Proyecto Visualización e-commerce Olist

Sergio Alexander Almanza Muñoz
Mónica María Castro Benitez
Sebastián Gómez Román
Daniela Rodríguez Vargas

Pontificia Universidad Javeriana
Facultad de Ingeniería
Departamento de Ingeniería de Sistemas
Bogotá D.C., Colombia
Octubre de 2025

I. Introducción

El comercio electrónico ha transformado la dinámica empresarial en el entorno digital. Este estudio se enfoca en Olist, una empresa brasileña que ofrece soluciones digitales para minoristas, integrando canales de venta y gestión en una sola plataforma.

II. Objetivos

II-A. Objetivo General

Descubrir cuál es la tendencia 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.

II-B. Objetivos Específicos

- 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.
- Comparar la distribución geográfica de ventas en regiones y así identificar las ciudades y estados que concentran mayor número de ventas.
- Observar los patrones de elección de método de pago para descubrir cuales son los más utilizados y sus proporciones en las transacciones.
- 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.
- 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.

III. Contexto y Caracterización

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.

En este contexto, Olist se presenta como un caso de estudio pertinente para explorar cómo las plataformas digitales influyen en la eficiencia operativa y en las dinámicas de consumo dentro del comercio electrónico.

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 tanto, ofrecen 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 y Marketplace que permite publicaciones y marketing optimizado de productos en los principales canales de venta de Brasil.

La información recopilada de esta e-commerce se encuentra en la plataforma Kaggle, donde se dispone un dataset público con mas de cien mil pedidos realizados entre los años 2016 y 2018 con información referente a procesos y patrones de compra, medios de pago, comportamiento del cliente, desempeño logístico y ubicación geográfica.

IV. Indicadores

En esta sección se definen las métricas, variables e indicadores asociados a las tareas analíticas planteadas en los objetivos del proyecto. 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. Adicionalmente se indica cómo se va a medir la calidad de cada indicador.

La tabla se encuentra al final del documento.

V. Diseños

V-A. Análisis Exploratorio (EDA)

La base de datos disponible de Olist en Kaggle, cuenta con 9 datasets en total, los cuales recopilan toda la información mencionada en el contexto de la empresa. Para fines del proyecto, seleccionamos 5 datasets que representan el ciclo completo de venta. A continuación de detalla cada dataset incluido [2]:

- **olist_orders_dataset.csv:** contiene información de las órdenes de compra, estados y fechas relevantes de cada orden.

	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 1: Diccionario de Datos

- **olist_order_items_dataset.csv:** contiene detalles de los productos, precios y costos de envíos.
- **olist_order_payments_dataset.csv:** contiene métodos de pago, monto y número de cuotas.
- **olist_customers_dataset.csv:** contiene información geográfica y demográfica de los clientes.
- **olist_geolocation_dataset.csv:** contiene datos de localización.

En la Figura 1 se observan todos los atributos por dataset:

El análisis aplicado a cada atributo se encuentra en el anexo, por lo tanto, a continuación indicamos los resultados generales con base al EDA:

- **Dominio Geográfico:** Se confirmó la alta concentración de órdenes en el estado de São Paulo, lo que sugiere la necesidad de un análisis regional específico para identificar oportunidades de crecimiento en otras zonas del país.
- **Comportamiento de Pago:** Se estableció que la tarjeta de crédito es el principal medio de pago, lo que justifica la creación de métricas específicas para analizar el valor de pago y el número de cuotas asociadas.

V-B. Preparación de datos

El proceso para la limpieza y la preparación de los datos se realizó en R. En esta primera parte preparamos los datos para el posterior modelamiento en Power Bi. Por lo tanto, se ejecutaron las siguientes acciones adjuntas al archivo de extensión .R:

- Normalización de nombres de variables con clean_names().
- Conversión de tipos de datos de fecha para order_purchase_timesmap, order_approved_at y order_delivered_customer_date.
- Eliminación y control de valores nulos o no válidos, evitando distorsiones en los cálculos derivados.
- Agregación de ítems y pagos por order_id para evitar duplicaciones y mantener un grano uniforme a nivel de orden.
- Homologación de códigos regionales (customer_state) mediante un diccionario (uf_region) que agrupa los estados brasileños por región geográfica (Norte, Nordeste, Sudeste, Sul y Centro-Oeste).

V-C. Ingeniería de Características

Una vez completada la limpieza de datos se derivaron nuevas variables a partir de los datos originales con el fin de enriquecer el análisis y facilitar la creación del modelo analítico. Las principales variables derivadas fueron:

- **total_order_value:** suma total del valor del pedido, obtenida a partir del precio de los productos más el costo del flete.
- **freight_ratio:** proporción entre el costo del flete y el valor total de los productos. Permite analizar el peso logístico del envío frente al costo de los bienes.
- **delivery_time:** número de días entre la fecha de compra y la fecha de entrega al cliente, utilizado para evaluar la eficiencia del proceso logístico.
- **approval_hours:** tiempo transcurrido (en horas) entre la creación del pedido y su aprobación, indicador del desempeño operativo.
- **is_repeated_customer:** variable binaria que identifica si un cliente realizó más de una compra dentro del periodo de análisis.
- **delivered_flag:** marca si la orden fue entregada exitosamente.
- **ontime_flag:** indica si la entrega se realizó dentro del tiempo estimado por la plataforma.
- **installments_category:** clasificación del número de cuotas de pago en rangos (1, 2–3, 4–6, 7–12, 13+).
- **main_payment_type:** método de pago principal por valor monetario en la transacción, útil para el análisis de preferencias de pago.
- **region:** variable categórica derivada del estado del cliente, agrupada en las regiones geográficas de Brasil (Norte, Nordeste, Sudeste, Sur, Centro-Oeste).

V-D. Modelo de datos

Para el modelado de datos usamos un esquema en estrella con grano orden (order_id). La tabla de

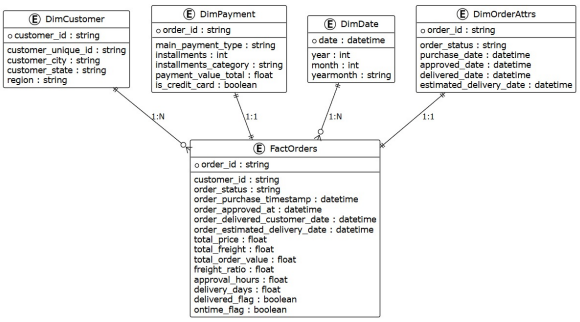

Figura 2: Diagrama del Modelo

hechos FactOrders concentra el valor de la orden, el flete, los flags de cumplimiento y los tiempos. Las dimensiones son: DimDate para calendario, DimCustomer para atributos de clienteUFregión y DimPayment agregada por order_id para método/cuotas. Las relaciones son 1 a muchos entre DimDate y DimCustomer con FactOrders, y 1 a 1 entre DimPayment y FactOrders.

Este modelo evita el doble conteo originado por múltiples filas por ítem y por pago en las fuentes, al consolidar todo a nivel orden. Alinea el análisis con los objetivos del proyecto: tendencias temporales, métodos de pago y cuotas, y evaluación logísticatiempos. Facilita filtros sencillos por fecha, cliente y pago, y deja medidas claras como tasa de finalización, entregas tardías y distribución del ratio fleteprecio.

La tabla de hechos FactOrders concentra las métricas numéricas y los indicadores derivados del proceso de compra y entrega, tales como el valor total de la orden, los costos logísticos y los tiempos de entrega.

A partir de esta tabla central, se establecen relaciones con las siguientes tablas de dimensión:

- **DimCustomer:** contiene información del cliente (identificadores, ciudad, estado y región).
- **DimPayment:** resume los métodos de pago dominantes, el total pagado y la categoría de cuotas por orden.
- **DimDate:** agrupa las fechas de compra, aprobación y entrega, permitiendo análisis temporales por año, mes o día.
- **DimOrderAttrs:** incluye el estado del pedido y las fechas clave del proceso logístico (compra, aprobación, entrega y estimado).

A continuación adjuntamos el diagrama del modelo:

V-E. Modelo de Visualización

Para la visualización de los datos de Olist se han seleccionado seis idioms correspondientes a los objetivos específicos.

VI. Visualizaciones clave del análisis

- **Tendencia mensual de ventas – Comparación de ventas (Linechart):** Permite identificar las tendencias de ventas a lo largo del tiempo, mostrando sus distintos comportamientos (aumentos, declives, etc), para poder obtener el proceso de crecimiento de la empresa.
- **Monto promedio de compra por mes – Ridgeline Plot de ticket promedio por usuarios (mensual):** El idiom nos ayuda a identificar la variabilidad y el comportamiento de los tíquets de compra de los usuarios según los meses del año, mostrando variaciones de manera ampliada. Facilita identificar ouliers, patrones y cambios drásticos en los montos promedio.
- **Ventas por estado o región – Idiom Geográfico:** Este gráfico geográfico representa la distribución de las ventas por ciudades y estados, permitiendo conocer el alcance de la plataforma y detectar regiones con mayor concentración de clientes o áreas con oportunidades de crecimiento.
- **Frecuencia de métodos de pago – Barras apiladas (Método de pago vs Estado de pedido):** Con este idiom podemos comprar la tasa de éxito de las entregas de pedidos con el método de pago realizado, así infiriendo cuál método de pago podría ser más confiable para el usuario. Las barras apiladas permiten comparar la tasa de éxito de las entregas según el método de pago, mostrando la proporción de órdenes completadas y canceladas, esto ayuda a inferir cuáles métodos son más confiables para los usuarios.
- **Relación entre precio del producto y costo de envío – Gráfico de dispersión (Scatterplot):** Gracias a este gráfico podemos identificar la relación entre los dos costos para identificar posibles desequilibrios o desproporciones entre precios.
- **Intensidad de ventas por trimestre – Heatmap trimestral:** Con este idiom lograremos identificar la intensidad de las ventas según los trimestres de un año específico lo que nos permite visualizar los comportamientos en cuanto a ventas que afectan al negocio a lo largo de un periodo

La plantilla del dashboard tendrá la siguiente estructura:

- Tendencia mensual de ventas: Observa el crecimiento o decrecimiento de las ventas en el tiempo.
- Monto promedio de compra por mes: Evalúa la variabilidad en los tickets de los clientes.

- Ventas por estado o región: Visualiza las regiones con mayor concentración de ventas.
- Frecuencia de métodos de pago: Muestra los métodos más utilizados y su proporción en las transacciones.
- Relación entre precio del producto y costo de envío: Identifica desproporciones entre ambos costos.
- Estado de las órdenes según método de pago: Compara la tasa de éxito de los pedidos y analiza la confiabilidad de cada método de pago.

La plantilla se encuentra adjunta a la carpeta de anexos.

VII. Herramientas

Se usaron las siguientes herramientas para el análisis inicial, limpieza y modelamiento de datos:

VIII. Herramientas y fuentes de datos

- **Python:** Utilizado para el análisis exploratorio de datos (EDA), permitiendo detectar patrones de comportamiento de los usuarios y *outliers* en los tickets de compra.
- **Power BI:** Empleado para el diseño e implementación de la plantilla para el dashboard interactivo y la creación de las variables derivadas en la ingeniería de características.
- **Kaggle / datasets de Olist:** Fuente principal de los datos, utilizada para obtener registros de pedidos, productos, pagos y clientes entre 2016 y 2018.
- **R:** Se empleó para la limpieza, depuración y preparación de los datasets de Olist. Esto incluyó manejo de valores nulos, derivación de nuevas variables y agrupaciones por periodo (mes, trimestre, año).

IX. Análisis de resultados

Este espacio se reserva para la última entrega del informe.

X. Conclusiones

Para la segunda entrega del proyecto se realizó el procesamiento de la base de datos junto con la ingeniería de Características, lo que permitió establecer el modelo con el cual se va a aplicar la analítica para la creación de la visualización enfocada a los objetivos de negocio especificados en los objetivos del proyecto.

XI. Lecciones aprendidas

Se aprendió que un proyecto de visualización no consiste únicamente en mostrar datos, sino en construir una narrativa coherente que abarque desde la limpieza hasta la presentación de los resultados. La preparación cuidadosa de los datos garantiza su confiabilidad; el análisis exploratorio permite identificar patrones y tendencias relevantes; y la elección adecuada de herramientas y tipos de visualización facilita la interpretación de la información. Además, documentar cada paso con claridad no solo mejora la comunicación dentro del equipo y con los evaluadores, sino que también refleja rigor y profesionalismo. Esta experiencia evidenció que una visualización efectiva requiere combinar precisión técnica con sensibilidad para comunicar los hallazgos de manera clara y significativa.

Referencias

[1] Kaggle, "Brazilian E-Commerce Public Dataset by Olist." [Online]. Available: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data

[2] Olist, "O que é Olist." [Online]. Available: https://olist.com/o-que-e-olist/

Idiom	Variables	Métrica	Indicador	Criterios de calidad
Línea temporal	payment_value, customer_id, order_purchase_timestamp	Promedio mensual	Monto promedio de compra por cliente	1. Precisión de tendencia respecto a datos reales. 2. Consistencia temporal. 3. Claridad en la variación mensual.
Mapa geográfico	customer_city, customer_state, order_id	Conteo de ventas	Concentración geográfica de ventas	1. Precisión geográfica. 2. Contraste entre regiones de alta y baja venta. 3. Representación proporcional correcta de las distintas áreas.
Gráfico de barras agrupadas	payment_type, payment_value, order_id	Proporción	Distribución de métodos de pago	1. Diferenciación visual entre categorías. 2. Escala balanceada entre los factores y sus ejes. 3. Claridad de etiquetas y porcentajes.
Gráfico de dispersión	freight_value, price	Correlación	Relación entre costo de envío y precio del producto	1. Correlación visual aproximada a la estadística. 2. Claridad en la densidad de puntos. 3. Detección de patrones y outliers.
Gráfico de barras apiladas	payment_type, order_status	Cumplimiento según tipo de pago	Entregas vs cancelaciones por método de pago	1. Correcta proporción entre entregados y cancelados. 2. Distinción cromática entre estados. 3. Fácil interpretación de las variables.

Cuadro I: Idioms, métricas, variables, indicadores y criterios de calidad de las visualizaciones analíticas del proyecto Olist.