Projekt Statystyka w Biznesie - Wszechstronna analiza klientów

Marek Polit 121564

Spis treści

Cel Raportu

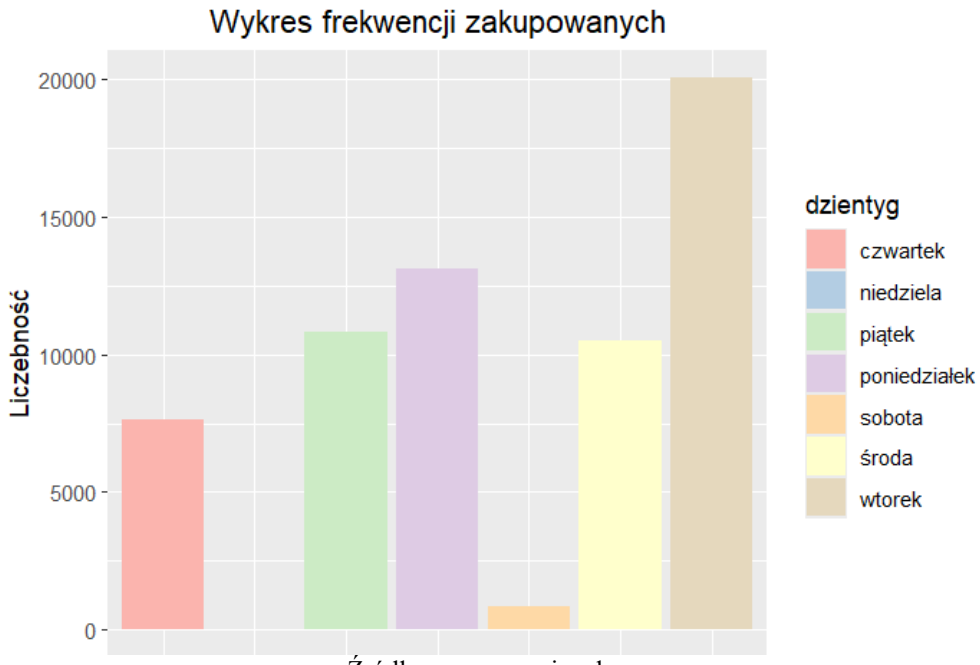
Celem niniejszego raportu jest przeanalizowanie zbioru danych pewnej firmy pod względem informacji o klientach oraz ich preferencjach. W raporcie dokonałem szczegółowej analizy bazy danych o koszykach klientów, wykorzystując narzędzia takie jak analiza Pareto, analiza koszykowa, czy segmentacja klientów. Dodatkowo, zidentyfikowałem najpopularniejsze wzorce zakupowe klientów, co pozwoliło na lepsze zrozumienie ich zachowań konsumenckich, z możliwością predykcji przyszłych koszyków zakupowych. Wykonana przeze mnie analiza pozwali na lepsze dostosowanie polityki biznesowej, tak aby poprawić wyniki finansowe firmy.

Opis Problemu

Używając języka R, przeanalizowałem zbiór danych o koszykach konsumenckich z kryteriami takimi jak m.in.: id_konsumenta, data zakupu, ilość i wartość pozycji w koszyku, czy kategoria zakupywanego produktu. Zbiór danych zawiera ponad 80 tys. koszyków konsumenckich z czego 17 tys. z nich zostało wyrzuconych ze względu na niekompletność, bądź niedostosowanie danych. Ponadto, ze względu na niską obecność zakupywanych produktów o kategoriach z literami od G do V, w niektórych z tych analiz te produkty zostały pogrupowane razem, co mogło istotnie wpłynąć na wyniki mojej analizy.

Wprowadzenie

Wykres 1. Frekwencja zakupowa konsumentów



Źródło: opracowanie własne

Przeprowadzona przeze mnie analiza klientów na podstawie ponad 60 tys. transakcji zakupowych z maja 2016 roku ukazuje, że spośród dni roboczych najbardziej popularnym

dniem robienia przez klientów zakupów był wtorek z około 20 tys. zaobserwowanych transakcji tego dnia. Z drugiej strony najmniej preferowanym dniem roboczym był czwartek, przy czym każda z ilości transakcji w dniach roboczych miała wartość przekraczającą przynajmniej 7 tys. obserwacji. W przypadku dni weekendowych sobota odnotowała porównywalnie dużo niższą liczbę obserwacji od dni roboczych na poziomie około 1000 obserwacji, z kolei w niedzielę sklep był prawdopodobnie nieczynny, gdyż na ten dzień nie przypada żadna obserwacja.

Grafika 1. Wordcloud dla rozmieszczenie sprzedaży według województwa

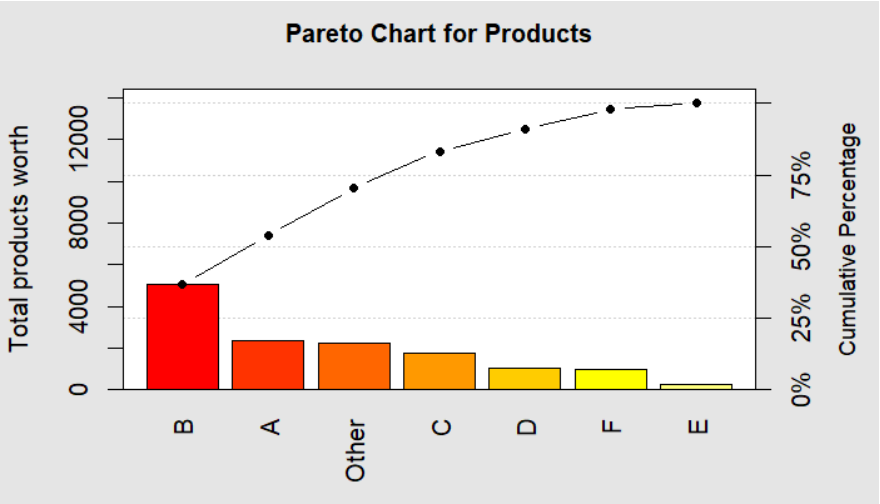
Źródło: opracowanie własne

Grafika Wordcloud dla rozmieszczenia sprzedaży z podziałem na regiony geograficzne dokonanych sprzedaży ukazuje, że wszystkie firma sprzedawała swoje produkty we wszystkich prowincjach, z czego największy odsetek zakupów dokonano w województwach takich jak: mazowieckie (10 tys. obs.), wielkopolskie (10 tys. obs.) i śląskie (8 tys. obs.). Najgorszym województwem pod względem liczebności sprzedaży okazało się województwo świętokrzyskie, gdzie liczba transakcji nie przekroczyła 1 tysiąca.

Wyniki analizy Pareto

Analiza Pareto dla grup produktów

Wykres 2. Pareto dla produktów

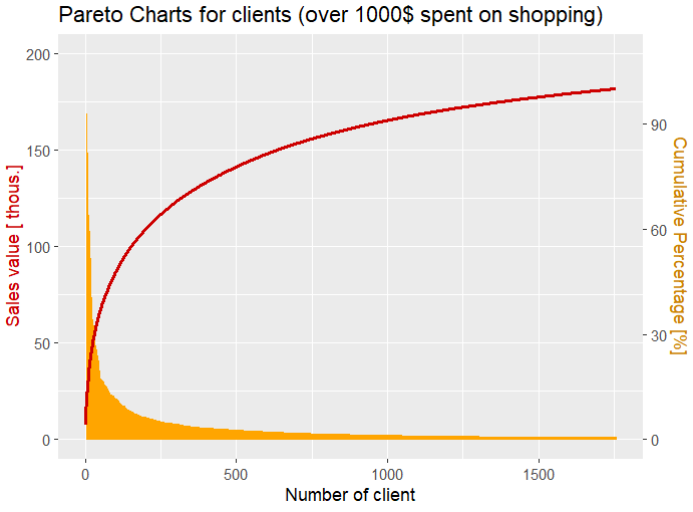


Źródło: opracowanie własne

Wykres Pareto dla pogrupowanych produktów według pierwszej litery nazwy pokazuje, że najbardziej dochodowym produktem dla firmy jest produkt kategorii B, który odpowiada za ponad 30% całkowitego dochodu z firmy. Produkt kategorii A ma podobne znaczenie w kontekście dochodu finansowego co produkty zgrupowane razem od G do V, które na wykresie widnieją pod nazwą Other. Najmniej dochodowymi produktami, które firma powinna przeanalizować pod względem rentowności są produkty kategorii E, które generują około 2% przychodów firmy. W przypadku grup produktów teoria Vilfred'a Pareto o proporcji 20/80 produktów do przychodu nie ma rzeczywistego odzwierciedlenia w przypadku tej firmy.

Analiza Pareto dla klientów

Wykres 3. Pareto dla zakupów z podziałem na klientów

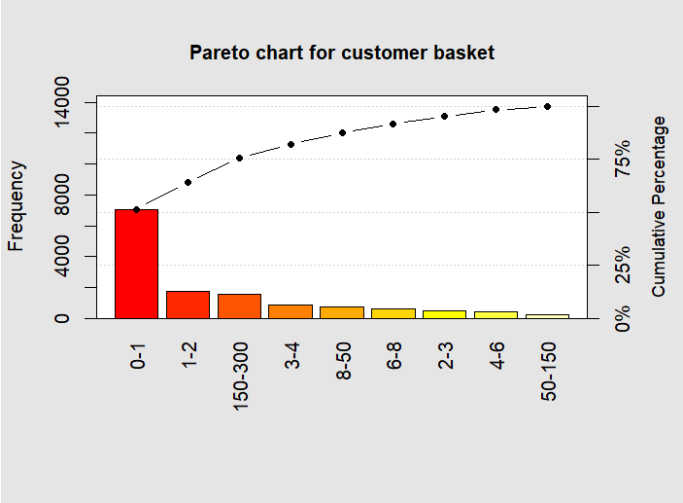


Źródło: opracowanie własne

Z powodu na dużą wariancję wartości zakupów dla klientów tej firmy, wykres został ograniczony na osi X do numerów klientów, których skumulowana wartość transakcji w tej firmie przekroczyła próg 1000 USD. Wykres Pareto dla zmiennej skumulowanej wartości zakupów z podziałem na klientów jest bardziej stromy w tym przypadku w porównaniu do poprzedniego. Całkowita skumulowana wartość zakupów klientów od klienta zaindeksowanego jako 0 do klienta znajdującego się w 20 percentylu wykresu wynosi 70.74% co oznacza, że 20% najbardziej dochodowych klientów odpowiada za 70% przychodów firmy i w tym przypadku teza Pareto ma lepsze odzwierciedlenie.

Analiza Pareto dla wartości koszyka konsumenckiego

Wykres 4. Pareto dla koszyka konsumenckiego

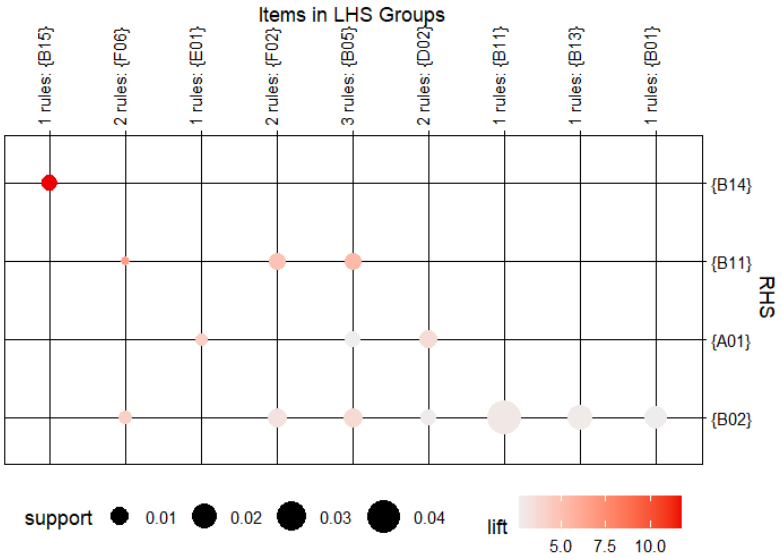


Źródło: opracowanie własne

Grafika przedstawia wykres Pareto dla koszyków konsumenckich z podziałem (na osi X) na różne przedziały wartości zakupowych w tysiącach. Zdecydowana większość wszystkich transakcji odpowiadająca za ponad 50% dochodów firmy to były transakcje nie przekraczające 1 tys. USD. Transakcje w przedziałach od 1-2 tys. USD. i 150-300 tys. USD. stanowiły dochody dla firmy na poziomie 4 mln USD. dla każdego z tych przedziałów i odpowiadały około 12% dochodów z pośród wszystkich kategorii. Najmniej dochodowe koszyki zakupowe dla firmy należały do przedziałów 2-3 tys. 4-6 tys. oraz 50-150 tys. USD. W ostatnim z omawianych przypadków wykres skumulowanej wartości procentowej jest bardziej wypłaszczony, lecz wartości zaczynając od pierwszego koszyka zaczynają się dosyć wysoko. Podsumowując wszystkie analizy Pareto, teza Vilfred'a okazuje się mieć najlepsze odzwierciedlenie w przypadku analizy konsumentów tej firmy.

Analiza reguł decyzyjnych dla klientów

Grafika 2. Reguły decyzyjne dla klientów



Źródło: opracowanie własne

Wykres bąbelkowy ilustruje wyniki analizy asocjacyjnej, ukazując reguły decyzyjne między grupami elementów z lewej strony (LHS) a grupami z prawej strony (RHS). Generalna większość produktów ma silne powiązanie pod względem wartości suport z produktem {B02}. Najbardziej znacząca reguła na wykresie to {B11} -> {B02}, charakteryzuje się największym wsparciem (support), lecz relatywnie niższą siłą asocjacji (lift). Inne reguły asocjacyjne przypisują również często innym koszykom produkty {A01} i {B11}. Reguła, która ma niższą wartość support na poziomie 9‰, lecz najwyższą wartość lift przekraczającą 10, to reguła {B15} -> {B14}.

Segmentacja klientów

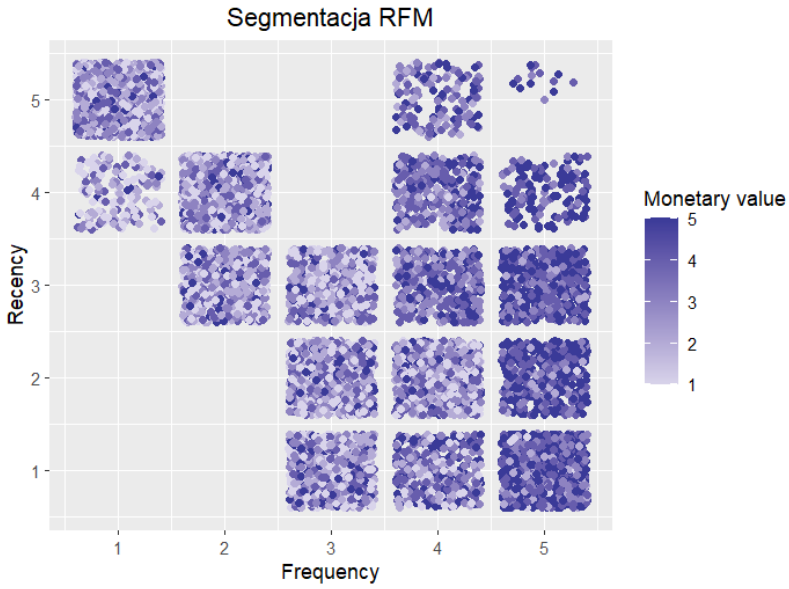
Segmentacja RFM

Segmentacja RFM (Recency, Frequency, Monetary) to technika marketingowa stosowana do analizy zachowań klientów na podstawie trzech kluczowych wskaźników: Recency, Frequency oraz Monetary. Recency odnosi się do ostatniego czasu zakupu, czyli jak dawno temu klient dokonał zakupu. Frequency to częstotliwość zakupów, czyli jak często klient dokonuje zakupów w określonym okresie. Monetary to wartość zakupów, czyli ile pieniędzy klient wydał w określonym czasie.

Wybrałem metodologię RFM ze względu na jej łatwość w interpretowalności i efektywność, co pozwoliło mi na uzyskanie wartościowych informacji o klientach. Metoda jest bezpośrednio powiązana z lojalnością i wartością klienta, co umożliwia na lepsze zrozumienie i

przewidywanie zachowań zakupowych. Ponadto, metodologia ta jest także skalowalna i może być stosowana dla różnych przedsiębiorstw, od małych firm po duże korporacje.

Wykres 5. Segmentacja klientów RFM



Źródło: opracowanie własne

Segmentacja RFM pokazała, że zmienna nie zawsze klienci, którzy robią zakupy często, robili je relatywnie niedawno. Widać to szczególnie w przypadku kategorii 5 dla zmiennej Frequency, gdzie wraz ze wzrostem zmiennej Recency, liczba obserwacji malała. Ponadto, wykres ten przedstawia, że generalna większość klientów, którzy robią zakupy o największych wartościach należą również do najwyższej kategorii pod względem częstotliwości zakupów, co z pewnością jest dobrą informacją dla przedsiębiorstwa. Obserwacje o niższych wartościach Monetary Value najczęściej były obserwowane wśród klientów, którzy robili zakupy relatywnie niedawno, co może oznaczać, że byli jednorazowymi gośćmi tej firmy. Najbardziej pożądani klienci tej firmy posiadający najwyższą wartość każdego z omawianych parametrów stanowili mniejszość całego odsetka klientów.

Ocena Segmentacji RFM

Tabela 1. Silhouette Score

Cluster	Size	Average Silhouette Width
1	73	0,42
2	4	0,29
3	16	0,53
4	10052	0,92
5	404	0,43

Źródło: opracowanie własne

Na podstawie analizy skupień przeprowadziłem analizę klastrów z podziałem na ich rozmiar i średnią szerokość. Klastry 2, 4 i 3 mają niższe średnie szerokości silhouette, co oznacza że obserwacje w tych klastrach mogą być mniej jednoznacznie przypisane do swoich klastrów niż w pozostałych klastrach. Klastry 1 i 5 mają średnio wyższe szerokości silhouette, w szczególności klaster 5 mający najwyższą wartość na poziomie 0.92, co sugeruje, że obserwacje w tym klastrze są dobrze zdefiniowane i oddzielone od innych klastrów. Podział na klastry obejmuje klastry o bardzo różnych rozmiarach, na przykład klastry 2 i 4 są znacznie mniejsze niż klastry 3 i 5.W przypadku tego podziału średnia wartość Silhouette Score wyniosła 0.89. Różnice w każdym z klastrów najprawdopodobniej odzwierciedlają naturalne zróżnicowanie w danych.

Podsumowanie

Przedstawiony raport opisuje szczegółową analizę danych koszyków zakupowych klientów firmy, wykorzystując narzędzia takie jak analiza Pareto, analiza koszykowa oraz segmentacja klientów. Analiza wykazała, że najbardziej dochodowymi produktami są te z kategorii B, a najważniejsi klienci generują większość przychodów. Wnioskiem jest konieczność skoncentrowania się na tych kluczowych klientach oraz dostosowanie oferty produktowej do ich preferencji w celu zwiększenia rentowności firmy. Dodatkowo, analiza skupień potwierdziła różnorodność w danych, sugerując potrzebę dalszego badania i dostosowywania strategii biznesowej do różnych segmentów klientów.