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)

2. **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.

3. **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).

4. **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, cena)
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

5. 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

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

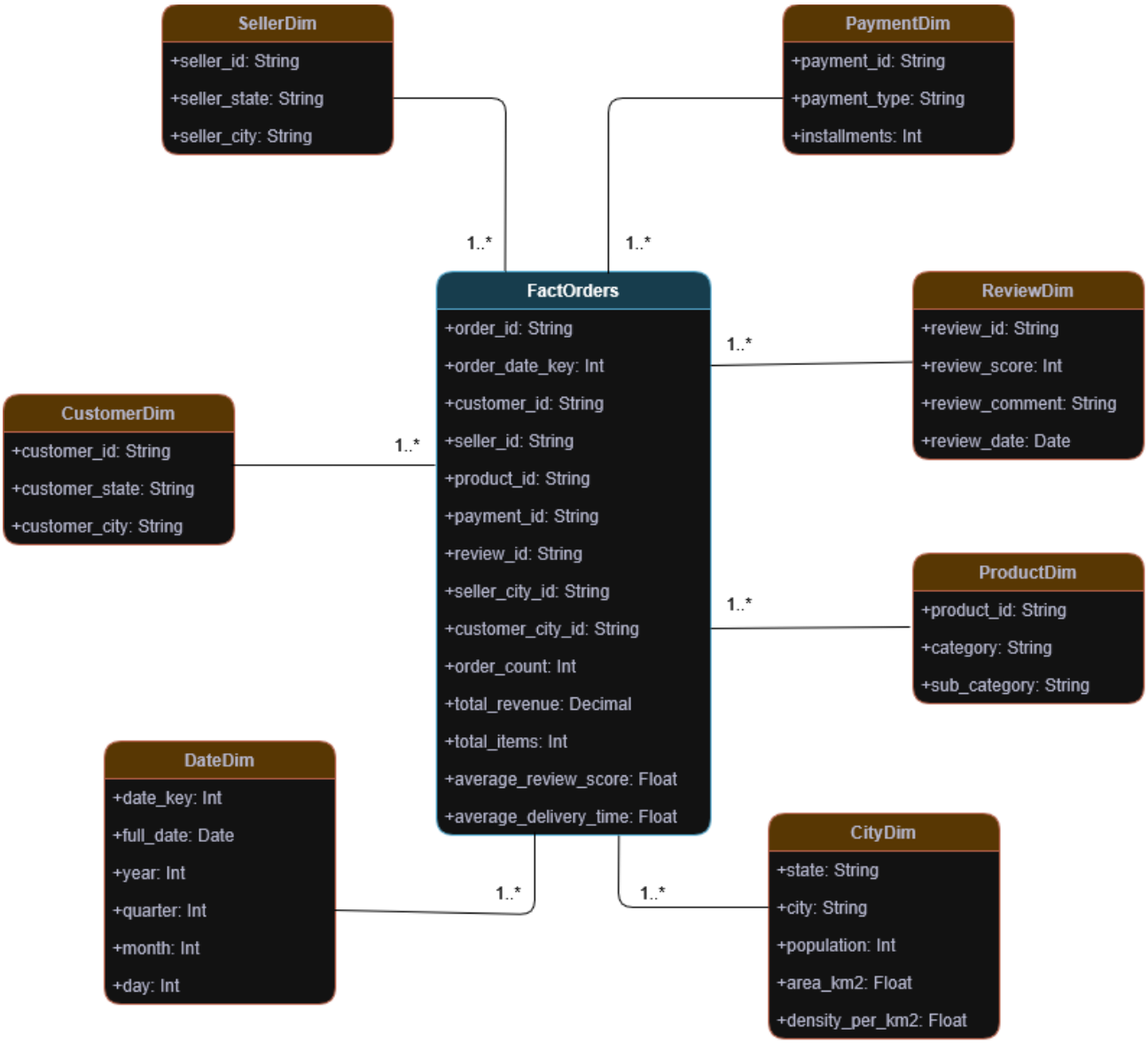
Encje wymiarów:

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:

7. 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:



7. 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

8. Implementacja bazy danych zgodnie z zaproponowanym konceptualnym modelem danych

```sql
-- schemat
CREATE SCHEMA Stepaniuk;

-- tabela pomocnicza dla miesięcy
CREATE TABLE Stepaniuk.MonthDim (
    month_key    SMALLINT    PRIMARY KEY,
    month_name   VARCHAR(20) NOT NULL
);

INSERT INTO Stepaniuk.MonthDim VALUES
(1,'January'),  (2,'February'), (3,'March'),    (4,'April'),
(5,'May'),      (6,'June'),     (7,'July'),     (8,'August'),
(9,'September'),(10,'October'), (11,'November'),(12,'December');

-- wymiar daty
CREATE TABLE Stepaniuk.DateDim (
    date_key      INT          PRIMARY KEY,
    full_date     DATE         NOT NULL,
    year_n        SMALLINT     NOT NULL,
    quarter_n     SMALLINT     NOT NULL,
    month_key     SMALLINT     NOT NULL       REFERENCES Stepaniuk.MonthDim(month_key),
    day_n         SMALLINT     NOT NULL
);

-- wymiar klienta
CREATE TABLE Stepaniuk.CustomerDim (
    customer_id   VARCHAR(50)  PRIMARY KEY,
    customer_state CHAR(2)     NOT NULL,
    customer_city  VARCHAR(100) NOT NULL
);

-- wymiar sprzedawcy
CREATE TABLE Stepaniuk.SellerDim (
    seller_id     VARCHAR(50)  PRIMARY KEY,
    seller_state  CHAR(2)      NOT NULL,
    seller_city   VARCHAR(100) NOT NULL
);
```

```sql
-- wymiar produktu
CREATE TABLE Stepaniuk.ProductDim (
    product_id     VARCHAR(50)  PRIMARY KEY,
    category       VARCHAR(100) NOT NULL,
    sub_category   VARCHAR(100)
);

-- wymiar płatności
CREATE TABLE Stepaniuk.PaymentDim (
    payment_id     VARCHAR(50)  PRIMARY KEY,
    payment_type   VARCHAR(30)  NOT NULL,
    installments   SMALLINT     NOT NULL
);

-- wymiar opinii
CREATE TABLE Stepaniuk.ReviewDim (
    review_id      VARCHAR(50) PRIMARY KEY,
    review_score   SMALLINT    NOT NULL,
    review_comment TEXT,
    review_date    DATE        NOT NULL
);

-- tabela faktów
CREATE TABLE Stepaniuk.FactOrders (
    order_id              VARCHAR(50)  PRIMARY KEY,
    order_date_key        INT          NOT NULL       REFERENCES Stepaniuk.DateDim(date_key),
    customer_id           VARCHAR(50)  NOT NULL       REFERENCES Stepaniuk.CustomerDim(customer_id),
    seller_id             VARCHAR(50)  NOT NULL       REFERENCES Stepaniuk.SellerDim(seller_id),
    product_id            VARCHAR(50)  NOT NULL       REFERENCES Stepaniuk.ProductDim(product_id),
    payment_id            VARCHAR(50)  NOT NULL       REFERENCES Stepaniuk.PaymentDim(payment_id),
    review_id             VARCHAR(50)                 REFERENCES Stepaniuk.ReviewDim(review_id),
    order_count           INT          NOT NULL DEFAULT 1,
    total_revenue         DECIMAL(18,2)NOT NULL,
    total_items           INT          NOT NULL,
    average_review_score  DECIMAL(3,2),
    average_delivery_time DECIMAL(10,2) -- dni
);
```

9. **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

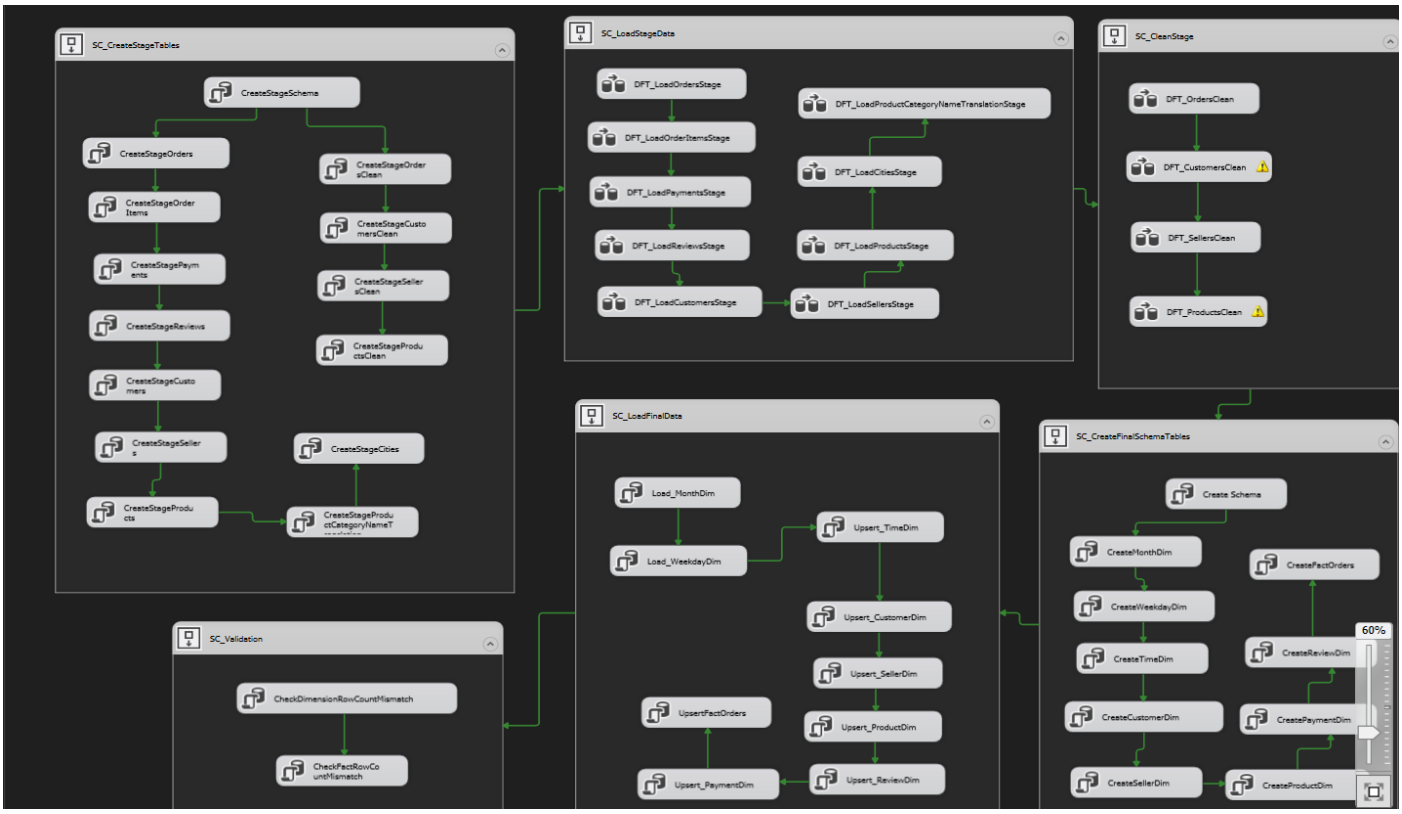
Ładuje dane do wymiarów i faktów, stosując przyrostowe upserty (insert where not exists) przy pomocy MERGE:

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

6. SC_Validation

Weryfikuje poprawność załadowanych danych i w razie błędów Raisuje błąd.

- CheckDimensionRowCountMismatch
- CheckFactRowCountMismatch



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

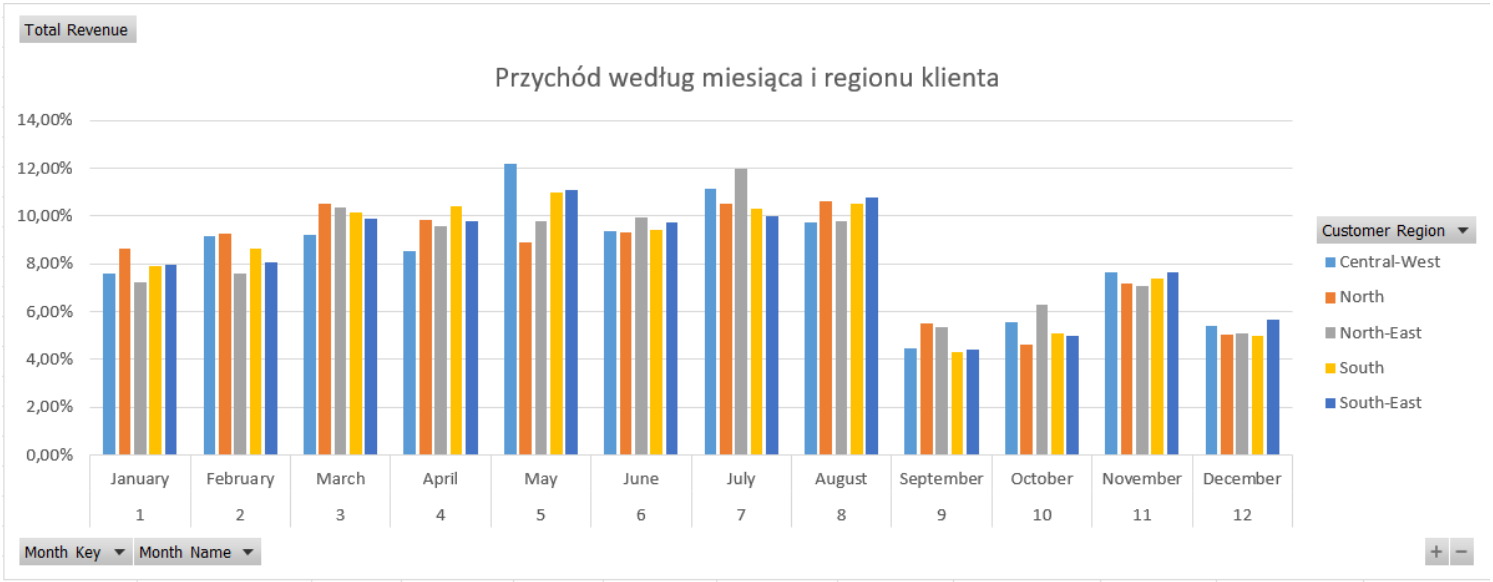
Etap 3 – 02.06.2025 r.

Kostka:

1. TimeDim – wymiar czasu
    - Rok, kwartał, miesiąc, dzień tygodnia, godzina
2. PaymentDim
    - Typ płatności
3. ProductDim
    - Kategoria produktu
4. SellerDim
    - Miasto, stan, region, gęstość zaludnienia miasta, powierzchnia, populacja
5. CustomerDim
    - Miasto, stan, region, gęstość zaludnienia miasta, powierzchnia, populacja, tag klienta mówiący czy jest stałym czy powracającym klientem
6. ReviewDim
    - Gwiazdkowa ocena zamówienia
7. 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

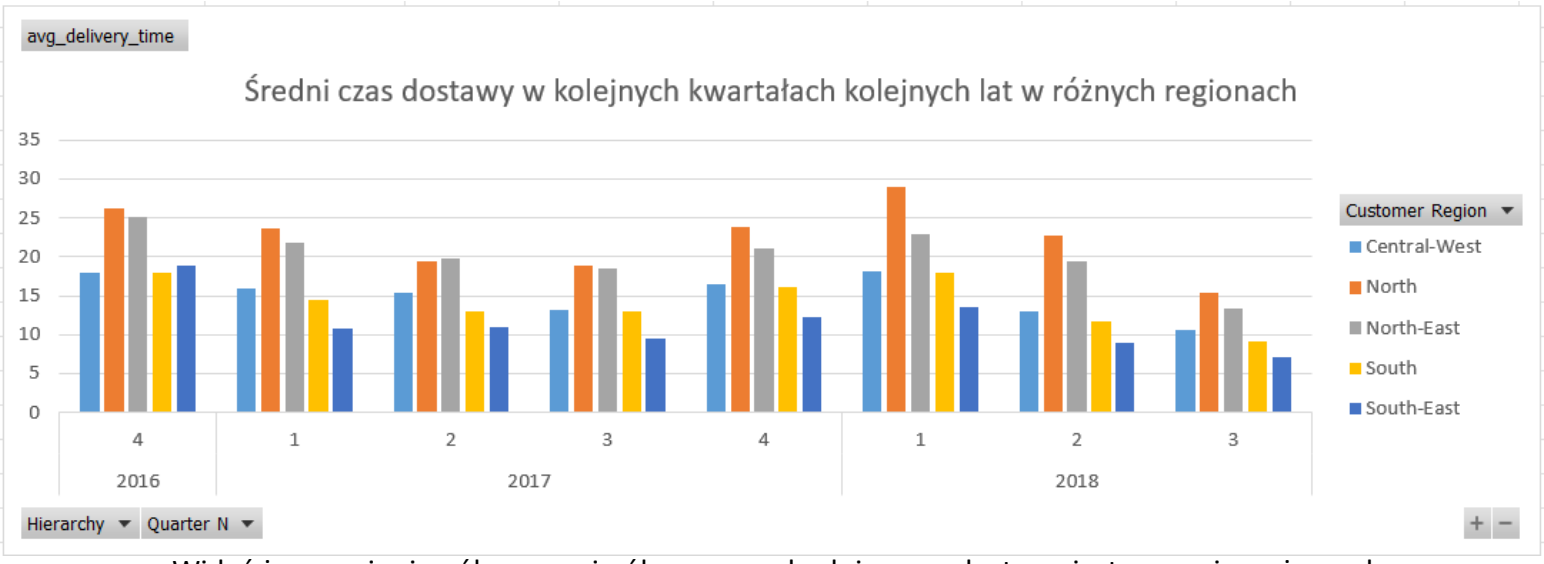
Zestawienia:

1. Procentowy przychód według miesiąca i regionu klienta



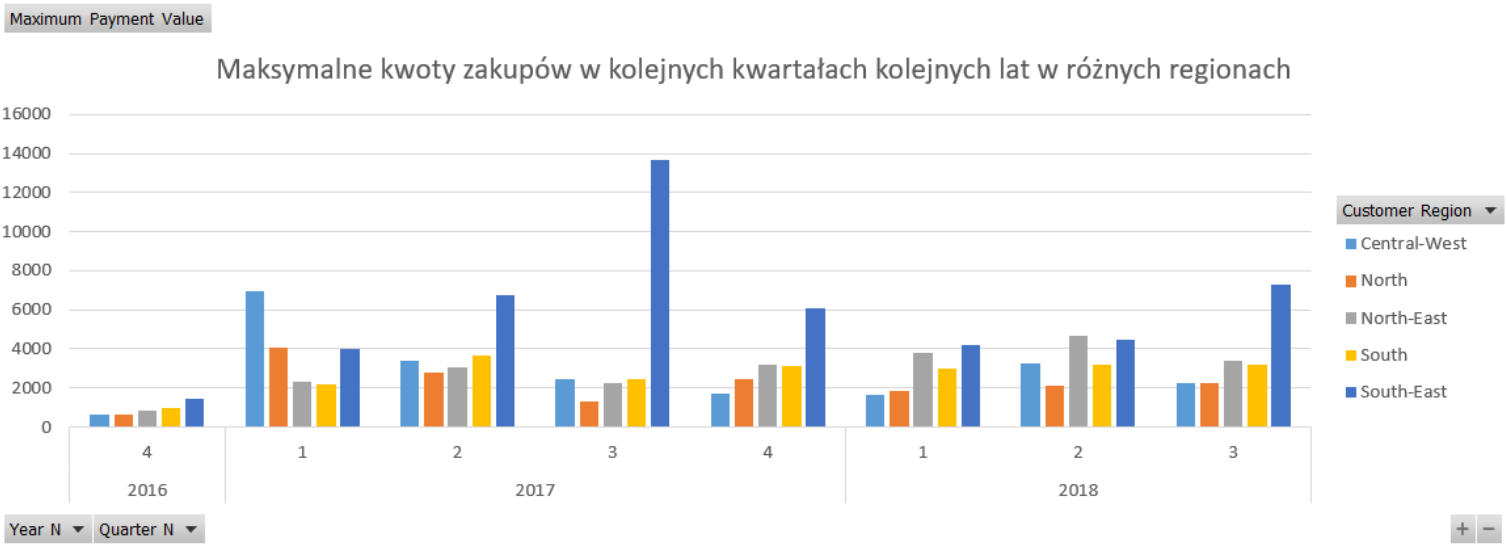
Widać większą sprzedaż w miesiącach od stycznia do sierpnia, niezależnie od obszaru (na co nie ma wpływu ograniczony zbiór danych, ponieważ dane zaczynają się od października 2016, a trwają do września 2018.

2. Średni czas dostawy w kolejnych kwartałach kolejnych lat w różnych regionach



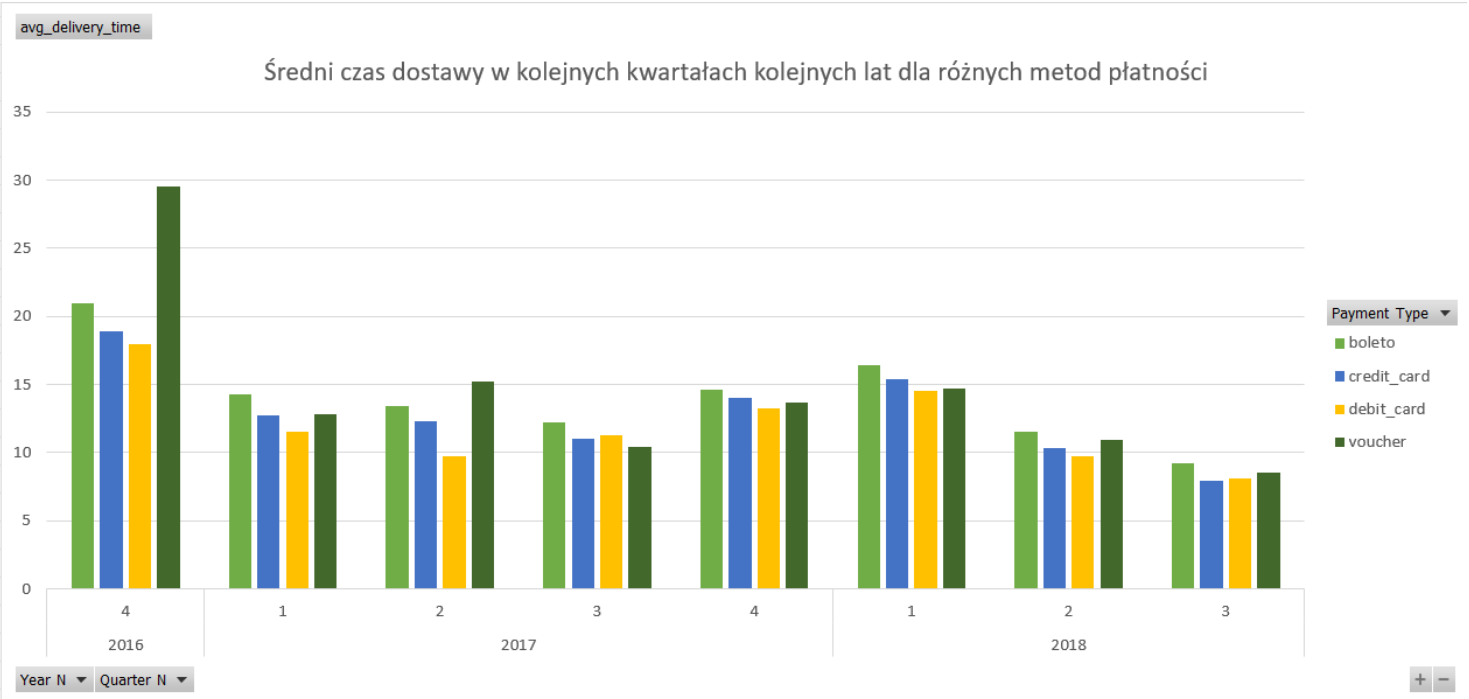
Widać że w regionie północnym i północno-wschodnim czas dostawy jest znacznie wyższy od pozostałych regionów, niezależnie od kwartału i roku są to zawsze regiony z najdłuższym czasem dostawy (około tydzień dłużej niż w innych regionach) - może to być spowodowane mniejszym zaludnieniem tego regionu oraz 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.

3. Maksymalne kwoty zakupów w kolejnych kwartałach kolejnych lat w różnych regionach



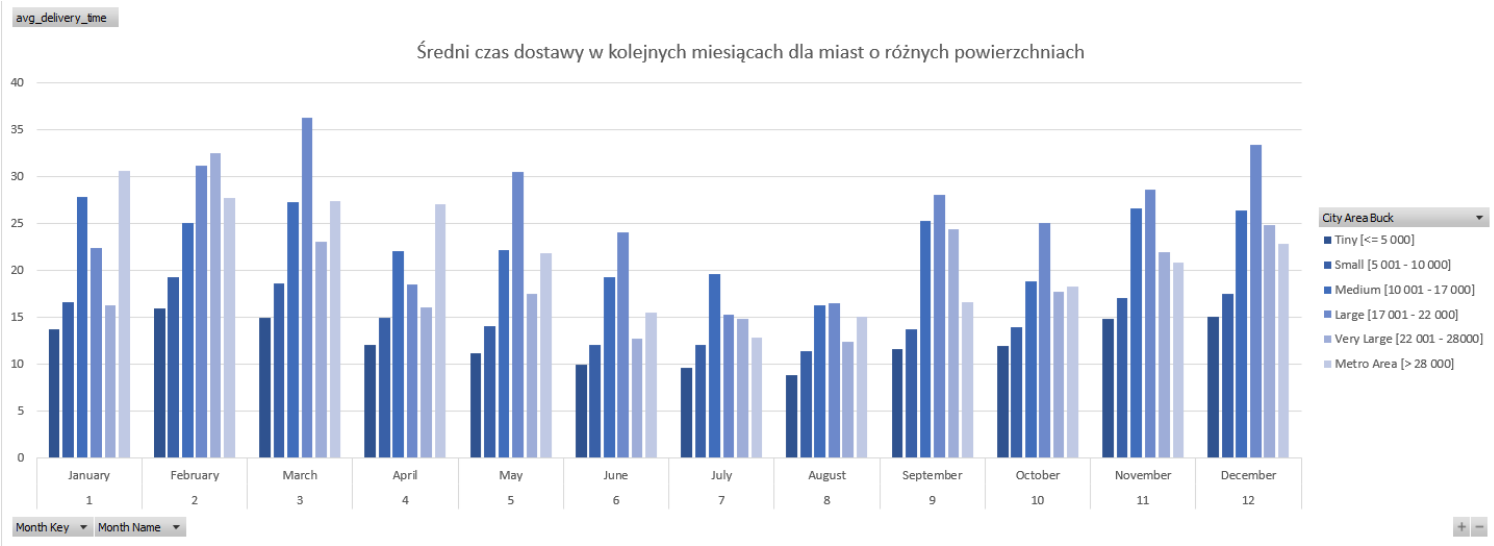
Widać, że największe kwoty padają w regionie południowo-wschodnim, co może być spowodowane większą zamożnością tego regionu jak i większą jego populacją (konsekwencją czego mają więcej różnych zamówień stamtąd i jest większa szansa na jakieś „duże” zamówienie) – rekord płatności padł w trzecim kwartale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

4. Średni czas dostawy w kolejnych kwartałach kolejnych lat dla różnych metod płatności



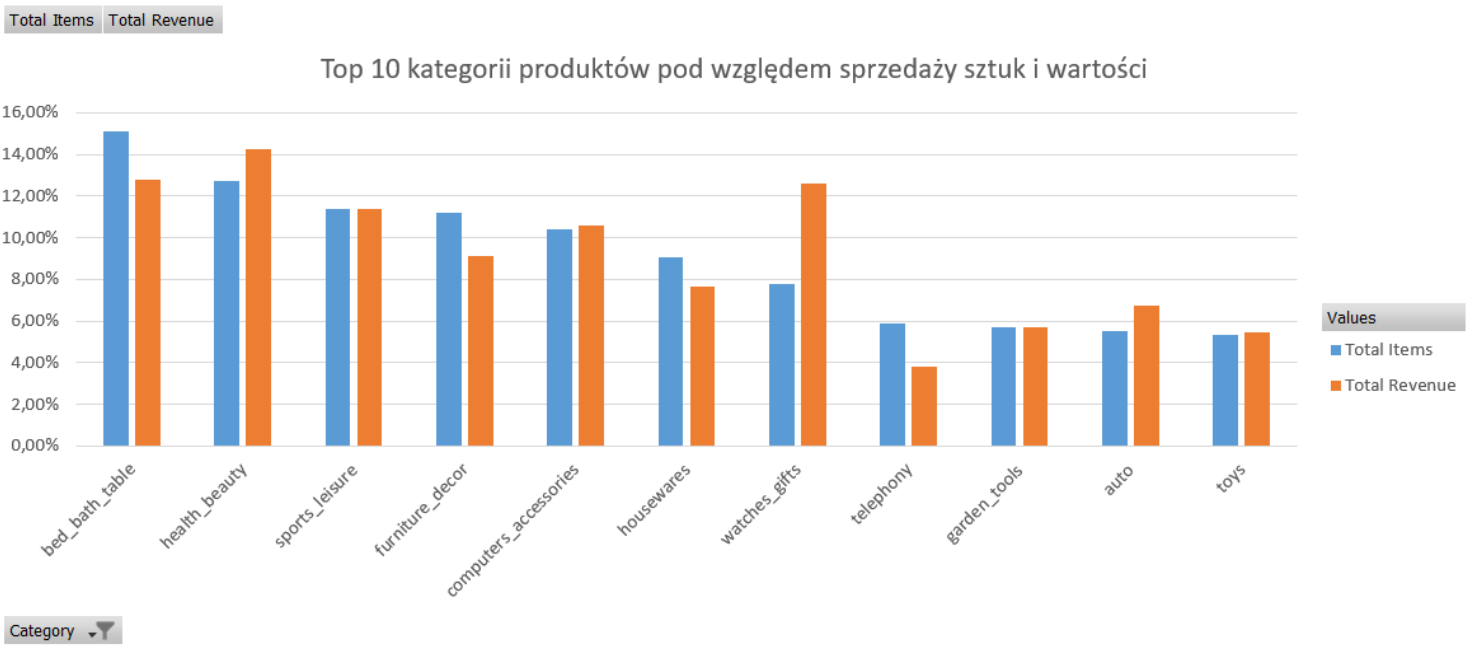
Widać, że czas dostawy spada wraz z biegiem czasu, jest też zbliżony do siebie niezależnie od metody płatności, z lekkim wskazaniem na większe wartości w przypadku boleto (co w wolnym tłumaczeniu znaczy bilet) – możliwe że zaksięgowanie tej formy płatności zajmuje czas, przez co wydłuża się łączna ilość 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.

5. Średni czas dostawy w kolejnych miesiącach dla miast o różnych powierzchniach



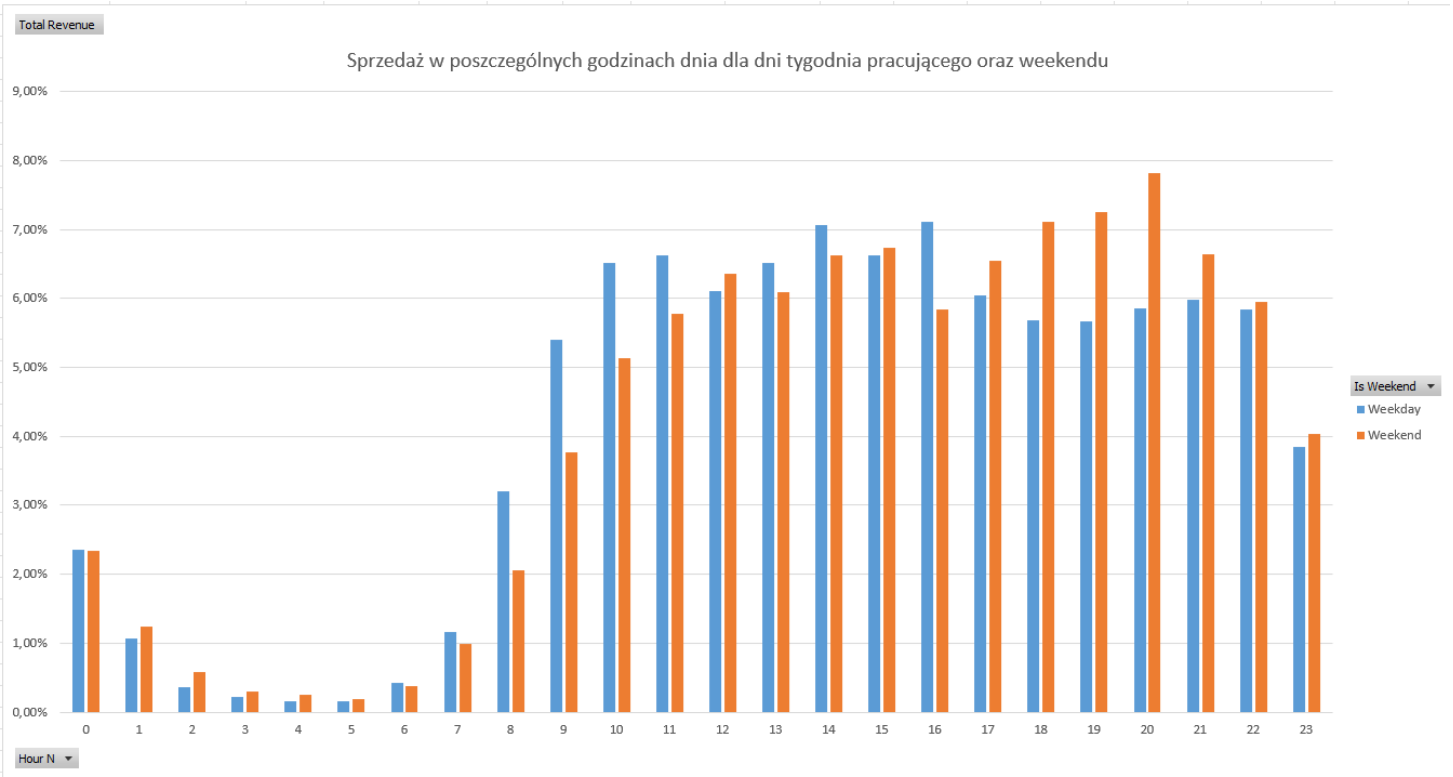
Widać, że dla miast o największej i najmniejszej gęstości zaludnienia czas dostawy jest najszybszy, natomiast dla miast o średniej gęstości zaludnienia widać wzrost w czasie dostawy. Dodatkowo najmniejszy czas dostawy jest w miesiącach letnich, gdzie nawet te największe wartości zazwyczaj mieszczą czas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

6. Top 10 kategorii produktów pod względem sprzedaży sztuk i wartości



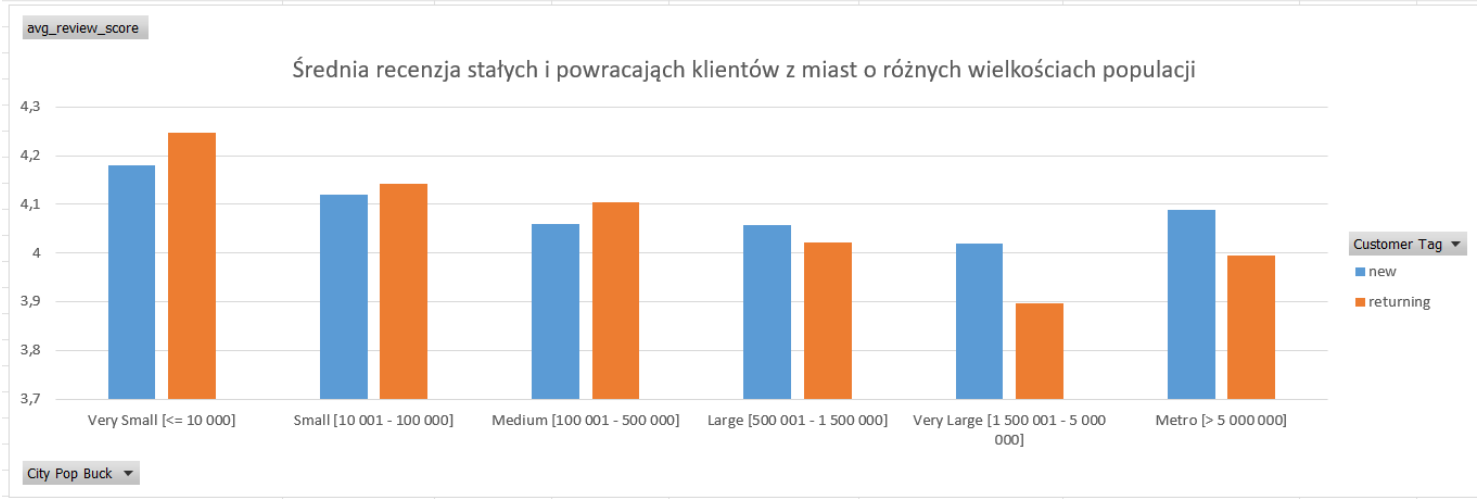
Widać, że sprzedaż ilościowa kategorii często idzie w parze z ich sprzedażą kapitałową, z wyjątkiem watches_gifts które to kosztują znacznie więcej niż sprzedawane jest ich ilościowo, widać więc że są to przedmioty o większej wartości, które klienci kupują rzadziej.

7. Sprzedaż w poszczególnych godzinach dnia dla dni tygodnia pracującego oraz weekendu



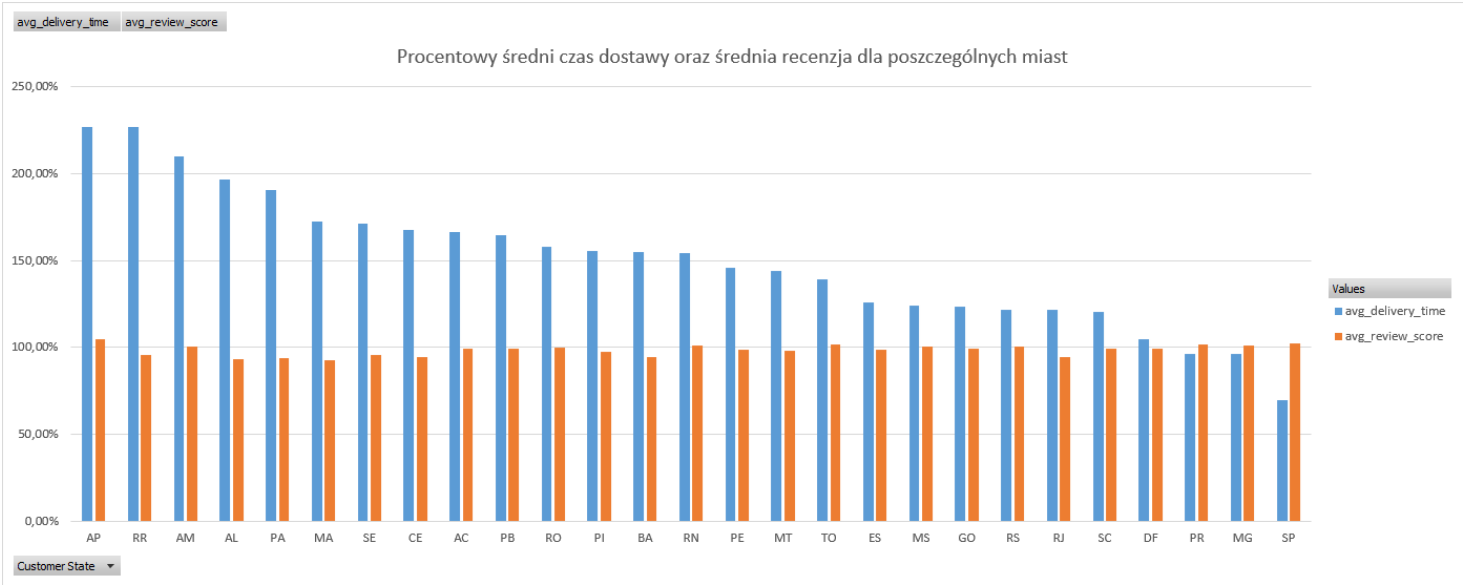
Widać, że godziny szczytu sprzedaży rozpoczynają się około godziny 9:00 rano w przypadku dni od poniedziałku do piątku, natomiast w weekendy widać nieco większy pik w godzinach wieczornych, około godziny 20:00 widać znaczący pik, którego nie ma w tygodniu, może to być spowodowane większą ilością wolnego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

8. Średnia recenzja stałych i powracających klientów z miast o różnych wielkościach populacji



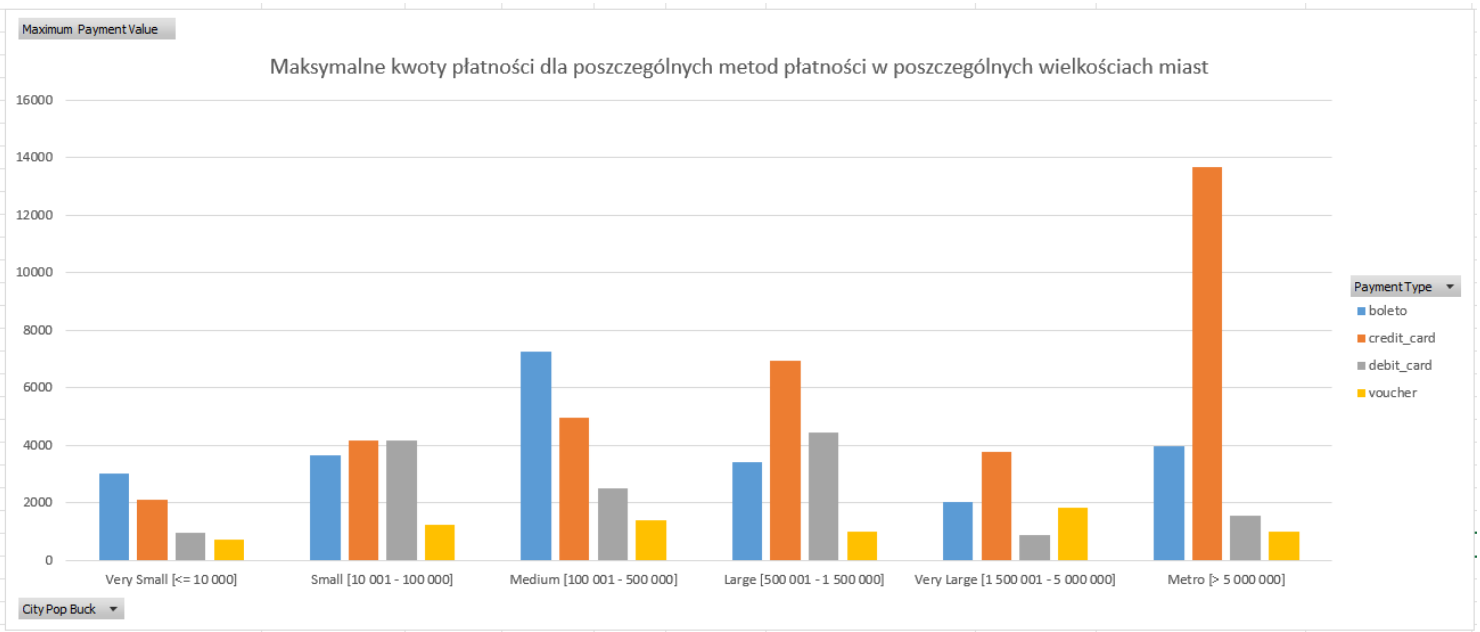
Widać, że nowi klienci zostawiają raczej tym gorsze noty, im w większym mieście mieszkają, natomiast klienci powracający czyli tacy którzy dokonali więcej niż jednej płatności mają na ogół lepsze opinie o sklepie w mniejszych miastach (do 500.000), a w tych większych trend się odwraca i stali klienci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

9. 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.

10. 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.