Sergio Alexander Almanza Muñoz
Monica Maria Castro Benitez
Sebastian Gomez Roman
Daniela Rodríguez Vargas

Proyecto Entrega Semana

1. Sector Seleccionado

Para el presente proyecto se ha decidido elegir el sector del comercio electrónico (e-commerce), debido a su relevancia en el entorno digital actual. Este sector permite analizar la trayectoria digital de diversas empresas, permitiendo evaluar su desempeño, crecimiento y viabilidad dentro del mercado competitivo. El análisis de estos aspectos es esencial para identificar patrones de comportamiento del consumidor y medir la productividad de la empresa.

2. Contexto y Caracterización

Olist es una empresa brasileña de tecnología que trabaja como un ecosistema de soluciones digitales que permiten a otras empresas minoristas potenciar sus negocios mediante la operación y conexión de canales de venta y gestión de procesos en una sola plataforma. Por lo tamto, ofrtecen los siguientes servicios [1]:

- Control centralizado de inventarios, pedidos y facturación a través del Sistema ERP.
- Centro de integración que sincroniza en tiempo real los pedidos, precios y stock.
- Sistema POS que conecta el punto de venta físico al entorno digital, lo cual permite realizar transacciones directamente en el punto de venta físico.
- Cuenta digital
- Marketplace que permite publicaciones y marketing

optimizado de productos en los principales canales de venta de Brasil.

3. Análisis Exploratorio de Datos

En la plataforma Kaggle se encuentra una base de datos que registra los pedidos realizados en Olist Store desde el año 2016 a 2018. De todo el conjunto de datos, se han seleccionado cinco datasets con el fin de analizar la compra del cliente, la orden generada y el método de pago. [2]

Para el análisis inicial de cada dataset, se realizó una revisión de los atributos, cada dataset contiene entre cinco y ocho atributos que abarcan desde identificadores, como fechas, categorías, entre otros. Estos atributos se pueden visualizar de manera más precisa en la Figura 1.

	Dataset	Atributo	Tipo	Rol	Descripción
0	Payments	order_id	texto	PK	ID de la orden
1	Payments	payment_sequential	numérico	atributo	Secuencia del pago
2	Payments	payment_type	categórica	atributo	Método de pago utilizado
3	Payments	payment_installments	numérico	atributo	Número de cuotas
4	Payments	payment_value	numérico	atributo	Monto del pago
5	Orders	order_id	texto	PK	ID de la orden
6	Orders	customer_id	texto	FK	ID del cliente
7	Orders	order_status	categórica	atributo	Estado del pedido
8	Orders	order_purchase_timestamp	fecha/hora	atributo	Fecha de compra
9	Orders	order_approved_at	fecha/hora	atributo	Fecha de aprobación
10	Orders	order_delivered_carrier_date	fecha/hora	atributo	Entrega al transportista
11	Orders	order_delivered_customer_date	fecha/hora	atributo	Entrega al cliente
12	Orders	order_estimated_delivery_date	fecha/hora	atributo	Fecha estimada entrega
13	Order_Items	order_id	texto	FK	ID de la orden
14	Order_Items	order_item_id	numérico	atributo	ID del ítem dentro de la orden
15	Order_Items	product_id	texto	FK	ID del producto
16	Order_Items	seller_id	texto	FK	ID del vendedor
17	Order_Items	shipping_limit_date	fecha/hora	atributo	Fecha límite de envío
18	Order_Items	price	numérico	atributo	Precio unitario
19	Order_Items	freight_value	numérico	atributo	Valor flete
20	Customers	customer_id	texto	PK	ID del cliente
21	Customers	customer_unique_id	texto	atributo	ID cliente único
22	Customers	customer_zip_code_prefix	numérico	atributo	Prefijo código postal
23	Customers	customer_city	categórica	atributo	Ciudad del cliente
24	Customers	customer_state	categórica	atributo	Estado del cliente
25	Geolocation	geolocation_zip_code_prefix	numérico	atributo	Prefijo código postal
26	Geolocation	geolocation_lat	numérico	atributo	Latitud
27	Geolocation	geolocation_lng	numérico	atributo	Longitud
28	Geolocation	geolocation_city	categórica	atributo	Ciudad
29	Geolocation	geolocation_state	categórica	atributo	Estado

Figura 1. *Diccionario de Datos*

En cuanto a la calidad de los datos, no existen valores nulos a excepción del dataset orders, el cual contiene nulos en campos de fechas de aprobación y entrega, lo cual puede explicarse con el ciclo de órdenes canceladas. A continuación se realiza el análisis para cada variable relevante:

- Payments

payment_type: Preferencia por pagos digitales a través de tarjetas de

crédito, puede deberse a la posibilidad de pagar en cuotas, lo que incentiva compras de mayor valor. Hay una clara preferencia por medios de pago electrónicos.

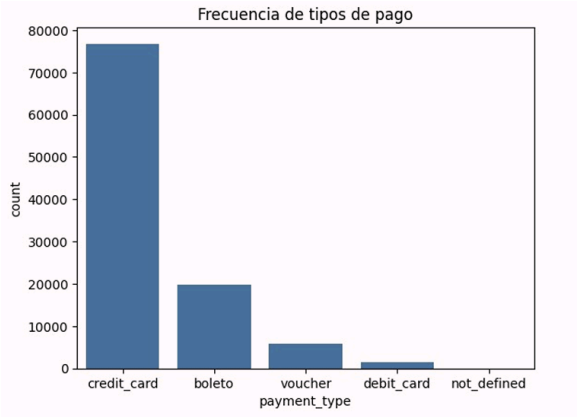

Figura 2. *Frecuencia de tipos de pago.*

payment_installments: Preferencia a realizar compras en una cuota, puede deberse a que el usuario decida no asumir gastos de valor alto

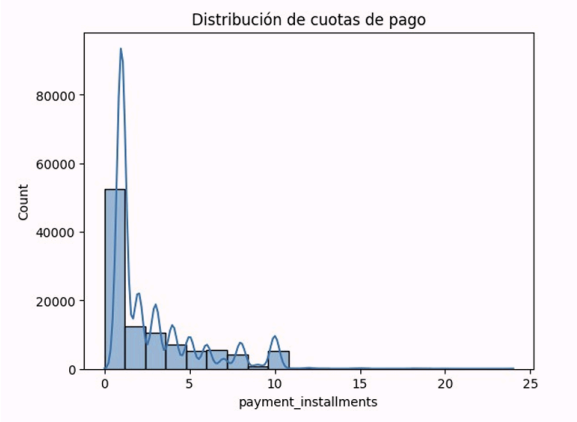

Figura 3. *Distribución de cuotas de pago.*

payment_value: La distribución demuestra un sesgo con pagos inferiores a $1000, lo que sugiere compras con montos inferiores. Existen casos atípicos de montos muy altos que podrían representar oportunidades, como por ejemplo, usuarios premium o riesgos.

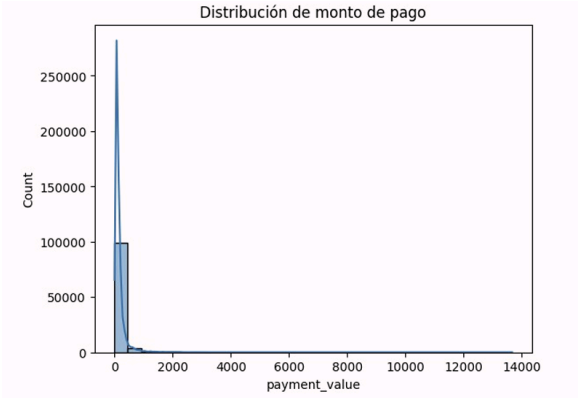

Figura 4. *Distribución de monto de pago.*

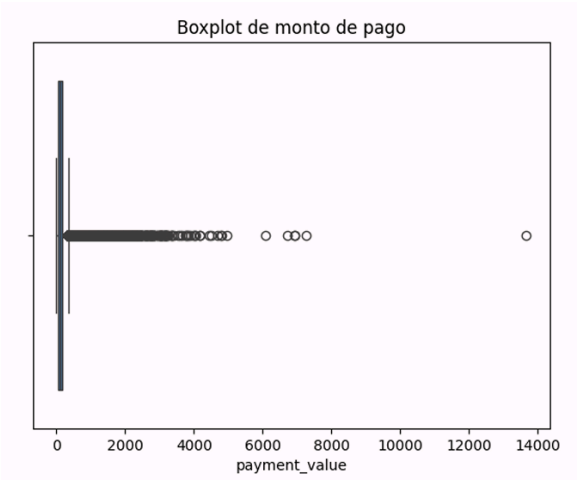

Figura 5. *Boxplot de monto de pago.*

- Orders

order_status: Casi todas las órdenes se encuentran en estado delivered, lo que sugiere un buen manejo logístico. La baja frecuencia de pedidos cancelados o no entregados puede indicar que los procesos dentro de la página son eficientes y confiables.

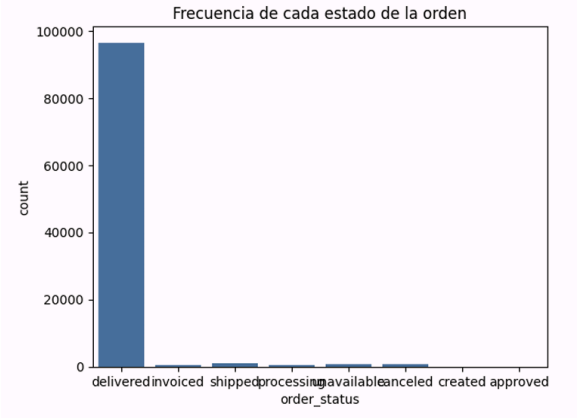

Figura 6. *Frecuencia de cada estado de orden.*

order_purchase_timestamp: Hay un crecimiento mayormente creciente entre el 2017 y 2018, lo que refleja una expansión de la empresa. Sin embargo, para el periodo de julio-octubre del 2018 se evidencia una caida abrupta en las órdenes, lo que sugiere un posible sesgo del dataset al no contar con los registros completos de la totalidad de ese año.

Figura 7. *Órdenes por mes de compra.*

- Items

Price: Casi todos los productos vendidos tienen un bajo precio. Esto puede

significar que Olist se centra en promocionar productos de consumo masivo y de fácil acceso económico, aunque existen casos atípicos con precios más altos. Estos outliers pueden corresponder a productos clasificados como premium que impactan de manera contundente en los ingresos.

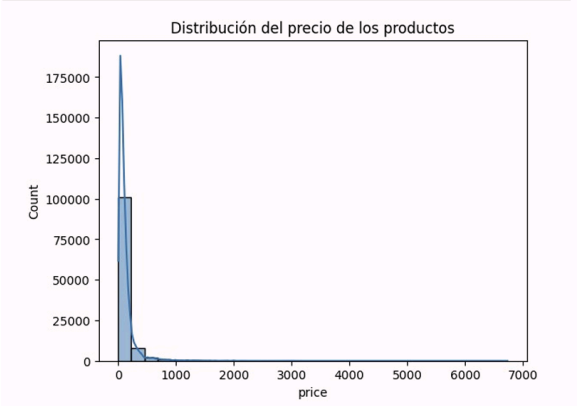

Figura 8. *Distribución del precio de los productos.*

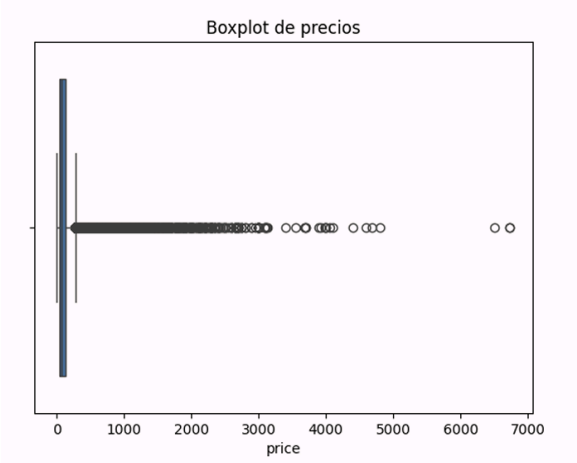

Figura 9. *Boxplot de precios.*

freight_value: La distribución del costo de envío demuestra que la mayoría de transacciones tiene un bajo valor, aunque existen valores de envío superiores al promedio. Esto puede sugerir que los pedidos se encuentran en áreas periféricas o el producto posee particularidades, como su tamaño.

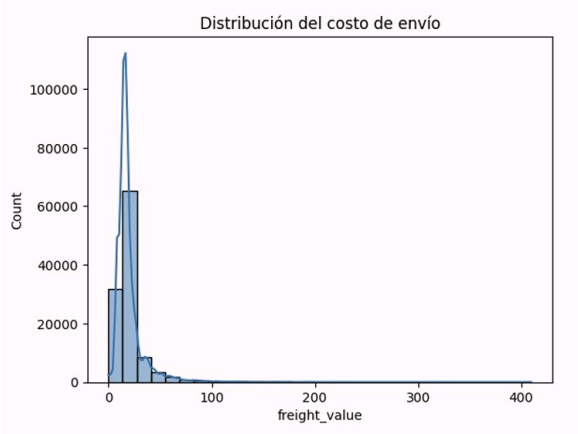

Figura 10. *Distribución del costo de envío.*

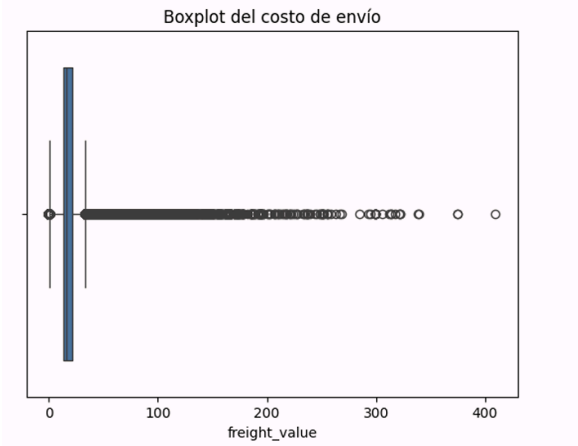

Figura 11. *Boxplot del costo de envío.*

- Customers

customer_city: Se evidencia mayor concentración de clientes en ciudades principales, lo cual indica que Olist posee mayor desarrollo en zonas urbanas con acceso a servicios lógisticos robustos. Podría haber una oportunidad para emprender en ciudades periféricas donde la presencia es baja.

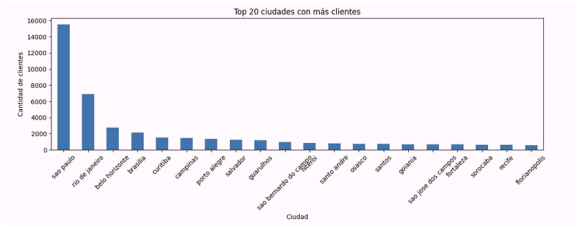

Figura 12. *Ciudades con más clientes.*

customer_state: Al haber mayor concentración de clientes en pocos estados, indica probablemente la centralización de los usuarios en la capital y ciudades importantes que responden a su peso poblacional y mayor desarrollo digital y logístico.

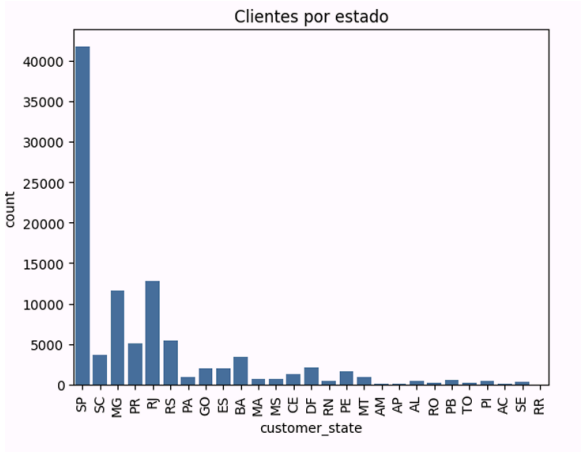

Figura 13. *Distribución de clientes por estado.*

- Geolocalitation

Conclusiones con base al EDA: Para cada dataset con base al análisis de cada atributo, podemos sugerir para la ingeniería de características las siguientes decisiones:

- Orders: Los datos nulos representan estados en proceso o cancelados, por lo que se considera que es relevante no imputar estos datos, sino cambiar la categoría a directamente null. Derivar los atributos de fecha para crear nuevos indicadores, por ejemplo:

delivery_time para analizar tiempos de entrega promedio y retrasos. Derivar con order_delivered_customer_date y order_purchase_timestamp.

approval_time para medir la aprobación de pagos. Derivar con order_approved_at y

y order_purchase_timestamp.

waitin_time para ver tiempos de entrega. Derivar con order_estimated_date y

y order_delivered_cutomer_date.

Considerar eliminar los siguientes atributos después de derivar en los campos sugeridos anteriormente: order_approved_at,

order_delivered_carrier_date,

order_delivered_customer_date,

order_estimated_delivery_date. O derivar en días, semanas, meses, trimestres y años.

- Customers: eliminar customer_id ya que es redundante con customer_unique_id, se decide dejar este atributo ya que permite detectar clientes recurrentes. Agrupar ciudades. Derivar los siguientes atributos:

  is_repeated_customer para identificar si el cliente aparece en más de un pedido, permitiendo segmentar clientes recurrentes y nuevos.

  customer_region para agrupar estados en una misma etiqueta.
- Items: eliminar shipping_limit_date ya que hemos calculado tiempos de entrega desde orders y order_item_id ya que es un consecutivo interno dentro de cada orden. Derivar en:

  total_order_value suamndo price con freight_value mediante order_id para saber el valor de cada pedido.

num_items contando la cantidad de productos distintos que hay mediante order_id para analizar si los clientes compran uno o varios productos.

freight_ratio dividiendo freight_value con price para ver los casos donde el envío cueste más que el producto.

- Payments: eliminar payment_sequential ya que es solo un identificador y no aporta al análisis. Derivar:

  is_credit_card con valores booleanos de 1 para tarjeta de crédito y 0 si no lo es.

  installment_category para categorizar las cuotas.

  Adicionalmente, derivar cada atributo de fecha por periodo (mes, trimestre, año)

4. What aplicado al dataset

El what aplicado a los dataset se resume en la Figura 14.

	Dataset	Campo	Tipo	Naturaleza	Descripción
0	Customers	customer_id	Nominal (texto)	Divergente	Identificador del usuario
1	Customers	unique_customer_id	Nominal (texto)	Divergente	Identificador único del usuario
2	Customers	customer_zip_code_prefix	Numérico	Divergente	Prefijo código postal del usuario
3	Customers	customer_city	Categórico nominal (texto)	Divergente	Ciudad del usuario
4	Customers	customer_state	Categórico nominal (texto)	Divergente	Estado de residencia del usuario
5	Geolocation	geolocation_zip_code_prefix	Numérico	Divergente	Prefijo código postal de la ubicación
6	Geolocation	geolocation_lat	Numérico	Secuencial	Latitud de la ubicación
7	Geolocation	geolocation_lng	Numérico	Secuencial	Longitud de la ubicación
8	Geolocation	geolocation_city	Categórico nominal (texto)	Divergente	Ciudad de la ubicación
9	Geolocation	geolocation_state	Categórico nominal (texto)	Divergente	Estado de la ubicación
10	Order Items	order_id	Nominal (texto)	Divergente	Identificador de la orden
11	Order Items	order_item_id	Numérico	Secuencial	Número de ítem en la orden
12	Order Items	product_id	Nominal (texto)	Divergente	Identificador del producto
13	Order Items	seller_id	Nominal (texto)	Divergente	Identificador del vendedor
14	Order Items	shipping_limit_date	Categórico ordinal (fecha)	Secuencial	Fecha límite de envío
15	Order Items	price	Numérico continuo	Secuencial	Precio del producto
16	Order Items	freight_value	Numérico continuo	Secuencial	Valor del flete
17	Order Payments	order_id	Nominal (texto)	Divergente	Identificador de la orden
18	Order Payments	payment_sequential	Numérico discreto	Secuencial	Pago secuencial
19	Order Payments	payment_type	Categórico nominal	Divergente	Tipo de pago
20	Order Payments	payment_installments	Numérico discreto	Secuencial	Pagos a plazos
21	Order Payments	payment_value	Numérico continuo	Secuencial	Valor del pago
22	Orders	order_id	Nominal (texto)	Divergente	Identificador de la orden
23	Orders	customer_id	Nominal (texto)	Divergente	Identificador del cliente
24	Orders	order_status	Categórico nominal	Divergente	Estado de la orden
25	Orders	order_purchase_timestamp	Categórico ordinal (fecha)	Secuencial	Fecha/hora de compra
26	Orders	order_approved_at	Categórico ordinal (fecha)	Secuencial	Fecha/hora de aprobación
27	Orders	order_delivered_carrier_date	Categórico ordinal (fecha)	Secuencial	Fecha entrega transportista
28	Orders	order_delivered_customer_date	Categórico ordinal (fecha)	Secuencial	Fecha entrega al cliente
29	Orders	order_estimated_delivery_date	Categórico ordinal (fecha)	Secuencial	Fecha entrega estimada

Figura 14. *What aplicado a la base de datos.*

5. Preguntas de Negocio

5.1. Tarea Principal

Descubrir cúal es la tendencia de temporal de ventas de Olist, analizando cómo la plataforma se ha expandido a lo largo de los años 2016 a 2018, para identificar patrones de consumo y ventas

5.2. Tareas Secundarias

- Analizar la evolución del monto promedio de compra por cliente en distintos periodos de tiempo, para saber cual es el monto

(ticked) promedio mensual por cliente y su variación en el tiempo.

Nota: puede responderse con total_order_value y is_repeated_customer.

- Comparar la distribución geográfica de ventas en regiones y así identificar las ciudades y estados que concentran mayor número de ventas

  Nota: puede responderse con customer_region.

- Observar los patrones de elección de método de pago para descubrir cuales son los más utilizados y sus proporciones en las transacciones.

  Nota: puede responderse con is_credit_card e installment_category.

- Hacer una correlación entre el costo del envío y el precio del producto vendido con el objetivo de encontrar la relación existente entre el costo del envío y el precio del producto vendido.

  Nota: puede responderse con freight_ratio.

- Examinar la relación entre el método de pago y la proporción de ventas entregadas frente a canceladas, identificando patrones de comportamiento y posibles diferencias en el desempeño de cada método.

  Nota: puede responderse con order_status y payment_type.

8. Definición de Métricas, Variables e Indicadores

En esta sección se definen las métricas, variables e indicadores asociados a

las tareas analíticas planteadas en la Sección 8. La construcción de estos elementos permite operacionalizar las preguntas de negocio y garantizar que cada tarea pueda ser evaluada a partir de la información disponible en los *datasets* de Olist.

La metodología empleada consistió en identificar las variables relevantes de cada *dataset*, asociarlas con una métrica cuantitativa y, finalmente, derivar un indicador que pueda implementarse en tableros de control y visualizaciones interactivas. De esta manera se asegura la trazabilidad entre los objetivos de negocio, los datos y los resultados esperados. En la **Tabla I** se resume la relación entre las tareas analíticas, las variables de entrada, las métricas de cálculo y los indicadores propuestos.

Tabla I. Variables, métricas e indicadores por tarea analítica

T Ar	**Va (data**	**Métri ca**	**Indica dor**	**Métri ca de valida ción**
Tende ncia de venta s de Olist (2016 –201 8)	order_i order_p estamp price, fi (order_	Volume por per de pric freight_	Serie te mensua totales; porcent interan	¿Se recon oce la tende ncia gener al (creci meint o y caída) . ? Utiliz ar escala de 1 a 10
Ticket mensu y su v tiempo	custom er_uni que_id (custo mers), price, freight _value (order _items), order_ purcha se_tim	Ticket (Σ valo / Nº de únicos	Ticket mensua porcent mes	¿El usuari o puede identif icar si el ticket prome dio sube o baja en un period
Ciuda que co mayor ventas	custom er_city , custom er_stat e (custo mers), order_i d (orders)	Númer por ubi	Rankin ciudade con ma de vent	¿Cuál es podría n ser las tres ciuda des con mayor volum en de ventas ?
Métoc más u propo transa	payme nt_typ e, payme nt_val ue (paym ents)	Distrib relativa de pag transac tipo)	Particip porcent método el total	¿Cual es el métod o de pago más usado y que propo rcione s existe n entre estos?
Relaci costo precio vendic	price, fi (order_	Ratio d freight_ price	Promec distribu ratio de proporc pedidos 1	¿El usuari o puede recon ocer casos donde el valor del
	estamp (orders)			o de tiemp o?

				envío es mayor al valor del produ cto?
Propo ventas vs. car métod	order_s (orders) paymer (payme	Tasa de = (# de total de por mé	Porcen taje de órdene s compl etadas vs. cancel adas discri minad o por métod o de pago	¿Se puede difere nciar fácilm ente qué métod o de pago tiene mayor tasa de cancel ación?

9. Referencias

[1] “Brazilian E-Commerce Public Dataset by Olist,” Kaggle, disponible en: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data. [Accedido: Sep. 22, 2025].

[2] “O que é Olist,” Olist, disponible en: https://olist.com/o-que-e-olist/. [Accedido: Sep. 22, 2025].