

Análisis preliminar de calidad de datos

A continuación se extiende con más detalle en análisis preliminar de la calidad de los datos, especificando los datos por cada tabla.

Nombre de la tabla: 'olist_closed_deals_dataset.csv'.

Descripción: Lista de clientes con contrato de Olist (venden sus productos usando los servicios de Olist).

Tamaño: 842 registros, 14 columnas.

Detalle columnas: (numeradas 0 a 13)

Columna 0:

Nombre: 'mql_id'.

Descripción: Id de 'marketing qualified lead' (potencial Vendedor en Olist, obtiene este id en su primer contacto).

Tipo de dato: 'object'.

NO Nulos: 842/842.

Valores únicos: 842 (todos).

Columna 1:

Nombre: 'seller_id'.

Descripción: Refiere id del vendedor.

Tipo de dato: 'object'

NO Nulos: 842/842.

Valores únicos: 842 (todos). Luego todos los vendedores también tienen su mql_id.

Columna 2:

Nombre: 'sdr_id'.

Descripción: Refiere a 'sales development representative'.

Tipo de dato: 'object'.

NO Nulos: 842/842.

Valores únicos: 32.

Columna 3:

Nombre: 'sr_id'.

Descripción: Id de 'sales representative'.

Tipo de dato: 'object'.

NO Nulos: 842/842.

Valores únicos: 22.

Columna 4:

Nombre: 'won_date'.

Descripción: Fecha de cierre de contrato.

Tipo de dato: 'object'.

NO Nulos: 842/842.

Valores únicos: 824. De tipo 'timestamp' a transformar.

Valor mínimo: '2017-12-05 02:00:00'

Valor máximo: '2018-11-14 18:04:19'.

Columna 5:

Nombre: 'business_segment'.

Descripción: Segmento al cual pertenece o en el cual publicará sus productos.

Tipo de dato: 'object'.

NO Nulos: 841/842.

Valores únicos: 34.

Columna 6:

Nombre: 'lead_type'.

Descripción: Tipo de cliente-vendedor.

Tipo de dato: 'object'.

NO Nulos: 836/842.

Valores únicos: 9: 'online_medium', 'industry', 'online_big', 'online_small', 'offline', 'online_top', 'online_beginner', 'other', nan.

Columna 7:

Nombre: 'lead_behaviour_profile'.

Descripción: Perfil de comportamiento del lead (vendedor). (consultar Olist por caracterización.)

Tipo de dato: 'object'.

NO Nulos: 665/842

Valores únicos: 10: 'cat', 'eagle', nan, 'wolf', 'shark', 'cat, wolf', 'eagle, wolf', 'shark, cat', 'eagle, cat', 'shark, wolf'.

Observación: Hay valores combinados.

Columna 8:

Nombre: 'has_company'.

Descripción: Refieren a si tiene compañía

Tipo de dato: 'object'.

NO Nulos: 63/842

Valores únicos: True, False, nan.

Observación: SE OBSERVAN GRAN MAYORÍA DE NULOS (>90%).

Columna 9:

Nombre: 'has_gtin'.

Descripción: se refiere a si tiene 'Global Trade Item Number'.

Tipo de dato: 'object'.

NO Nulos: 64/842

Valores únicos: True, False, nan.

Observación: SE OBSERVAN GRAN MAYORÍA DE NULOS (>90%).

Columna 10:

Nombre: 'average_stock'.

Descripción: Stock promedio.

Tipo de dato: 'object'.

NO Nulos: 66/842

Valores únicos: Se presentan segmentados de esta manera: (nan, '20-50', '1-5', '5-20', '200+', '50-200', 'unknown').

Observación: SE OBSERVAN GRAN MAYORÍA DE NULOS (>90%).

Columna 11:

Nombre: 'business_type'.

Descripción: Tipo de negocio, del lead (vendedor): fabricante o revendedor (retail).

Tipo de dato: 'object'.

NO Nulos: 832/842

Valores únicos: 'reseller':587 , 'manufacturer':242, 'other':3, nan:10.

Columna 12:

Nombre: 'declared_product_catalog_size'.

Descripción: Tamaño del catálogo del producto.

Tipo de dato: 'object'.

NO Nulos: 69/842

Valores únicos: 34

Observación: SE OBSERVAN GRAN MAYORÍA DE NULOS (>90%).

Columna 13:

Nombre: 'declared_monthly_revenue'.

Descripción: Ganancia o facturación mensual declarada por el lead (vendedor o seller)

Tipo de dato: 'object'.

NO Nulos: 832/842

Valores únicos: 34, pero con 797='0.0'.

Observación: NO APORTA BUENA INFORMACIÓN, MUCHOS 0.

Nombre de la tabla: 'olist_customers_dataset.csv'.

Descripción: Listado de clientes.

Tamaño: 99441 registros, 5 columnas.

Detalle columnas: (numeradas 0 a 4)

Columna 0:

Nombre: 'customer_id'.

Descripción: Identificación de cliente.

Tipo de dato: 'object'.

NO Nulos: 99441/99441.

Valores únicos: 99441.

Observación: No hay Ids duplicados, contiene caracteres alfanuméricos.

Columna 1:

Nombre: 'customer_unique_id'.

Descripción: Otra identificación de cliente?.

Tipo de dato: 'object'.

NO Nulos: 99441/99441.

Valores únicos: 96096.

Observación: Referencia a cliente recurrente , contiene caracteres alfanuméricos.

Columna 2:

Nombre: 'customer_zip_code_prefix'.

Descripción: Prefijo el código postal de Brasil.

Tipo de dato: 'int64'.

NO Nulos: 99441/99441.

Valores únicos: 14994.

Observación: Primeros 5 dígitos (PREFIX): región, sub-región, sector, subsector, división de subsector. Últimos 3 dígitos (SUFIX), NO INCLUIDOS EN ESTA TABLA, corresponden a identificadores de distribución.

Columna 3:

Nombre: 'customer_city'.

Descripción: Referencia la ciudad para el customer_id.

Tipo de dato: 'object'.

NO Nulos: 99441/99441.

Valores únicos: 4119.

Observación: Idioma: Portugues.

Columna 4:

Nombre: 'customer_state'.

Descripción: Referencia al estado, dos letras mayúsculas. Ej: SP= São Paulo.

Tipo de dato: 'object'.

NO Nulos: 99441/99441.

Valores únicos: 27.

Observación: Brasil tiene 26 estados, y la capital 'DF'. Mayor ocurrencia: SP (Sao Paulo), 41746 veces, menor ocurrencia: RR (Roraima), 46 veces.

Nombre de la tabla: 'olist_geolocation_dataset.csv'

Descripción: Datos de geolocalización, por ciudad, prefijo de cod postal y estado.

Tamaño: 1000163 registros, 5 columnas.

Detalle columnas: (numeradas 0 a 4). Ninguna presenta elementos nulos.

Observación: Tiene un tamaño MUY superior al resto de las tablas.

Columna 0:

Nombre: 'geolocation_zip_code_prefix'.

Descripción: Prefijo código postal para geolocalización.

Tipo de dato: 'int64'.

NO Nulos: 1000163/1000163.

Valores únicos: 19015.

Columna 1:

Nombre: 'geolocation_lat'.

Descripción: Referencia a LATITUD, geolocalización.

Tipo de dato: 'float'.

NO Nulos: 1000163/1000163.

Valores únicos: 717360.

Columna 2:

Nombre: 'geolocation_lng'.

Descripción: Referencia a LONGITUD, geolocalización.

Tipo de dato: 'float'.

NO Nulos: 1000163/1000163.

Valores únicos: 717613.

Columna 3:

Nombre: 'geolocation_city'.

Descripción: Ciudad.

Tipo de dato: 'object'.

NO Nulos: 1000163/1000163.

Valores únicos: 8011.

Observación: Son el DOBLE de valores de 'customer_city' que en la tabla anterior, Idioma: Portugues.

Columna 4:

Nombre: 'geolocation_state'.

Descripción: Datos de geolocalización de cada Estado Brasileiro.

Tipo de dato: 'object'.

NO Nulos: 1000163/1000163.

Valores únicos: 27.

Observación: condice con 'customer_state', tabla 2. Mayor ocurrencia: SP (Sao Paulo), 404268 veces. Menor ocurrencia: RR (Roraima), 646 veces.

Nombre de la tabla: 'olist_marketing_qualified_leads_dataset.csv'.

Descripción: Lista de potenciales clientes/sellers (vendedores de productos en Olist)

Tamaño: 8000 registros, 4 columnas.

Detalle columnas: (numeradas 0 a 3)

Columna 0:

Nombre: 'mql_id'.

Descripción: Id de 'marketing qualified lead'. Referencia a col 0, tabla 1

Tipo de dato: 'object'. Caracteres alfanuméricos.

NO Nulos: 8000/8000.

Valores únicos: 8000. No presenta duplicados.

Columna 1:

Nombre: 'first_contact_date'.

Descripción: Fecha de Primer contacto. Formato: FECHA, sin horario.

Tipo de dato: 'object'. Caracteres alfanuméricos.

NO Nulos: 8000/8000.

Valores únicos: 8000. No presenta duplicados. Valor mínimo: '2017-06-14'.

Valor máximo: '2018-05-31'.

Observación: Convertir a date.

Columna 2:

Nombre: 'landing_page_id'.

Descripción: Página desde la cual realiza el primer contacto el 'mql'.

Tipo de dato: 'object'. Caracteres alfanuméricos.

NO Nulos: 8000/8000.

Valores únicos: 495.

Observación: TOP 5 CON MAYOR OCURRENCIA:

id:	Counts:
b76ef37428e6799c421989521c0e5077	912
22c29808c4f815213303f8933030604c	883
58326e62183c14b0c03085c33b9fdc44	495
88740e65d5d6b056e0cda098e1ea6313	445
ce1a65abd0973638f1c887a6efcfa82d	394

Columna 3:

Nombre: 'origin'.

Descripción: Origen del contacto.

Tipo de dato: 'object'.

NO Nulos: 7940/8000.

Valores únicos: 11.

Observación: Se desconocen los detalles de algunas denominaciones.

OCURRENCIAS:	origen:	counts:	Descripción:
	organic_search	2296	búsqueda natural en la web.
	paid_search	1586	llega mediante publicidad dirigida?
	social	1350	redes sociales?
	unknown	1099	-
	direct_traffic	499	-
	email	493	lista de email
	referral	284	referencia por tercera persona.
	other	150	
	display	118	-
	other_publicities	65	otras publicidades
	nan	60	-

Nombre de la tabla: 'olist_order_items_dataset.csv'.

Descripción: Ítems en cada orden.

Tamaño: 112650 registros, 7 columnas.

Detalle columnas: (numeradas 0 a 6). Ninguna presenta elementos nulos.

Columna 0:

Nombre: 'order_id'.

Descripción: Id de orden compra.

Tipo de dato: 'object'. Caracteres alfanuméricos.

NO Nulos: 112650/112650.

Valores únicos: 98666. Presenta duplicados.

Columna 1:

Nombre: 'order_item_id'.

Descripción: Cantidad de unidades en la orden.

Tipo de dato: 'object'. Caracteres alfanuméricos.

NO Nulos: 112650/112650.

Valores únicos: Valor: Conteo:

1	98666
2	9803
3	2287
4	965
5	460
6	256
7	58
8	36
9	28
10	25
11	17
12	13
13	8
14	7
15	5
16	3
17	3
18	3
19	3
20	3
21	1

Columna 2:

Nombre: 'product_id'.

Descripción: Id de producto.

Tipo de dato: 'object'. Caracteres alfanuméricos.

NO Nulos: 112650/112650.

Valores únicos: 32951. Top 5 más repetidos:

product_id:	counts:
aca2eb7d00ea1a7b8ebd4e68314663af	527
99a4788cb24856965c36a24e339b6058	488
422879e10f46682990de24d770e7f83d	484
389d119b48cf3043d311335e499d9c6b	392
368c6c730842d78016ad823897a372db	388

Columna 3:

Nombre: 'seller_id'.

Descripción: Id del vendedor.

Tipo de dato: 'object'. Caracteres alfanuméricos.

NO Nulos: 112650/112650.

Valores únicos: 3095. Top 5 mas repetidos:

seller_id:	counts:
6560211a19b47992c3666cc44a7e94c0	2033
4a3ca9315b744ce9f8e9374361493884	1987
1f50f920176fa81dab994f9023523100	1931
cc419e0650a3c5ba77189a1882b7556a	1775
da8622b14eb17ae2831f4ac5b9dab84a	1551

Columna 4:

Nombre: 'shipping_limit_date'.

Descripción: Fecha límite de envío.

Tipo de dato: 'object'. De tipo TIMESTAMP.

NO Nulos: 112650/112650.

Valores únicos: 93318. Valor mínimo: '2016-09-19 00:15:34'

Valor máximo: '2020-04-09 22:35:08'.

Observación: Está fuera de rango (fecha) con máximos de otras tablas.

Columna 5:

Nombre: 'price'.

Descripción: Precio del producto. Moneda: REALES (\$RS). Decimales: 2.

Tipo de dato: 'float64'.

NO Nulos: 112650/112650.

Valores únicos: 5968. Valores mínimo, máximo: (0.85, 6735.0).

Observación: Identificar producto con precio = 0.85.

Columna 6:

Nombre: 'freight_value'.

Descripción: Costo del envío. Moneda: REALES (\$RS). Decimales: 2.

Tipo de dato: 'float64'.

NO Nulos: 112650/112650.

Valores únicos: 6999. Valores mínimo, máximo: (0.0, 409.68).

Observación: Identificar producto con precio = 0.85.

Nombre de la tabla: 'olist_order_payments_dataset.csv'.

Descripción: Detalles sobre pagos.

Tamaño: 103886 registros, 5 columnas. OBS: Cantidad similar a la tabla anterior.

Detalle columnas: (numeradas 0 a 4). Ninguna presenta elementos nulos.

Columna 0:

Nombre: 'order_id'.

Descripción: Id de orden compra.

Tipo de dato: 'object'. Caracteres alfanuméricos.

NO Nulos: 103886 /103886.

Valores únicos: 99440. Presenta duplicados. (pago en más de 1 cuota.)

Columna 1:

Nombre: 'payment_sequential'.

Descripción: Pago secuencial.

Tipo de dato: 'int64'.

NO Nulos: 103886 /103886.

Valores únicos: 28. Toma valores del 1 al 29. Top 5 mayores ocurrencias:

valor: counts:

1	99360	(aprox 90% total.)
2	3039	(aprox 3% total.)
3	581	
4	278	
5	170	
6	118	

Columna 2:

Nombre: 'payment_type'.

Descripción: Tipo de Pago.

Tipo de dato: 'object'.

NO Nulos: 103886 /103886.

Valores únicos: 5. Ocurrencias: payment_type: counts:

credit_card	76795	(aprox 75% total.)
boleto	19784	(aprox 20% total.)
voucher	5775	
debit_card	1529	
not_defined	3	

Columna 3:

Nombre: 'payment_installments'.

Descripción: Pago de cuotas.

Tipo de dato: 'int64'.

NO Nulos: 103886 /103886.

Valores únicos: 24. Top 5 mayores ocurrencias: valor: counts:

1	52546
---	-------

2	12413
3	10461
4	7098
10	5328

Columna 4:

Nombre: 'payment_value'.

Descripción: Valor del pago, incluye envío.

Tipo de dato: 'float64'.

NO Nulos: 103886 /103886.

Valores únicos: 29077. Valores mínimo, máximo: (0.0, 13664.08).

Nombre de la tabla: 'olist_order_reviews_dataset.csv'.

Descripción: Opiniones/Reviews sobre compras.

Tamaño: 99224 registros, 7 columnas.

Detalle columnas: (numeradas 0 a 6).

Columna 0:

Nombre: 'review_id'.

Descripción: Id de las reviews de los clientes con respecto a sus compras.

Tipo de dato: 'object'.

NO Nulos: 99224/99224.

Valores únicos: 98410. Valores más repetidos:

7b606b0d57b078384f0b58eac1d41d78 (3 apariciones),
dbdf1ea31790c8ecfcc6750525661a9b (3 apariciones),
32415bbf6e341d5d517080a796f79b5c (3 apariciones),
0c76e7a547a531e7bf9f0b99cba071c1 (3 apariciones),
4219a80ab469e3fc9901437b73da3f75 (3 apariciones).

Columna 1:

Nombre: 'order_id'.

Descripción: Id de las órdenes de compra.

Tipo de dato: 'object'.

NO Nulos: 99224/99224.

Valores únicos: 98673. Valores más repetidos:

88b1d1b157a9999ce368f218a407141 (3 apariciones),
8e17072ec97ce29f0e1f111e598b0c85 (3 apariciones),
df56136b8031ecd28e200bb18e6ddb2e (3 apariciones),
03c939fd7fd3b38f8485a0f95798f1f6 (3 apariciones),
5cb890a68b91b6158d69257e4e2bc359 (2 apariciones).

Columna 2:

Nombre: 'review_score'.

Descripción: Puntuación dada por el cliente en su review.

Tipo de dato: 'int64'.

NO Nulos: 99224/99224.

Valores únicos: 5. Valores más repetidos:

5 (57382 apariciones),
4 (19142 apariciones),
1 (11424 apariciones),
3 (8179 apariciones),
2 (3151 apariciones).

Columna 3:

Nombre: 'review_comment_title'.

Descripción: Título de la review.

Tipo de dato: 'object'.

NO Nulos: 87656/99224.

Valores únicos: 4527. Valores más repetidos:

Recomendo (423 apariciones),
recomendo (345 apariciones),
Bom (293 apariciones),
super recomendo (270 apariciones),
Excelente (248 apariciones).

Observación: Alta cantidad de nulos.

Columna 4:

Nombre: 'review_comment_message'.

Descripción: Mensaje de la review del cliente.

Tipo de dato: 'object'.

NO Nulos: 58247/99224.

Valores únicos: 36159. Valores más repetidos:

Muito bom (230 apariciones),
Bom (189 apariciones),
muito bom (122 apariciones),
bom (107 apariciones),
Recomiendo (100 apariciones).

Columna 5:

Nombre: 'review_creation_date'.

Descripción: Momento de creación de la review.

Tipo de dato: 'object'. Convertir TIMESTAMP.

NO Nulos: 99224/99224.

Valores únicos: 636. Valores más repetidos:

2017-12-19 00:00:00 (463 apariciones),
2018-08-28 00:00:00 (452 apariciones),
2017-12-20 00:00:00 (440 apariciones),
2018-05-19 00:00:00 (434 apariciones),
2018-05-15 00:00:00 (425 apariciones).

Columna 6:

Nombre: 'review_answer_timestamp'.

Descripción: Refiere al momento en que se respondió la review.

Tipo de dato: 'object'. Convertir TIMESTAMP.

NO Nulos: 99224/99224.

Valores únicos: 98248. Valores más repetidos:

2017-06-15 23:21:05 (4 apariciones),
2018-02-26 15:53:18 (3 apariciones),
2018-04-07 21:13:05 (3 apariciones),
2017-02-21 23:30:22 (3 apariciones),
2017-08-12 14:35:35 (3 apariciones).

Nombre de la tabla: 'olist_order_dataset.csv'.

Descripción: Ordenes/Compras.

Tamaño: 99441 registros, 8 columnas.

Detalle columnas: (numeradas 0 a 7).

Columna 0:

Nombre: 'order_id'.

Descripción: Refiere a el id de las órdenes de compra.

Tipo de dato: 'object'. Alfanumérico.

NO Nulos: 99441/99441.

Valores únicos: 99441. (Todos)

Columna 1:

Nombre: 'customer_id'.

Descripción: Refiere al id del cliente.

Tipo de dato: 'object'. Alfanumérico.

NO Nulos: 99441/99441.

Valores únicos: 99441. (Todos)

Columna 2:

Nombre: 'order_status'.

Descripción: Refiere al estado de la orden de compra.

Tipo de dato: 'object'.

NO Nulos: 99441/99441.

Valores únicos: 8. Valores más repetidos:

delivered (96478 apariciones),
shipped(1107 apariciones),
canceled (625 apariciones),
unavaible (609 apariciones),
invoiced (314 apariciones) .

Columna 3:

Nombre: 'order_purchase_timestamp'.

Descripción: Refiere al estado de la orden de compra.

Tipo de dato: 'object'. Cambiar a TIMESTAMP.

NO Nulos: 99441/99441.

Valores únicos: 98875. Valor máximo registrado: 2018-10-17 17:30:18

Valor mínimo registrado: 2016-09-04 21:15:19.

Valores más repetidos:

2018-04-11 10:48:14 (3 apariciones),
2018-07-28 13:11:22 (3 apariciones),
2017-11-20 10:59:08 (3 apariciones),
2018-08-02 12:05:26 (3 apariciones),
2018-08-02 12:06:09 (3 apariciones).

Columna 4:

Nombre: 'order_approved_at'.

Descripción: Refiere al momento de aprobación de la compra.

Tipo de dato: 'object'. Cambiar a TIMESTAMP.

NO Nulos: 99281/99441.

Valores únicos: 90733. Valor máximo: 2018-09-03 17:40:06

Valor mínimo: 2016-09-15 12:16:38

Valores más repetidos:

2018-02-27 04:31:10 (9 apariciones),
2017-11-07 07:30:38 (7 apariciones),
2018-02-27 04:31:01 (7 apariciones),
2018-02-06 05:31:52 (7 apariciones),
2017-11-07 07:30:29 (7 apariciones).

Columna 5:

Nombre: 'order_delivered_carrier_date'.

Descripción: Refiere al momento en que comienza el transporte por el 'carrier'.

Tipo de dato: 'object'. Cambiar a TIMESTAMP.

NO Nulos: 97658/99441.

Valores únicos: 81018. Valor máximo : 2018-09-11 19:48:28

Valor mínimo: 2016-10-08 10:34:01

Valores más repetidos:

2018-05-09 15:48:00 (47 apariciones),
2018-05-10 18:29:00 (32 apariciones),
2018-05-07 12:31:00 (21 apariciones),
2018-05-02 15:15:00 (16 apariciones),
2018-07-24 16:07:00 (16 apariciones).

Columna 6:

Nombre: 'order_delivered_customer_date'.

Descripción: Momento en el que la compra llega al cliente (fin de recorrido).

Tipo de dato: 'object'. Cambiar a TIMESTAMP.

NO Nulos: 96476/99441.

Valores únicos: 95664. Valor máximo: 2018-10-17 13:22:46

Valor mínimo: 2016-10-11 13:46:32

Valores más repetidos:

2018-05-08 23:38:46 (3 apariciones),
2016-10-27 17:32:07 (3 apariciones),
2018-02-14 21:09:19 (3 apariciones),
2017-06-19 18:47:51 (3 apariciones),
2018-05-08 19:36:48 (3 apariciones)

Columna 7:

Nombre: 'estimated_delivery_date'.

Descripción: Fecha estimada de llegada del producto al cliente.

Tipo de dato: 'object'. Cambiar a TIME. (no tiene horas)

NO Nulos: 99441/99441.

Valores únicos: 459. Valor máximo: 2018-11-12 00:00:00

Valor mínimo: 2016-09-30 00:00:00

Valores más repetidos:

2017-12-20 00:00:00 (522 apariciones),
2018-03-12 00:00:00 (516 apariciones),
2018-05-29 00:00:00 (513 apariciones),
2018-03-13 00:00:00 (513 apariciones),
2018-02-14 00:00:00 (507 apariciones)

Observación: Se desconoce método de estimación.

Nombre de la tabla: 'olist_products_datasets.csv'.

Descripción: Listado de productos.

Tamaño: 32951 registros, 9 columnas.

Detalle columnas: (numeradas 0 a 8).

Columna 0:

Nombre: 'product_id'.

Descripción: Id de producto.

Tipo de dato: 'object'. Alfanumérico.

NO Nulos: 32951/32951.

Valores únicos: 32951 (todos).

Columna 1:

Nombre: 'product_category_name'.

Descripción: Nombre de la categoría del producto.

Tipo de dato: 'object'.

NO Nulos: 32341/32951.

Valores únicos: 73. Valores más repetidos:

cama_mesa_banho (3029 apariciones),
esporte_lazer (2867 apariciones),

moveis_decoracao (2657 apariciones),
beleza_saude (2444 apariciones),
Utilidades_domesticas (2335 apariciones).

Columna 2:

Nombre: 'product_name_lenght'.

Descripción: Se refiere a la cantidad de caracteres del nombre del producto.

Tipo de dato: 'float64'.

NO Nulos: 32341/32951.

Valores únicos: 66. Valor máximo registrado: 76.0, Valor mínimo registrado: 5.0

Valores más repetidos:

60.0 (2182 apariciones),
59.0 (2025 apariciones),
58.0 (1887 apariciones),
57.0 (1719 apariciones),
55.0 (1683 apariciones).

Columna 3:

Nombre: 'product_description_lenght'.

Descripción: Se refiere a la cantidad de caracteres de la descripción del producto.

Tipo de dato: 'float64'.

NO Nulos: 32341/32951.

Valores únicos: 2960. Valor máximo registrado: 3992.0,

Valor mínimo registrado: 4.0

Valores más repetidos:

404.0 (94 apariciones),
729.0(86 apariciones),
651.0(66 apariciones),
703.0(66 apariciones),
236.0(65 apariciones).

Columna 4:

Nombre: 'product_photos_qty'.

Descripción: Cantidad de fotos del producto publicadas por el vendedor.

Tipo de dato: 'float64'.

NO Nulos: 32341/32951.

Valores únicos: 19. Valor máximo registrado: 20.0, Valor mínimo registrado: 1.0

Valores más repetidos:

1.0 (16489 apariciones),
2.0 (6263 apariciones),
3.0 (3860 apariciones),
4.0 (2428 apariciones),
5.0 (1484 apariciones).

Columna 5:

Nombre: 'product_weight_g'.

Descripción: Refiere al peso del producto en gramos.

Tipo de dato: 'float64'.

NO Nulos: 32949/32951.

Valores únicos: 2204. Valor máximo registrado: 40425.0,

Valor mínimo registrado: 0.0

Valores más repetidos:

200.0 (2048 apariciones),
300.0 (1561 apariciones),
150.0 (1259 apariciones),
400.0 (1206 apariciones),
100.0 (1188 apariciones).

Columna 6:

Nombre: 'product_length_cm'.

Descripción: Refiere al largo del producto en cm.

Tipo de dato: 'float64'.

NO Nulos: 32949/32951.

Valores únicos: 99. Valor máximo registrado: 105.0, Valor mínimo registrado: 7.0

Valores más repetidos:

16.0 (5520 apariciones),
20.0 (2816 apariciones),
30.0 (2029 apariciones),
18.0 (1502 apariciones),
25.0 (1387 apariciones).

Columna 7:

Nombre: 'product_length_cm'.

Descripción: Refiere al largo del producto en cm.

Tipo de dato: 'float64'.

NO Nulos: 32949/32951.

Valores únicos: 102. Valor máximo registrado: 105.0, Valor mínimo registrado: 2.0

Valores más repetidos:

10.0 (2548 apariciones),
15.0 (2022 apariciones),
20.0 (1991 apariciones),
16.0 (1595 apariciones),
11.0 (1551 apariciones).

Columna 8:

Nombre: 'product_width_cm'.

Descripción: Refiere al ancho del producto en cm.

Tipo de dato: 'float64'.

NO Nulos: 32949/32951.

Valores únicos: 95. Valor máximo registrado: 118.0, Valor mínimo registrado: 6.0

Valores más repetidos:

11.0 (3718 apariciones),
20.0 (3053 apariciones),
16.0 (2808 apariciones),
15.0 (2393 apariciones),
30.0 (1786 apariciones).

Nombre de la tabla: 'olist_sellers_dataset.csv'.

Descripción: Listado de vendedores.

Tamaño: 3095 registros, 4 columnas.

Detalle columnas: (numeradas 0 a 3).

Columna 0:

Nombre: 'sellers_id'.

Descripción: Refiere a id vendedores.

Tipo de dato: 'object'. Alfanumérico.

NO Nulos: 3095/3095.

Valores únicos: 3095 (todos).

Columna 1:

Nombre: 'sellers_zip_code_prefix'.

Descripción: Prefijo del código postal de los vendedores.

Tipo de dato: int64'.

NO Nulos: 3095/3095.

Valores únicos: 2246. Valores mas repetidos:

14940 (40 apariciones),
13660(10 apariciones),
13920(9 apariciones),
16200(9 apariciones).

Columna 2:

Nombre: 'seller_city'.

Descripción: Ciudad de los vendedores.

Tipo de dato: 'object'.

NO Nulos: 3095/3095.

Valores únicos: 611. Valores más repetidos:

sao pablo (694 apariciones),
curitiba (127 apariciones),
rio de janeiro (96 apariciones),
belo horizonte (68 apariciones),
ribeirao preto (52 apariciones).

Columna 3:

Nombre:'seller_state'.

Descripción: Nomenclatura del Estado/Provincia de los vendedores.

Tipo de dato: 'object'.

NO Nulos: 3095/3095.

Valores únicos: 23. Valores más repetidos:

SP (1849 apariciones),
PR (349 apariciones),

MG (244 apariciones),
SC (190 apariciones),
RJ (151 apariciones).

Nombre de la tabla: 'product_category_name_translation.csv'.

Descripción: Traducción al inglés de categoría de productos.

Tamaño: 71 registros, 2 columnas.

Detalle columnas: (numeradas 0 y 1).

Columna 0:

Nombre: 'product_category_name'.

Descripción: Categorías de productos, idioma: Portugues.

Tipo de dato: 'object'.

NO Nulos: 71/71.

Valores únicos: 71 (todos).

Columna 1:

Nombre: 'product_category_name_english'.

Descripción: Categorías de productos, idioma: Inglés.

Tipo de dato: 'object'.

NO Nulos: 71/71.

Valores únicos: 71 (todos).