

Hurtownie danych

Projekt – Analiza danych platformy e-commerce Olist (Brazylia)



Aleksander stepaniuk

nr. indeksu: 272644

Politechnika Wrocławska, Informatyka Stosowana

**Etap I – 19.05.2025 r.**

1. **Tytuł projektu**

Analiza danych platformy e-commerce Olist (Brazylia)

1. **Charakterystyka dziedziny problemowej, krótki opis obszaru analizy, problemy i potrzeby**

Celem projektu jest analiza brazylijskiego rynku e-commerce, gdzie platforma Olist pośredniczy między tysiącami sprzedawców, a setkami tysięcy klientów, gromadząc dane o zamówieniach, płatnościach, produktach i opiniach.

W obszarze analizy ważne jest śledzenie przebiegu transakcji od momentu złożenia zamówienia aż po dostawę i ocenę satysfakcji użytkownika. Rozproszone w różnych plikach typu CSV źródła danych utrudniają szybkie agregowanie takich informacji. Problemy pojawiają się także przy zapewnieniu spójnej jakości danych – niekompletne lub niespójne rekordy recenzji czy płatności mogą zafałszować raporty, a brak zintegrowanego modelu danych wydłuża czas przygotowania analiz.

Potrzeby biznesowe koncentrują się na możliwości błyskawicznego generowania wielowymiarowych raportów (np. przychód wg regionu i miesiąca, ocena sprzedawcy według kwartału), monitorowaniu najważniejszych wskaźników jakości obsługi klienta (czas dostawy, liczba reklamacji) oraz elastycznej segmentacji klientów i produktów, co wymaga wdrożenia hurtowni danych z jasno zdefiniowanymi wymiarami i faktami.

1. **Cel przedsięwzięcia (oczekiwania) oraz zakres analizy – badane aspekty**

Cel główny: zbudowanie hurtowni danych, która pozwoli na:

* Monitorowanie kluczowych wskaźników sprzedaży (przychód, liczba zamówień) w czasie.
* Analizę jakości dostaw i satysfakcji klienta (recenzje, czasy realizacji).
* Podział klientów i sprzedawców wg regionów i zachowań zakupowych.

Zakres:

* Dane sprzedażowe, zamówienia, płatności, opinie, dane klientów, sprzedawców, produktów.
* Analizy czasowe (miesiąc, kwartał), geograficzne (stan, miasto), produktowe (kategorie).

1. **Źródła danych (lokalizacja, format, dostępność), wstępna analiza źródeł danych**

Dane zostały pobrane ze strony kaggle.com:

* <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Lp.** | **Plik** | **Typ** | **Liczba rekordów** | **Rozmiar [MB]** | **Opis** |
| 1 | olist\_customers\_dataset | .csv | 99441 | 9.03 | zamówienia (daty: złożenia, zatwierdzenia, dostawy) |
| 2 | olist\_geolocation \_dataset | .csv | 1000164 | 61.27 | pozycje zamówień (produkt, ilość, cenа) |
| 3 | olist\_order\_items \_dataset | .csv | 112650 | 15.44 | klienci (id, miasto, stan, kod pocztowy) |
| 4 | olist\_order\_payments \_dataset | .csv | 103887 | 5.78 | płatności (metoda, rata, wartość) |
| 5 | olist\_order\_reviews\_dataset | .csv | 103887 | 14.45 | opinie klientów (ocena, komentarz, data) |
| 6 | olist\_orders\_dataset | .csv | 99442 | 17.65 | produkty (id, kategoria, wymiary) |
| 7 | olist\_products | .csv | 32952 | 2.38 | sprzedawcy (id, lokalizacja) |
| 8 | olist\_sellers | .csv | 3096 | 0.17 | tłumaczenia kategorii produktowych |
| 9 | product\_category\_name\_translation | .csv | 72 | 0.26 | geolokalizacja kodów pocztowych |

1. **Profilowanie danych (analiza jakości danych oraz ich przydatności w projekcie)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Lp.** | **Atrybut** | **Typ danych** | **Zakres wartości** | **Uwagi – ocena jakości danych** |
| 1 | **order\_purchase\_timestamp** | datetime | 2016-09-04 – 2018-09-03 | brak nulli, format ISO spójny w 100 % wierszy |
| 2 | **order\_approved\_at** | datetime | 2016-09-04 – 2018-09-05 | ~0,1 % null (zamówienia anulowane) |
| 3 | **order\_delivered\_carrier\_date** | datetime | 2016-09-07 – 2018-09-17 | ~0,2 % null (problemy logistyczne) |
| 4 | **order\_delivered\_customer\_date** | datetime | 2016-09-09 – 2018-09-23 | ~0,3 % null (zwrócone lub nie dostarczone) |
| 5 | **order\_status** | string | delivered, shipped, invoiced, created, approved… | brak nulli; wartości spójne |
| 6 | **price** | decimal(10,2) | 0.01 – 9999.00 | brak wartości ujemnych, ~0,01 % skrajnie niskich cen (promocje) |
| 7 | **payment\_type** | string | credit\_card, boleto, voucher, debit\_card | ~5 % null (zwroty/refundy); pozostałe wartości zgodne z dokumentacją |
| 8 | **payment\_installments** | integer | 1 – 12 | brak nulli, realistyczny rozkład (najwięcej 1–3 raty) |
| 9 | **review\_score** | integer | 1 – 5 | ~0,3 % null (brak opinii), średnia ocena ≈ 4,09 |
| 10 | **review\_creation\_date** | datetime | 2016-10-01 – 2018-10-15 | ~0,3 % null, daty recenzji mieszczą się do 30 dni po dostawie |
| 11 | **customer\_state** | string (2) | SP, RJ, MG, BA, CE, … | brak nulli, 27 kodów stanów (BR-XX), wszystkie poprawne zgodnie z ISO 3166-2:BR |
| 12 | **customer\_city** | string | São Paulo, Rio de Janeiro, Salvador, … | ~0,05 % literówek (akcenty), można ujednolicić wielkość liter |
| 13 | **seller\_state** | string (2) | SP, RJ, MG, PR, RS, … | brak nulli; ~3 095 unikalnych sprzedawców, wszystkie stany pokryte |
| 14 | **seller\_city** | string | São Paulo, Curitiba, Porto Alegre, … | ok, podobnie jak w kliencie: drobne literówki/różnice w zapisie |
| 15 | **product\_category\_name** | string | bed\_bath\_table, health\_beauty, sports\_leisure, … | 71 kategorii, wszystkie występują min. raz, brak nulli |
| 16 | **product\_weight\_g** | integer | 50 – 40000 | ~0,1 % null, wartości realistyczne, możliwe outliery do weryfikacji |
| 17 | **product\_length\_cm** | integer | 5 – 200 | ~0,1 % null, typowe zakresy dla e-commerce |
| 18 | **product\_height\_cm** | integer | 1 – 150 | analogicznie do długości |
| 19 | **product\_width\_cm** | integer | 2 – 150 | ok, można obliczyć objętość |
| 20 | **geolocation\_lat** | decimal(9,6) | -33.868820 – 5.193082 | ~2 % błędnych koordynat (poza granicami BR) – wymaga filtrowania |
| 21 | **geolocation\_lng** | decimal(9,6) | -73.985506 – -34.793129 | jak wyżej |

1. **Definicja typów encji/klas (wraz z właściwościami) oraz związków pomiędzy nimi, diagram klas (propozycja wymiarów, hierarchii, miar addytywnych i nieaddytywnych)**

**Encje wymiarów:**

1. DateDim
   * Klucz: date\_key (INT, YYYYMMDD)
   * Atrybuty: full\_date, year, quarter, month, day, weekday
   * Hierarchia: Year -> Quarter -> Month -> Day -> Weekday
2. CustomerDim
   * Klucz: customer\_id (VARCHAR)
   * Atrybuty: customer\_city\_key, customer\_zip\_code, customer\_segment
   * Hierarchia: State -> City -> Zip Code -> Segment
3. SellerDim
   * Klucz: seller\_id (VARCHAR)
   * Atrybuty: seller\_city\_key, seller\_zip\_code, seller\_segment
   * Hierarchia: State -> City -> Zip Code -> Segment
4. ProductDim
   * Klucz: product\_id (VARCHAR)
   * Atrybuty: category, sub\_category, weight\_g, height\_cm, width\_cm, length\_cm
   * Hierarchia: Category -> Sub-category -> Product
5. PaymentDim
   * Klucz: payment\_id (VARCHAR)
   * Atrybuty: payment\_type, installments, payment\_value
   * Hierarchia: Type -> Installments
6. ReviewDim
   * Klucz: review\_id (VARCHAR)
   * Atrybuty: review\_score, review\_date, review\_comment
   * Hierarchia: Score -> Date
7. CityDim
   * Klucz: city\_key (INT, autoincrement)
   * Atrybuty: state, city, population, area\_km2, density\_per\_km2
   * Hierarchia: State -> City

**Encja faktu:**

1. FactOrders
   * Klucz główny: order\_id (VARCHAR)
   * Klucze obce:
     + order\_date\_key -> DateDim
     + customer\_id -> CustomerDim
     + seller\_id -> SellerDim
     + product\_id -> ProductDim
     + payment\_id -> PaymentDim
     + review\_id -> ReviewDim
     + customer\_city\_key -> CityDim
     + seller\_city\_key -> CityDim
   * Miary addytywne:
     + order\_count (INT) – liczba zamówień,
     + total\_revenue (DECIMAL) – suma przychodu,
     + total\_items (INT) – liczba produktów.
   * Miary nieaddytywne:
     + average\_review\_score (DECIMAL) – średnia ocena;
     + payment\_type (VARCHAR), order\_status (VARCHAR) – opisowe, nie sumują się.

**Diagram klas:**

A screenshot of a computer screen

AI-generated content may be incorrect.

1. **Min. 10 wielowymiarowych zestawień, które zostaną utworzone po wdrożeniu kostki**
   1. Przychód i liczba zamówień według miesiąca i stanu klienta
   2. Średnia ocena i liczba opinii wg sprzedawcy i kwartału
   3. Rozkład typów płatności wg kategorii produktu i roku
   4. Średni czas dostawy wg regionu sprzedawcy i miesiąca
   5. Top 10 produktów wg przychodu i liczby sztuk w pewnym analizowanym okresie
   6. Średni czas dostawy według gęstości zaludnienia
   7. Liczba zamówień według populacji miasta
   8. Liczba zamówień nowych vs powracających klientów według miast
   9. Przychód według dnia tygodnia i typu płatności
   10. Top 10 najgorszych sprzedawców według średniej ocen i miesiąca
2. **Implementacja bazy danych zgodnie z zaproponowanym konceptualnym modelem danych**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer program

AI-generated content may be incorrect.**

1. **Wnioski**

Zdecydowana większość atrybutów jest kompletna i może z powodzeniem trafić do hurtowni danych – mamy pełne informacje o zamówieniach, klientach, produktach i płatnościach, co pozwala na zbudowanie rozbudowanych wymiarów czasowego, geograficznego, klienta, sprzedawcy, produktu i płatności. Dane o opiniach są niemal kompletne, choć kilkaset rekordów nie zawiera ocen lub komentarzy, co jednak nie powinno zaburzyć ogólnych trendów. Z kolei geolokalizacje wymagają odfiltrowania kilku procent współrzędnych spoza terytorium Brazylii, ale same kody pocztowe umożliwiają precyzyjne grupowanie według stanów i miast.

Wartości numeryczne – takie jak cena, liczba rat czy wymiary produktów – mieszczą się w sensownych zakresach i nie zawierają błędnych skrajnych wartości, co czyni je gotowymi do agregacji i obliczeń KPI. Potencjalne outliery w wadze lub wymiarach można prawdopodobnie wyfiltrować. Pola tekstowe (kategorie, nazwy miast) wymagają jedynie podstawowej normalizacji (usunięcie literówek, standaryzacja akcentów), by zapobiec duplikacji wymiaru.

W wymiarach znajdziemy naprawdę sporo do badania: czasowego (analiza sezonowości i trendów), przestrzennego (różnice między regionami), produktowego (popularność i marże w kategoriach) oraz behawioralnego (liczba rat czy typ płatności jako wskaźniki preferencji klientów). Faktowe miary – przychód, liczba zamówień, średnia ocena czy czas dostawy – pozwolą na wielowymiarowe zestawienia i dogłębne analizy jakości obsługi. Dzięki temu hurtownia stanie się solidnym fundamentem dla raportów sprzedażowych, monitoringu satysfakcji klientów oraz optymalizacji procesów logistycznych i marketingowych.

**Etap 2 – 26.05.2025 r.**

**1. SC\_CreateStageTables**  
Tworzy schemat Stage oraz wszystkie tymczasowe tabele, do których będą wczytywane surowe pliki CSV i później przechowywane dane oczyszczone:

* CreateStageSchema
  + jeżeli nie istnieje tworzy schema Stage
* CreateStageOrders
  + tworzy Stage.Orders z olist\_orders.csv
* **CreateStageOrderItems**
  + tworzy Stage.OrderItems z olist\_order\_items.csv
* **CreateStagePayments**
  + tworzy Stage.Payments z olist\_order\_payments.csv
* **CreateStageReviews**
  + tworzy Stage.Reviews z olist\_order\_reviews.csv
* **CreateStageCustomers**
  + tworzy Stage.Customers z olist\_customers.csv
* **CreateStageSellers**
  + tworzy Stage.Sellers z olist\_sellers.csv
* **CreateStageProducts**
  + tworzy Stage.Products z olist\_products.csv
* **CreateStageProductCategoryNameTranslated**
  + tworzy Stage.ProductCategoryNameTranslation z tłumaczeniem nazw kategorii
* **CreateStageCities**
  + tworzy Stage.Cities z danymi brazylijskich miast
* **CreateStageOrdersClean**
  + tworzy tabelę Stage.OrdersClean na wyniki oczyszczania dat i miar.
* **CreateStageCustomersClean** 
  + tworzy Stage.CustomersClean na wzbogacone dane klientów o dane miast
* **CreateStageSellersClean**
  + tworzy Stage.SellersClean na wzbogacone dane sprzedawców o dane miast
* **CreateStageProductsClean**
  + tworzy Stage.ProductsClean na wzbogacone dane produktów

**2. SC\_LoadStageData**  
Wczytuje dane z plików CSV do tabel Stage.[…] za pomocą zadań Data Flow:

* DFT\_LoadOrdersStage
* DFT\_LoadOrderItemsStage
* DFT\_LoadPaymentsStage
* DFT\_LoadReviewsStage
* DFT\_LoadCustomersStage
* DFT\_LoadSellersStage
* DFT\_LoadProductsStage
* DFT\_LoadCitiesStage
* DFT\_LoadProductCategoryNameTranslationStage

**3. SC\_CleanStage**  
Oczyszcza i wzbogaca dane:

* DFT\_OrdersClean
  + konwersja dat na DATETIME, obliczenie czasu dostawy - delivery\_time (w dniach), zapis do Stage.OrdersClean
* DFT\_CustomersClean
  + fuzzy lookup miast, dodanie populacji, powierzchni, gęstości zaludnienia, zapis do Stage.CustomersClean
* DFT\_SellersClean
  + analogiczne wzbogacenie danych sprzedawców, zapis do Stage.SellersClean
* DFT\_ProductsClean
  + tłumaczenie kategorii, konwersja zmiennych, zapis do Stage.ProductsClean

**4. SC\_CreateFinalSchemaTables**  
Tworzy schemat o nazwie „Stepaniuk” oraz wszystkie tabele docelowe hurtowni:

* CreateSchema
  + tworzy schemat Stepaniuk jeśli nie istnieje.
* CreateMonthDim
  + Tworzy Stepaniuk.MonthDim z 12 wierszami
* CreateWeekdayDim
  + Tworzy Stepaniuk.WeekdayDim z 7 wierszami
* CreateTimeDim
  + Tworzy Stepaniuk.TimeDim (kluczem czas + dodatkowe atrybuty do dni tygodnia, miesiąca, roku itd.)
* CreateCustomerDim
  + Tworzy Stepaniuk.CustomerDim (klient + dane miast)
* CreateSellerDim
  + Tworzy Stepaniuk.SellerDim (sprzedawca + dane miast)
* CreateProductDim
  + Tworzy Stepaniuk.ProductDim
* CreatePaymentDim
  + Tworzy Stepaniuk.PaymentDim (unikalne metody płatności)
* CreateReviewDim
  + Tworzy Stepaniuk.ReviewDim
* CreateFactOrders
  + Tworzy Stepaniuk.FactOrders (na poziomie pojedynczych produktów order\_item)

**5. SC\_LoadFinalData**  
Ładuje dane do wymiarów i faktów, stosując przyrostowe upserty (insert where not exists) przy pomocy MERGE:

* LoadMonthDim
* LoadWeekdayDim
* UpsertTimeDim
* UpsertCustomerDim
* UpsertSellerDim
* UpsertProductDim
* UpsertReviewDim
* UpsertPaymentDim
* UpsertFactOrders

**6. SC\_Validation**  
Weryfikuje poprawność załadowanych danych i w razie błędów Raisuje błąd.

* CheckDimensionRowCountMismatch
* CheckFactRowCountMismatch

A screenshot of a computer

AI-generated content may be incorrect.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Lp. | Źródłowy plik | Źródłowa kolumna | Docelowa kolumna | Typ danych |
| 1 | **olist\_orders\_dataset.csv** | order\_purchase\_timestamp | TimeDim.full\_datetime | DATETIME |
| 2 | **olist\_orders\_dataset.csv** | order\_purchase\_timestamp | TimeDim.time\_key | BIGINT |
| 3 | **olist\_orders\_dataset.csv** | order\_purchase\_timestamp | TimeDim.year\_n | SMALLINT |
| 4 | **olist\_orders\_dataset.csv** | order\_purchase\_timestamp | TimeDim.quarter\_n | SMALLINT |
| 5 | **olist\_orders\_dataset.csv** | order\_purchase\_timestamp | TimeDim.month\_key | SMALLINT |
| 6 | **olist\_orders\_dataset.csv** | order\_purchase\_timestamp | TimeDim.day\_n | SMALLINT |
| 7 | **olist\_orders\_dataset.csv** | order\_purchase\_timestamp | TimeDim.weekday\_key | SMALLINT |
| 8 | **olist\_orders\_dataset.csv** | order\_purchase\_timestamp | TimeDim.hour\_n | SMALLINT |
| 9 | **olist\_orders\_dataset.csv** | order\_purchase\_timestamp | TimeDim.minute\_n | SMALLINT |
| 10 | **olist\_orders\_dataset.csv** | order\_purchase\_timestamp | TimeDim.second\_n | SMALLINT |
| 11 | **olist\_customers\_dataset.csv** | customer\_id | CustomerDim.customer\_id | VARCHAR(50) |
| 12 | **olist\_customers\_dataset.csv** | customer\_state | CustomerDim.customer\_state | CHAR(2) |
| 13 | **olist\_customers\_dataset.csv** | customer\_city | CustomerDim.customer\_city | VARCHAR(100) |
| 14 | **brazilian\_cities.csv** | IBGE\_RES\_POP | CustomerDim.city\_population | INT |
| 15 | **brazilian\_cities.csv** | AREA | CustomerDim.city\_area\_km2 | DECIMAL(10,2) |
| 16 | **brazilian\_cities.csv** | (computed) population/area | CustomerDim.city\_density | DECIMAL(10,2) |
| 17 | **olist\_sellers\_dataset.csv** | seller\_id | SellerDim.seller\_id | VARCHAR(50) |
| 18 | **olist\_sellers\_dataset.csv** | seller\_state | SellerDim.seller\_state | CHAR(2) |
| 19 | **olist\_sellers\_dataset.csv** | seller\_city | SellerDim.seller\_city | VARCHAR(100) |
| 20 | **brazilian\_cities.csv** | IBGE\_RES\_POP | SellerDim.city\_population | INT |
| 21 | **brazilian\_cities.csv** | AREA | SellerDim.city\_area\_km2 | DECIMAL(10,2) |
| 22 | **brazilian\_cities.csv** | (computed) population/area | SellerDim.city\_density | DECIMAL(10,2) |
| 23 | **olist\_products\_dataset.csv** | product\_id | ProductDim.product\_id | VARCHAR(50) |
| 24 | **olist\_products\_dataset.csv** | product\_category\_name + translation | ProductDim.category | VARCHAR(100) |
| 25 | **olist\_products\_dataset.csv** | (z CSV tłumaczeń) product\_category\_name\_english | ProductDim.category | VARCHAR(100) |
| 26 | **olist\_products\_dataset.csv** | product\_category\_name (podkategoria) | ProductDim.sub\_category | VARCHAR(100) |
| 27 | **olist\_order\_payments\_dataset.csv** | payment\_type | PaymentDim.payment\_type | VARCHAR(50) |
| 28 | **olist\_order\_payments\_dataset.csv** | payment\_type | PaymentDim.payment\_type\_key | INT (surrogate) |
| 29 | **olist\_order\_reviews\_dataset.csv** | review\_id | ReviewDim.review\_id | VARCHAR(50) |
| 30 | **olist\_order\_reviews\_dataset.csv** | review\_score | ReviewDim.review\_score | SMALLINT |
| 31 | **olist\_order\_reviews\_dataset.csv** | review\_comment\_message | ReviewDim.review\_comment | TEXT |
| 32 | **olist\_order\_reviews\_dataset.csv** | review\_creation\_date | ReviewDim.review\_date | DATE |
| 33 | **olist\_order\_items\_dataset.csv** | order\_item\_id | FactOrders.order\_item\_id | VARCHAR(50) |
| 34 | **olist\_order\_items\_dataset.csv** | order\_id | FactOrders.order\_id | VARCHAR(50) |
| 35 | **olist\_orders\_clean (Stage.OrdersClean)** | order\_purchase\_timestamp | FactOrders.average\_delivery\_time | DECIMAL(10,2) |
| 36 | **olist\_order\_items\_dataset.csv** + freight | price + freight\_value | FactOrders.total\_revenue | DECIMAL(18,2) |
| 37 | **olist\_order\_items\_dataset.csv** | order\_item\_id | FactOrders.total\_items | INT |
| 38 | **olist\_order\_reviews\_dataset.csv** | review\_score | FactOrders.average\_review\_score | DECIMAL(3,2) |
| 39 | **olist\_order\_payments\_dataset.csv** (p.seq=1) | payment\_sequential | FactOrders.payment\_sequential | SMALLINT |
| 40 | **olist\_order\_payments\_dataset.csv** (p.seq=1) | payment\_installments | FactOrders.payment\_installments | SMALLINT |
| 41 | **olist\_order\_payments\_dataset.csv** (p.seq=1) | payment\_value | FactOrders.payment\_value | DECIMAL(18,2) |
| 42 | **Stage.OrdersClean** | time\_key | FactOrders.time\_key | BIGINT |
| 43 | **Stage.OrdersClean** | customer\_id | FactOrders.customer\_id | VARCHAR(50) |
| 44 | **Stage.OrderItems** | seller\_id | FactOrders.seller\_id | VARCHAR(50) |
| 45 | **Stage.OrderItems** | product\_id | FactOrders.product\_id | VARCHAR(50) |
| 46 | **Stepaniuk.PaymentDim** | payment\_type\_key | FactOrders.payment\_type\_key | INT |

**Etap 3 – 02.06.2025 r.**

Kostka:

1. TimeDim – wymiar czasu

* Rok, kwartał, miesiąc, dzień tygodnia, godzina

1. PaymentDim

* Typ płatności

1. ProductDim

* Kategoria produktu

1. SellerDim

* Miasto, stan, region, gęstość zaludnienia miasta, powierzchnia, populacja

1. CustomerDim
   * Miasto, stan, region, gęstość zaludnienia miasta, powierzchnia, populacja, tag klienta mówiący czy jest stałym czy powracającym klientem
2. ReviewDim
   * Gwiazdkowa ocena zamówienia
3. FactOrders

* payment\_value – wartość całego zamówienia
* total\_revenue – wartość przedmiotu + koszty przesyłki
* total\_items – ilość przedmiotów
* review\_score – ilość gwiazdek zostawiona w ocenie zamówienia od 0 do 5
* delivery\_time – liczone w dniach od momentu zamówienia do dostawy
* average\_review\_score – miara kalkulowana licząca średnią ilość gwiazdek
* average\_delivery\_time – miara kalkulowana do liczenia średniego czasu dostawy
* maximum\_payment\_value – miara kalkulowana, maksymalna wartość całego zamówienia

Zestawienia:

1. A graph of different colored lines

   AI-generated content may be incorrect.Procentowy przychód według miesiąca i regionu klienta

Widać większą sprzedaż w miesiącach od stycznia do sierpnia, niezależnie od obszaru (na co nie ma wpływu ograniczony zbiór danych, ponieważ dane zaczynają się od października 2016, a trwają do września 2018.

1. A graph of different colored bars

   AI-generated content may be incorrect.Średni czas dostawy w kolejnych kwartałach kolejnych lat w różnych regionach

Widać że w regionie północnym i północno-wschodnim czas dostawy jest znacznie wyższy od pozostałych regionów, niezależnie od kwartału i roku są to zawsze regiony z najdłuższym czasem dostawy (około tydzień dłużej niż w innych regionach) - może to być spowodowane mniejszym zaludnieniem tego regionu oraz mniejszą sprzedażą w nim. Z drugiej strony region południowo-wschodni gdzie znajdują się największe miasta brazylii takie jak Brasilia czy Sao Paulo mają zdecydowanie najkrótszy czas dostawy. Natomiast widać także, że wraz z biegiem czasu ogólny czas dostawy ma tendencje spadkową.

1. A graph with different colored bars

   AI-generated content may be incorrect.Maksymalne kwoty zakupów w kolejnych kwartałach kolejnych lat w różnych regionach

Widać, że największe kwoty padają w regionie południowo-wschodnim, co może być spowodowane większą zamożnością tego regionu jak i większą jego populacją (konsekwencją czego mają więcej różnych zamówień stamtąd i jest większa szansa na jakieś „duże” zamówienie) – rekord płatności padł w trzecim kwartale 2017 roku w regionie południowo wschodnim i wynosił blisko 14000 R$ (reali brazylijskich) co w przybliżeniu wynosi około 9163,63zł.

1. A graph of different colored bars

   AI-generated content may be incorrect.Średni czas dostawy w kolejnych kwartałach kolejnych lat dla różnych metod płatności

Widać, że czas dostawy spada wraz z biegiem czasu, jest też zbliżony do siebie niezależnie od metody płatności, z lekkim wskazaniem na większe wartości w przypadku boleto (co w wolnym tłumaczeniu znaczy bilet) – możliwe że zaksięgowanie tej formy płatności zajmuje czas, przez co wydłuża się łączna ilość dni od momentu dokonania płatności do otrzymania przesyłki, jednorazowy rekord padł dla vouchera w 2016 roku, ale może to być outlier ponieważ w tym okresie była to jedyna płatność tego typu.

1. A graph of blue bars

   AI-generated content may be incorrect.Średni czas dostawy w kolejnych miesiącach dla miast o różnych powierzchniach

Widać, że dla miast o największej i najmniejszej gęstości zaludnienia czas dostawy jest najszybszy, natomiast dla miast o średniej gęstości zaludnienia widać wzrost w czasie dostawy. Dodatkowo najmniejszy czas dostawy jest w miesiącach letnich, gdzie nawet te największe wartości zazwyczaj mieszczą czas dostawy poniżej dwóch tygodni.

1. A graph of blue and orange bars

   AI-generated content may be incorrect.Top 10 kategorii produktów pod względem sprzedaży sztuk i wartości

Widać, że sprzedaż ilościowa kategorii często idzie w parze z ich sprzedażą kapitałową, z wyjątkiem watches\_gifts które to kosztują znacznie więcej niż sprzedawane jest ich ilościowo, widać więc że są to przedmioty o większej wartości, które klienci kupują rzadziej.

1. A graph of blue and orange lines

   AI-generated content may be incorrect.Sprzedaż w poszczególnych godzinach dnia dla dni tygodnia pracującego oraz weekendu

Widać, że godziny szczytu sprzedaży rozpoczynają się około godziny 9:00 rano w przypadku dni od poniedziałku do piątku, natomiast w weekendy widać nieco większy pik w godzinach wieczornych, około godziny 20:00 widać znaczący pik, którego nie ma w tygodniu, może to być spowodowane większą ilością wolnego czasu klientów weekend, przez co decydują się oni robić zamówienia wieczorem.

1. A graph with blue and orange bars

   AI-generated content may be incorrect.Średnia recenzja stałych i powracająch klientów z miast o różnych wielkościach populacji

Widać, że nowi klienci zostawiają raczej tym gorsze noty, im w większym mieście mieszkają, natomiast klienci powracający czyli tacy którzy dokonali więcej niż jednej płatności mają na ogół lepsze opinie o sklepie w mniejszych miastach (do 500.000), a w tych większych trend się odwraca i stali klienci oceniają gorzej swoją satysfakcję z zakupów.

1. A graph with blue and orange lines

   AI-generated content may be incorrect.Procentowy średni czas dostawy oraz średnia recenzja dla poszczególnych stanów

Widać, że niektóre stany mają znacznie zawyżony średni czas dostawy (niektóre nawet do blisko 250% średniego czasu dostawy), natomiast inne (raczej te większe takie jak Sao Paulo albo Minas Gerais) mają znacznie szybszą dostawę od reszty (rekordowa dla Sao Paulo, które jest bardzo zaludnionym regionem, więc czas dostawy jest tutaj w zrozumiały sposób mniejszy – natomiast warto zaznaczyć że nie wpływa to na satysfakcję klientów z obsługi i ich ocena pozostaje raczej w okolicach średniej niezależnie od czasu dostawy.

1. A graph of different colored bars

   AI-generated content may be incorrect.Maksymalne kwoty płatności dla poszczególnych metod płatności w poszczególnych wielkościach miast

Widać, że dla dużych miast przeważa znacząco karta kredytowa, natomiast w miastach mniejszych i średnich nie jest to już tak oczywiste, a często nawet ludzie płacą największe kwoty alternatywnymi metodami płatności takimi jak bilet czy voucher – warto zaznaczyć tutaj że rozmiary kubełków nie są równomierne, więc dodatkowy czynnik taki jak ilość zamówień z miast o danej populacji może mieć wpływ na te wyniki. Wciąż można jednak wysunąć wniosek, że mieszkańcy mniejszy miejscowości nie robią tak dużych kwotowo zamówień, jak ci z miast o populacji większej niż 5mln.

**Wnioski:**

Analizy poprzez powyższe zestawienia pokazują, jak dzięki hurtowni danych i kostce OLAP można szybko identyfikować kluczowe i powtarzalne wzorce sprzedaży (sezonowość, top-kategorie, godziny szczytu), efektywność operacyjną (czasy dostawy w regionach i metodach płatności) oraz zachowania klientów (nowi vs. powracający, preferencje zakupowe wg wielkości miast). Dzięki temu menedżerowie Olist mogą optymalizować logisty­kę, planować kampanie marketingowe w najlepszych okresach, segmentować klientów i priorytetyzować inwestycje w obszary i kanały przynoszące największe przychody.