Analiza struktury sprzedaży i segmentacja klientów na podstawie danych z systemu transakcyjnego przedsiębiorstwa

Projekt z przedmiotu Statystyka w Biznesie

Mateusz Dziendziel 129745

Warszawa, maj 2025

Spis treści

Wstęp – krótkie wprowadzenie do analizy

W niniejszym raporcie przedstawiam analizę danych sprzedażowych pochodzących z systemu transakcyjnego przedsiębiorstwa. Zbiór danych zawiera informacje na temat liczby oraz wartości produktów (SKU) zakupionych przez klientów detalicznych (PSD) w określonym przedziale czasowym.

Celem analizy jest lepsze zrozumienie struktury sprzedaży oraz przygotowanie segmentacji klientów na podstawie ich zachowań zakupowych. Chcę sprawdzić, czy możliwe jest wyróżnienie grup klientów o podobnych potrzebach i preferencjach, a następnie opisać cechy charakterystyczne dla poszczególnych segmentów.

Tego typu segmentacja może okazać się przydatna przy lepszym dopasowaniu oferty, usprawnieniu działań marketingowych oraz podejmowaniu trafniejszych decyzji operacyjnych. Analizę przeprowadzono w środowisku R, a wyniki zaprezentowano w formie wykresów i tabel.

1. Opis zbioru danych

Zbiór danych obejmuje informacje dotyczące 9795 unikalnych klientów (punktów sprzedaży detalicznej) oraz wartości zakupów dokonanych przez nich na poszczególne produkty (SKU). Dla wygody kolumna identyfikująca klientów została oznaczona jako client_ID (zamiast oryginalnego oznaczenia PSD).

Początkowo w analizie uwzględniano 15 produktów, jednak po wstępnym oczyszczeniu danych pozostawiono 10 SKU. Produkty SKU3 i SKU6 zostały usunięte, ponieważ zawierały wyłącznie zera w całych kolumnach, co uniemożliwiało ich dalsze wykorzystanie. Ponadto, ze względu na bardzo niską sprzedaż, z analizy wykluczono również produkty SKU4, SKU7 oraz SKU10 — ich udział był znikomy i mógł zaburzać wyniki segmentacji.

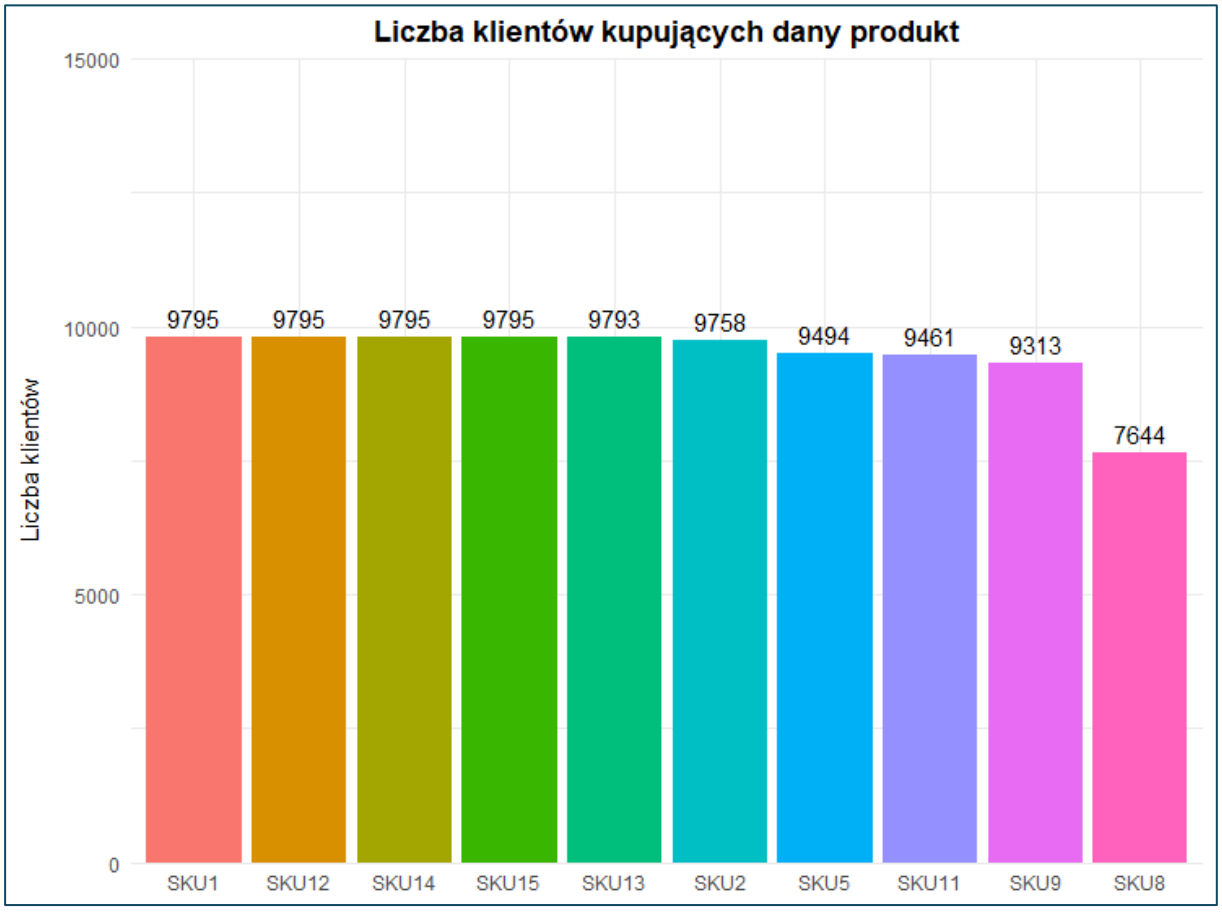
2. Podstawowa analiza danych

2.1 Przedstawienie podstawowych statystyk

We wstępnej analizie skupiłem się na ogólnym obrazie aktywności zakupowej klientów. Łączna wartość zakupów dokonanych przez wszystkich klientów wyniosła około 1,086 mld zł. Przeciętny klient wydał średnio 111 tys. zł, podczas gdy mediana wydatków była niższa i wyniosła 75,86 tys. zł — co sugeruje obecność klientów o znacznie wyższych wydatkach. Największa kwota wydana przez pojedynczego klienta sięgnęła 2,1 mln zł (ID klienta: 820).

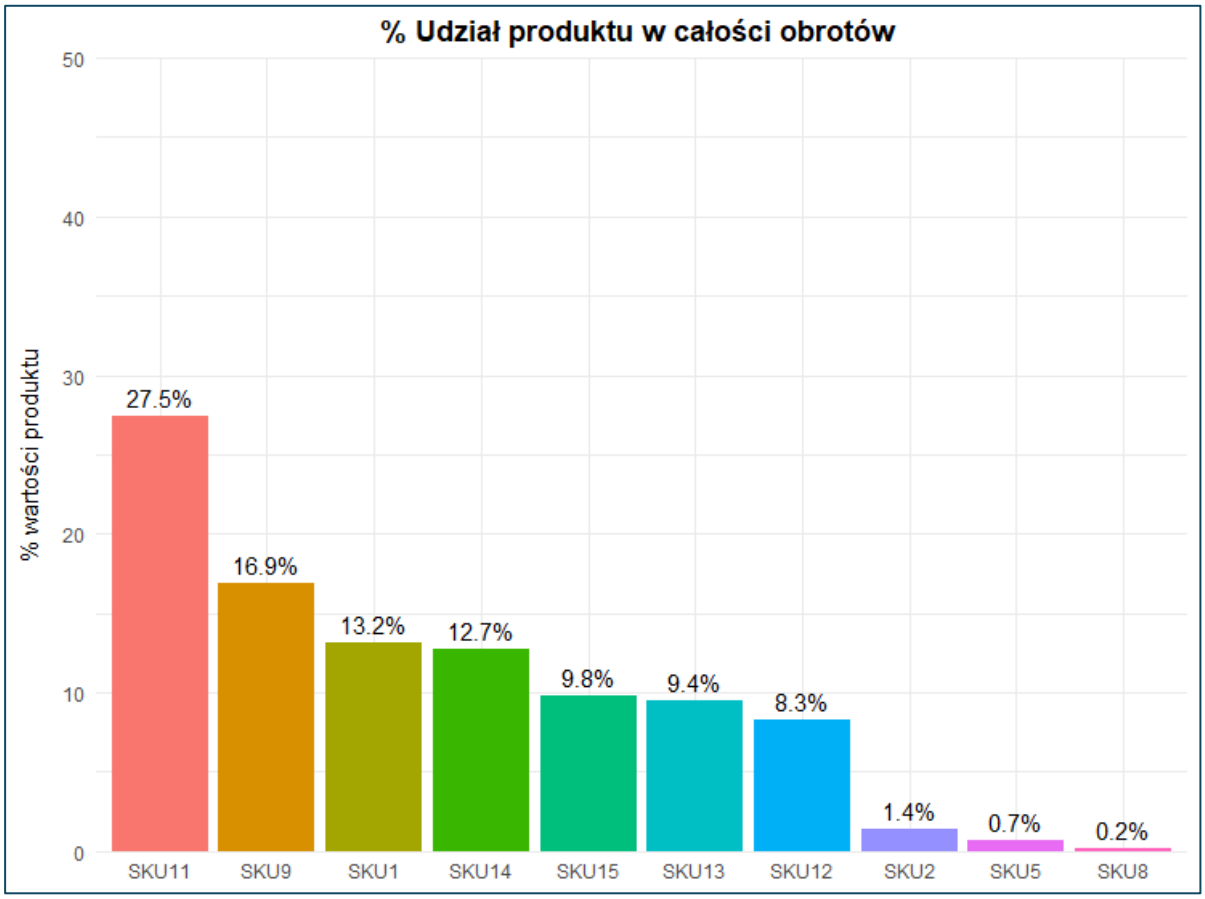
Pod względem szerokości koszyka zakupowego klienci kupowali średnio 9,67 różnych produktów. Minimalna liczba unikalnych SKU wyniosła 7, a maksymalna 10, co świadczy o zróżnicowanym profilu klientów — od bardziej wyspecjalizowanych po tych o szerszym zakresie zakupów.

Dokładna liczba klientów kupujących poszczególne produkty przedstawiona jest poniżej. Widać, że SKU1, SKU12, SKU14 i SKU15 są nabywane praktycznie przez wszystkich, natomiast SKU8 zdecydowanie odstaje pod względem liczby klientów.



Wykres 1. Liczba klientów kupujących dany produkt

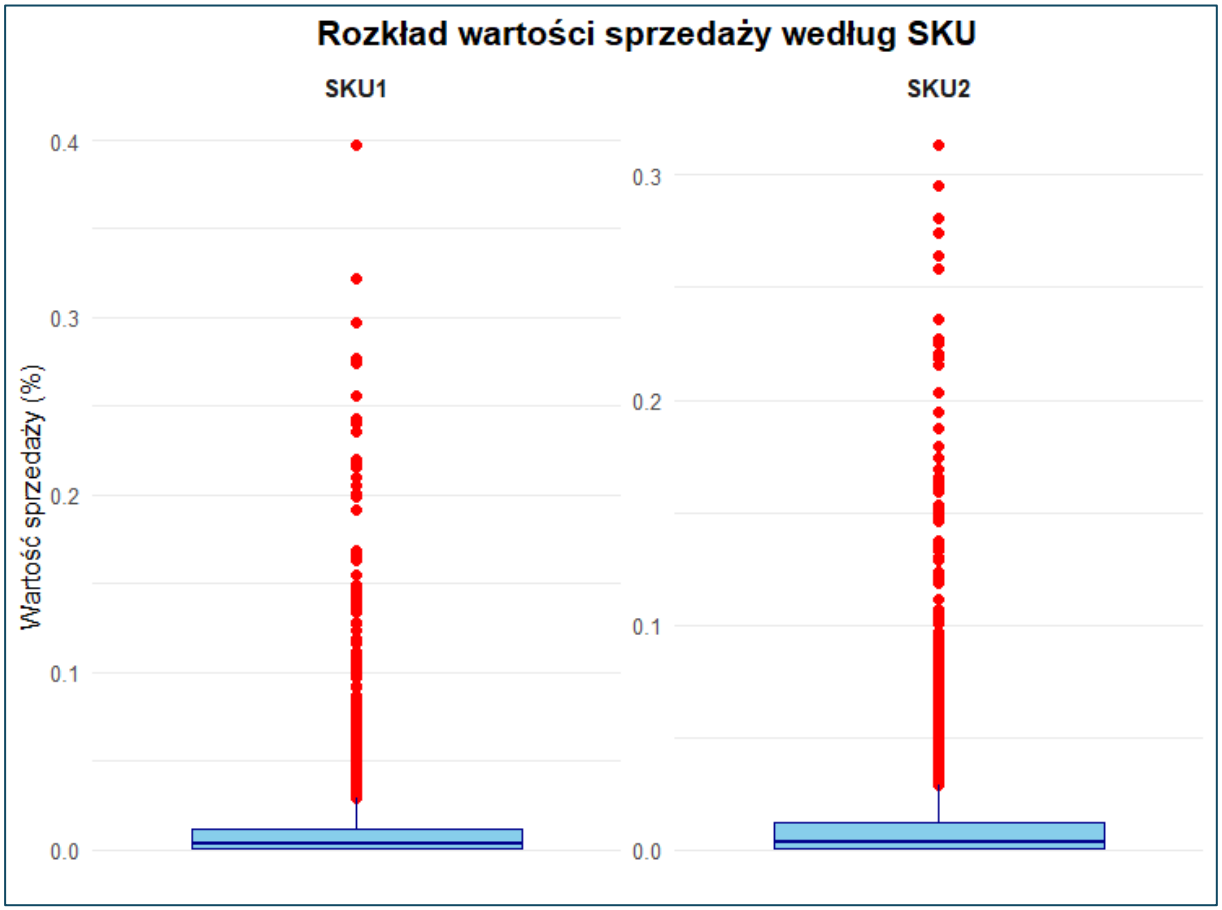
Warto też zauważyć, jaki procent całkowitego obrotu przypada na poszczególne produkty. Szczegóły pokazuje wykres poniżej. Widać wyraźnie, że SKU11 wyróżnia się na tle innych, odpowiadając za ponad jedną czwartą całkowitej sprzedaży. Z kolei produkty SKU2, SKU5 i SKU8 mają udział poniżej 3%.



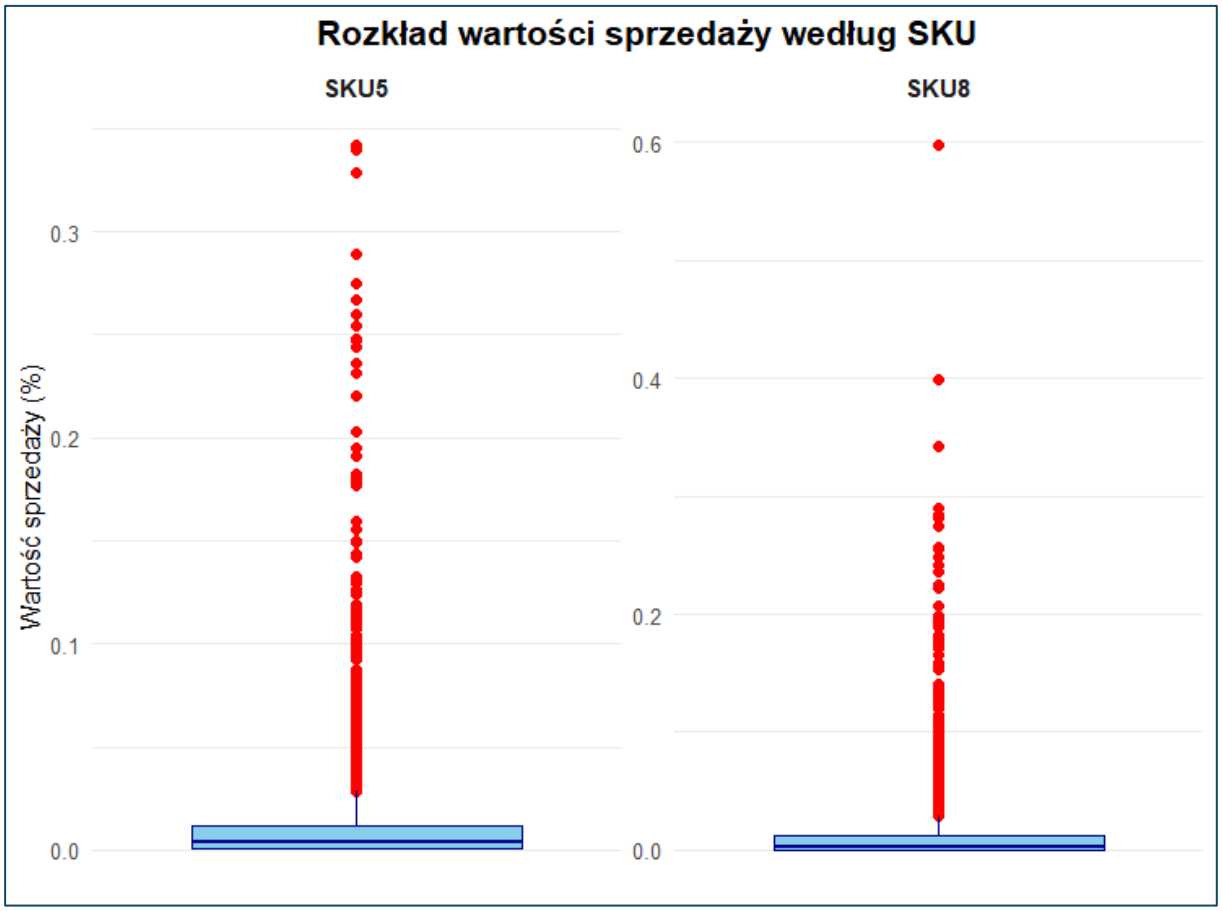
Wykres 2. Procentowy udział danego produktu w całości obrotów przedsiębiorstwa

Ważnym elementem analizy jest też poznanie rozkładu udziału poszczególnych produktów w zakupach klientów. Dzięki temu lepiej zrozumiemy wstępną strukturę klientów. W przypadku każdego produktu około 75% klientów ma bardzo niski udział – zazwyczaj w granicach 0–0,01%. To pokazuje duże zróżnicowanie klientów i szeroką bazę zakupową.

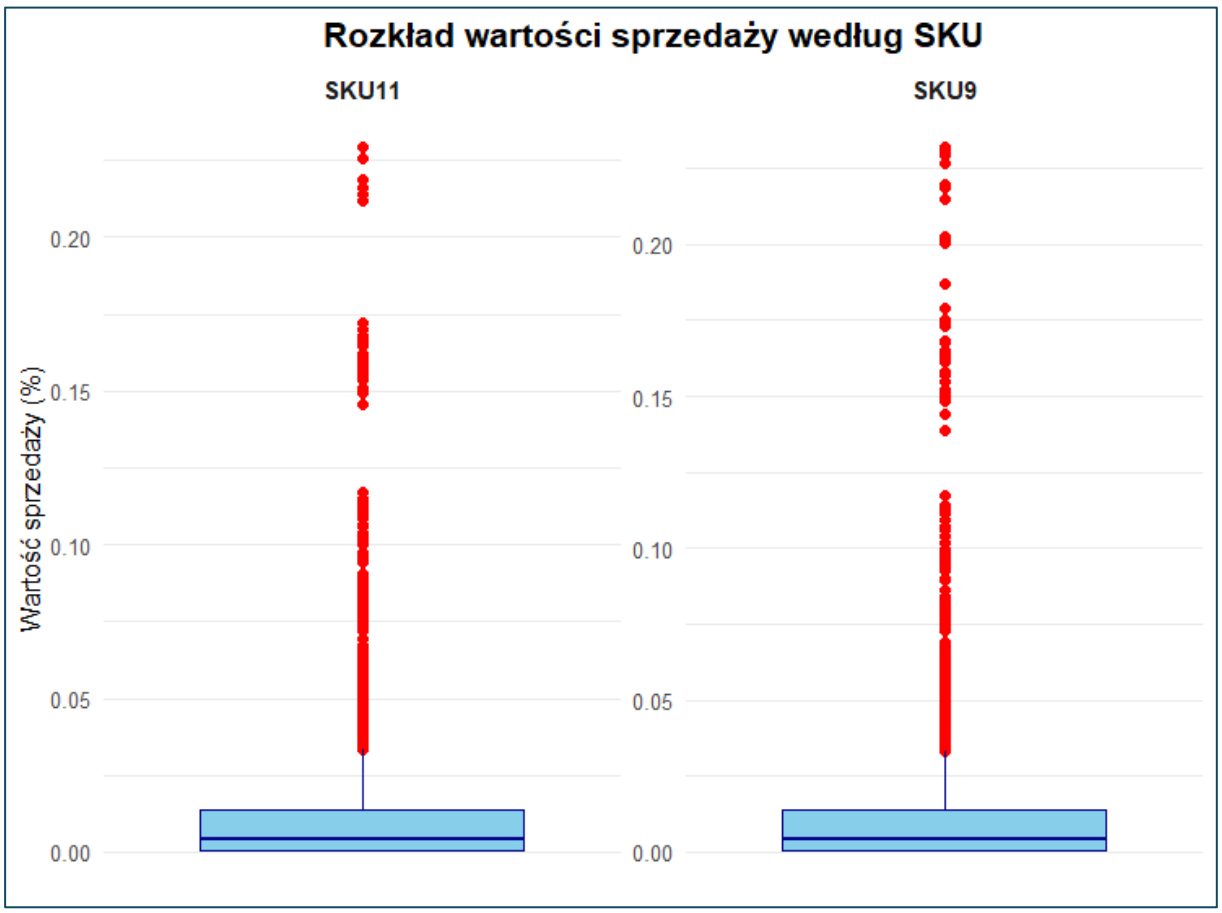
Dodatkowo widać, że niektórzy klienci wyraźnie odstają nawet od reszty outlierów, zwłaszcza przy produktach SKU1, SKU8, SKU12 i SKU15. Mogą to być tzw. klienci specjalni, działający na indywidualnych warunkach. Z braku pełnych informacji na ten temat, na razie zostawiamy ich w analizie bez wyróżnienia



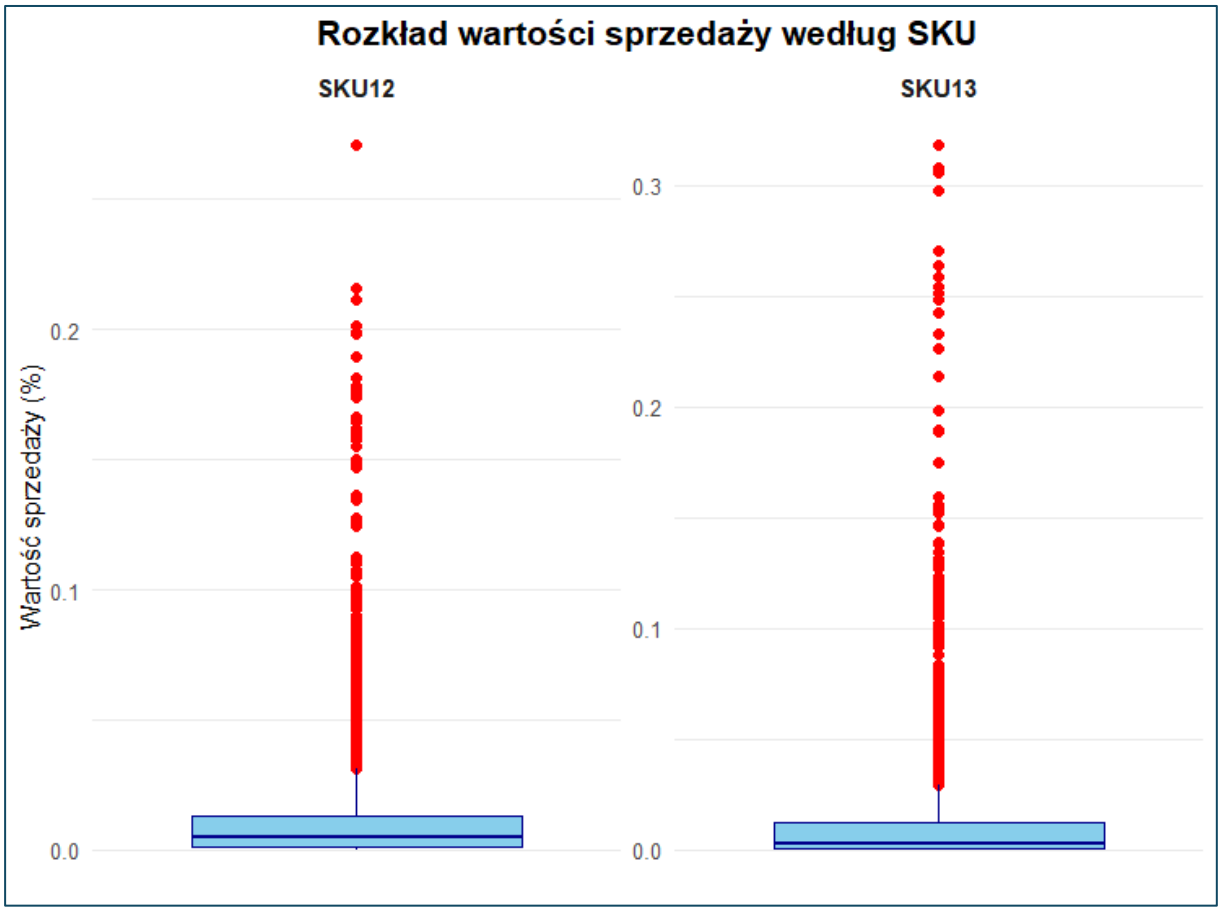
Wykres 3. Rozkład wartości sprzedaży produktów SKU1 i SKU2



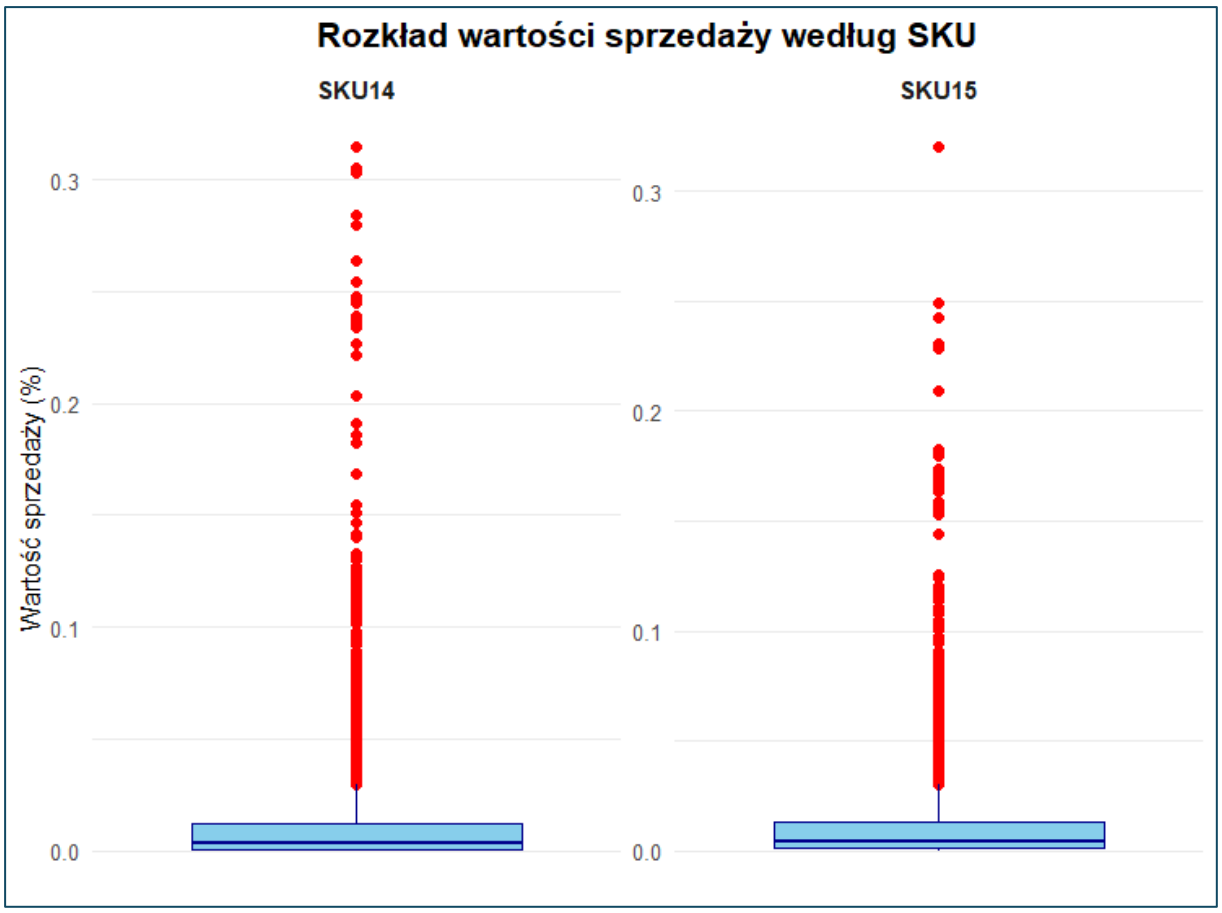
Wykres 4. Rozkład wartości sprzedaży produktów SKU5 i SKU8



Wykres 5. Rozkład wartości sprzedaży produktów SKU9 i SKU11



Wykres 6. Rozkład wartości sprzedaży produktów SKU12 i SKU13

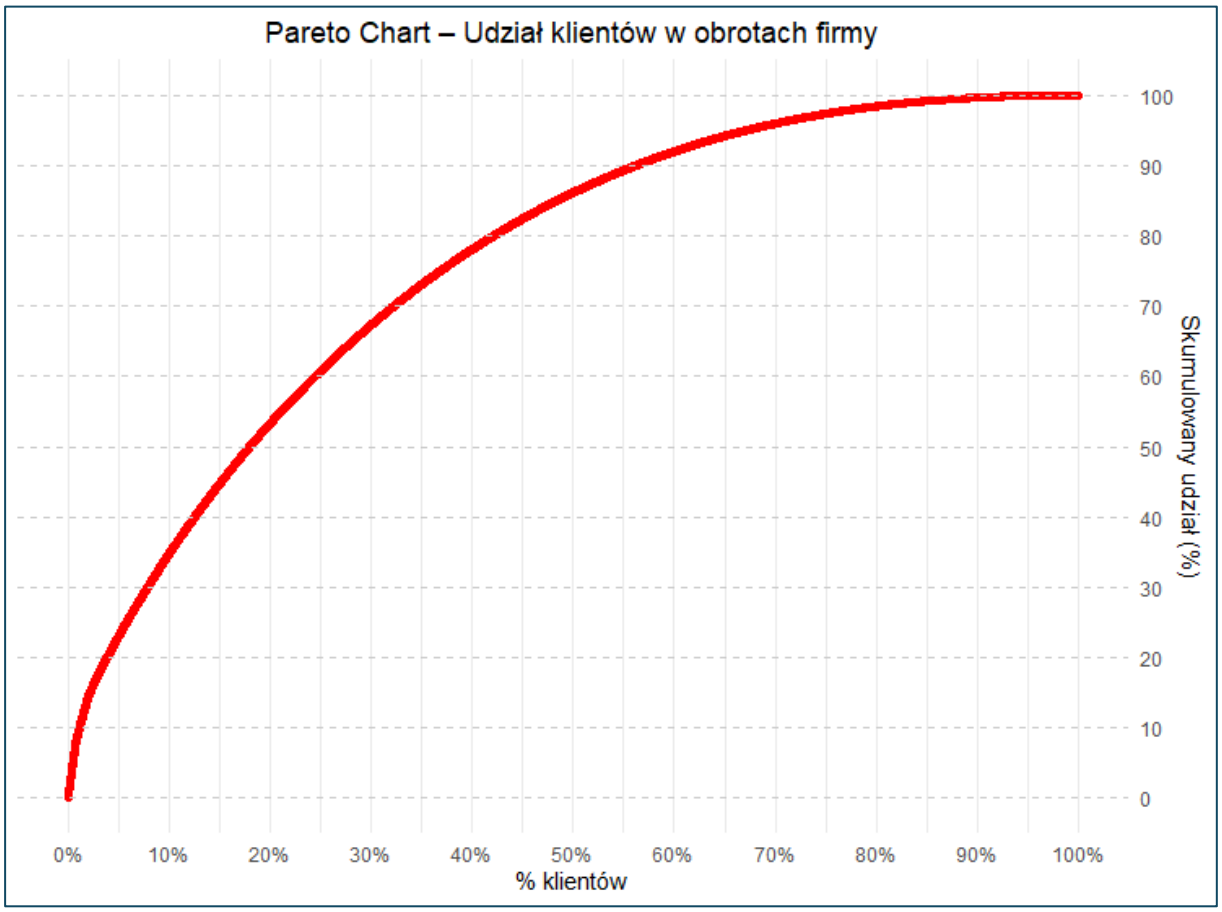


Wykres 7. Rozkład wartości sprzedaży produktów SKU14 i SKU15

2.2 Analiza PARETO klientów

Aby lepiej zrozumieć strukturę sprzedaży, przeprowadziłem analizę PARETO, która pokazuje, jaki procent klientów odpowiada za określony procent całości dochodów. Szczegółowe wartości można zobaczyć na poniższym wykresie.

Za 50% obrotów odpowiada zaledwie 15% klientów, a niemal 45% klientów generuje aż 80% sprzedaży. To jasno pokazuje, że kluczowe dla firmy jest dbanie o dobre relacje właśnie z tą grupą klientów

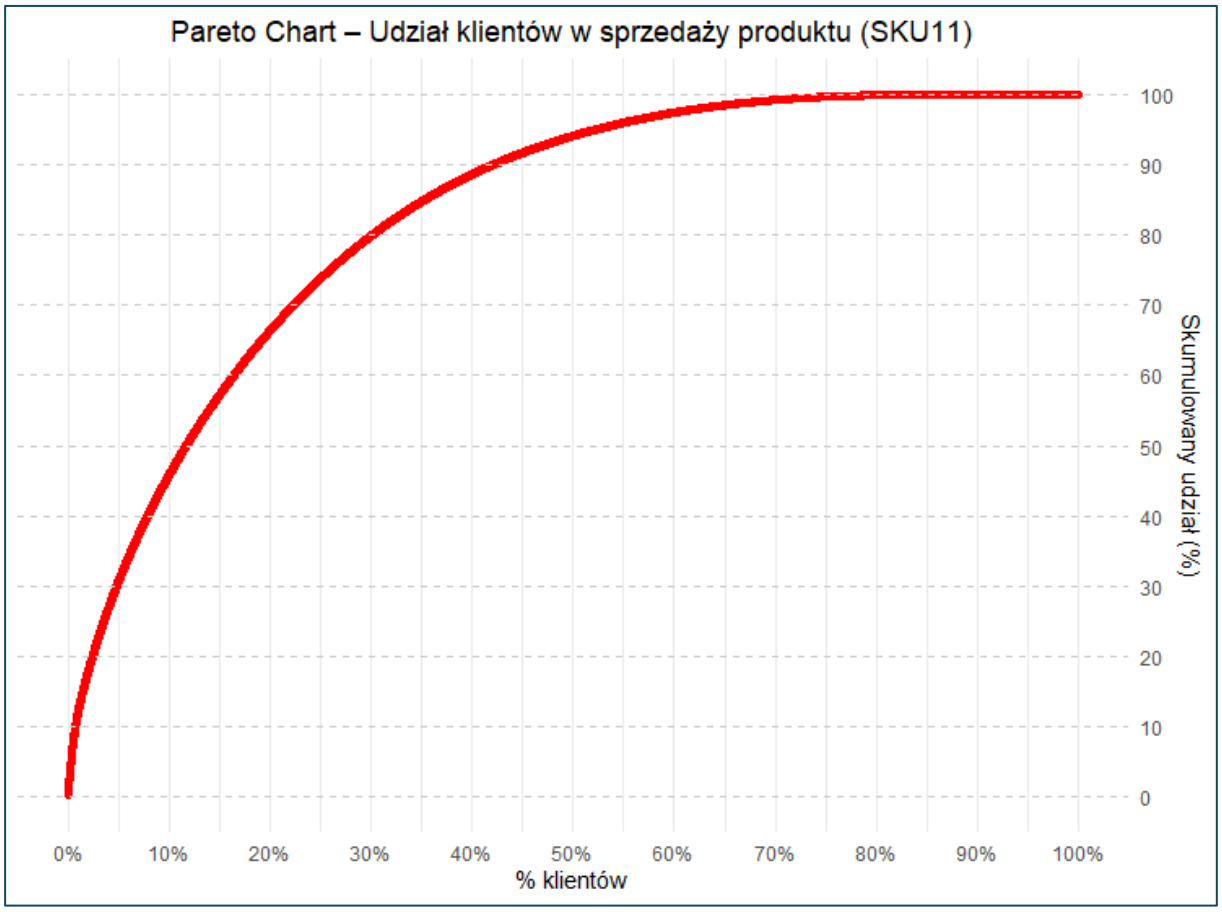


Wykres 8. Pareto chart pokazujący procentowy udział klientów w obrotach firmy

2.3 Analiza PARETO produktów

Przeprowadzono również analizę Pareto dla poszczególnych produktów, aby lepiej poznać strukturę sprzedaży. Ze względu na podobieństwo rozkładów udziału różnych SKU, w głównej części raportu skupiam się na szczegółowej analizie produktu SKU11, który wyróżnia się najwyższą wartością sprzedaży. Pozostałe wykresy dotyczące innych produktów umieściłem w aneksie, gdzie dostępny jest pełen zestaw danych.

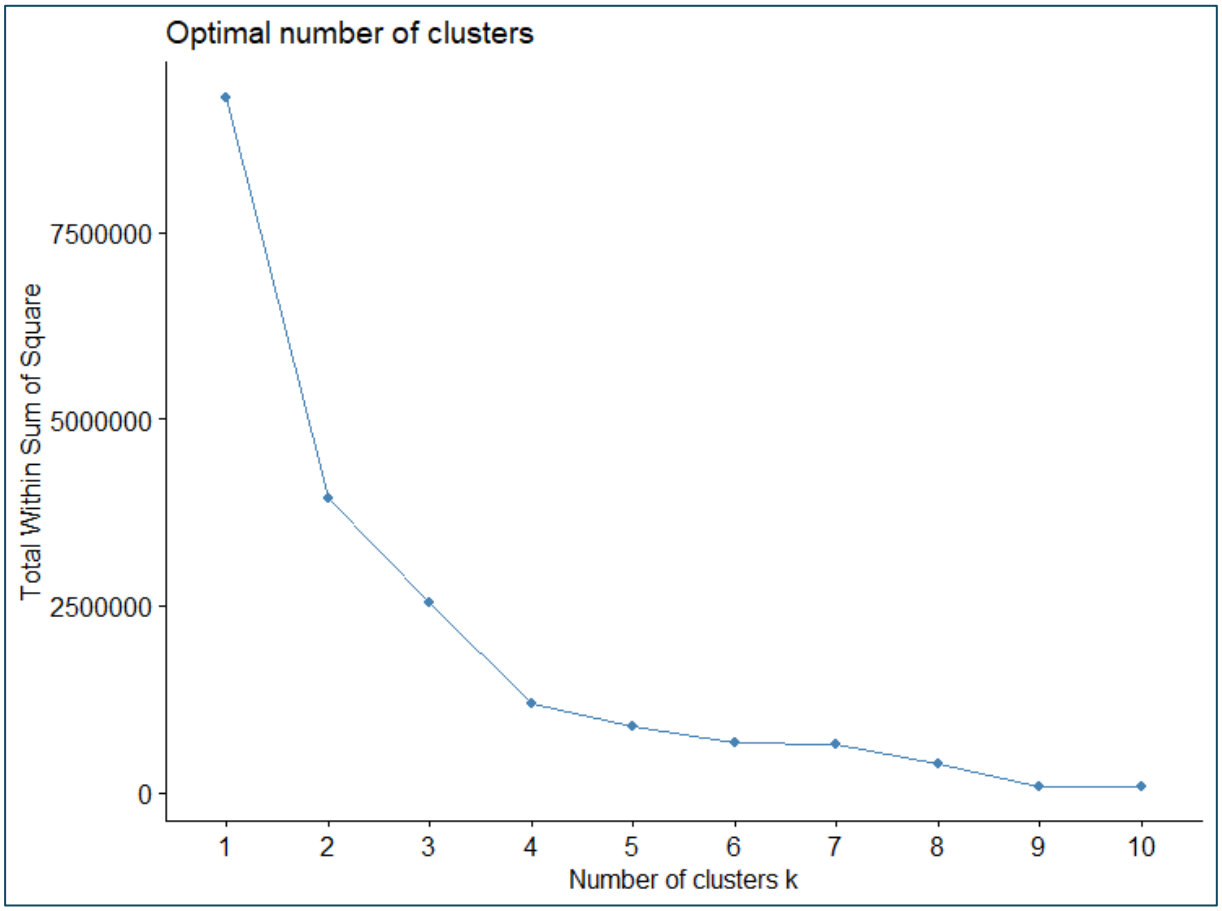
Z wykresu możemy między innymi odczytać, że za 80% wartości naszego najważniejszego pod względem obrotów produktu odpowiada zaledwie 30% klientów. Sugeruje to, że warto skupić szczególną uwagę i zaangażowanie na tej konkretnej grupie klientów.



Wykres 9.1 Pareto chart pokazujący procentowy udział klientów w sprzedaży produktu SKU11

3. Opis metodologii segmentacji

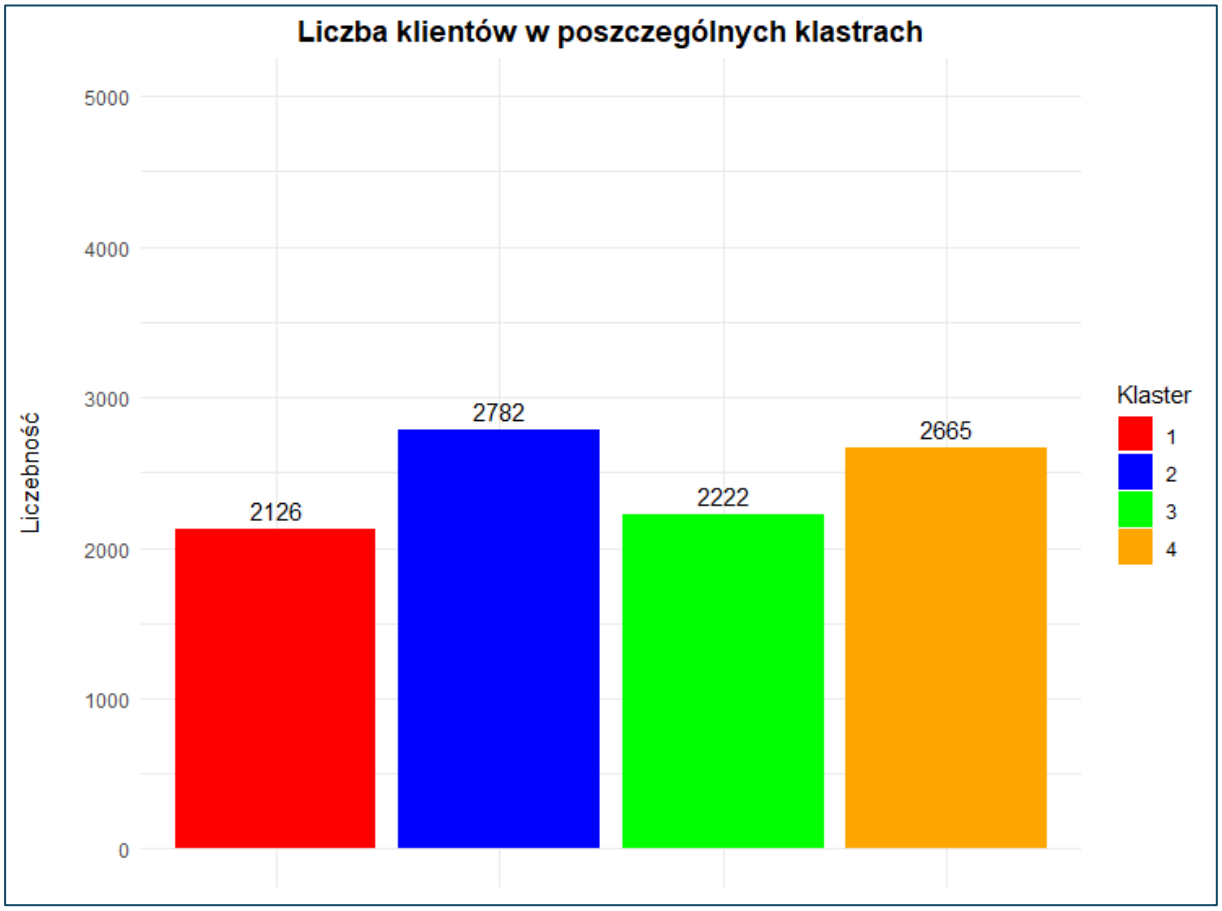
Do podziału klientów na grupy użyłem metody k-średnich, ponieważ dobrze sprawdza się przy tego typu danych i jest dość prosta w interpretacji. Aby wybrać optymalną liczbę klastrów, zastosowałem metodę łokcia — analizowałem, jak zmienia się błąd wewnątrz grup wraz ze wzrostem ich liczby. Wykres pokazał, że po osiągnięciu czterech klastrów dalsze zwiększanie ich liczby nie przynosi już znaczącej poprawy (linie zaczęły się wygładzać), dlatego uznałem cztery klastry za najlepszy wybór.



Wykres 10. Wykres „łokcia" pokazujący optymalną liczbę klastrów

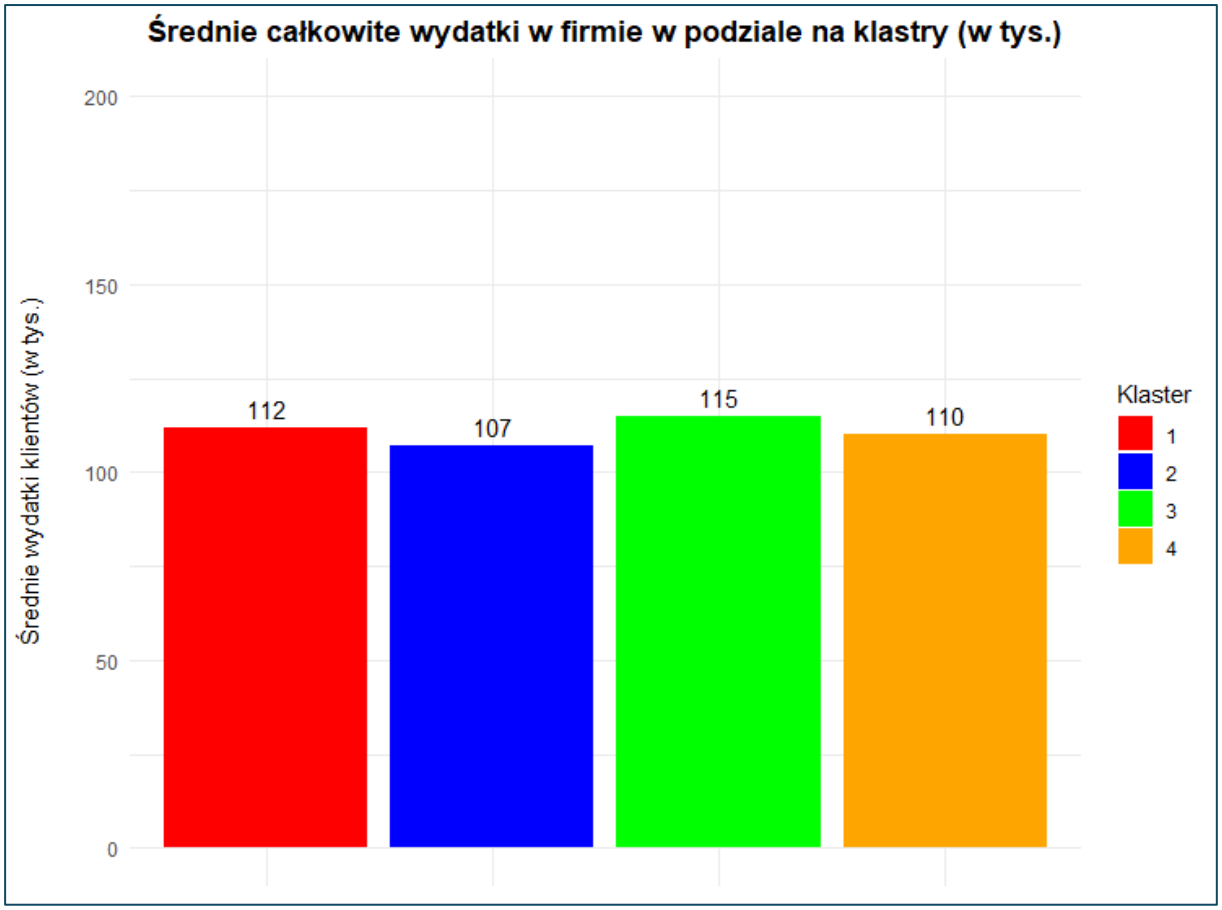
4. Wyniki segmentacji

Po przeprowadzeniu segmentacji klientów otrzymaliśmy cztery grupy. Ich liczebność przedstawia poniższy wykres. Widać, że klastry 2 i 4 są największe, podczas gdy klastry 1 i 3 mają nieco mniej klientów. Mimo to różnice w liczebności nie są znaczące — każdy klaster liczy ponad 2000 klientów.



Wykres 11. Wykres pokazujący całkowitą liczbę klientów w każdym z klastrów

Jednym z kluczowych pytań po segmentacji było, czy któryś z klastrów wyróżnia się pod względem średnich obrotów. Na wykresie poniżej widać, że wartości te są do siebie bardzo zbliżone. Aby to sprawdzić, przeprowadziłem test ANOVA, który nie wykazał podstaw do odrzucenia hipotezy zerowej o równości średnich. Oznacza to, że pod względem obrotów klastry nie różnią się istotnie.

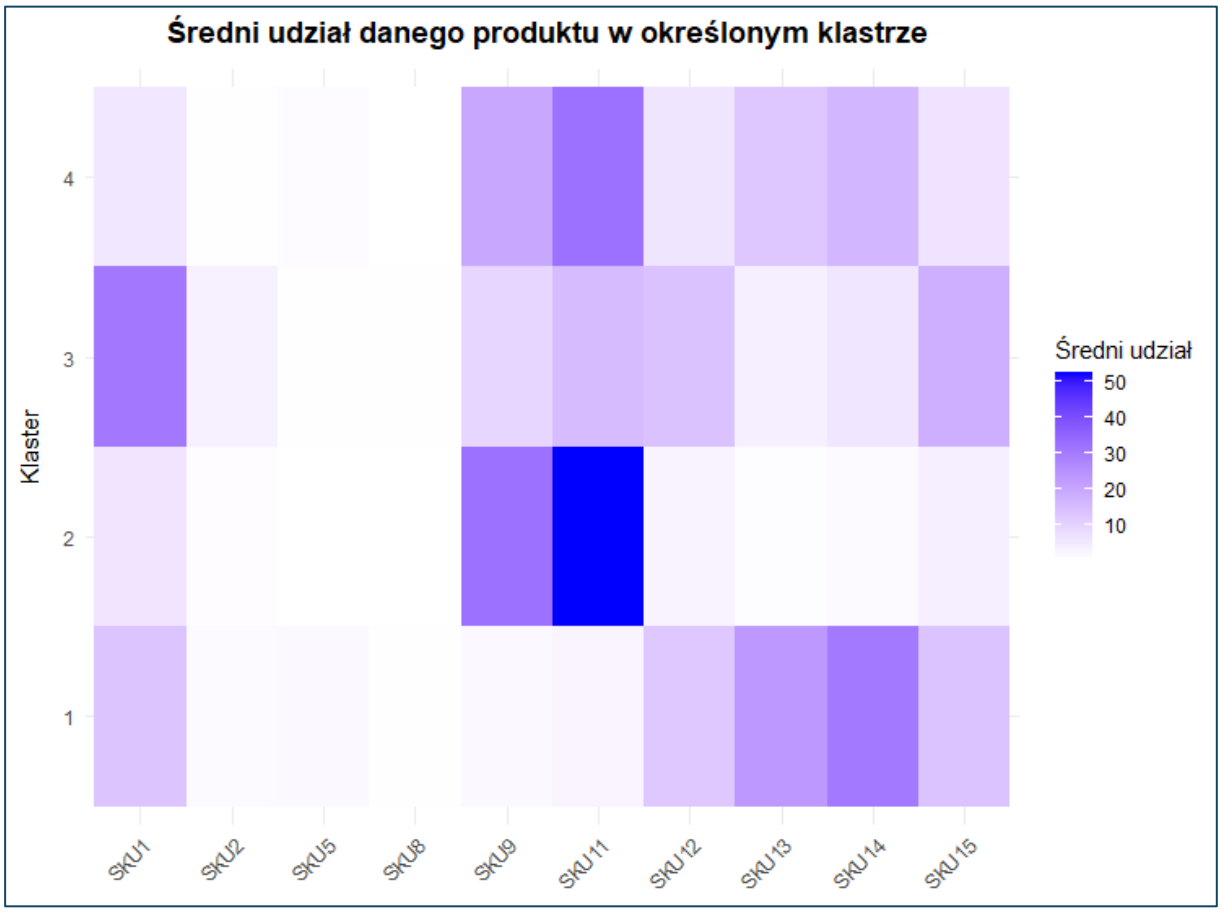


Wykres 12. Wykres pokazujący średnie całkowite wydatki w firmie zależnie od klastra

Na koniec, aby lepiej zrozumieć podobieństwa w nawykach zakupowych klientów, obliczyłem średni procentowy udział każdego produktu w poszczególnych klastrach i przedstawiłem to na heatmapie.

Analiza pokazała, że klaster 1 wyróżnia się największym udziałem produktu SKU14, a osoby z tej grupy koncentrują się głównie na czterech ostatnich produktach. Klaster 2 skupia się przede wszystkim na SKU11, ale ma też wyraźny udział w SKU9. Może to sugerować, że klienci kupujący dużo SKU11 często są zainteresowani także SKU9 — i odwrotnie.

Klaster 3 jest najbardziej zróżnicowany, z wyraźnym naciskiem na SKU1. Natomiast klaster 4 ma podobną strukturę do klastra 2, ale jest bardziej zrównoważony — ma mniejszy udział w SKU9 i SKU11, ale większy w SKU13 i SKU14.



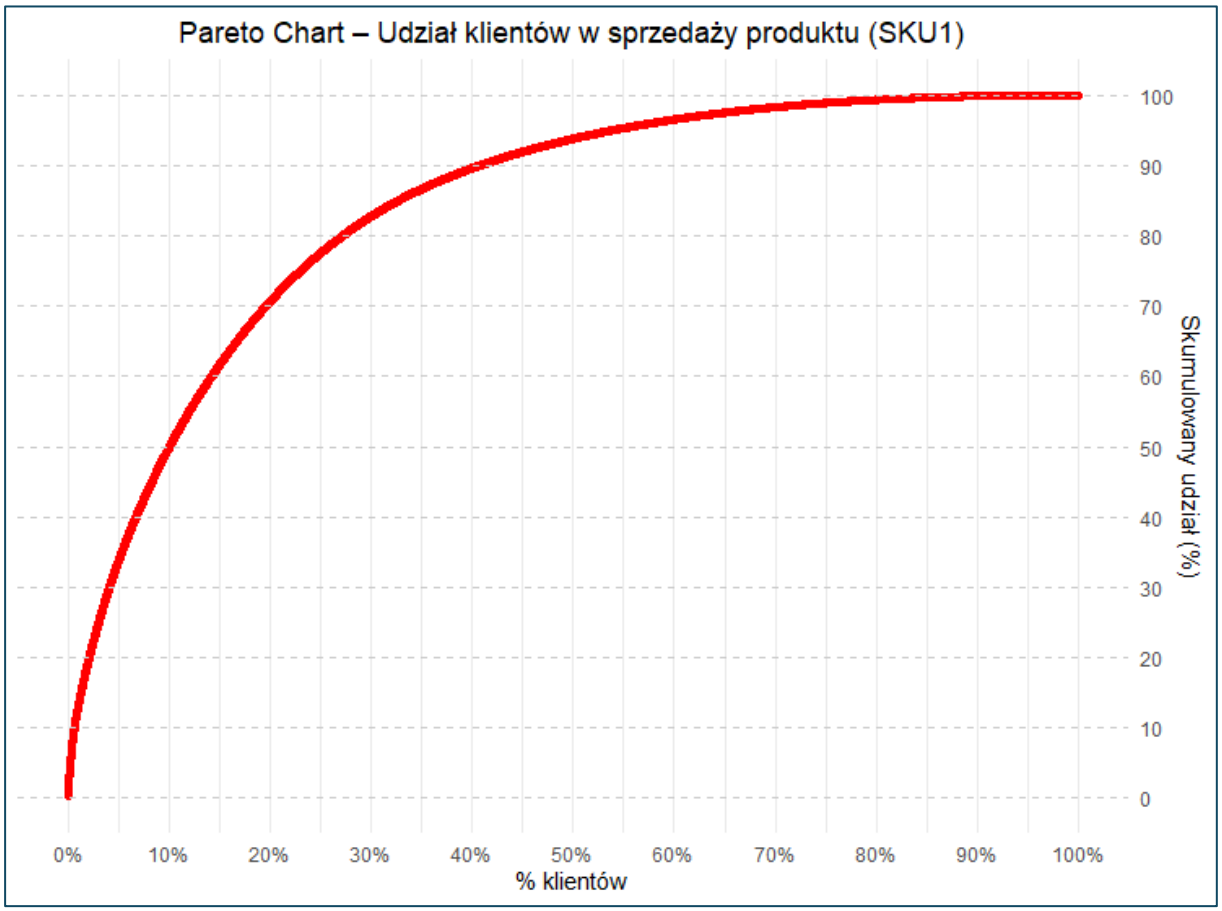
Wykres 13. Wykres pokazujący średni udział danego produktu w danym klastrze

5. Podsumowanie

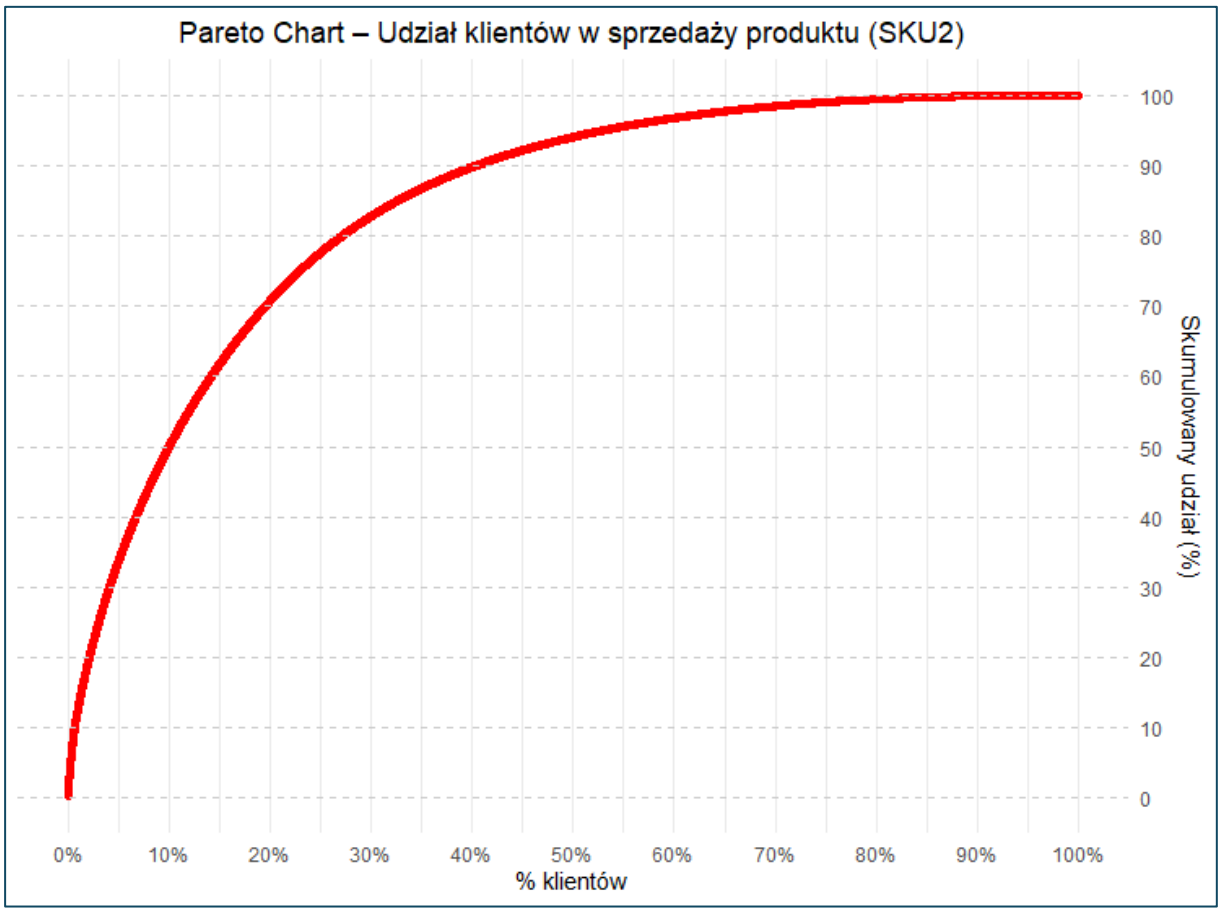
Na podstawie przeprowadzonej analizy można sformułować kilka ważnych wniosków: Produkt SKU11 generuje największy obrót i jest uznawany za kluczowy dla całej sprzedaży. Z kolei produkty SKU2, SKU5 oraz SKU8 mają marginalne znaczenie – każdy z nich odpowiada za mniej niż 3% całkowitego obrotu. Warto także zwrócić uwagę na rozkład udziałów poszczególnych produktów – dla większości klientów mieści się on w przedziale 0–0,01%, co świadczy o dużym rozproszeniu. Zidentyfikowano również pojedynczych klientów, których udział znacznie odstaje od reszty – mogą to być tzw. klienci VIP, jednak ze względu na brak dodatkowych informacji pozostają w analizie bez specjalnego wyróżnienia.

Segmentacja została przeprowadzona na cztery klastry o zbliżonej liczebności. Test ANOVA nie wykazał istotnych różnic w średnich obrotach między grupami, co sugeruje, że żaden klaster nie generuje wyraźnie większych przychodów niż pozostałe. Mimo to klastry różnią się pod względem struktury zakupowej, co może stanowić cenną wskazówkę przy planowaniu działań marketingowych lub tworzeniu ofert dopasowanych do konkretnych grup klientów.

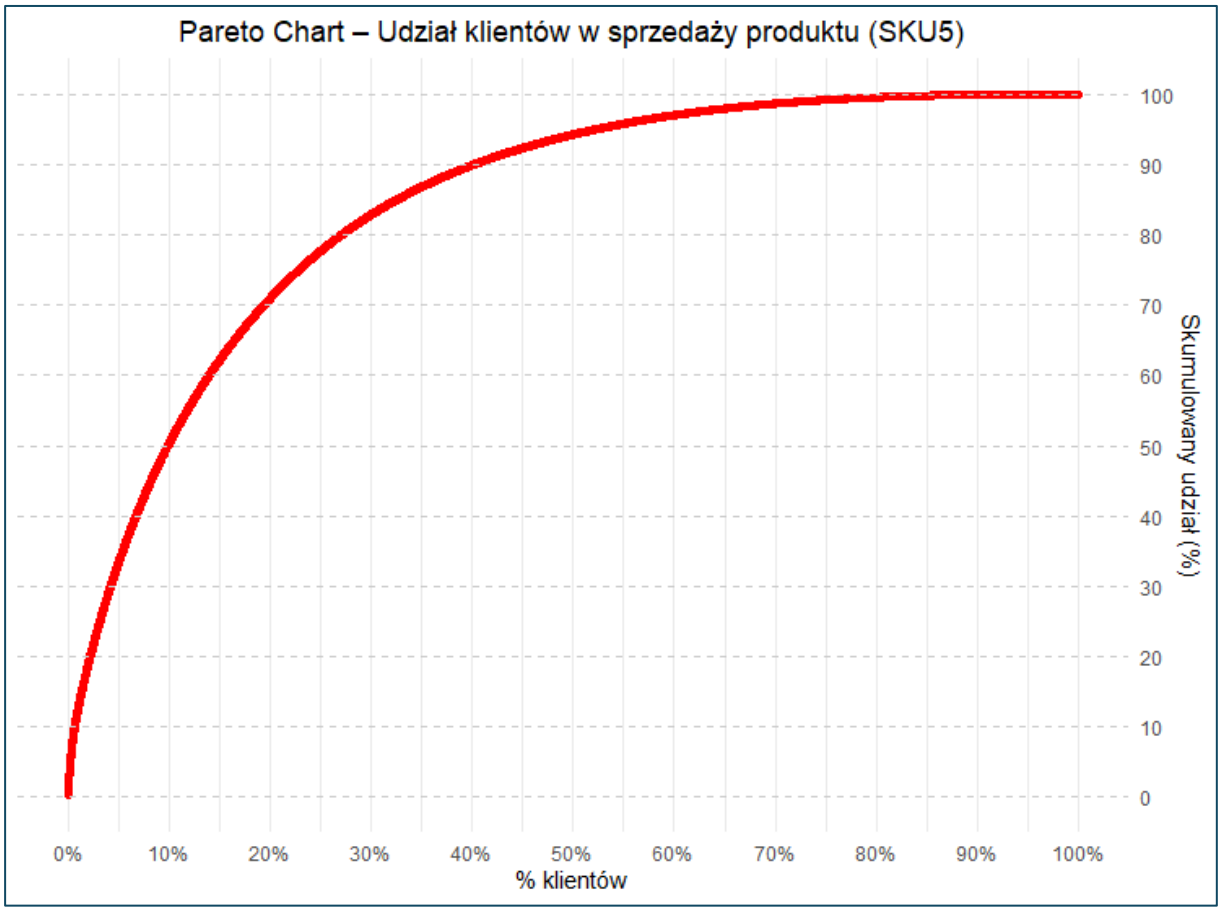
Aneks



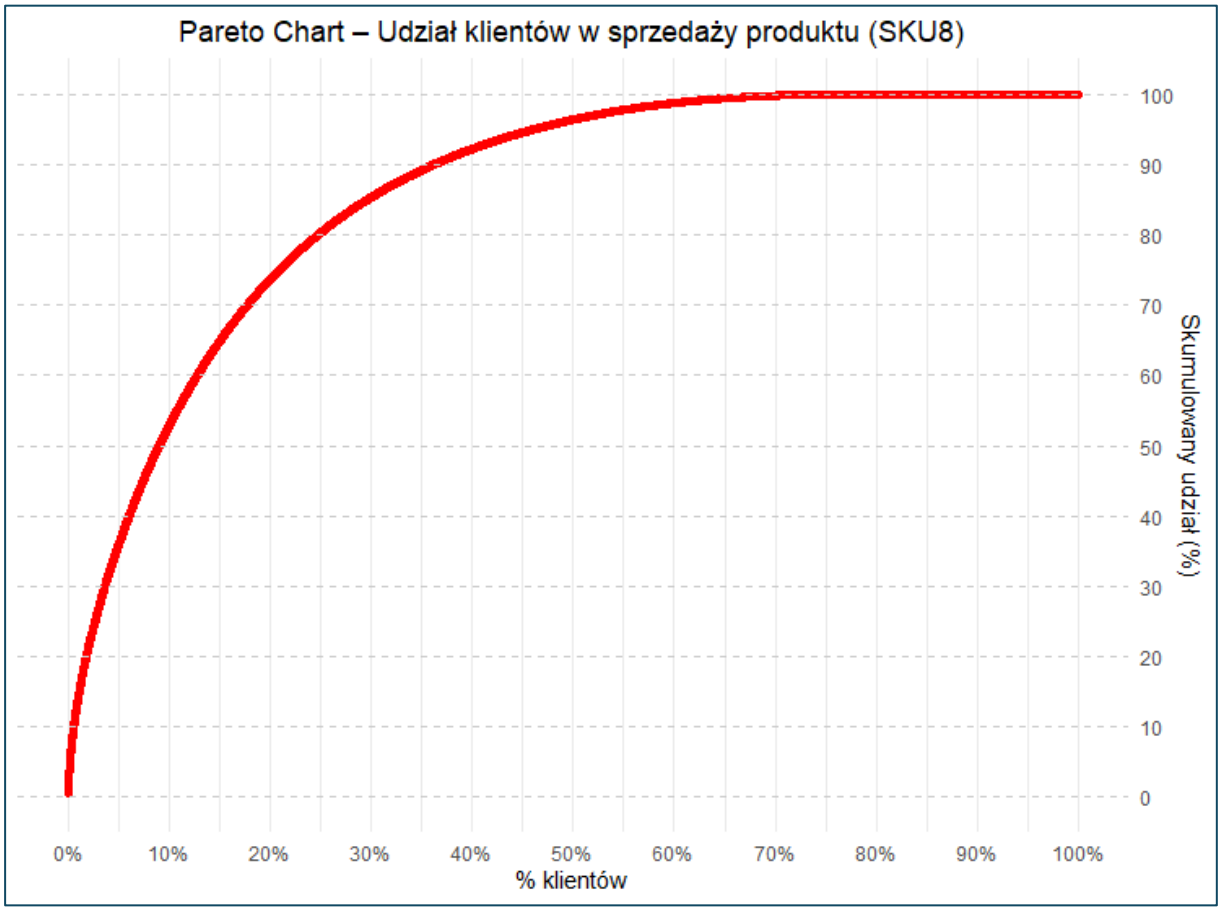
Wykres 9.2 Pareto chart pokazujący procentowy udział klientów w sprzedaży produktu SKU1



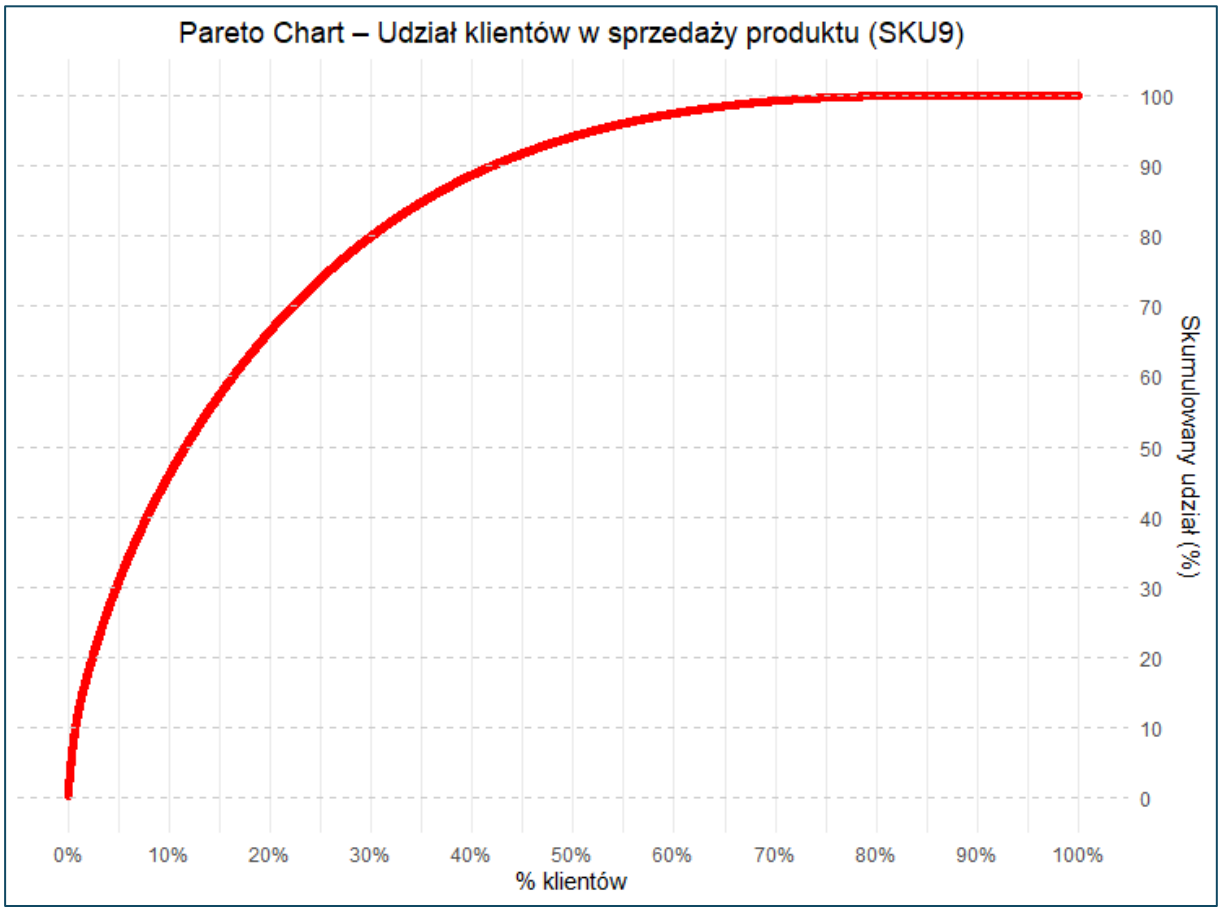
Wykres 9.3 Pareto chart pokazujący procentowy udział klientów w sprzedaży produktu SKU2



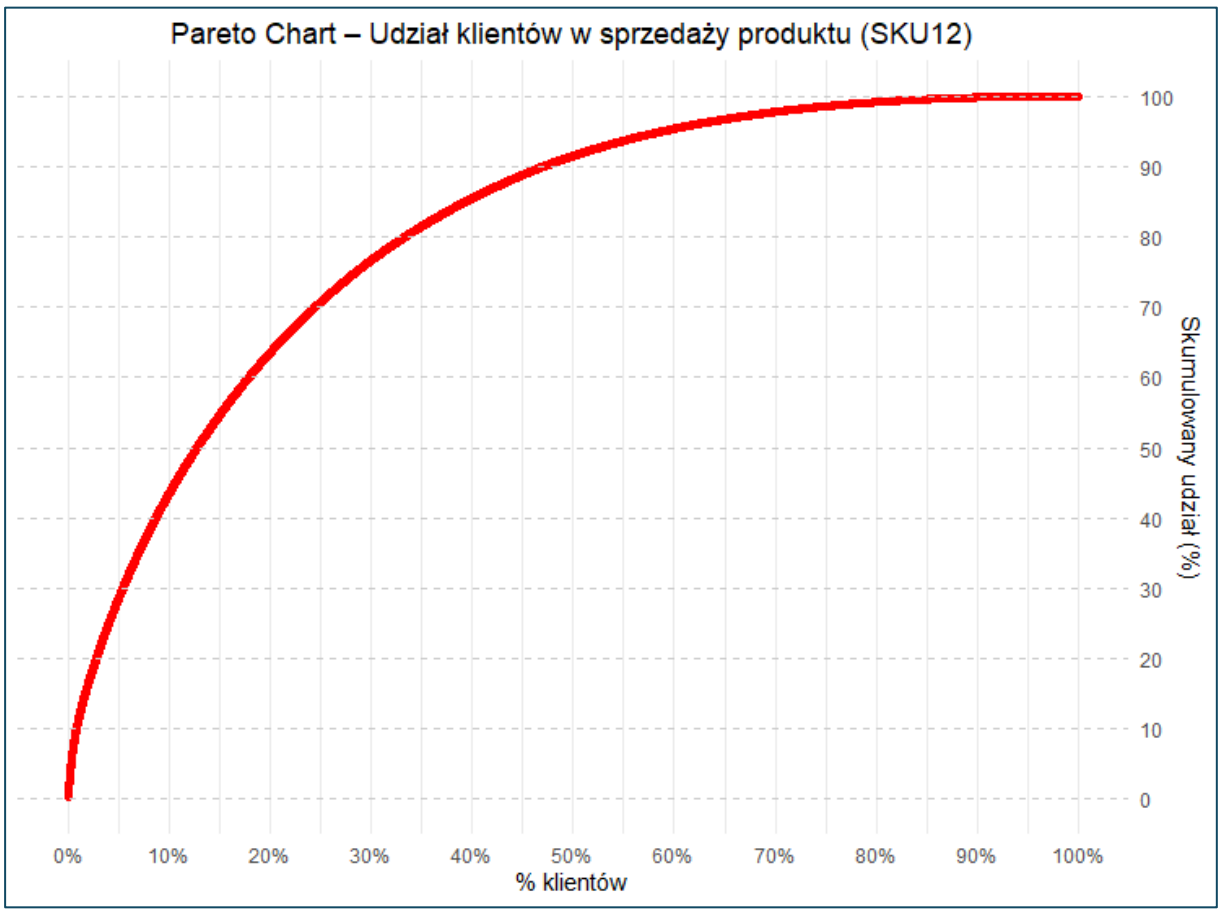
Wykres 9.4 Pareto chart pokazujący procentowy udział klientów w sprzedaży produktu SKU5



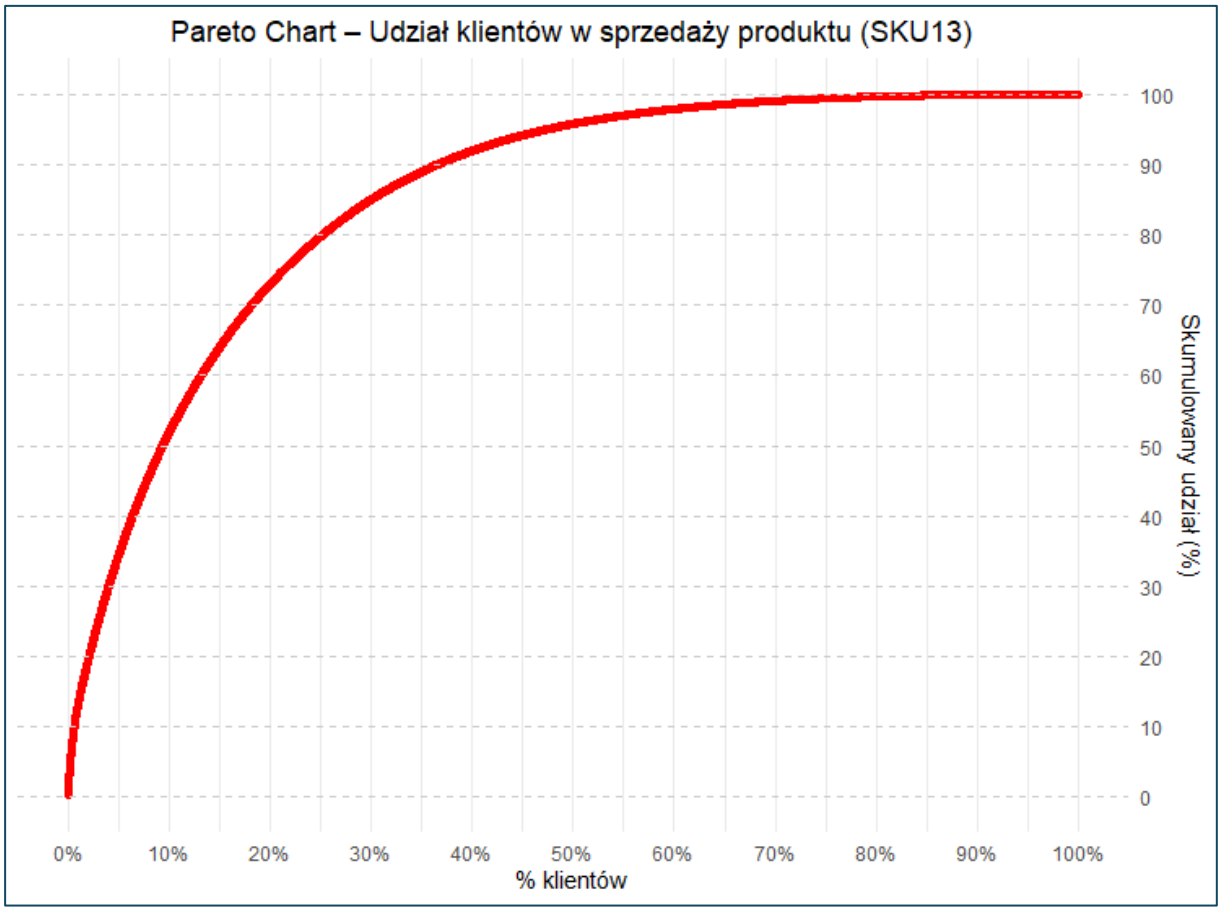
Wykres 9.5 Pareto chart pokazujący procentowy udział klientów w sprzedaży produktu SKU8



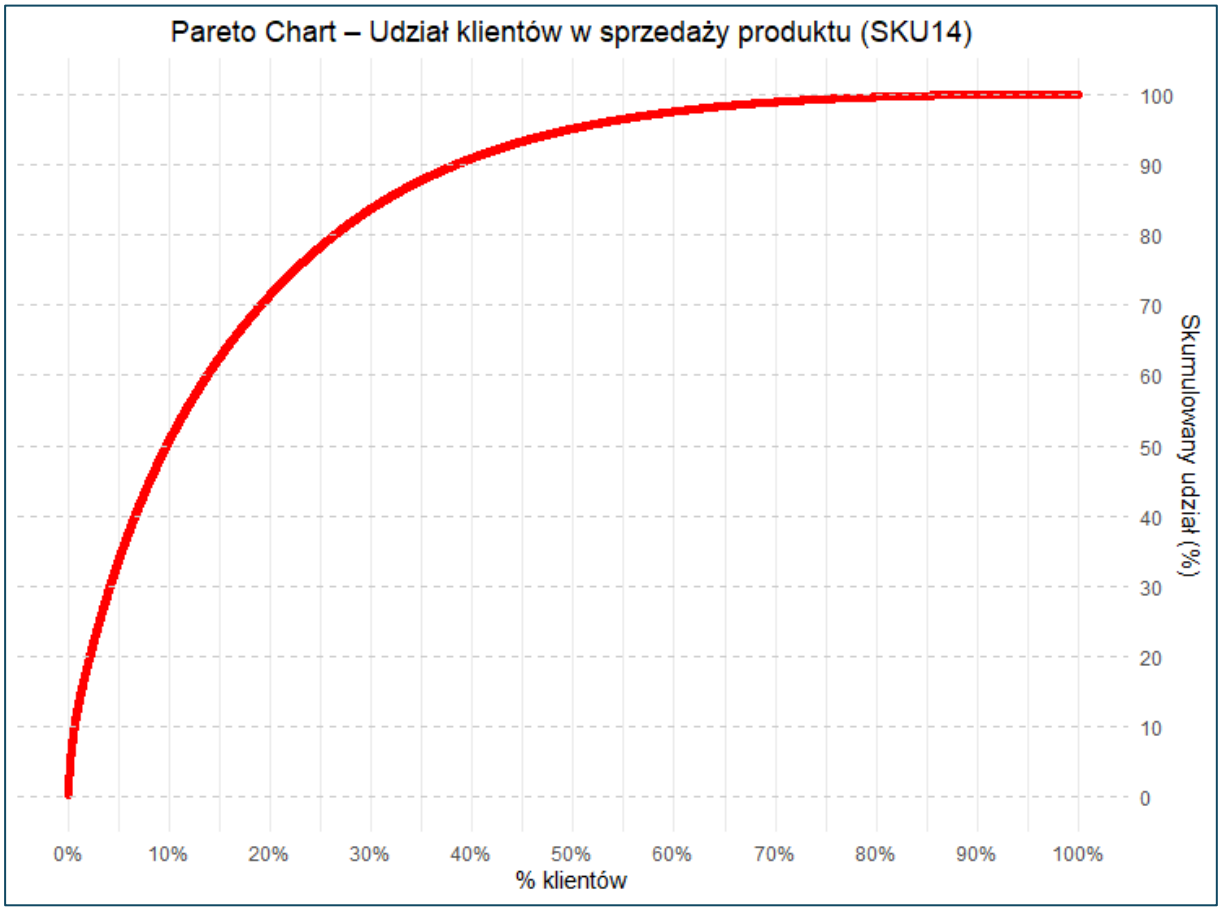
Wykres 9.6 Pareto chart pokazujący procentowy udział klientów w sprzedaży produktu SKU9



Wykres 9.7 Pareto chart pokazujący procentowy udział klientów w sprzedaży produktu SKU12



Wykres 9.8 Pareto chart pokazujący procentowy udział klientów w sprzedaży produktu SKU13



Wykres 9.9 Pareto chart pokazujący procentowy udział klientów w sprzedaży produktu SKU14