



---

# OLIST ECOMMERCE DATA ANALYSIS

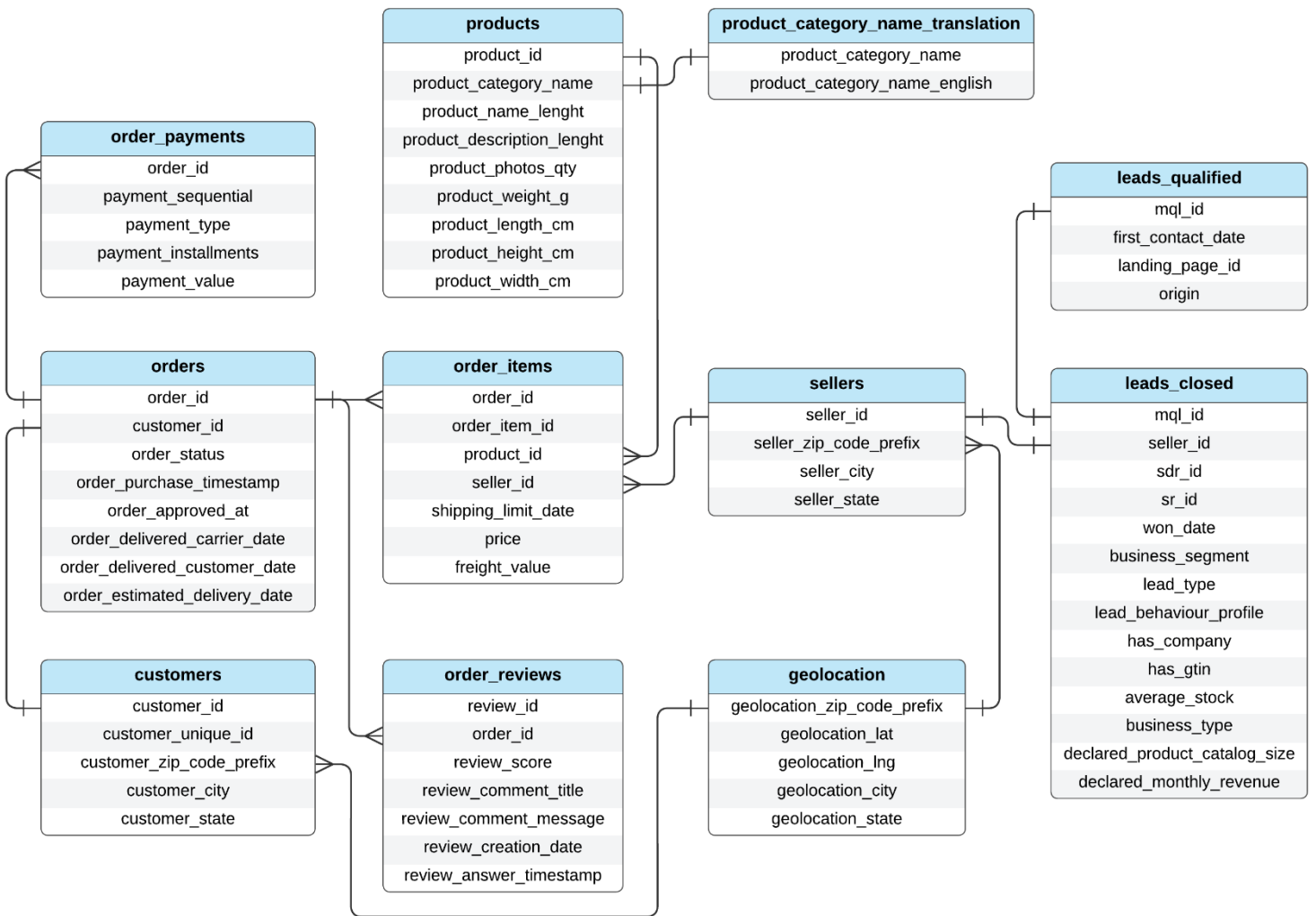
---

Olist ecommerce is a Brazilian Ecommerce Business Organization.



# Database Schema

## DATABASE SCHEMA



## Customer Analysis

1. What are the top 10 cities with the highest number of customers?

```
SELECT customer_city, count_of_customer FROM  
(  
SELECT customer_city, COUNT(DISTINCT customer_unique_id) AS count_of_customer,  
DENSE_RANK()OVER(ORDER BY COUNT(customer_id) DESC) AS rnk  
FROM olist_ecommerce.customers  
GROUP BY 1) customer_count  
WHERE rnk <= 10;
```

	customer_city	count_of_customer	
1	sao paulo	14984	
2	rio de janeiro	6620	
3	belo horizonte	2672	
4	brasilvia	2069	
5	curitiba	1465	
6	campinas	1398	
7	porto alegre	1326	
8	salvador	1209	
9	guarulhos	1153	
10	sao bernardo d...	908	

2. What are the top 10 states with the highest number of customers?

```
SELECT customer_state, count_of_customer FROM
(
SELECT customer_state, COUNT(DISTINCT customer_unique_id) AS count_of_customer,
DENSE_RANK()OVER(ORDER BY COUNT(customer_id) DESC) AS rnk
FROM olist_ecommerce.customers
GROUP BY 1
) customer_count
WHERE rnk <= 10;
```

	customer_state	count_of_customer
1	SP	40302
2	RJ	12384
3	MG	11259
4	RS	5277
5	PR	4882
6	SC	3534
7	BA	3277
8	DF	2075
9	ES	1964
10	GO	1952

### 3. What is the top city with highest number of customers in each state

```
SELECT customer_state, customer_city, highest_customer_count_by_city
FROM
(
SELECT customer_state, customer_city,
COUNT(DISTINCT customer_unique_id) AS highest_customer_count_by_city,
DENSE_RANK()OVER(PARTITION BY customer_state ORDER BY COUNT(DISTINCT customer_unique_id) DESC) AS rnk
FROM olist_ecommerce.customers
GROUP BY 1,2) customer_count
WHERE rnk = 1;
```

	customer_state	customer_city	highest_customer_count_by_city
1	AC	rio branco	66
2	AL	maceio	243
3	AM	manaus	135
4	AP	macapa	53
5	BA	salvador	1209
6	CE	fortaleza	643
7	DF	brasilia	2069
8	ES	vitoria	366
9	GO	goiania	671
10	MA	sao luis	341
11	MG	belo horizonte	2672
12	MS	campo grande	315
13	MT	cuiaba	242
14	PA	belem	432
15	PB	joao pessoa	241
16	PE	recife	590
17	PI	teresina	271
18	PR	curitiba	1465
19	RJ	rio de janeiro	6620

## Order Analysis

4. How has the order status distribution changed over time?

```
-- This query is written as per estimated delivery time
SELECT
EXTRACT( YEAR FROM (CAST(order_estimated_delivery_date AS TIMESTAMP))) AS YEAR,
order_status, COUNT(order_id) AS order_count
FROM olist_ecommerce.orders
GROUP BY 1,2
ORDER BY 1;
```

	year	order_status	order_count
1	2016	delivered	264
2	2016	canceled	26
3	2016	unavailable	7
4	2016	processing	2
5	2016	shipped	9
6	2016	invoiced	18
7	2017	approved	2
8	2017	created	2
9	2017	unavailable	425
10	2017	shipped	487
11	2017	delivered	38767
12	2017	processing	209
13	2017	canceled	258
14	2017	invoiced	167
15	2018	created	3
16	2018	shipped	611
17	2018	delivered	57447
18	2018	unavailable	177
19	2018	canceled	341
20	2018	processing	90
21	2018	invoiced	129

## Order Timing Analysis

5. What is the average time between order purchase and approval?

```
SELECT  
concat(  
(round(  
(AVG(EXTRACT(epoch FROM (order_approved_at::TIMESTAMP - order_purchase_timestamp::TIMESTAMP))  
) / 60 / 60), 2))::TEXT, ' ', 'minutes'  
)  
AS average_time_between_order_purchase_and_approval  
FROM olist_ecommerce.orders ;
```

	average_time_between_order_purchase_and_approval
1	10.42 minutes

6. What is the average delivery time from order approval to customer delivery?

```
SELECT  
concat(  
(round((((AVG(EXTRACT(epoch FROM (order_delivered_customer_date::TIMESTAMP -  
order_purchase_timestamp::TIMESTAMP))) / 60) / 60) / 24), 2))::TEXT,  
' ', 'days'  
)  
AS average_time_from_order_approval_to_customer_delivery  
FROM olist_ecommerce.orders  
WHERE order_purchase_timestamp IS NOT NULL  
AND order_delivered_customer_date IS NOT NULL ;
```

	average_time_from_order_approval_to_customer_delivery
1	12.56 days

## Order Volume Trends

7. How does the number of orders vary by month and year?

```
SELECT *,
DENSE_RANK()OVER(PARTITION BY YEAR ORDER BY count_of_orders DESC) AS "rank"
FROM(
  SELECT
  EXTRACT(YEAR FROM order_purchase_timestamp::TIMESTAMP ) as YEAR
  ,to_char(order_purchase_timestamp::TIMESTAMP, 'Mon') as MONTH
  ,COUNT(*) AS count_of_orders
  FROM olist_ecommerce.orders
  GROUP BY 1,2
  ORDER BY 3 DESC)
;
```



	year	month	count_of_orders	rank	
1	2016	Oct	324	1	
2	2016	Sep	4	2	
3	2016	Dec	1	3	
4	2017	Nov	7544	1	
5	2017	Dec	5673	2	
6	2017	Oct	4631	3	
7	2017	Aug	4331	4	
8	2017	Sep	4285	5	
9	2017	Jul	4026	6	
10	2017	May	3700	7	
11	2017	Jun	3245	8	
12	2017	Mar	2682	9	
13	2017	Apr	2404	10	
14	2017	Feb	1780	11	
15	2017	Jan	800	12	
16	2018	Jan	7269	1	
17	2018	Mar	7211	2	
18	2018	Apr	6939	3	
19	2018	May	6873	4	
20	2018	Feb	6728	5	
21	2018	Aug	6512	6	

## 8. What are the peak months for order placements?

```
SELECT YEAR, MONTH, count_of_orders
FROM
(
  SELECT *,
  DENSE_RANK()OVER(PARTITION BY YEAR ORDER BY count_of_orders DESC) AS "rank"
  FROM(
    SELECT
    EXTRACT(YEAR FROM order_purchase_timestamp::TIMESTAMP ) as YEAR
    ,to_char(order_purchase_timestamp::TIMESTAMP, 'Mon') as MONTH
    ,COUNT(*) AS count_of_orders
    FROM olist_ecommerce.orders
    GROUP BY 1,2
    ORDER BY 3 DESC)
  )
WHERE "rank" = 1
;
```

	year	month	count_of_orders	
1	2016	Oct	324	
2	2017	Nov	7544	
3	2018	Jan	7269	

## Order Item Insights

9. What is the average number of items per order?

```
SELECT AVG(COUNT) AS avg_number_of_items
FROM(
    SELECT order_id, COUNT(order_item_id) AS COUNT
    FROM olist_ecommerce.order_items
    GROUP BY 1)
;
```

	avg_number_of_items
1	1.1417306873695093

10. Which products have the highest sales volume?

```
SELECT product_id, product_category_name_english, sales_volumn
FROM(
    SELECT oi.product_id
    ,pcnt.product_category_name_english
    ,COUNT(*) AS sales_volumn
    ,DENSE_RANK()OVER(ORDER BY COUNT(*)DESC) AS rnk
    from olist_ecommerce.order_items oi
    JOIN olist_ecommerce.products p
    on oi.product_id = p.product_id
    JOIN olist_ecommerce.product_category_name_translation pcnt
    on p.product_category_name = pcnt.product_category_name
    GROUP BY 1,2
    ORDER BY 3 DESC)
WHERE rnk = 1
;
```

	product_id	product_category_name_english	sales_volumn
1	aca2eb7d00ea1a7b8ebd4e68314663af	furniture_decor	527

## Payment Analysis

11. What is the distribution of payment types used by customers?

```
SELECT payment_type, COUNT(payment_type) AS no_of_customer_used
FROM olist_ecommerce.order_payments
GROUP BY 1
ORDER BY 2 DESC;
```

	payment_type	no_of_customer_used
1	credit_card	76795
2	boleto	19784
3	voucher	5775
4	debit_card	1529
5	not_defined	3

12. What is the average payment value and how does it vary by payment type?

```
WITH avg_payment AS(
SELECT AVG(payment_value) AS average_payment FROM olist_ecommerce.order_payments
)
SELECT payment_type,
AVG(payment_value) AS average_payment_by_payment_type,
(SELECT average_payment FROM avg_payment) AS average_payment
FROM olist_ecommerce.order_payments
WHERE payment_type != 'not_defined'
GROUP BY 1
ORDER BY 2 DESC;
```

	payment_type	average_payment_by_payment_type	average_payment
1	credit_card	163.31902063935578	154.10038041698365
2	boleto	145.03443540234412	154.10038041698365
3	debit_card	142.57017004578162	154.10038041698365
4	voucher	65.70335411255424	154.10038041698365

## Review Score Analysis

13. What is the distribution of review scores?

```
SELECT english_review_title, review_score, COUNT(1) AS count_of_reviews  
FROM olist_ecommerce.order_reviews  
GROUP BY 1, 2  
ORDER BY 3 DESC  
;
```

	english_review_title	review_score	count_of_reviews
1	Super Recommended	5	57328
2	Recommended	4	19142
3	Not recommended	1	11424
4	Satisfied	3	8179
5	Regular	2	3151

## 14. How do review scores correlate with delivery times?

```
WITH review_score_delivery_time AS(  
    SELECT _or.order_id  
    , _or.review_score  
    , _or.english_review_title  
    ,  
    round((((EXTRACT  
    (epoch FROM (o.order_delivered_customer_date::TIMESTAMP - o.order_purchase_timestamp::TIMESTAMP)) /  
    60) / 60) / 24)  
    , 2) AS delivery_time_in_days  
    FROM olist_ecommerce.order_reviews _or  
    JOIN olist_ecommerce.orders o  
    on _or.order_id = o.order_id  
)  
SELECT english_review_title, review_score, round(AVG(delivery_time_in_days),2) AS avg_delivery_time_in_days  
FROM review_score_delivery_time  
GROUP BY 1,2  
ORDER BY 3  
;
```

	english_review_title	review_score	avg_delivery_time_in_days
1	Super Recommended	5	10.69
2	Recommended	4	12.31
3	Satisfied	3	14.26
4	Regular	2	16.66
5	Not recommended	1	21.31

## Product Category Analysis

15. Which product categories have the highest sales?

```
SELECT product_category_name_english, total_sales
FROM
(
    SELECT pcnt.product_category_name_english,
    SUM(oi.price) AS total_sales,
    DENSE_RANK()OVER(ORDER BY SUM(oi.price) DESC) AS "rank"
    FROM olist_ecommerce.products p
    JOIN olist_ecommerce.order_items oi
    on p.product_id = oi.product_id
    JOIN olist_ecommerce.product_category_name_translation pcnt
    on p.product_category_name = pcnt.product_category_name
    GROUP BY 1)
WHERE "rank" = 1
;
```

	product_category_name_english	total_sales
1	health_beauty	1258681.3400000164

16. What is the average product rating by category?

```
SELECT pcnt.product_category_name_english,
round(AVG(ore.review_score),2) AS avg_reveiw_score
FROM olist_ecommerce.order_items oi
JOIN olist_ecommerce.order_reviews ore
on oi.order_id = ore.order_id
JOIN olist_ecommerce.products p
on oi.product_id = p.product_id
JOIN olist_ecommerce.product_category_name_translation pcnt
on p.product_category_name = pcnt.product_category_name
GROUP BY 1
ORDER BY 2 DESC
;
```

	product_category_name_english	avg_reveiw_score
1	cds_dvds_musicals	4.64
2	fashion_childrens_clothes	4.50
3	books_general_interest	4.45
4	costruction_tools_tools	4.44
5	flowers	4.42
6	books_imported	4.40
7	books_technical	4.37
8	luggage_accessories	4.32
9	food_drink	4.32
10	small_appliances_home_oven_and_c...	4.30
11	fashion_sport	4.26
12	fashion_shoes	4.23
13	food	4.22
14	music	4.21
15	cine_photo	4.21
16	pet_shop	4.19
17	stationery	4.19
18	computers	4.18
19	home_appliances	4.17
20	toys	4.16