SPOŁECZNA AKADEMIA NAUK W ŁODZI

KIERUNEK STUDIÓW: INFORMATYKA

Mateusz Koniuch

Numer albumu: 110893

# Analiza i wizualizacja danych e-commerce przy użyciu Power BI

Praca inżynierska napisana pod kierunkiem
dr inż. Konrada Grzanka

**2025**

# Spis treści

# Wstęp

Postęp technologiczny i coraz większa dostępność narzędzi analitycznych przyczyniają się do zwiększenia wykorzystania danych w procesach podejmowania decyzji przez organizacje. Szczególną rolę odgrywają systemy wspomagania analizy danych, które umożliwiają przetwarzanie, analizowanie oraz wizualizowanie informacji, wspierając użytkowników w szybkim pozyskiwaniu istotnych wniosków. Niniejsza praca inżynierska koncentruje się na projektowaniu i wdrażaniu raportu analitycznego w środowisku Power BI, wykorzystującego dane z sektora e-commerce.

Celem pracy jest stworzenie kompleksowego raportu, który umożliwia analizę danych sprzedażowych, identyfikację trendów oraz ocenę zachowań klientów na podstawie rzeczywistych danych. W tym celu zastosowano różnorodne metody przetwarzania i transformacji informacji, w tym procedury ekstrakcji, przekształcania i ładowania danych, a także zasady modelowania struktur analitycznych oraz tworzenia interaktywnych pulpitów menedżerskich. Dodatkowo omówiono techniki optymalizacji raportów, najlepsze praktyki w zakresie wizualizacji danych oraz metody analizy kohortowej i segmentacji klientów, co pozwoliło na uzyskanie istotnych spostrzeżeń biznesowych.

Zakres pracy obejmuje zarówno teoretyczne aspekty związane z analizą danych, jak i praktyczną realizację raportu w Power BI. Pierwsza część zawiera przegląd narzędzi i technologii stosowanych w analizie informacji, ze szczególnym uwzględnieniem możliwości oferowanych przez Power BI. Następnie opisano proces planowania i przygotowania danych, obejmujący identyfikację wymagań, dobór źródeł informacji oraz ich przekształcenie na potrzeby analizy. Kolejna część skupia się na modelowaniu danych, ich transformacji i testowaniu wyników, co stanowiło istotny etap w zapewnieniu poprawności i wydajności raportu. Następnie omówiono zasady projektowania dashboardów, uwzględniając aspekty związane z ograniczeniem obciążenia poznawczego, percepcją wizualną i organizacją treści. W końcowej części pracy przedstawiono rezultaty przeprowadzonych analiz, podsumowując uzyskane wyniki oraz formułując wnioski dotyczące praktycznego zastosowania raportu.

Przedstawiona praca dostarcza kompleksowego przeglądu procesu budowy raportu analitycznego w Power BI, wskazując zarówno na istotne aspekty techniczne, jak i zasady efektywnej wizualizacji danych. Zaprezentowane rozwiązania mogą stanowić punkt wyjścia do dalszych badań nad skutecznością narzędzi analitycznych w przetwarzaniu danych biznesowych oraz ich wpływem na podejmowanie strategicznych decyzji w organizacjach.

# Rozdział I. Charakterystyka narzędzi i technologii

## 1.1. Power BI – krótki opis i funkcjonalności

Power BI to nowoczesne narzędzie Business Intelligence (BI) stworzone przez firmę Microsoft, służące do analizy danych oraz ich wizualizacji. Jest to wszechstronne oprogramowanie umożliwiające przekształcanie surowych danych w łatwe do interpretacji raporty oraz interaktywne wizualizacje. Power BI oferuje użytkownikom szeroką gamę funkcji, które wspierają proces podejmowania decyzji biznesowych, co sprawia, że jest chętnie wykorzystywane przez organizacje na całym świecie.

Jednym z kluczowych atutów Power BI jest jego zdolność do integracji z ekosystemem Microsoft, w tym takimi produktami jak Excel, SharePoint czy Azure. To umożliwia łatwe łączenie się z istniejącymi źródłami danych i wykorzystanie ich w analizie. Dzięki wbudowanym konektorom Power BI może także łączyć się z zewnętrznymi bazami danych, aplikacjami w chmurze czy systemami ERP.

Główne funkcjonalności Power BI obejmują:

- **Modelowanie danych**: Power BI pozwala na tworzenie relacji między danymi pochodzącymi z różnych źródeł, co umożliwia ich spójną analizę. Narzędzie oferuje intuicyjny interfejs oraz bogaty zestaw narzędzi do czyszczenia i przekształcania danych.
- **Zaawansowane wizualizacje**: Użytkownicy mogą tworzyć złożone wizualizacje danych, takie jak wykresy, tabele czy mapy. Narzędzie udostępnia wiele predefiniowanych wizualizacji oraz umożliwia tworzenie własnych.
- **Interaktywność**: Raporty stworzone w Power BI są w pełni interaktywne, co pozwala użytkownikom na eksplorowanie danych i dynamiczne filtrowanie wyników. Funkcjonalność ta pozwala na szczegółową analizę i odkrywanie ukrytych zależności.
- **Analiza w czasie rzeczywistym**: Power BI wspiera analizy na żywo dzięki zdolności do łączenia się z aktywnymi źródłami danych, co umożliwia monitorowanie zmiennych wskaźników w czasie rzeczywistym.

Power BI jest narzędziem niezwykle intuicyjnym, co sprawia, że może być wykorzystywane nie tylko przez ekspertów ds. analizy danych, ale również przez osoby bez

zaawansowanej wiedzy technicznej. W efekcie umożliwia to szeroką adopcję narzędzia w organizacjach różnej wielkości.

## 1.2. Porównanie Power BI z innymi narzędziami BI

Power BI jest jednym z wiodących narzędzi BI i często porównywany jest z innymi popularnymi rozwiązaniami, takimi jak Tableau, QlikView czy Excel. Każde z tych narzędzi oferuje unikalne cechy, które mogą być istotne w zależności od potrzeb użytkowników.

Tableau słynie z zaawansowanych możliwości wizualizacyjnych oraz intuicyjnego interfejsu. Jest szczególnie cenione za możliwość tworzenia estetycznych i skomplikowanych wizualizacji. Jednakże Power BI wyróżnia się integracją z ekosystemem Microsoft oraz konkurencyjną ceną, co czyni go bardziej dostępnym, zwłaszcza dla małych i średnich firm.

QlikView oferuje asocjacyjne modelowanie danych, co pozwala na odkrywanie ukrytych zależności w danych. Mimo to Power BI jest bardziej intuicyjne w obsłudze, a także oferuje szeroki zakres konektorów umożliwiających integrację z różnymi źródłami danych. Ponadto, Power BI zapewnia większą elastyczność w zakresie publikowania i udostępniania raportów.

Excel, choć wciąż popularny w analizie danych, ustępuje Power BI w zakresie nowoczesnych możliwości wizualizacyjnych oraz interaktywności. Power BI pozwala na łatwe importowanie danych z Excela, a następnie ich zaawansowaną analizę i prezentację w nowoczesnej, dynamicznej formie.

W kontekście narzędzi programistycznych warto wspomnieć o rozwiązaniach takich jak Shiny, które umożliwia tworzenie interaktywnych aplikacji analitycznych w języku R, czy narzędziach powiązanych z Pythonem, takich jak Dash lub Plotly. Te platformy oferują dużą elastyczność w tworzeniu wizualizacji i analiz danych, jednak wymagają większych umiejętności programistycznych. W porównaniu do Power BI, które charakteryzuje się gotowymi do użycia funkcjonalnościami, narzędzia te są bardziej odpowiednie dla zaawansowanych użytkowników poszukujących niestandardowych rozwiązań.

Power BI wyróżnia się również konkurencyjną ceną, co sprawia, że jest atrakcyjnym rozwiązaniem dla organizacji o ograniczonym budżecie. Dodatkowo, Microsoft regularnie aktualizuje narzędzie, wprowadzając nowe funkcje, co zapewnia jego stały rozwój. Duża społeczność użytkowników oraz bogate zasoby edukacyjne wspierają szybkie wdrożenie narzędzia i ułatwiają rozwiązywanie ewentualnych problemów.

Podsumowując, Power BI łączy w sobie intuicyjność, zaawansowane możliwości oraz przystępną cenę, co czyni go jednym z najbardziej uniwersalnych i popularnych narzędzi BI na rynku.

| Narzędzie | Zalety | Wady | Docelowi użytkownicy |
|---|---|---|---|
| Power BI | Integracja z Microsoft, intuicyjny interfejs, konkurencyjna cena | Ograniczona personalizacja w porównaniu do Tableau | Organizacje każdej wielkości, szczególnie korzystające z ekosystemu Microsoft |
| Tableau | Zaawansowane wizualizacje, intuicyjny interfejs | Wyższy koszt licencji | Firmy wymagające skomplikowanych analiz |
| QlikView | Asocjacyjne modelowanie danych, interaktywność | Trudniejsza obsługa, wysoki koszt | Duże organizacje i analitycy danych |
| Excel | Prostota, dostępność, znajomość wśród użytkowników | Brak zaawansowanych wizualizacji | Małe firmy i użytkownicy indywidualni |
| Shiny (R) | Tworzenie aplikacji analitycznych, elastyczność | Wymaga znajomości programowania | Programiści i analitycy danych |
| Dash (Python) | Integracja z Pythonem, dynamiczne wizualizacje | Wymaga znajomości programowania | Zaawansowani analitycy danych |

**Tabela 1.** Porównanie kluczowych cech narzędzi Business Intelligence (źródło: opracowanie własne)

# Rozdział II. Planowanie i przetwarzanie danych

## 2.1. Planowanie i zbieranie wymagań

### 2.1.1. Identyfikacja celu raportu

Tworzenie dashboardu biznesowego (z ang. tablica rozdzielcza lub panel kontrolny – narzędzie wizualne do prezentacji danych) wymaga precyzyjnego określenia celu, który ma zostać osiągnięty. Dashboardy odgrywają istotną rolę w organizacjach, umożliwiając integrację danych, monitorowanie kluczowych wskaźników efektywności (KPI, ang. Key Performance Indicators) oraz podejmowanie decyzji w oparciu o dane. Proces definiowania celów raportu jest istotny dla stworzenia skutecznych narzędzi analitycznych, takich jak dashboard w Power BI.

Cele biznesowe powinny być jednoznaczne, mierzalne i dostosowane do konkretnych potrzeb organizacji. Dashboardy zostały zaprojektowane w celu spełniania specyficznych wymagań użytkowników, np. w zakresie monitorowania wyników sprzedaży, analizy trendów rynkowych czy identyfikacji problemów operacyjnych. Raport interaktywny opracowany w ramach niniejszego projektu w Power BI ma na celu realizację następujących założeń:

- Analiza wyników sprzedaży, pozwalająca zespołowi sprzedaży na bieżące śledzenie wyników w podziale na produkty, regiony lub metody płatności.
- Identyfikacja problemów operacyjnych, takich jak opóźnienia w dostawach, niska efektywność procesów logistycznych czy negatywne zmiany w poziomie zadowolenia klientów.
- Wspieranie decyzji strategicznych, poprzez dostarczanie menedżerom kompleksowych danych dotyczących wydajności przedsiębiorstwa, wspierających procesy decyzyjne.

Fundamentalnym etapem w definiowaniu celów jest precyzyjne określenie problemów, które dashboard ma za zadanie rozwiązać. Przykładowe cele obejmują poprawę efektywności operacyjnej, m.in. poprzez identyfikację wąskich gardeł w procesach logistycznych lub problemów z dostawcami. Inne istotne założenia to zwiększenie wydajności sprzedaży dzięki analizie segmentacji klientów i metod płatności oraz optymalizacja doświadczeń klientów poprzez analizę czasu dostawy i poziomu satysfakcji.

W analizowanym projekcie, opartym na Power BI, zastosowano podejście łączące dogłębną analizę potrzeb użytkowników końcowych z oceną możliwości narzędzia. Główne cele projektu obejmują stworzenie intuicyjnego i prostego w obsłudze narzędzia, dostępnego dla użytkowników o różnym poziomie zaawansowania technicznego. Narzędzie to ma zapewniać możliwość kompleksowej analizy wyników firmy w branży e-commerce, a także cechować się elastycznością, umożliwiającą dostosowanie do przyszłych wymagań biznesowych.

Określenie celów raportu stanowi fundament dla powodzenia każdego projektu analitycznego. Jasno sprecyzowane cele oraz dogłębne zrozumienie problemów, które dashboard ma rozwiązać, stanowią solidną podstawę dla kolejnych etapów, takich jak zbieranie wymagań, przetwarzanie danych i projektowanie wizualizacji. Dzięki temu proces tworzenia dashboardu staje się bardziej precyzyjny, a końcowy efekt lepiej odpowiada na potrzeby użytkowników[1].

### 2.1.2. Zdefiniowanie grupy docelowej i określenie kluczowych wskaźników

Aby opracować efektywny i użyteczny raport w Power BI, zasadnicze znaczenie ma dokładne zrozumienie odbiorców oraz ich potrzeb. W kontekście projektu można wyróżnić trzy przykładowe grupy użytkowników końcowych:

1. **Kadra zarządzająca**: Potrzebują przejrzystych raportów prezentujących kluczowe wskaźniki, takie jak całkowite przychody czy trendy sprzedaży. W raporcie dla tej grupy uwzględniono karty KPI na stronie przeglądowej, takie jak „Total Revenue" i „Customer Lifetime Value".

2. **Menedżerowie operacyjni**: Skupiają się na szczegółowych danych dotyczących wydajności, czasu realizacji zamówień i wskaźników zwrotów. Raport dla nich zawiera strony, takie jak „Customer Analysis" czy „Delivery Performance Analysis", które umożliwiają dokładną analizę i identyfikację problematycznych obszarów.

3. **Analitycy danych i specjaliści ds. sprzedaży**: Wymagają dostępu do szczegółowych danych do analizy trendów, wyników produktów oraz efektywności kampanii sprzedażowych. Raport zapewnia wykresy segmentacyjne, np. „Customer Segmentation by RFM Scores", oraz opcje eksploracji danych klientów i produktów.

[1] Chrabski B., Zmitrowicz K.: *Inżynieria wymagań w praktyce*, Wydawnictwo Naukowe PWN, Warszawa 2015, s. 121-128

Identyfikacja potrzeb tych grup pozwala na optymalne dostosowanie układu raportu, jego interaktywności oraz prezentowanych wskaźników, co zwiększa jego wartość i przydatność dla użytkowników[2].

Wskaźniki KPI to fundament każdego efektywnego dashboardu. Ich wybór powinien być ściśle powiązany z celami biznesowymi oraz potrzebami grup docelowych. W ramach raportu opracowanego w tym projekcie zdefiniowano następujące kluczowe wskaźniki:

- **Total Revenue** (Całkowity przychód):
    - Cel: Monitorowanie ogólnej wydajności finansowej firmy. Ważne dla kadry zarządzającej.
    - Prezentacja: Karta KPI oraz wykres trendu przychodów miesięcznych.
- **Customer Lifetime Value** (Wartość klienta w czasie):
    - Cel: Ocena wartości klientów w dłuższej perspektywie czasowej oraz identyfikacja kluczowych segmentów klientów.
    - Prezentacja: Strona „Customer Analysis” z wykresami porównującymi grupy klientów według CLV.
- **On-Time Delivery Rate** (Wskaźnik terminowej dostawy):
    - Cel: Pomiar skuteczności działań logistycznych. Istotne dla menedżerów operacyjnych.
    - Prezentacja: Strona „Delivery Performance Analysis” z rozbiciem na czasy dostawy oraz wskaźniki terminowości.
- **Total Orders** (Całkowita liczba zamówień):
    - Cel: Ocena popularności i aktywności w sklepie e-commerce.
    - Prezentacja: Karta KPI oraz wykres liczby zamówień w czasie.

[2] Chrabski B., Zmitrowicz K.: *Inżynieria wymagań w praktyce*, s. 128-130

- **Retention Rate** (Wskaźnik retencji klientów):
    - Cel: Pomiar zdolności firmy do utrzymania klientów i ich ponownej aktywacji.
    - Prezentacja: Strona „Customer Retention Analysis" z analizą kohortową.
- **Average Order Value** (Średnia wartość zamówienia):
    - Cel: Analiza wzorców zakupowych i efektywności działań sprzedażowych.
    - Prezentacja: Porównanie metod płatności oraz segmentacja klientów.
- **Low-Rated Orders** (% zamówień z niską oceną):
    - Cel: Ocena jakości obsługi klienta i identyfikacja problematycznych obszarów.
    - Prezentacja: Wskaźnik w „Customer Retention Analysis" oraz możliwość filtrowania według niskich ocen.
- **Top Products and Categories** (Najlepsze produkty i kategorie):
    - Cel: Analiza najbardziej dochodowych produktów i segmentów.
    - Prezentacja: Wykres Pareto w sekcji „Sales Analysis".

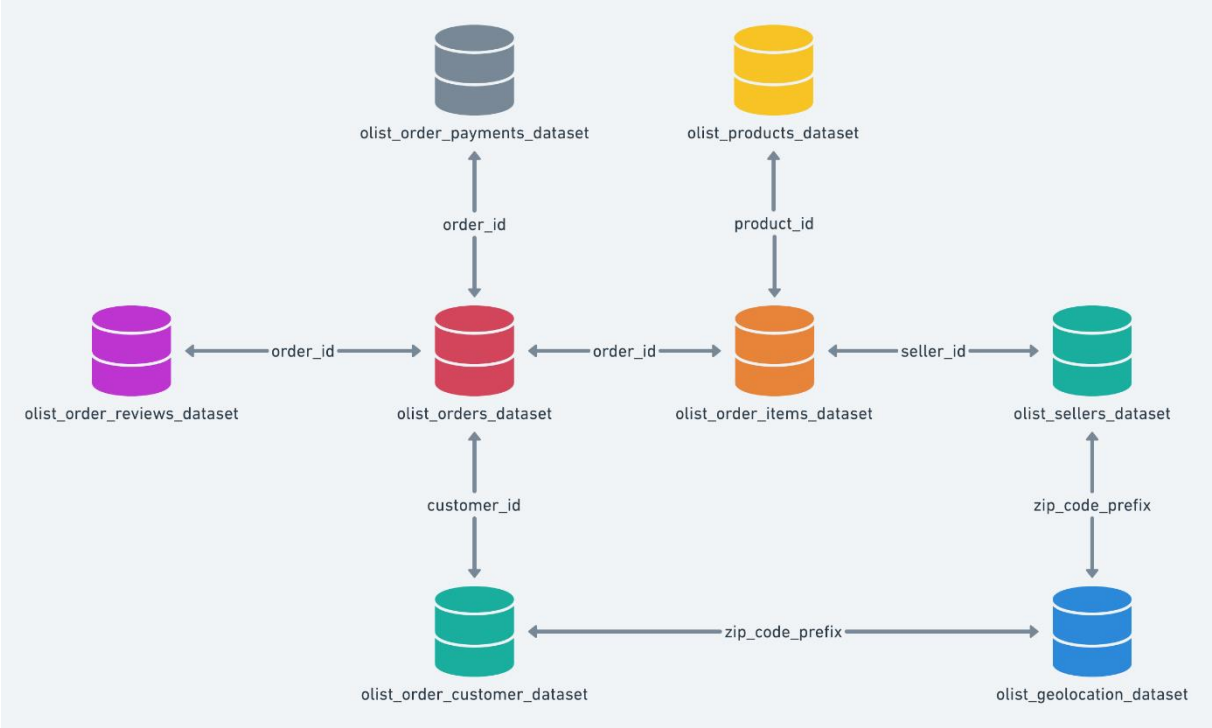
Dzięki tym wskaźnikom raport dostarcza wszechstronnego wglądu w krytyczne aspekty działalności biznesowej – od wydajności operacyjnej, przez lojalność klientów, aż po wyniki finansowe i sprzedażowe. Taka struktura raportu wspiera zarówno działania operacyjne, jak i podejmowanie decyzji strategicznych, dostarczając kompleksowego wsparcia w zarządzaniu organizacją[3].

## 2.2. Identyfikacja i pozyskanie źródeł danych

Dane wykorzystane w projekcie pochodzą z publicznie dostępnego zestawu „Brazilian E-Commerce Dataset", opublikowanego na platformie Kaggle. Jest to kompleksowy zbiór danych przedstawiający różnorodne aspekty działalności e-commerce w Brazylii, obejmujące zamówienia, klientów, sprzedawców, produkty, płatności, opinie oraz geolokalizację. Dane udostępniono w formacie CSV, co umożliwiło ich łatwy import do Power BI i przetwarzanie na potrzeby analizy.

[3] https://www.grow.com/blog/how-to-select-metrics-for-your-kpi-dashboard, [dostęp 30.12.2024]

Zbiór danych składa się z kilku podstawowych tabel, z których każda opisuje inne elementy procesu e-commerce. Tabela „orders” zawiera dane o statusach zamówień, datach ich zakupu i dostawy. Tabela „products” prezentuje szczegóły dotyczące produktów, takie jak kategorie, wymiary czy liczba zdjęć. Informacje o klientach znajdują się w tabeli „customers”, umożliwiając segmentację na podstawie lokalizacji i unikalnych identyfikatorów. Z kolei tabela „sellers” gromadzi dane o sprzedawcach, w tym ich lokalizacji i efektywności. Szczegółowe informacje dotyczące poszczególnych zamówień, takie jak ceny i koszty wysyłki, są dostępne w tabeli „order_items”. Dane płatności zawiera tabela „order_payments”, obejmująca metody płatności, liczby rat oraz kwoty transakcji, a tabela „order_reviews” dostarcza opinii klientów, w tym ocen i komentarzy. Dodatkowo tabela „geolocation” zawiera dane lokalizacyjne sprzedawców i klientów, co pozwala na przeprowadzanie analiz regionalnych, takich jak badania efektywności logistyki dostaw. Dane te zostały zaimportowane do Power BI i zintegrowane w modelu relacyjnym, co umożliwiło przeprowadzenie zaawansowanych analiz i wizualizacji.




**Rysunek 1.** Relacyjny model danych dla zestawu „Brazilian E-Commerce Dataset” (źródło: [16])

W projektach komercyjnych proces pozyskiwania danych jest bardziej złożony niż w przypadku wykorzystania publicznych zbiorów. Pierwszym krokiem jest określenie wymagań projektowych wynikających z celów biznesowych i oczekiwań interesariuszy.

Należy zidentyfikować, jakie dane są niezbędne do realizacji założonych celów oraz które wskaźniki KPI będą kluczowe.

Po sprecyzowaniu wymagań następuje identyfikacja źródeł danych. W środowisku biznesowym dane mogą pochodzić zarówno z wewnętrznych systemów organizacji, takich jak CRM (ang. Customer Relationship Management – zarządzanie relacjami z klientami), ERP (ang. Enterprise Resource Planning – planowanie zasobów przedsiębiorstwa) czy bazy transakcyjne, jak i z zewnętrznych źródeł, np. informacji demograficznych, danych rynkowych, prognoz ekonomicznych czy danych pogodowych. Ważnym etapem jest ocena wiarygodności, dostępności oraz zgodności wybranych źródeł z istniejącymi systemami organizacji.

Kolejnym etapem jest ocena jakości danych, obejmująca sprawdzenie ich kompletności, spójności i aktualności. Braki w danych lub ich niespójność mogą prowadzić do błędnych wyników analizy. Dlatego niezbędne jest przeprowadzenie procesów czyszczenia i transformacji danych przed ich dalszym wykorzystaniem. W tym celu często stosowane są narzędzia ETL (ang. Extract, Transform, Load – ekstrakcja, transformacja, ładowanie), takie jak Apache NiFi, Talend czy Power Query, które umożliwiają automatyzację tych działań.

Integracja danych z różnych źródeł wymaga ich ujednolicenia, zmapowania pól i ustalenia relacji między nimi. Na przykład dane sprzedażowe z systemu ERP można łączyć z danymi marketingowymi, aby przeanalizować wpływ kampanii reklamowych na wyniki sprzedaży. Na tym etapie istotne jest również zapewnienie zgodności z regulacjami dotyczącymi ochrony danych, takimi jak RODO, co stanowi nieodłączny element w środowisku biznesowym[4].

## 2.3. Eksploracyjna analiza danych

Eksploracyjna analiza danych (EDA, ang. Exploratory Data Analysis) jest kluczowym etapem w procesie przygotowywania danych do analizy i wizualizacji. W kontekście raportu e-commerce realizowanego w Power BI profilowanie danych odegrało zasadniczą rolę w zrozumieniu ich struktury, jakości oraz charakterystyki, co pozwoliło na identyfikację potencjalnych problemów.

W projekcie wykorzystano różnorodne techniki eksploracyjnej analizy danych, które zostały opisane w literaturze, w tym w książce „Analiza danych z wykorzystaniem SQL-a”

[4] Hand D., Mannila H., Smyth P.: *Eksploracja danych*, Wydawnictwa Naukowo-Techniczne, Warszawa 2005, s. 82-89

autorstwa Cathy Tanimura. Jednym z pierwszych kroków było szczegółowe zapoznanie się z tabelami i ich zawartością. Dla każdej z tabel wykonano podsumowanie zmiennych opisujących zbiór za pomocą języka DAX[5].

| | [Column] | [Count] | [Distinct Values] | [Null Count] | [Min] | [Max] | [Zeros] |
|---|---|---|---|---|---|---|---|
| 1 | geolocation_zip_code_prefix | 1000163 | 19015 | 0 | 1001 | 99990 | 0 |
| 2 | geolocation_lat | 1000163 | 717372 | 0 | -0.000043673... | -9.999731658... | N/A |
| 3 | geolocation_lng | 1000163 | 717615 | 0 | -101.4667664... | -98.48412074... | N/A |
| 4 | geolocation_city | 1000163 | 8010 | 0 | * cidade | zortea | N/A |
| 5 | geolocation_state | 1000163 | 27 | 0 | AC | TO | N/A |

**Tabela 2.** Podsumowanie statystyk tabeli „olist_geolocation_dataset" (źródło: opracowanie własne)

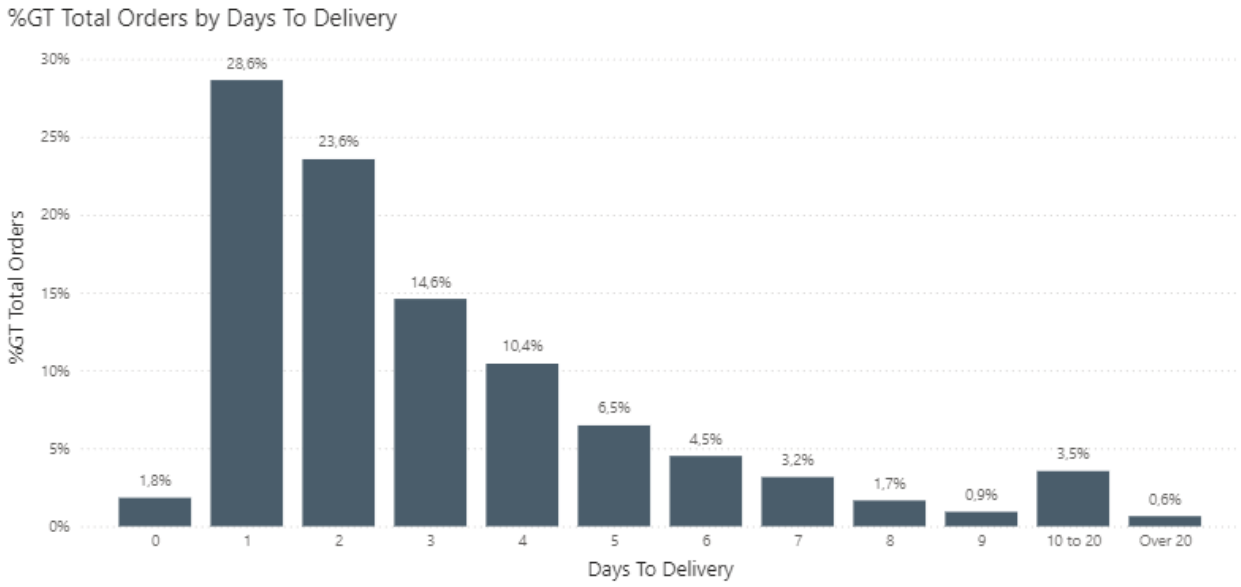
Na podstawie analizy statystyk z tabeli o nazwie „olist_geolocation_dataset" (Tab. 2) można wysnuć następujące wnioski:

- **Kompletność danych**: Dla żadnej z kolumn nie występują puste rekordy, co oznacza, że zbiór jest kompletny.
- **Konsolidacja danych**: W kontekście projektu postanowiono stworzyć tabelę wymiarów z kluczem głównym „zip_code_prefix". Analiza wykazała dużą liczbę powielonych wierszy, w których dla tego samego „zip_code_prefix" występowały różne współrzędne geograficzne. Taki poziom szczegółowości nie był potrzebny, dlatego zaplanowano oczyszczenie tabeli.
- **Zakres wartości**: Wartości w kolumnie „zip_code_prefix" mieszczą się w zakresie od 1001 do 99990 i nie zawierają wartości 0, co wskazuje na ich poprawność.
- **Jakość danych**: Kolumna „city" zawiera 8010 unikalnych wartości, co przekracza liczbę rzeczywistych miast w Brazylii. Ponadto jedna z widocznych nazw miast zawiera znak specjalny „*", co sugeruje potrzebę oczyszczenia i standaryzacji danych.

Proces ten pozwolił na identyfikację brakujących wartości, duplikatów oraz innych potencjalnych problemów. Dla zmiennych numerycznych warto również obliczyć średnią, medianę, kwartyle i odchylenie standardowe. Przebadanie całego zbioru umożliwia wstępne zaplanowanie modelu powiązań między tabelami i stanowi punkt wejściowy do dalszej analizy.

[5] Tanimura C.: *Analiza danych z wykorzystaniem SQL-a: Zaawansowane techniki przekształcania danych we wnioski*, Helion, Gliwice 2022, s. 36

Kolejnym istotnym etapem zapoznania się ze zbiorem danych jest analiza rozkładu danych. W projekcie zastosowano histogramy w celu wizualizacji częstotliwości występowania różnych wartości w danych. Przykładowo, grupowanie zamówień według czasu dostawy przy użyciu techniki tworzenia kubełków (ang. binning) umożliwiło podział danych ciągłych na przedziały (np. 10-20 dni, 20+ dni), co ułatwiło interpretację wyników. W wyniku tej analizy (Rys. 2) ustalono, że 3,5% zamówień było dostarczanych w ciągu 10–20 dni[6].




**Rysunek 2.** Rozkład zamówień według czasu dostawy (źródło: opracowanie własne)

Histogramy stanowią efektywne narzędzie do identyfikacji nieoczekiwanych wzorców oraz wykrywania danych rzadkich. W wielu przypadkach zastosowanie przedziałów okazuje się niezbędne dla zwiększenia czytelności wyników oraz umożliwienia formułowania trafnych wniosków. Kluczowym elementem tego procesu jest wybór odpowiedniej techniki podziału, choć często trudno wskazać jeden obiektywnie najlepszy sposób. Przykładowo, w przypadku zbiorów danych, w których wartości maksymalne są o rzędy wielkości większe od wartości minimalnych, efektywnym rozwiązaniem może być tworzenie kubełków oparte na logarytmach. Z kolei w innych sytuacjach bardziej odpowiednią metodą może okazać się zastosowanie techniki podziału na równe przedziały (ang. n przedziałów)[7].

Zapewnienie wysokiej jakości danych było kolejnym istotnym etapem eksploracyjnej analizy danych. Dane o niskiej jakości mogą prowadzić do błędnych wniosków

[6] Tanimura C.: *Analiza danych z wykorzystaniem SQL-a*, s. 39
[7] Tanimura C.: *Analiza danych z wykorzystaniem SQL-a*, s. 36-42

i niewiarygodnych raportów. Proces analizy jakości danych obejmował identyfikację braków, duplikatów, anomalii oraz innych niespójności[8].

Na przykład analiza zamówień według kwartałów (Tab. 3) wykazała:

- **Bardzo niską liczbę zamówień w wybranych kwartałach**: W Q3 2016, Q4 2016 oraz Q4 2018 liczba zamówień była znacząco niższa w porównaniu do innych okresów, co budziło podejrzenia o niekompletność danych.
- **Nieprawidłowe dane dotyczące czasu dostawy**: W kwartałach Q3 2016 i Q4 2016 średni czas dostawy był nieproporcjonalnie wysoki w porównaniu do innych kwartałów, co mogło wskazywać na błędy w zapisie lub procesie gromadzenia danych. W Q4 2018 brakowało danych dotyczących czasu dostawy, co również wskazywało na problemy z jakością danych.
- **Niekompletne dane przychodowe**: W Q3 2016 i Q4 2016 przychody były niewielkie w porównaniu do innych kwartałów, co mogło sugerować błędy w rejestrowaniu transakcji. W Q4 2018 całkowicie brakowało danych dotyczących przychodów, co uniemożliwiało ich uwzględnienie w analizie.

| Quarter | Total Orders | Average Delivery Time | Total Revenue |
|---|---|---|---|
| Q3 2016 | 4 | 55 days, 23:45:60 | R$ 143,46 |
| Q4 2016 | 325 | 20 days, 10:12:56 | R$ 45 690,32 |
| Q1 2017 | 5262 | 13 days, 00:25:29 | R$ 803 905,48 |
| Q2 2017 | 9349 | 12 days, 11:49:03 | R$ 1 437 005,81 |
| Q3 2017 | 12642 | 11 days, 12:45:53 | R$ 1 896 394,70 |
| Q4 2017 | 17848 | 14 days, 09:18:19 | R$ 2 718 511,32 |
| Q1 2018 | 21208 | 15 days, 17:56:48 | R$ 3 135 140,19 |
| Q2 2018 | 19979 | 10 days, 18:33:13 | R$ 3 258 038,57 |
| Q3 2018 | 12820 | 8 days, 08:09:07 | R$ 1 996 523,46 |
| Q4 2018 | 4 | None | |
| **Total** | **99441** | **12 days, 13:40:23** | **R$ 15 291 353,31** |

**Tabela 3.** Zestawienie zamówień, średnich czasów dostaw i przychodów według kwartałów (źródło: opracowanie własne)

W rezultacie podjęto decyzję o ograniczeniu zakresu analizy do okresu od Q1 2017 do Q3 2018 w celu zapewnienia spójności i wiarygodności analizy. Należy jednak zauważyć, że

[8] Tanimura C.: *Analiza danych z wykorzystaniem SQL-a*, s. 43-45

ograniczenie zakresu czasowego miało swoje konsekwencje. Wykluczenie części danych mogło prowadzić do utraty pełnego obrazu analizowanych trendów i zjawisk. Mimo to decyzja ta była uzasadniona, ponieważ zapewniła wysoką jakość i rzetelność końcowych wyników.

Podsumowując, eksploracyjna analiza danych odegrała kluczową rolę w zrozumieniu struktury, jakości i potencjalnych problemów w zbiorze danych, stanowiąc fundament dla dalszych etapów analizy w ramach raportu e-commerce. Dzięki zastosowaniu różnorodnych technik, takich jak profilowanie danych, histogramy czy analiza jakości, możliwe było nie tylko zidentyfikowanie braków, duplikatów i niespójności, ale również podjęcie świadomych decyzji naprawczych, takich jak oczyszczenie tabel czy ograniczenie zakresu czasowego analizy.

# Rozdział III. Projektowanie i budowa modelu danych

## 3.1. Wprowadzenie do modelu danych

Model danych odgrywa nadrzędną rolę w każdym projekcie analitycznym realizowanym w Power BI, ponieważ jego konstrukcja bezpośrednio wpływa na wydajność, skuteczność oraz dokładność generowanych raportów i wizualizacji. Starannie zaprojektowany model umożliwia nie tylko efektywne przetwarzanie dużych zbiorów danych, ale również logiczne ich powiązanie, co jest fundamentem procesów analitycznych. Poprawna konstrukcja modelu danych minimalizuje ryzyko błędów interpretacyjnych dzięki zapewnieniu spójności oraz jednoznaczności w analizie danych.

Jednym z istotnych aspektów dobrze zaprojektowanego modelu jest jego zdolność do adaptacji do zmieniających się wymagań biznesowych. W pracy z Power BI szczególną uwagę należy zwrócić na optymalizację modelu pod kątem wydajności, co jest szczególnie istotne w przypadku dużych zbiorów danych. Wykorzystanie odpowiednich technik modelowania, takich jak schemat gwiazdy, pozwala na uproszczenie struktury danych, co przekłada się na większą przejrzystość i łatwość użytkowania projektu. Dzięki temu tworzone raporty są bardziej intuicyjne dla użytkowników końcowych i wspierają ich w podejmowaniu trafnych decyzji.

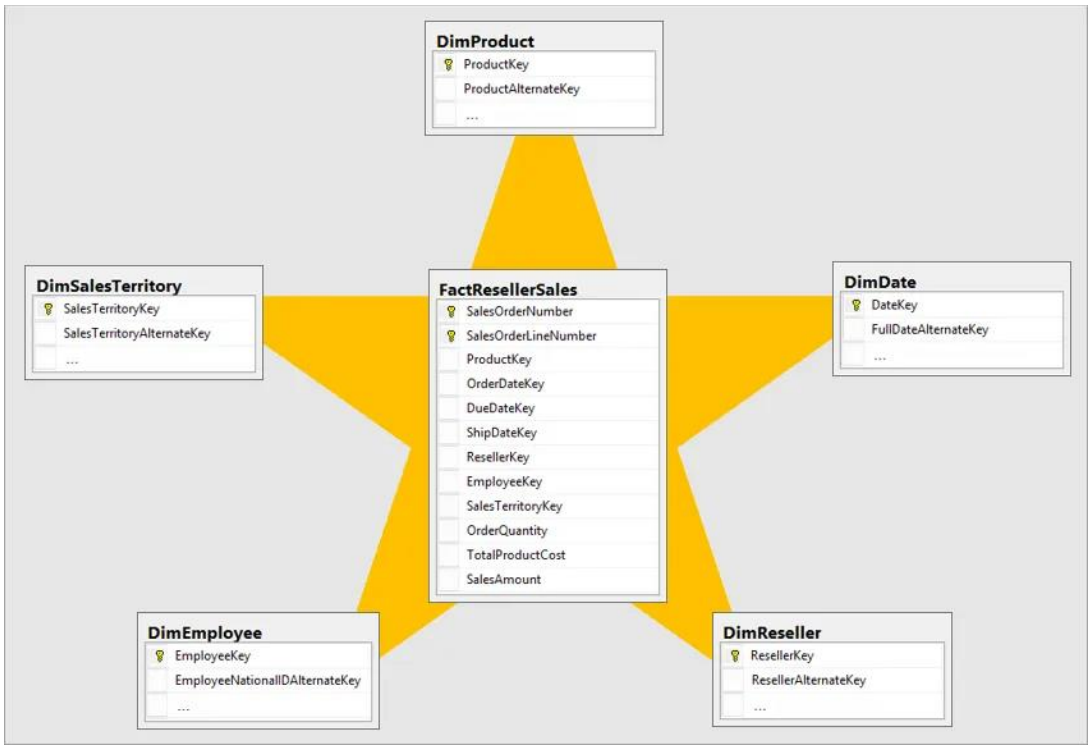
W procesie budowy modelu danych istotnym etapem jest również zastosowanie narzędzi ETL, które umożliwiają przygotowanie danych poprzez ich oczyszczenie, standaryzację i zapewnienie spójności. Dzięki temu załadowane dane są nie tylko poprawne, ale także gotowe do wykorzystania w analizie, co znacząco zwiększa jakość i wartość końcowych wyników analitycznych.

## 3.2. Zaplanowanie schematu gwiazdy

### 3.2.1. Co to jest schemat gwiazdy?

Schemat gwiazdy to jedna z najbardziej popularnych metod organizacji modeli danych stosowana w analityce biznesowej, w tym w Power BI. Charakteryzuje się prostą, a zarazem efektywną strukturą, która wspiera zarówno zarządzanie danymi, jak i ich wydajną analizę. Układ schematu przypomina kształt gwiazdy, ponieważ w centrum znajduje się tabela faktów, otoczona tabelami wymiarów.

Tabela faktów gromadzi dane ilościowe, takie jak sprzedaż, przychody czy czas realizacji zamówień, które stanowią podstawę analiz biznesowych. Z kolei tabele wymiarów zawierają dane opisowe, takie jak informacje o produktach, lokalizacjach geograficznych czy danych klienta, co pozwala na szczegółowe grupowanie i analizę danych z tabeli faktów. Kluczowym elementem schematu gwiazdy jest relacja między tabelą faktów a tabelami wymiarów, umożliwiająca skuteczne filtrowanie i agregowanie danych w procesie analitycznym.




**Rysunek 3.** Struktura schematu gwiazdy z tabelą faktów i tabelami wymiarów (źródło: [17])

Schemat gwiazdy jest uznawany w literaturze i praktyce analityki danych za standardową i rekomendowaną strukturę modelu danych. Jak zauważa Alberto Ferrari w artykule „The Importance of Star Schemas in Power BI", schemat gwiazdy nie tylko zwiększa wydajność narzędzi analitycznych, takich jak Power BI, ale również upraszcza analizę danych. Według Ferrariego, ta struktura minimalizuje problemy związane z redundancją danych, poprawiając ich czytelność i zrozumiałość dla użytkowników końcowych. Dzięki uproszczonej konstrukcji i ograniczeniu liczby relacji między tabelami, analitycy mogą efektywniej przeprowadzać zapytania oraz lepiej interpretować wyniki.

Ferrari podkreśla również, że Power BI, podobnie jak inne narzędzia analityczne, jest zoptymalizowane pod kątem pracy z modelami opartymi na schemacie gwiazdy. Tego rodzaju

modele pozwalają na bardziej efektywne przetwarzanie dużych zbiorów danych, unikając problemów z wydajnością. Co więcej, struktura ta zmniejsza ryzyko błędnej interpretacji danych przez użytkowników końcowych, co ma kluczowe znaczenie w kontekście podejmowania decyzji biznesowych[9].

Wykorzystanie schematu gwiazdy w praktyce nie tylko poprawia wydajność analizy danych, ale również zwiększa ich spójność i dokładność. W przypadku dużych zbiorów danych, gdzie złożoność modeli może wpływać na czas przetwarzania zapytań oraz jakość wyników, schemat gwiazdy jest uznawany za najlepszą praktykę w branży. Dzięki swoim zaletom, struktura ta jest preferowanym rozwiązaniem w projektach wymagających zaawansowanej analizy danych, co potwierdzają zarówno literatura naukowa, jak i doświadczenia ekspertów z branży analitycznej.

### 3.2.2. Planowanie schematu gwiazdy w projekcie

Proces planowania schematu gwiazdy w analizowanym projekcie stanowił kluczowy etap w przygotowaniu danych do dalszej analizy w Power BI. Celem tego etapu było zdefiniowanie docelowych tabel modelu i ich kolumn, które powstaną podczas procesu ETL.

Model danych został zaplanowany z trzema tabelami faktów:

- **Orders** – zawiera główne informacje związane z realizacją zamówień.
- **Payments** – dostarcza szczegółowych danych o płatnościach.
- **Reviews** – przechowuje informacje na temat ocen i opinii klientów.

Wokół tych tabel faktów rozplanowano zestaw tabel wymiarów, które umożliwią analizę danych w różnych przekrojach. Do najważniejszych tabel wymiarów należą:

- **Order** – zawiera szczegółowe informacje dotyczące danego zamówienia, takie jak identyfikator klienta, data zakupu, status zamówienia.
- **Customer** – zawiera informacje o klientach, takie jak unikalne identyfikatory, kody pocztowe.
- **Product** – przechowuje szczegóły dotyczące produktów, w tym kategorię, wagę, wymiary, ilość dostępnych zdjęć.

[9] https://www.sqlbi.com/articles/the-importance-of-star-schemas-in-power-bi/, [dostęp 30.12.2024]

- **Seller** – dostarcza danych o sprzedawcach, takich jak kody pocztowe.
- **Geolocation** – dostarcza informacji o lokalizacjach geograficznych na podstawie kodów pocztowych.
- **State** – zawiera dane o stanach, takie jak skrót czy region.
- **Payment Type** – zawiera informacje o metodzie płatności.

Ten podział na tabele faktów i wymiarów pozwolił na optymalizację struktury modelu oraz sprawne zarządzanie relacjami między danymi.

Podczas planowania schematu gwiazdy w projekcie przyjęto zasady nazewnictwa zaproponowane przez Marco Russo w artykule „Data Import Best Practices in Power BI". Zastosowanie spójnych i intuicyjnych nazw tabel oraz pól jest niezwykle istotne, gdyż wpływa na czytelność i efektywność modelu danych. W modelu przyjęto następujące konwencje:

- **Jednoznaczność i precyzja** – nazwy tabel i kolumn dokładnie odzwierciedlają ich zawartość. Na przykład kolumna „Height (cm)" w tabeli Product jednoznacznie wskazuje na wysokość produktu podaną w centymetrach.
- **Krótkie, lecz opisowe nazwy** – unikanie zbędnych skrótów lub niejasnych terminów. Na przykład nazwa „Product Id" jednoznacznie określa identyfikator produktu, bez potrzeby dodatkowych wyjaśnień.
- **Użycie wielkich liter dla nazw tabel** – np. „Customer", „Orders", co zwiększa przejrzystość modelu w interfejsie Power BI.
- **Nadanie nazw tabelom wymiarów w liczbie pojedynczej** – np. „Customer", „Product", co jest zgodne z najlepszymi praktykami przedstawionymi w literaturze.
- **Nadanie nazw tabelom faktów w liczbie mnogiej** – np. „Orders", „Payments", „Reviews", co wskazuje na zawieranie wielu rekordów danych związanych z faktami analitycznymi.
- **Unikanie powtórzeń** – każda kolumna w modelu została nazwana w sposób, który eliminuje redundancję[10].

[10] https://www.sqlbi.com/articles/data-import-best-practices-in-power-bi/, [dostęp 30.12.2024]

Przyjęte zasady pozwalają na uzyskanie czytelnego modelu zarówno dla analityków, jak i użytkowników końcowych, co znacząco przyspiesza proces budowy wizualizacji.

## 3.3. Proces ETL (Ekstrakcja, Transformacja, Ładowanie)

### 3.3.1. Wprowadzenie do procesu ETL

Proces ETL stanowi główny etap w przygotowywaniu danych do analizy w systemach Business Intelligence, takich jak Power BI. Skrót ETL odnosi się do trzech głównych faz procesu: ekstrakcji danych (ang. Extract), ich transformacji (ang. Transform) oraz ładowania (ang. Load) do modelu danych. Każdy z tych etapów pełni istotną rolę w przekształcaniu surowych danych w uporządkowaną strukturę gotową do analizy i wizualizacji.

Pierwszym krokiem procesu ETL jest ekstrakcja, polegająca na pobieraniu danych z różnorodnych źródeł, takich jak bazy danych, pliki tekstowe, systemy ERP czy usługi chmurowe. Ważnym aspektem tego etapu jest zapewnienie kompletności i integralności zbieranych danych, tak aby wszystkie niezbędne informacje zostały prawidłowo przechwycone. Następnie dane poddawane są transformacji, podczas której są czyszczone, standaryzowane i dostosowywane do wymagań projektu. Przykładowe operacje na tym etapie obejmują ujednolicenie formatów danych, eliminację błędów i braków w rekordach oraz tworzenie nowych kolumn obliczeniowych w celu wzbogacenia analizy. Ostatni etap, ładowanie, polega na zapisaniu przetworzonych danych w modelu, co umożliwia ich efektywne wykorzystanie w analizach i wizualizacjach.

Proces ETL pełni kluczową funkcję w zapewnieniu kompletności, spójności i optymalizacji danych, które są podstawą dalszych działań analitycznych. W kontekście Power BI prawidłowo zaprojektowany proces ETL nie tylko wpływa na szybkość generowania raportów i wizualizacji, ale również ułatwia zarządzanie modelem danych w dłuższym horyzoncie czasowym, co przyczynia się do zwiększenia efektywności i elastyczności analizy biznesowej[11].

### 3.3.2. Optymalizacja modelu danych

Wydajność modelu danych w Power BI w dużej mierze zależy od zastosowanych technik optymalizacyjnych oraz zasad projektowych i transformacyjnych, przyjętych podczas

[11] Alexander M., Decker J., Wehbe B.: *Analizy Business Intelligence: Zaawansowane wykorzystanie Excela*, Helion, Gliwice 2019, s. 141-165

przygotowywania danych. Jeszcze przed rozpoczęciem procesu przekształcania tabel do finalnego modelu danych kluczowe jest zrozumienie zasad, których przestrzeganie pozwala na zbudowanie modelu zoptymalizowanego pod kątem wydajności. Proces optymalizacji obejmuje działania mające na celu zmniejszenie złożoności modelu, poprawę szybkości obliczeń oraz ograniczenie zajmowanej pamięci. W tym kontekście istotne znaczenie mają ustawienia Power BI, zarządzanie danymi w modelu oraz odpowiednie przekształcenia tabel i kolumn.

Jednym z pierwszych kroków optymalizacyjnych jest dostosowanie ustawień Power BI. Na przykład, wyłączenie funkcji „Auto Date/Time", która automatycznie tworzy ukryte tabele kalendarzowe dla każdego pola daty w modelu, jest istotne w dużych projektach. Choć funkcja ta może być przydatna w mniejszych analizach, w przypadku większych zbiorów danych prowadzi do niepotrzebnego wzrostu rozmiaru modelu i obciążenia systemu. Wyłączenie tej opcji pozwala na wdrożenie własnej tabeli kalendarzowej, dostosowanej do specyficznych wymagań projektu, co zwiększa efektywność modelu.

Kolejną powszechną praktyką jest rozdzielanie pól typu „DateTime" na oddzielne kolumny dla daty i czasu. Taki podział zmniejsza złożoność modelu, ogranicza rozmiar słownika i kardynalność danych, co w efekcie poprawia wydajność obliczeń DAX.

Optymalizacja obejmuje również przegląd i ocenę każdej tabeli pod kątem jej przydatności w analizie. Kolumny oraz tabele, które nie są wykorzystywane w raportach ani analizach, powinny zostać usunięte, aby zmniejszyć rozmiar modelu i ograniczyć obciążenie pamięci. Ma to szczególne znaczenie w przypadku dużych zbiorów danych, gdzie każda zbędna kolumna zwiększa wielkość słownika i czas przetwarzania zapytań.

Tabela "Order" przed wykonaniem transformacji

| | order_id | customer_id | order_status | order_purchase_timestamp |
|---|---|---|---|---|
| 1 | 00010242fe8c5a6d1ba2dd792cb162... | 3ce436f183e68e07877b285a838db1... | delivered | 13.09.2017 08:59:02 |
| 2 | 00018f77f2f0320c557190d7a144bdd3 | f6dd3ec061db4e3987629fe6b26e5cce | delivered | 26.04.2017 10:53:06 |
| 3 | 000229ec398224ef6ca0657da4fc703e | 6489ae5e4333f3693df5ad4372dab6... | delivered | 14.01.2018 14:33:31 |

Tabela "Order" po wykonaniu transformacji

| | Order Id | Customer Unique Id | Status Id | Purchase Date | Purchase Time |
|---|---|---|---|---|---|
| 1 | 1 | 68585 | 1 | 02.10.2017 | 10:56:33 |
| 2 | 2 | 74977 | 1 | 24.07.2018 | 20:41:37 |
| 3 | 3 | 555 | 1 | 08.08.2018 | 08:38:49 |

**Rysunek 4.** Porównanie struktury tabeli „Order" przed i po transformacji (źródło: opracowanie własne)

Podobnie, włączenie do modelu wyłącznie niezbędnych danych znacząco poprawia wydajność. Usuwanie zbędnych tabel oraz kolumn pozwala na zmniejszenie całkowitego rozmiaru modelu, co z kolei skraca czas odświeżania raportów oraz optymalizuje zarządzanie pamięcią. Dzięki temu model staje się bardziej efektywny i łatwiejszy w utrzymaniu.

Równie ważne jest planowanie odświeżania danych w modelu. Zaleca się dostosowanie częstotliwości odświeżania do rzeczywistych potrzeb biznesowych, aby uniknąć zbędnego obciążenia systemu. W przypadku dużych zbiorów danych można rozważyć odświeżanie tylko najnowszych fragmentów tabel, zamiast przetwarzać całość zbioru, co znacząco obniża zużycie zasobów systemowych.

| Tabela / Kolumna | Kardynalność | Rozmiar kolumny (bajt) | Rozmiar danych (bajt) | Rozmiar słownika (bajt) | Rozmiar hierarchii (bajt) | Rozmiar relacji (bajt) |
|---|---|---|---|---|---|---|
| **Order przed transformacją** | **99 441** | **28 880 706** | **1 587 728** | **22 767 762** | **4 525 216** | **530 384** |
| customer_id | 99 441 | 5 570 621 | 265 312 | 4 509 773 | 795 536 | |
| order_approved_at | 90 734 | 3 908 252 | 264 920 | 2 917 444 | 725 888 | |
| order_delivered_carrier_date | 81 019 | 3 720 040 | 260 560 | 2 811 320 | 648 160 | |
| order_delivered_customer_date | 95 665 | 3 999 804 | 257 424 | 2 977 052 | 765 328 | |
| order_estimated_delivery_date | 459 | 32 600 | 8 624 | 20 296 | 3 680 | |
| order_id | 99 441 | 5 570 381 | 265 312 | 4 509 533 | 795 536 | |
| order_purchase_timestamp | 98 875 | 6 061 232 | 265 312 | 5 004 912 | 791 008 | |
| order_status | 8 | 17 520 | 136 | 17 304 | 80 | |
| **Order po transformacji** | **99 441** | **16 501 956** | **1 798 120** | **11 750 780** | **2 953 056** | **256 280** |
| Approval Purchase Date | 612 | 131 008 | 104 560 | 21 536 | 4 912 | |
| Approval Purchase Time | 41 747 | 1 938 696 | 198 720 | 1 405 992 | 333 984 | |
| Carrier Delivery Date | 548 | 72 660 | 47 224 | 21 036 | 4 400 | |
| Carrier Delivery Time | 37 003 | 1 849 292 | 195 456 | 1 357 804 | 296 032 | |
| Customer Delivery Date | 646 | 149 224 | 122 208 | 21 832 | 5 184 | |
| Customer Delivery Time | 41 100 | 1 918 584 | 193 104 | 1 396 664 | 328 816 | |
| Customer Unique Id | 96 096 | 3 515 920 | 265 312 | 2 481 824 | 768 784 | |
| Estimated Delivery Date | 459 | 116 448 | 92 472 | 20 296 | 3 680 | |
| Order Id | 99 441 | 3 556 052 | 265 312 | 2 495 204 | 795 536 | |
| Purchase Date | 634 | 141 256 | 114 472 | 21 696 | 5 088 | |
| Purchase Time | 50 818 | 3 111 000 | 199 016 | 2 505 424 | 406 560 | |
| Status Id | 8 | 1 560 | 136 | 1 344 | 80 | |
| **Order Status po transformacji** | **8** | **19 336** | **400** | **18 776** | **160** | |
| Status | 8 | 17 520 | 136 | 17 304 | 80 | |
| Status Id | 8 | 1 560 | 136 | 1 344 | 80 | |

**Tabela 4.** Porównanie parametrów wydajnościowych tabeli „Order” przed i po optymalizacji (źródło: opracowanie własne)

Przykład zastosowania powyższych zasad w projekcie stanowi tabela „Order", która została zoptymalizowana poprzez zastosowanie następujących działań:

- **Zmiana indeksów**: Wartości identyfikatorów, takich jak „Order Id", zostały przekształcone z ciągów znaków na liczby całkowite, co zmniejszyło rozmiary słownika danych.
- **Rozdzielenie kolumn typu „DateTime"**: Pole „Order Purchase Timestamp" zostało podzielone na kolumny „Purchase Date" i „Purchase Time", co zmniejszyło kardynalność danych oraz rozmiar słownika.
- **Normalizacja kolumny „Order Status"**: W celu poprawy organizacji danych utworzono osobną tabelę wymiarów „Order Status". Normalizacja tej kolumny nie wpłynęła bezpośrednio na redukcję rozmiaru tabeli, ponieważ silnik VertiPaq automatycznie generuje słownik dla każdej kolumny, eliminując koszty związane z duplikowaniem statusów. Jednakże należy podkreślić, że denormalizacja nie zawsze jest korzystnym rozwiązaniem – szczególnie w przypadku bardzo dużych zbiorów danych może mieć negatywny wpływ na wydajność modelu.

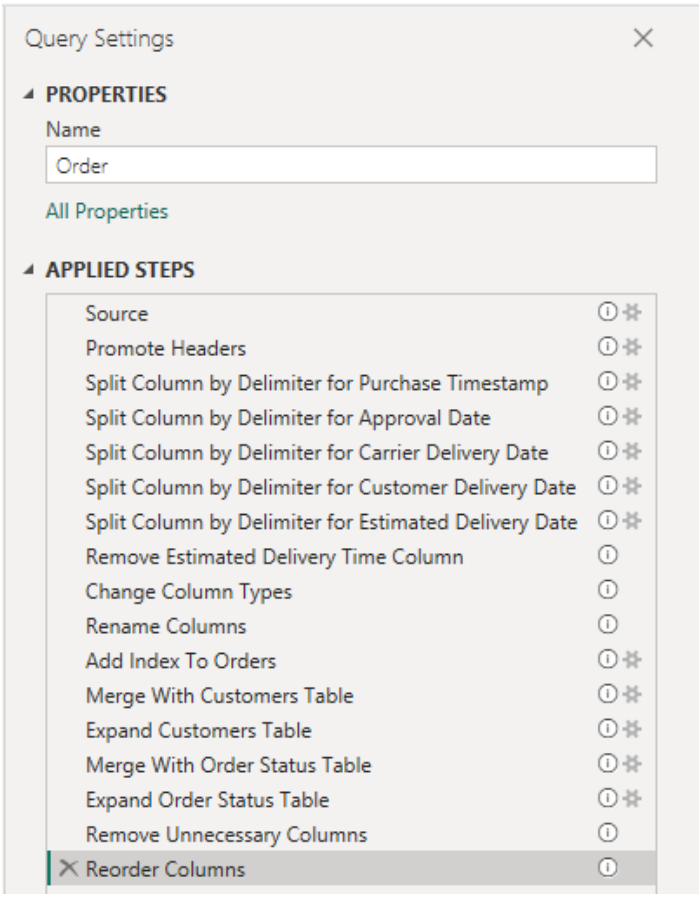
Efektem tych działań było zmniejszenie całkowitego rozmiaru tabeli o 43%. Ponadto kardynalność kolumn również uległa znacznemu zmniejszeniu, co istotnie wpłynęło na poprawę wydajności obliczeń DAX. Wyniki te dowodzą, że odpowiednie techniki optymalizacji przynoszą realne korzyści, zarówno w kontekście wydajności, jak i efektywności zarządzania modelem danych[12].

### 3.3.3. Przekształcanie danych w Power Query

Power Query to zaawansowane narzędzie ETL dostępne w programie Power BI, które umożliwia przekształcanie i przygotowanie danych do analizy w sposób intuicyjny i wydajny. Dzięki funkcji interfejsu graficznego użytkownicy mogą bez konieczności pisania kodu realizować skomplikowane operacje, takie jak czyszczenie danych, ujednolicanie formatów czy scalanie tabel. Power Query oferuje również zaawansowane możliwości automatyzacji, filtrowania oraz dostosowania danych, co czyni go niezwykle przydatnym w procesach modelowania danych i przygotowywania raportów.

[12] https://ashwinijain.medium.com/how-to-optimize-your-power-bi-data-model-1da03f48ec8e, [dostęp 31.12.2024]

Jednym z kluczowych elementów pracy w Power Query jest możliwość stosowania kroków transformacji, które są automatycznie zapisywane i mogą być dynamicznie odświeżane przy zmianach w źródłach danych. To narzędzie jest szczególnie cenione w środowisku analityki danych za swoją elastyczność i integrację z różnorodnymi źródłami danych.




**Rysunek 5.** Kroki transformacji danych w tabeli „Order” w Power Query (źródło: opracowanie własne)

Tabela „Order”, która była szczegółowo opisana w poprzednim podrozdziale, przeszła kompleksowy proces transformacji w Power Query. Na załączonym zrzucie ekranu (Rys. 5) przedstawiono poszczególne kroki tego procesu, których łączna liczba wynosi 17. Transformacje te były niezbędne, aby dostosować dane do wymagań modelu analitycznego oraz poprawić ich strukturę i jakość. Kluczowe kroki transformacji:

1. **Source i Promote Headers**: Pierwsze kroki obejmowały załadowanie danych źródłowych oraz promowanie nagłówków kolumn. Zapewniło to odpowiednie zdefiniowanie struktury tabeli na etapie początkowym.

2. **Split Column by Delimiter**: Rozdzielenie kolumn typu „DateTime”, takich jak „order_purchase_timestamp”, na osobne kolumny „Purchase Date” oraz „Purchase

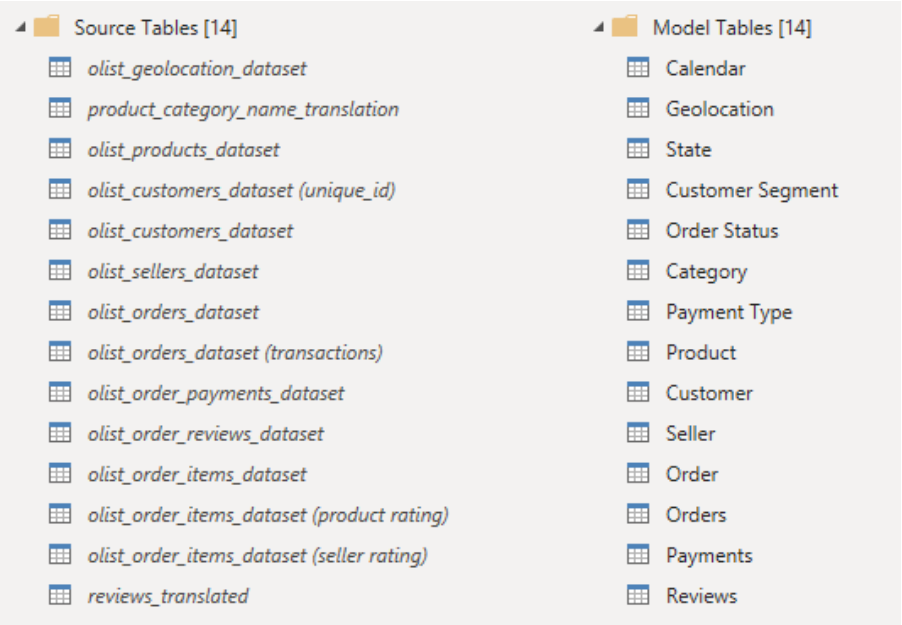
Time”. Podobne operacje przeprowadzono na innych kolumnach dat i godzin, takich jak „Approval Date” czy „Carrier Delivery Date”. Taki zabieg pozwolił zmniejszyć kardynalność danych oraz rozmiar słowników, co pozytywnie wpłynęło na wydajność modelu.

3. **Change Column Types i Rename Columns**: Zmiana typów danych w kolumnach, takich jak indeksy, oraz dostosowanie nazw kolumn do konwencji przyjętej w modelu. Te kroki zapewniły spójność i zgodność danych ze standardami modelu.
4. **Merge Steps**: Kilka kroków związanych było ze scalaniem tabel źródłowych, takich jak „Customer” czy „Order Status”. Scalanie tych tabel było konieczne, aby przekształcić oryginalne indeksy oparte na ciągach znaków w nowe indeksy liczbowe, co w znacznym stopniu zredukowało rozmiar słowników w modelu.
5. **Remove Unnecessary Columns**: Usunięcie zbędnych kolumn, które nie były wykorzystywane w analizach, co zmniejszyło całkowity rozmiar tabeli oraz zwiększyło czytelność danych.
6. **Reorder Columns**: Ostateczne kroki obejmowały uporządkowanie kolejności kolumn, co poprawiło przejrzystość tabeli.

Każdy z tych kroków był niezbędny dla uzyskania ostatecznej struktury tabeli „Order” i przygotowania jej do załadowania do modelu danych. Proces ten pokazuje, jak ważne jest precyzyjne planowanie i realizacja transformacji w Power Query, aby osiągnąć maksymalną wydajność oraz jakość danych.

Optymalizacja modelu danych w Power BI wymaga odpowiedniej konfiguracji ładowania i odświeżania tabel. W analizowanym projekcie zastosowano następujące zasady:

- Wyłączenie ładowania zbędnych tabel: Tabele źródłowe, takie jak „olist_order_status_dataset” czy „olist_customers_dataset”, nie zostały załadowane do modelu, co zmniejszyło obciążenie raportu.
- Wyłączenie odświeżania dla stałych tabel wymiarów: W przypadku tabel, takich jak „Geolocation” czy „State”, których dane są niezmienne, wyłączono opcję odświeżania podczas procesu aktualizacji raportu, co przyczyniło się do skrócenia czasu odświeżania.




**Rysunek 6.** Tabele źródłowe i docelowe w procesie transformacji danych w Power Query (źródło: opracowanie własne)

### 3.3.4. Oczyszczanie danych w Power Query

Tabela „Geolocation" została poddana szczegółowemu procesowi oczyszczania i przekształcania danych w Power Query, aby uzyskać strukturę dostosowaną do wymagań modelu analitycznego. Proces ten obejmował zarówno zaawansowane techniki grupowania, jak i redukcji danych, co pozwoliło na poprawę jakości oraz spójności zbioru danych.

```
(...)
#"Add Longitude Range Start" = Table.AddColumn(
  #"Add Latitude Range End",
  "lng_range_1",
  each
    if [geolocation_lng] < 0 then
      (Number.RoundUp([geolocation_lng] / 3)) * 3
    else
      (Number.RoundDown([geolocation_lng] / 3)) * 3,
  Int64.Type
),
#"Add Longitude Range End" = Table.AddColumn(
  #"Add Longitude Range Start",
  "lng_range_2",
  each if [lng_range_1] < 0 then [lng_range_1] - 3 else [lng_range_1] + 3,
  Int64.Type
),
(...)
```

**Listing 1.** Zaokrąglanie współrzędnych geograficznych

Pierwszym etapem była normalizacja współrzędnych geograficznych poprzez zaokrąglenie wartości do pełnych liczb będących 0 lub dzielnikami liczby 3 w celu uproszczenia analizy i umożliwienia efektywnego grupowania (Listing 1). Ten krok miał na celu zredukowanie liczby unikalnych wartości współrzędnych i ułatwienie dalszych etapów grupowania.

```
(...)
#"Combine Longitude Ranges" = Table.AddColumn(
  #"Combine Latitude Ranges",
  "lng_range",
  each
    if [lng_range_1] < 0 then
      "(" & Number.ToText([lng_range_1]) & ") - (" & Number.ToText([lng_range_2]) & ")"
    else
      Number.ToText([lng_range_1]) & " - " & Number.ToText([lng_range_2]),
  Text.Type
),
#"Create Coordinate Range" = Table.AddColumn(
  #"Combine Longitude Ranges",
  "coordinates_range",
  each "[" & [lng_range] & "] ; [" & [lat_range] & "]",
  Text.Type
),
(...)
```

**Listing 2.** Łączenie zakresów współrzędnych geograficznych

Następnie zaokrąglone zakresy długości i szerokości geograficznej zostały połączone w jedną kolumnę, która reprezentowała pełen zakres współrzędnych geograficznych (Listing 2). Operacja ta pozwoliła na stworzenie jednoznacznego identyfikatora dla każdego zakresu współrzędnych, który odpowiadał kwadratowi o wymiarach 3x3 na siatce geograficznej.

Kolejnym krokiem było grupowanie danych według kodów pocztowych („geolocation_zip_code_prefix”) oraz utworzonego zakresu współrzędnych (Listing 3). Grupowanie to miało na celu obliczenie liczby wystąpień kodu pocztowego („geolocation_zip_code_prefix”) w danym kwadracie 3x3 na siatce oraz wyliczenie średnich wartości szerokości („latitude”) i długości („longitude”) geograficznej dla tego kodu pocztowego w obrębie danego kwadratu. To podejście pozwoliło na zidentyfikowanie najbardziej reprezentatywnych współrzędnych dla każdego kodu pocztowego, jednocześnie uwzględniając częstotliwość jego występowania w określonych zakresach współrzędnych.

```
(...)
#"Group By Zip And Coordinates" = Table.Group(
  #"Select Necessary Columns",
  {"geolocation_zip_code_prefix", "coordinates_range"},
  {
    {"latitude", each List.Average([geolocation_lat]), type nullable number},
    {"longitude", each List.Average([geolocation_lng]), type nullable number},
    {"number_of_coordinates", each Table.RowCount(_), Int64.Type}
  }
),
(...)
```

**Listing 3.** Obliczanie średnich współrzędnych dla kodów pocztowych w obrębie kwadratów 3x3

Ostatecznym etapem było usunięcie duplikatów kodów pocztowych („zip_code_prefix"). W tym celu wiersze zostały posortowane według liczby wystąpień danego kodu pocztowego w konkretnym kwadracie 3x3 na siatce geograficznej, w porządku malejącym. Następnie, na podstawie najbardziej reprezentatywnego kwadratu, obliczono ostateczne średnie współrzędne („latitude" i „longitude") dla każdego kodu pocztowego (Listing 4). Po tej operacji usunięto duplikaty na podstawie kolumny „geolocation_zip_code_prefix", zachowując dane, które najlepiej odzwierciedlały lokalizację danego kodu pocztowego.

```
(...)
#"Sort By Coordinate Density" = Table.Sort(
  #"Group By Zip And Coordinates",
  {{"number_of_coordinates", Order.Descending}}
),
#"Remove Duplicate Zip Codes" = Table.Distinct(
  #"Sort By Coordinate Density",
  {"geolocation_zip_code_prefix"}
),
(...)
```

**Listing 4.** Sortowanie i usuwanie duplikatów kodów pocztowych

Proces oczyszczania i przekształcania danych w tabeli „Geolocation" pozwolił na znaczną redukcję złożoności zbioru danych, jednocześnie zachowując jego użyteczność analityczną. Dzięki zastosowaniu technik, takich jak zaokrąglanie współrzędnych, grupowanie i usuwanie duplikatów, tabela została przygotowana do efektywnego załadowania do modelu Power BI.

## 3.4. Budowanie modelu danych

### 3.4.1. Tworzenie modelu danych

Po odpowiednim oczyszczeniu i załadowaniu tabel do modelu danych, tworzenie jego struktury staje się jednym z kluczowych etapów procesu analizy, szczególnie w narzędziach takich jak Power BI. Prawidłowo zaprojektowany model w znaczący sposób wpływa na efektywność raportów oraz ich poprawną interpretację. Na przedstawionym diagramie modelu danych zobrazowano powiązania między tabelami faktów i tabelami wymiarów, które stanowią fundament całego projektu (Rys. 7).


**Rysunek 7.** Schemat modelu danych w projekcie (źródło: opracowanie własne)

Schemat opiera się na relacjach, które łączą tabele faktów, takie jak „Orders", „Payments" i „Reviews", z tabelami wymiarów, m.in. „Product", „Customer" czy „Geolocation". Podstawowym aspektem tej struktury jest wykorzystanie kluczy głównych i obcych do budowania relacji, co zapewnia integralność danych i umożliwia efektywne filtrowanie oraz analizę.

Przykładowo, tabela faktów „Orders” w modelu projektu jest połączona z tabelą wymiarów „Product” za pomocą kolumny „Product Id”, co pozwala na analizę zamówień w kontekście szczegółowych informacji o produktach. Analogicznie, tabela „Orders” jest połączona z tabelą „Customer” za pomocą kolumny „Customer Unique Id”, co umożliwia analizowanie zakupów w odniesieniu do danych o klientach.

W Power BI relacje między tabelami mogą być jednostronne lub dwustronne, co określa kierunek propagacji filtrów w modelu danych. Relacje jednostronne są najbardziej powszechne i efektywne pod względem wydajności. W przypadku relacji jednostronnych filtr propaguje się tylko w jednym kierunku – od tabeli wymiarów do tabeli faktów. Taki układ jest zgodny z klasycznym schematem gwiazdy i zapewnia spójność analizy, eliminując ryzyko nieoczekiwanych wyników. Z kolei relacje dwustronne umożliwiają propagację filtrów w obu kierunkach, co może być konieczne w bardziej złożonych modelach, gdzie wymagane jest filtrowanie danych zarówno od tabeli faktów do wymiarów, jak i w odwrotnym kierunku. Należy jednak stosować relacje dwustronne ostrożnie, ponieważ mogą one prowadzić do problemów z wydajnością modelu oraz do potencjalnych konfliktów w propagacji filtrów. W dobrze zaprojektowanym modelu opartym na schemacie gwiazdy relacje jednostronne są preferowanym rozwiązaniem, a relacje dwustronne są używane wyłącznie w sytuacjach, gdzie ich zastosowanie jest niezbędne.

Dodatkowo celowe ułożenie tabel w modelu odzwierciedla propagację kontekstu filtru. W zaprojektowanym schemacie tabele wymiarów zostały umieszczone w górnej części modelu, podczas gdy tabele faktów znajdują się w dolnej części. Takie rozplanowanie wizualne modelu nie tylko podkreśla hierarchię relacji, ale także odzwierciedla kierunek propagacji filtrów – od tabel wymiarów u góry do tabel faktów na dole. Taki układ sprzyja lepszemu zrozumieniu modelu zarówno przez analityków, jak i użytkowników końcowych, wspierając bardziej przejrzyste budowanie raportów i zapytań. Dzięki temu model jest nie tylko funkcjonalny, ale również intuicyjny w obsłudze, co znacząco ułatwia pracę z narzędziem Power BI.

Końcowym etapem budowy modelu danych było testowanie jego funkcjonalności. W ramach projektu przeprowadzono szeroko zakrojone testy, obejmujące między innymi:

- Weryfikację, czy dane w tabelach faktów są poprawnie filtrowane w oparciu o tabele wymiarów.
- Sprawdzenie, czy obliczenia DAX, takie jak sumy i średnie, dają spójne wyniki w kontekście filtrów.

- Upewnienie się, że relacje dwustronne są stosowane wyłącznie tam, gdzie są absolutnie konieczne.

Testowanie pozwoliło na wykrycie i eliminację potencjalnych błędów, a także na upewnienie się, że model danych spełnia zarówno wymagania projektowe, jak i potrzeby analityczne.

Tworzenie modelu danych wymaga uwzględnienia wielu aspektów, takich jak definiowanie relacji, zarządzanie kontekstem filtru, właściwe oznaczanie kluczy głównych oraz testowanie działania modelu. Model oparty na schemacie gwiazdy, zaprojektowany na potrzeby projektu, stanowi przykład optymalnej struktury dla analizy danych w Power BI. Dzięki temu modelowi możliwe było stworzenie wydajnego, skalowalnego i intuicyjnego raportu, który spełnia wszystkie założenia biznesowe[13].

### 3.4.2. Tworzenie tabeli kalendarzowej

W procesie optymalizacji modelu danych w ramach realizowanego projektu zdecydowano o wyłączeniu funkcji „Auto Date/Time", która jest domyślnie aktywna w Power BI. Choć funkcja ta może być użyteczna w prostych analizach, niesie za sobą istotne ograniczenia, szczegółowo opisane w książce „Kompletny przewodnik po DAX. Analiza biznesowa przy użyciu Microsoft Power BI, SQL Server Analysis Services i Excel" autorstwa Marca Russo i Alberto Ferrari. Autorzy podkreślają, że „Auto Date/Time" automatycznie generuje odrębne, ukryte tabele dat dla każdej kolumny typu „Date" lub „DateTime" w modelu. Tabele te nie są edytowalne, co uniemożliwia ich dostosowanie do specyficznych potrzeb analitycznych, takich jak dodanie tygodni fiskalnych czy przedziałów czasowych. Ponadto, nadmiarowość tych tabel zwiększa obciążenie modelu i negatywnie wpływa na jego wydajność w dużych projektach. Z tych powodów wyłączenie tej funkcji było świadomą decyzją, zgodną z najlepszymi praktykami modelowania danych w Power BI.

Tabela kalendarzowa pełni fundamentalną rolę w analizach czasowych i jest niezbędnym elementem w modelu danych. Jak wskazują Marco Russo i Alberto Ferrari, powinna ona zawierać pełny zakres dni w analizowanym okresie, niezależnie od tego, czy dane źródłowe obejmują wszystkie daty. Takie podejście zapewnia poprawność grupowania danych według lat, kwartałów, miesięcy czy tygodni. Na przykład zakres tabeli kalendarzowej

[13] Knight D., Pearson M., Schacht B., Ostrowsky E.: *Microsoft Power BI: Jak modelować i wizualizować dane oraz budować narracje cyfrowe*, Helion, Gliwice 2022, s. 62-82

powinien być dostosowany do minimalnej i maksymalnej daty występującej w analizowanych danych, aby zapewnić kompletną analizę czasową. Ponadto tabela kalendarzowa powinna zawierać unikalną kolumnę typu „Date" oraz kolumny opisujące dodatkowe atrybuty, takie jak dni tygodnia, miesiące czy lata fiskalne. Oznaczenie tabeli jako tabeli dat w Power BI umożliwia korzystanie z funkcji analizy szeregów czasowych (ang. Time Intelligence) w języku DAX, co znacząco upraszcza tworzenie zapytań i poprawia ich efektywność[14].

W Projekcie zdecydowano się na utworzenie tabeli kalendarzowej za pomocą Power Query, zgodnie z zasadą, że przekształcenia danych powinny być wykonywane jak najbliżej ich źródła. Takie podejście pozwala odciążyć model danych i zwiększyć jego wydajność. Choć język DAX oferuje łatwe metody tworzenia tabel kalendarzowych, w dużych projektach może powodować spowolnienia operacyjne. Tworzenie tabeli w Power Query zapewnia większą wydajność, a w przypadku kalendarzy pochodzących z baz danych może być optymalnym rozwiązaniem.

| Date | Date Key | Day of Mont | Week Day Number | Week Day | Year Number | Year | Quarter Number | Quarter | Quarter Year | Month Number | Month |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 01.01.2016 | 20160101 | 1 | 5 | Fri | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 02.01.2016 | 20160102 | 2 | 6 | Sat | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 03.01.2016 | 20160103 | 3 | 7 | Sun | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 04.01.2016 | 20160104 | 4 | 1 | Mon | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 05.01.2016 | 20160105 | 5 | 2 | Tue | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 06.01.2016 | 20160106 | 6 | 3 | Wed | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 07.01.2016 | 20160107 | 7 | 4 | Thu | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 08.01.2016 | 20160108 | 8 | 5 | Fri | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 09.01.2016 | 20160109 | 9 | 6 | Sat | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 10.01.2016 | 20160110 | 10 | 7 | Sun | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 11.01.2016 | 20160111 | 11 | 1 | Mon | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 12.01.2016 | 20160112 | 12 | 2 | Tue | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 13.01.2016 | 20160113 | 13 | 3 | Wed | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 14.01.2016 | 20160114 | 14 | 4 | Thu | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 15.01.2016 | 20160115 | 15 | 5 | Fri | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 16.01.2016 | 20160116 | 16 | 6 | Sat | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 17.01.2016 | 20160117 | 17 | 7 | Sun | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 18.01.2016 | 20160118 | 18 | 1 | Mon | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 19.01.2016 | 20160119 | 19 | 2 | Tue | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 20.01.2016 | 20160120 | 20 | 3 | Wed | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 21.01.2016 | 20160121 | 21 | 4 | Thu | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 22.01.2016 | 20160122 | 22 | 5 | Fri | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 23.01.2016 | 20160123 | 23 | 6 | Sat | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |
| 24.01.2016 | 20160124 | 24 | 7 | Sun | 2016 | 2016 | 1 | Q1 | Q1 2016 | 1 | Jan |

**Tabela 5.** Fragment tabeli kalendarzowej wykorzystanej w projekcie (źródło: opracowanie własne)

Tabela kalendarzowa została opracowana zgodnie z poniższymi etapami:

1. **Określenie zakresu dat**: Wyznaczono zakres dat obejmujący minimalną i maksymalną wartość dat w analizowanych danych, uwzględniając wszystkie dni w tym przedziale. Zakres ten pozwolił na stworzenie pełnej tabeli kalendarzowej, obejmującej cały okres analizy.

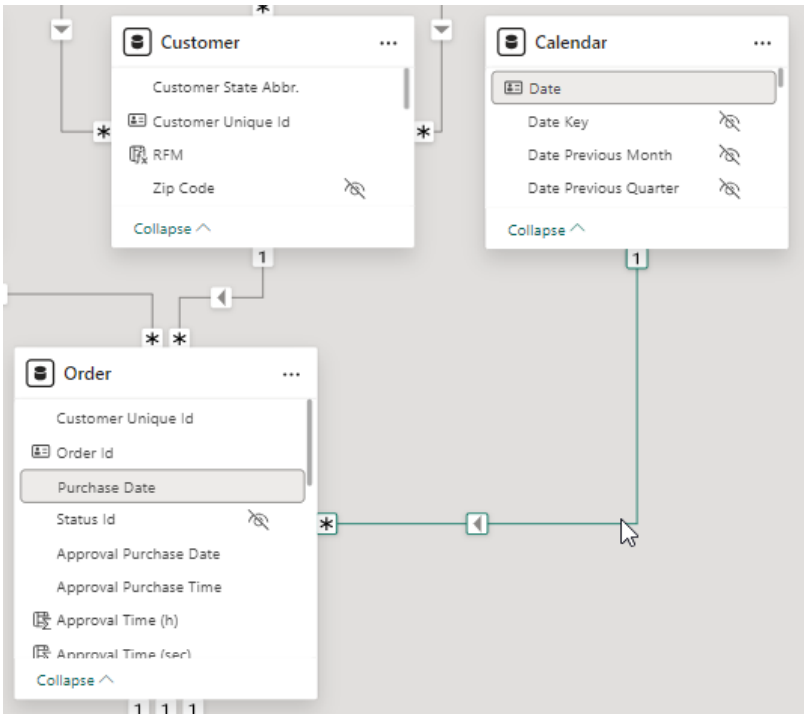
[14] Russo M., Ferrari A.: *Kompletny przewodnik po DAX: Analiza biznesowa przy użyciu Microsoft Power BI, SQL Server Analysis Services i Excel*, APN Promise, Warszawa 2019, s. 241-247

2. **Generowanie ciągłości dat**: Wygenerowano listę wszystkich dni w określonym zakresie, zapewniając ciągłość danych czasowych. Ten etap był kluczowy dla dokładnego odwzorowania analizowanego przedziału czasowego..

3. **Dodanie atrybutów czasowych**: W tabeli dodano kolumny zawierające różnorodne atrybuty czasowe, takie jak numery dni tygodnia, miesięcy, kwartałów czy lat. Atrybuty te wzbogaciły tabelę, umożliwiając realizację zaawansowanych analiz czasowych.

4. **Ostateczna struktura tabeli**: Tabela została uzupełniona o kolumny zawierające podstawowe informacje czasowe, takie jak „Date", „Year", „Month", „Quarter", a także dodatkowe atrybuty, które pozwoliły na bardziej szczegółowe analizy danych w kontekście czasu..

Końcowym etapem było załadowanie tabeli kalendarzowej do modelu danych oraz ustanowienie relacji z odpowiednimi tabelami (Rys. 8). W analizowanym projekcie tabela kalendarzowa została połączona z kolumną „Purchase Date" w tabeli „Order" poprzez relację typu jeden-do-wielu. Dzięki wykorzystaniu tabeli kalendarzowej w tej konfiguracji możliwe było pełne zastosowanie funkcji Time Intelligence w języku DAX, co znacząco rozszerzyło zakres i jakość analiz w projekcie. W rezultacie model danych stał się bardziej wydajny, elastyczny i dostosowany do wymagań biznesowych.



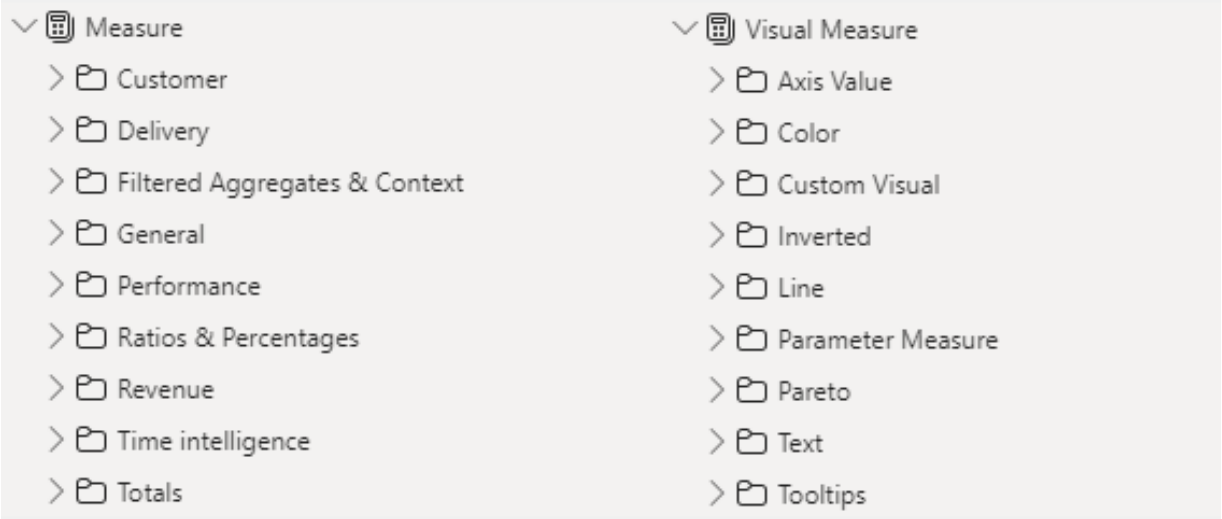

**Rysunek 8.** Schemat relacji tabeli kalendarzowej z tabelą „Order" w modelu danych (źródło: opracowanie własne)

# Rozdział IV. Tworzenie miar do wizualizacji

## 4.1. Wprowadzenie

Miary pełnią istotną funkcję w analizie danych realizowanej w Power BI, będąc podstawą do wyciągania wartościowych wniosków oraz przedstawiania wyników w raportach i wizualizacjach. Są to formuły obliczeniowe, tworzone przy użyciu języka DAX (ang. Data Analysis Expressions), które umożliwiają dynamiczne kalkulacje wartości w zależności od kontekstu zapytania oraz interakcji użytkownika. Dzięki temu miary dostarczają elastyczne i precyzyjne wyniki, odpowiadając na zmieniające się potrzeby analityczne.

Proces tworzenia miar w Power BI wymaga uwzględnienia specyfiki danych, założeń analitycznych oraz struktury modelu danych. Miary znajdują zastosowanie w obliczeniach takich jak sumy, średnie, wskaźniki procentowe czy zmiany w czasie. Kluczowym aspektem ich projektowania jest dokładne uwzględnienie relacji między tabelami w modelu, ponieważ mają one bezpośredni wpływ na poprawność i precyzję obliczeń.




**Rysunek 9.** Struktura tabel miar z folderami grupującymi (źródło: opracowanie własne)

Jedną z zalecanych praktyk w analityce danych, zgodnie z literaturą i doświadczeniami ekspertów, jest tworzenie specjalnej tabeli przeznaczonej wyłącznie na miary, zwanej często „Measure Table". Tabela ta nie zawiera danych, a jedynie miary, co poprawia organizację modelu i zwiększa jego przejrzystość. W celu jeszcze lepszej strukturyzacji miary można grupować w foldery tematyczne. Na przykład miary związane z klientami można umieścić w folderze „Customer", a te dotyczące czasu dostawy w folderze „Delivery" (Rys. 9). Takie

podejście szczególnie dobrze sprawdza się w dużych projektach, gdzie liczba miar jest znaczna i wymaga logicznego uporządkowania.

Ważnym elementem tworzenia miar jest ich testowanie i walidacja. Pierwszym krokiem powinno być sprawdzenie wyników miar w podstawowych tabelach modelu danych, co pozwala zweryfikować poprawność obliczeń. Następnie miary należy przetestować w kontekście wizualizacji, takich jak tabele przestawne czy wykresy, aby upewnić się, że działają poprawnie w połączeniu z relacjami i filtrowaniem w modelu. Proces ten pozwala zidentyfikować ewentualne błędy lub niespójności jeszcze przed etapem tworzenia ostatecznych raportów.

Tworzenie i testowanie miar to proces iteracyjny, który umożliwia analizę danych z różnych perspektyw oraz optymalizację działania modelu. Jest to niezmiernie ważny etap przygotowania modelu danych, ponieważ jakość miar bezpośrednio wpływa na wiarygodność wyników analitycznych i ich prezentację użytkownikom końcowym. Odpowiednio zaprojektowane miary stanowią fundament dalszych działań związanych z tworzeniem raportów i wizualizacji w Power BI[15].

## 4.2. Analiza szeregów czasowych

Cathy Tanimura w swojej książce „Analiza danych z wykorzystaniem SQL-a” zwraca uwagę na istotną rolę analizy szeregów czasowych jako jednego z najczęściej stosowanych podejść analitycznych. Metoda ta pozwala na badanie danych w kontekście upływu czasu, co umożliwia lepsze zrozumienie dynamiki zmian w danym zjawisku. Szeregi czasowe znajdują zastosowanie w różnorodnych sektorach, takich jak finanse, analiza rynkowa czy prognozowanie w biznesie. Jak podkreśla autorka, dane tego rodzaju mają fundamentalne znaczenie w identyfikacji trendów, analizie sezonowości oraz przewidywaniu przyszłych zachowań na podstawie danych historycznych.

Autorka zaznacza, że analiza szeregów czasowych nie tylko pozwala na zrozumienie zmian zachodzących w czasie, ale także umożliwia identyfikację kluczowych okresów wzmożonej aktywności, trendów rynkowych czy momentów strategicznych decyzji. Prognozowanie na podstawie takich danych może dostarczyć organizacjom wartościowych informacji, zwłaszcza w kontekście dynamicznie zmieniających się warunków rynkowych. Tanimura podkreśla jednak, że analiza wyłącznie historycznych danych może być ograniczona,

[15] https://www.sqlbi.com/articles/calculated-columns-and-measures-in-dax/, [dostęp 31.12.2024]

dlatego zaleca uwzględnienie zewnętrznych czynników, które mogły wpłynąć na dane z przeszłości[16].

W analizowanym projekcie przeprowadzono szczegółową analizę szeregów czasowych, której celem była ocena dynamiki liczby unikalnych klientów w ujęciu miesięcznym oraz porównanie tych wartości w perspektywie czasowej. Analiza opierała się na zestawie miar zdefiniowanych w języku DAX, które umożliwiły precyzyjne obliczenia i wizualizację wyników.

Miara „Total Customers" oblicza liczbę unikalnych klientów w określonym przedziale czasowym, wykorzystując funkcję CALCULATE do zliczania unikalnych wartości w kolumnie „Customer[Customer Unique Id]", uwzględniając filtry nałożone przez tabelę „Orders" (Listing 5). Wyniki tej miary stanowiły podstawę do dalszych obliczeń.

```
Total Customers =
CALCULATE (
    DISTINCTCOUNT ( Customer[Customer Unique Id] ),
    KEEPFILTERS ( Orders )
)
```

**Listing 5.** Miara „Total Customers" - całkowita liczba unikalnych klientów

Miara „PM Total Customers" oblicza liczbę unikalnych klientów w poprzednim miesiącu, korzystając z funkcji DATEADD, która przesuwa datę o jeden miesiąc wstecz w tabeli kalendarzowej. Dodatkowo zastosowano warunek „_ShowValueForDates", aby wyniki były prezentowane wyłącznie dla wybranych dat (Listing 6).

```
PM Total Customers =
IF (
    [_ShowValueForDates],
    CALCULATE (
        [Total Customers],
        CALCULATETABLE (
            DATEADD ( 'Calendar2'[Date], -1, MONTH ),
            'Calendar2'[DateWithTransactions] = TRUE
        )
    )
)
```

**Listing 6.** Miara „PM Total Customers" - liczba klientów w poprzednim miesiącu

[16] Tanimura C.: *Analiza danych z wykorzystaniem SQL-a*, s. 66-67

Miara „MOM Total Customers” oblicza różnicę w liczbie klientów między bieżącym a poprzednim miesiącem, wykorzystując zmienne „__ValueCurrentPeriod” oraz „__ValuePreviousPeriod” (Listing 7). Miara ta pozwala na precyzyjne określenie dynamiki zmian.

```
MOM Total Customers =
VAR __ValueCurrentPeriod = [Total Customers]
VAR __ValuePreviousPeriod = [PM Total Customers]
VAR __Result =
    IF (
        NOT ISBLANK ( __ValueCurrentPeriod ) && NOT ISBLANK ( __ValuePreviousPeriod ),
        __ValueCurrentPeriod - __ValuePreviousPeriod
    )
RETURN
    __Result
```

**Listing 7.** Miara „MOM Total Customers” - różnica w liczbie klientów między miesiącami

Miara „MOM % Total Customers” kalkuluje procentową zmianę liczby klientów w stosunku do poprzedniego miesiąca, dzieląc wynik miary „MOM Total Customers” przez „PM Total Customers”. Użycie funkcji DIVIDE eliminuje ryzyko błędów wynikających z dzielenia przez zero (Listing 8).

```
MOM % Total Customers =
DIVIDE ( [MOM Total Customers], [PM Total Customers] )
```

**Listing 8.** Miara „MOM % Total Customers” - procentowa zmiana liczby klientów w stosunku do poprzedniego miesiąca

Rezultaty analizy zostały zaprezentowane w formie tabelarycznej, obejmując dane miesięczne z okresu od lutego 2017 do sierpnia 2018 roku (Tab. 6). Wyniki te posłużyły jako podstawa do tworzenia dynamicznych wizualizacji, takich jak wykresy liniowe i kolumnowe, które przedstawiają zarówno wartości bezwzględne liczby klientów, jak i procentowe zmiany w czasie. Tego rodzaju prezentacja danych w interaktywnych raportach wspiera interesariuszy w szybkim identyfikowaniu trendów, sezonowości oraz anomalii w danych, co może znacząco wspomóc podejmowanie decyzji biznesowych i strategicznych.

| Month Year | Total Customers | PM Total Customers | MOM Total Customers | MOM % Total Customers |
|---|---|---|---|---|
| Feb 2017 | 1 708 | 755 | 953 | 126,23% |
| Mar 2017 | 2 601 | 1 708 | 893 | 52,28% |
| Apr 2017 | 2 359 | 2 601 | -242 | -9,30% |
| May 2017 | 3 588 | 2 359 | 1 229 | 52,10% |
| Jun 2017 | 3 154 | 3 588 | -434 | -12,10% |
| Jul 2017 | 3 894 | 3 154 | 740 | 23,46% |
| Aug 2017 | 4 211 | 3 894 | 317 | 8,14% |
| Sep 2017 | 4 170 | 4 211 | -41 | -0,97% |
| Oct 2017 | 4 501 | 4 170 | 331 | 7,94% |
| Nov 2017 | 7 342 | 4 501 | 2 841 | 63,12% |
| Dec 2017 | 5 557 | 7 342 | -1 785 | -24,31% |
| Jan 2018 | 7 120 | 5 557 | 1 563 | 28,13% |
| Feb 2018 | 6 537 | 7 120 | -583 | -8,19% |
| Mar 2018 | 7 096 | 6 537 | 559 | 8,55% |
| Apr 2018 | 6 878 | 7 096 | -218 | -3,07% |
| May 2018 | 6 795 | 6 878 | -83 | -1,21% |
| Jun 2018 | 6 121 | 6 795 | -674 | -9,92% |
| Jul 2018 | 6 211 | 6 121 | 90 | 1,47% |
| Aug 2018 | 6 411 | 6 211 | 200 | 3,22% |

**Tabela 6.** Wyniki analizy szeregów czasowych (źródło: opracowanie własne)

## 4.3. Analiza kohortowa

Analiza kohortowa to technika analityczna umożliwiająca badanie zachowań grup jednostek, zwanych kohortami, w określonym przedziale czasowym. Cathy Tanimura w swojej książce „Analiza danych z wykorzystaniem SQL-a" definiuje kohortę jako grupę jednostek, które na początku obserwacji wykazują wspólne cechy, takie jak moment rozpoczęcia korzystania z produktu, wykonanie zakupu czy podjęcie określonych działań. Autorka podkreśla, że kohorty mogą obejmować nie tylko ludzi, ale również inne jednostki analityczne, takie jak firmy, produkty czy określone zdarzenia.

Metoda ta jest szczególnie cenna w analizie długoterminowych trendów i zachowań użytkowników. Dzięki analizie kohortowej możliwe jest zidentyfikowanie różnic między grupami klientów lub produktów oraz zrozumienie, jak zmieniają się ich zachowania w czasie. Takie podejście wspiera ocenę skuteczności działań marketingowych, identyfikację potencjalnych problemów oraz odkrywanie wzorców kluczowych dla strategii rozwoju organizacji.

Tanimura wskazuje trzy główne elementy składające się na analizę kohortową: grupowanie kohort, szereg czasowy oraz agregowanie wskaźników opisujących zachowania członków kohort. Grupowanie kohort odbywa się na podstawie wspólnego momentu początkowego, takiego jak data pierwszego zakupu lub rejestracji, co umożliwia szczegółowe

porównanie ich dynamiki rozwoju. Szeregi czasowe pozwalają na analizę aktywności lub interakcji w określonych przedziałach czasowych, takich jak tygodnie czy miesiące, co zapewnia pełniejszy obraz zmian w zachowaniach kohort. Agregowane wskaźniki, takie jak retencja, powtarzalność działań lub przeżywalność w obrębie kohort, dostarczają szczegółowych informacji na temat ich długoterminowych zachowań i efektywności działań podejmowanych w ich kontekście[17].

W ramach omawianego projektu przeprowadzono analizę kohortową, opierając się na zestawie miar umożliwiających szczegółowe badanie zachowań klientów w różnych okresach czasu. Wśród tych miar wyróżniono:

- **Active Customers**: Określa liczbę unikalnych klientów, którzy złożyli zamówienie w danym okresie. Dzięki temu możliwe jest bieżące monitorowanie aktywności klientów w poszczególnych miesiącach.
- **New Customers**: Określa liczbę nowych klientów, którzy zrealizowali swoje pierwsze zamówienie w analizowanym miesiącu. Ta miara pozwala na ocenę skuteczności działań marketingowych skierowanych na pozyskiwanie nowych użytkowników.
- **Lost Customers**: Określa liczbę klientów aktywnych w poprzednim miesiącu, którzy nie złożyli zamówienia w bieżącym okresie. Analiza tej grupy dostarcza informacji na temat potencjalnych problemów z utrzymaniem klientów.
- **Churned Customers**: Określa liczbę klientów, którzy przestali być aktywni w dłuższym horyzoncie czasowym, biorąc pod uwagę brak transakcji w określonym oknie czasowym. Miara ta uwzględnia dodatkowe kryteria, takie jak wskaźnik retencji.
- **Returning Customers**: Określa liczbę klientów, którzy zrealizowali zakupy zarówno w poprzednim, jak i bieżącym miesiącu, wskazując na stabilność bazy lojalnych klientów.
- **Recovered Customers**: Określa liczbę klientów, którzy po okresie braku aktywności ponownie złożyli zamówienie. Miara ta jest szczególnie przydatna w ocenie skuteczności kampanii mających na celu re-aktywację klientów.

[17] Tanimura C.: *Analiza danych z wykorzystaniem SQL-a*, s. 117-120

- **Cohort Performance**: Monitoruje aktywność klientów w wybranym przedziale czasowym, grupując ich według daty pierwszej transakcji. Dzięki tej mierze możliwe jest szczegółowe badanie trendów aktywności kohort w różnych okresach.

Zastosowanie tych miar pozwala na wielowymiarowe badanie zachowań klientów, umożliwiając zarówno ocenę lojalności, jak i identyfikację problemów związanych z retencją oraz skuteczność działań marketingowych.

```
Recovered Customers =
VAR __CustomersThisMonth =
    VALUES ( 'Order'[Customer Unique Id] )
VAR __CustomersLastMonth =
    CALCULATETABLE (
        VALUES ( 'Order'[Customer Unique Id] ),
        PREVIOUSMONTH ( Calendar[Start of Month] )
    )
VAR __NewCustomers =
    CALCULATETABLE (
        VALUES ( 'Order'[Customer Unique Id] ),
        'Order'[Months Since First Transaction] = 0
    )
VAR __ResurrectedCustomers =
    EXCEPT (
        -- remove last month's customers
        EXCEPT (
            __CustomersThisMonth,
            __CustomersLastMonth
        ),
        -- remove new customers
        __NewCustomers
    )
VAR __Result =
    COUNTROWS ( __ResurrectedCustomers ) + 0
RETURN
    __Result
```

**Listing 9.** Miara „Recovered Customers” - klienci odzyskani

Miara „Recovered Customers” została zaprojektowana w celu identyfikacji klientów, którzy powrócili do aktywności po okresie braku zamówień. Kod opiera się na wykorzystaniu funkcji takich jak VALUES, CALCULATETABLE, EXCEPT oraz COUNTROWS.

W pierwszym kroku tworzone są dwie kolekcje klientów: obecnych w bieżącym miesiącu oraz tych, którzy byli aktywni w poprzednim miesiącu. Dodatkowo, za pomocą filtru na kolumnie „Months Since First Transaction”, identyfikowani są nowi klienci. Głównym elementem tej miary jest użycie podwójnego operatora EXCEPT, który umożliwia wykluczenie

klientów aktywnych w poprzednich okresach oraz nowych klientów. Ostatecznie pozostają wyłącznie klienci, którzy przeszli od nieaktywności do aktywności, a ich liczba jest zliczana za pomocą COUNTROWS (Listing 9). Ta technika pozwala na precyzyjne wyodrębnienie grupy klientów odzyskanych, co dostarcza cennych informacji w kontekście analizy skuteczności działań marketingowych.

```
Cohort Performance =
VAR __MinDate =
    MIN ( Calendar[Start of Month] )
VAR __MaxDate =
    MAX ( Calendar[Start of Month] )
VAR __Performance =
    CALCULATE (
        [Active Customers],
        REMOVEFILTERS ( Calendar[Start of Month] ),
        REMOVEFILTERS ( Calendar[Month Year] ),
        RELATEDTABLE ( Customer ),
        Customer[First Transaction Date] >= __MinDate
            && Customer[First Transaction Date] <= __MaxDate
    )
VAR __MonthsAfter =
    CALCULATE (
        SELECTEDVALUE ( 'Order'[Months Since First Transaction] ),
        REMOVEFILTERS ( Calendar[Start of Month] ),
        REMOVEFILTERS ( Calendar[Month Year] )
    ) + 1
VAR __MonthsAfterLimit =
    DATEDIFF ( __MinDate, [End Date], MONTH )
VAR __Result =
    IF ( __MonthsAfterLimit > __MonthsAfter, __Performance + 0, BLANK () )
RETURN
    __Result
```

**Listing 10.** Miara „Cohort Performance” – aktywność kohort w czasie

Miara „Cohort Performance” została stworzona w celu analizy aktywności klientów w ramach określonych kohort czasowych. Kluczowym elementem tego kodu jest zastosowanie zmiennych, które upraszczają i porządkują obliczenia. Miara korzysta z funkcji MIN i MAX, aby określić minimalną i maksymalną datę dla danej kohorty. Następnie za pomocą funkcji CALCULATE i REMOVEFILTERS obliczana jest liczba aktywnych klientów w zadanym zakresie dat, uwzględniając warunek, że data pierwszej transakcji klienta musi zawierać się w przedziale wyznaczonym przez kohortę. Kolejnym krokiem jest ustalenie liczby miesięcy od pierwszej transakcji, co realizowane jest za pomocą funkcji DATEDIFF. Ostatecznie miara zwraca wynik aktywności kohorty, filtrując dane dla określonych ram czasowych.

Zastosowanie technik takich jak CALCULATE, REMOVEFILTERS i RELATEDTABLE pozwala na elastyczne operowanie danymi, umożliwiając precyzyjną analizę trendów zachowań klientów w czasie (Listing 10).

Tabela wynikowa kohort przedstawia zebrane dane, które pozwalają na analizę zmian aktywności klientów w poszczególnych okresach czasu (Tab. 7). Dzięki niej możliwe jest nie tylko zrozumienie dynamiki zachowań kohort, ale także identyfikacja obszarów wymagających dodatkowych działań optymalizacyjnych.

| Month Year | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Jan 2017 | 754 | 3 | 2 | 1 | 3 | 1 | 3 | 1 | 1 | 0 | 3 | 1 | 5 | 3 | 1 | 1 | 2 | 3 | 1 |
| Feb 2017 | 1 705 | 4 | 5 | 2 | 7 | 2 | 4 | 3 | 3 | 4 | 2 | 5 | 3 | 3 | 2 | 1 | 1 | 4 | |
| Mar 2017 | 2 595 | 13 | 9 | 10 | 9 | 4 | 4 | 8 | 9 | 2 | 10 | 3 | 6 | 3 | 4 | 6 | 2 | 4 | |
| Apr 2017 | 2 339 | 14 | 5 | 4 | 8 | 6 | 8 | 7 | 7 | 4 | 6 | 2 | 2 | 1 | 2 | 2 | 5 | | |
| May 2017 | 3 560 | 18 | 18 | 14 | 11 | 12 | 15 | 6 | 9 | 10 | 9 | 12 | 9 | 1 | 7 | 9 | | | |
| Jun 2017 | 3 113 | 15 | 11 | 12 | 8 | 12 | 12 | 7 | 4 | 7 | 10 | 11 | 5 | 4 | 6 | | | | |
| Jul 2017 | 3 842 | 20 | 14 | 10 | 11 | 8 | 12 | 4 | 7 | 10 | 9 | 11 | 5 | 10 | | | | | |
| Aug 2017 | 4 149 | 29 | 14 | 11 | 15 | 22 | 12 | 11 | 6 | 6 | 10 | 8 | 4 | | | | | | |
| Sep 2017 | 4 089 | 28 | 22 | 12 | 18 | 9 | 9 | 10 | 12 | 7 | 11 | 3 | | | | | | | |
| Oct 2017 | 4 410 | 31 | 11 | 4 | 10 | 9 | 10 | 16 | 12 | 9 | 9 | | | | | | | | |
| Nov 2017 | 7 216 | 40 | 28 | 13 | 14 | 13 | 8 | 14 | 10 | 4 | | | | | | | | | |
| Dec 2017 | 5 441 | 14 | 15 | 19 | 15 | 11 | 9 | 3 | 12 | | | | | | | | | | |
| Jan 2018 | 6 981 | 24 | 26 | 20 | 20 | 11 | 12 | 16 | | | | | | | | | | | |
| Feb 2018 | 6 421 | 25 | 25 | 19 | 17 | 14 | 13 | | | | | | | | | | | | |
| Mar 2018 | 6 947 | 32 | 22 | 20 | 9 | 8 | | | | | | | | | | | | | |
| Apr 2018 | 6 707 | 39 | 21 | 16 | 9 | | | | | | | | | | | | | | |
| May 2018 | 6 603 | 35 | 18 | 14 | | | | | | | | | | | | | | | |
| Jun 2018 | 5 934 | 25 | 16 | | | | | | | | | | | | | | | | |
| Jul 2018 | 6 052 | 31 | | | | | | | | | | | | | | | | | |
| Aug 2018 | 6 237 | | | | | | | | | | | | | | | | | | |

**Tabela 7.** Wyniki analizy kohortowej klientów (źródło: opracowanie własne)

## 4.4. Segmentacja RFM

Zgodnie z definicją przedstawioną w książce „Analiza danych z wykorzystaniem SQL-a" autorstwa Cathy Tanimura, segmenty i kohorty to pojęcia często stosowane w podobnych kontekstach analitycznych, jednak ich charakterystyka istotnie się różni. Kohorta to grupa użytkowników (lub innych jednostek analitycznych), których łączy wspólna data początkowa – oznacza to, że ich analiza jest związana z konkretnym momentem w czasie, takim jak data pierwszego zakupu lub rejestracji. W odróżnieniu od kohort segment to grupa użytkowników mających wspólną cechę (lub zestaw cech), niezależnie od daty początkowej.

Innymi słowy, segmenty mogą obejmować użytkowników z różnych kohort, ale łączą ich podobne zachowania lub cechy demograficzne[18].

Segmenty użytkowników są zatem bardziej statyczne w swojej strukturze, ponieważ odnoszą się do określonych cech użytkowników w danym momencie, takich jak częstotliwość zakupów czy preferencje produktowe. W kontekście analizy marketingowej segmenty są przydatne do badania bieżących zachowań użytkowników, niezależnie od ich historii, podczas gdy kohorty pozwalają zrozumieć, jak zmieniają się zachowania w czasie, począwszy od określonego punktu początkowego.

Jednym z rodzajów wykorzystania segmentów w analizie jest segmentacja RFM, która pozwala na podział klientów na podstawie ich zachowań zakupowych. Metoda ta opiera się na trzech kluczowych wymiarach:

1. **Recency** (z ang. aktualność) – Mierzy, jak niedawno klient dokonał zakupu, co stanowi wskaźnik jego aktualnego zaangażowania.
2. **Frequency** (z ang. częstotliwość) – Określa, jak często klient dokonuje zakupów, co pozwala na ocenę lojalności klienta i jego nawyków zakupowych.
3. **Monetary** (z ang. wartość monetarna) – Analizuje, jak dużo klient wydał, pomagając określić jego wartość dla organizacji.

Segmentacja RFM umożliwia tworzenie segmentów klientów na podstawie kombinacji tych trzech wymiarów, co pozwala zidentyfikować grupy o różnym poziomie wartości dla firmy, takie jak „najlepsi klienci", „klienci lojalni" czy „klienci o niskiej wartości". Metoda ta znajduje szerokie zastosowanie w marketingu i CRM, ponieważ dostarcza konkretnych informacji o zachowaniach klientów, co z kolei pozwala na skuteczniejsze prowadzenie działań marketingowych i personalizację oferty.

Analiza RFM różni się od klasycznej segmentacji demograficznej tym, że koncentruje się na bieżących interakcjach klientów z marką, a nie na ich stałych cechach, takich jak wiek czy lokalizacja. Dzięki temu możliwe jest bardziej dynamiczne i praktyczne podejście do segmentacji, które wspiera podejmowanie decyzji biznesowych i alokację zasobów na działania skierowane do najbardziej wartościowych grup klientów.

[18] Tanimura C.: *Analiza danych z wykorzystaniem SQL-a*, s. 118

W projekcie segmentacja RFM została przeprowadzona przy użyciu języka DAX, co umożliwiło opracowanie odpowiednich miar dla każdego z trzech elementów modelu RFM. Głównym etapem analizy było przypisanie każdemu klientowi wartości w każdej z kategorii oraz określenie ogólnego wyniku RFM, który pozwala na klasyfikację klientów do odpowiednich segmentów.

Wartość „Recency (R)" określono jako liczbę dni od daty ostatniej transakcji klienta do ustalonej daty końcowej w analizie. Miara ta została zrealizowana za pomocą funkcji DATEDIFF, która precyzyjnie wyznacza odstęp czasowy między ostatnią transakcją a punktem odniesienia. „Frequency (F)" zdefiniowano jako liczbę transakcji dokonanych przez klienta, przy czym do jej obliczenia wykorzystano funkcję DISTINCTCOUNT, liczącą unikalne identyfikatory zamówień dla każdego klienta. Z kolei „Monetary (M)" określała sumaryczną wartość przychodów wygenerowanych przez klienta, bazując na funkcji SUM operującej na kolumnie zawierającej wartości przychodów.

```
R Score =
SWITCH (
    TRUE (),
    Customer[R Value] <= PERCENTILE.INC ( Customer[R Value], 0.20 ), 5,
    Customer[R Value] <= PERCENTILE.INC ( Customer[R Value], 0.40 ), 4,
    Customer[R Value] <= PERCENTILE.INC ( Customer[R Value], 0.60 ), 3,
    Customer[R Value] <= PERCENTILE.INC ( Customer[R Value], 0.80 ), 2,
    1
)
```

**Listing 11.** Miara „R Score" - aktualność klientów

Kolejnym krokiem było przypisanie wyników dla każdej kategorii RFM za pomocą percentyli, co pozwoliło podzielić klientów na pięć grup w każdej kategorii – od najniższej (1) do najwyższej (5). Na przykład dla „Recency (R)" opracowano miarę „R Score", korzystając z funkcji PERCENTILE.INC (Listing 11). Klienci mieszczący się w najniższym 20% pod względem liczby dni od ostatniej transakcji otrzymali wartość 5, co oznaczało ich wysoką aktualność, natomiast ci w najwyższym 20% uzyskali wartość 1.

| Customer Unique Id | R Value | F Value | M Value | R Score | F Score | M Score | RFM |
|---|---|---|---|---|---|---|---|
| 77 | 57 | 1 | R$ 361,20 | 5 | 1 | 5 | 515 |
| 148 | 56 | 1 | R$ 35,65 | 5 | 1 | 1 | 511 |
| 76272 | 53 | 1 | R$ 201,82 | 5 | 1 | 4 | 514 |
| 322 | 72 | 1 | R$ 195,54 | 5 | 1 | 4 | 514 |
| 334 | 78 | 1 | R$ 72,57 | 5 | 1 | 2 | 512 |
| 504 | 52 | 1 | R$ 53,48 | 5 | 1 | 1 | 511 |
| 663 | 70 | 1 | R$ 85,79 | 5 | 1 | 2 | 512 |
| 679 | 73 | 1 | R$ 141,92 | 5 | 1 | 4 | 514 |
| 63751 | 75 | 1 | R$ 33,34 | 5 | 1 | 1 | 511 |
| 808 | 62 | 1 | R$ 394,31 | 5 | 1 | 5 | 515 |
| 1004 | 66 | 1 | R$ 36,94 | 5 | 1 | 1 | 511 |

**Tabela 8.** Wyniki analizy RFM w tabeli „Customer" (źródło: opracowanie własne)

Dla każdego klienta obliczono ogólny wynik RFM, będący kombinacją wartości R, F i M. Wyniki te zapisano w tabeli „Customer", zawierającej zarówno szczegółowe informacje o poszczególnych składowych, jak i pełen wynik RFM przedstawiony w postaci trzycyfrowego kodu (Tab. 8). Tabela ta stała się podstawą dalszej analizy.

| Index | Segment | Score | R | F | M | Activity |
|---|---|---|---|---|---|---|
| 1 | Champions | 555 | 5 | 5 | 5 | Bought recently, buy often and spend the most! |
| 1 | Champions | 554 | 5 | 5 | 4 | Bought recently, buy often and spend the most! |
| 1 | Champions | 544 | 5 | 4 | 4 | Bought recently, buy often and spend the most! |
| 1 | Champions | 545 | 5 | 4 | 5 | Bought recently, buy often and spend the most! |
| 1 | Champions | 454 | 4 | 5 | 4 | Bought recently, buy often and spend the most! |
| 1 | Champions | 455 | 4 | 5 | 5 | Bought recently, buy often and spend the most! |
| 1 | Champions | 445 | 4 | 4 | 5 | Bought recently, buy often and spend the most! |
| 2 | Loyal | 543 | 5 | 4 | 3 | Spend good money with us often. Responsive to promotions. |
| 2 | Loyal | 444 | 4 | 4 | 4 | Spend good money with us often. Responsive to promotions. |
| 2 | Loyal | 435 | 4 | 3 | 5 | Spend good money with us often. Responsive to promotions. |
| 2 | Loyal | 355 | 3 | 5 | 5 | Spend good money with us often. Responsive to promotions. |

**Tabela 9.** Fragment tabeli „Customer Segment" (źródło: opracowanie własne)

Na podstawie przypisanych wartości RFM każdy klient został sklasyfikowany do konkretnego segmentu, takiego jak „Champions" (najbardziej wartościowi klienci) czy „Loyal" (klienci lojalni). Szczegółowe informacje na temat aktywności klientów w poszczególnych segmentach zawarto w tabeli „Customer Segment" (Tab. 9). Takie podejście umożliwiło bardziej spersonalizowaną analizę oraz skierowanie odpowiednich działań marketingowych do różnych grup klientów.

Podsumowując, analiza segmentacji RFM w projekcie stanowi przykład efektywnego wykorzystania Power BI do analizy danych. Wykorzystanie podejścia opartego na percentylach pozwoliło na precyzyjne przypisanie wartości każdemu klientowi. Dzięki temu możliwe było skuteczne zidentyfikowanie kluczowych grup klientów, takich jak najbardziej wartościowi nabywcy czy klienci wymagający dodatkowych działań reaktywacyjnych[19].

[19] Cheverton P.: *Zarządzanie kluczowymi klientami: Jak uzyskać status głównego dostawcy*, Oficyna Ekonomiczna, Kraków 2001, s. 204-211

# Rozdział V. Projektowanie wizualizacji i tworzenie raportu

## 5.1. Zasady projektowania wizualizacji – psychologia percepcji i estetyka

Przed przystąpieniem do projektowania dashboardu niezwykle istotne jest zrozumienie zasad, które pomogą stworzyć raporty przyjazne dla użytkownika, zapewniając efektywne przyswajanie informacji. Podstawowym elementem jest redukcja obciążenia poznawczego, która pozwala odbiorcom łatwiej skupić się na istotnych danych. W tym kontekście warto zwrócić uwagę na reguły gestaltu, które są podstawą psychologii percepcji i wspierają organizację wizualną informacji. Równie istotne jest wykorzystanie atrybutów przetwarzanych mimowolnie, które wpływają na intuicyjne rozpoznawanie kluczowych elementów wizualizacji. Ponadto, uwzględnienie zasad afordancji, przystępności i estetyki w projektowaniu dashboardu nie tylko poprawia funkcjonalność raportów, ale także sprawia, że stają się one bardziej atrakcyjne i użyteczne dla odbiorców.

### 5.1.1. Obciążenie poznawcze i zasady gestaltu

Obciążenie poznawcze to ilość wysiłku umysłowego wymaganego od odbiorcy, aby przyswoić prezentowane informacje. Jak podkreśla Cole Nussbaumer Knaflic w książce „Storytelling danych", nadmiar informacji wizualnych może skutkować przeciążeniem odbiorcy, który traci zdolność do skupienia uwagi na głównych aspektach komunikatu. Każdy element wizualny, który nie wnosi wartości informacyjnej, stanowi dodatkowy „szum", odciągający uwagę od istotnych treści.

W kontekście wizualizacji danych istotne jest, aby eliminować zbędne elementy i maksymalizować stosunek „sygnału do szumu". Oznacza to, że wszystkie elementy wizualne powinny służyć przekazywaniu kluczowych informacji, a nie stanowić dekoracji, która utrudnia interpretację danych. Jak zauważa Edward Tufte, cytowany w książce Nussbaumer Knaflic, umiejętne ograniczenie obciążenia poznawczego pozwala w pełni wykorzystać potencjał intelektualny odbiorców i poprawia jakość przekazu[20].

[20] Nussbaumer Knaflic C.: *Storytelling danych: Poradnik wizualizacji danych dla profesjonalistów*, Helion, Gliwice 2019, s. 83-85

Jednym z narzędzi redukcji obciążenia poznawczego są zasady gestaltu, które wywodzą się z psychologii percepcji. Te reguły opisują, w jaki sposób ludzki umysł organizuje bodźce wizualne, tworząc spójne i logiczne całości. Dzięki ich zastosowaniu możliwe jest projektowanie wizualizacji, które są bardziej czytelne i łatwiejsze do interpretacji.

Nussbaumer Knaflic w swojej książce wyróżnia sześć głównych zasad gestaltu:

1. **Bliskość**: Elementy znajdujące się blisko siebie w przestrzeni są postrzegane jako należące do tej samej grupy. Przykładowo, odpowiednie rozmieszczenie danych w tabeli za pomocą odpowiednich odstępów między kolumnami może pomóc użytkownikom w intuicyjnym odczytywaniu informacji.

2. **Podobieństwo**: Obiekty o podobnym kolorze, kształcie, rozmiarze czy teksturze są automatycznie grupowane. Zasada ta pozwala na wyróżnienie ważnych informacji w wykresach i diagramach – np. przez zastosowanie koloru, który zwraca uwagę na wybrane elementy.

3. **Zamknięcie w przestrzeni**: Elementy wizualne zamknięte w wyraźnych granicach, takich jak ramki czy cieniowanie, są postrzegane jako należące do tej samej grupy. W wizualizacjach dane mogą być grupowane na przykład w sekcjach z delikatnym tłem, co ułatwia ich odróżnienie od pozostałych treści.

4. **Domknięcie**: Umysł ludzki ma tendencję do uzupełniania brakujących elementów obrazu, aby stworzyć spójną całość. Ta zasada pozwala na uproszczenie wykresów i usunięcie zbędnych elementów, takich jak nadmiarowe linie siatki, co podkreśla kluczowe dane.

5. **Ciągłość**: Elementy rozmieszczone w sposób tworzący wizualną ścieżkę są postrzegane jako powiązane. Na przykład, w wykresach liniowych ciągłe linie łączące punkty danych kierują wzrok odbiorcy wzdłuż naturalnego przepływu informacji.

6. **Połączenie**: Elementy połączone liniami są silniej kojarzone ze sobą niż te, które mają podobny kształt lub kolor. Połączenia te są szczególnie przydatne w diagramach przepływu czy wykresach sieciowych, gdzie relacje między różnymi częściami danych są niezbędne[21].

[21] Nussbaumer Knaflic C.: *Storytelling danych*, s. 85-91

Zasady gestaltu i świadomość wpływu obciążenia poznawczego stanowią fundament projektowania skutecznych wizualizacji danych. Redukując niepotrzebne elementy i stosując odpowiednie techniki percepcyjne, można stworzyć raporty i dashboardy, które są nie tylko estetyczne, ale przede wszystkim użyteczne dla użytkowników. Dzięki takim praktykom wizualizacje stają się narzędziem ułatwiającym podejmowanie decyzji, a odbiorcy mogą w pełni skupić się na analizie kluczowych informacji.

### 5.1.2. Atrybuty przetwarzane mimowolnie

W procesie projektowania wizualizacji danych istotne jest nie tylko odpowiednie zaprezentowanie informacji, ale również skierowanie uwagi odbiorców na najważniejsze elementy przedstawianych treści. Jednym z narzędzi, które pozwala osiągnąć ten cel, jest wykorzystanie atrybutów przetwarzanych mimowolnie. Atrybuty te odnoszą się do cech wizualnych, które ludzki umysł wychwytuje automatycznie, bez potrzeby świadomego skupienia uwagi. Ich strategiczne zastosowanie może znacznie poprawić klarowność prezentowanych danych, podczas gdy ich niewłaściwe użycie może wprowadzić chaos i utrudnić odbiór informacji.

Atrybuty przetwarzane mimowolnie to elementy wizualne, które przyciągają wzrok odbiorców w sposób automatyczny. W książce „Storytelling danych" autorstwa Cole Nussbaumer Knaflic podkreślono, że wykorzystanie tych atrybutów może skutecznie pokierować uwagą odbiorców i wskazać, na które elementy diagramu lub wykresu powinni zwrócić szczególną uwagę. Kluczowe jest jednak ich odpowiednie zastosowanie w celu uniknięcia nieporozumień lub dezorientacji.

Cole Nussbaumer Knaflic w swojej publikacji omawia podstawowe atrybuty przetwarzane mimowolnie, które można efektywnie wykorzystać w procesie wizualizacji danych:

- **Kolor** jest jednym z najważniejszych narzędzi wizualnych, które naturalnie przyciąga uwagę odbiorców. Zastosowanie wyróżniającego się koloru pozwala podkreślić najistotniejsze elementy w zestawieniu danych. Knaflic zaleca stosowanie szarej palety jako podstawy i użycie jednego wyrazistego koloru do wyróżnienia kluczowych informacji, co zapobiega chaosowi wizualnemu i ułatwia interpretację.

- **Rozmiar** odzwierciedla hierarchię informacji – większe elementy są automatycznie odbierane jako bardziej istotne w porównaniu z mniejszymi. Projektując wizualizację, ważne jest, aby elementy o podobnym znaczeniu miały zbliżony rozmiar, a kluczowe informacje były podkreślone większymi proporcjami.
- **Grubość linii** może wskazywać na znaczenie danych w odniesieniu do innych elementów wykresu. Linie o większej grubości szybciej przyciągają uwagę i mogą być wykorzystane do podkreślenia istotnych trendów lub kontrastów w prezentowanych danych.
- **Orientacja i kształt** elementów odgrywają istotną rolę w przyciąganiu wzroku. Nietypowe kształty lub orientacje mogą skutecznie zwracać uwagę na konkretne części wizualizacji. Należy jednak unikać nadmiernego użycia niestandardowych elementów, aby nie wprowadzać zbędnego zamieszania.
- **Pozycja w przestrzeni** również ma kluczowe znaczenie. Odbiorcy zwykle rozpoczynają analizę wizualizacji od lewego górnego rogu, przesuwając wzrok w kierunku dolnej prawej części. Umieszczanie najważniejszych informacji w strategicznych miejscach, takich jak górna część wykresu, może znacząco wpłynąć na skuteczność komunikacji danych.
- **Nasycenie kolorów** wpływa na odbiór wizualizacji – bardziej nasycone kolory są bardziej widoczne i mogą być wykorzystane do podkreślenia krytycznych danych. Zróżnicowanie nasycenia pozwala wskazać priorytety w prezentowanych informacjach.
- **Długość i krzywizna linii** mogą również pełnić funkcję informacyjną. Długość linii odzwierciedla proporcje między danymi, podczas gdy zakrzywione linie zwracają uwagę ze względu na swój niestandardowy charakter. Linie proste są postrzegane jako bardziej uporządkowane i przejrzyste, natomiast linie zakrzywione mogą być stosowane w celu wyróżnienia trendów[22].

Odpowiednie zastosowanie atrybutów przetwarzanych mimowolnie może znacząco zwiększyć czytelność wizualizacji, kierując uwagę odbiorców na kluczowe elementy oraz ułatwiając interpretację danych. Jednak niewłaściwe korzystanie z tych elementów może prowadzić do szeregu problemów, takich jak przeciążenie wizualne wynikające z nadmiaru

[22] Nussbaumer Knaflic C.: *Storytelling danych*, s. 112-135

kolorów, kształtów czy linii, co wprowadza chaos i utrudnia odbiór najistotniejszych informacji. Brak spójności w stosowaniu atrybutów, takich jak kolor czy rozmiar, może dezorientować odbiorców, którzy nie będą w stanie zidentyfikować najważniejszych elementów wizualizacji. Ponadto niewłaściwe rozmieszczenie elementów na wykresie może skutkować utratą klarowności i trudnościami w interpretacji danych. Dlatego każda decyzja dotycząca ich zastosowania powinna być świadoma, oparta na jasno określonym celu komunikacyjnym i uwzględniająca potrzeby odbiorców.

### 5.1.3. Eliminacja zbędnych elementów i upraszczanie wizualizacji danych

Wizualizowanie danych wymaga nie tylko zdolności analitycznych, lecz także umiejętności efektywnego projektowania graficznego. Cole Nussbaumer Knaflic w książce „Storytelling danych" zwraca uwagę na istotność eliminacji zbędnych elementów oraz upraszczania prezentacji danych w celu zwiększenia ich przystępności dla odbiorców. Autorka przedstawia szereg wskazówek, które mają na celu poprawę przejrzystości komunikacji wizualnej, co prowadzi do bardziej intuicyjnego odbioru przekazywanych informacji.

Pierwszą podstawową zasadą jest identyfikacja i usuwanie elementów, które odciągają uwagę odbiorców od głównego przekazu. W kontekście wizualizacji danych oznacza to, że nie należy obciążać wykresów i raportów niepotrzebnymi szczegółami, które nie wspierają kluczowego przesłania. Wskazówki autorki obejmują:

- **Selekcja istotnych danych**: Nie wszystkie informacje mają jednakową wartość. W praktyce należy skupić się na najważniejszych danych, a mniej istotne szczegóły zepchnąć na dalszy plan lub całkowicie je wyeliminować.
- **Redukcja szczegółów**: Autorka podkreśla, że szczegóły są wartościowe tylko wówczas, gdy wspierają przekaz. Nadmiar informacji może powodować niepotrzebne przeciążenie poznawcze.
- **Porządkowanie informacji**: Elementy, które są niezbędne, lecz mają mniejsze znaczenie dla przekazu, powinny być przedstawione w sposób subtelny, np. za pomocą jaśniejszych kolorów lub mniejszych rozmiarów[23].

[23] Nussbaumer Knaflic C.: *Storytelling danych*, s. 143

Jednym z głównych celów wizualizacji danych jest ich łatwa interpretacja. Wskazówki autorki dotyczące upraszczania przekazu obejmują:

- **Unikanie nadmiernej komplikacji**: Knaflic wskazuje, że skomplikowane diagramy lub nieczytelne czcionki mogą powodować trudności w odbiorze informacji. Eksperymenty wykazały, że osoby czytające tekst zapisany prostym fontem, takim jak Arial, szybciej i efektywniej przyswajają informacje niż osoby, którym przedstawiono treści w bardziej złożonych krojach pisma.
- **Czytelność i przejrzystość**: Wybór prostych czcionek, odpowiednie marginesy oraz zachowanie porządku w układzie wizualnym są kluczowe dla poprawy czytelności.
- **Użycie nieskomplikowanego języka**: Wizualizacje danych powinny być zrozumiałe zarówno dla ekspertów, jak i laików. W związku z tym warto unikać nadmiaru technicznego żargonu, zamiast tego korzystając z prostego języka, który klarownie wyjaśnia przekaz[24].

Knaflic zwraca również uwagę na znaczenie estetyki w wizualizacji danych. Estetyczne projekty są lepiej odbierane przez odbiorców i zwiększają ich otwartość na przedstawiane informacje. Elementy wizualne powinny być uporządkowane w harmonijny sposób, z zachowaniem podstawowych zasad projektowania graficznego:

- **Wyrównanie pionowe i poziome** wprowadza porządek na wykresach i diagramach, poprawiając ich czytelność.
- **Wykorzystanie pustej przestrzeni**, takiej jak marginesy czy odstępy wokół głównych elementów, pozwala podkreślić najważniejsze informacje i uniknąć wizualnego chaosu.
- **Kolor powinien być stosowany oszczędnie i strategicznie**, aby zwracać uwagę na najistotniejsze aspekty wizualizacji oraz ułatwiać interpretację danych[25].

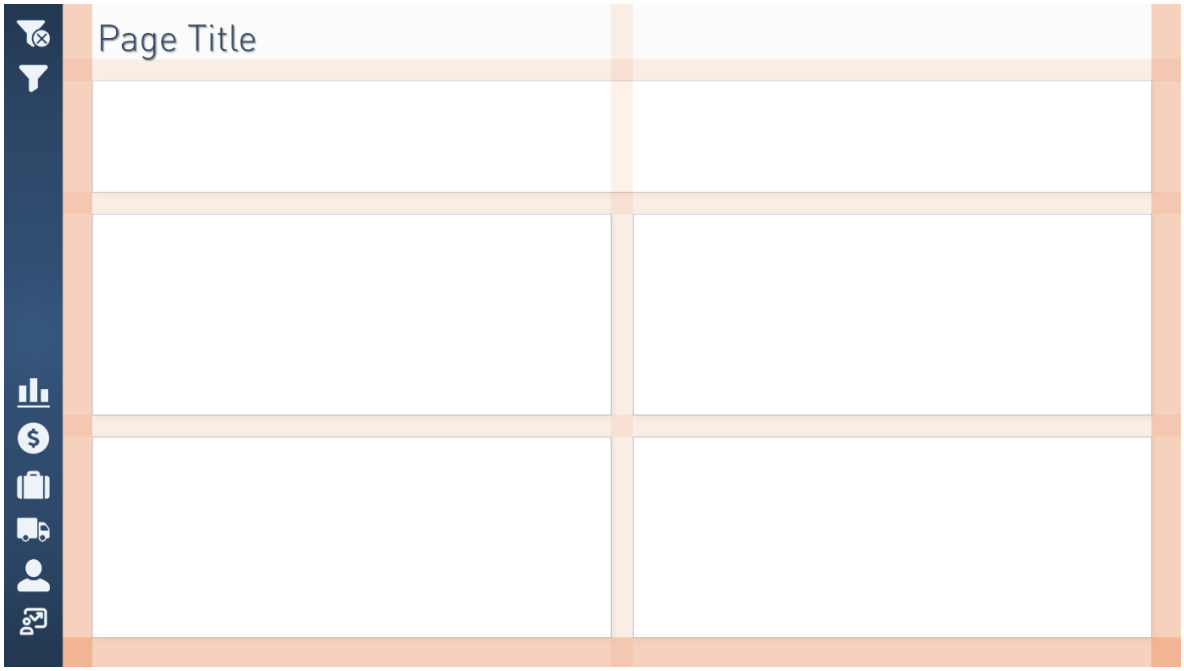
Proponowane przez autorkę podejście do wizualizacji danych opiera się na minimalizmie i przemyślanym projektowaniu. Niezmiernie ważne jest wyeliminowanie elementów odwracających uwagę oraz unikanie niepotrzebnej złożoności. Dzięki temu możliwe jest stworzenie materiałów wizualnych, które nie tylko skutecznie komunikują dane,

[24] Nussbaumer Knaflic C.: *Storytelling danych*, s. 151
[25] Nussbaumer Knaflic C.: *Storytelling danych*, s. 157

lecz także są estetyczne i przyjazne dla odbiorców. Tego rodzaju podejście wspiera efektywne przekazywanie informacji, zwiększając ich zrozumienie i zapamiętywanie.

## 5.2. Projektowanie dashboardu i planowanie stron

Projektowanie dashboardu odgrywa istotną rolę w procesie tworzenia raportów wizualnych, ponieważ bezpośrednio wpływa na ich czytelność oraz funkcjonalność. Kluczowymi elementami tego procesu są odpowiednie zaplanowanie stron, logiczne rozmieszczenie komponentów oraz opracowanie spójnego schematu kolorystycznego. Dzięki takim działaniom możliwe jest efektywne przekazywanie informacji użytkownikom.

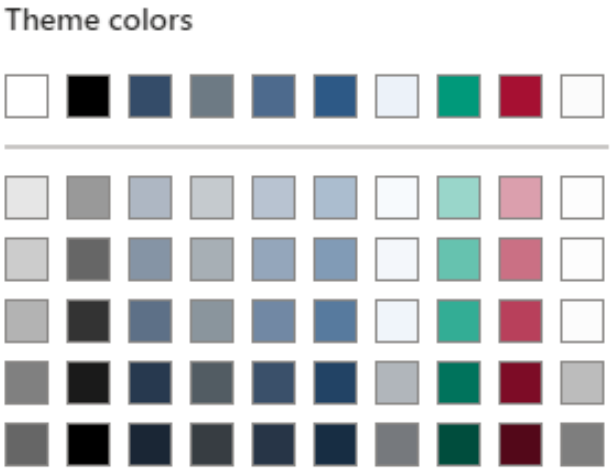



**Rysunek 10.** Szablon strony raportu zaprojektowany w PowerPoint (źródło: opracowanie własne)

Pierwszym krokiem w projektowaniu dashboardu jest precyzyjne określenie zawartości każdej strony raportu. Wymaga to wcześniejszego zdefiniowania miar i wskaźników KPI, które będą prezentowane w raporcie. Na podstawie wstępnych wizualizacji podejmowane są decyzje dotyczące rozmieszczenia najważniejszych wskaźników na poszczególnych stronach, co umożliwia dostosowanie treści do oczekiwań użytkowników. Wykorzystanie dedykowanego szablonu stron (Rys. 10), ułatwia kontrolę nad układem elementów wizualnych. Tego rodzaju szablon obejmuje stałe obszary na wizualizacje, marginesy zapewniające odpowiednią przestrzeń oraz pasek nawigacyjny pozwalający na szybkie przechodzenie między sekcjami

raportu. Dzięki takiemu podejściu użytkownicy z łatwością mogą odnaleźć interesujące ich dane, co pozytywnie wpływa na komfort pracy z raportem[26].

Organizacja wizualna komponentów powinna być zgodna z zasadami przejrzystości i intuicyjności. Elementy wizualne, takie jak wykresy, tabele czy filtry, muszą być rozmieszczone w sposób spójny i logiczny. Warto przy tym uwzględniać zasady projektowania wizualnego, takie jak hierarchia wizualna, która polega na umieszczaniu najważniejszych informacji w najbardziej widocznych miejscach, na przykład w lewym górnym rogu strony. Należy także dbać o proporcje – elementy o podobnym znaczeniu powinny mieć zbliżony rozmiar i być równomiernie rozmieszczone. Ważnym aspektem jest również zastosowanie odpowiednich marginesów i przestrzeni negatywnej, co pomaga w interpretacji treści i nadaje raportowi uporządkowany charakter. Tak zaprojektowany układ nie tylko poprawia estetykę raportu, ale również zwiększa jego funkcjonalność, wspierając użytkowników w procesie podejmowania decyzji.




**Rysunek 11.** Paleta kolorów wykorzystana w projekcie (źródło: opracowanie własne)

Kolory w wizualizacji danych pełnią kluczową funkcję, pomagając w kierowaniu uwagi na istotne elementy oraz ułatwiając zrozumienie informacji. Opracowanie spójnego schematu kolorystycznego powinno uwzględniać minimalizm, co oznacza unikanie nadmiernej liczby kolorów, aby zapobiec chaosowi wizualnemu. Kontrast i czytelność są równie ważne – istotne elementy należy wyróżniać za pomocą wyrazistych kolorów, takich jak czerwony lub niebieski, zachowując jednocześnie stonowane tło. Należy również pamiętać o dostępności dla osób

[26] https://medium.com/@alaa511str/beyond-the-grid-exploring-dashboard-layouts-and-their-superpowers-74707809552b, [dostęp 31.12.2024]

z zaburzeniami widzenia kolorów, stosując kombinacje odcieni, które są czytelne dla osób z daltonizmem, na przykład szarości i niebieskiego. Konsekwentne wykorzystanie tego schematu na wszystkich stronach raportu pozwala na stworzenie spójnego wrażenia estetycznego oraz ułatwia rozpoznawanie najważniejszych informacji[27].

Podsumowując, skuteczne projektowanie dashboardu opiera się na przemyślanym planowaniu stron raportu, intuicyjnym rozmieszczeniu komponentów oraz harmonijnym schemacie kolorystycznym. Przestrzeganie tych zasad umożliwia użytkownikom łatwiejsze poruszanie się po raportach i szybsze przyswajanie prezentowanych informacji.

## 5.3. Wybór wykresów i wizualizacji

Dobór właściwych wykresów i wizualizacji odgrywa istotną rolę w procesie projektowania efektywnych dashboardów. Jest to nie tylko kwestia estetyki, lecz przede wszystkim sposób na efektywne przekazanie informacji. Trafnie dobrana wizualizacja upraszcza złożone dane, uwydatnia istotne wzorce i zależności, a także ogranicza ryzyko błędnych interpretacji.

Podstawą właściwego wyboru wizualizacji jest jej dostosowanie do specyfiki prezentowanych danych oraz celu, który ma zostać osiągnięty. Jak omówiono w poprzednich rozdziałach, niezwykle istotne jest uwzględnienie zasad projektowania, takich jak redukcja obciążenia poznawczego, stosowanie reguł gestaltu czy wykorzystanie atrybutów przetwarzanych mimowolnie. Każdy element wizualizacji powinien być projektowany z myślą o oczekiwaniach i możliwościach odbiorców.

Podstawowe rodzaje wykresów, takie jak słupkowe, liniowe, punktowe czy pudełkowe, oferują wiele opcji w zakresie prezentacji danych ilościowych, jakościowych i relacji między nimi. Wykresy liniowe doskonale nadają się do ukazywania zmian w czasie, podczas gdy wykresy słupkowe ułatwiają porównywanie wartości w różnych kategoriach. Dobór odpowiedniego rodzaju wykresu powinien być podyktowany dążeniem do maksymalnej czytelności oraz skuteczności w przekazywaniu informacji.

Równie istotne jest przestrzeganie zasad projektowania omówionych wcześniej w pracy. Wykresy powinny być spójne pod względem kolorystyki, zgodne z przyjętymi standardami estetycznymi i funkcjonalnymi. Warto unikać nadmiernego skomplikowania

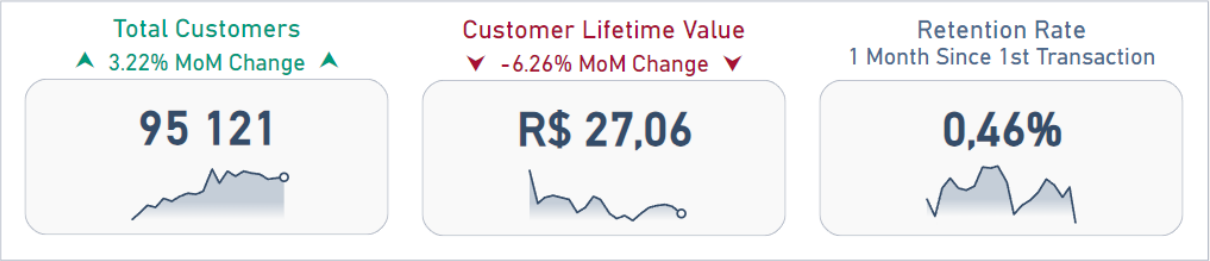
[27] https://dashboards.mysidewalk.com/style-guide-for-dashboards/color, [dostęp 31.12.2024]

wizualizacji i stosować hierarchię wizualną, aby wyróżnić najważniejsze elementy. Ponadto należy zadbać o odpowiednie marginesy i logiczne rozmieszczenie komponentów, co pozwala na zachowanie przejrzystości i porządku.

W kolejnych częściach tego rozdziału zostaną zaprezentowane konkretne przykłady wizualizacji wykorzystanych w raporcie, wraz z ich szczegółowym omówieniem w kontekście omówionych zasad projektowania. Każda z omawianych sekcji będzie uzupełniona o rysunek ilustrujący dany przykład, co pozwoli lepiej zobrazować opisywane rozwiązania oraz podkreślić ich zastosowanie w praktyce.

### 5.3.1. Karty KPI

Na stronie „Customer" kluczowe wskaźniki efektywności zostały przedstawione za pomocą kart, które umieszczono w górnej części raportu. To strategiczne rozmieszczenie wynika z dążenia do ułatwienia użytkownikom dostępu do najważniejszych informacji już na pierwszy rzut oka, zgodnie z zasadą umieszczania najistotniejszych danych w najbardziej widocznych obszarach.




**Rysunek 12.** Karty KPI na stronie „Customer" (źródło: opracowanie własne)

Każda karta reprezentuje jeden ze wskaźników, takich jak liczba klientów („Total Customers"), wartość klienta w czasie („Customer Lifetime Value") oraz wskaźnik retencji („Retention Rate"). Oprócz prezentacji bieżących wartości KPI, karty wzbogacono o podtytuły, które są dynamiczne i informują o zmianie wskaźnika miesiąc do miesiąca (MoM Change). Podtytuły te zostały utworzone za pomocą języka DAX, wykorzystując miary analizy czasowej, co zapewnia ich automatyczną aktualizację na podstawie bieżących danych.

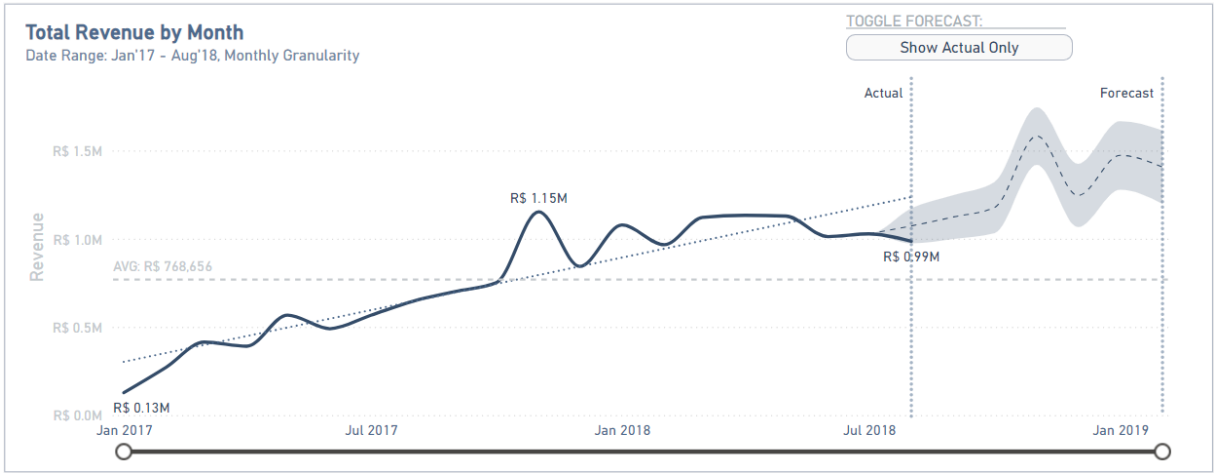
Istotnym elementem jest również formatowanie kolorystyczne podtytułów, które pomaga użytkownikom w szybkim rozpoznawaniu trendów. Wskaźniki wzrostu są prezentowane w kolorze zielonym, co intuicyjnie sugeruje pozytywne zmiany, natomiast spadki wyróżniono kolorem czerwonym, sygnalizując potencjalne obszary wymagające uwagi. Takie

podejście wspiera percepcję wizualną użytkownika, umożliwiając natychmiastowe wychwycenie kluczowych informacji bez potrzeby ich głębszej analizy.

Dodatkowo każda karta zawiera miniaturową linię trendu, która dostarcza użytkownikom wizualnej reprezentacji historycznych zmian wskaźnika. Linie te zostały stworzone przy użyciu języka DAX oraz kodu SVG, co pozwala na ich dynamiczne dostosowywanie w zależności od zmieniających się danych. Dzięki tej funkcjonalności użytkownicy mogą szybko ocenić, czy dany wskaźnik podlega stabilnym wzrostom, spadkom, czy też fluktuacjom w czasie.

### 5.3.2. Wykres liniowy

Wykres liniowy przedstawiający miesięczne przychody stanowi doskonały przykład odpowiedniego dopasowania typu wizualizacji do rodzaju danych. Dzięki swojej konstrukcji idealnie nadaje się do analizy trendów w czasie, umożliwiając odbiorcom szybkie rozpoznanie wzrostów, spadków oraz okresów stabilizacji. W omawianym przykładzie dodatkowo wprowadzono liczne elementy wspierające interpretację danych oraz zwiększające wartość analityczną wykresu.




**Rysunek 13.** Wykres liniowy miesięcznych przychodów (źródło: opracowanie własne)

Podtytuł umieszczony nad wykresem precyzyjnie określa zakres czasowy prezentowanych danych („Date Range: Jan'17 – Aug'18") oraz szczegółowość ich agregacji („Monthly Granularity"). Jest to przydatne dla użytkowników, którzy dzięki tym informacjom od razu wiedzą, czego mogą oczekiwać od wizualizacji.

Sam wykres wzbogacono o szereg dodatkowych elementów, które wspierają analizę. Zastosowano linię trendu, która przedstawia ogólną tendencję przychodów w analizowanym

okresie, a także średnią wartość przychodów zaznaczoną jako poziomą linię. W celu przewidywania przyszłych wyników dodano sekcję prognozy, która znajduje się po prawej stronie wykresu i jest wyraźnie odseparowana od rzeczywistych danych za pomocą przerywanych linii pionowych. Sekcja prognozy nie tylko odróżnia się wizualnie, ale także zawiera cieniowaną strefę, która ilustruje przedział ufności prognozy.

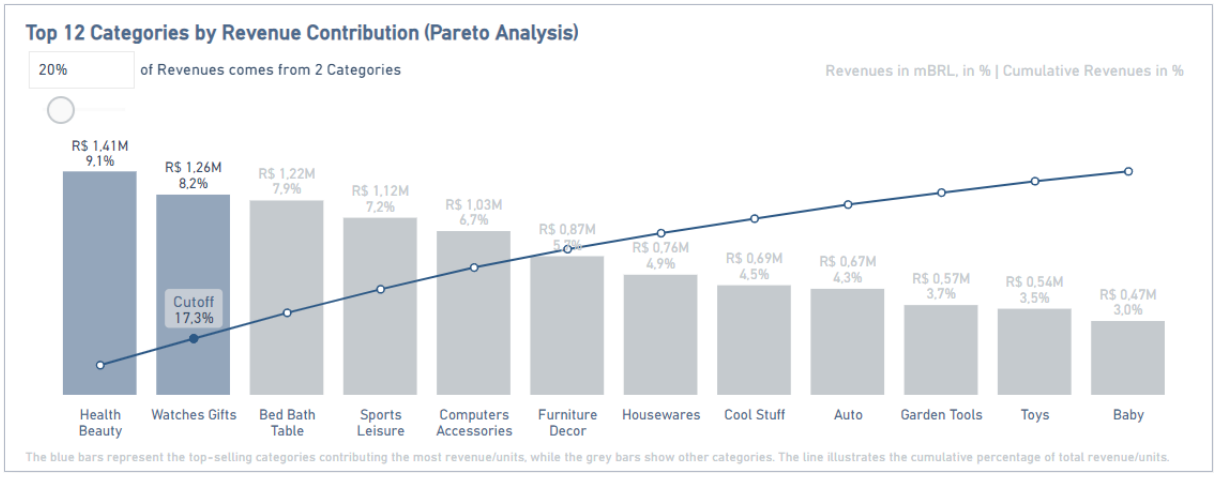
Wykres uwzględnia wiele wzorców projektowych, które wspierają jego czytelność i estetykę. Oś y oraz linia średniej zostały zaprojektowane w kolorze szarym, co skutecznie zepchnęło je na dalszy plan, nie odwracając uwagi od kluczowych informacji. Linia wartości rzeczywistych wyróżnia się swoją ciągłością i wyrazistością w stosunku do przerywanej linii prognozy, co podkreśla różnicę między danymi historycznymi a przewidywaniami. Wartości przedstawione na wykresie ograniczono do tych najistotniejszych – początkowej, końcowej i maksymalnej – co minimalizuje przeładowanie informacyjne i umożliwia łatwiejsze zrozumienie głównych punktów.

Dostosowanie tego typu wykresu wymagało zaawansowanego podejścia, które wiąże się z wykorzystaniem języka DAX do tworzenia niestandardowych elementów. Każdy dodatek, taki jak pionowe linie stref czy dynamiczne wartości na osiach, został wygenerowany za pomocą dodatkowego kodu, co pozwoliło osiągnąć pełną dynamikę i elastyczność wizualizacji.

### 5.3.3. Wykres kolumnowy

Wykres kolumnowy przedstawiający przychody w podziale na kategorie jest szczególnie odpowiedni do prezentacji danych tego rodzaju, ponieważ umożliwia jednoczesne porównanie wartości liczbowych. Dzięki swojej konstrukcji pozwala na szybkie wychwycenie kategorii generujących największe przychody, co wspiera procesy analityczne i decyzyjne.

Wzbogacenie wykresu o linię wartości skumulowanych wprowadza dodatkowy wymiar analizy, wskazując, w jakim stopniu poszczególne kategorie przyczyniają się do całkowitych przychodów. Linia ta wizualizuje narastający udział procentowy, dzięki czemu odbiorcy mogą z łatwością zidentyfikować punkt, w którym określona liczba kategorii generuje dominującą część przychodów (np. 17.3% przychodów pochodzi z dwóch kategorii). Interaktywny suwak, umieszczony w lewym górnym rogu wizualizacji, pozwala użytkownikowi dynamicznie dostosowywać punkt odcięcia, co znacząco podnosi funkcjonalność wykresu i umożliwia personalizację analizy.




**Rysunek 14.** Wykres kolumnowy z analizą Pareto dla przychodów w podziale na kategorie (źródło: opracowanie własne)

Wizualizacja zawiera liczne elementy wspierające interpretację. Każda kolumna posiada etykietę liczbową przedstawiającą przychód oraz procentowy udział danej kategorii w całkowitych przychodach. Dodatkowo, w prawym górnym rogu umieszczono szczegółowy opis formatu prezentowanych danych. Na dole wykresu znajduje się natomiast dodatkowy opis wyjaśniający znaczenie kolorów i kluczowych elementów wizualizacji, co wzmacnia klarowność przekazu.

Zastosowane wzorce projektowe znacząco podnoszą estetykę oraz czytelność wykresu. Wszystkie dodatkowe opisy, zarówno w prawym górnym, jak i dolnym obszarze wizualizacji, zostały zaprojektowane w szarych odcieniach, co sprawia, że nie odciągają uwagi od głównego przekazu. Kolumny i etykiety należące do kategorii znajdujących się poniżej ustalonego punktu odcięcia również są utrzymane w szarym kolorze, co dodatkowo wzmacnia wizualny nacisk na najważniejsze dane. Linia skumulowana jest wyraźnie zaznaczona i zawiera dynamiczną etykietę wskazującą aktualną wartość punktu odcięcia, co czyni wykres interaktywnym i intuicyjnym.

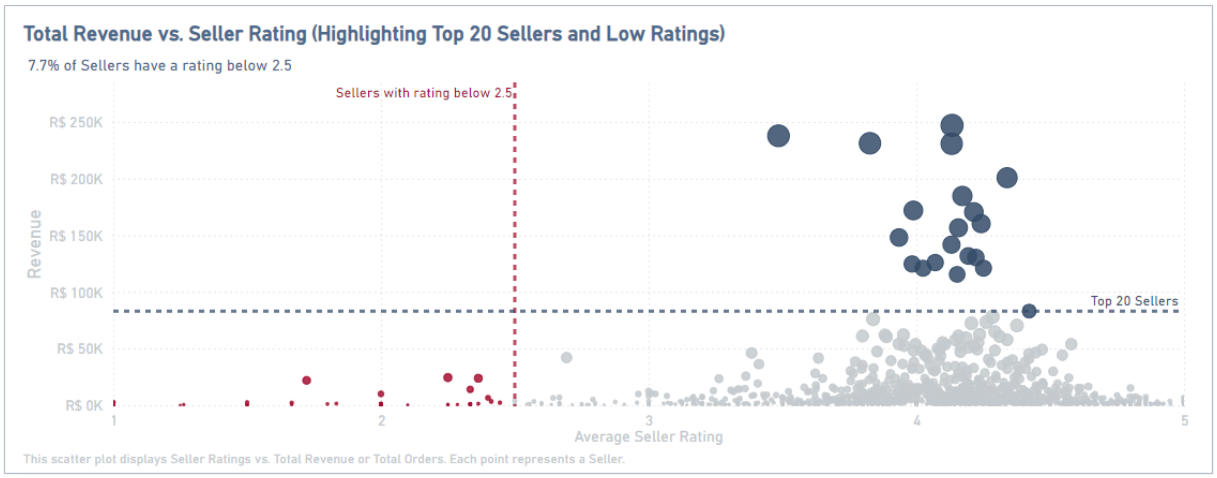
Stworzenie tak zaawansowanej wizualizacji wymagało napisania wielu niestandardowych miar w języku DAX. Miary te odpowiadają zarówno za obliczanie wartości skumulowanych, jak i za dynamiczne aktualizowanie punktu odcięcia oraz formatowania warunkowego.

### 5.3.4. Wykres punktowy

Wykres punktowy stanowi doskonały wybór do prezentacji relacji pomiędzy dwiema zmiennymi ilościowymi, w tym przypadku całkowitymi przychodami generowanymi przez

sprzedawców oraz ich średnią oceną. Dzięki swojej strukturze umożliwia on identyfikację zarówno ogólnych trendów, jak i szczególnych grup w analizowanych danych, co czyni go niezwykle użytecznym narzędziem w procesach decyzyjnych i optymalizacyjnych.

Prezentowana wizualizacja została wzbogacona o dodatkowe elementy, które znacząco ułatwiają interpretację danych. Pozioma linia przerywana wyznacza próg najlepszych dwudziestu sprzedawców pod względem generowanych przychodów, natomiast pionowa linia przerywana wskazuje sprzedawców z oceną poniżej 2.5, co czyni ich potencjalnym źródłem problemów dla organizacji. Te progi pozwalają na szybkie zidentyfikowanie dwóch kluczowych grup: sprzedawców, którzy przynoszą największe korzyści, oraz tych, którzy mogą wymagać poprawy jakości swoich usług. Dodatkowo podtytuł informuje o procentowym udziale sprzedawców z oceną poniżej 2.5, co pozwala odbiorcy natychmiast zorientować się w skali problemu. Opis na dole wykresu zawiera szczegółowe informacje dotyczące prezentowanych danych oraz zasad ich wizualizacji, co wspiera klarowność przekazu.

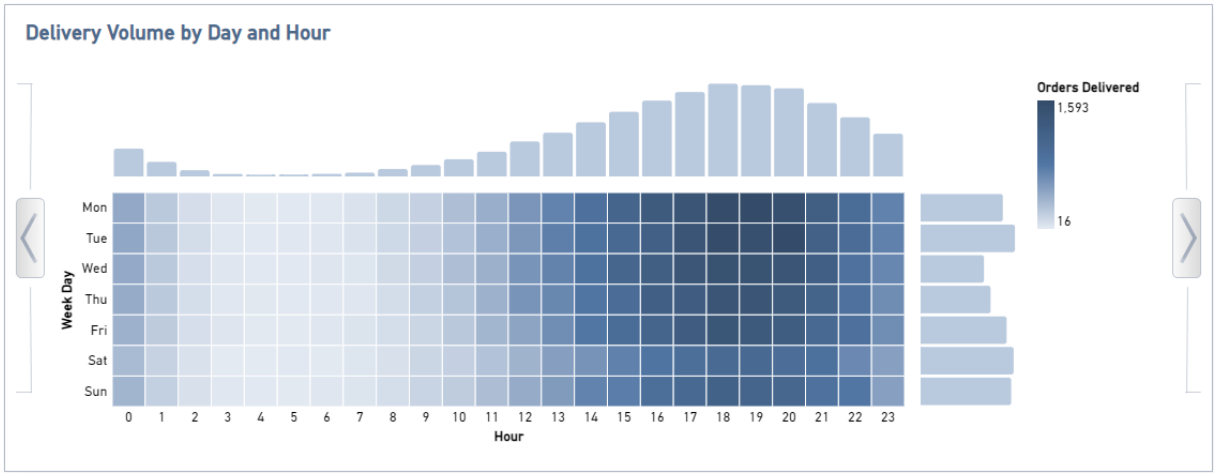



**Rysunek 15.** Wykres punktowy przedstawiający sprzedawców (źródło: opracowanie własne)

Wykorzystane zasady projektowania zwiększają zarówno użyteczność, jak i estetykę wykresu. Osie wykresu oraz ich etykiety zostały utrzymane w neutralnym, szarym kolorze, co minimalizuje ich wizualny wpływ i pozwala skupić się na danych. Podobne podejście zastosowano w przypadku sprzedawców, którzy nie należą do kluczowych grup – ich punkty na wykresie są wyświetlane w szarym kolorze, dzięki czemu nie odciągają uwagi od sprzedawców oznaczonych jako najważniejsi. Najlepsi sprzedawcy zostali wyróżnieni kolorem ciemnoniebieskim, a sprzedawcy z najniższą oceną – kolorem czerwonym. Takie intuicyjne wykorzystanie kolorów wspiera szybkie rozpoznanie istotnych grup.

### 5.3.5. Mapa cieplna

Mapa cieplna jest doskonałym narzędziem wizualizacyjnym do przedstawienia intensywności danych w dwóch wymiarach. Prezentowany przykład dotyczy analizy liczby dostarczonych zamówień w podziale na dni tygodnia oraz godziny. Każde pole na wykresie reprezentuje kombinację określonego dnia tygodnia i godziny, a intensywność koloru wskazuje na liczbę zamówień dostarczonych w danym przedziale czasowym. Wartości są dodatkowo przedstawione na skali kolorystycznej po prawej stronie, co umożliwia szybkie i intuicyjne zrozumienie danych.



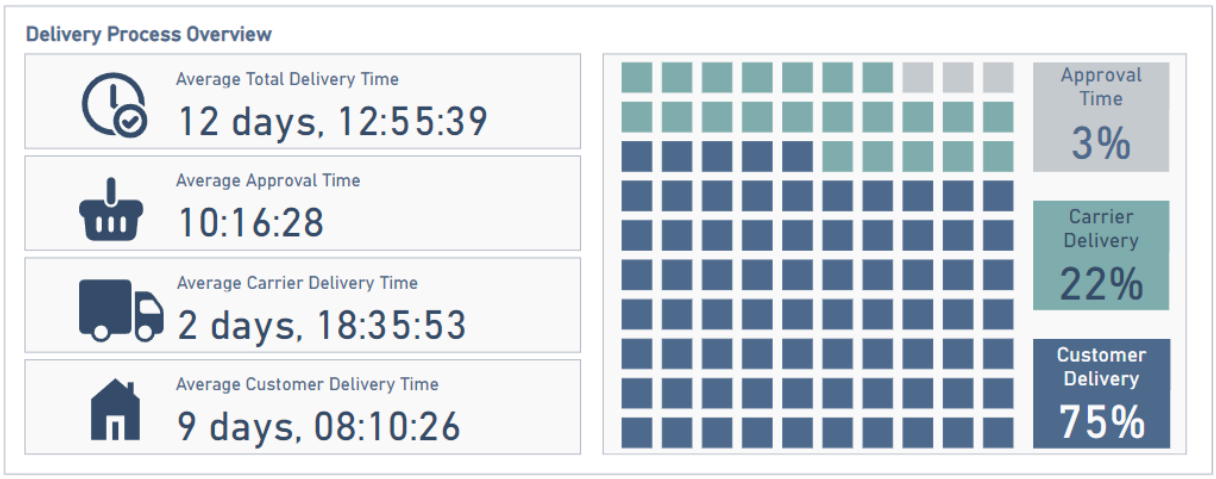

**Rysunek 16.** Mapa cieplna przedstawiająca ilość dostarczonych zamówień w podziale na dni tygodnia i godziny (źródło: opracowanie własne)

Wizualizacja ta została stworzona za pomocą niestandardowego narzędzia „Deneb: Declarative Visualization", co wyróżnia ją na tle natywnych rozwiązań Power BI. Deneb wykorzystuje deklaratywną składnię JSON, opartą na językach Vega i Vega-Lite, co pozwala na dużą elastyczność i precyzyjne dostosowanie wykresu do specyficznych potrzeb analizy. Takie podejście okazuje się często prostsze, niż próby dostosowania standardowych wizualizacji dostępnych w Power BI.

Zaletą tego podejścia jest możliwość szybkiego stworzenia wizualizacji, która dokładnie spełnia założenia projektu, z zachowaniem przejrzystości i estetyki. Deneb umożliwia pełną kontrolę nad strukturą wizualizacji, co pozwala na uniknięcie kompromisów związanych z ograniczeniami natywnych rozwiązań Power BI. W tym przypadku mapę cieplną wzbogacono o dodatkowe elementy, takie jak sumy marginesowe dla dni tygodnia oraz godzin, co umożliwia użytkownikowi analizę zarówno szczegółowych, jak i zbiorczych wyników.

### 5.3.6. Diagram prostokątny

Diagram prostokątny stanowi efektywne narzędzie do wizualizacji podziału całości na części, gdy pomiędzy poszczególnymi częściami występują znaczące różnice. W omawianym przykładzie diagram obrazuje podział całkowitego czasu dostawy na trzy główne etapy: czas zatwierdzenia, czas dostawy realizowanej przez przewoźnika oraz czas dostawy do klienta. Każdy etap jest oznaczony odrębnym kolorem, co pozwala na intuicyjne zrozumienie danych.




**Rysunek 17.** Diagram prostokątny przedstawiający podział czasu dostawy na etapy w procesie dostawy (źródło: opracowanie własne)

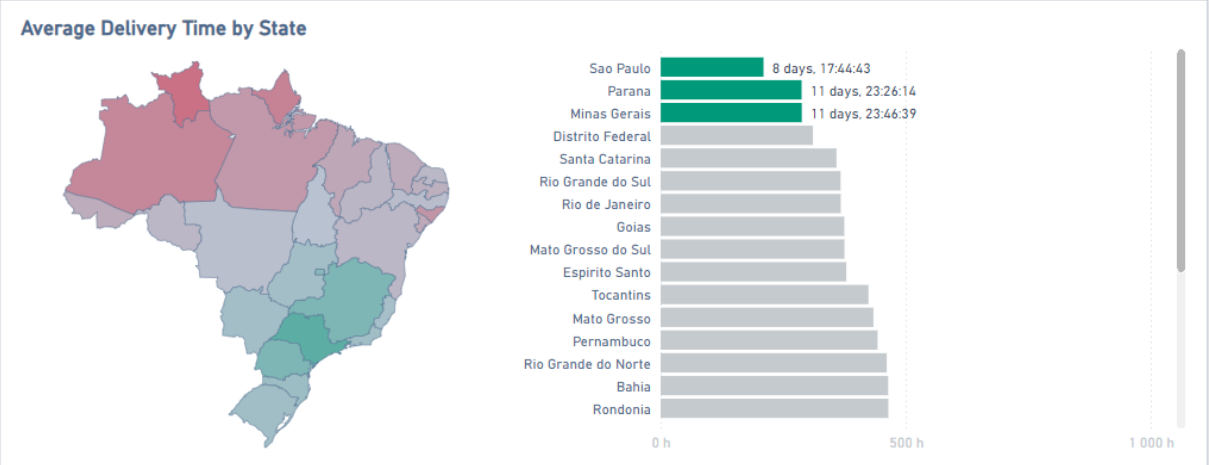
Warto zauważyć, że Power BI nie posiada natywnej opcji tworzenia tego rodzaju diagramów prostokątnych. Przedstawiona wizualizacja została opracowana na bazie wykresu punktowego, który został odpowiednio dostosowany w celu osiągnięcia finalnego efektu. To podejście wymagało zaawansowanej konfiguracji oraz wykorzystania miar do dopasowania kolorów poszczególnych elementów, aby uzyskać czytelny i estetyczny rezultat.

Dodatkową zaletą tego rozwiązania jest jego zdolność do zapewnienia klarownej komunikacji, szczególnie w przypadku wizualizacji danych hierarchicznych lub proporcjonalnych. Diagram prostokątny doskonale pokazuje, że etap dostawy do klienta zajmuje aż 75% całkowitego czasu, co zostało dodatkowo podkreślone za pomocą wyróżnionego koloru i dynamicznej etykiety procentowej. Dzięki temu odbiorcy mogą natychmiast zidentyfikować najważniejsze elementy procesu.

### 5.3.7. Kartogram

Kartogram stanowi jedno z najbardziej efektywnych narzędzi wizualizacji danych przestrzennych, które umożliwia identyfikację regionalnych wzorców oraz różnic

w intensywności danych. W przedstawionej wizualizacji kartogram ilustruje średni czas dostawy zamówień w podziale na poszczególne stany Brazylii, co pozwala na szybkie wychwycenie obszarów, gdzie proces realizacji zamówień jest bardziej efektywny, a gdzie wymaga usprawnienia. Wartość dodaną w tym przypadku stanowi połączenie kartogramu z wykresem słupkowym, który dostarcza precyzyjnych danych liczbowych dla każdego regionu, umożliwiając bardziej szczegółowe porównania między stanami.




**Rysunek 18.** Kartogram prezentujący średni czas dostawy w podziale na stany wraz z dodatkiem w postaci wykresu słupkowego (źródło: opracowanie własne)

Zastosowane wzorce projektowe zapewniają zarówno wysoką estetykę, jak i przejrzystość wizualizacji. Na kartogramie użyto odpowiednich odcieni kolorów, aby zasygnalizować różnice w średnich czasach dostawy. Zielone odcienie wskazują na krótsze czasy realizacji zamówień, podczas gdy czerwone odcienie reprezentują dłuższe czasy. Dzięki temu odbiorca może łatwo zidentyfikować regiony, które wymagają większej uwagi, bez konieczności wczytywania się w szczegółowe wartości.

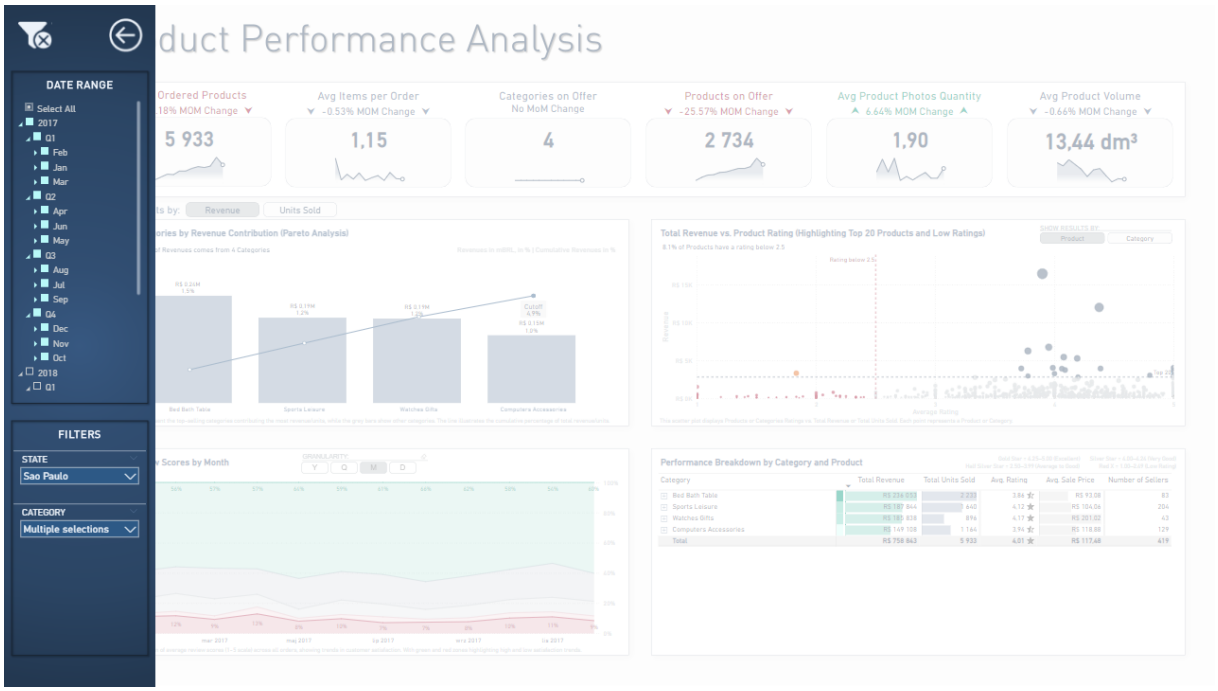
Wykres słupkowy, zlokalizowany obok kartogramu, służy jako element wspierający. Taka kombinacja umożliwia szczegółowe porównania między regionami, które na samym kartogramie mogłyby być trudne do uchwycenia. Na wykresie wyraźnie wyróżniono stany o skrajnych wartościach, co pomaga odbiorcy w szybkiej identyfikacji kluczowych obszarów. Przyjęto zasadę minimalizmu, dzięki której mniej istotne elementy, takie jak osie wykresu czy stany o średnich wynikach, zostały zepchnięte na dalszy plan za pomocą szarego koloru.

## 5.4. Interaktywność i funkcjonalność raportu

Jednym z fundamentalnych aspektów projektowania raportów w narzędziach analitycznych, takich jak Power BI, jest stworzenie środowiska sprzyjającego intuicyjnemu

poruszaniu się po danych. Interaktywność i użyteczność raportu odgrywają niezmiernie ważną rolę, szczególnie w kontekście umożliwienia pogłębionej eksploracji informacji oraz dostosowania treści do wymagań użytkownika. Elementy takie jak nawigacja, panele filtrowania, etykiety wizualizacji oraz interaktywne przyciski stanowią integralną część nowoczesnych raportów.

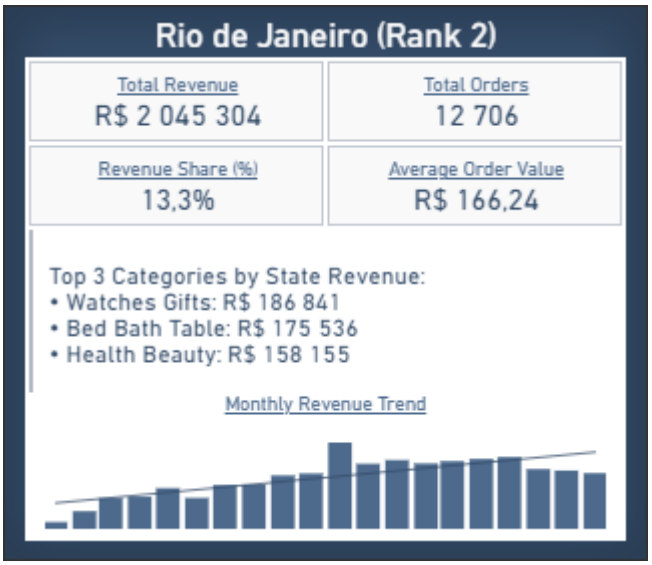
Panel filtrowania to jedno z głównych narzędzi interaktywnych, które umożliwia użytkownikom ograniczenie zakresu prezentowanych danych. Na przykład, w jednym z widoków raportu panel pozwala na filtrowanie danych według zakresu dat, lokalizacji lub kategorii produktów (Rys. 19). Filtry te są zorganizowane i zintegrowane w sposób pozwalający na szybkie wprowadzenie zmian. Dzięki precyzyjnie zaprojektowanemu interfejsowi użytkownicy mogą analizować dane w czasie rzeczywistym, bez potrzeby modyfikacji struktury raportu. Tego typu rozwiązania nie tylko upraszczają dostęp do szczegółowych informacji, ale również oszczędzają czas, eliminując konieczność przygotowywania wielu statycznych widoków.




**Rysunek 19.** Interaktywny panel filtrowania w projekcie (źródło: opracowanie własne)

Etykiety wizualizacji, takie jak dynamiczne podsumowania danych, odgrywają ważną rolę w prezentacji informacji. Przykładowo, etykiety dla danych sprzedażowych w podziale na lokalizacje nie tylko przedstawiają podstawowe statystyki, takie jak całkowite przychody, liczba zamówień czy średnia wartość zamówienia, ale również akcentują najważniejsze

kategorie produktów (Rys. 20). Takie rozwiązanie pozwala użytkownikom raportu szybko identyfikować istotne trendy i koncentrować się na kluczowych aspektach analizy. Etykiety te są uzupełnione o wykresy trendów miesięcznych, które w przejrzysty sposób przedstawiają zmiany w czasie. Tworzenie dynamicznych etykiet wymaga zastosowania zaawansowanych funkcji, takich jak kalkulacje w języku DAX, które umożliwiają dostosowanie treści do kontekstu prezentowanych danych.



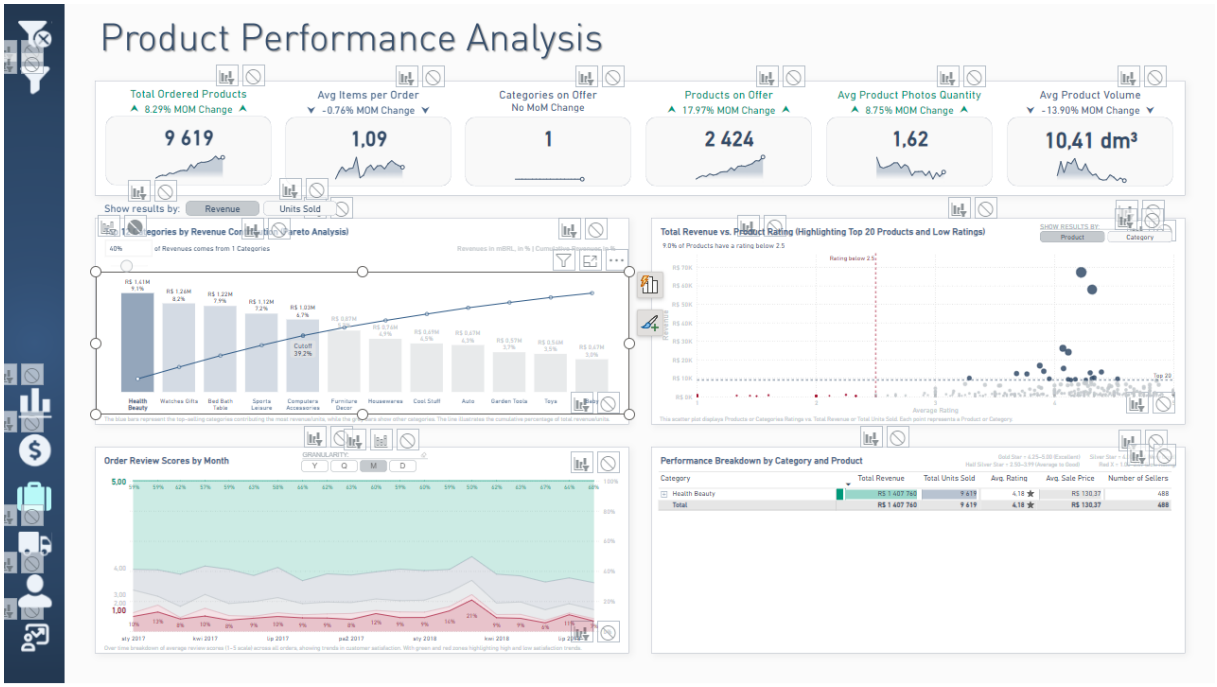

**Rysunek 20.** Dynamiczna etykieta dla wybranego stanu prezentująca dane sprzedażowe (źródło: opracowanie własne)

Intuicyjne przyciski nawigacyjne, w tym zakładki, stanowią kolejny element zwiększający użyteczność raportu. Zakładki umożliwiają łatwe przechodzenie między różnymi widokami raportu lub aktywowanie wybranych konfiguracji filtrów i wizualizacji. Takie rozwiązania sprawiają, że korzystanie z raportu jest nie tylko prostsze, ale także bardziej spójne i intuicyjne. Dzięki nim użytkownicy mogą natychmiast przechodzić do interesujących ich sekcji, bez potrzeby każdorazowego dostosowywania ustawień.

Zintegrowanie takich elementów, jak panele filtrowania, dynamiczne etykiety i przyciski nawigacyjne, tworzy środowisko raportowe, które jest jednocześnie estetyczne i funkcjonalne. Takie podejście zwiększa komfort użytkowania raportu oraz wspiera efektywność analizy danych. Projektowanie tego rodzaju rozwiązań wymaga jednak dogłębnej znajomości możliwości narzędzi analitycznych oraz zrozumienia potrzeb użytkowników raportu.

## 5.5. Testowanie i udostępnianie raportu w środowisku produkcyjnym

W końcowym etapie projektowania raportu ważnym krokiem jest sprawdzenie poprawności interakcji i przetestowanie funkcji w środowisku produkcyjnym. Skutecznie działające mechanizmy filtrowania oraz filtrowania krzyżowego, a także odpowiednio zaprojektowane interakcje między wizualizacjami, są fundamentem, aby użytkownicy mogli w pełni wykorzystywać możliwości raportu i podejmować decyzje oparte na dostarczanych danych.

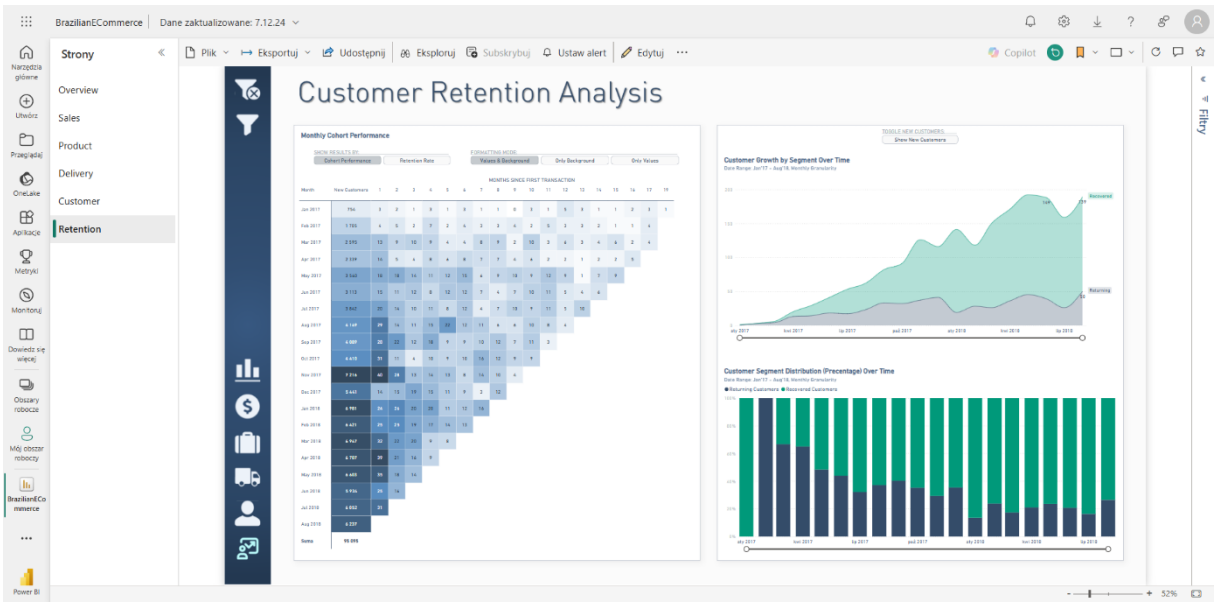



**Rysunek 21.** Konfiguracja interakcji między wizualizacjami w raporcie (źródło: opracowanie własne)

Pierwszym zadaniem jest dokładna weryfikacja działania filtrów i mechanizmów filtrowania krzyżowego. Testowanie obejmuje analizę poprawności działania filtrów, ich zastosowania w danych oraz interakcji pomiędzy różnymi wizualizacjami zgodnie z zamierzeniami projektowymi. Należy uwzględnić różnorodne scenariusze interakcji, w tym sytuacje, gdy użytkownicy dokonują wielokrotnych wyborów filtrów, sprawdzając, jak wpływają one na inne elementy raportu. Jest to szczególnie ważne w przypadku zaawansowanych raportów, w których zależności między wizualizacjami odgrywają znaczącą rolę w budowaniu pozytywnego doświadczenia użytkownika.

Jeżeli raport ma być dostosowany do potrzeb różnych grup użytkowników, istotnym krokiem jest wdrożenie mechanizmu ról użytkowników. Role te pozwalają personalizować dostęp do danych i wizualizacji w zależności od specyficznych wymagań i kompetencji poszczególnych grup odbiorców. Podczas definiowania ról należy dokładnie przeanalizować potrzeby organizacji i zapewnić, że każdy użytkownik ma dostęp wyłącznie do informacji niezbędnych do wykonywania swoich obowiązków. Takie podejście zwiększa bezpieczeństwo danych oraz wspiera przejrzystość w dostępie do informacji.

Kolejnym etapem jest publikacja raportu w środowisku produkcyjnym, takim jak Power BI Service. Po publikacji raportu należy przeprowadzić testy w tym środowisku, aby upewnić się, że raport działa prawidłowo w rzeczywistych warunkach użytkowania. Weryfikacja obejmuje analizę interfejsu użytkownika, dostępności danych, szybkości działania oraz poprawności wyświetlania wizualizacji na różnych urządzeniach. Dodatkowo należy sprawdzić, czy role użytkowników i ich uprawnienia do danych funkcjonują zgodnie z założeniami.




**Rysunek 22.** Widok publikowanego raportu w Power BI Service (źródło: opracowanie własne)

Testowanie raportu w środowisku produkcyjnym pozwala również zidentyfikować potencjalne problemy z wydajnością lub dostępnością danych. W przypadku wykrycia jakichkolwiek nieprawidłowości projektant powinien wprowadzić odpowiednie poprawki i ponownie opublikować raport. Taki iteracyjny proces testowania i doskonalenia raportu gwarantuje wysoką jakość produktu końcowego oraz satysfakcję jego użytkowników.

# Podsumowanie

Realizacja projektu analitycznego w Power BI umożliwiła kompleksowe opracowanie interaktywnego raportu opartego na danych z sektora e-commerce. W ramach pracy przeprowadzono transformację surowych danych, zoptymalizowano model oraz wdrożono szereg miar analitycznych, co pozwoliło na efektywne badanie zachowań klientów, identyfikację trendów sprzedażowych oraz ocenę wskaźników biznesowych. Uzyskane wyniki potwierdziły, że przyjęta metodologia pozwala na szybkie i intuicyjne pozyskiwanie wartościowych informacji, a interaktywność raportu znacząco zwiększa możliwości eksploracji danych. Dzięki dynamicznym wskaźnikom oraz analizom opartym na segmentacji klientów raport dostarcza praktycznych wniosków wspierających proces podejmowania decyzji biznesowych.

Zastosowanie Power BI przyniosło liczne korzyści, w tym intuicyjną obsługę, integrację z wieloma źródłami danych oraz szerokie możliwości wizualizacyjne. Wbudowane mechanizmy automatycznej aktualizacji oraz rozbudowane funkcje analizy czasowej ułatwiły implementację dynamicznych wskaźników. Niemniej jednak podczas realizacji projektu ujawniły się również pewne ograniczenia tego narzędzia. Praca z dużymi zbiorami danych wymagała optymalizacji modelu oraz eliminacji zbędnych elementów w celu poprawy wydajności. Dodatkowo, ograniczenia związane z zaawansowanym dostosowaniem wizualizacji oraz ich formatowaniem stanowiły istotne wyzwanie. W niektórych przypadkach konieczne było zastosowanie alternatywnych rozwiązań w celu uzyskania pożądanej estetyki i czytelności raportu. Aspekty te mogą stanowić obszar dalszego doskonalenia projektu oraz eksploracji bardziej zaawansowanych technik dostosowywania wizualizacji dostępnych w Power BI.

W kontekście przyszłych usprawnień raportu możliwe jest wdrożenie dodatkowych źródeł danych, takich jak informacje rynkowe czy demograficzne, co mogłoby wzbogacić zakres analizy. Ponadto wykorzystanie modeli predykcyjnych pozwoliłoby na prognozowanie trendów sprzedaży i dokładniejsze przewidywanie zachowań klientów. Istotnym obszarem dalszego rozwoju jest także analiza opinii użytkowników poprzez zastosowanie metod przetwarzania języka naturalnego, umożliwiających automatyczną klasyfikację oraz ocenę sentymentu zawartego w recenzjach produktów. Takie podejście pozwoliłoby na identyfikację najczęściej zgłaszanych problemów oraz określenie czynników wpływających na satysfakcję klientów.

Dalsza optymalizacja modelu danych i procesów ekstrakcji, transformacji i ładowania informacji mogłaby przyczynić się do zwiększenia wydajności raportu, zwłaszcza w kontekście obsługi dużych wolumenów danych. Dodatkowo wdrożenie bardziej zaawansowanych metod wizualizacji, takich jak niestandardowe wykresy, mogłoby zwiększyć użyteczność raportu dla różnych grup użytkowników. Kolejnym etapem rozwoju może być również integracja raportu z innymi systemami analitycznymi.

Podsumowując, projekt wykazał wysoką użyteczność Power BI jako narzędzia do analizy i wizualizacji danych e-commerce. Pomimo pewnych ograniczeń, opracowane rozwiązania umożliwiły skuteczną prezentację istotnych wskaźników oraz analizę trendów, co stanowi solidną podstawę do dalszego rozwoju i rozszerzenia zakresu analizy w przyszłości.

# Bibliografia


1. Chrabski B., Zmitrowicz K.: *Inżynieria wymagań w praktyce*, Wydawnictwo Naukowe PWN, Warszawa 2015
2. Hand D., Mannila H., Smyth P.: *Eksploracja danych*, Wydawnictwa Naukowo-Techniczne, Warszawa 2005
3. Tanimura C.: *Analiza danych z wykorzystaniem SQL-a: Zaawansowane techniki przekształcania danych we wnioski*, Helion, Gliwice 2022
4. Alexander M., Decker J., Wehbe B.: *Analizy Business Intelligence: Zaawansowane wykorzystanie Excela*, Helion, Gliwice 2019
5. Knight D., Pearson M., Schacht B., Ostrowsky E.: *Microsoft Power BI: Jak modelować i wizualizować dane oraz budować narracje cyfrowe*, Helion, Gliwice 2022
6. Russo M., Ferrari A.: *Kompletny przewodnik po DAX: Analiza biznesowa przy użyciu Microsoft Power BI, SQL Server Analysis Services i Excel*, APN Promise, Warszawa 2019
7. Cheverton P.: *Zarządzanie kluczowymi klientami: Jak uzyskać status głównego dostawcy*, Oficyna Ekonomiczna, Kraków 2001
8. Nussbaumer Knaflic C.: *Storytelling danych: Poradnik wizualizacji danych dla profesjonalistów*, Helion, Gliwice 2019
9. *How to Select Metrics for Your KPI Dashboard*, https://www.grow.com/blog/how-to-select-metrics-for-your-kpi-dashboard, [dostęp 30.12.2024]
10. Jain A.: *How to Optimize Your Power BI Data Model*, https://ashwinijain.medium.com/how-to-optimize-your-power-bi-data-model-1da03f48ec8e, [dostęp 31.12.2024]
11. Ferrari A.: *The Importance of Star Schemas in Power BI*, https://www.sqlbi.com/articles/the-importance-of-star-schemas-in-power-bi/, [dostęp 30.12.2024]

12. Russo M.: *Data Import Best Practices in Power BI*, https://www.sqlbi.com/articles/data-import-best-practices-in-power-bi/, [dostęp 30.12.2024]

13. Russo M.: *Calculated Columns and Measures in DAX*, https://www.sqlbi.com/articles/calculated-columns-and-measures-in-dax/, [dostęp 31.12.2024]

14. *Beyond the Grid: Exploring Dashboard Layouts and Their Superpowers*, https://medium.com/@alaa511str/beyond-the-grid-exploring-dashboard-layouts-and-their-superpowers-74707809552b, [dostęp 31.12.2024]

15. *Color*, https://dashboards.mysidewalk.com/style-guide-for-dashboards/color, [dostęp 31.12.2024]

16. *Data schema*, https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce, [dostęp 07.10.2024]

17. *Star schema*, https://medium.com/@nimanthaF/data-modelling-techniques-star-schema-f1077a1cced7, [dostęp 30.12.2024]

# Spis tabel



# Spis rysunków

# Spis kodu