John W. Foreman

MISTRZ ANALIZY DANYCH

Od danych do wiedzy

„Dzięki lekturze książki *Mistrz analizy danych. Od danych do wiedzy* współczesne metody statystyczne i algorytmy stają się zrozumiałe i łatwe do zaimplementowania. Po jej przeczytaniu nie będziesz już musiał mozolnie przedzierać się przez inne podręczniki i opracowania!" — **Patrick Crosby**

Helion

Tytuł oryginału: Data Smart: Using Data Science to Transform Information into Insight
Tłumaczenie: Konrad Matuk
ISBN: 978-83-283-3357-4

Wydawnictwo HELION
ul. Kościuszki 1c, 44-100 GLIWICE
tel. 32 231 22 19, 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *http://helion.pl* (księgarnia internetowa, katalog książek)

Pliki z przykładami omawianymi w książce można znaleźć pod adresem:
ftp://ftp.helion.pl/przyklady/mianda.zip

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
http://helion.pl/user/opinie/mianda
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Printed in Poland.

Spis treści

2

Analiza skupień Część I — zastosowanie algorytmu centroidów do segmentowania bazy klientów

Pracuję w branży zajmującej się marketingiem za pomocą poczty elektronicznej dla serwisu MailChimp.com. Pomagamy naszym klientom w wysyłaniu newsletterów do ich odbiorców. Za każdym razem, gdy któryś z klientów używa określenia „uderzenie e-mailem", coś we mnie umiera.

Dlaczego? Ponieważ adresy e-mail nie są już czarnymi skrzynkami, do których można kierować zbiorcze „uderzenie". W e-mail marketingu (tak jak w przypadku innych form generowania zainteresowania za pomocą internetu, np. przy użyciu postów publikowanych w serwisach takich jak Twitter, Facebook czy Pinterest) firmy otrzymują informację zwrotną na temat tego, jak odbiorcy angażują się w ich kampanię na *poziomie indywidualnym* poprzez śledzenie kliknięć, zakupów elektronicznych czy udostępnień w serwisach społecznościowych. Te dane nie są szumem. One charakteryzują Twoich odbiorców, ale na początku mogą wydawać się zawiłe i niezrozumiałe.

Jak zebrać dane o transakcji od klientów (odbiorców, użytkowników, abonentów, mieszkańców itd.) i zastosować je w celu ich zrozumienia? Mając do czynienia z dużą grupą ludzi, trudno jest zrozumieć każdego klienta, zwłaszcza gdy każdy z nich komunikuje się z Tobą w inny sposób. Nawet jeżeli byś rozumiał każdego klienta na poziomie osobistym, to i tak trudno byłoby to zastosować w praktyce.

Musisz skorzystać z bazy klientów i znaleźć złoty środek pomiędzy „uderzaniem" do każdego tak, jakby był bezosobową jednostką, a rozumieniem wszystkiego o każdym i tworzeniem zindywidualizowanych ofert. Pomoc w znalezieniu tego złotego środka

może stanowić analiza skupień, umożliwiająca podział klientów na grupy (**segmentację rynku**), do których można kierować wybrane oferty.

Analiza skupień polega na zebraniu pewnej liczby obiektów i podzieleniu ich na grupy składające się z obiektów podobnych do siebie. Przyglądając się różnym grupom — analizując ich podobieństwa i różnice pomiędzy nimi — możesz wyciągnąć sporo praktycznych informacji z danych, które na początku wyglądały dość bezpostaciowo. Informacje te mogą pomóc Ci w podjęciu lepszych i bardziej szczegółowych decyzji.

Taką analizę skupień określamy mianem **eksploracyjnej analizy danych** — ma ona na celu ustalenie trudnych do zauważenia zależności w dużym zbiorze danych. Ustalanie zależności pomiędzy klientami przydaje się w rozmaitych przedsięwzięciach: umożliwia polecanie filmów na podstawie zachowań innych osób z grupy o podobnym guście, pozwala na identyfikację miejsc, w których najczęściej dochodzi do przestępstw, a także sprawia, że łatwiej jest zaproponować środek finansowy gronu osób, które mogą w niego potencjalnie zainwestować.

Jednym z moich ulubionych zastosowań analizy skupień jest tworzenie algorytmów umożliwiających komputerowi rozpoznawanie podobnych do siebie obrazów. Algorytmy takie mogą się okazać przydatne w serwisach przeznaczonych do udostępniania zdjęć, takich jak np. Flickr. Ich użytkownicy mogą umieszczać tak dużo zdjęć, że standardowe nawigowanie pomiędzy nimi staje się trudne. Techniki analizy skupień pozwalają na grupowanie obrazów podobnych do siebie. Dzięki nim użytkownicy mogą łatwiej znaleźć interesujący ich obraz.

NADZOROWANE I NIENADZOROWANE UCZENIE MASZYNOWE

Rozpoczynając eksploracyjną analizę danych, nie wiesz jeszcze, czego szukasz. Jesteś odkrywcą, jak Krzysztof Kolumb. Możesz powiedzieć, że jakichś dwóch klientów wygląda podobnie, ale nie wiesz, jak najlepiej podzielić bazę klientów. W związku z tym proces, w którym prosisz komputer o posegmentowanie bazy klientów, określamy mianem **nienadzorowanego uczenia maszynowego** — nie „nadzorujesz" komputera, informując go o tym, jak ma wykonać swoją pracę.

Przeciwieństwem tego procesu jest **nadzorowane uczenie maszynowe**, z którego korzystamy, gdy chcemy, aby sztuczna inteligencja wygenerowała pierwszą stronę naszego raportu. Jeżeli wiem, że chcę podzielić klientów na dwie grupy: tych, którzy prawdopodobnie coś kupią, i tych, którzy prawdopodobnie nie dokonają zakupu, i dostarczam komputerowi przykłady takich klientów z przeszłości, a następnie proszę go o podzielenie nowych klientów na takie dwie grupy, to praca komputera jest wtedy nadzorowana.

Nadzór polega na podzieleniu się z komputerem dotychczasową wiedzą na temat klientów i przedstawieniu mu sposobu pomiaru różnic i podobieństw zachodzących pomiędzy klientami.

W tym rozdziale chciałbym się przyjrzeć najpopularniejszej technice analizy skupień — **algorytmowi centroidów** (***k*-średnich**) opracowanemu w latach 50. XX w. Stał się on najpopularniejszym algorytmem analizy skupień służącym do *odkrywania wiedzy z baz danych* (ang. *knowledge discovery in databases* — KDD) stosowanym przez podmioty prywatne, a także agendy rządowe.

Z punktu widzenia matematyki algorytm centroidów nie jest najbardziej rygorystyczną techniką. Opiera się on na praktyczności i logice, które można dostrzec w kuchni *soul food*. Nie jest ona tak wyrafinowana jak kuchnia francuska, ale niektóre jej dania są przepyszne. Analiza skupień za pomocą algorytmu centroidów to mieszanina matematyki i opisowości. Jej ogromną zaletą jest intuicyjna prostota.

Aby zrozumieć działanie tego algorytmu, przeanalizujmy prosty przykład.

Dziewczyny tańczą z dziewczynami, a chłopcy drapią się po łokciach

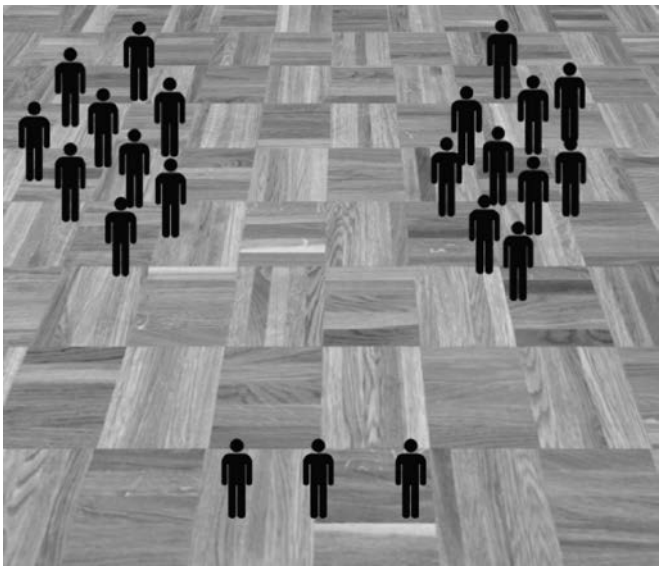
Celem analizy skupień za pomocą algorytmu centroidów jest wyznaczenie pewnych punktów w przestrzeni i podzielenie ich na k grup (k jest dowolnie wybraną liczbą). Każda z tych k grup jest definiowana przez punkt znajdujący się w jej środku. Punkt ten można porównać do flagi, na której ktoś wyhaftował napis: „Hej, to środek mojej grupy, przyłącz się do mnie, jeżeli jest ci bliżej do tej flagi niż do innych flag". Taki środek grupy (formalnie określany mianem **centroidu klasy**) jest średnią. W związku z tym algorytm centroidów często określany jest mianem algorytmu ***k*-średnich**.

Przyjrzyjmy się przykładowej potańcówce szkolnej. Bardzo przepraszam, jeżeli przypomniałem Ci o horrorze licealnych potańcówek.

Na rysunku 2.1 przedstawiono uczniów McAcne Middle School uczestniczących w szkolnej potańcówce typu „podmorska gala". Abyś mógł sobie lepiej wyobrazić tę sytuację, umieściłem też zdjęcie parkietu.

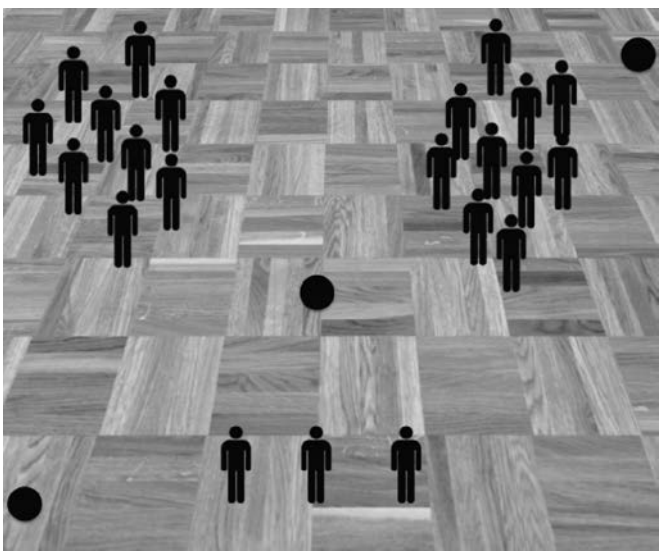
Oto przykładowa lista piosenek, przy których będą się bawić ci przyszli liderzy wolnego świata (możesz je znaleźć w serwisie Spotify):

- Styx — *Come Sail Away*.
- Everything But the Girl — *Missing*.
- Ace of Base — *All that She Wants*.
- Soft Cell — *Tainted Love*.
- Montell Jordan — *This is How We Do It*.
- Eiffel 65 — *Blue*.



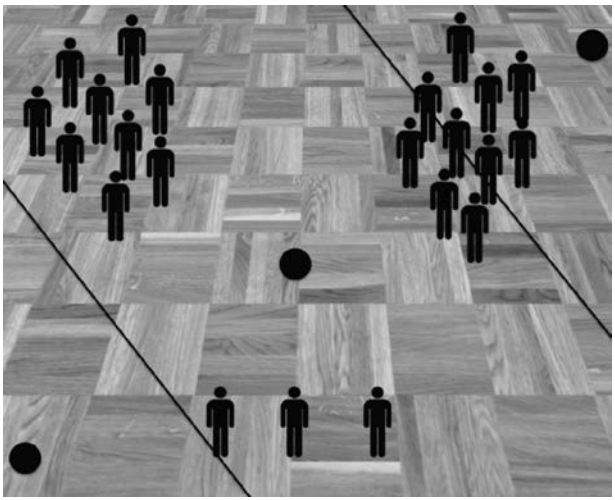
Rysunek 2.1. Uczniowie na parkiecie dyskoteki szkolnej

Algorytm centroidów wymaga określenia liczby skupień, na które chcemy posegregować uczestników potańcówki. Na początek spróbujmy podzielić ich na trzy skupienia (w dalszej części tego rozdziału dowiesz się, jak dobrać liczbę skupień — parametr *k*). Algorytm umieści na podłodze trzy flagi — zacznie pracę od jakiegoś początkowego, wykonywalnego rozwiązania. Na rysunku 2.2 przedstawiono takie rozwiązanie w postaci trzech czarnych kółek.



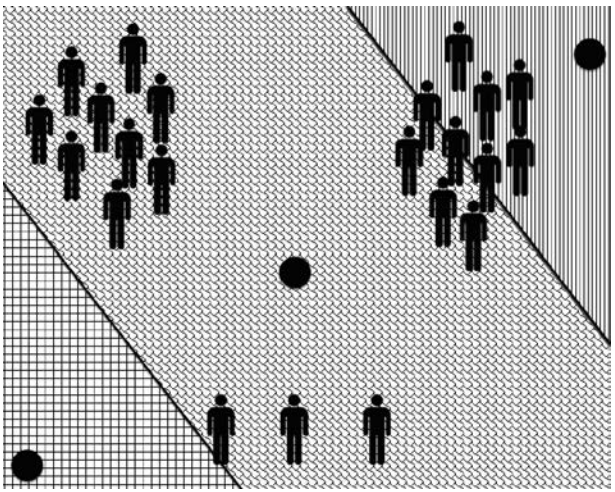
Rysunek 2.2. Początkowe ułożenie środkowych punktów skupień

Algorytm centroidów przypisuje tancerzy do najbliższego punktu, a więc pomiędzy dowolnie wybranymi skupieniami możliwe jest narysowanie linii oddzielających osoby przypisane do jednego skupienia od osób przyporządkowanych do innego skupienia (rysunek 2.3).



Rysunek 2.3. Linie symbolizujące granice skupień

Na podstawie tych trzech linii demarkacyjnych możesz przypisać tancerzy do poszczególnych grup i odpowiednio ich pocieniować (rysunek 2.4). Diagram, na którym przestrzeń podzielono na obszary przypisane do środków skupień, nazywamy **diagramem Woronoja**.



Rysunek 2.4. Podział na obszary skupień można przedstawić za pomocą różnych sposobów cieniowania na diagramie Woronoja

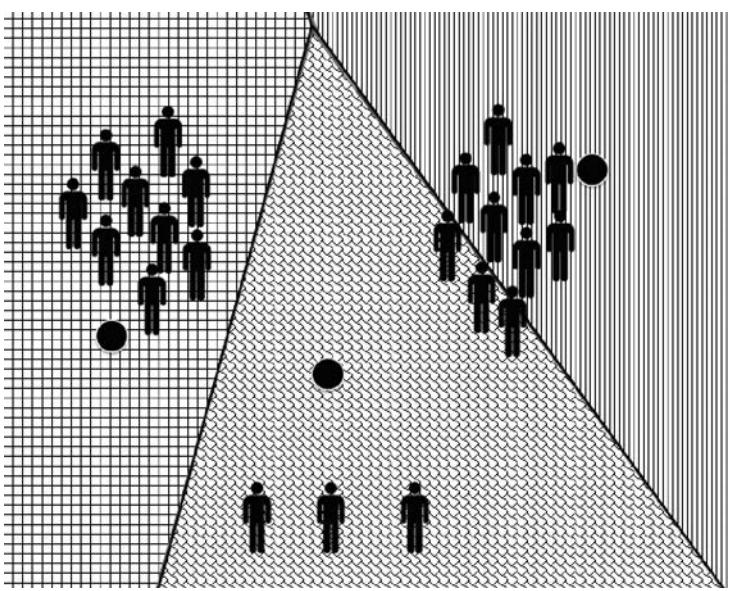
Ten początkowy podział nie jest najlepszy, prawda? Przestrzeń została podzielona w dziwny sposób — lewa dolna grupa jest pusta i bardzo dużo osób znalazło się na granicy grup środkowej i górnej.

Algorytm centroidów dokonujący analizy skupień będzie dzielił parkiet na trzy części, aż uzyska najlepszy podział.

Czym charakteryzuje się „najlepszy podział”? Każdy z uczestników imprezy znajduje się w jakiejś odległości od środka skupienia. Za najlepszy podział można uznać taki, przy którym średnia odległość uczestnika od przypisanego do niego środka skupienia jest najmniejsza.

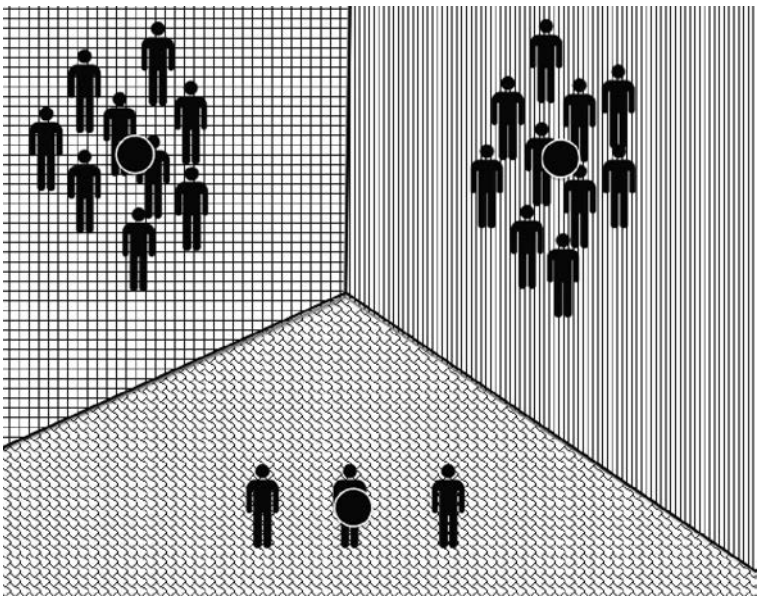
Zgodnie z tym, co napisałem w rozdziale 1., słowo „minimalizacja” oznacza konieczność zastosowania modelu optymalizacji w celu uzyskania optymalnego rozmieszczenia środków skupień, a więc w tym rozdziale będziesz korzystać z narzędzia Solver w celu przesuwania środków skupień. Solver znajduje właściwe pozycje środków skupień, przesuwając je w sposób uporządkowany i inteligentny — dokonując tych operacji, rejestruje liczbę dobrych przyporządkowań i znajduje najlepsze pozycje, łącząc dane uzyskane w wyniku tych operacji.

Diagram widoczny na rysunku 2.4 wygląda nieszczególnie, ale Solver mógłby przestawić środki skupień tak, aby uzyskać diagram pokazany na rysunku 2.5. Na tym rysunku zmniejszono nieco średnie odległości środków skupień od przyporządkowanych do nich tancerzy.



Rysunek 2.5. Środki skupień zostały odrobinę przesunięte

Na koniec Solver zapewne doszedłby do wniosku, że środki skupień powinny znajdować się w środku trzech grup uczestników imprezy (rysunek 2.6).



Rysunek 2.6. Optymalny podział uczestników dyskoteki na grupy

Super! Tak właśnie powinien wyglądać idealny efekt analizy skupień. Poszczególne środki skupień znajdują się w środku poszczególnych grup tancerzy, dzięki czemu zminimalizowano średnią odległość tancerza od środka skupienia, do którego jest przypisany. Po uzyskaniu podziału na grupy czas przejść do najciekawszego etapu pracy — spróbuj zrozumieć, co oznacza każde skupienie.

Gdybyś analizował kolor włosów tańczących, ich poglądy polityczne czy prędkość, z jaką potrafią biegać, to uzyskane przed chwilą skupienia nie miałyby sensu, ale gdybyś zwrócił uwagę na płeć i wiek uczestników przyporządkowanych do poszczególnych grup, zauważyłbyś pewne prawidłowości. Mała grupa znajdująca się w dolnej części parkietu to osoby starsze — prawdopodobnie są to opiekunowie osób bawiących się na dyskotece. Grupa po lewej stronie to sami młodzi mężczyźni, a grupa po prawej to same młode kobiety. Wszyscy boją się tańczyć z osobami o przeciwnej płci.

No dobrze. Algorytm centroidów umożliwił podział osób znajdujących się na parkiecie na grupy, a także zrozumienie przyczyn takiego podziału.

Być może myślisz sobie: „To głupie, rozwiązanie zaprezentowanego problemu znałem od początku". Masz rację. Tak było w tym przypadku. To tylko przykładowe zagadnienie, które można zrozumieć, patrząc na rysunek. Wszystko jest przedstawione w przestrzeni dwuwymiarowej, którą da się z łatwością ogarnąć wzrokiem.

A teraz wyobraź sobie, że prowadzisz sklep mający w ofercie kilka tysięcy produktów. Niektórzy z Twoich klientów kupili w ciągu kilku ostatnich lat jedną lub dwie rzeczy, a inni zakupili w tym czasie kilkadziesiąt artykułów. Co zrobić w sytuacji, w której klienci kupowali różne towary?

Jak rozplanować je na „sklepowym parkiecie"? Przestrzeń, w której musiałbyś dokonać takiej analizy, nie jest dwu- ani trójwymiarowa. Składa się ona z tysięcy wymiarów utworzonych przez produkty, które mogły być kupione lub nie przez każdego z klientów. Zapewne widzisz już problem związany z tym przykładem — nie da się go rozwiązać „na oko".

Prawdziwy problem: implementacja algorytmu centroidów w e-mail marketingu

Przejdźmy do jakiegoś bardziej praktycznego zastosowania algorytmu *k*-średnich. Zajmuję się e-mail marketingiem, a więc przedstawię Ci przykład zadania, z którym zmierzyłem się, pracując w MailChimp.com. Zaprezentowany przeze mnie przykład można zastosować w przypadku pracy z danymi ze sprzedaży detalicznej, danymi dotyczącymi akcji reklamowej, danymi wyciągniętymi z mediów społecznościowych itd. Można z niego skorzystać, analizując w zasadzie dowolny typ danych w celu dotarcia do klienta z odpowiednim materiałem marketingowym.

Handel winem

Wyobraź sobie, że mieszkasz w New Jersey i pracujesz w firmie Joey Bag O' Donuts Wholesale Wine Emporium zajmującej się importem win i dostarczaniem ich do wybranych sklepów na terytorium USA. Biznes opiera się na tym, że Joey Bag podróżuje po świecie, szukając niezwykle opłacalnych ofert sprzedaży dużych ilości wina. Joey dostarcza je do New Jersey, a Ty masz za zadanie sprzedać je sklepom, uzyskując jak największy dochód.

Docierasz do klientów na różne sposoby: korzystasz z serwisów Facebook i Twitter, a nawet wysyłasz oferty bezpośrednio za pośrednictwem poczty elektronicznej. Z ostatniego rozwiązania korzysta większość firm. W ubiegłym roku wysyłałeś do swoich klientów comiesięczny newsletter. Każda taka wiadomość przedstawia dwie lub trzy oferty związane ze sprzedażą wina — jedna z nich może dotyczyć np. szampana, a druga francuskiego wina malbec. Niektóre z ofert są naprawdę korzystne: umożliwiają osiągnięcie zysku ze sprzedaży na poziomie 80%. W sumie w tym roku zaproponowałeś 32 oferty. Wszystkie spotkały się z zainteresowaniem klientów.

To, że interes idzie dobrze, nie oznacza, że nie może iść lepiej. Warto byłoby poznać nieco bliżej swoich klientów. Oczywiście analizując dane dotyczące określonej transakcji, możesz się dowiedzieć, że osoba o nazwisku Adams kupiła w lipcu pewną ilość wina espumante ze zniżką 50%, ale nie wiesz, czy zakup został dokonany, ponieważ kupującemu

spodobało się to, że mógł kupić tylko jedno pudełko z sześcioma butelkami, czy może uznał, że cena jest atrakcyjna, czy też doszedł do wniosku, że cena tego produktu ma tendencję wzrostową.

Warto byłoby podzielić klientów na grupy skupiające podmioty dokonujące podobnych transakcji. Mógłbyś wtedy wysyłać do każdej grupy newsletter zoptymalizowany pod kątem danego segmentu rynku. Taki newsletter mógłby prezentować w pierwszej kolejności te oferty, które potencjalnie bardziej interesują daną grupę odbiorców, co mogłoby zwiększyć sprzedaż.

Jak podzielić listę klientów na segmenty? Od czego zacząć?

Podziału listy można dokonać za pomocą komputera. Analiza skupień umożliwia uzyskanie optymalnego podziału klientów na grupy, a wtedy będziesz mógł odkryć przyczynę takiego podziału i wybrać najlepsze grupy docelowe dla przygotowanych ofert.

Początkowy zbiór danych

UWAGA

W tym rozdziale będę korzystał ze skoroszytu programu Excel o nazwie *Wina.xlsx*. Możesz go pobrać ze strony: *ftp://ftp.helion.pl/przyklady/mianda.zip*. Plik zawiera dane, które będę przetwarzał w tym rozdziale (możesz na nich pracować podczas lektury), a także arkusze z przetworzonymi danymi prezentujące wyniki opisanych przeze mnie operacji (możesz je przeglądać bez konieczności samodzielnego wprowadzania formuł).

Zacznijmy od przyjrzenia się dwóm interesującym zbiorom danych:

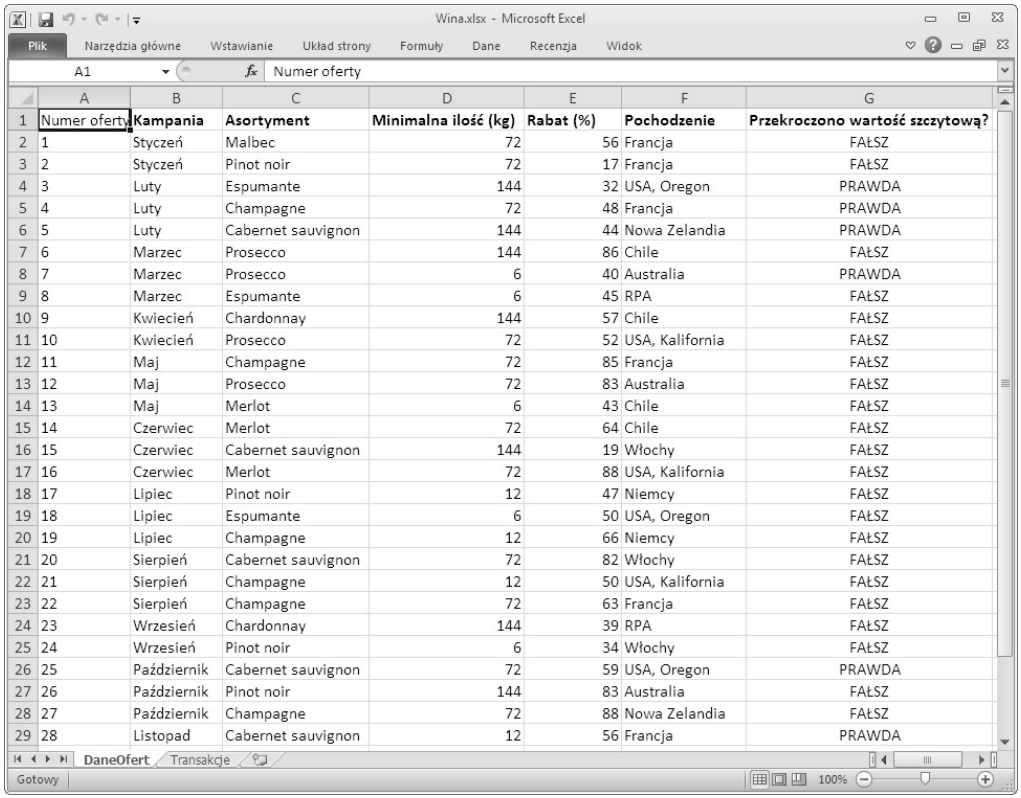
- Metadane każdej oferty zapisane w formie arkusza kalkulacyjnego. Zawierają one informacje określające rodzaj wina, minimalną liczbę butelek, jaką można kupić, wartość udzielonego rabatu, kraj pochodzenia oraz to, czy cena danego wina przekroczyła swoją wartość szczytową. Dane te zapisano w zakładce *DaneOfert* (rysunek 2.7).
- Jako pracownik wiesz, którzy klienci skorzystali z danej oferty, a więc możesz wpisać te dane do kolejnego arkusza. W zakładce *Transakcje* umieszczono nazwiska klientów wraz z informacją o tym, z której oferty skorzystali (rysunek 2.8).

Określanie tego, co chcemy mierzyć

Teraz musimy zmierzyć się z pewnym problemem. W kwestii dyskoteki szkolnej pomiar odległości pomiędzy osobami i grupami był prosty — wystarczyło rozwinąć metrówkę.

Co możemy zrobić w przypadku sprzedaży win?

Wiesz, że w zeszłym roku złożono 32 oferty, w zakładce *Transakcje* znajdują się dane 324 operacji zakupu podzielonych na klientów. W celu dokonania pomiaru odległości pomiędzy klientami i wyznaczenia środków grup musisz umieścić klientów w przestrzeni 32 transakcji. Innymi słowy: musisz określić transakcje, *których nie dokonali*, i stworzyć



Wina.xlsx - Microsoft Excel

Plik | Narzędzia główne | Wstawianie | Układ strony | Formuły | Dane | Recenzja | Widok

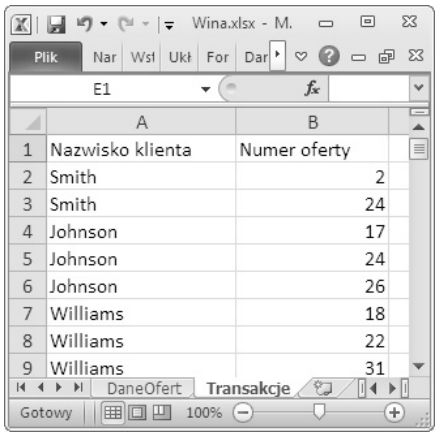
A1 | Numer oferty

	A	B	C	D	E	F	G
1	Numer oferty	Kampania	Asortyment	Minimalna ilość (kg)	Rabat (%)	Pochodzenie	Przekroczono wartość szczytową?
2	1	Styczeń	Malbec	72	56	Francja	FAŁSZ
3	2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ
4	3	Luty	Espumante	144	32	USA, Oregon	PRAWDA
5	4	Luty	Champagne	72	48	Francja	PRAWDA
6	5	Luty	Cabernet sauvignon	144	44	Nowa Zelandia	PRAWDA
7	6	Marzec	Prosecco	144	86	Chile	FAŁSZ
8	7	Marzec	Prosecco	6	40	Australia	PRAWDA
9	8	Marzec	Espumante	6	45	RPA	FAŁSZ
10	9	Kwiecień	Chardonnay	144	57	Chile	FAŁSZ
11	10	Kwiecień	Prosecco	72	52	USA, Kalifornia	FAŁSZ
12	11	Maj	Champagne	72	85	Francja	FAŁSZ
13	12	Maj	Prosecco	72	83	Australia	FAŁSZ
14	13	Maj	Merlot	6	43	Chile	FAŁSZ
15	14	Czerwiec	Merlot	72	64	Chile	FAŁSZ
16	15	Czerwiec	Cabernet sauvignon	144	19	Włochy	FAŁSZ
17	16	Czerwiec	Merlot	72	88	USA, Kalifornia	FAŁSZ
18	17	Lipiec	Pinot noir	12	47	Niemcy	FAŁSZ
19	18	Lipiec	Espumante	6	50	USA, Oregon	FAŁSZ
20	19	Lipiec	Champagne	12	66	Niemcy	FAŁSZ
21	20	Sierpień	Cabernet sauvignon	72	82	Włochy	FAŁSZ
22	21	Sierpień	Champagne	12	50	USA, Kalifornia	FAŁSZ
23	22	Sierpień	Champagne	72	63	Francja	FAŁSZ
24	23	Wrzesień	Chardonnay	144	39	RPA	FAŁSZ
25	24	Wrzesień	Pinot noir	6	34	Włochy	FAŁSZ
26	25	Październik	Cabernet sauvignon	72	59	USA, Oregon	PRAWDA
27	26	Październik	Pinot noir	144	83	Australia	FAŁSZ
28	27	Październik	Champagne	72	88	Nowa Zelandia	FAŁSZ
29	28	Listopad	Cabernet sauvignon	12	56	Francja	PRAWDA

DaneOfert | Transakcje

Gotowy | 100%

Rysunek 2.7. Szczegóły ostatnich 32 ofert



Wina.xlsx - M.

Plik | Nar | Wst | Ukł | For | Dar

E1

	A	B
1	Nazwisko klienta	Numer oferty
2	Smith	2
3	Smith	24
4	Johnson	17
5	Johnson	24
6	Johnson	26
7	Williams	18
8	Williams	22
9	Williams	31

DaneOfert | Transakcje

Gotowy | 100%

Rysunek 2.8. Lista ofert, z których skorzystali klienci

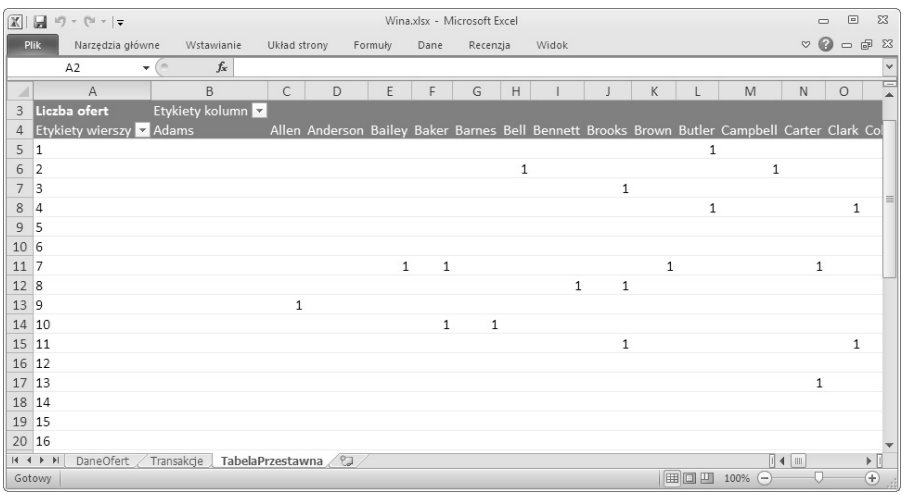
tabelę zawierającą transakcje przyporządkowane do danego klienta, w której do każdego klienta zostaną przypisane 32 kolumny wypełnione wartościami 1 (dokonano transakcji) lub 0 (nie dokonano transakcji).

Krótko mówiąc: musisz przenieść dane z zakładki *Transakcje* do tabeli klientów, w której każdej transakcji zostanie przydzielony oddzielny rząd komórek, a każdemu klientowi zostanie przydzielona oddzielna komórka. Operację taką najlepiej wykonać za pomocą tabeli przestawnej.

UWAGA

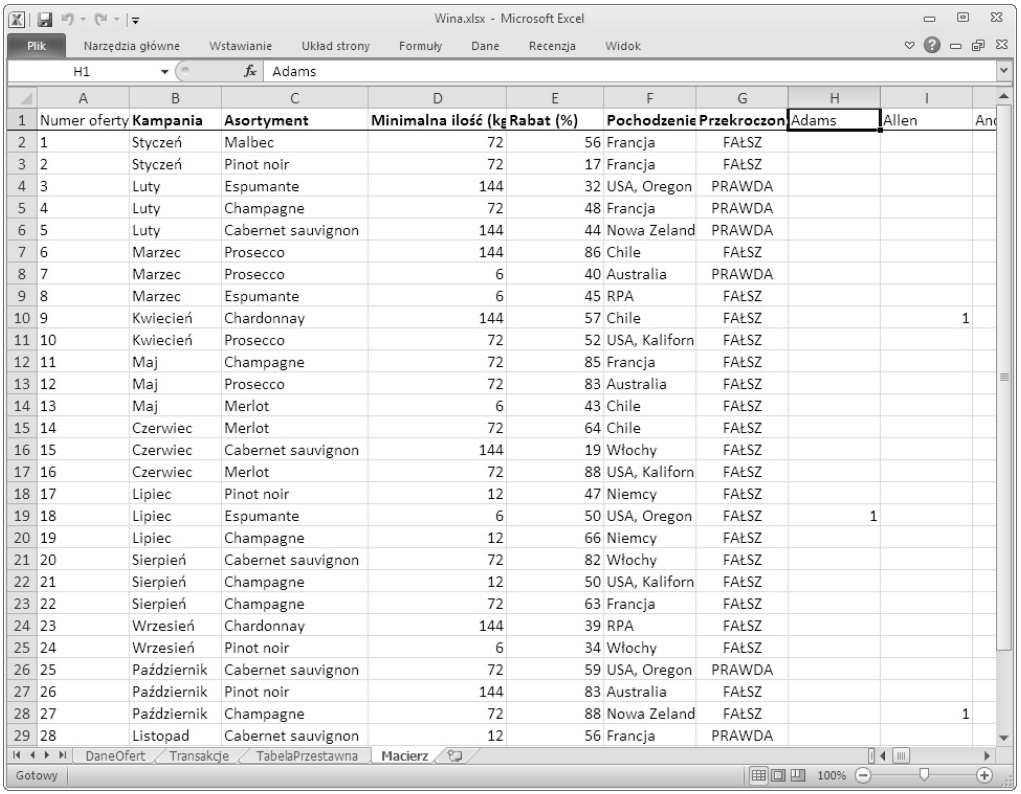
Podstawowe informacje dotyczące tabel przestawnych znajdziesz w rozdziale 1.

Oto czynności, które powinieneś wykonać. Zaznacz kolumny A i B znajdujące się w zakładce *Transakcje*, a następnie wstaw tabelę przestawną. W oknie *Lista pól tabeli przestawnej* jako etykiety wierszy wybierz kolumnę z ofertami, a jako etykiety kolumn wybierz kolumnę z danymi klientów. Następnie przypisz wartości do transakcji (w polu *Wartości* wybierz opcję *Liczba ofert*) — komórki wypełnione wartością *1* oznaczają skorzystanie przez danego klienta z wybranej oferty, a *0* lub (jak w tym przypadku) pusta komórka oznacza nieskorzystanie z oferty. Na rysunku 2.9 przedstawiono tabelę przestawną utworzoną przeze mnie.



Rysunek 2.9. Tabela przypisująca klientów do transakcji

Gdy masz już dane zakupów przedstawione w formie macierzy, skopiuj zawartość zakładki *DaneOfert* do nowej zakładki (nazwij ją *Macierz*). Do nowego arkusza skopiuj wartości znajdujące się w tabeli przestawnej (nie musisz kopiować numeru transakcji, ponieważ znajduje się on w danych ofert). Dane umieść w kolejnych kolumnach za danymi ofert (zacznij od kolumny H). W ten sposób utworzysz tabelę zawierającą informacje o ofertach oraz dane transakcji (rysunek 2.10).



Wina.xlsx - Microsoft Excel

Plik | Narzędzia główne | Wstawianie | Układ strony | Formuły | Dane | Recenzja | Widok

H1 | Adams

	A	B	C	D	E	F	G	H	I
1	Numer oferty	Kampania	Asortyment	Minimalna ilość (kg	Rabat (%)	Pochodzenie	Przekroczon	Adams	Allen
2	1	Styczeń	Malbec	72	56	Francja	FAŁSZ		
3	2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ		
4	3	Luty	Espumante	144	32	USA, Oregon	PRAWDA		
5	4	Luty	Champagne	72	48	Francja	PRAWDA		
6	5	Luty	Cabernet sauvignon	144	44	Nowa Zeland	PRAWDA		
7	6	Marzec	Prosecco	144	86	Chile	FAŁSZ		
8	7	Marzec	Prosecco	6	40	Australia	PRAWDA		
9	8	Marzec	Espumante	6	45	RPA	FAŁSZ		
10	9	Kwiecień	Chardonnay	144	57	Chile	FAŁSZ		1
11	10	Kwiecień	Prosecco	72	52	USA, Kaliforn	FAŁSZ		
12	11	Maj	Champagne	72	85	Francja	FAŁSZ		
13	12	Maj	Prosecco	72	83	Australia	FAŁSZ		
14	13	Maj	Merlot	6	43	Chile	FAŁSZ		
15	14	Czerwiec	Merlot	72	64	Chile	FAŁSZ		
16	15	Czerwiec	Cabernet sauvignon	144	19	Włochy	FAŁSZ		
17	16	Czerwiec	Merlot	72	88	USA, Kaliforn	FAŁSZ		
18	17	Lipiec	Pinot noir	12	47	Niemcy	FAŁSZ		
19	18	Lipiec	Espumante	6	50	USA, Oregon	FAŁSZ	1	
20	19	Lipiec	Champagne	12	66	Niemcy	FAŁSZ		
21	20	Sierpień	Cabernet sauvignon	72	82	Włochy	FAŁSZ		
22	21	Sierpień	Champagne	12	50	USA, Kaliforn	FAŁSZ		
23	22	Sierpień	Champagne	72	63	Francja	FAŁSZ		
24	23	Wrzesień	Chardonnay	144	39	RPA	FAŁSZ		
25	24	Wrzesień	Pinot noir	6	34	Włochy	FAŁSZ		
26	25	Październik	Cabernet sauvignon	72	59	USA, Oregon	PRAWDA		
27	26	Październik	Pinot noir	144	83	Australia	FAŁSZ		
28	27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ		1
29	28	Listopad	Cabernet sauvignon	12	56	Francja	PRAWDA		

DaneOfert | Transakcje | TabelaPrzestawna | Macierz

Gotowy | 100%

Rysunek 2.10. Tabela, która powstała w wyniku połączenia danych ofert oraz transakcji

STANDARYZACJA DANYCH

W tym rozdziale każdy wymiar danych jest typu binarnego, ale nie jest to reguła dotycząca każdego problemu rozwiązywanego za pomocą analizy skupień. Wyobraź sobie sytuację, w której ludzie są dzieleni na grupy na podstawie wzrostu, wagi i zarobków. Wzrost może wahać się od 150 do 200 centymetrów, a waga od 45 do 140 kilogramów.

W takim przypadku pomiar odległości pomiędzy klientami staje się o wiele bardziej skomplikowany od pomiaru odległości pomiędzy uczestnikami dyskoteki. W związku z tym dane są bardzo często *standaryzowane* — obliczana jest średnia danych znajdujących się w każdej kolumnie, a także wykonywana jest operacja dzielenia przez wartość charakteryzującą rozkład danych, czyli odchylenie standardowe (parametr ten opiszę w rozdziale 4.). Taki zabieg umożliwia przeskalowanie danych znajdujących się w każdej kolumnie tak, aby wartości oscylowały w granicach zera.

Dane, które przetwarzamy w rozdziale 2., nie wymagają standaryzacji. Praktyczne zastosowanie standaryzacji do wykrywania elementów odstających przedstawię w rozdziale 9.

Zacznij od czterech grup

Dysponujesz danymi skonsolidowanymi w formacie umożliwiającym ich dalsze przetwarzanie. Aby rozpocząć proces klastryzacji, musisz określić parametr *k* (liczbę klastrów algorytmu centroidów). Często, korzystając z tego algorytmu, testuje się jego działanie dla różnych wartości tego parametru (w dalszej części książki dowiesz się, jak je dobrać), ale na początek przyjmijmy tylko jedną jego wartość.

Najpierw musisz określić liczbę grup, na które chcesz podzielić swoich klientów, a to zależy od strategii marketingowej, jaką chcesz przyjąć. Możesz stworzyć 50 grup (klastrów), do których będziesz wysyłał 50 spersonalizowanych ofert (przeprowadzał 50 kampanii), ale takie rozwiązanie sprawi, że to ćwiczenie stanie się bezsensowne. Lepiej będzie podzielić klientów na względnie małą liczbę grup. Zacznij od podziału na cztery klastry — być może uzyskasz listę podzieloną na grupy składające się z 25 klientów, których preferencje da się łatwo zrozumieć (w rzeczywistości będzie to mało prawdopodobne).

No dobrze, jeżeli chcesz podzielić klientów na cztery grupy, to jakie grupy najlepiej byłoby uzyskać?

Zamiast zaśmiecać zakładkę *Macierz*, skopiuj znajdujące się w niej dane do nowej zakładki (nazwij ją `4MC`). W kolumnach od `H` do `K` wstaw cztery puste kolumny. W celu wstawienia nowej kolumny kliknij prawym przyciskiem myszy kolumnę `H`, a następnie wybierz opcję *Wstaw*. Spowoduje to wstawienie pustej kolumny po lewej. Po wstawieniu kolumn nadaj im etykiety od *Klaster 1.* do *Klaster 4.* Możesz je sformatować za pomocą opcji *Formatowanie warunkowe* — pozwoli Ci to obserwować zmiany wartości umieszczonych w komórkach tych kolumn podczas przesuwania środkowych punktów klastrów.

Zakładka *4MC* powinna teraz wyglądać tak, jak pokazano na rysunku 2.11.

Na razie środki wszystkich klastrów mają wartość 0, ale mogą przyjąć dowolną wartość. Twoim celem (podobnie jak w przykładzie dyskoteki szkolnej) będzie przesunięcie ich w położenia, przy których średnia odległość klientów przypisanych do danego klastra od jego środka będzie jak najmniejsza.

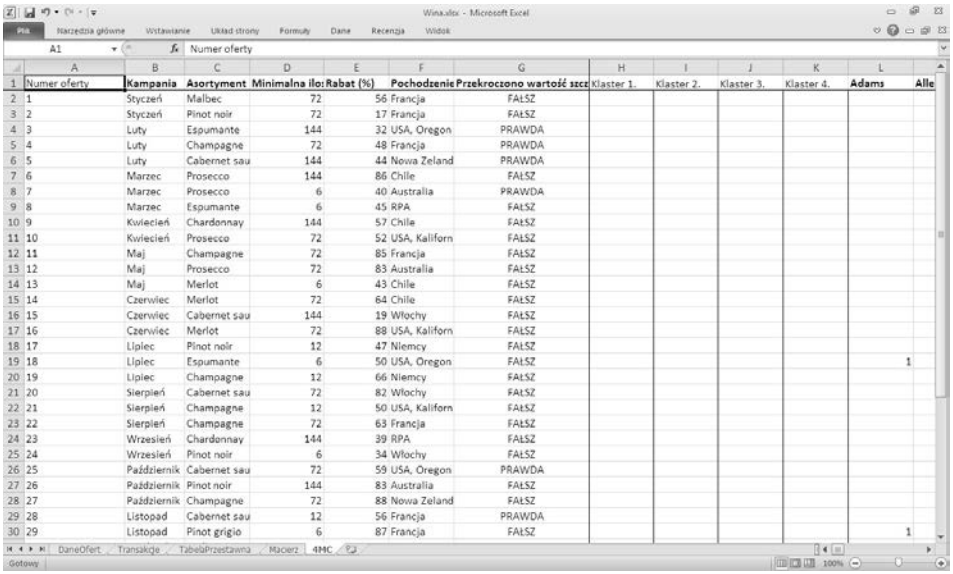
Oczywiście środki przyjmą dla każdej transakcji wartości znajdujące się w zakresie od 0 do 1, ponieważ wszystkie wektory klientów są binarne.

Co tak naprawdę oznacza pomiar odległości pomiędzy środkiem klastra a klientem?

Odległość euklidesowa — pomiar odległości w linii prostej

Teraz każdy klient jest opisany za pomocą jednej kolumny. Jak zmierzyć odległość pomiędzy nimi? Trzeba określić najkrótszą drogę, jaką trzeba pokonać, aby dojść z jednego punktu do drugiego (tzw. **odległość euklidesową**), a następnie ją zmierzyć.

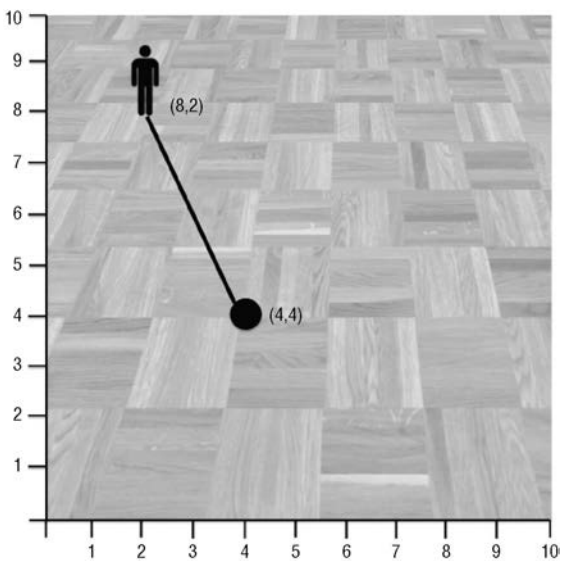
Aby zrozumieć sposób obliczania tej odległości, wróćmy na chwilę do przykładu potańcówki.



	A	B	C	D	E	F	G	H	I	J	K	L	
1	Numer oferty	Kampania	Asortyment	Minimalna ilo	Rabat (%)	Pochodzenie	Przekroczono wartość szcz	Klaster 1.	Klaster 2.	Klaster 3.	Klaster 4.	Adams	Alle
2	1	Styczeń	Malbec	72	56	Francja	FAŁSZ						
3	2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ						
4	3	Luty	Espumante	144	32	USA, Oregon	PRAWDA						
5	4	Luty	Champagne	72	48	Francja	PRAWDA						
6	5	Luty	Cabernet sau	144	44	Nowa Zeland	PRAWDA						
7	6	Marzec	Prosecco	144	86	Chile	FAŁSZ						
8	7	Marzec	Prosecco	6	40	Australia	PRAWDA						
9	8	Marzec	Espumante	6	45	RPA	FAŁSZ						
10	9	Kwiecień	Chardonnay	144	57	Chile	FAŁSZ						
11	10	Kwiecień	Prosecco	72	52	USA, Kaliforn	FAŁSZ						
12	11	Maj	Champagne	72	85	Francja	FAŁSZ						
13	12	Maj	Prosecco	72	83	Australia	FAŁSZ						
14	13	Maj	Merlot	6	43	Chile	FAŁSZ						
15	14	Czerwiec	Merlot	72	64	Chile	FAŁSZ						
16	15	Czerwiec	Cabernet sau	144	19	Włochy	FAŁSZ						
17	16	Czerwiec	Merlot	72	88	USA, Kaliforn	FAŁSZ						
18	17	Lipiec	Pinot noir	12	47	Niemcy	FAŁSZ						
19	18	Lipiec	Espumante	6	50	USA, Oregon	FAŁSZ					1	
20	19	Lipiec	Champagne	12	66	Niemcy	FAŁSZ						
21	20	Sierpień	Cabernet sau	72	82	Włochy	FAŁSZ						
22	21	Sierpień	Champagne	12	50	USA, Kaliforn	FAŁSZ						
23	22	Sierpień	Champagne	72	63	Francja	FAŁSZ						
24	23	Wrzesień	Chardonnay	144	39	RPA	FAŁSZ						
25	24	Wrzesień	Pinot noir	6	34	Włochy	FAŁSZ						
26	25	Październik	Cabernet sau	72	59	USA, Oregon	PRAWDA						
27	26	Październik	Pinot noir	144	83	Australia	FAŁSZ						
28	27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ						
29	28	Listopad	Cabernet sau	12	56	Francja	PRAWDA						
30	29	Listopad	Pinot grigio	6	87	Francja	FAŁSZ					1	

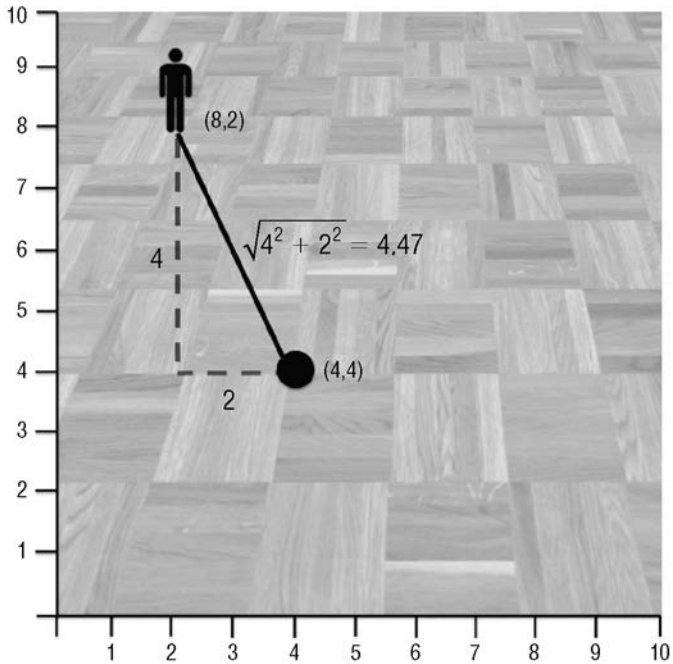
Rysunek 2.11. Zakładka 4MC z pustymi kolumnami definiującymi środki klastrów

Na rysunku 2.12 przedstawiłem parkiet dyskoteki opatrzony osiami układu współrzędnych. Jak widzisz, osoba tańcząca znajduje się w punkcie o współrzędnych (8,2), a środek klastra (grupy) leży w punkcie o współrzędnych (4,4). W celu obliczenia odległości euklidesowej pomiędzy tymi punktami należy przypomnieć sobie twierdzenie Pitagorasa, z którego często korzystałeś w gimnazjum i szkole średniej.



Rysunek 2.12. Osoba tańcząca znajduje się w punkcie o współrzędnych (8,2), a środek klastra jest położony w punkcie o współrzędnych (4,4)

Punkty te są od siebie oddalone o 8 – 4 = 4 metry w płaszczyźnie pionowej i 4 – 2 = 2 metry w płaszczyźnie poziomej. Zgodnie z twierdzeniem Pitagorasa, w celu obliczenia odległości w linii prostej wyznaczone wcześniej odległości należy podnieść do kwadratu i zsumować: $4^2 + 2^2 = 16 + 4 = 20$ metrów, a następnie obliczyć pierwiastek kwadratowy — pierwiastek z 20 to ok. 4,47 (rysunek 2.13).



Rysunek 2.13. Odległość euklidesowa równa jest pierwiastkowi kwadratowemu z sumy podniesionych do kwadratu odległości mierzonych w poszczególnych kierunkach

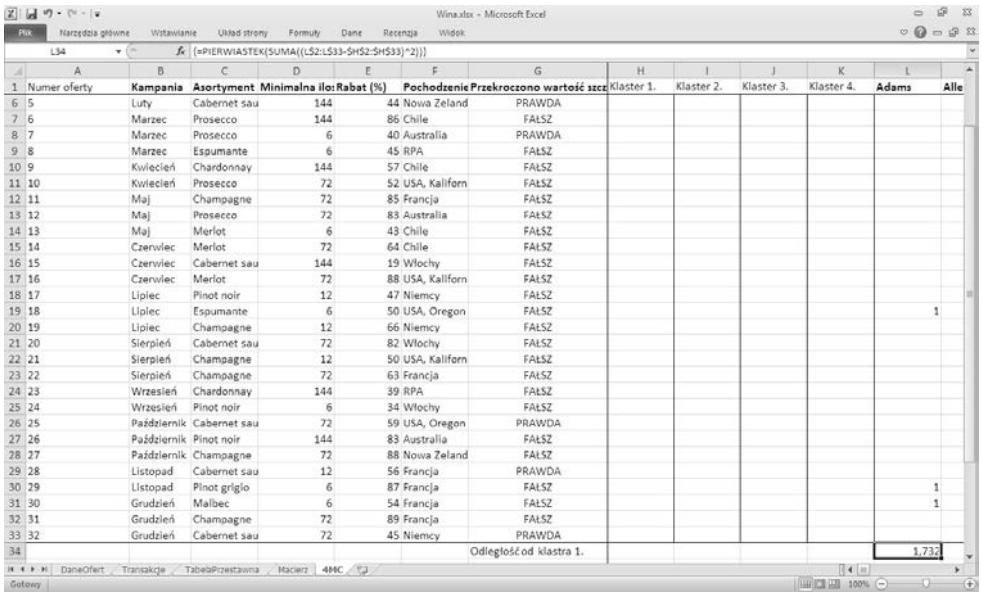
W przypadku osób zamawiających Twój newsletter masz do czynienia z więcej niż dwoma wymiarami, ale zasada pomiaru odległości w obu przypadkach jest taka sama: odległość pomiędzy klientem a środkiem klastra jest określana w wyniku zmierzenia odległości pomiędzy nimi w każdym wymiarze, podniesienia tych odległości do kwadratu, zsumowania, a następnie wyciągnięcia z nich pierwiastka kwadratowego.

Załóżmy, że chcesz obliczyć odległość euklidesową pomiędzy środkiem klastra 1. (kolumna `H`) a klientem o nazwisku *Adams* (kolumna `L`).

W komórce `L34` (pod zakupami wspomnianego klienta) możesz obliczyć różnicę wektora wybranego kupującego i środka klastra we wszystkich wymiarach, podnieść ją do kwadratu, zsumować, a następnie wyciągnąć pierwiastek z tej sumy za pomocą poniższej formuły tablicowej (zauważ, że zastosowano w niej odwołania bezwzględne, a więc możesz ją przenosić do innych pól, nie zmieniając zdefiniowanych odwołań):

```
{=PIERWIASTEK(SUMA((L$2:L$33-$H$2:$H$33)^2))}
```

Musisz skorzystać z formuły tablicowej (wprowadź kod formuły, a następnie wciśnij kombinację klawiszy *Ctrl+Shift+Enter* lub *Cmd+Return*, o czym pisałem w rozdziale 1.), ponieważ część `(L$2:L$33-$H$2:$H$33)^2` tej funkcji musi odczytywać pojedyncze wartości z kolejnych komórek i podnosić je do kwadratu. Formuła zwróci wartość *1,732* (rysunek 2.14) — osoba o nazwisku Adams dokonała trzech transakcji, początkowe współrzędne środków wszystkich klastrów to same zera, a pierwiastek kwadratowy z 3 to 1,732.



Rysunek 2.14. Odległość pomiędzy klientem o nazwisku Adams a środkiem pierwszego klastra

W arkuszu widocznym na rysunku 2.14 zablokowałem początkowe kolumny (kolumny od `A` do `G` są stale wyświetlane pomimo przewijania przez kolejne kolumny). Ponadto w komórce `G34` umieściłem etykietę *Odległość od klastra 1.* Zabiegi te ułatwiają śledzenie danych podczas przeglądania kolumn znajdujących się po prawej stronie arkusza.

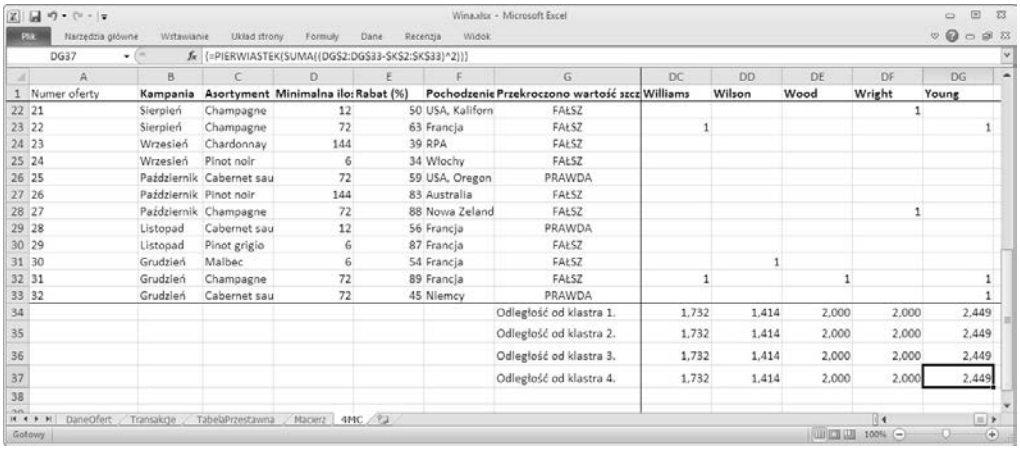
Odległość dla wszystkich!

Już wiesz, jak obliczyć odległość pomiędzy wektorem zakupu a środkiem klastra.

Czas obliczyć odległość pomiędzy klientem o nazwisku *Adams* a pozostałymi środkami. W tym celu przeciągnij zawartość komórki `L34` w dół aż do komórki `L37`, a następnie *ręcznie* zmień w kolejnych wierszach odwołania z kolumny `H` na kolumny `I`, `J` i `K`. W wyniku tych operacji w komórkach `L34:L37` znajdą się następujące formuły:

```
{=PIERWIASTEK(SUMA((L$2:L$33-$H$2:$H$33)^2))}
{=PIERWIASTEK(SUMA((L$2:L$33-$I$2:$I$33)^2))}
{=PIERWIASTEK(SUMA((L$2:L$33-$J$2:$J$33)^2))}
{=PIERWIASTEK(SUMA((L$2:L$33-$K$2:$K$33)^2))}
```

W powyższych formułach zastosowałeś odwołania bezwzględne (znak `$`; więcej informacji na ten temat znajdziesz w rozdziale 1.) do komórek definiujących środki klastrów, a więc możesz przeciągnąć zawartość komórek `L34:L37` do wszystkich komórek aż do `DG34:DG37`, co pozwoli obliczyć odległości pomiędzy każdym klientem a środkami wszystkich czterech klastrów. W kolumnie `G` (w rzędach od 35. do 37.) wprowadź etykiety `Odległość od klastra 2.` itd. Na rysunku 2.15 przedstawiono arkusz, w którym wprowadzono te etykiety.



	A	B	C	D	E	F	G	DC	DD	DE	DF	DG
1	Numer oferty	Kampania	Asortyment	Minimalna ilo:	Rabat (%)	Pochodzenie	Przekroczono wartość szcz	Williams	Wilson	Wood	Wright	Young
22	21	Sierpień	Champagne	12	50	USA, Kaliforn	FAŁSZ				1	
23	22	Sierpień	Champagne	72	63	Francja	FAŁSZ	1				1
24	23	Wrzesień	Chardonnay	144	39	RPA	FAŁSZ					
25	24	Wrzesień	Pinot noir	6	34	Włochy	FAŁSZ					
26	25	Październik	Cabernet sau	72	59	USA, Oregon	PRAWDA					
27	26	Październik	Pinot noir	144	83	Australia	FAŁSZ					
28	27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ				1	
29	28	Listopad	Cabernet sau	12	56	Francja	PRAWDA					
30	29	Listopad	Pinot grigio	6	87	Francja	FAŁSZ					
31	30	Grudzień	Malbec	6	54	Francja	FAŁSZ		1			
32	31	Grudzień	Champagne	72	89	Francja	FAŁSZ	1		1		1
33	32	Grudzień	Cabernet sau	72	45	Niemcy	PRAWDA					1
34							Odległość od klastra 1.	1,732	1,414	2,000	2,000	2,449
35							Odległość od klastra 2.	1,732	1,414	2,000	2,000	2,449
36							Odległość od klastra 3.	1,732	1,414	2,000	2,000	2,449
37							Odległość od klastra 4.	1,732	1,414	2,000	2,000	2,449
38												

Rysunek 2.15. Obliczanie odległości każdego klienta od wszystkich klastrów

Znasz odległość pomiędzy klientami a czterema klastrami. Każdy klient powinien być przypisany do najbliższego klastra. Operację tę możesz wykonać w dwóch krokach.

Wróć do klienta o nazwisku *Adams* (kolumna `L`) i oblicz minimalną odległość pomiędzy nim a środkiem klastra. W komórce `L38` wprowadź następującą formułę:

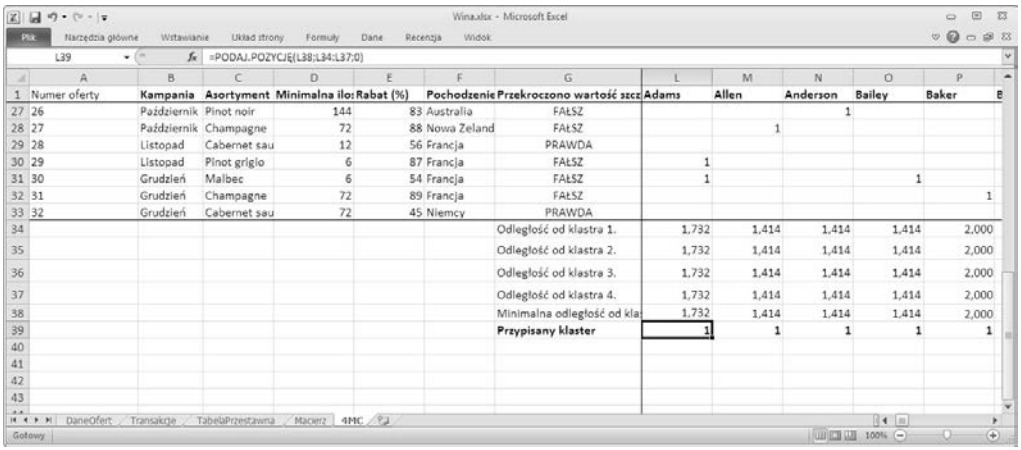
```
=MIN(L34:L37)
```

Teraz trzeba określić klaster, którego środek odpowiada tej minimalnej wartości. Możesz to zrobić za pomocą formuły `PODAJ.POZYCJĘ` (opisałem ją w rozdziale 1.). Umieszczając ją w komórce `L39`, możesz określić indeks komórki z zakresu od `L34` do `L37`, której zawartość pokrywa się z wartością minimalnej odległości:

```
=PODAJ.POZYCJĘ(L38,L34:L37,0)
```

W tym przypadku odległość pomiędzy klientem a wszystkimi czterema klastrami jest identyczna, a więc funkcja `PODAJ.POZYCJĘ` zwróci indeks pierwszej znalezionej wartości (rysunek 2.16).

Przeciągnij te dwie formuły do komórek znajdujących się po prawej stronie (aż do kolumny `DG`). Dodaj etykiety rzędów: `Minimalna odległość od klastra` i `Przypisany klaster`.



	A	B	C	D	E	F	G	L	M	N	O	P
1	Numer oferty	Kampania	Asortyment	Minimalna ilo:	Rabat (%)	Pochodzenie	Przekroczono wartość szcz	Adams	Allen	Anderson	Bailey	Baker
27	26	Październik	Pinot noir	144	83	Australia	FAŁSZ			1		
28	27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ		1			
29	28	Listopad	Cabernet sau	12	56	Francja	PRAWDA					
30	29	Listopad	Pinot griglo	6	87	Francja	FAŁSZ	1				
31	30	Grudzień	Malbec	6	54	Francja	FAŁSZ	1			1	
32	31	Grudzień	Champagne	72	89	Francja	FAŁSZ					1
33	32	Grudzień	Cabernet sau	72	45	Niemcy	PRAWDA					
34							Odległość od klastra 1.	1,732	1,414	1,414	1,414	2,000
35							Odległość od klastra 2.	1,732	1,414	1,414	1,414	2,000
36							Odległość od klastra 3.	1,732	1,414	1,414	1,414	2,000
37							Odległość od klastra 4.	1,732	1,414	1,414	1,414	2,000
38							Minimalna odległość od kla	1,732	1,414	1,414	1,414	2,000
39							**Przypisany klaster**	1	1	1	1	1

Rysunek 2.16. Arkusz zawierający formuły przypisujące indeks klastra do klientów

Określanie położenia środków klastrów

W arkuszu masz już formuły obliczające odległości i przypisujące najbliższy klaster. W celu określenia najlepszych położeń środków klastrów musisz określić wartości dla kolumn od `H` do `K` minimalizujące całkowitą odległość pomiędzy klientami i klastrami, do których są oni przypisani (39. rząd komórek znajdujący się pod danymi konsumentów).

Jeżeli przeczytałeś rozdział 1., to powinieneś wiedzieć, co należy zrobić, gdy słyszysz słowo „zminimalizować". Musisz przeprowadzić optymalizację, a to wiąże się z koniecznością uruchomienia narzędzia Solver.

Narzędzie to wymaga komórki celu, a więc w komórce `A36` zsumuj wszystkie odległości pomiędzy klientami a środkami:

```
=SUMA(L38:DG38)
```

W zaprezentowanym wcześniej przykładzie dyskoteki szkolnej również sumowaliśmy odległości za pomocą tej funkcji. Jednak odległość euklidesowa (wymagająca podnoszenia do kwadratu i obliczania pierwiastków) jest wysoce nieliniowa, dlatego określenie położenia środków klastrów wymaga zastosowania metody ewolucyjnej, a nie prostego algorytmu LP simpleks.

W rozdziale 1. korzystałeś z algorytmu LP simpleks. W sytuacjach, w których można go stosować, działa on o wiele szybciej od innych metod. Niestety, nie możesz z niego korzystać podczas podnoszenia do kwadratu i pierwiastkowania — wtedy, gdy podjęcie decyzji wymaga rozwiązania funkcji nieliniowej. Również zaawansowana wersja tego algorytmu dostępna w dodatku OpenSolver (opisałem go w rozdziale 1.) nie nadaje się do rozwiązania tego problemu.

W takiej sytuacji będziemy musieli użyć algorytmu ewolucyjnego wbudowanego w narzędzie Solver, który łączy wyniki losowego poszukiwania z „hodowaniem" dobrych rozwiązań (działa podobnie do ewolucji biologicznej).

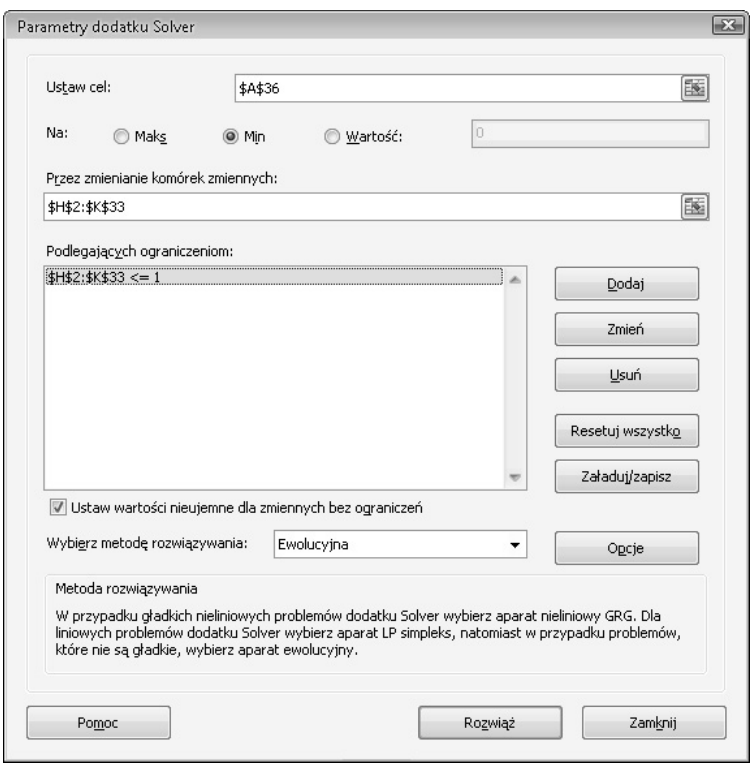
UWAGA

Zagadnienia związane z optymalizacją opisałem szerzej w rozdziale 4.

Zauważ, że masz już wszystkie elementy niezbędne do rozwiązania problemu za pomocą narzędzia Solver:

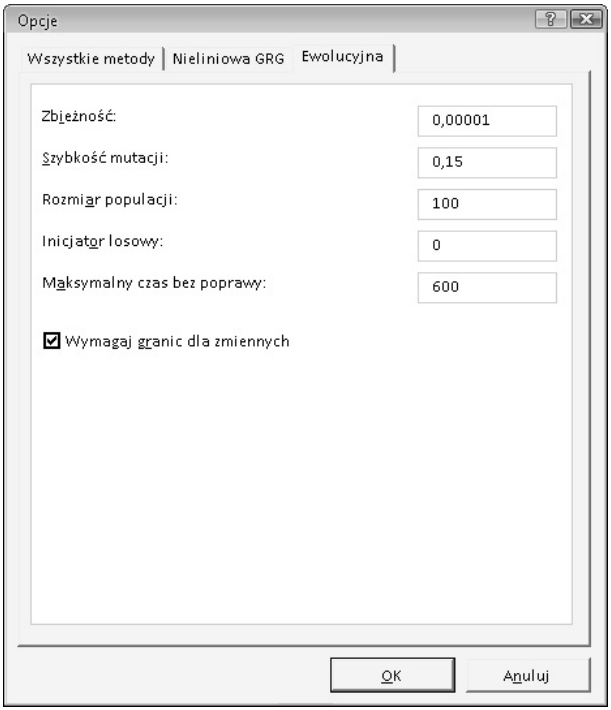
- **Cel** — zminimalizowanie sumy odległości klientów od środków klastrów, do których są przypisani (A36).
- **Zmienne decyzyjne** — modyfikowane wartości definiujące środki klastrów (H2:K33).
- **Ograniczenia** — środki klastrów powinny przyjmować wartości w zakresie od 0 do 1.

Otwórz okno narzędzia Solver i umieść w nim powyższe parametry. Skonfiguruj Solver tak, aby zminimalizował wartość zapisaną w komórce A36, modyfikując komórki H2:K33 i biorąc pod uwagę ograniczenie wartości wpisywanych do tych komórek (muszą być one, podobnie jak dane, mniejsze lub równe 1). Zaznacz opcję *Ustaw wartości nieujemne dla zmiennych bez ograniczeń* i wybierz metodę rozwiązania o nazwie *Ewolucyjna* (rysunek 2.17).



Rysunek 2.17. Konfiguracja Solvera do przeprowadzenia analizy skupień dzielącej klientów na cztery grupy

Zadanie to wcale nie jest dla Solvera takie proste, jak by się mogło wydawać, a więc zajrzyj do opcji metody ewolucyjnej (kliknij przycisk *Opcje* i przejdź do karty *Ewolucyjna*). Warto zmodyfikować parametr *Maksymalny czas bez poprawy* — wybierz czas ok. 30 sekund (zależnie od tego, ile czasu możesz poświęcić na czekanie na rozwiązanie problemu). Ustawiłem czas 600 sekund (10 minut) — rysunek 2.18. Dzięki temu po uruchomieniu Solvera będę mógł pójść na kawę. Jeżeli chciałbyś zatrzymać działanie Solvera, to wciśnij klawisz *Escape* i zamknij go, zachowując najlepsze znalezione dotychczas rozwiązanie.



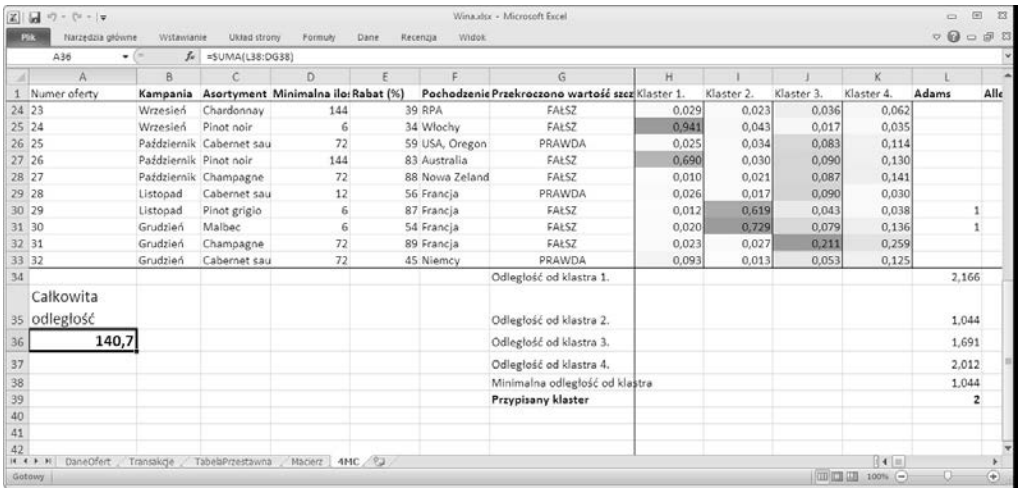
Rysunek 2.18. Zakładka opcji algorytmu ewolucyjnego

Jeżeli ciekawi Cię sposób działania algorytmu ewolucyjnego, to opis jego funkcjonowania znajdziesz w rozdziale 4. i na stronie: *http://www.solver.com/*.

Kliknij przycisk *Rozwiąż* i poczekaj na zakończenie pracy algorytmu ewolucyjnego.

Analiza uzyskanych wyników

Zabawa zacznie się dopiero po wygenerowaniu przez Solvera optymalnych klastrów. Przyjrzyj się danym, które uzyskałem (rysunek 2.19). Jak widzisz, Solver obliczył optymalną odległość całkowitą (*140,7*) i dzięki formatowaniu warunkowemu określił środki czterech klastrów.



	A	B	C	D	E	F	G	H	I	J	K	L	
1	Numer oferty	Kampania	Asortyment	Minimalna ilo	Rabat (%)	Pochodzenie	Przekroczono wartość szcz	Klaster 1.	Klaster 2.	Klaster 3.	Klaster 4.	Adams	Alle
24	23	Wrzesień	Chardonnay	144	39	RPA	FAŁSZ	0,029	0,023	0,036	0,062		
25	24	Wrzesień	Pinot noir	6	34	Włochy	FAŁSZ	0,941	0,043	0,017	0,035		
26	25	Październik	Cabernet sau	72	59	USA, Oregon	PRAWDA	0,025	0,034	0,083	0,114		
27	26	Październik	Pinot noir	144	83	Australia	FAŁSZ	0,690	0,030	0,090	0,130		
28	27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ	0,010	0,021	0,087	0,141		
29	28	Listopad	Cabernet sau	12	56	Francja	PRAWDA	0,026	0,017	0,090	0,030		
30	29	Listopad	Pinot grigio	6	87	Francja	FAŁSZ	0,012	0,619	0,043	0,038	1	
31	30	Grudzień	Malbec	6	54	Francja	FAŁSZ	0,020	0,729	0,079	0,136	1	
32	31	Grudzień	Champagne	72	89	Francja	FAŁSZ	0,023	0,027	0,211	0,259		
33	32	Grudzień	Cabernet sau	72	45	Niemcy	PRAWDA	0,093	0,013	0,053	0,125		
34							Odległość od klastra 1.					2,166	
35	Całkowita odległość						Odległość od klastra 2.					1,044	
36	140,7						Odległość od klastra 3.					1,691	
37							Odległość od klastra 4.					2,012	
38							Minimalna odległość od klastra					1,044	
39							**Przypisany klaster**					**2**	

Rysunek 2.19. Optymalne środki czterech klastrów

Twoja wersja Excela mogła wygenerować inne wartości środków klastrów. Mogło tak się stać, ponieważ algorytm ewolucyjny pracuje z wartościami losowymi i nie zawsze generuje takie same wyniki. Twoje klastry mogą być zupełnie inne, mogą nieco przypominać moje lub ich środki mogą być ułożone w różnej kolejności (środek mojego klastra 1. może znajdować się bardzo blisko środka Twojego klastra 4. itd.).

W kolumnach `B` – `G` umieściłeś informacje o ofertach, a więc teraz możesz z łatwością z nich korzystać, ponieważ dane te mogą być bardzo ważne z punktu widzenia środków klastrów (rysunek 2.19).

W przypadku klastra 1. znajdującego się w kolumnie `H` formatowanie warunkowe wyróżnia oferty 24., 26., 17. i w mniejszym stopniu ofertę 2. Jeżeli zagłębisz się w szczegóły tych ofert, to mają one jedną rzecz wspólną — wszystkie są związane z winem pinot noir.

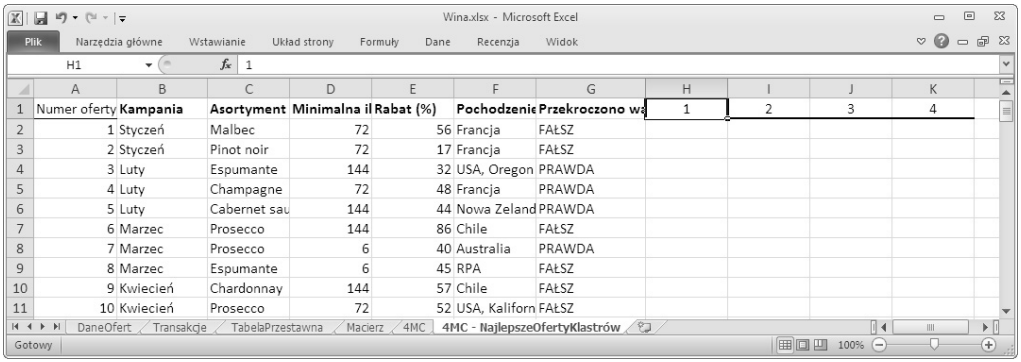
Jeżeli przyjrzysz się kolumnie `I`, to zauważysz w niej pola wyróżnione na zielono — są to oferty, które łączy niska minimalna ilość kupowanego wina. W klastrze tym znajdują się nabywcy, którzy nie chcą kupować dużych ilości trunku.

Będę szczery: interpretacja dwóch ostatnich klastrów jest trudna. Może zamiast interpretować położenie środka klastra, lepiej przyjrzeć się preferencjom przyporządkowanych do niego klientów? Być może w ten sposób wysnujesz jakieś sensowniejsze wnioski?

Ustalanie najlepszej oferty dla danego klastra

Zamiast szukać wymiarów, które przyjmują dla danego klastra wartości bliższe 1, sprawdźmy, kto jest przypisany do danego klastra i jakie oferty preferuje.

W tym celu skopiujmy zawartość zakładki *DaneOfert* do nowej zakładki (nazwij ją `4MC – NajlepszeOfertyKlastrów`). W nowej zakładce przypisz kolumnom `H` – `K` etykiety *1*, *2*, *3*, i *4* (rysunek 2.20).



	A	B	C	D	E	F	G	H	I	J	K
1	Numer oferty	Kampania	Asortyment	Minimalna il	Rabat (%)	Pochodzenie	Przekroczono wa	1	2	3	4
2	1	Styczeń	Malbec	72	56	Francja	FAŁSZ				
3	2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ				
4	3	Luty	Espumante	144	32	USA, Oregon	PRAWDA				
5	4	Luty	Champagne	72	48	Francja	PRAWDA				
6	5	Luty	Cabernet sau	144	44	Nowa Zeland	PRAWDA				
7	6	Marzec	Prosecco	144	86	Chile	FAŁSZ				
8	7	Marzec	Prosecco	6	40	Australia	PRAWDA				
9	8	Marzec	Espumante	6	45	RPA	FAŁSZ				
10	9	Kwiecień	Chardonnay	144	57	Chile	FAŁSZ				
11	10	Kwiecień	Prosecco	72	52	USA, Kaliforn	FAŁSZ				

Rysunek 2.20. Przygotowywanie tabeli zliczającej oferty najpopularniejsze w danym klastrze

W 39. wierszu zakładki *4MC* uzyskałeś listę klastrów przypisanych do poszczególnych klientów. W celu określenia liczby transakcji dotyczących danego klastra potrzebne jest sprawdzenie tytułu kolumn `H` – `K` zakładki *4MC — NajlepszeOfertyKlastrów*, a następnie użycie tej informacji w celu określenia, kto został przypisany do danego klastra w 39. wierszu zakładki *4MC*. Na koniec wystarczy zsumować wartości każdego rzędu transakcji. W ten sposób określisz liczbę klientów danego klastra, którzy skorzystali z danej oferty.

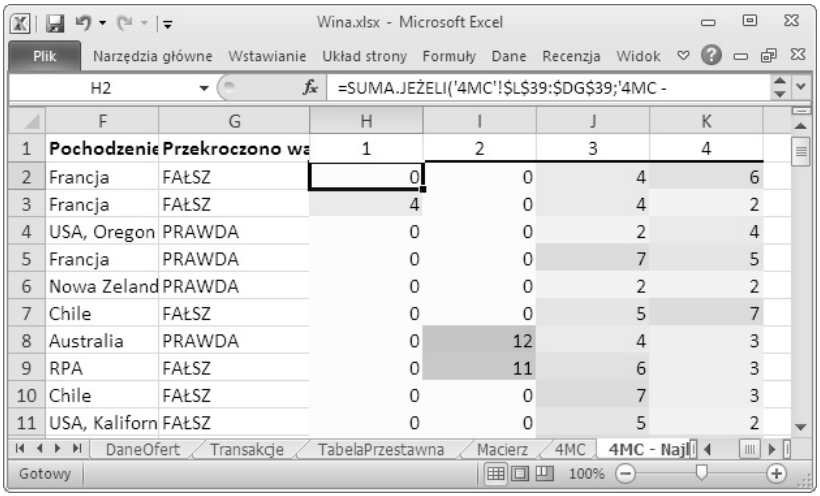
Zacznij od komórki `H2` — liczby klientów zaklasyfikowanych do klastra 1., którzy skorzystali z oferty 1. (styczniowa oferta na wino malbec). Chcesz zsumować wartości wpisane w komórkach `L2:DG2` zakładki *4MC*, ale musisz wziąć pod uwagę tylko klientów przypisanych do klastra 1. Jest to klasyczny przykład zastosowania formuły `SUMA.JEŻELI`. W komórce `H2` umieść następującą formułę:

```
=SUMA.JEŻELI('4MC'!$L$39:$DG$39,'4MC — NajlepszeOfertyKlastrów'!H$1,'4MC'!$L2:$DG2)
```

Formuła `SUMA.JEŻELI` wymaga zadeklarowania wartości (`'4MC'!$L$39:$DG$39`), które będą porównywane z wartością 1 zadeklarowaną w nagłówku kolumny (`'4MC — NajlepszeOfertyKlastrów'!H$1`). W przypadku znalezienia odpowiednich wartości dochodzi do sumowania wiersza 2. (zobacz trzeci element deklaracji formuły — `'4MC'!$L2:$DG2`).

Zauważ, że w formule zastosowałem odwołania bezwzględne — znak `$` umieściłem przed wszystkimi elementami sekcji przypisania wiersza klastra, przed numerem wiersza nagłówków kolumny i przed literą kolumny analizowanych transakcji. Dzięki zastosowaniu odwołań bezwzględnych formułę tę możesz przeciągnąć do komórek `H2:K33` w celu uzyskania liczby transakcji, do których doszło przy wszystkich kombinacjach transakcji i klastrów (rysunek 2.21). Aby zwiększyć czytelność danych zaprezentowanych w kolumnach `H` – `K`, uruchom formatowanie warunkowe.

Dane możesz posegregować, korzystając z opcji automatycznego filtrowania kolumn `A` – `K` (zob. rozdział 1.). Sortując dane znajdujące się w kolumnie `H` w kolejności od największych do najmniejszych, możesz zobaczyć, które oferty były najpopularniejsze wśród klientów przyporządkowanych do klastra 1. (rysunek 2.22).



Wina.xlsx - Microsoft Excel

Plik | Narzędzia główne | Wstawianie | Układ strony | Formuły | Dane | Recenzja | Widok

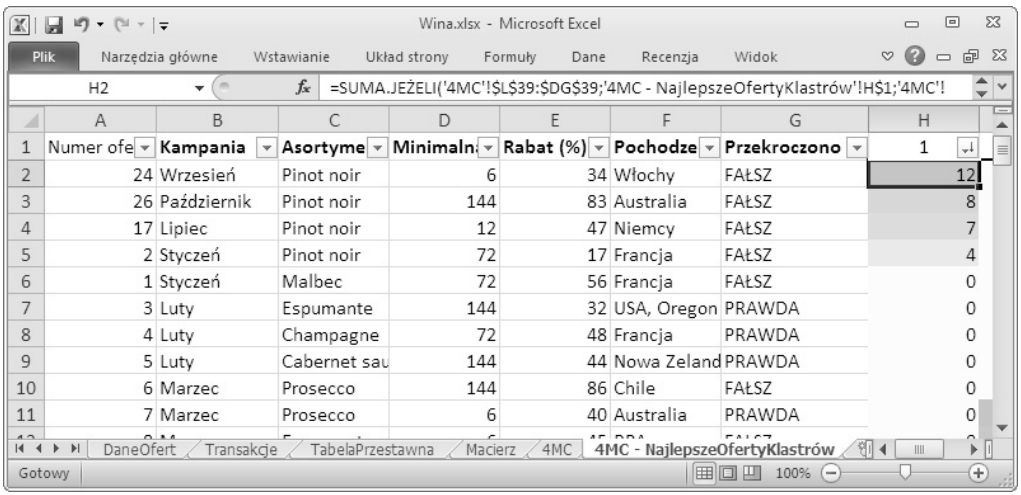
H2 =SUMA.JEŻELI('4MC'!$L$39:$DG$39;'4MC -

	F	G	H	I	J	K
1	Pochodzenie	Przekroczono wa	1	2	3	4
2	Francja	FAŁSZ	0	0	4	6
3	Francja	FAŁSZ	4	0	4	2
4	USA, Oregon	PRAWDA	0	0	2	4
5	Francja	PRAWDA	0	0	7	5
6	Nowa Zeland	PRAWDA	0	0	2	2
7	Chile	FAŁSZ	0	0	5	7
8	Australia	PRAWDA	0	12	4	3
9	RPA	FAŁSZ	0	11	6	3
10	Chile	FAŁSZ	0	0	7	3
11	USA, Kalifor	FAŁSZ	0	0	5	2

DaneOfert | Transakcje | TabelaPrzestawna | Macierz | 4MC | 4MC - Najl

Gotowy 100%

Rysunek 2.21. Sumy liczby transakcji podzielone na klastry



Wina.xlsx - Microsoft Excel

Plik | Narzędzia główne | Wstawianie | Układ strony | Formuły | Dane | Recenzja | Widok

H2 =SUMA.JEŻELI('4MC'!$L$39:$DG$39;'4MC - NajlepszeOfertyKlastrów'!H$1;'4MC'!

	A	B	C	D	E	F	G	H
1	Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	1
2	24	Wrzesień	Pinot noir	6	34	Włochy	FAŁSZ	12
3	26	Październik	Pinot noir	144	83	Australia	FAŁSZ	8
4	17	Lipiec	Pinot noir	12	47	Niemcy	FAŁSZ	7
5	2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ	4
6	1	Styczeń	Malbec	72	56	Francja	FAŁSZ	0
7	3	Luty	Espumante	144	32	USA, Oregon	PRAWDA	0
8	4	Luty	Champagne	72	48	Francja	PRAWDA	0
9	5	Luty	Cabernet sau	144	44	Nowa Zeland	PRAWDA	0
10	6	Marzec	Prosecco	144	86	Chile	FAŁSZ	0
11	7	Marzec	Prosecco	6	40	Australia	PRAWDA	0

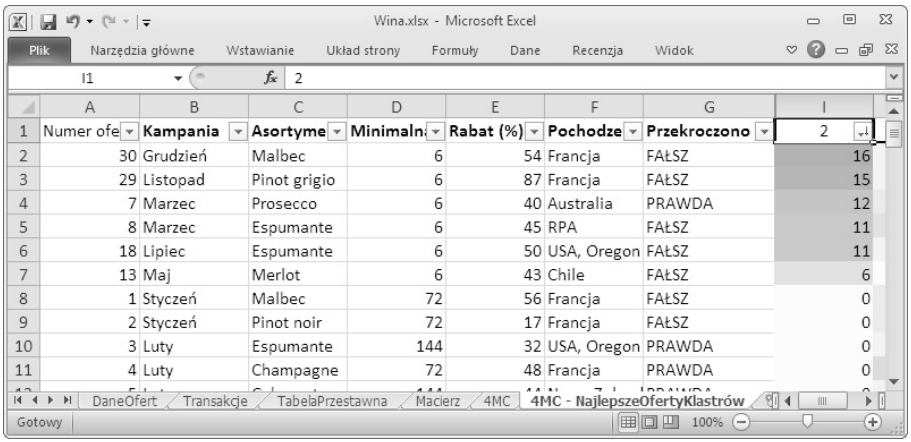
DaneOfert | Transakcje | TabelaPrzestawna | Macierz | 4MC | 4MC - NajlepszeOfertyKlastrów

Gotowy 100%

Rysunek 2.22. Sortowanie klastra 1. — pinot, pinot, pinot!

Zgodnie z tym, co zauważyłem wcześniej, wszystkie oferty w tym klastrze dotyczą wina pinot. Ci klienci chyba za dużo naoglądali się filmu *Bezdroża*. Po posortowaniu kolumny klastra 2. okazuje się, że znajdują się w nim klienci, którzy preferują zakup małej ilości wina (rysunek 2.23).

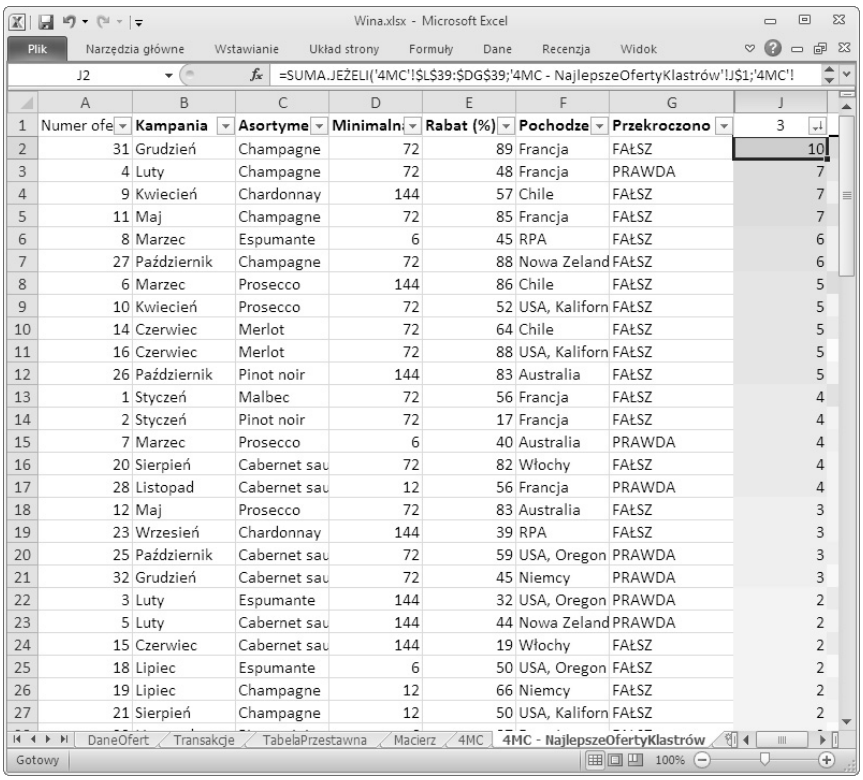
Niestety, w wyniku sortowania klastra 3. nie da się wyciągnąć tak oczywistych wniosków. W tym klastrze jest wiele bardzo popularnych ofert i nie widać wyraźnej granicy pomiędzy nimi a ofertami, z których nikt nie skorzystał. Najpopularniejsze oferty w tej grupie wydają się mieć coś wspólnego — wszystkie charakteryzują się dużą zniżką. Pięć z sześciu



	A	B	C	D	E	F	G	I
1	Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	2
2	30	Grudzień	Malbec	6	54	Francja	FAŁSZ	16
3	29	Listopad	Pinot grigio	6	87	Francja	FAŁSZ	15
4	7	Marzec	Prosecco	6	40	Australia	PRAWDA	12
5	8	Marzec	Espumante	6	45	RPA	FAŁSZ	11
6	18	Lipiec	Espumante	6	50	USA, Oregon	FAŁSZ	11
7	13	Maj	Merlot	6	43	Chile	FAŁSZ	6
8	1	Styczeń	Malbec	72	56	Francja	FAŁSZ	0
9	2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ	0
10	3	Luty	Espumante	144	32	USA, Oregon	PRAWDA	0
11	4	Luty	Champagne	72	48	Francja	PRAWDA	0

Rysunek 2.23. Sortowanie klastra 2. — osoby preferujące małe zakupy

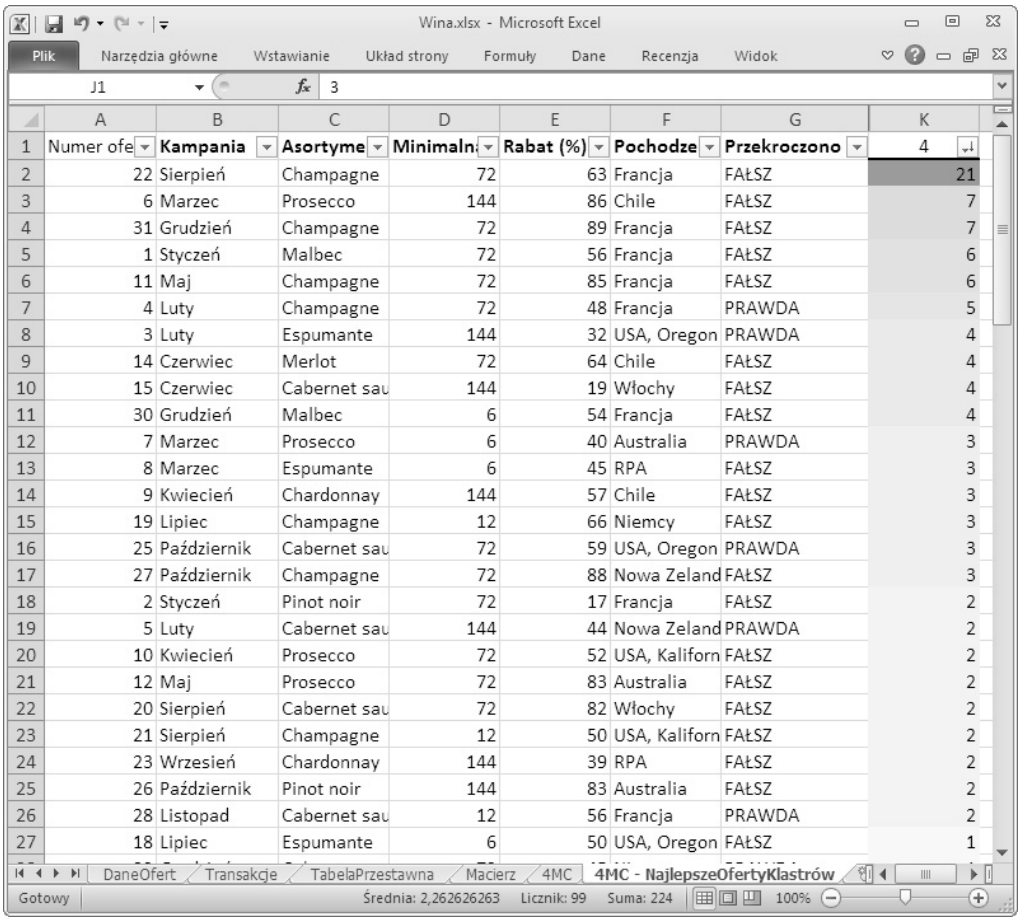
najpopularniejszych ofert to prawdziwe okazje, a wśród czterech najpopularniejszych aż trzy dotyczą win z Francji. Niestety, trudno jest tu wysnuć jakieś jednoznaczne wnioski (rysunek 2.24).



	A	B	C	D	E	F	G	J
1	Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	3
2	31	Grudzień	Champagne	72	89	Francja	FAŁSZ	10
3	4	Luty	Champagne	72	48	Francja	PRAWDA	7
4	9	Kwiecień	Chardonnay	144	57	Chile	FAŁSZ	7
5	11	Maj	Champagne	72	85	Francja	FAŁSZ	7
6	8	Marzec	Espumante	6	45	RPA	FAŁSZ	6
7	27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ	6
8	6	Marzec	Prosecco	144	86	Chile	FAŁSZ	5
9	10	Kwiecień	Prosecco	72	52	USA, Kalifor	FAŁSZ	5
10	14	Czerwiec	Merlot	72	64	Chile	FAŁSZ	5
11	16	Czerwiec	Merlot	72	88	USA, Kalifor	FAŁSZ	5
12	26	Październik	Pinot noir	144	83	Australia	FAŁSZ	5
13	1	Styczeń	Malbec	72	56	Francja	FAŁSZ	4
14	2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ	4
15	7	Marzec	Prosecco	6	40	Australia	PRAWDA	4
16	20	Sierpień	Cabernet sau	72	82	Włochy	FAŁSZ	4
17	28	Listopad	Cabernet sau	12	56	Francja	PRAWDA	4
18	12	Maj	Prosecco	72	83	Australia	FAŁSZ	3
19	23	Wrzesień	Chardonnay	144	39	RPA	FAŁSZ	3
20	25	Październik	Cabernet sau	72	59	USA, Oregon	PRAWDA	3
21	32	Grudzień	Cabernet sau	72	45	Niemcy	PRAWDA	3
22	3	Luty	Espumante	144	32	USA, Oregon	PRAWDA	2
23	5	Luty	Cabernet sau	144	44	Nowa Zeland	PRAWDA	2
24	15	Czerwiec	Cabernet sau	144	19	Włochy	FAŁSZ	2
25	18	Lipiec	Espumante	6	50	USA, Oregon	FAŁSZ	2
26	19	Lipiec	Champagne	12	66	Niemcy	FAŁSZ	2
27	21	Sierpień	Champagne	12	50	USA, Kalifor	FAŁSZ	2

Rysunek 2.24. Sortowanie klastra 3. nie daje jednoznacznych wniosków

Przyglądając się klastrowi 4., można wysnuć wniosek, że wszystkie zakwalifikowane do niego osoby z jakiegoś powodu lubią dokonywać zakupów w sierpniu. Ponadto pięć z sześciu najpopularniejszych ofert dotyczy win pochodzących z Francji, a dziesięć najpopularniejszych — zakupu dużej ilości win (rysunek 2.25). Być może w tym klastrze znajdują się klienci preferujący zakup dużej ilości francuskich win? Nakładanie się klastrów 3. i 4. jest dość kłopotliwe.



	A	B	C	D	E	F	G	K
1	Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	4
2	22	Sierpień	Champagne	72	63	Francja	FAŁSZ	21
3	6	Marzec	Prosecco	144	86	Chile	FAŁSZ	7
4	31	Grudzień	Champagne	72	89	Francja	FAŁSZ	7
5	1	Styczeń	Malbec	72	56	Francja	FAŁSZ	6
6	11	Maj	Champagne	72	85	Francja	FAŁSZ	6
7	4	Luty	Champagne	72	48	Francja	PRAWDA	5
8	3	Luty	Espumante	144	32	USA, Oregon	PRAWDA	4
9	14	Czerwiec	Merlot	72	64	Chile	FAŁSZ	4
10	15	Czerwiec	Cabernet sau	144	19	Włochy	FAŁSZ	4
11	30	Grudzień	Malbec	6	54	Francja	FAŁSZ	4
12	7	Marzec	Prosecco	6	40	Australia	PRAWDA	3
13	8	Marzec	Espumante	6	45	RPA	FAŁSZ	3
14	9	Kwiecień	Chardonnay	144	57	Chile	FAŁSZ	3
15	19	Lipiec	Champagne	12	66	Niemcy	FAŁSZ	3
16	25	Październik	Cabernet sau	72	59	USA, Oregon	PRAWDA	3
17	27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ	3
18	2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ	2
19	5	Luty	Cabernet sau	144	44	Nowa Zeland	PRAWDA	2
20	10	Kwiecień	Prosecco	72	52	USA, Kaliforn	FAŁSZ	2
21	12	Maj	Prosecco	72	83	Australia	FAŁSZ	2
22	20	Sierpień	Cabernet sau	72	82	Włochy	FAŁSZ	2
23	21	Sierpień	Champagne	12	50	USA, Kaliforn	FAŁSZ	2
24	23	Wrzesień	Chardonnay	144	39	RPA	FAŁSZ	2
25	26	Październik	Pinot noir	144	83	Australia	FAŁSZ	2
26	28	Listopad	Cabernet sau	12	56	Francja	PRAWDA	2
27	18	Lipiec	Espumante	6	50	USA, Oregon	FAŁSZ	1

Rysunek 2.25. Sortowanie klastra 4. — czy ta grupa klientów po prostu lubi pić szampana w sierpniu?

Dochodzimy do pewnej wątpliwości: czy cztery to odpowiednia liczba grup, na które dzielimy naszych klientów za pomocą algorytmu centroidów? Być może nie. Jak zatem określić właściwą liczbę grup?

Sylwetka podziału — dobry sposób na określenie optymalnej liczby klastrów

Nie ma niczego złego w dokonywaniu podziału na różne liczby grup aż do momentu dojścia do wniosku, że któryś podział wydaje się sensowny. Oczywiście czasami wynik podziału okazuje się bezsensowny, ponieważ w analizowanych danych brakuje informacji, które umożliwiłyby sensowny podział na klastry.

Czy istnieje jakiś sposób na ocenienie doboru liczby grup, na które dzielimy nasz zbiór danych, poza dokonywaniem podziału i analizowaniem go „gołym okiem"?

Owszem, istnieje. Możesz obliczyć parametr określający jakość klastrów — **sylwetkę podziału** (ang. *silhouette*). Zaletą tego parametru jest to, że na jakość otrzymanych wyników nie wpływa liczba grup, na które dokonujemy podziału.

Sylwetka podziału na wysokim poziomie — jak daleko od Ciebie są Twoi sąsiedzi?

Możesz porównać średnią odległość pomiędzy każdym klientem a jego sąsiadami przyporządkowanymi do tego samego klastra ze średnią odległością do klientów w klastrze, którego centrum znajduje się najbliżej.

Jeżeli znajduję się bliżej osób należących do mojej grupy niż osób należących do sąsiedniej grupy, to chyba zostałem przydzielony do właściwej grupy, mam rację? A co, jeżeli osoby z sąsiedniej grupy znajdują się praktycznie tak blisko mnie jak osoby z grupy, do której zostałem przyporządkowany? Mogę wtedy dojść do wniosku, że moje przydzielenie nie jest do końca przemyślane.

W sposób formalny można to obliczyć za pomocą wzoru:

> *(średnia odległość od elementów sąsiedniego klastra – średnia odległość od elementów mojego klastra) / maksimum tych dwóch średnich*

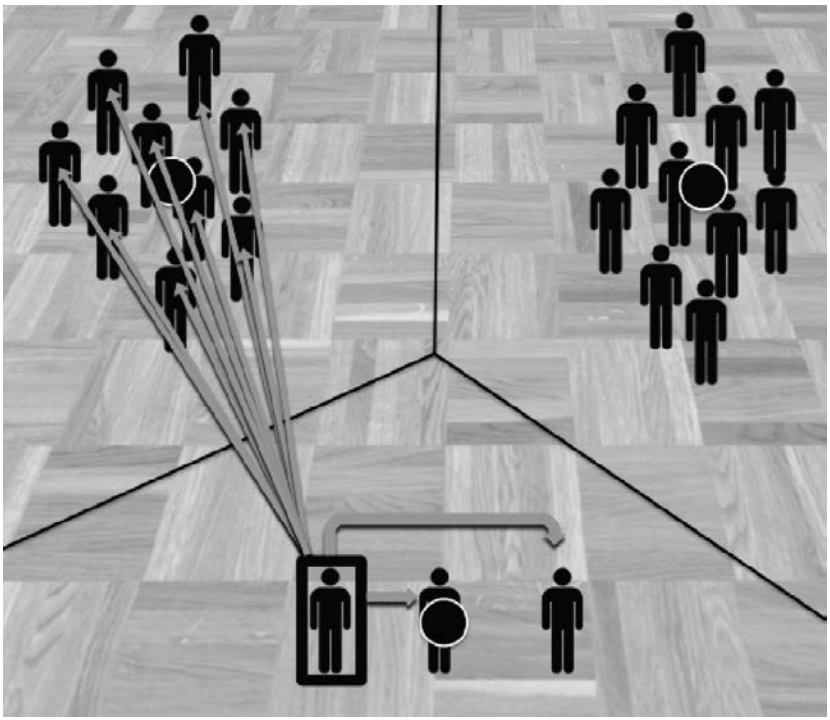
Mianownik tego wzoru sprawia, że umożliwia on otrzymanie wartości z zakresu od –1 do 1.

Przeanalizuj ten wzór. Gdy osoby z innej grupy oddalają się ode mnie (bardziej się ode mnie różnią), to wzór zwraca wartość bliższą 1. Jeżeli obie średnie odległości są podobne, to wzór zwraca wartość zbliżoną do 0.

Obliczając średnią wartość takich liczb charakteryzujących każdego klienta, uzyskamy sylwetkę podziału. Gdy sylwetka podziału przyjmuje wartość 1, to podział jest idealny. Zero oznacza zły podział klastrów, a wartość mniejsza od zera oznacza, że wielu klientów pasowałoby lepiej do innego klastra.

Obliczając wartości sylwetki podziałów dla różnych liczb grup, można stwierdzić, czy dodanie grupy poprawia podział.

Aby wyjaśnić działanie tej techniki, wróćmy jeszcze raz do przykładu dyskoteki szkolnej. Na rysunku 2.26 przedstawiono obliczenia odległości użyte w celu określenia sylwetki podziału. Porównano odległość pomiędzy jednym z opiekunów a dwoma pozostałymi opiekunami z odległością pomiędzy tym opiekunem a osobami przyporządkowanymi do najbliższego klastra (grupy chłopców).



Rysunek 2.26. Odległości, które należy wziąć pod uwagę w przypadku określania wpływu przyporządkowania osoby pilnującej na wartość określającą sylwetkę podziału

W sytuacji widocznej na rysunku osoba pilnująca znajduje się o wiele bliżej pozostałych opiekunów niż nastolatków, a więc wartość współczynnika odległości będzie w tym przypadku wyraźnie większa od zera.

Tworzenie macierzy odległości

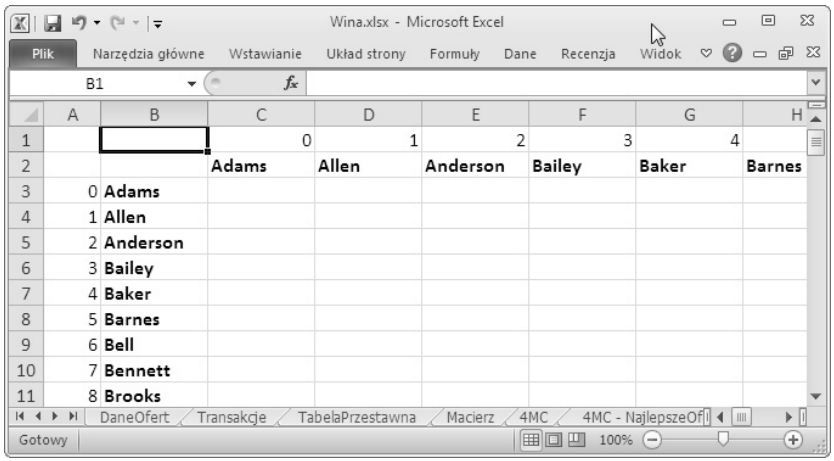
W celu zaimplementowania metody sylwetki podziału musisz uzyskać ważne dane — odległość pomiędzy klientami. Środki klastrów mogą zmieniać położenie, ale odległości pomiędzy klientami są stałe. W związku z tym możesz stworzyć jedną tabelę z odległościami i korzystać z niej przy wszystkich obliczeniach sylwetki podziału (przy podziale na różne liczby grup i przy różnych współrzędnych ich środków).

Zacznij od utworzenia pustego arkusza o nazwie *Odległości*. Skopiuj do niego nazwiska klientów w pionie i w poziomie. Każda komórka tej tabeli będzie definiowała odległość pomiędzy klientem, którego nazwisko jest umieszczone w kolumnie, a klientem, którego nazwisko zapisano w wierszu. W celu wklejenia danych klientów do wierszy tej zakładki skopiuj komórki `H1:DC1` z zakładki *Macierz* i skorzystaj z polecenia *Wklej specjalnie…* Pamiętaj o wybraniu opcji wklejania *Wartości* i zaznaczeniu pola *Transpozycja* w oknie *Wklejanie specjalne*.

Aby śledzić położenia klientów w tej dużej tabeli, ponumeruj ich (przypisz im numery od 0 do 99 w obu kierunkach). Numery te umieść w kolumnie A i wierszu 1. W tym celu wstaw pusty wiersz i pustą kolumnę od góry i na lewo od wiersza i kolumny, do których wkleiłeś nazwiska (kliknij kolumnę A i wiersz 1. prawym przyciskiem myszy, a następnie wybierz opcję wstawiania nowej kolumny i nowego wiersza).

UWAGA

Kolejne liczby z zakresu 0 – 99 możesz wstawić na wiele różnych sposobów. Na przykład zacznij od wpisania cyfr 0, 1, 2, 3 w pierwsze komórki. Następnie zaznacz je i przeciągnij dolny róg zaznaczonego obszaru na pozostałe komórki. Excel powinien zrozumieć Twój zamiar i automatycznie dokończyć sekwencję. Na rysunku 2.27 przedstawiono tabelę gotową do zapełnienia danymi.



	A	B	C	D	E	F	G	H
1			0	1	2	3	4	
2			Adams	Allen	Anderson	Bailey	Baker	Barnes
3	0	Adams						
4	1	Allen						
5	2	Anderson						
6	3	Bailey						
7	4	Baker						
8	5	Barnes						
9	6	Bell						
10	7	Bennett						
11	8	Brooks						

Rysunek 2.27. Pusta tabela odległości

Spójrz na komórkę C3 — powinna zawierać odległość pomiędzy klientem o nazwisku *Adams* a klientem o nazwisku *Adams*, a więc odległość pomiędzy tą samą osobą. Będzie ona wynosić 0 — nikt nie może być bliżej siebie od siebie samego.

Jak obliczyć tę odległość? W kolumnie H zakładki *Macierz* znajduje się wektor transakcji dokonanych przez klienta o nazwisku *Adams*. W celu obliczenia odległości euklidesowej pomiędzy tym użytkownikiem a nim samym wystarczy od kolumny H odjąć kolumnę H, uzyskaną różnicę podnieść do kwadratu, zsumować otrzymane wartości, a następnie wyciągnąć z nich pierwiastek kwadratowy.

Jak przeciągnąć te obliczenia do każdej komórki arkusza? Ręczne wpisywanie każdej formuły byłoby bardzo czasochłonne. W związku z tym w komórce C3 musisz skorzystać z formuły PRZESUNIĘCIE. Więcej informacji na temat tej formuły znajdziesz w rozdziale 1.

Formuła PRZESUNIĘCIE przyjmuje określony zakres komórek (w tym wypadku wektor transakcji dokonanych przez klienta o nazwisku *Adams* — Macierz!$H$2:$H$33), a następnie przesuwa cały zakres o określoną liczbę rzędów i kolumn.

Na przykład formuła `PRZESUNIĘCIE(Macierz!$H$2:$H$33,0,0)` zwraca wektor transakcji dokonanych przez klienta o nazwisku *Adams*, ponieważ przesuwamy podany zakres o 0 wierszy w dół i 0 kolumn w prawo.

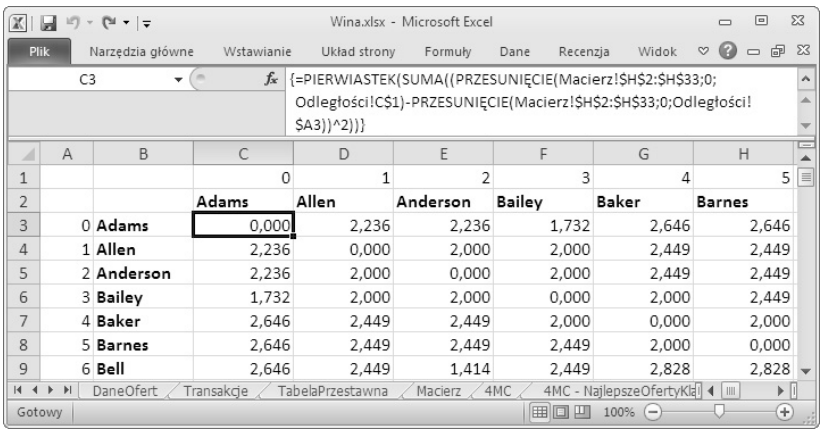
Natomiast formuła `PRZESUNIĘCIE(Macierz!$H$2:$H$33,0,1)` zwróci kolumnę z transakcjami klienta o nazwisku *Allen*, formuła `PRZESUNIĘCIE(Macierz!$H$2:$H$33,0,2)` zwróci kolumnę z transakcjami klienta o nazwisku *Anderson* itd.

Właśnie do tego przydadzą się wartości z zakresu od 0 do 99, które wpisaliśmy do wiersza 1. i kolumny `A`. Możemy dzięki nim zbudować np. następującą formułę:

```
{=PIERWIASTEK(SUMA((PRZESUNIĘCIE(Macierz!$H$2:$H$33,0,Odległości!C$1)-PRZESUNIĘCIE
(Macierz!$H$2:$H$33,0,Odległości!$A3))^2))}
```

W ten sposób określimy odległość pomiędzy klientem o nazwisku *Adams* a nim samym. Zauważ, że `Odległości!C$1` definiuje przesunięcie kolumny w pierwszym wektorze transakcji, a `Odległości!$A3` definiuje przesunięcie kolumny w drugim wektorze transakcji.

Dzięki temu po przeciągnięciu tej formuły na cały arkusz wszystko będzie zakotwiczone na wektorze transakcji klienta o nazwisku *Adams*, ale formuła `PRZESUNIĘCIE` będzie przesuwała wektor we właściwe miejsca wskazywane przez indeksy umieszczone w kolumnie `A` i wierszu 1. W ten sposób będziemy przetwarzać właściwe wektory transakcji klientów. Na rysunku 2.28 pokazano tabelę wypełnioną wartościami odległości.



	A	B	C	D	E	F	G	H
1			0	1	2	3	4	5
2			Adams	Allen	Anderson	Bailey	Baker	Barnes
3	0	Adams	0,000	2,236	2,236	1,732	2,646	2,646
4	1	Allen	2,236	0,000	2,000	2,000	2,449	2,449
5	2	Anderson	2,236	2,000	0,000	2,000	2,449	2,449
6	3	Bailey	1,732	2,000	2,000	0,000	2,000	2,449
7	4	Baker	2,646	2,449	2,449	2,000	0,000	2,000
8	5	Barnes	2,646	2,449	2,449	2,449	2,000	0,000
9	6	Bell	2,646	2,449	1,414	2,449	2,828	2,828

Rysunek 2.28. Tabela odległości wypełniona danymi

Pamiętaj, że formuły wpisywane w zakładce *Odległości* muszą być formułami tablicowymi, tak jak w przypadku zakładki *4MC*.

Excel i implementacja sylwetki podziału

Przygotowałeś dane w zakładce *Odległości*, a więc możesz przystąpić do wykonywania obliczeń niezbędnych do określenia wartości sylwetki podziału. Utwórz nowy arkusz i nazwij go `4MC Sylwetka`.

Na początek skopiuj z zakładki *4MC* nazwiska klientów i numery grup, do których zostali przypisani — skorzystaj z opcji *Wklej specjalnie…* i umieść nazwiska w kolumnie `A`, a numery grup umieść w kolumnie `B` (nie zapomnij zaznaczyć opcji *Transponuj* w oknie *Wklejanie specjalne*).

Teraz będziesz mógł skorzystać z arkusza *Odległości* i obliczyć średnią odległość pomiędzy każdym klientem a klientami przyporządkowanymi do tej samej grupy. W kolumnach od `C` do `F` umieść etykiety `Odległość od członków 1. grupy` – `Odległość od członków 4. grupy`.

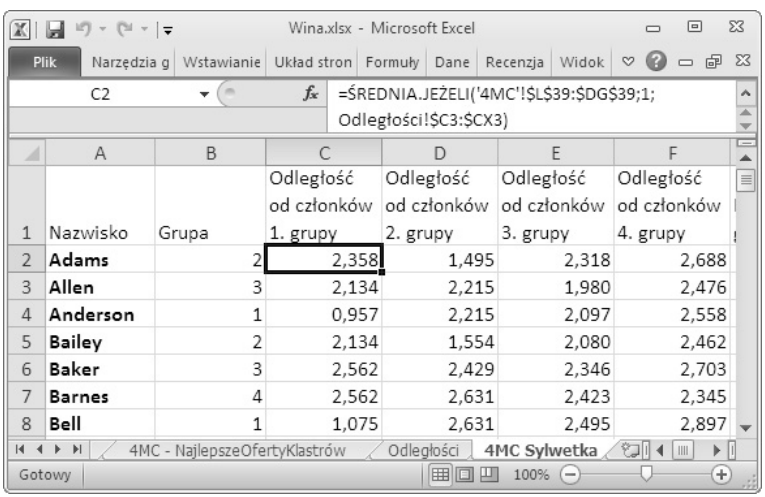
W moim skoroszycie klient o nazwisku *Adams* został przypisany do klastra 2., a więc w komórce `C2` będziesz musiał obliczyć odległość pomiędzy nim a wszystkimi klientami przypisanymi do klastra 1. Musisz przeszukać listę klientów i wybrać tych, którzy zostali przypisani do klastra 1., następnie obliczyć średnią odległość pomiędzy nimi a klientem o nazwisku *Adams* (możesz skorzystać z wiersza 3. arkusza *Odległości*).

Brzmi to jak typowe zastosowanie formuły `ŚREDNIA.JEŻELI`:

```
=ŚREDNIA.JEŻELI('4MC'!$L$39:$DG$39,1,Odległości!$C3:$CX3)
```

Formuła `ŚREDNIA.JEŻELI` sprawdza przypisania do klastrów i dobiera je do klastra 1. przed określeniem średniej odpowiednich odległości z komórek `C3:CX3`.

Formuły wpisywane w kolumnach `D` – `F` są identyczne, ale zamiast do klastra 1. odwołujemy się do klastrów 2., 3. i 4. Po umieszczeniu formuł w odpowiednich kolumnach kliknij je dwukrotnie w celu skopiowania do pozostałych komórek klientów. W ten sposób uzyskasz tabelę pokazaną na rysunku 2.29.



	A	B	C	D	E	F
1	Nazwisko	Grupa	Odległość od członków 1. grupy	Odległość od członków 2. grupy	Odległość od członków 3. grupy	Odległość od członków 4. grupy
2	**Adams**	2	2,358	1,495	2,318	2,688
3	**Allen**	3	2,134	2,215	1,980	2,476
4	**Anderson**	1	0,957	2,215	2,097	2,558
5	**Bailey**	2	2,134	1,554	2,080	2,462
6	**Baker**	3	2,562	2,429	2,346	2,703
7	**Barnes**	4	2,562	2,631	2,423	2,345
8	**Bell**	1	1,075	2,631	2,495	2,897

Rysunek 2.29. Średnie odległości pomiędzy poszczególnymi klientami a klientami przypisanymi do każdego z klastrów

W kolumnie G możesz dokonać obliczeń dla najbliższej grupy klientów — skorzystaj z formuły MIN. W przypadku klienta o nazwisku *Adams* zastosuj formuły:

```
=MIN(C2:F2)
```

W kolumnie H za pomocą formuły MIN.K możesz obliczyć wartości dla drugiej najbliższej grupy klientów (w podanym przykładzie zastosowaliśmy parametr 2, ponieważ formuła ma określać drugie najbliższe miejsce):

```
=MIN.K(C2:F2,2)
```

W podobny sposób możesz obliczyć w kolumnie I odległość od klientów przyporządkowanych do tej samej grupy (prawdopodobnie będzie to wartość identyczna z tą, która znalazła się w kolumnie G, ale nie jest to reguła):

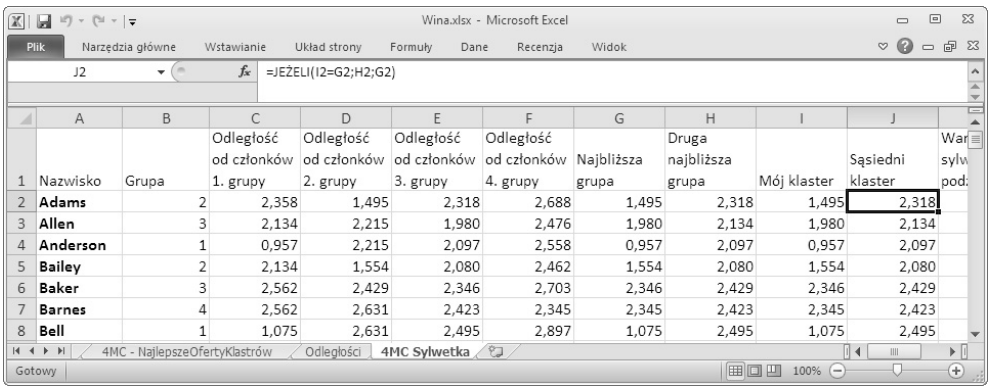
```
=INDEKS(C2:F2,B2)
```

Formuła INDEKS jest używana w celu określenia właściwej odległości zapisanej w kolumnach C – F przy użyciu wartości zapisanej w kolumnie B, która to kolumna pełni funkcję indeksu.

W celu obliczenia wartości sylwetki podziału musisz również określić odległość od najbliższej grupy klientów, którzy *nie* należą do klastra, do którego przyporządkowany został analizowany klient (zwykle wartość ta będzie równa tej, którą umieszczono w kolumnie H, ale nie jest to regułą). W celu określenia tej wartości w kolumnie J musisz porównać odległość od własnego klastra umieszczoną w kolumnie I z odległością od najbliższego klastra. Jeżeli te wartości są identyczne, to w kolumnie J wpisujemy wartość odczytaną z kolumny H, w przeciwnym wypadku wpisujemy wartość odczytaną z kolumny G:

```
=JEŻELI(I2=G2,H2,G2)
```

Po skopiowaniu tych formuł w dół uzyskasz arkusz przedstawiony na rysunku 2.30.



	A	B	C	D	E	F	G	H	I	J	
1	Nazwisko	Grupa	Odległość od członków 1. grupy	Odległość od członków 2. grupy	Odległość od członków 3. grupy	Odległość od członków 4. grupy	Najbliższa grupa	Druga najbliższa grupa	Mój klaster	Sąsiedni klaster	War sylw pod
2	**Adams**	2	2,358	1,495	2,318	2,688	1,495	2,318	1,495	2,318	
3	**Allen**	3	2,134	2,215	1,980	2,476	1,980	2,134	1,980	2,134	
4	**Anderson**	1	0,957	2,215	2,097	2,558	0,957	2,097	0,957	2,097	
5	**Bailey**	2	2,134	1,554	2,080	2,462	1,554	2,080	1,554	2,080	
6	**Baker**	3	2,562	2,429	2,346	2,703	2,346	2,429	2,346	2,429	
7	**Barnes**	4	2,562	2,631	2,423	2,345	2,345	2,423	2,345	2,423	
8	**Bell**	1	1,075	2,631	2,495	2,897	1,075	2,495	1,075	2,495	

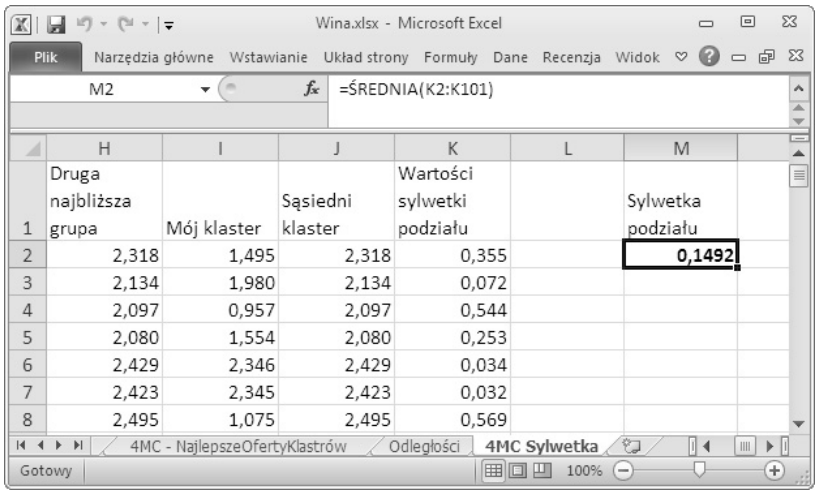
Rysunek 2.30. Średnie odległości pomiędzy klientami przyporządkowanymi do tej samej grupy a klientami przyporządkowanymi do najbliższej grupy, w której dany klient się nie znajduje

Dysponując obliczonymi wcześniej wartościami, określenie wartości sylwetki podziału dla każdego klienta nie stanowi żadnego problemu. Wystarczy zastosować formułę:

```
=(J2-I2)/MAX(J2,I2)
```

Skopiuj tę formułę do komórek znajdujących się poniżej, a uzyskasz współczynniki wartości sylwetki podziału każdego klienta.

Przyglądając się uzyskanym wynikom, zauważysz, że niektóre z nich są bliskie 1. Na przykład wartość sylwetki podziału klienta o nazwisku *Anderson* wynosi w przypadku mojego podziału na grupy 0,544 (rysunek 2.31). Nieźle! Niestety, w przypadku klientów takich jak np. *Collins* wartość ta jest mniejsza od 0, co świadczy o tym, że charakterystyki tego klienta są bliższe sąsiedniemu klastrowi. Biedak.



	H	I	J	K	L	M
1	Druga najbliższa grupa	Mój klaster	Sąsiedni klaster	Wartości sylwetki podziału		Sylwetka podziału
2	2,318	1,495	2,318	0,355		0,1492
3	2,134	1,980	2,134	0,072		
4	2,097	0,957	2,097	0,544		
5	2,080	1,554	2,080	0,253		
6	2,429	2,346	2,429	0,034		
7	2,423	2,345	2,423	0,032		
8	2,495	1,075	2,495	0,569		

Rysunek 2.31. Współczynnik sylwetki podziału przy podziale na cztery klastry

Teraz możesz obliczyć średnią tych wartości, która będzie równa ogólnemu współczynnikowi sylwetki podziału. W moim przypadku (rysunek 2.31) wynosi on *0,1492*. Wartość ta jest wyraźnie bardziej zbliżona do 0 niż 1. To smutne, ale wcale nie zaskakujące. W końcu dwa z czterech klastrów były trudne do jednoznacznego zinterpretowania i opisania za pomocą preferowanych ofert.

No dobrze, co dalej?

Współczynnik sylwetki podziału wynosi *0,1492*. Co to znaczy? Jak można z tego skorzystać? Wypróbuj podział na inną liczbę grup. Później będziesz mógł porównać współczynniki sylwetki tych podziałów i dowiedzieć się, czy dzieląc klientów na większą liczbę klastrów, poprawiasz jakość podziału.

A może potrzebujesz pięciu klastrów?

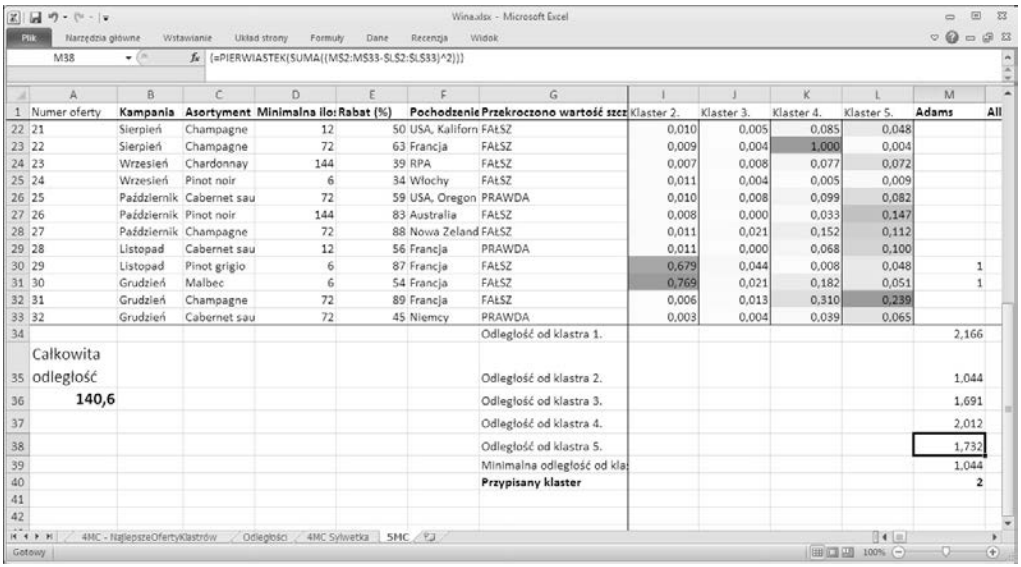
Spróbuj podzielić klientów na pięć grup i zobacz, co się stanie.

Mam dla Ciebie dobrą wiadomość: opracowałeś już arkusz dla czterech klastrów, a więc nie musisz zaczynać pracy nad kolejnym arkuszem od podstaw. Ponadto nie musisz w żaden sposób modyfikować arkusza *Odległość*. Czas rozpocząć pracę.

Utwórz kopię arkusza *4MC* i nazwij go `5MC`. Wystarczy, że dodasz do niego piąty klaster i uwzględnisz go w obliczeniach.

Kliknij prawym przyciskiem myszy kolumnę `L` i wstaw obok niej nową kolumnę o nazwie `Klaster 5`. Musisz również dodać wiersz `Odległość od klastra 5.` — kliknij 38. rząd komórek i wstaw nad nim pusty wiersz. Możesz skopiować zawartość wiersza *Odległość od klastra 4.* i zmienić w jego formułach odwołania do kolumny `K` na odwołania do kolumny `L`. Ponadto formuły wierszy *Minimalna odległość od klastra* i *Przypisany klaster* muszą zawierać odwołania do wiersza 38., a nie 37. (zwiększ zakres formuł o nowy klaster).

Po wykonaniu tych czynności uzyskasz arkusz przedstawiony na rysunku 2.32.



	A	B	C	D	E	F	G	I	J	K	L	M	
1	Numer oferty	Kampania	Asortyment	Minimalna ilo:	Rabat (%)	Pochodzenie	Przekroczono wartość szcz	Klaster 2.	Klaster 3.	Klaster 4.	Klaster 5.	Adams	All
22	21	Sierpień	Champagne	12	50	USA, Kaliforn	FAŁSZ	0,010	0,005	0,085	0,048		
23	22	Sierpień	Champagne	72	63	Francja	FAŁSZ	0,009	0,004	1,000	0,004		
24	23	Wrzesień	Chardonnay	144	39	RPA	FAŁSZ	0,007	0,008	0,077	0,072		
25	24	Wrzesień	Pinot noir	6	34	Włochy	FAŁSZ	0,011	0,004	0,005	0,009		
26	25	Październik	Cabernet sau	72	59	USA, Oregon	PRAWDA	0,010	0,008	0,099	0,082		
27	26	Październik	Pinot noir	144	83	Australia	FAŁSZ	0,008	0,000	0,033	0,147		
28	27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ	0,011	0,021	0,152	0,112		
29	28	Listopad	Cabernet sau	12	56	Francja	PRAWDA	0,011	0,000	0,068	0,100		
30	29	Listopad	Pinot grigio	6	87	Francja	FAŁSZ	0,679	0,044	0,008	0,048	1	
31	30	Grudzień	Malbec	6	54	Francja	FAŁSZ	0,769	0,021	0,182	0,051	1	
32	31	Grudzień	Champagne	72	89	Francja	FAŁSZ	0,006	0,013	0,310	0,239		
33	32	Grudzień	Cabernet sau	72	45	Niemcy	PRAWDA	0,003	0,004	0,039	0,065		
34							Odległość od klastra 1.					2,166	
35	Całkowita odległość						Odległość od klastra 2.					1,044	
36	140,6						Odległość od klastra 3.					1,691	
37							Odległość od klastra 4.					2,012	
38							Odległość od klastra 5.					1,732	
39							Minimalna odległość od kla:					1,044	
40							**Przypisany klaster**					2	

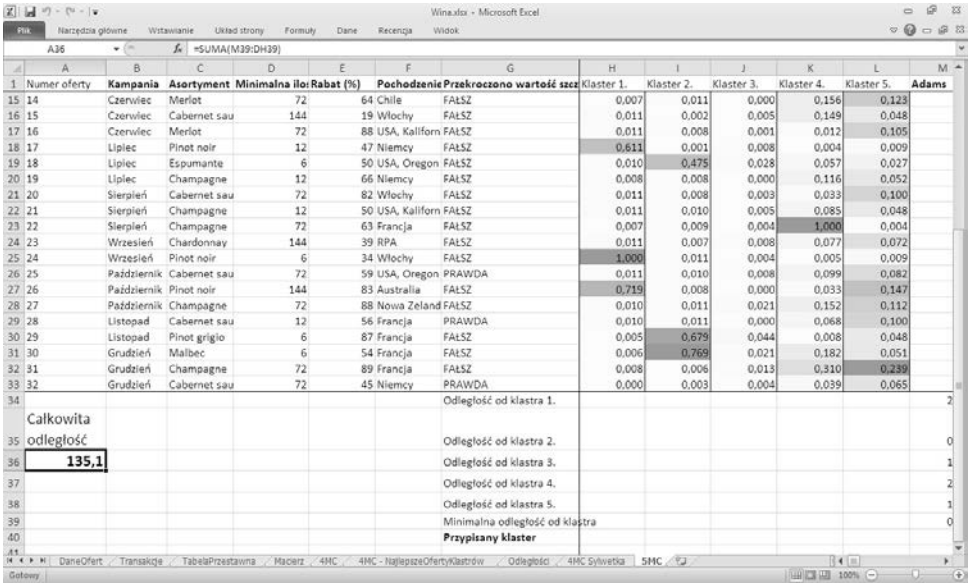
Rysunek 2.32. Tworzenie podziału na pięć grup

Dzielenie klientów na pięć klastrów za pomocą narzędzia Solver

Otwórz narzędzie Solver. W zmiennych decyzyjnych i ograniczeniach musisz zmienić `$H$2:$K$33` na `$H$2:$L$33`. Teraz będą one uwzględniać nowy (piąty) klaster. Pozostałe opcje narzędzia Solver pozostają bez zmian.

Kliknij przycisk *Rozwiąż* i poczekaj na rozwiązanie problemu.

W moim przypadku Solver zakończył pracę, uzyskując całkowitą odległość równą *135,1* (rysunek 2.33).



Numer oferty	Kampania	Asortyment	Minimalna ilo	Rabat (%)	Pochodzenie	Przekroczono wartość szcz	Klaster 1.	Klaster 2.	Klaster 3.	Klaster 4.	Klaster 5.	Adams
14	Czerwiec	Merlot	72	64	Chile	FAŁSZ	0,007	0,011	0,000	0,156	0,123	
15	Czerwiec	Cabernet sau	144	19	Włochy	FAŁSZ	0,011	0,002	0,005	0,149	0,048	
16	Czerwiec	Merlot	72	88	USA, Kalifor	FAŁSZ	0,011	0,008	0,001	0,012	0,105	
17	Lipiec	Pinot noir	12	47	Niemcy	FAŁSZ	0,611	0,001	0,008	0,004	0,009	
18	Lipiec	Espumante	6	50	USA, Oregon	FAŁSZ	0,010	0,475	0,028	0,057	0,027	
19	Lipiec	Champagne	12	66	Niemcy	FAŁSZ	0,008	0,008	0,000	0,116	0,052	
20	Sierpień	Cabernet sau	72	82	Włochy	FAŁSZ	0,011	0,008	0,003	0,033	0,100	
21	Sierpień	Champagne	12	50	USA, Kalifor	FAŁSZ	0,011	0,010	0,005	0,085	0,048	
22	Sierpień	Champagne	72	63	Francja	FAŁSZ	0,007	0,009	0,004	1,000	0,004	
23	Wrzesień	Chardonnay	144	39	RPA	FAŁSZ	0,011	0,007	0,008	0,077	0,072	
24	Wrzesień	Pinot noir	6	34	Włochy	FAŁSZ	1,000	0,011	0,004	0,005	0,009	
25	Październik	Cabernet sau	72	59	USA, Oregon	PRAWDA	0,011	0,010	0,008	0,099	0,082	
26	Październik	Pinot noir	144	83	Australia	FAŁSZ	0,719	0,008	0,000	0,033	0,147	
27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ	0,010	0,011	0,021	0,152	0,112	
28	Listopad	Cabernet sau	12	56	Francja	PRAWDA	0,010	0,011	0,000	0,068	0,100	
29	Listopad	Pinot grigio	6	87	Francja	FAŁSZ	0,005	0,679	0,044	0,008	0,048	
30	Grudzień	Malbec	6	54	Francja	FAŁSZ	0,006	0,769	0,021	0,182	0,051	
31	Grudzień	Champagne	72	89	Francja	FAŁSZ	0,008	0,006	0,013	0,310	0,239	
32	Grudzień	Cabernet sau	72	45	Niemcy	PRAWDA	0,000	0,003	0,004	0,039	0,065	

Całkowita odległość	
135,1	
Odległość od klastra 1.	2
Odległość od klastra 2.	0
Odległość od klastra 3.	1
Odległość od klastra 4.	2
Odległość od klastra 5.	1
Minimalna odległość od klastra	0
Przypisany klaster	

Rysunek 2.33. Optymalny podział klientów na pięć klastrów

Ustalanie najlepszych ofert dla wszystkich pięciu klastrów

Czas sprawdzić sensowność podziału klientów na pięć grup. Utwórz kopię arkusza *4MC — NajlepszeOfertyKlastrów* i nazwij ją `5MC – NajlepszeOfertyKlastrów`. Dodatkowo musisz poddać ją kilku modyfikacjom.

Zacznij od upewnienia się, że arkusz jest posortowany według numeru oferty (kolumna `A`). Następnie nadaj etykietę `5` kolumnie `L` i przeciągnij formuły z kolumny `K` do kolumny `L`. Zaznacz kolumny od `A` do `L` i zastosuj ponownie automatyczne filtrowanie, tak aby móc sortować zakupy dokonane przez klientów z klastra 5.

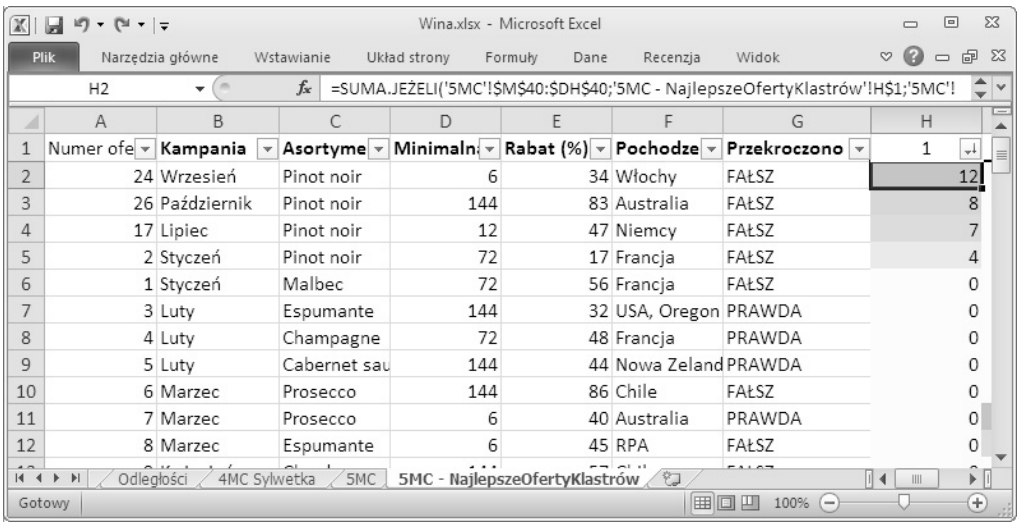
Obecnie wszystkie formuły tego arkusza odwołują się do arkusza *4MC*, a więc czas skorzystać z dobrze znanego Ci okna *Znajdowanie i zamienianie*. Dane klastrów w arkuszu *5MC* są przesunięte o jeden wiersz w dół i jedną kolumnę w prawo, a więc odwołanie do komórki `'4MC'!$L$39:$DG$39` w formułach `SUMA.JEŻELI` powinno zostać zastąpione odwołaniem `'5MC'!$M$40:$DH$40`. Odwołania te można zmienić za pomocą okna *Znajdowanie i zamienianie* (rysunek 2.34).

Rysunek 2.34. Zastępowanie odwołań do podziału na cztery klastry odwołaniami do podziału na pięć klastrów

UWAGA

Pamiętaj, że wyniki widoczne w Twoim arkuszu mogą różnić się od tych uzyskanych przeze mnie z powodu zastosowania algorytmu ewolucyjnego.

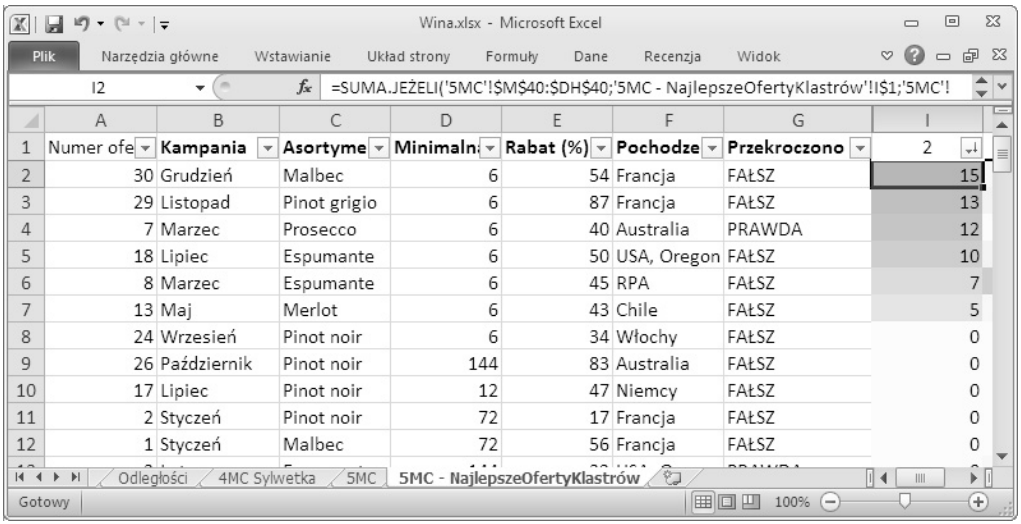
W wyniku posortowania klastra 1. ponownie wyraźnie widać, że przyporządkowani do niego kupujący preferują wino pinot noir (rysunek 2.35).



Rysunek 2.35. W wyniku posortowania klastra 1. ponownie widać preferencje do zakupu wina pinot noir

Klaster 2. ponownie zawiera osoby kupujące małe ilości wina (rysunek 2.36).

Klaster 3. przyprawia mnie o ból głowy. Z jakiegoś powodu znajdujący się w nim klienci preferują wino espumante pochodzące z RPA (rysunek 2.37).



Wina.xlsx - Microsoft Excel

Plik | Narzędzia główne | Wstawianie | Układ strony | Formuły | Dane | Recenzja | Widok

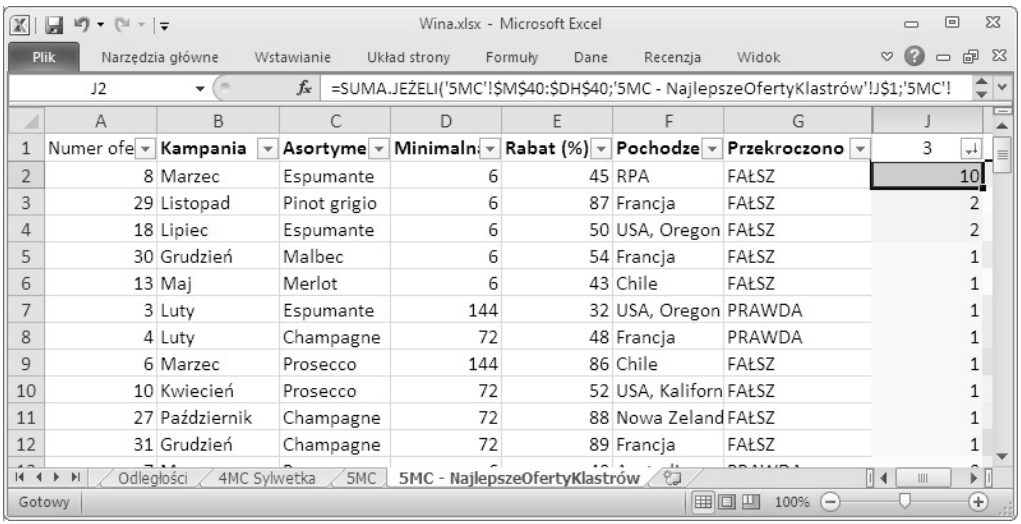
I2 =SUMA.JEŻELI('5MC'!$M$40:$DH$40;'5MC - NajlepszeOfertyKlastrów'!I$1;'5MC'!

	A	B	C	D	E	F	G	I
1	Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	2
2	30	Grudzień	Malbec	6	54	Francja	FAŁSZ	15
3	29	Listopad	Pinot grigio	6	87	Francja	FAŁSZ	13
4	7	Marzec	Prosecco	6	40	Australia	PRAWDA	12
5	18	Lipiec	Espumante	6	50	USA, Oregon	FAŁSZ	10
6	8	Marzec	Espumante	6	45	RPA	FAŁSZ	7
7	13	Maj	Merlot	6	43	Chile	FAŁSZ	5
8	24	Wrzesień	Pinot noir	6	34	Włochy	FAŁSZ	0
9	26	Październik	Pinot noir	144	83	Australia	FAŁSZ	0
10	17	Lipiec	Pinot noir	12	47	Niemcy	FAŁSZ	0
11	2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ	0
12	1	Styczeń	Malbec	72	56	Francja	FAŁSZ	0

Odległości | 4MC Sylwetka | 5MC | 5MC - NajlepszeOfertyKlastrów

Gotowy 100%

Rysunek 2.36. Sortowanie klastra 2. — osoby kupujące tylko małe ilości wina



Wina.xlsx - Microsoft Excel

Plik | Narzędzia główne | Wstawianie | Układ strony | Formuły | Dane | Recenzja | Widok

J2 =SUMA.JEŻELI('5MC'!$M$40:$DH$40;'5MC - NajlepszeOfertyKlastrów'!J$1;'5MC'!

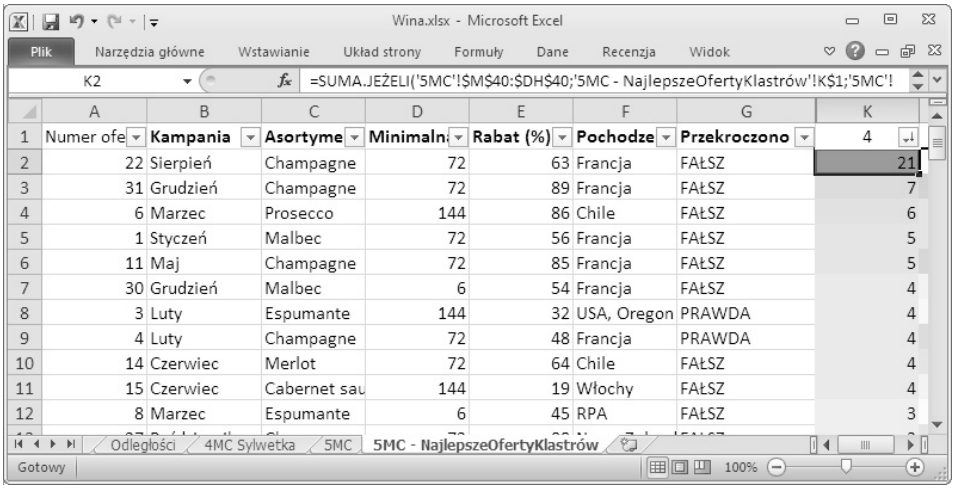
	A	B	C	D	E	F	G	J
1	Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	3
2	8	Marzec	Espumante	6	45	RPA	FAŁSZ	10
3	29	Listopad	Pinot grigio	6	87	Francja	FAŁSZ	2
4	18	Lipiec	Espumante	6	50	USA, Oregon	FAŁSZ	2
5	30	Grudzień	Malbec	6	54	Francja	FAŁSZ	1
6	13	Maj	Merlot	6	43	Chile	FAŁSZ	1
7	3	Luty	Espumante	144	32	USA, Oregon	PRAWDA	1
8	4	Luty	Champagne	72	48	Francja	PRAWDA	1
9	6	Marzec	Prosecco	144	86	Chile	FAŁSZ	1
10	10	Kwiecień	Prosecco	72	52	USA, Kaliforn	FAŁSZ	1
11	27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ	1
12	31	Grudzień	Champagne	72	89	Francja	FAŁSZ	1

Odległości | 4MC Sylwetka | 5MC | 5MC - NajlepszeOfertyKlastrów

Gotowy 100%

Rysunek 2.37. Sortowanie klastra 3. — czy wino espumante jest naprawdę tak ważne?

Osoby zakwalifikowane do klastra 4. preferują zakup dużych ilości win, które pochodzą głównie z Francji, a ich ceny charakteryzują się dużym rabatem. Zauważalna jest również pewna preferencja win musujących. Trudno odczytać informacje zawarte w tym klastrze — jest ich tak wiele (rysunek 2.38).



	A	B	C	D	E	F	G	K
1	Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	4
2	22	Sierpień	Champagne	72	63	Francja	FAŁSZ	21
3	31	Grudzień	Champagne	72	89	Francja	FAŁSZ	7
4	6	Marzec	Prosecco	144	86	Chile	FAŁSZ	6
5	1	Styczeń	Malbec	72	56	Francja	FAŁSZ	5
6	11	Maj	Champagne	72	85	Francja	FAŁSZ	5
7	30	Grudzień	Malbec	6	54	Francja	FAŁSZ	4
8	3	Luty	Espumante	144	32	USA, Oregon	PRAWDA	4
9	4	Luty	Champagne	72	48	Francja	PRAWDA	4
10	14	Czerwiec	Merlot	72	64	Chile	FAŁSZ	4
11	15	Czerwiec	Cabernet sau	144	19	Włochy	FAŁSZ	4
12	8	Marzec	Espumante	6	45	RPA	FAŁSZ	3

Rysunek 2.38. Sortowanie klastra 4. — zróżnicowane preferencje kupujących

Sortowanie klastra 5. daje podobne rezultaty do sortowania klastra 4., ale tym razem główne preferencje wydają się związane z dużą ilością i dużymi rabatami (rysunek 2.39).

Określanie sylwetki podziału na pięć klastrów

Zapewne zastanawiasz się, czy podział na pięć klastrów jest lepszy od podziału na cztery. Na pierwszy rzut oka nie widać większej różnicy. Obliczmy wartość sylwetki podziału na pięć klastrów i zobaczmy, co myśli o tym podziale komputer.

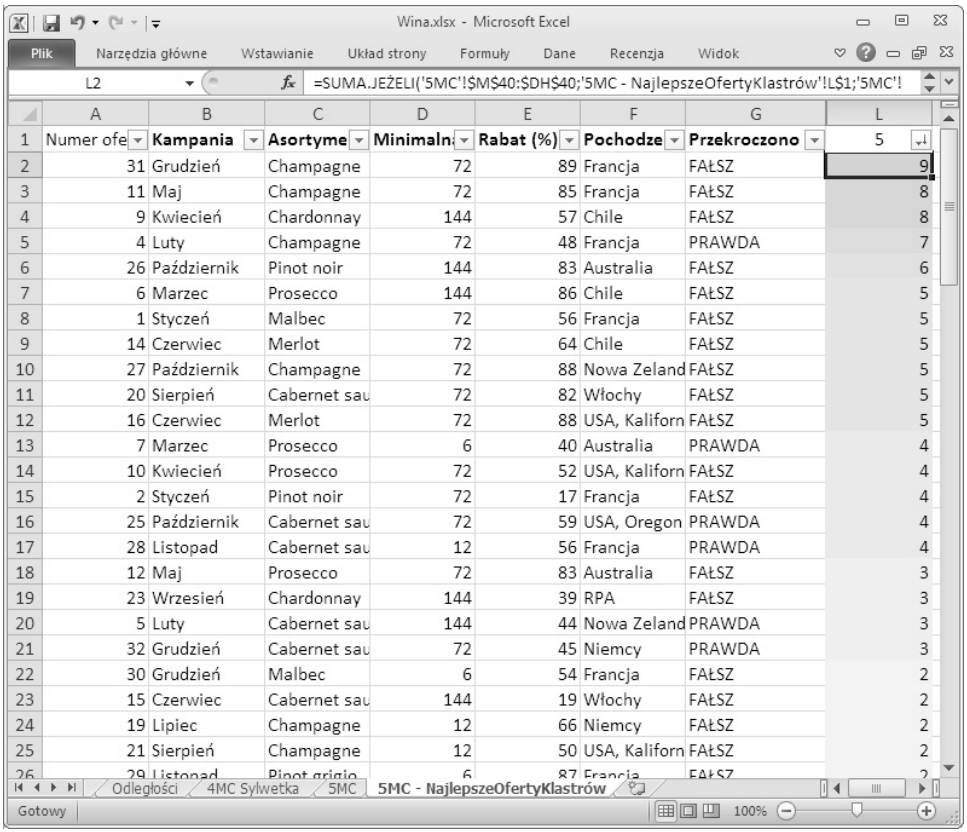
Zacznij od skopiowania zawartości arkusza *4MC Sylwetka* do nowego arkusza o nazwie `5MC Sylwetka`. Kliknij prawym przyciskiem myszy kolumnę `G` i wstaw nową kolumnę. Nadaj jej etykietę `Odległość od członków 5. grupy`. Przeciągnij formułę z komórki `F2` do komórki `G2`, zmień numer sprawdzanego klastra (z `4` na `5`), a następnie kliknij tę komórkę dwukrotnie, aby skopiować znajdującą się w niej formułę do wszystkich komórek tej kolumny znajdujących się poniżej.

Tak samo jak w poprzednim podrozdziale skorzystaj z funkcji *Znajdowanie i zamienianie* i zastąp `'4MC'!$L$39:$DG$39` następującym fragmentem formuły: `'5MC'!$M$40:$DH$40`.

W kolumnach `H2`, `I2` i `J2` powinieneś również wziąć pod uwagę odległość od osób przypisanych do klastra 5., a więc zakresy komórek kończące się na `F2` powinny być rozszerzone do `G2`. Po wykonaniu tych modyfikacji możesz zaznaczyć komórki `H2:J2` i kliknąć dwukrotnie ich prawy dolny róg w celu zaktualizowania formuł wpisanych do komórek znajdujących się niżej.

Na koniec skopiuj wartości przypisania klastra znajdujące się w 40. wierszu arkusza *5MC*. Umieść je w kolumnie `B` arkusza *5MC Sylwetka*. Skorzystaj z opcji *Wklej specjalnie*, zaznaczając opcję *Transponuj*.

Po wykonaniu tych modyfikacji Twój arkusz powinien wyglądać tak, jak pokazano na rysunku 2.40.



Wina.xlsx - Microsoft Excel

Plik | Narzędzia główne | Wstawianie | Układ strony | Formuły | Dane | Recenzja | Widok

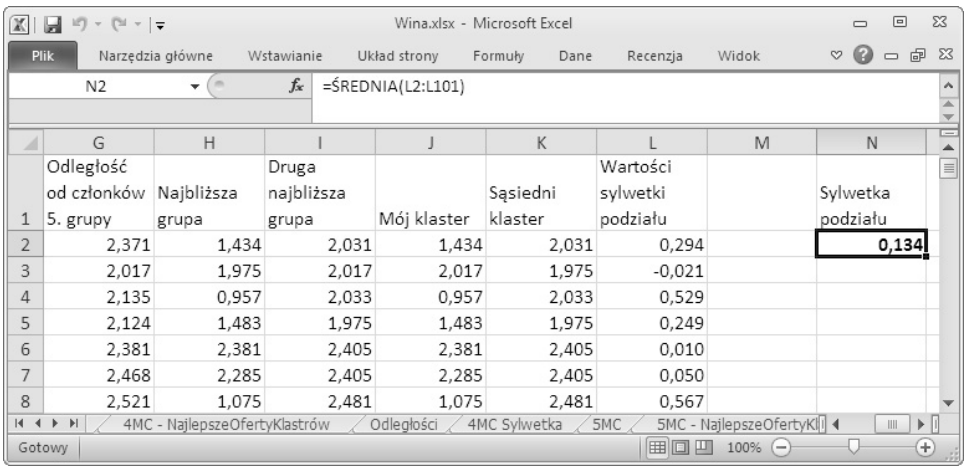
L2 =SUMA.JEŻELI('5MC'!$M$40:$DH$40;'5MC - NajlepszeOfertyKlastrów'!L$1;'5MC'!

	A	B	C	D	E	F	G	L
1	Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	5
2	31	Grudzień	Champagne	72	89	Francja	FAŁSZ	9
3	11	Maj	Champagne	72	85	Francja	FAŁSZ	8
4	9	Kwiecień	Chardonnay	144	57	Chile	FAŁSZ	8
5	4	Luty	Champagne	72	48	Francja	PRAWDA	7
6	26	Październik	Pinot noir	144	83	Australia	FAŁSZ	6
7	6	Marzec	Prosecco	144	86	Chile	FAŁSZ	5
8	1	Styczeń	Malbec	72	56	Francja	FAŁSZ	5
9	14	Czerwiec	Merlot	72	64	Chile	FAŁSZ	5
10	27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ	5
11	20	Sierpień	Cabernet sau	72	82	Włochy	FAŁSZ	5
12	16	Czerwiec	Merlot	72	88	USA, Kalifor	FAŁSZ	5
13	7	Marzec	Prosecco	6	40	Australia	PRAWDA	4
14	10	Kwiecień	Prosecco	72	52	USA, Kalifor	FAŁSZ	4
15	2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ	4
16	25	Październik	Cabernet sau	72	59	USA, Oregon	PRAWDA	4
17	28	Listopad	Cabernet sau	12	56	Francja	PRAWDA	4
18	12	Maj	Prosecco	72	83	Australia	FAŁSZ	3
19	23	Wrzesień	Chardonnay	144	39	RPA	FAŁSZ	3
20	5	Luty	Cabernet sau	144	44	Nowa Zeland	PRAWDA	3
21	32	Grudzień	Cabernet sau	72	45	Niemcy	PRAWDA	3
22	30	Grudzień	Malbec	6	54	Francja	FAŁSZ	2
23	15	Czerwiec	Cabernet sau	144	19	Włochy	FAŁSZ	2
24	19	Lipiec	Champagne	12	66	Niemcy	FAŁSZ	2
25	21	Sierpień	Champagne	12	50	USA, Kalifor	FAŁSZ	2
26	29	Listopad	Pinot grigio	6	87	Francja	FAŁSZ	2

Odległości | 4MC Sylwetka | 5MC | 5MC - NajlepszeOfertyKlastrów

Gotowy 100%

Rysunek 2.39. Sortowanie klastra 5. — preferowane są ilości hurtowe



Wina.xlsx - Microsoft Excel

Plik | Narzędzia główne | Wstawianie | Układ strony | Formuły | Dane | Recenzja | Widok

N2 =ŚREDNIA(L2:L101)

	G	H	I	J	K	L	M	N
1	Odległość od członków 5. grupy	Najbliższa grupa	Druga najbliższa grupa	Mój klaster	Sąsiedni klaster	Wartości sylwetki podziału		Sylwetka podziału
2	2,371	1,434	2,031	1,434	2,031	0,294		0,134
3	2,017	1,975	2,017	2,017	1,975	-0,021		
4	2,135	0,957	2,033	0,957	2,033	0,529		
5	2,124	1,483	1,975	1,483	1,975	0,249		
6	2,381	2,381	2,405	2,381	2,405	0,010		
7	2,468	2,285	2,405	2,285	2,405	0,050		
8	2,521	1,075	2,481	1,075	2,481	0,567		

4MC - NajlepszeOfertyKlastrów | Odległości | 4MC Sylwetka | 5MC | 5MC - NajlepszeOfertyKl

Gotowy 100%

Rysunek 2.40. Sylwetka podziału na pięć klastrów

Czy nie uważasz, że to smutne? Sylwetka podziału niemal wcale się nie zmieniła. Wartość *0,134* wskazuje, że podział na pięć grup jest nawet nieco gorszy! Nie jest to nic dziwnego. W obu przypadkach uzyskano trzy sensowne klastry, a pozostałe były zaszumione. Może obraliśmy zły kierunek i trzeba sprawdzić podział na trzy klastry? Jeżeli chcesz wypróbować ten podział, potraktuj to jako ćwiczenie i zrób to samodzielnie.

W kolejnym podrozdziale chciałbym zwrócić uwagę na coś, co być może powoduje generowanie zaszumionych i kłopotliwych klastrów.

Podział na grupy za pomocą algorytmu k-medioidów i asymetryczny pomiar odległości

W większości przypadków sprawdza się standardowy algorytm *k*-średnich i pomiar odległości euklidesowej, ale teraz trafiłeś na problem, który trapi osoby analizujące małe zbiory danych (zbiory takie są generowane przez handel, klasyfikację tekstu i bioinformatykę).

Podział na grupy za pomocą metody k-medioidów

Pierwszy problem jest dość oczywisty i wynika z tego, że środki klastrów są wartościami dziesiętnymi, mimo że wektory transakcji dokonanych przez klientów składają się z samych zer i jedynek. Co tak naprawdę oznacza 0,113 transakcji? Chciałbym, aby środki klastrów określały dokonanie lub niedokonanie transakcji.

Jeżeli zmodyfikujesz algorytm dzielący klientów na klastry, to możesz korzystać tylko z wartości znajdujących się w wektorach transakcji dokonanych przez klientów. Taką metodę nazywamy algorytmem ***k*-medioidów** (wcześniej korzystaliśmy z algorytmu ***k*-średnich**).

Gdybyś chciał w dalszym ciągu korzystać z odległości euklidesowej, to wystarczyłoby, abyś dodał w opcjach narzędzia Solver ograniczenie współrzędnych położenia środków klastrów do wartości binarnych (`binarna`).

Warto zastanowić się nad tym, co uzyskamy, wyrażając odległość euklidesową za pomocą wartości binarnych.

Stosowanie lepszego sposobu pomiaru odległości

Zwykle po przejściu z algorytmu *k*-średnich na algorytm *k*-medioidów nie korzysta się z metryki euklidesowej i przechodzi się na **metrykę miejską**, zwaną również **metryką Manhattanu**.

Wrona może przelecieć z punktu A do punktu B w linii prostej, ale taksówka na Manhattanie musi korzystać z siatki ulic, a więc może jechać tylko na północ, południe, wschód lub zachód. Na przedstawionym wcześniej przykładzie pomiaru odległości na dyskotece szkolnej odległość euklidesowa wynosiła 4,47 metra, ale odległość mierzona zgodnie z metryką miejską wynosi 6 metrów (4 metry w dół plus 2 metry w bok).

W przypadku danych binarnych, takich jak np. dane transakcji, odległość mierzona w metryce miejskiej jest odległością pomiędzy środkiem klastra a wektorem zakupów klienta będącą sumą rozbieżności. Jeżeli środek klastra przyjął wartość 0 i zakupy klienta również przyjęły wartość 0, to odległość w danym kierunku jest równa 0. W przypadku rozbieżnych wartości (0 i 1) odległość w danym kierunku wynosi 1. Po zsumowaniu odległości w poszczególnych kierunkach otrzymamy odległość całkowitą, która jest w zasadzie liczbą rozbieżności. Odległość miejską implementowaną podczas pracy z danymi binarnymi określa się często mianem **odległości Hamminga**.

Czy pomiar odległości zgodnie z metryką miejską rozwiązuje problem?

Zanim rozpoczniesz dzielenie klientów za pomocą algorytmu *k*-medioidów i metryki miejskiej, spójrz jeszcze raz na dane transakcji.

Co oznacza dokonanie przez klienta transakcji? Oznacza chęć nabycia produktu przez klienta.

Co oznacza niedokonanie transakcji? Czy oznacza niechęć zakupu produktu tak samo silną jak chęć zakupu produktu wyrażona przez wykonanie transakcji? Czy sygnał negacji jest tak samo silny jak wykonanie transakcji? Być może ktoś lubi szampana, ale ma już jego zapas. Może pewna grupa klientów nie przeczytała akurat treści newslettera. Istnieje wiele powodów, dla których ktoś nie wykonał jakiejś czynności, ale czynności zakupu są wykonywane z pewnych określonych powodów.

Innymi słowy: powinieneś analizować zakupy, a nie ich brak.

Można stwierdzić, że analizowane przez Ciebie dane są „asymetryczne" — jedynki są warte więcej niż zera. Jeżeli jakichś dwóch klientów jest podobnych z powodu trzech identycznych zakupów, znaczy to więcej niż podobieństwo innych klientów ustalone na podstawie identycznego niewykonania trzech zakupów. Dane, w których jedynki są ważne, ale rzadko występują w zgromadzonym zbiorze, określamy przymiotnikiem „rzadkie".

Zastanówmy się jeszcze, co oznacza to, że jakiś klient jest blisko środka klastra, z punktu widzenia metryki euklidesowej. Jeżeli mamy klienta z dużą liczbą jedynek dla jakiejś transakcji i dużą liczbą zer dla innej transakcji, to obie te informacje wpływają tak samo na odległość od środka klastra.

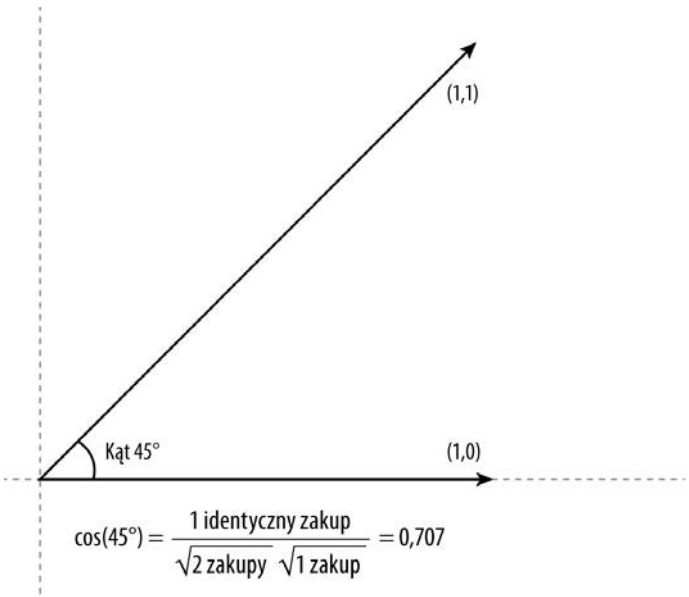
W opisywanym przykładzie potrzebujesz metody **asymetrycznego obliczania odległości**. Istnieje wiele takich metod, które można stosować w przypadku danych transakcji zapisanych za pomocą wartości binarnych.

Chyba najczęściej używaną metodą obliczania odległości asymetrycznej dla danych binarnych jest metoda **odległości kosinusowej**.

Odległość kosinusowa wcale nie jest czymś strasznym

Odległość kosinusową najłatwiej jest wyjaśnić za pomocą jej przeciwieństwa — podobieństwa kosinusowego.

Załóżmy, że dysponujesz dwoma binarnymi wektorami transakcji: (1,1) i (1,0). W pierwszym wektorze dokonano zakupu obu produktów, a w drugim dokonano zakupu tylko pierwszego produktu. Możesz dokonać wizualizacji tych wektorów w przestrzeni. Zobaczysz wtedy, że pomiędzy nimi powstaje kąt 45° (rysunek 2.41). Śmiało! Weź kątomierz i zmierz go.



Rysunek 2.41. Podobieństwo kosinusowe dwóch wektorów binarnych zawierających dane transakcji

Możemy więc stwierdzić, że podobieństwo kosinusowe tych wektorów wynosi cos(45°) = 0,707. Dlaczego? Okazuje się, że kosinus kąta pomiędzy dwoma binarnymi wektorami transakcji jest równy:

> *liczbie identycznych zakupów w obu wektorach podzielonej przez iloczyn pierwiastka kwadratowego z liczby zakupów pierwszego wektora i pierwiastka kwadratowego z liczby zakupów drugiego wektora.*

W przypadku wektorów (1,1) i (1,0) jeden zakup jest identyczny, a więc wartość 1 jest dzielona przez pierwiastek kwadratowy z 2 (wykonano dwie transakcje) pomnożony przez pierwiastek kwadratowy z 1 (zrealizowano jedną transakcję). Wykonując to działanie, uzyskasz wynik 0,707 (rysunek 2.41).

Dlaczego ten wynik jest tak interesujący?

Są ku temu trzy powody:

- Licznik bierze pod uwagę tylko liczbę identycznych dokonanych transakcji, a więc miara ta jest asymetryczna (szukaliśmy właśnie takiego mechanizmu).
- Dzieląc przez pierwiastek kwadratowy z liczby transakcji dokonanych w każdym wektorze, bierzesz pod uwagę to, że wektor, w którym dokonano *wszystkich możliwych zakupów* (nazwijmy go wektorem dostatku), jest bardziej oddalony od innego wektora niż wektor, w którym dokonano tych samych transakcji, ale nie wykonano wielu innych transakcji. Chcesz zgrupować wektory klientów o podobnych gustach, a nie znaleźć wektor, który zawiera wektor innego klienta.
- W przypadku danych binarnych wartości podobieństw znajdują się w przedziale od 0 do 1, a dwa wektory nie uzyskują wartości 1, o ile opisywane przez te wektory transakcje nie są identyczne. W związku z tym *1 – podobieństwo kosinusowe* można określić mianem odległości kosinusowej, która również przyjmuje wartość 0 lub 1.

Implementacja za pomocą Excela

Czas zaimplementować podział na grupy za pomocą techniki *k*-medioidów i obsługę odległości kosinusowej w Excelu.

UWAGA

Dzielenie na grupy za pomocą odległości kosinusowej jest określane również mianem **sferycznego algorytmu *k*-średnich**. W rozdziale 10. przyjrzysz się implementacji tego algorytmu w języku R.

Dla zachowania ciągłości przyjmijmy podział na pięć klastrów ($k = 5$).

Skopiuj arkusz *5MC*. Utworzoną kopię nazwij `5MedK`. Możesz skasować dane wygenerowane przez narzędzie Solver, ponieważ tym razem będziemy korzystać z binarnych danych definiujących środki klastrów.

Poza dodaniem ograniczenia do wartości binarnych w oknie narzędzia Solver musisz tylko zmodyfikować formuły określające odległość umieszczone w wierszach 34. – 38. Zacznij od komórki `M34`, w której ma się znaleźć odległość od klienta o nazwisku *Adams* i środka 1. klastra.

Aby policzyć liczbę wykonanych przez tego klienta transakcji, które pokrywają się z transakcjami opisującymi 1. klaster, musisz skorzystać z formuły `SUMA.ILOCZYNÓW` i policzyć sumę iloczynów dla komórek tych dwóch kolumn. Jeżeli obie porównywane komórki przyjmują wartość 0 lub obie te komórki przyjmują różne wartości, oznacza to, że porównywane dane transakcji różnią się, ale jeżeli w obu porównywanych komórkach znajdzie się wartość 1, to zostanie ona zsumowana przez funkcję `SUMA.ILOCZYNÓW` (w końcu 1 razy 1 to 1).

Pierwiastek kwadratowy z liczby transakcji, do których doszło w danym wektorze, można obliczyć za pomocą funkcji PIERWIASTEK przyjmującej jako argument sumę danego wektora (SUMA), a więc formuła obliczająca odległość ma następującą postać:

```
=1-SUMA.ILOCZYNÓW(M$2:M$33,$H$2:$H$33)/(PIERWIASTEK (SUMA(M$2:M$33))*PIERWIASTEK (SUMA($H$2:$H$33)))
```

Zauważ, że na początku formuły znajduje się zapis 1-, który zamienia podobieństwo kosinusowe na odległość kosinusową. Obliczenie tej odległości, w przeciwieństwie do odległości euklidesowej, nie wymaga zastosowania formuł tablicowych.

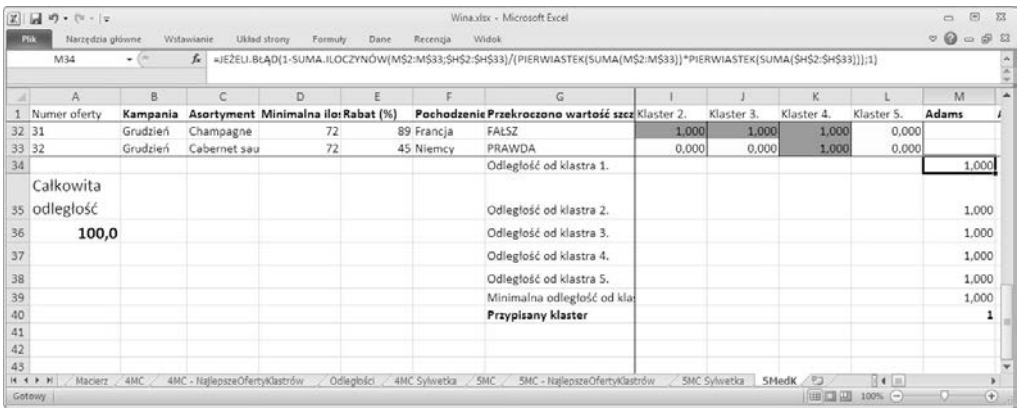
W komórce M34 powinieneś dodać jeszcze instrukcję sprawdzania błędu, na wypadek gdyby środek klastra był zdefiniowany za pomocą samych zer:

```
=JEŻELI.BŁĄD(1-SUMA.ILOCZYNÓW(M$2:M$33,$H$2:$H$33)/(PIERWIASTEK (SUMA(M$2:M$33))*PIERWIASTEK
(SUMA($H$2:$H$33)));1)
```

Dodanie formuły JEŻELI.BŁĄD zapobiega dzieleniu przez 0. Jeżeli z jakiegoś powodu Solver dobierze środki klastrów tak, że jeden z nich będzie opisywany przez same zera, to możesz założyć, że taki środek jest oddalony od wszystkiego o 1 (1 jest największą odległością, jaką możemy zadeklarować w systemie binarnym).

Teraz możesz skopiować zawartość komórki M34 aż do komórki M38, a następnie zmienić odwołania z kolumny H na odwołania do kolejnych kolumn (I, J, K i L). Podobnie jak w przypadku obliczania odległości euklidesowej możesz zastosować odwołania bezwzględne ($), dzięki czemu będziesz mógł przeciągnąć formuły na pozostałe komórki wiersza, a adresy kolumn klastrów nie będą zmieniane.

W ten sposób uzyskasz arkusz *5MedK* (rysunek 2.42), który jest bardzo podobny do utworzonego wcześniej arkusza *5MC*.



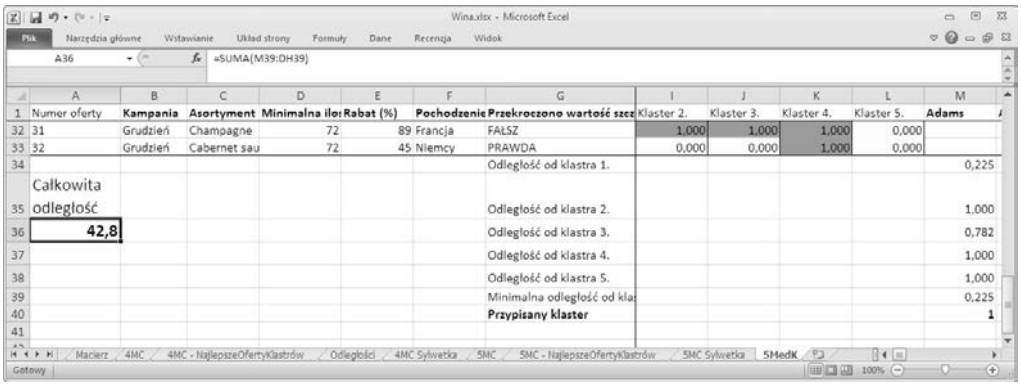
Rysunek 2.42. Arkusz 5MedK przed optymalizacją

W celu znalezienia środków klastrów otwórz narzędzie Solver i zmień ograniczenie <= 1 komórek H2:L33 na binarna.

Kliknij przycisk *Rozwiąż*. Ustalanie optymalnych klastrów może komputerowi zająć nawet pół godziny. Wygenerowane klastry będą przyjmowały tylko wartości binarne, a więc po sformatowaniu warunkowym komórki będą miały tylko dwa kolory.

Najlepsze oferty przy podziale na pięć klastrów za pomocą median

Po zakończeniu pracy narzędzia Solver uzyskasz parametry środków pięciu klastrów. Tym razem jedynki będą oznaczały oferty preferowane przez osoby przyporządkowane do danego klastra. Otrzymałem całkowitą odległość na poziomie *42,8*, ale wyniki uzyskane przez narzędzie Solver uruchomione na Twoim komputerze mogą oczywiście odbiegać od moich (rysunek 2.43).



Rysunek 2.43. Mediany pięciu klastrów

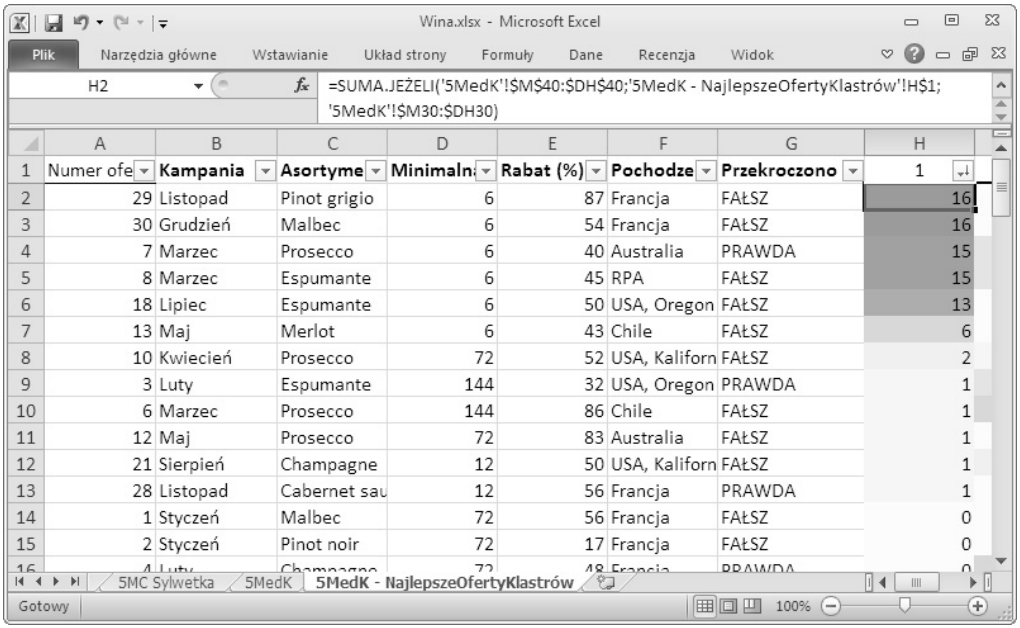
Przeanalizujmy te klastry tak jak w przypadku klastrów uzyskanych za pomocą algorytmu *k*-średnich. W tym celu skopiuj arkusz *5MC — NajlepszeOfertyKlastrów* i nazwij go `5MedK – NajlepszeOfertyKlastrów`.

W nowej zakładce otwórz okno *Znajdowanie i zamienianie* i zmień `5MC` na `5MedK`. Rozkład wierszy i kolumn jest w obu tych arkuszach identyczny, a więc formuły po modyfikacji odwołań będą działały poprawnie.

Korzystamy z algorytmu ewolucyjnego, zatem uzyskane przez Ciebie klastry mogą być nieco inne od moich, jeżeli chodzi o ich kolejność i skład, ale nie będą to z pewnością znaczące różnice. Czas przyjrzeć się klastrom utworzonym przez algorytm.

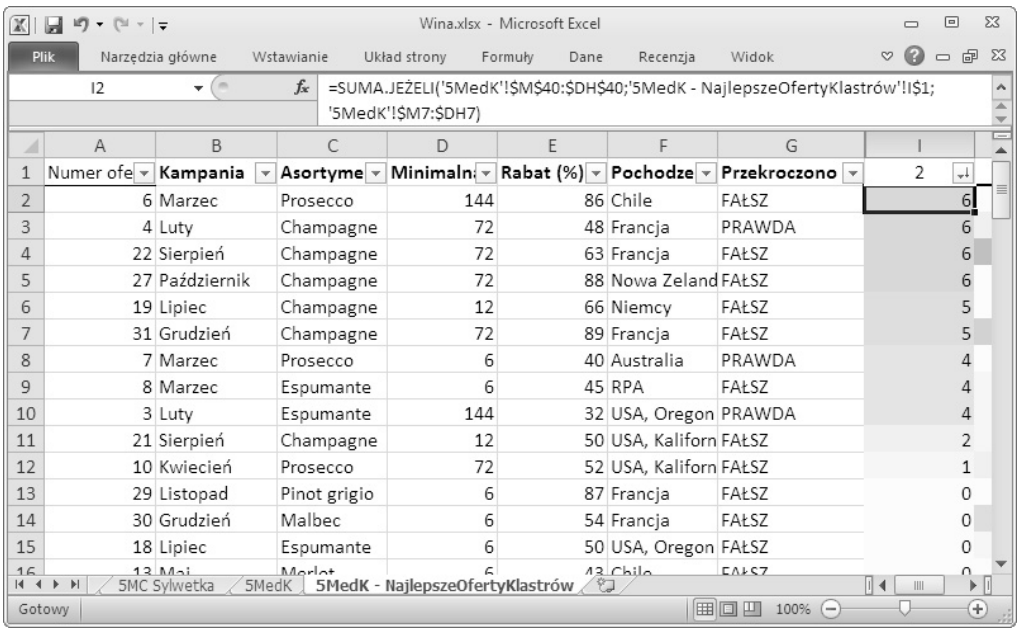
Po posortowaniu klastra 1. widać, że zawiera on klientów, którzy kupują małe ilości win (rysunek 2.44).

Do klastra 2. przypisano klientów, którzy kupują wina musujące — 11 najpopularniejszych ofert dotyczyło win takich jak champagne, prosecco i espumante (rysunek 2.45). Warto zauważyć, że grupowanie za pomocą algorytmu *k*-średnich (przy podziale na cztery i pięć grup) nie wygenerowało klastra miłośników wina musującego.



	A	B	C	D	E	F	G	H
1	Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	1
2	29	Listopad	Pinot grigio	6	87	Francja	FAŁSZ	16
3	30	Grudzień	Malbec	6	54	Francja	FAŁSZ	16
4	7	Marzec	Prosecco	6	40	Australia	PRAWDA	15
5	8	Marzec	Espumante	6	45	RPA	FAŁSZ	15
6	18	Lipiec	Espumante	6	50	USA, Oregon	FAŁSZ	13
7	13	Maj	Merlot	6	43	Chile	FAŁSZ	6
8	10	Kwiecień	Prosecco	72	52	USA, Kaliforn	FAŁSZ	2
9	3	Luty	Espumante	144	32	USA, Oregon	PRAWDA	1
10	6	Marzec	Prosecco	144	86	Chile	FAŁSZ	1
11	12	Maj	Prosecco	72	83	Australia	FAŁSZ	1
12	21	Sierpień	Champagne	12	50	USA, Kaliforn	FAŁSZ	1
13	28	Listopad	Cabernet sau	12	56	Francja	PRAWDA	1
14	1	Styczeń	Malbec	72	56	Francja	FAŁSZ	0
15	2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ	0

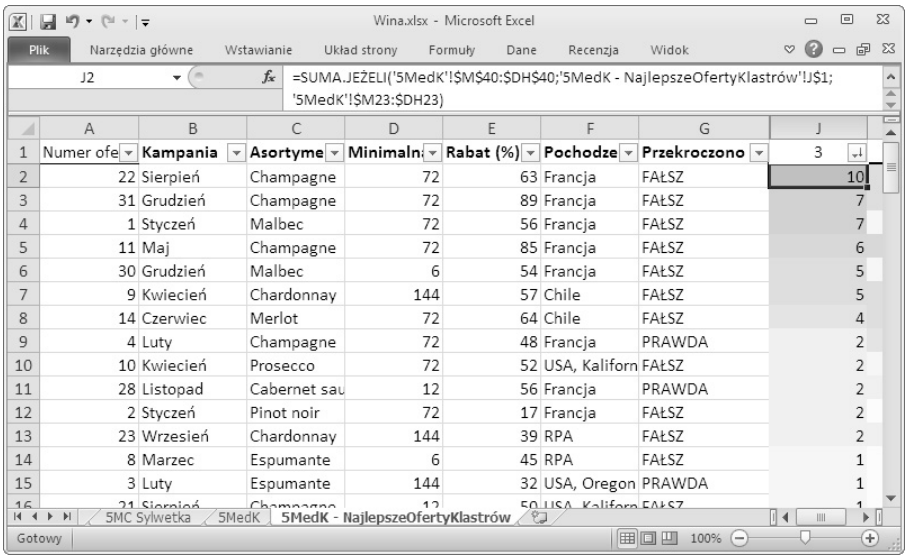
Rysunek 2.44. Sortowanie klastra 1. — klienci kupujący małe ilości wina



	A	B	C	D	E	F	G	I
1	Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	2
2	6	Marzec	Prosecco	144	86	Chile	FAŁSZ	6
3	4	Luty	Champagne	72	48	Francja	PRAWDA	6
4	22	Sierpień	Champagne	72	63	Francja	FAŁSZ	6
5	27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ	6
6	19	Lipiec	Champagne	12	66	Niemcy	FAŁSZ	5
7	31	Grudzień	Champagne	72	89	Francja	FAŁSZ	5
8	7	Marzec	Prosecco	6	40	Australia	PRAWDA	4
9	8	Marzec	Espumante	6	45	RPA	FAŁSZ	4
10	3	Luty	Espumante	144	32	USA, Oregon	PRAWDA	4
11	21	Sierpień	Champagne	12	50	USA, Kaliforn	FAŁSZ	2
12	10	Kwiecień	Prosecco	72	52	USA, Kaliforn	FAŁSZ	1
13	29	Listopad	Pinot grigio	6	87	Francja	FAŁSZ	0
14	30	Grudzień	Malbec	6	54	Francja	FAŁSZ	0
15	18	Lipiec	Espumante	6	50	USA, Oregon	FAŁSZ	0

Rysunek 2.45. Sortowanie klastra 2. — miłośnicy win musujących

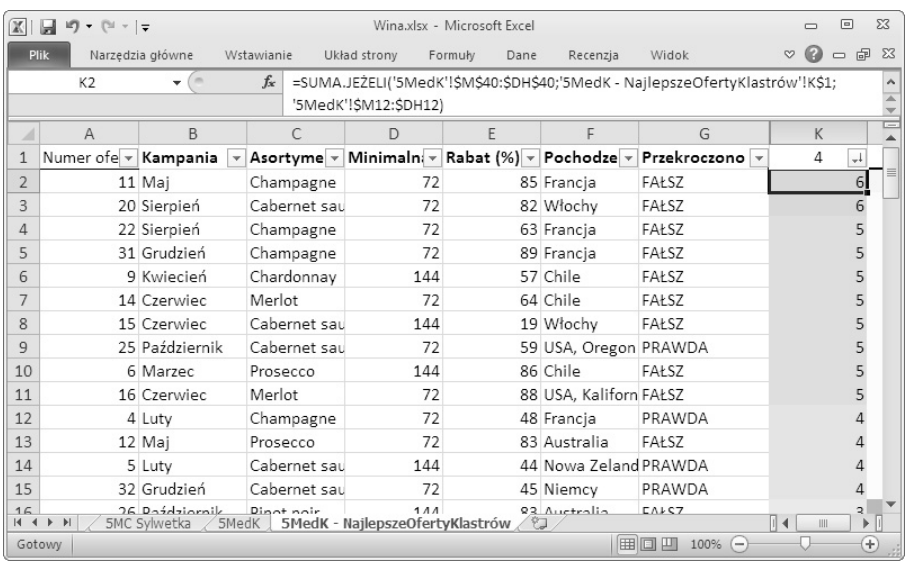
Klaster 3. zrzesza frankofilów. Pięć najpopularniejszych ofert dotyczyło win pochodzących z Francji (rysunek 2.46). Czy oni naprawdę nie wiedzą, że najlepsze wina produkuje się w Kalifornii?



Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	3
22	Sierpień	Champagne	72	63	Francja	FAŁSZ	10
31	Grudzień	Champagne	72	89	Francja	FAŁSZ	7
1	Styczeń	Malbec	72	56	Francja	FAŁSZ	7
11	Maj	Champagne	72	85	Francja	FAŁSZ	6
30	Grudzień	Malbec	6	54	Francja	FAŁSZ	5
9	Kwiecień	Chardonnay	144	57	Chile	FAŁSZ	5
14	Czerwiec	Merlot	72	64	Chile	FAŁSZ	4
4	Luty	Champagne	72	48	Francja	PRAWDA	2
10	Kwiecień	Prosecco	72	52	USA, Kalifor	FAŁSZ	2
28	Listopad	Cabernet sau	12	56	Francja	PRAWDA	2
2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ	2
23	Wrzesień	Chardonnay	144	39	RPA	FAŁSZ	2
8	Marzec	Espumante	6	45	RPA	FAŁSZ	1
3	Luty	Espumante	144	32	USA, Oregon	PRAWDA	1

Rysunek 2.46. Sortowanie klastra 3. — frankofile

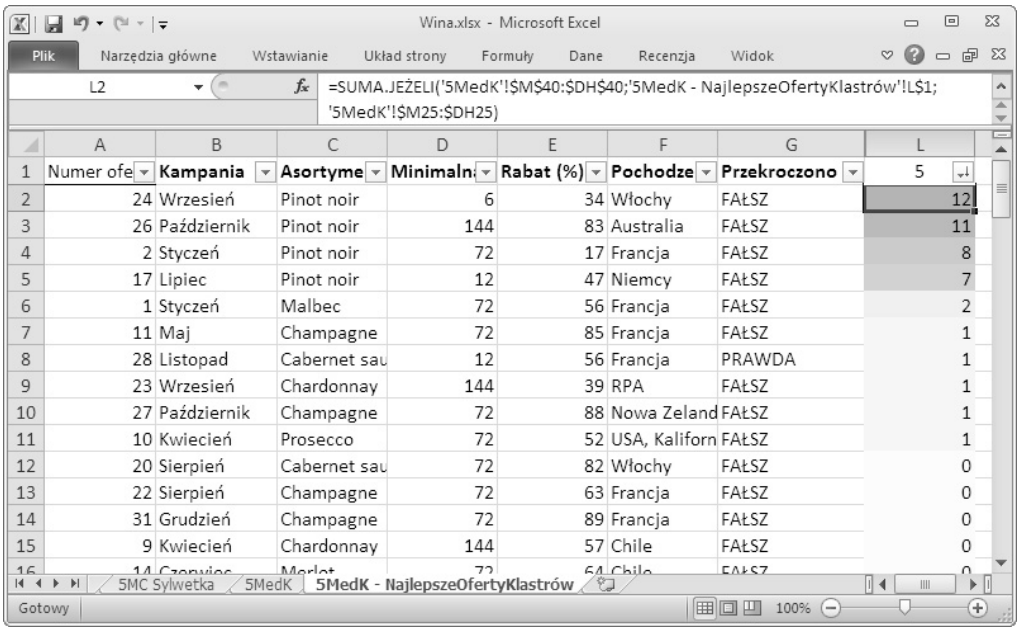
Wszystkie transakcje w klastrze 4. dotyczą sprzedaży dużej ilości wina. Najpopularniejsze oferty charakteryzuje duży rabat i to, że cena tych win ma tendencję wzrostową (rysunek 2.47).



Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	4
11	Maj	Champagne	72	85	Francja	FAŁSZ	6
20	Sierpień	Cabernet sau	72	82	Włochy	FAŁSZ	6
22	Sierpień	Champagne	72	63	Francja	FAŁSZ	5
31	Grudzień	Champagne	72	89	Francja	FAŁSZ	5
9	Kwiecień	Chardonnay	144	57	Chile	FAŁSZ	5
14	Czerwiec	Merlot	72	64	Chile	FAŁSZ	5
15	Czerwiec	Cabernet sau	144	19	Włochy	FAŁSZ	5
25	Październik	Cabernet sau	72	59	USA, Oregon	PRAWDA	5
6	Marzec	Prosecco	144	86	Chile	FAŁSZ	5
16	Czerwiec	Merlot	72	88	USA, Kalifor	FAŁSZ	5
4	Luty	Champagne	72	48	Francja	PRAWDA	4
12	Maj	Prosecco	72	83	Australia	FAŁSZ	4
5	Luty	Cabernet sau	144	44	Nowa Zeland	PRAWDA	4
32	Grudzień	Cabernet sau	72	45	Niemcy	PRAWDA	4

Rysunek 2.47. Sortowanie klastra 4. — 19 najpopularniejszych transakcji dotyczyło dużych ilości wina

Klaster 5. ponownie zrzesza osoby kupujące wino pinot noir (rysunek 2.48).



	A	B	C	D	E	F	G	L
1	Numer ofe	Kampania	Asortyme	Minimaln	Rabat (%)	Pochodze	Przekroczono	5
2	24	Wrzesień	Pinot noir	6	34	Włochy	FAŁSZ	12
3	26	Październik	Pinot noir	144	83	Australia	FAŁSZ	11
4	2	Styczeń	Pinot noir	72	17	Francja	FAŁSZ	8
5	17	Lipiec	Pinot noir	12	47	Niemcy	FAŁSZ	7
6	1	Styczeń	Malbec	72	56	Francja	FAŁSZ	2
7	11	Maj	Champagne	72	85	Francja	FAŁSZ	1
8	28	Listopad	Cabernet sau	12	56	Francja	PRAWDA	1
9	23	Wrzesień	Chardonnay	144	39	RPA	FAŁSZ	1
10	27	Październik	Champagne	72	88	Nowa Zeland	FAŁSZ	1
11	10	Kwiecień	Prosecco	72	52	USA, Kalifor	FAŁSZ	1
12	20	Sierpień	Cabernet sau	72	82	Włochy	FAŁSZ	0
13	22	Sierpień	Champagne	72	63	Francja	FAŁSZ	0
14	31	Grudzień	Champagne	72	89	Francja	FAŁSZ	0
15	9	Kwiecień	Chardonnay	144	57	Chile	FAŁSZ	0

Rysunek 2.48. Sortowanie klastra 5. — najpopularniejszym winem jest pinot noir

Tym razem uzyskałeś bardziej czytelne wyniki. Stało się tak, ponieważ zastosowałeś algorytm *k*-medioidów i mierzyłeś odległość w sposób asymetryczny — obliczałeś odległość kosinusową. Dzięki tym rozwiązaniom posegregowałeś klientów na podstawie ich zainteresowań, a nie tego, co ich nie interesuje, i o to właśnie nam chodziło.

Metoda obliczania odległości ma duży wpływ na uzyskane klastry.

Teraz możesz dokonać importu podziału klientów do serwisu MailChimp.com i stworzyć spersonalizowane wersje newslettera skierowane do klientów przyporządkowanych do określonych klastrów. Powinno to pomóc w lepszym dotarciu do kupujących i doprowadzić do zwiększenia sprzedaży.

Podsumowanie

W tym rozdziale opisałem wiele praktycznych rzeczy. Przyjrzałeś się:

- odległości euklidesowej;
- optymalizacji centroidów za pomocą algorytmu *k*-średnich i narzędzia Solver;
- procesowi analizy uzyskanych klastrów;
- obliczaniu sylwetki podziału na daną liczbę klastrów;
- dzieleniu na klastry za pomocą algorytmu *k*-medioidów;

- odległości mierzonej zgodnie z metryką miejską (metryką Manhattanu) — odległości Hamminga;
- mierze odległości opartej na podobieństwie kosinusowym.

Jeżeli przebrnąłeś przez ten rozdział, to powinieneś umieć dzielić dane na klastry, a także określać problemy biznesowe, które można rozwiązać za pomocą grupowania. Dodatkowo nauczyłeś się przygotowywać dane do dzielenia na klastry.

Dzielenie na klastry za pomocą algorytmu centroidów (*k*-średnich) jest używane od kilku dziesięcioleci. Analizę danych klientów warto zacząć od segmentacji za pomocą tej metody. Niestety, nie jest to najbardziej „współczesna" metoda grupowania. W rozdziale 5. opiszę zastosowanie teorii grafów do znajdowania podobieństw klientów w tym samym zbiorze danych. Ponadto wyjdę na chwilę poza Excela i dokonam wizualizacji danych.

Jeżeli chcesz rozwijać swoją umiejętność korzystania z algorytmu *k*-średnich, to pamiętaj, że narzędzie Solver dostępne w Excelu może pracować tylko z 200 zmiennymi decyzyjnymi, a więc warto, żebyś zaczął używać lepszego nieliniowego narzędzia Solver (np. z wersji Premium Solver oferowanej przez serwis *www.solver.com*). Możesz również zacząć pracować w nieliniowej wersji Solvera, dostępnej w pakiecie LibreOffice, która umożliwia dzielenie wielowymiarowych danych na dużą liczbę grup.

Większość narzędzi statystycznych umożliwia analizę skupień. W języku R jest to funkcja `skmeans()`, aczkolwiek możliwości pakietu *fastcluster* (zawiera on m.in. algorytm *k*-medioidów i zestaw różnych funkcji przeznaczonych do obliczania odległości) sprawiają, że korzysta się z niego częściej. W rozdziale 10. opiszę zastosowanie pakietu *skmeans* do wykonania sferycznego algorytmu *k*-średnich.

Skorowidz

E

F

G

H

I

J

K

L

M

N

O

P

R

S

Ś

T

U

W

Z

WYCIŚNIJ Z DANYCH KAŻDĄ KROPLĘ WIEDZY!

Nauka o danych, znana również pod nazwą data science, jest stosunkowo nową, interdyscyplinarną dziedziną, obejmującą techniki analizy danych oraz zagadnienia związane z ich implementacją i wykorzystaniem do różnych celów. Zalety nauki o danych doceniają specjaliści z wielu branż: analitycy biznesowi, statystycy, architekci oprogramowania i osoby zajmujące się sztuczną inteligencją. Właściwie ta dziedzina nie koncentruje się na kodowaniu i bazach danych, raczej na metodach wyłuskiwania z nich najróżniejszych cennych informacji. Wartość tej wiedzy niejednokrotnie okazuje się ogromna.

Niniejsza książka stanowi przystępne wprowadzenie do nauki o danych. Jest przeznaczona dla osób, które chcą stosować techniki analizy danych w biznesie. Te techniki, opisane na podstawie praktycznych przypadków, to m.in. optymalizacja, prognozowanie i symulacja, a także sztuczna inteligencja, teoria grafów, analiza skupień i wykrywanie anomalii. Dzięki lekturze nie tylko zrozumiesz zasady analizowania danych, ale i nauczysz się wybierać technikę właściwą do rozwiązania danego problemu. Poznasz też techniki pracy z prototypami. Co ciekawe, niemal wszystkie opisane tu metody zostały zaprezentowane w arkuszu kalkulacyjnym.

W książce opisano m.in.:

— optymalizację za pomocą programowania liniowego i całkowitoliczbowego
— szereg czasowy, wykrywanie trendów i wahań sezonowych
— przewidywanie za pomocą wygładzania wykładniczego
— metodę symulacji Monte Carlo
— test Tukeya i lokalne czynniki odstające
— język R — zaawansowane techniki analizy danych

John W. Foreman — jest głównym analitykiem danych w MailChimp. Udziela również porad dotyczących analizy danych takim podmiotom jak Coca-Cola czy InterContinental Hotels, a także amerykańskim agendom rządowym, w tym DoD, IRS, DHS i FBI. Często wygłasza prelekcje o rozwiązaniach analitycznych w biznesie.

Helion

księgarnia internetowa
http://helion.pl
zamówienia telefoniczne
0 801 339900
0 601 339900

Helion SA
ul. Kościuszki 1c, 44-100 Gliwice
tel.: 32 230 98 63
e-mail: helion@helion.pl
http://helion.pl

Sprawdź najnowsze promocje:
http://helion.pl/promocje
Książki najchętniej czytane:
http://helion.pl/bestsellery
Zamów informacje o nowościach:
http://helion.pl/nowosci

Informatyka w najlepszym wydaniu

sięgnij po WIĘCEJ
KOD KORZYŚCI

ISBN 978-83-283-3357-4

cena: 77,00 zł