

1. Wstęp

Segmentacja klientów jest w obecnych czasach częstym zabiegiem przeprowadzanym przez firmy w celu jak najdokładniejszego określenia potencjalnej grupy docelowej dla swoich produktów i usług. Wyznaczenie jej ma ogromne znaczenie dla działań marketingowych, gdyż bez niej ciężko prawidłowo ulokować kanał komunikacji z klientem, a co za tym idzie reklama nie będzie trafiała idealnie w odpowiednią grupę odbiorców. Reklama jest bardzo ważnym narzędziem kontaktu z potencjalnym konsumentem, dlatego też bardzo istotne jest, by docierała do tych, którzy mogą z niej skorzystać. Istotne jest to również z perspektywy odpowiedniego wykorzystania budżetu, gdyż segmentacja może sprawić, że środki na cele marketingowe nie są marnowane na reklamy, które nie mają żadnej grupy docelowej. W tym kontekście celem pracy jest zaprezentowanie możliwości i wykorzystanie eksploracji danych do przeprowadzenia segmentacji klientów. Badania przeprowadzone w artykule sprawdzą, czy wybrane analizy z zakresu *data mining* będą prowadziły do istotnych wniosków, które mogłyby zostać wykorzystane do określenia potencjalnych grup danego typu reklam. Do realizacji tego celu zostały przeprowadzone studia literaturowe oraz zastosowane podstawowe oraz zaawansowane metody zgłębiania danych. Pierwszy rozdział zawiera analizę pojęć i historii takich zagadnień, jak reklama, targetowanie klientów czy segmentacja klientów. W drugim rozdziale przedstawiono wybrane zagadnienia metod statystycznych oraz eksploracji danych oparte na studiach literaturowych. Na ich podstawie w trzecim rozdziale

zawarto rezultaty badań z wykorzystaniem metod *data mining* na danych, które zawierają wyniki kampanii marketingowej dla sklepu spożywczego. Do przetwarzania i analizy danych oraz ich wizualizacji posłuży pakiet R.

2. Reklama a określenie grupy docelowej – segmentacja klientów

2.1. Reklama – skuteczny czynnik w promocji marki

2.1.1. Definicja i historia reklamy

Jedną z możliwych definicji reklamy jest ta zaproponowana przez Grzybczyka (2012): „Każdy sposób rozpowszechniania wiadomości o świadczonych usługach i sprzedawanych towarach, który ma na celu wpłynięcie na kształtujący się popyt". Reklama polega na przekazywaniu informacji od reklamodawcy do potencjalnego konsumenta, a pochodzenie słowa wywodzi się z łacińskiego czasownika *reclamare*, który oznacza hałasowanie, robienie wrzawy i odnosi się do zwracania uwagi na reklamowaną usługę lub przedmiot.

Promocja wyrobów i usług sięga starożytności. W tamtych czasach głównym źródłem przekazu były dyskusje i rekomendacje, które można porównać do dzisiejszego *refferal marketingu*, w którym zadowoleni klienci, jak pisał Berman (2016), motywują innych do korzystania z usług lub produktów poprzez polecenia. Współcześnie firmy często korzystają z programów poleceniowych, które zachęcają klientów do promowania produktów lub usług.

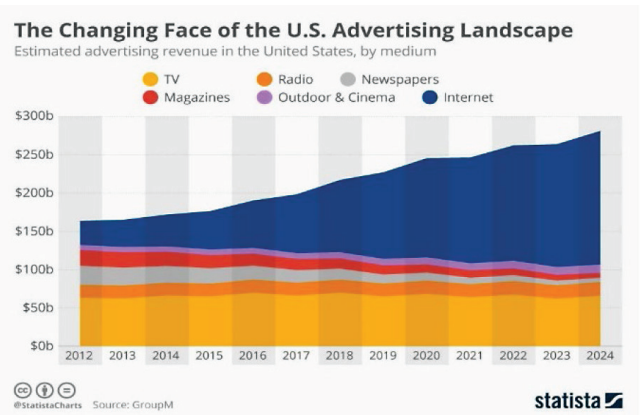
Już w starożytności zauważono, że dla rozwoju biznesu konieczne jest dotarcie do większej grupy potencjalnych klientów. Zaczęto eksperymentować z różnymi sposobami przyciągania uwagi. Pierwsze dokumentowane próby promowania produktów inaczej niż przez polecenia ustne miały miejsce w starożytnych Chinach. Tam producenci lizaków zachęcali potencjalnych nabywców do zapoznania się z ich ofertą, grając na bambusowych fletach, które przyciągały uwagę zainteresowanych klientów (Eckhardt i Bengtsson, 2010). Reklamy pisane pojawiły się wkrótce po wynalezieniu pisma w jego pierwotnej postaci. Już w czasach antycznych używano ogłoszeń, na przykład w formie opisów zbiegłych niewolników i nagród za ich odnalezienie (Kalmane, 2010).

Z czasem, wraz z rozwojem druku i technologii, reklama stała się coraz bardziej powszechna. W XX wieku nastąpił prawdziwy przełom w usługach reklamowych. Pojawienie się radia i telewizji umożliwiło dotarcie do ogromnej liczby odbiorców za pomocą dźwięku i obrazu. Pierwsza komercyjna reklama telewizyjna została wyemitowana w Stanach Zjednoczonych, kiedy firma produkująca zegarki o nazwie Bulova wypuściła reklamę swoich produktów w 1941 roku w stacji NBC (Fennis i Stroebe, 2015). Właściciele firmy Bulova, by dotrzeć do jak największej liczby odbiorców, jako moment emisji wybrali dziesięciosekundowy blok tuż przed meczem futbolowym Philadelphia Phillies i Brooklyn Dodgers. Reklamy radiowe i telewizyjne stały się popularnym sposobem promocji produktów i usług.

2.1.2. Reklamy obecnie – okres nowych możliwości

Wraz z rozwojem Internetu i mediów społecznościowych reklama cyfrowa zyskała na znaczeniu. Firmy mogą dotrzeć do swoich klientów za pomocą banerów reklamowych na stronach internetowych, reklam w wyszukiwarkach oraz reklam w mediach społecznościowych. Reklama online umożliwia także bardziej precyzyjne docieranie do określonych grup odbiorców przez personalizację i targeting reklam.

Wiele marek przenosi swoje reklamy do sieci, gdzie wydaje coraz więcej pieniędzy. Pokazano to na rysunku 1, na wykresie stworzonym na podstawie badań dla firmy GroupM, która jest największą na świecie mediową firmą inwestycyjną i odpowiada za około 30% wszystkich reklam internetowych, które widzą konsumenci w Polsce. Badanie miało na celu zobrazować szacunkowe przychody z reklam i utworzyć prognozę na kolejne lata (Richter, 2019).



Rysunek 1. Szacunkowe przychody z reklam w USA z podziałem na media

Źródło: (Richter, 2019).

Według prognoz reklama internetowa odgrywa coraz większą rolę w globalnym rynku reklamowym. Przychody z reklam internetowych mają wzrosnąć o blisko 60 mld dol., osiągając szacunkowe 284 mld dol. w 2024 roku. To oznacza, że w ciągu najbliższego roku reklama internetowa będzie odpowiadała za około 67% wszystkich przychodów reklamowych. Jednocześnie tradycyjne media, takie jak telewizja i radio, nadal utrzymają znaczący udział w wydatkach reklamowych.

W obliczu tych zmian jednym z najważniejszych aspektów marketingu internetowego jest *Social Media Marketing*. Liczba użytkowników mediów społecznościowych na całym świecie stale rośnie, osiągnęła imponującą liczbę 4,65 mld w kwietniu 2022 roku (Kepios, 2022). *Social Media Marketing* jest uważany za najbardziej wpływowy sposób komunikacji od czasów pojawienia się telewizji. Wiele firm koncentruje się na tworzeniu wysokiej jakości treści reklamowych na platformach społecznościowych, skierowanych głównie do młodszej grupy odbiorców, która stanowi

znaczną część użytkowników tych platform. Według Evansa (2011) w dzisiejszych czasach firmy muszą dobrze słuchać i wchodzić w interakcje z odbiorcami, tak by w najlepszy sposób dojść do swoich grup docelowych.

2.2. Jak dotrzeć do odpowiedniego klienta – targetowanie reklam i segmentacja klientów

2.2.1. Czym jest targetowanie reklam?

Targetowanie reklamy polega na skierowaniu jej do konkretnie dobranej grupy odbiorców o określonych cechach w celu dotarcia promocji do potencjalnych nabywców określonego produktu lub usługi (Maciorowski, 2013). Dzięki targetowaniu możliwe jest odpowiednie wykorzystanie budżetu reklamowego przez uniknięcie strat wynikających z dotarcia do osób, których preferencje nie pokrywają się z atrybutami oferowanego produktu (Dibb i Simkin, 1991). Segmentacja pozwala również na elastyczną reakcję na zmiany rynkowe przy uwzględnieniu faktu, że żadne dobro ani usługa nie przemawiają do wszystkich konsumentów, a różni klienci mogą mieć różne powody dokonywania zakupów (Wind i Bell, 2008). Reklamodawcy często wykorzystują segmentację w sieci, gdzie można łatwo śledzić aktywność konsumentów, jednakże dostosowane reklamy do profilu odbiorcy są stosowane również w innych mediach, na przykład reklama wędki w magazynie wędkarskim, która trafia w większym stopniu do pasjonatów wędkarstwa niż do przeciętnego czytelnika gazety codziennej, czy też reklamy sklepu budowlanego emitowane na kanale, który koncentruje się na treściach związanych z remontami i budową domów. Mimo to targetowanie reklam w Internecie jest najbardziej opłacalne, ponieważ pozwala na skierowanie treści do mniejszych, bardziej niszowych grup na podstawie zachowań konsumenckich, a także umożliwia bezpośredni kontakt klienta z firmą przez stronę internetową lub sklep internetowy, co przyspiesza proces zakupowy i zapewnia większą wygodę odbiorcy (Jaska i Werenowska, 2016).

W kontekście targetowania reklam wyróżnia się kilka rodzajów. Targetowanie kontekstowe opiera się na słowach kluczowych obecnych na stronach internetowych, które warunkują pojawienie się odpowiednich pozycji reklamowych. Natomiast retargeting polega na kierowaniu reklam do klientów, którzy wcześniej mieli kontakt z daną marką.

2.2.2. Segmentacja klientów

Segmentacja klientów odgrywa kluczową rolę w celu skutecznego targetowania reklam. Aby osiągnąć największą precyzję, konieczne jest opracowanie rozbudowanego i wielowarstwowego profilu potencjalnego odbiorcy na podstawie danych klientów i ich wcześniejszych zachowań konsumenckich. Firmy zwykle nie mają możliwości tworzenia niestandardowych ofert dla każdego klienta, dlatego ważne jest stworzenie grup, w których znajdują się klienci o podobnych cechach. Segmen-

tacja klientów polega na podziale rynku na części lub segmenty, które spełniają cztery kluczowe cechy: dostępność, mierzalność, obszerność i reaktywność.

Segmentacja klientów rozwinęła się od czasu jej pierwszego wprowadzenia przez Wendella R. Smitha w 1956 roku. Pierwsze próby segmentacji opierały się głównie na demograficznych cechach, takich jak wiek, płeć, wykształcenie i klasa społeczna (Yankelovich i Meer, 2006). Obecnie uwzględnia się również sytuację finansową, a także czynniki behawioralne i geograficzne.

Do przeprowadzenia badań segmentacyjnych można zastosować wiele metodyk, a stosowanie kombinacji metod często zapewnia najbardziej użyteczny wynik. Przykładem może być rozpoczęcie badań od ankiety wśród klientów, a następnie uzupełnienie wyników badaniami jakościowymi w celu lepszego obliczenia korzyści i barier powiązanych z danym zachowaniem. Poza wyżej wspomnianymi ankietami, przeprowadzanymi przez markę, innymi źródłami dostępnymi wewnątrz firmy są raporty sprzedażowe oraz zestawienia dotyczące przeprowadzanych z klientami transakcji.

Podczas segmentacji klientów firma może napotkać wiele problemów, z których jeden to jakość wybranych danych. Niedokładne dane źródłowe zazwyczaj prowadzą do słabego grupowania. Na przykład mogą pojawić się nieprawdziwe informacje, które są wprowadzane w celu ukrycia danych, takie jak wiek, płeć i stan cywilny. Na szczęście wartości odstające można łatwo wykryć, a także często pojawiające się się braki w danych związane z nieodpowiednio przygotowanymi ankietami. Dlatego ważne jest regularne sprawdzanie i oczyszczanie danych.

Po zastosowaniu segmentacji ważny jest również wybór właściwych segmentów, na których należy się skoncentrować. Metoda TARPARE opracowana przez Donovana i Henleya w 2003 roku pozwala na łatwe dopasowanie wag i istotności do utworzonych grup. Pierwotnie została opracowana do wyboru docelowych odbiorców interwencji w zakresie zdrowia publicznego. Metoda TARPARE składa się z sześciu części:

T – całkowita liczba osób. Im więcej osób znajduje się w grupie, tym istotniejsza powinna być dla firmy;

AR – odsetek osób narażonych na ryzyko. Następuje podział na małe, średnie i duże ryzyko. Ta część jest ściśle powiązana z pierwotnym celem metody TARPARE. Przy segmentacji klientów ryzyko zamieniane jest na zaangażowanie finansowe. Im wyższy odsetek zaangażowania, tym większy potencjalny zwrot, a zatem wyższy priorytet dla tej kategorii;

P – możliwość przekonania docelowej grupy odbiorców. Im bardziej możliwa jest zmiana postaw lub zachowań segmentu, tym większy priorytet ma ta klasa;

A – dostępność docelowej grupy odbiorców. Określa, jak łatwo (i opłacalnie) można dotrzeć do każdego segmentu za pośrednictwem dostępnych kanałów;

R – zasoby potrzebne do zaspokojenia potrzeb docelowej grupy. Czy do dotarcia do klienta są potrzebne zasoby finansowe, ludzkie i strukturalne oraz czy zasoby te są dostępne, czy trzeba je pozyskać;

E – sprawiedliwość społeczna. Czy małe grupy wymagają specjalnych programów ze względu na sprawiedliwość społeczną.

Segmentacja klientów może być przeprowadzana na różnych poziomach, a w przypadku dużych grup można tworzyć subsegmenty. Małe grupy, które wyróżniają się znacząco spośród innych, nazywane są niszami rynkowymi (Hajdas, 2014). Przykładem niszy rynkowej może być zapotrzebowanie na produkty bezglutenowe ze względu na uczulenie potencjalnych odbiorców.

3. Wybrane zagadnienia *data mining* – metodyka badań

3.1. *Data mining* – wprowadzenie

Problematyka zgłębiania danych, zwana także *data mining*, jest bezpośrednio powiązana z metodyką CRISP. CRISP, czyli *Cross Industry Standard Process for Data Mining*, został zaprezentowany po raz pierwszy na IV Warsztatach CRISP-DM SIG w 1999 roku (Chapman, Clinton, Khabaza, Reinartz i Wirth, 1999; Shearer, 2000). Metodyka ta dzieli proces eksploracji danych na sześć faz: zrozumienie uwarunkowań biznesowych; zrozumienie danych; przygotowanie danych; modelowanie; ewaluacja i wdrożenie.

Eksploracja danych ma długą historię. Termin *data mining* został użyty już w artykule M. C. Lovella z lat osiemdziesiątych XX wieku (Lovell, 1983). *Data mining* łączy w sobie statystykę, sztuczną inteligencję, uczenie maszynowe i badanie baz danych. Proces ten polega na zastosowaniu algorytmów analizy danych w celu odkrycia wzorców i modeli w danych. Kluczowym etapem jest eksploracja danych, w której stosuje się różne metody w celu wyodrębnienia prawidłowości (Fayyad, Piatetsky, Shapiro i Smyth, 1996).

Proces odkrywania wiedzy można podzielić na siedem kroków, które mają bezpośrednie odniesienie do trzeciego, czwartego i piątego etapu metodyki CRISP (Han, Kamber i Pei, 2012):

1. Czyszczenie danych (w celu usunięcia brakujących i niespójnych danych).
2. Integracja danych (łączenie danych w istotne kombinacje).
3. Selekcja danych (gdzie dane istotne dla analizy są pobierane z bazy danych).
4. Transformacja danych (gdzie dane są przekształcane do postaci odpowiednich do eksploracji).
5. Eksploracja danych (proces właściwy, gdzie wybrane metody są stosowane do wyodrębniania prawidłowości wśród danych).
6. Ewaluacja wzorców (w celu zidentyfikowania naprawdę interesujących wzorców reprezentujących wiedzę na podstawie pomiarów ciekawości).
7. Prezentacja wiedzy (gdzie wykorzystuje się techniki wizualizacji wiedzy do zaprezentowania wydobytych informacji).

Sam etap eksploracji danych może być nadzorowany lub nienadzorowany. Wykryte interesujące wzorce mogą być wizualizowane i przechowywane jako nowe zmienne w bazach danych (Han i in., 2012).

Data mining ma zastosowanie w różnych dziedzinach, takich jak gospodarka, nauki przyrodnicze i technologia. Przykładowo jest obiecującą technologią w dziedzinie uczenia się systemów autonomicznych, w tym robotów humanoidalnych i samochodów. Istnieje wiele źródeł danych, w tym bazy danych, hurtownie danych, sieć WWW, inne repozytoria informacji oraz dane dynamicznie przesyłane strumieniowo do systemu (Dillmann, 2004).

3.2. Charakterystyka wykorzystywanych danych

Praca wykorzystuje bazę danych opracowaną przez doktora Omara Romero-Hernandeza, obecnie profesora na UC. Berkeley's Haas School of Business oraz w Hult International Business School. Baza danych przedstawia wyniki kampanii marketingowej w sklepie spożywczym i składa się z 28 zmiennych oraz 2240 wyników. Zmienne zostały podzielone na cztery grupy: *People*, *Products*, *Promotion i Place*.

Grupa *People* zawiera informacje dotyczące konsumentów, takie jak numer identyfikacyjny klienta, rok urodzenia, poziom wykształcenia, stan cywilny, roczny dochód gospodarstwa domowego, liczba dzieci w gospodarstwie domowym, data rejestracji klienta, liczba dni od ostatniego zakupu oraz informacja o reklamacji w ciągu ostatnich dwóch lat.

Grupa *Products* zawiera informacje dotyczące wydatków konsumentów na różne kategorie produktów w ciągu ostatnich dwóch lat, takie jak wydatki na wino, owoce, produkty mięsne, ryby, słodycze oraz złoto.

Grupa *Promotion* zawiera informacje dotyczące responsywności klientów na proponowane promocje, w tym liczbę zakupów wykonanych z użyciem promocji oraz informacje o przyjęciu ofert w pięciu różnych kampaniach.

Ostatnia grupa *Place*, dotyczy miejsc, w których klienci dokonywali zakupów, i zawiera informacje o liczbie zakupów dokonanych za pośrednictwem strony internetowej, katalogu sklepowego oraz bezpośrednio w sklepie, a także o liczbie wejść na stronę firmy w ciągu ostatniego miesiąca.

Baza danych zapewnia szeroki zakres informacji na temat klientów sklepu spożywczego i ich zachowań zakupowych, co umożliwia przeprowadzenie analizy i eksploracji w artykule.

3.3. Opis zastosowanych metod analitycznych.

Jak podkreślono w podpunkcie 3.1, pierwszym krokiem do prawidłowego przeprowadzenia eksploracji jest wstępne przetwarzanie zbioru, w skład czego wchodzi czyszczenie, integracja, selekcja i transformacja danych. W celu przygotowania danych do interpretacji należało zająć się brakującymi danymi. Wybrany został jeden z najpopularniejszych sposobów, czyli usunięcie niepełnych wierszy (Szeliga, 2017). Nie zostały również wzięte pod uwagę dane z wartościami odstającymi, które zostały wyznaczone na podstawie wyznaczonych kwartyli Q_1 i Q_3 (Han i in., 2012). Kolejnym krokiem, który należało wykonać w celu lepszego przeprowadzenia segmentacji

klientów, było przekształcenie zmiennych oraz wybór tylko tych, które są rzeczywiście istotne dla analizy. W ten sposób przekształcono lub usunięto siedem zmiennych:

1. *Wiek* – liczba lat klienta (przekształcenie zmiennej Year_Birth).
2. *Wykształcenie* – wartości „2n Cycle" zostały zamienione na „Master", aby odzwierciedlić ten sam poziom wykształcenia, co w Stanach Zjednoczonych.
3. *Liczba_dzieci* – zmienna Kidhome i Teenhome zostały zastąpione przez jedną zmienną, która zawiera informację o liczbie wszystkich dzieci w gospodarstwie domowym.
4. *Wszystkie_kampanie* – informacje o wcześniejszym skorzystaniu z kampanii marketingowych zostały połączone.
5. *Wydatki* – koszty wszystkich rodzajów produktów zostały dodane do siebie.
6. *Ile_klientem* – utworzono zmienną, która informuje, jak długo klient jest zapisany w bazie danych firmy. Oblicza się ją przez odjęcie daty dopisania do kartoteki klienta od daty ostatniego dopisanego rekordu (datowanej na 06.12.2014).
7. Zmienna *Complain* została usunięta jako mało istotna dla przeprowadzanych badań. W efekcie przeprowadzonych operacji pozostało 2208 wyników dla 21 zmiennych.

W sytuacji, gdy analizuje się wiele zmiennych, istotnym czynnikiem może być występowanie powiazań (zależności) między nimi. Do najczęściej wykorzystywanej miary tych zależności należy współczynnik korelacji Pearsona, który pozwala opisać stopień powiązania liniowego między dwiema zmiennymi. Został on opracowany przez angielskiego matematyka, Karla Pearsona, w 1896 roku na podstawie pracy między innymi Francisa Galtona z 1889 roku, który jako pierwszy wprowadził pojęcie korelacji. Współczynnik korelacji Pearsona przyjmuje wartości z zakresu [–1, 1]. Wartość 0 oznacza całkowity brak korelacji, natomiast im bliżej 1 bądź –1 tym silniejsza jest zależność liniowa między zmiennymi. Znak przed liczbą wskazuje na kierunek zależności. (Dziechciarz, 2003). Sposób wyliczenia korelacji Pearsona zaprezentowany został na wzorze (1):

$$r_{xy} = \frac{C(X,Y)}{s_x s_y} \tag{1}$$

gdzie: r_{xy} – korelacja Pearsona między cechami X i Y, $C(X, Y)$ – kowariancja między cechami X i Y, s_x – odchylenie standardowe zmiennej X, s_y – odchylenie standardowe zmiennej Y.

Aby wartości dla różnych zmiennych mogły być względem siebie porównywalne, konieczne jest przeprowadzenie normalizacji. Polega ona na przeskalowaniu danych, tak by mieściły się w bardzo małym zakresie, przykładowo [–1, 1]. Znanych jest wiele metod normalizacji obserwacji, w tym: metoda min-max (transformacja liniowa), metoda Z-score (standaryzacja), metoda przez skalowanie funkcją eksponencjalną, metoda doprowadzania do rozkładu logarytmiczno-normalnego (Szeliga, 2017).

Rodzajem normalizacji przeprowadzonym w rozdziale trzecim jest standaryzacja, czyli metoda Z-score. Jej celem jest takie przeskalowanie danych, by otrzymać rozkład ze średnią 0, a odchyleniem standardowym równym 1. Po przeprowadzeniu standaryzacji wartości wcześniej mniejsze od średniej dla danej zmiennej będą miały wartość ujemną, zaś większe od średniej – dodatnią.

Ustandaryzowane dane wykorzystać można do segmentacji, czyli podzielenia na klastry. Algorytmy grupowania danych należą do analizy skupień, czyli metody klasyfikacji bez nadzoru. Za ojca analizy skupień uważany jest Robert Tryon (2008). Algorytmy grupowań można podzielić na pięć kategorii: hierarchiczne, niehierarchiczne, inaczej nazywane iteracyjno-optymalizacyjnymi, gęstościowe, gridowe i modelowe (Han i in., 2012). Najczęściej wykorzystywane algorytmy grupowań to dwa wymienione jako pierwsze.

Grupowanie hierarchiczne można zwizualizować za pomocą dendrogramu klastrowego, który ma dwa podejścia: aglomeracyjne oraz deglomeracyjne. W tym pierwszym każda obserwacja początkowo tworzy swój własny klaster, który następnie łączy się z najbliższym klastrem, aż do zakończenia działania algortymu. Efektem końcowym jest drzewo hierarchiczne, czyli dendrogram. Zakończenie algorytmu następuje, gdy każdy obiekt znajduje się w pojedynczym osobnym klastrze. Ze względu na charakter hierarchiczny do zastosowania wyżej wymienionych algorytmów nie jest konieczne wstępne określenie liczby tworzonych klastrów. Aby poprawnie przeprowadzić grupowanie hierarchiczne, należy wyznaczyć odległości między każdym z punktów. Dzięki temu, klastry będzie można dobrać w taki sposób, by odległości między punktami w tym samym klastrze były jak najmniejsze, a pomiędzy różnymi klastrami największe (Kassambara, 2017). Głównymi miarami odległości są: odległość euklidesowa, kwadratowa odległość euklidesowa, maksymalna odległość, odległość Manhattan.

W rozdziale trzecim miary odległości zostały obliczone za pomocą odległości euklidesowej przy pomocy wzoru (2):

$$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2} \tag{2}$$

gdzie: $\|a-b\|_2$ – odległość euklidesowa między punktami *a* i *b*, a_i – *i*-ta współrzędna punktu *a,* b_i – *i*-ta współrzędna punktu *b.*

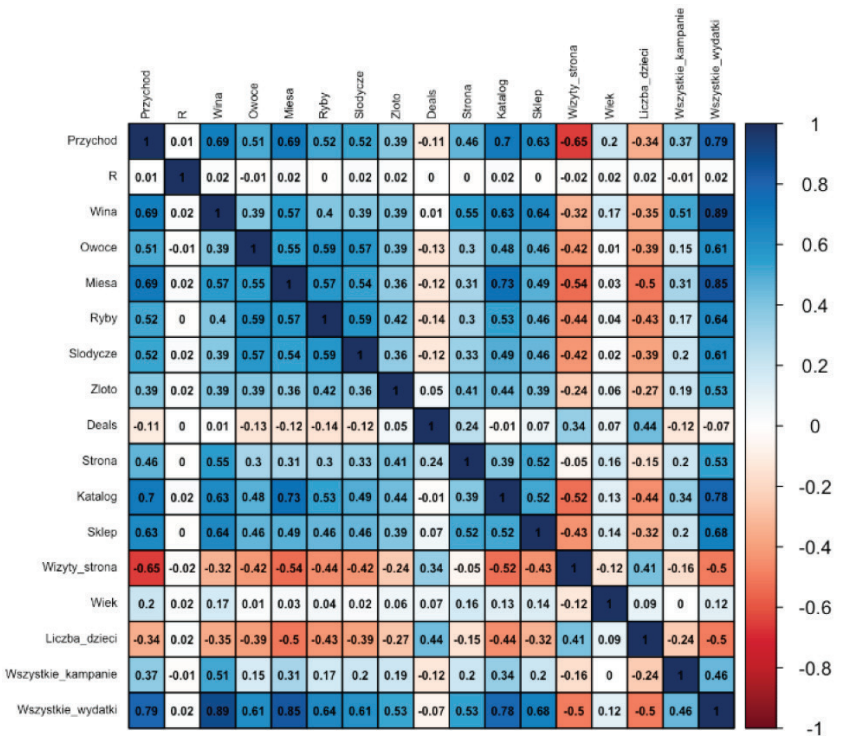
Następnie może zostać wyznaczona miara odległości między dwoma różnymi klastrami. Metodą wykorzystaną w rozdziale trzecim jest połączenie metody Warda, dla którego odległość między dwoma klastrami jest sumą kwadratów odchyleń od punktów do centroidów (reprezentujących geometryczne uściślenie środka obszaru) (Pacheco, 2015). Dendrogram więc przedstawia na wykresie końcowe połączenia klastrów dla obserwacji, czego efektem jest drzewo hierarchiczne.

Wykres osypiska, przedstawiony w rozdziale trzecim, pomimo bycia kolejnym sposobem wyboru liczby klas należących do grupowania hierarchicznego, opiera się na metodzie *k*-średnich, która należy do grupowania iteracyjno-optymalizacyjnego.

Wykres osypiska wizualizuje porównanie wariancji wewnątrzgrupowych wyliczonych za jej pomocą. Termin *k*-średnie został po raz pierwszy użyty przez Jamesa McQueena w 1967 roku, choć sam pomysł należy do Hugo Steinhausa (Steinhaus i Zubrzycki, 1957). Litera *k* w nazwie metody określa założoną z góry liczbę klastrów. Aby w poprawny sposób zastosować ten algorytm, potrzebne jest kolejno: ustalenie liczby skupień, ustalenie wstępnych środków skupień, obliczenie odległości obiektów od środków skupień, przypisanie obiektów do skupień, ustalenie nowego środka skupień, rekurencyjne powtarzanie kroków 3-5 do momentu wykonania liczby zadanych iteracji (Hartigan i Wong, 1979). Ustalone wstępnie środki skupień, czyli ich centroidy, decydują o końcowym przypisaniu do danego klastra.

4. Rezultaty badań

Duża część zmiennych to zmienne numeryczne mierzalne, dlatego zdecydowano się również na obliczenie korelacji dla wszystkich z nich. W ten sposób sprawdzone zostanie istnienie możliwych niespodziewanych zależności. Wyniki zostały przedstawione w formie macierzy korelacji na rysunku 2.



Rysunek 2. Macierz korelacji zmiennych numerycznych

Źródło: opracowanie własne z wykorzystaniem funkcji corrplot w programie R.

Kolor czerwony dobrany został do korelacji ujemnej, natomiast niebieski do dodatniej. Na pierwszy rzut oka widać, że wszystkie wydatki konsumenta są silnie skorelowane z wieloma zmiennymi, takimi jak wydatki na wina lub mięso, przychody, a także wizyty na stronie internetowej. Zmienna *R*, która określa liczbę dni od ostatniego zakupu klienta (*Recency*), nie ma praktycznie żadnej korelacji z pozostałymi zmiennymi. Na podstawie macierzy korelacji można utworzyć trzy profile klienta bazujące na wzajemnych zależnościach.

Konsumenci o wysokich przychodach dla gospodarstwa domowego:

- wydają więcej,
- mają tendencję do odwiedzania strony internetowej firmy zdecydowanie rzadziej niż inni klienci,
- często korzystają ze zniżek podczas robienia zakupów.

Osoby mające dzieci w domu:

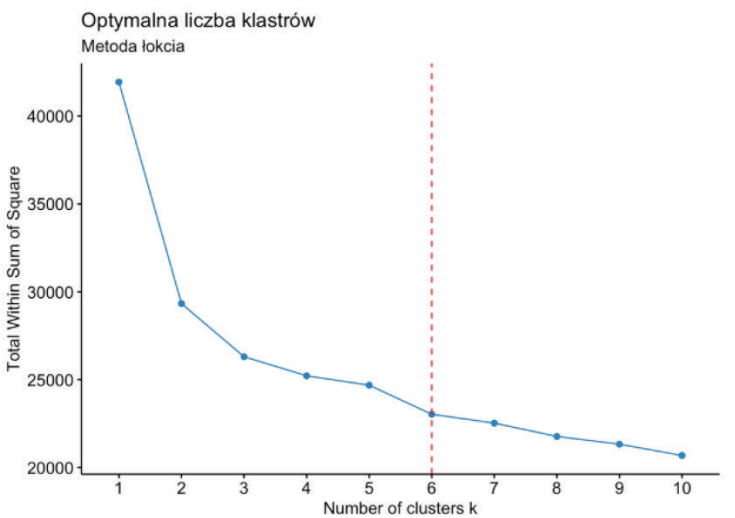
- wydają mniej,
- mają tendencję do dużej liczby zakupów z rabatem,

Klienci często korzystający z kampanii:

- kupują więcej win i produktów mięsnych,
- nie odwiedzają strony internetowej firmy,
- rzadko mają dzieci.

4.1. Segmentacja metodą *k*-średnich

Pierwszym sposobem, który wybrano, aby dokonać segmentacji klientów jest metoda *k*-średnich. Na danych przeprowadzono standaryzację, korzystając z funkcji orderNorm z pakietu bestNormalize. Na rysunku 3 porównano wariancje wewnątrzgrupowe

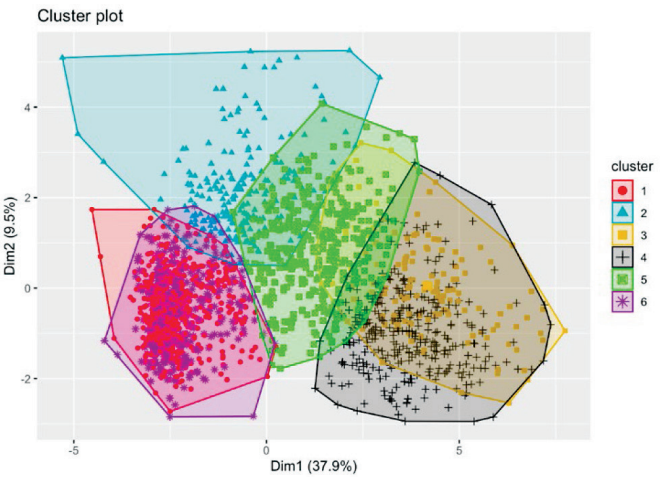


Rysunek 3. Wykres osypiska

Źródło: opracowanie własne z wykorzystaniem funkcji fviz_nbclust w programie R.

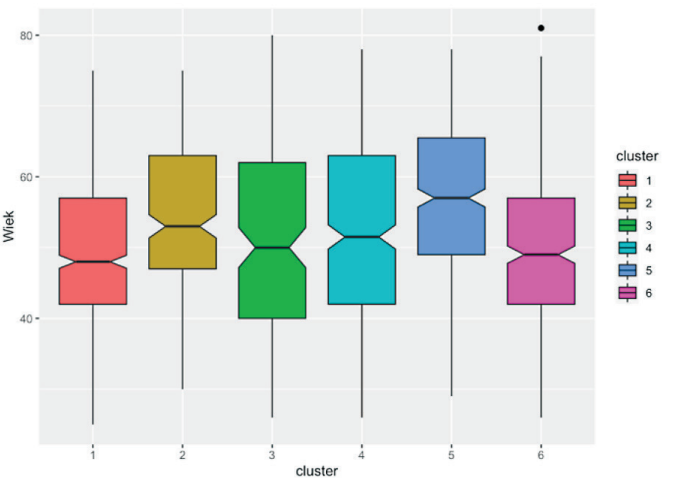
wyliczone za pomocą metody *k*-średnich dla różnej liczby grup. Wybrana została liczba klastrów, po której nie następuje już stromy spadek wartości wariancji, czyli sześć.

Na podstawie wyżej przeprowadzonych analiz, za pomocą funkcji kmeans, przeprowadzone zostało grupowanie na sześć klastrów. Zawierają one odpowiednio: 659, 234, 149, 378, 415 oraz 373 wartości. Po przeprowadzeniu tego działania, dzięki funkcji fviz_nbclust z pakietu factoextra, zwizualizowany został podział grup na rysunku 4. Przestawiony został na wykresie dwuwymiarowym.



Rysunek 4. Wykres podziału klastrów

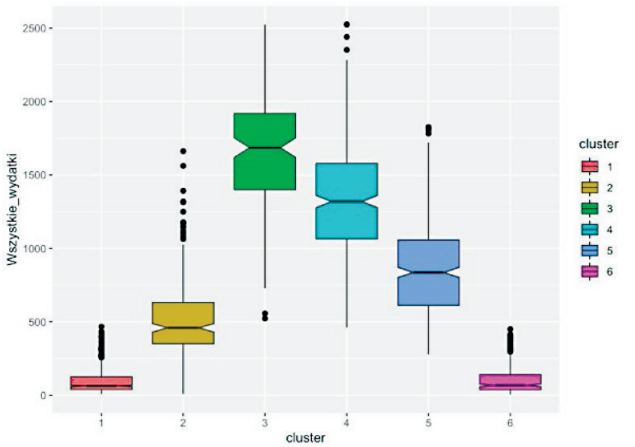
Źródło: opracowanie własne z wykorzystaniem funkcji fviz_nbclust w programie R.



Rysunek 5. Rozkład wieku dla klastrów

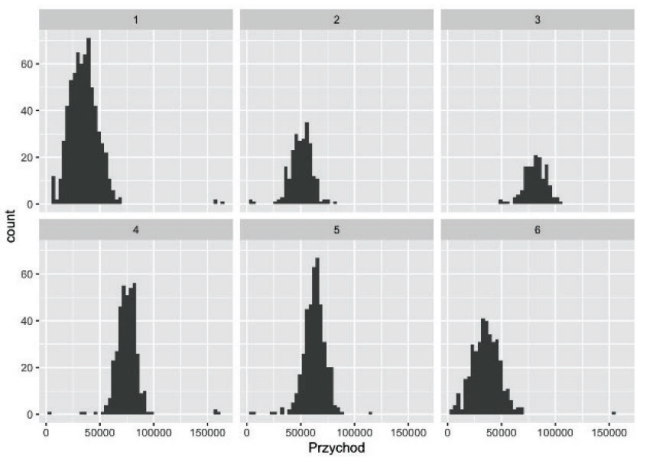
Źródło: opracowanie własne z wykorzystaniem funkcji ggplot i geom_boxplot w programie R.

Aby lepiej zwizualizować charakterystyki danych grup przygotowane zostały wykresy przedstawiające rozkład wartości w klastrach dla wybranych zmiennych: wiek, wszystkie wydatki, przychód i edukacja. Zaprezentowane zostały na rysunkach 5-8.



Rysunek 6. Rozkład wszystkich wydatków dla klastrów

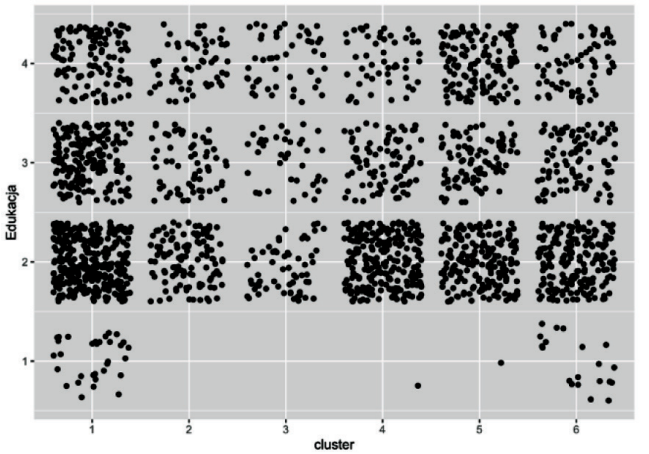
Źródło: opracowanie własne z wykorzystaniem funkcji ggplot i geom_boxplot w programie R.



Rysunek 7. Rozkład przychodów dla klastrów

Źródło: opracowanie własne z wykorzystaniem funkcji ggplot i geom_histogram w programie R.

Na podstawie wyżej zamieszczonych rozkładów i średnich wyróżnić można istotne charakterystyki dla poszczególnych grup: klaster pierwszy – młody, biedny klient z niskim poziomem wykształcenia, klaster drugi – średni wiek, średnie zarobki, niskiej



Rysunek 8. Wizualizacja rozkładu klientów pod względem poziomu edukacji dla wszystkich klastrów

Źródło: opracowanie własne z wykorzystaniem funkcji ggplot i geom_jitter w programie R.

wartości wydatki, klaster trzeci – zróżnicowany wiek, wysokie zarobki, bardzo duże wydatki, klaster czwarty – zróżnicowany wiek, średnie zarobki, za to bardzo duże wydatki, klaster piąty – starszy klient, średnie zarobki, niskiej wartości wydatki, duża liczba doktorów, klaster szósty – młody konsument, niskie zarobki, niskiej wartości wydatki, wykształcenie raczej na niskim poziomie.

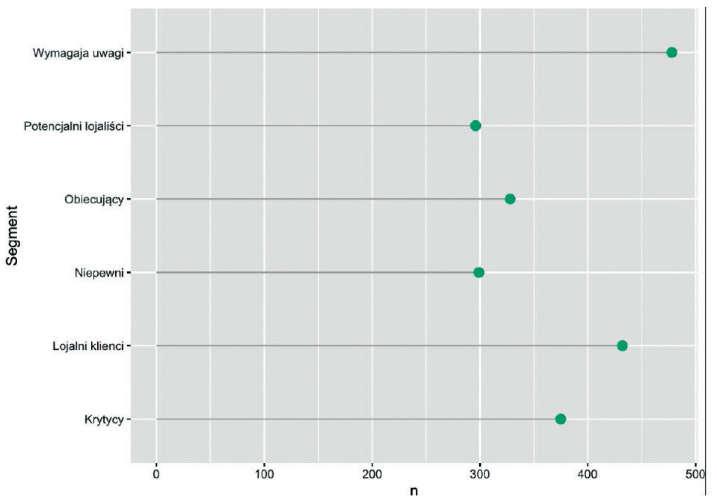
4.2. Segmentacja metodą RFM

Drugą przeprowadzoną metodą segmentacji jest metoda RFM. Polega ona na przydzieleniu każdego z konsumentów ze względu na jego preferencje i zachowania w ramach trzech wskaźników. Pierwszy z nich nosi nazwę *Recency* i informuje o tym, jak dawno przeprowadzona została ostatnia transakcja z klientem. Zmienna dla tego wskaźnika znajduje się w bazie pod nazwą *R*. Drugim jest *Frequency*, obrazujący, jak często klient korzysta ze sklepu. W tym celu utworzona została zmienna *F* przez dodanie do siebie wartości zmiennych, określających liczbę zakupów dokonanych za pośrednictwem strony internetowej firmy, przez katalog oraz bezpośrednio w sklepie. Ostatni wskaźnik nosi nazwę *Monetary* i przedstawia, jaką kwotę konsument wydaje w sklepie. Zmienna ta znajduje się już w bazie danych pod nazwą *Wszystkie_wydatki*. Klienci w poszczególnych grupach są kategoryzowani i otrzymują wartości od 1 do 4, z czego 1 jest wartością najlepszą. *Recency* jest jedyną destymulantą, ponieważ im krótszy czas minął od ostatniego zakupu tym lepiej. Przydzielenie odpowiednich wartości dokonywane jest na podstawie kwartyli każdej ze zmiennych.

Na podstawie wyników podziału rozdzielono klientów według następującego wzoru:

- Lojalni klienci: to najlepsi klienci, kupujący bardzo często, bez dużych odstępów czasowych i na duże kwoty.

- Potencjalni lojaliści: kupują często, ostatnie zakupy względnie niedawno, średnie kwoty.
- Obiecujący: ostatni zakup odbył się dawno, kupują często, warto im o sobie przypomnieć.
- Niepewni: zakupu dokonano niedawno, zakupy są dokonywane rzadko.
- Wymaga uwagi: ostatni zakup przeprowadzony dawno, wcześniej zakupy odbywały się często i na kwoty średniej wartości.
- Krytycy: klienci, którzy kupowali rzadko i na małe kwoty, a dodatkowo od dawna nie mają żadnej relacji z firmą.



Rysunek 9. Podział konsumentów na segmenty

Źródło: opracowanie własne z wykorzystaniem funkcji geom_point i geom_segment w programie R.

Na rysunku 9 przedstawiono wykres liczebności wszystkich klas. Najwięcej konsumentów zalicza się do grupy lojalnych klientów, to znaczy tych często powracających, oraz grupy wymagających uwagi. Jest to bardzo dobra informacja dla firmy przeprowadzającej badania, ponieważ przedstawia jasne sygnały, do kogo powinni kierować swoje reklamy i promocje. Klienci wymagający uwagi są związani ze sklepem, jednakże muszą przekonać się na nowo do kupowania produktów. Natomiast lojalnych konsumentów firma może nagrodzić swego rodzaju rabatami, tak by zachować z nimi dobre stosunki.

4.3. Podsumowanie

Obie przeprowadzone segmentacje niosą za sobą istotne i zróżnicowane wnioski. Dzięki zastosowaniu metody *k*-średnich baza danych została podzielona na klientów o różnych cechach. W oparciu o te dane firma może lepiej dobierać reklamy do

swoich odbiorców, dostosowując je przykładowo ze względu na wiek, liczbę dzieci czy liczbę wydawanych pieniędzy. Grupą najbardziej interesującą dla firmy może okazać się grupa trzecia, ponieważ konsumenci do niej należący zarówno dużo zarabiali, jak i dużo wydawali. Najmniej istotne klasy to pierwsza i szósta, ponieważ kupują zdecydowanie najmniej produktów, a ich zarobki są na bardzo niskim poziomie. Jednak metoda RFM przedstawiła klastry, które pozwolą na ukierunkowane kampanie marketingowe na podstawie zachowań klientów. Dwoma najważniejszymi klasami są lojalni klienci oraz ci wymagający uwagi. Firma może za to pominąć targetowanie reklam pod grupę krytyków. Dzięki tej analizie również utworzone zostały fundamenty dla przyszłych badań jakościowych, których celem jest dowiedzieć się więcej o motywacjach stojących za zachowaniem konsumentów oraz o ich potrzebach.

5. Zakończenie

Powyższa praca miała na celu wykorzystanie eksploracji danych w segmentacji klientów. We wstępnym zarysie dotyczącym zagadnień zarówno reklamy, jak i samej segmentacji zostało przedstawione, jak ważną rolę niesie właściwe dobranie grupy docelowej swoich ogłoszeń. Na podstawie przeprowadzonych studiów literaturowych w rozdziale drugim przedstawiono zagadnienia *data mining*, jak i algorytmów odpowiadających za metody statystyczne oraz metody zgłębiania danych. W rozdziale trzecim przeprowadzone zostały analizy na wybranym zbiorze obserwacji. Wykazały one interesujące wnioski, między innymi podział konsumentów na grupy rzeczywiście różniące się od siebie cechami bądź zachowaniem. Dzięki takiemu rozkładowi danych możliwe było nakreślenie grup, które mogły okazać się istotne dla firmy przeprowadzającej badania. Na przykład dzięki segmentacji metodą *k*-średnich odkryta została klasa konsumentów bogatych, cechujących się dużymi wydatkami. Jest to bardzo dobra grupa docelowa dla reklam sklepu. Jednak dwie wytworzone klasy łączące bardzo biednych klientów, o bardzo znikomych wydatkach, są mało interesujące z perspektywy prowadzenia kampanii marketingowych. Metoda RFM natomiast w prosty sposób podzieliła konsumentów ze względu na ich zachowanie względem firmy. Przedstawiła istotną dla przyszłego targetowania reklam klasę lojalnych klientów i tych potrzebujących uwagi. Nakreśliła również grupę krytyków marki, od dawna niemających nic wspólnego z jej usługami.

Podsumowując całe badanie, nie ma wątpliwości, że zastosowanie formalnych metod segmentacji klientów umożliwia wskazanie grup klienckich, na które należy zwrócić specjalną uwagę i do których należy kierować reklamę i promocję.

Literatura

Berman, B. (2016). Referral Marketing: Harnessing the Power of Your Customers. *Business Horizons*, *59*(1), 19-28.

Chapman, P., Clinton, J., Khabaza, T., Reinartz, T. i Wirth, R. (1999). *The CRISP-DM Process Model*. The CRIP–DM Consortium.

Dibb, S. i Simkin, L. (1991). Targeting, Segments and Positioning. *International Journal of Retail & Distribution Management, 19*(3), 4-10.

Dillmann, R. (2004). Teaching and Learning of Robot Tasks via Observation of Human Performance. *Robotics and Autonomous Systems*, *47*(2-3), 109-116.

Donovan, R. J. i Henley, N. (2003). *Social Marketing: Principles and Practice*. IP Communications.

Dziechciarz, J. (2003). *Ekonometria: metody, przykłady, zadania*. Wydawnictwo AE im. Oskara Langego we Wrocławiu.

Eckhardt, G. M. i Bengtsson, A. (2010). A Brief History of Branding in China. *Journal of Macromarketing*, *30*(3), 210-221.

Evans, L. (2011). *Social media marketing: Odkryj potencjał Facebooka, Twittera i innych portali społecznościowych*. Wydawnictwo Helion.

Fayyad, U. M., Piatetsky-Shapiro, G. i Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a Unifying Framework. W: *KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (s. 82-88). AAAI Press.

Fennis, B. M. i Stroebe, W. (2015). *The Psychology of Advertising*. Psychology Press.

Galton, F. (1889). I. Co-relations and Their Measurement, Chiefly from Anthropometric Data. *Proceedings of the Royal Society of London*, 45(273-279), 135-145.

Grzybczyk, K. (2012). *Prawo reklamy*. Wydawnictwo Wolters Kluwer.

Han, J., Kamber, M. I Pei, J. (2011). *Data Mining. Concepts and Techniques*. Elsevier.

Hartigan, J. A. i Wong, M. A. (1979). Algorithm AS 136. A k-means Clustering Algorithm. *Journal of the Royal Statistical Society*. Se*ries C. Applied Statistics, 28*(1), 100-108.

Jaska, E. i Werenowska, A. (2016). Promocja marki w mediach społecznościowych. *Handel Wewnętrzny*, *62*(2), 205-215.

Kalmane, M. R. (2010). Advertising: Using Words as Tools for Selling. Lulu.com

Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R. Unsupervised Machine Learning.* CreateSpace Independent Publishing Platform.

Kepios, (2022). *Digital 2022 Global Digital Overview*. Pobrane z https://datareportal.com/reports/digital-2022-global-overview-report

Lovell, M. C. (1983). Data Mining. *The Review of Economics and Statistics*, *65*(1), 1-12.

Maciorowski A. (2013), *E-marketing w praktyce. Strategie skutecznej promocji online*. Samo Sedno.

Pacheco, E. R. (2015). *Unsupervised Learning with R*. Packt Publishing.

Richter, F. (2019, 9 grudnia). *The Changing Face of the U.S. Advertising Landscape* [Digital image]. Pobrane z https://www.statista.com/chart/10269/us-advertising-revenue

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, *5*(4), 13-22.

Smith, W. R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of marketing*, *21*(1), 3-8.

Steinhaus, H. i Zubrzycki, S. (1957). On the Comparison of Two Production Processes and the Rule of Dualism. *Colloquium Mathematicum 5*(1), 103-115.

Szeliga, M. (2017). *Data science i uczenie maszynowe*. Wydawnictwo Naukowe PWN.

Troy, A. (2008). Geodemographic Segmentation. W: S. Shekhar i H. Xiong (red.), *Encyclopedia of GIS* (s. 347-355), Springer.

Tungate, M. (2007). *Adland: A Global History of Advertising*. Kogan Page Publishers.
Wind, Y. i Bell, D. R. (2008). Market Segmentation. W: M. J. Baker i S. Hart (red.), The Marketing Book (s. 222-244). Elsevier.
Yankelovich, D. i Meer, D. (2006). Rediscovering Market Segmentation. *Harvard Business Review*, *84*(2), 122-131.

Application of Selected Data Mining Methods in Customer Segmentation

Abstract: The aim of this Bachelor thesis is the application of data mining in customer segmentation. The paper firstly outlines the issues of segmentation as well as advertising, where it is used very often nowadays. Advertising is now a very important factor in building and promoting a brand. Then, after conducted literature studies on topics related to data mining algorithms, database-based analyzes were carried out. They divided consumers into groups that differed from each other in terms of their characteristics, as well as behaviour. These analyzes also outlined which customers cluster may be important for future marketing campaigns.

Keywords: data analysis, RFM, *k*-means, client segmentation, advertising, targeting, clusters