

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.