Informe del Proyecto de Análisis de Datos

Análisis y Visualización de Datos del Brazilian E-Commerce

Integrantes:

- Almirón Pedro Augusto
- Altamirano Axel Adrian
- Cardozo Carola Guillermina

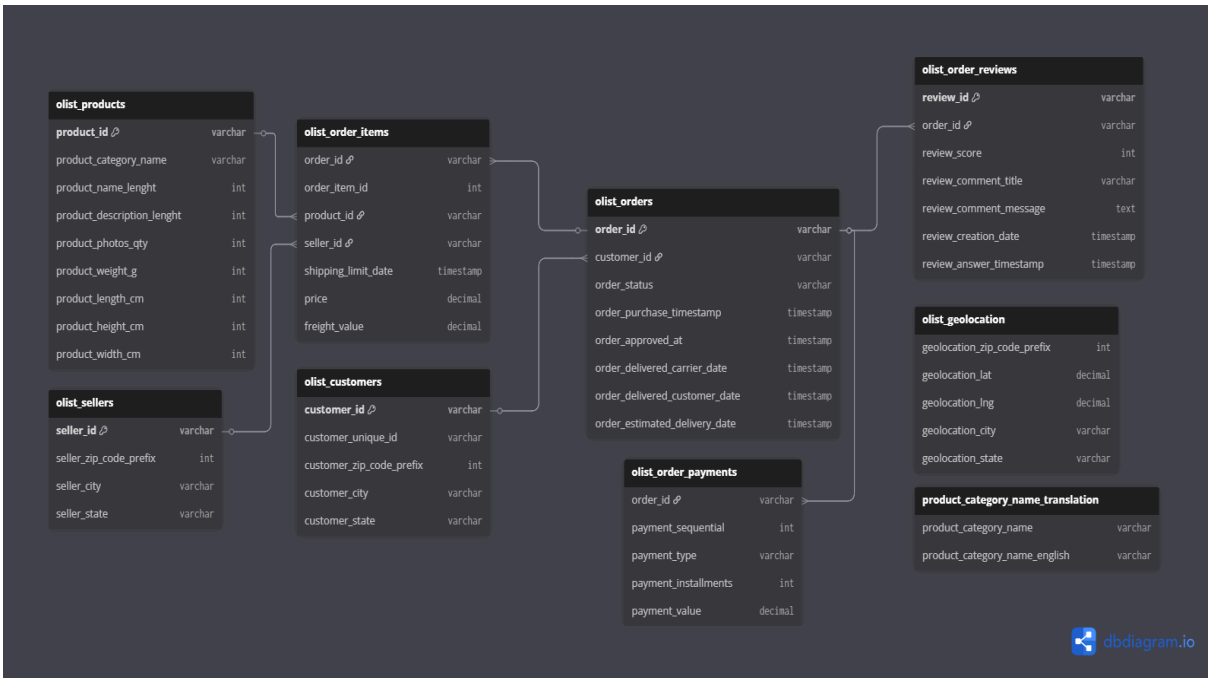
ÍNDICE

1. Descripción del dataset

El dataset utilizado en este proyecto corresponde al **Brazilian E-Commerce Public Dataset (Olist)**, un conjunto de datos públicos que contiene información real de un marketplace brasileño. El mismo registra transacciones realizadas entre clientes y vendedores, incluyendo pedidos, productos, pagos, vendedores, clientes y ubicaciones geográficas. Los datos se presentan en archivos separados (formato CSV), cada uno representando una entidad del negocio.

La estructura original responde a un modelo relacional normalizado, orientado a operaciones (OLTP), y no a análisis analítico directo.

Diagrama entidad relación (DER)



Descripción de las tablas

<u>olist_customers_dataset</u>:

Registra los pedidos realizados en la plataforma y su estado.

Columnas:

- order_id: Identificador único del pedido.
- customer_id: Identificador del cliente.
- order_status: Estado del pedido (delivered, shipped, canceled, etc.).
- order_purchase_timestamp: Fecha y hora de compra.

- order_approved_at: Fecha de aprobación del pago.
- order_delivered_carrier_date: Fecha de envío al transportista.
- order_delivered_customer_date: Fecha de entrega al cliente.
- order_estimated_delivery_date: Fecha estimada de entrega.

olist_orders_dataset:

Registra los pedidos realizados en la plataforma y su estado.

Columnas:

- order_id: Identificador único del pedido.
- customer_id: Identificador del cliente.
- order_status: Estado del pedido (delivered, shipped, canceled, etc.).
- order_purchase_timestamp: Fecha y hora de compra.
- order_approved_at: Fecha de aprobación del pago.
- order_delivered_carrier_date: Fecha de envío al transportista.
- order_delivered_customer_date: Fecha de entrega al cliente.
- order_estimated_delivery_date: Fecha estimada de entrega.

olist_order_items_dataset:

Detalle de los productos incluidos en cada pedido.

Columnas:

- order_id: Identificador del pedido.
- order_item_id: Número del ítem dentro del pedido.
- product_id: Identificador del producto.
- seller_id: Identificador del vendedor.
- shipping_limit_date: Fecha límite de envío.
- price: Precio del producto.
- freight_value: Costo de envío.

olist_products_dataset:

Información descriptiva de los productos.

Columnas:

- product_id: Identificador del producto.
- product_category_name: Categoría del producto.
- product_name_lenght: Longitud del nombre del producto.
- product_description_lenght: Longitud de la descripción.
- product_photos_qty: Cantidad de fotos.

olist_sellers_dataset:

Información sobre los vendedores del marketplace.

Columnas:

- seller_id: Identificador del vendedor.
- seller_zip_code_prefix: Prefijo del código postal.
- seller_city: Ciudad del vendedor.
- seller_state: Estado del vendedor.

olist_order_payments_dataset:

Información relacionada a los pagos de los pedidos.

Columnas:

- order_id: Identificador del pedido.
- payment_sequential: Número de pago (en caso de pagos múltiples).
- payment_type: Tipo de pago (credit_card, boleto, etc.).
- payment_installments: Cantidad de cuotas.
- payment_value: Monto pagado.

olist_order_reviews_dataset:

Opiniones y calificaciones realizadas por los clientes.

Columnas:

- review_id: Identificador de la reseña.
- order_id: Identificador del pedido.
- review_score: Puntaje de satisfacción (1 a 5).
- review_comment_title: Título del comentario.
- review_comment_message: Comentario del cliente.
- review_creation_date: Fecha de creación de la reseña.
- review_answer_timestamp: Fecha de respuesta.

olist_geolocation_dataset:

Información geográfica basada en códigos postales.

Columnas:

- geolocation_zip_code_prefix: Prefijo del código postal.
- geolocation_lat: Latitud.
- geolocation_lng: Longitud.
- geolocation_city: Ciudad.
- geolocation_state: Estado.

Este dataset permite estudiar el comportamiento de ventas, la performance de vendedores, la distribución geográfica del comercio electrónico y la evolución temporal de las transacciones.

El objetivo principal del proyecto es **transformar los datos transaccionales en una base analítica**, permitiendo la obtención de métricas clave de negocio y su posterior visualización en dashboards.

2. Análisis de tablas

Análisis de la tabla Customers

Descripción general

La tabla Customers contiene información básica de los clientes de la plataforma de e-commerce.
Cuenta con **99.441 registros y 5 columnas**, representando clientes únicos identificados por distintos atributos geográficos.

Estructura de la tabla

- **Filas:** 99.441
- **Columnas:** 5

Esto indica que cada fila representa un cliente registrado en el sistema, sin particiones temporales ni duplicaciones aparentes.

Decisión:
Los **tipos de datos** son correctos y coherentes con su significado.
No fue necesario realizar conversiones de tipo.

Análisis de valores nulos

Todas las columnas presentan **0 valores nulos** (0%).

Interpretación:
La información de clientes está completa, sin registros incompletos.

Decisión de limpieza:
No se aplicaron tratamientos sobre valores nulos ya que la calidad del dataset es óptima en este aspecto.

Análisis de duplicados

- **Filas duplicadas:** 0

Interpretación:
No existen registros duplicados a nivel fila completa.

Decisión:
No fue necesario eliminar ni consolidar registros duplicados.

Decisiones para el modelo analítico

- Se conserva **customer_id** como clave primaria para relaciones con pedidos.
- Se mantiene **customer_unique_id** para análisis de clientes recurrentes.
- Las variables geográficas (**zip_code_prefix**, **city**, **state**) se preservan para análisis regionales.

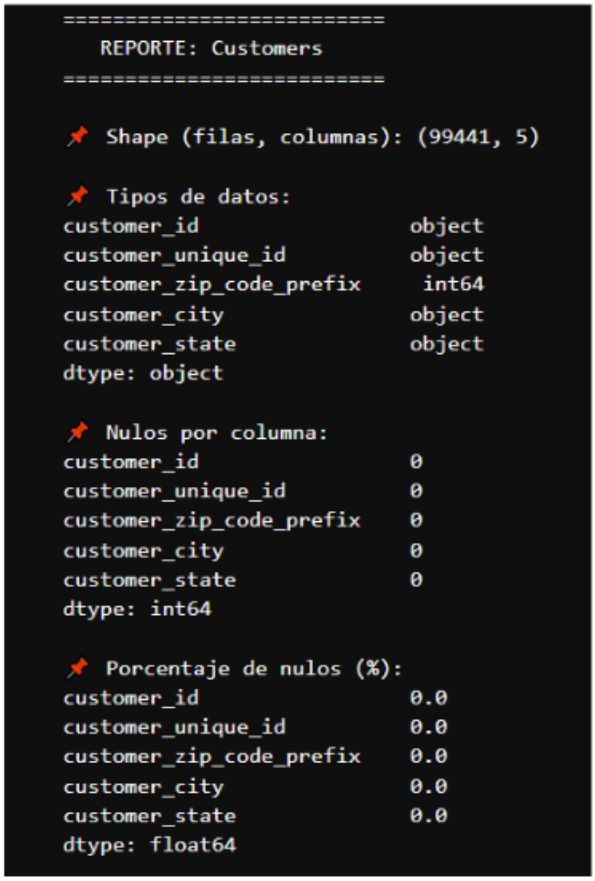
Resultado:
La tabla **Customers** está lista para ser utilizada como **dimensión (`dim_customer`)** dentro del modelo estrella, sin necesidad de transformaciones adicionales.

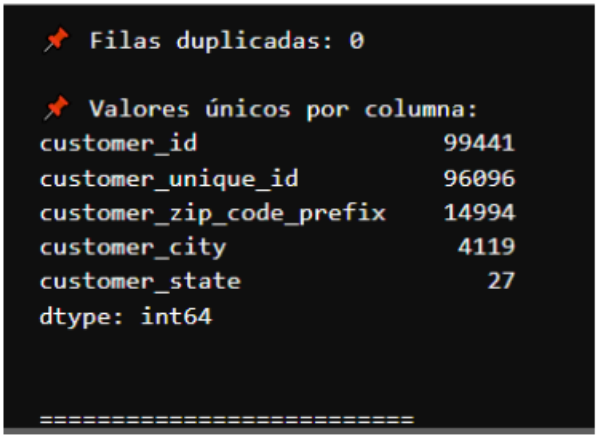
Conclusión

La tabla presenta:

- Alta calidad de datos
- Sin valores nulos
- Sin duplicados
- Tipos de datos correctos.

Por lo tanto, **no se requirió limpieza**, solo una **selección y renombrado de columnas** para su uso en la base de datos analítica.





Análisis de la tabla Orders

Descripción general

La tabla **Orders** contiene la información transaccional principal de los pedidos realizados en la plataforma de e-commerce.
Cuenta con **99.441 registros** y **8 columnas**, representando pedidos únicos asociados a clientes.

Estructura de la tabla

- **Filas:** 99.441
- **Columnas:** 8

Cada fila corresponde a **un pedido único**, identificado por **order_id**.

Tipos de datos

Todas las columnas relacionadas con fechas y tiempos se encuentran originalmente como **object (string)**.

Decisión de limpieza:
Todas las columnas de fechas fueron convertidas a tipo **datetime** para permitir:

- análisis temporales,
- cálculos de duración,
- agregaciones por fecha (día, mes, año).

Análisis de valores nulos

Interpretación clave:
Los valores nulos no representan errores, sino **eventos del negocio**:

- pedidos cancelados,
- pedidos aún en proceso,
- pedidos devueltos o no entregados.

Decisión de limpieza:
No se eliminaron filas
Los valores nulos se conservaron como **NULL**
Se utilizó **errors="coerce"** al convertir fechas para respetar estos estados

Eliminar estos registros hubiese implicado **perder información relevante del proceso de negocio**.

Análisis de duplicados

- **Filas duplicadas:** 0

Decisión:
No se realizaron eliminaciones ni consolidaciones.

Análisis de valores únicos

Interpretación:
El número reducido de valores únicos en **order_status** y fechas estimadas facilita:

- Segmentación por estado,
- Análisis de performance logística.

Decisiones para el modelo analítico

- La tabla **Orders** se utiliza como **fuente temporal y de estado**, no como tabla de hechos principal.
- Se mantiene **order_id** como clave para relacionar con:
    - ítems del pedido,
    - pagos,
    - reseñas.
- Las fechas se usarán para construir una **dimensión de tiempo (dim_date)**.

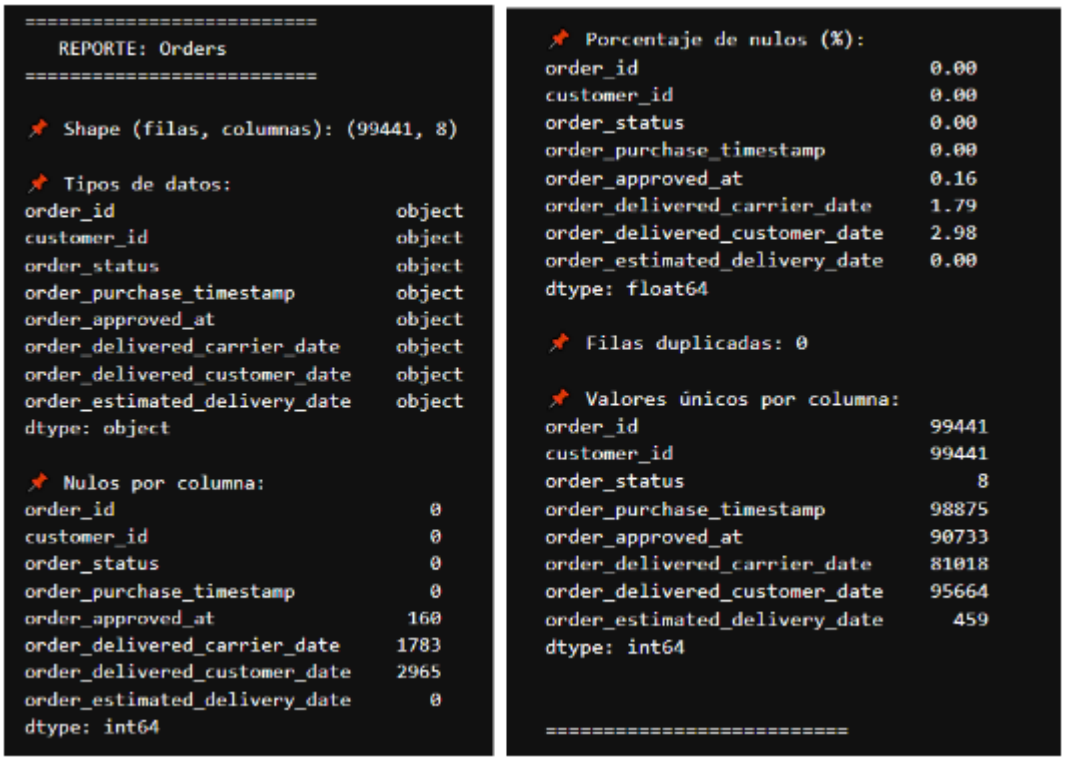
Conclusión

La tabla **Orders** presenta:

- baja proporción de valores nulos,
- nulos explicables por lógica de negocio,
- estructura consistente.

Resultado:
Se conserva completa, se transforma el tipo de datos y se utiliza como base para:

- análisis de procesos,
- construcción del modelo estrella,
- métricas temporales.



Análisis de la tabla Order Items

Descripción general

La tabla **Order Items** contiene el **detalle de los productos vendidos en cada pedido**. Cuenta con **112.650 registros** y **7 columnas**, representando la relación **pedido–producto–vendedor**.

A diferencia de la tabla *Orders*, aquí cada fila corresponde a **un ítem dentro de un pedido**, lo que explica que tenga **más filas que pedidos**.

Estructura de la tabla

- **Filas:** 112.650
- **Columnas:** 7
  Un pedido puede contener **uno o más ítems**, identificados por order_item_id.

Tipos de datos

Decisión de limpieza:
La columna **shipping_limit_date** fue convertida a **datetime** para permitir análisis logísticos y temporales.
Análisis de valores nulos

Todas las columnas presentan **0 valores nulos**.

Interpretación:
Los datos de ítems vendidos están completos, lo cual es fundamental para métricas financieras.

Decisión:
No fue necesario aplicar imputaciones ni eliminaciones.

Análisis de duplicados

- **Filas duplicadas:** 0

Decisión:
No se detectaron duplicados, por lo tanto no se realizaron ajustes.

Análisis de valores únicos

Interpretación clave:
La presencia de múltiples **order_item_id** por **order_id** confirma que esta tabla representa una **relación uno a muchos** entre pedidos e ítems.
Rol en el modelo analítico

La tabla **Order Items** es utilizada como la **tabla de hechos principal (fact_order_items)** porque:

- Contiene las **métricas numéricas**:
  - **price**
  - **freight_value**
- Permite agregaciones:
  - ventas totales,
  - ingresos por producto,
  - ingresos por vendedor,
  - costos de envío.
- Se relaciona directamente con todas las dimensiones:
  - pedidos,
  - productos,
  - clientes,
  - vendedores,
  - tiempo.

Decisiones para el modelo estrella

- Se preservaron todas las filas.
- Se mantuvieron las claves:
    - order_id
    - product_id
    - seller_id
- Se usará esta tabla como **hecho central** del modelo estrella.

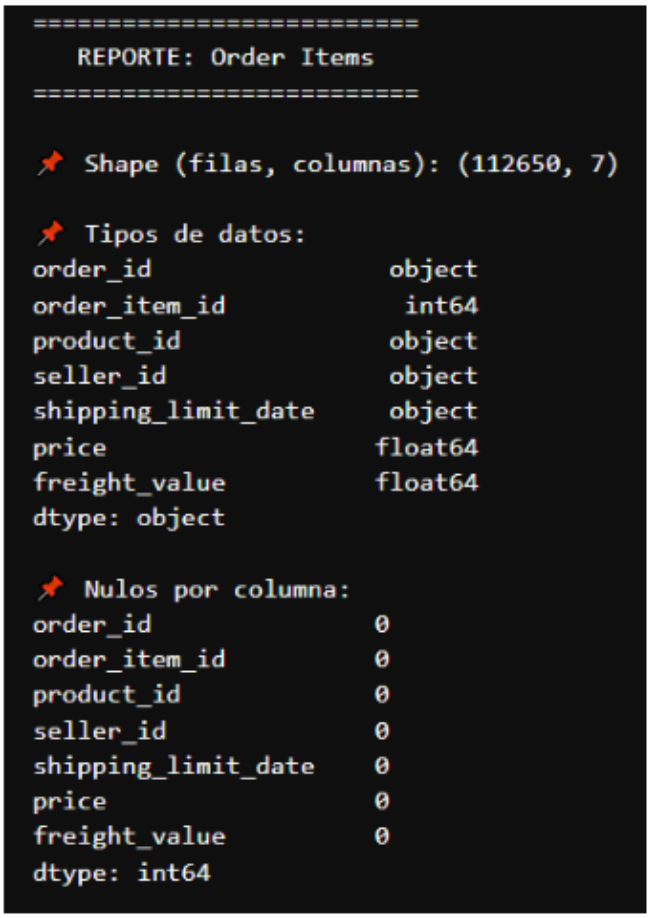
Resultado:
Order Items → fact_order_items

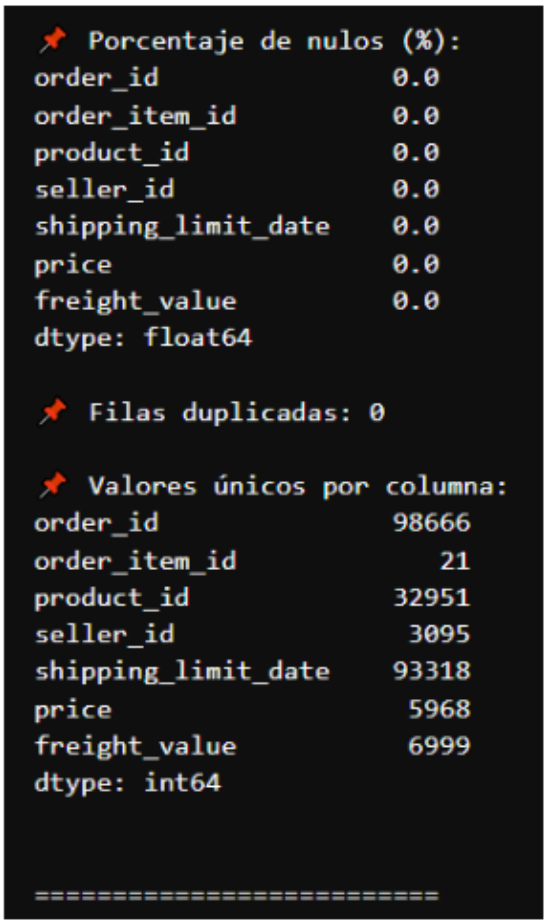
Conclusión

La tabla **Order Items** presenta:

- datos completos,
- estructura consistente,
- métricas financieras confiables.

Por estas razones, **no requirió limpieza profunda** y fue seleccionada como el **núcleo del análisis analítico**.





Análisis de la tabla Products

Descripción general

La tabla **Products** contiene la información descriptiva y física de los productos comercializados en la plataforma.
Cuenta con **32.951 registros** y **9 columnas**, donde cada fila representa **un producto único**.

Estructura de la tabla

- **Filas:** 32.951
- **Columnas:** 9

La granularidad es a nivel **producto**, lo que la convierte en una tabla candidata natural para una **dimensión del modelo estrella**.

Tipos de datos

Decisión de limpieza:
Las columnas numéricas se mantienen como `float64`, ya que pueden contener valores decimales o nulos.

Análisis de valores nulos

Interpretación clave:
Los nulos se concentran principalmente en **metadatos descriptivos**, no en identificadores ni claves de negocio.

Decisiones de limpieza sobre nulos

Categoría del producto

- Los valores nulos en **product_category_name** se imputaron como **"unknown"** o se mantuvieron como **NULL** (según necesidad analítica).
- Esto evita perder productos en análisis agregados por categoría.

Campos descriptivos y métricos

- Las columnas relacionadas con longitud de texto y cantidad de fotos se mantuvieron con nulos, ya que:
    - no afectan cálculos financieros,
    - representan ausencia real de información.

Dimensiones físicas

- Los pocos nulos (2 registros) se mantuvieron como **NULL** para no introducir valores artificiales.

Análisis de duplicados

- **Filas duplicadas:** 0

Decisión:
No se realizaron eliminaciones ni consolidaciones.

Análisis de valores únicos

Interpretación:
La diversidad de categorías y dimensiones permite realizar análisis por:

- tipo de producto,
- características físicas,
- logística y envío.

Rol en el modelo analitico

La tabla **Products** se utiliza como **dimensión de producto (dim_product)** en el modelo estrella.

Se relaciona directamente con:

- **fact_order_items** mediante **product_id.**

Permite análisis como:

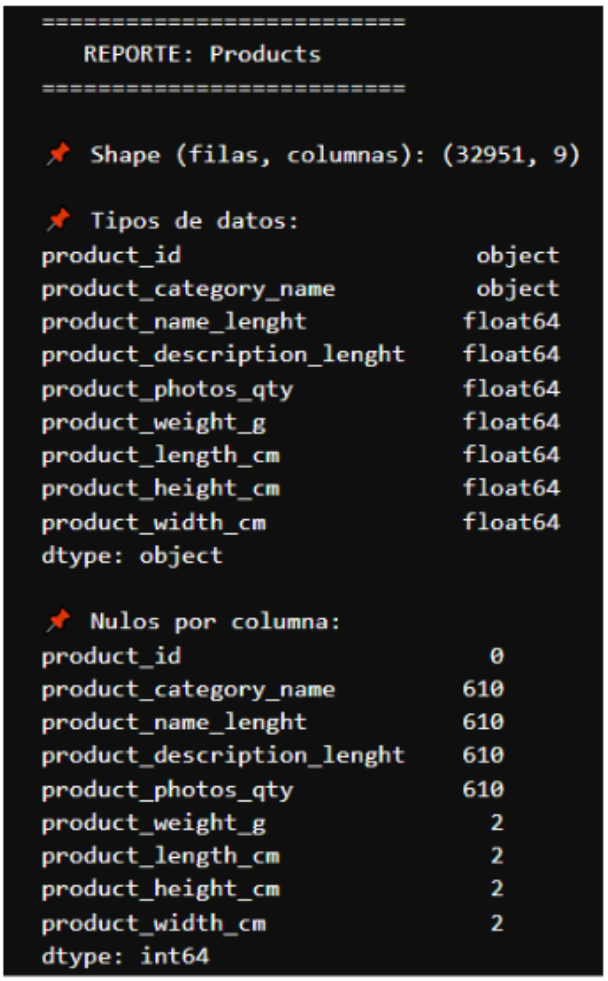
- ventas por categoría,
- ingresos por tipo de producto,
- impacto del peso o tamaño en costos logísticos.

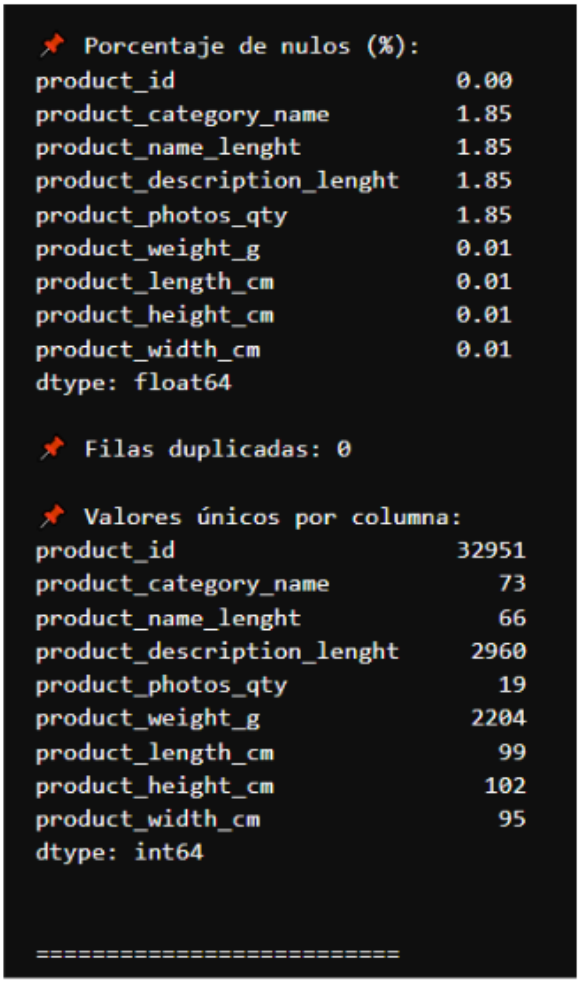
Conclusión

La tabla **Products** presenta:

- una baja proporción de valores nulos,
- nulos concentrados en atributos no críticos,
- estructura adecuada para análisis dimensional.

Resultado:
Se incorpora al modelo estrella como **dimensión**, aplicando imputaciones mínimas y conservando la integridad de los datos.





Análisis de la tabla Sellers

Descripción general

La tabla **Sellers** contiene la información básica de los **vendedores** que operan en la plataforma de e-commerce.
Cuenta con **3.095 registros** y **4 columnas**, donde cada fila representa **un vendedor único**.

Estructura de la tabla

- **Filas:** 3.095
- **Columnas:** 4

La granularidad es a nivel **vendedor**, lo que la convierte en una tabla adecuada para ser utilizada como **dimensión** dentro del modelo analítico.

Tipos de datos

Decisión:
Los tipos de datos son correctos y no requirieron transformaciones.

Análisis de valores nulos

Todas las columnas presentan **0 valores nulos**.

Interpretación:
La información de los vendedores está completamente informada, lo cual facilita análisis regionales sin necesidad de imputaciones.

Decisión de limpieza:
No fue necesario realizar ninguna acción correctiva.

Análisis de duplicados

- **Filas duplicadas:** 0

Decisión:
No se detectaron duplicados, por lo tanto no se realizaron eliminaciones ni consolidaciones.

Análisis de valores únicos

Interpretación:
Los vendedores se encuentran distribuidos en una amplia variedad de regiones, lo que permite:

- análisis de ventas por región,
- evaluación de desempeño por estado o ciudad.

Rol en el modelo analítico

La tabla **Sellers** se utiliza como **dimensión de vendedores (dim_seller)** en el modelo estrella.

Se relaciona directamente con:

- **fact_order_items** mediante **seller_id**.

Permite análisis como:

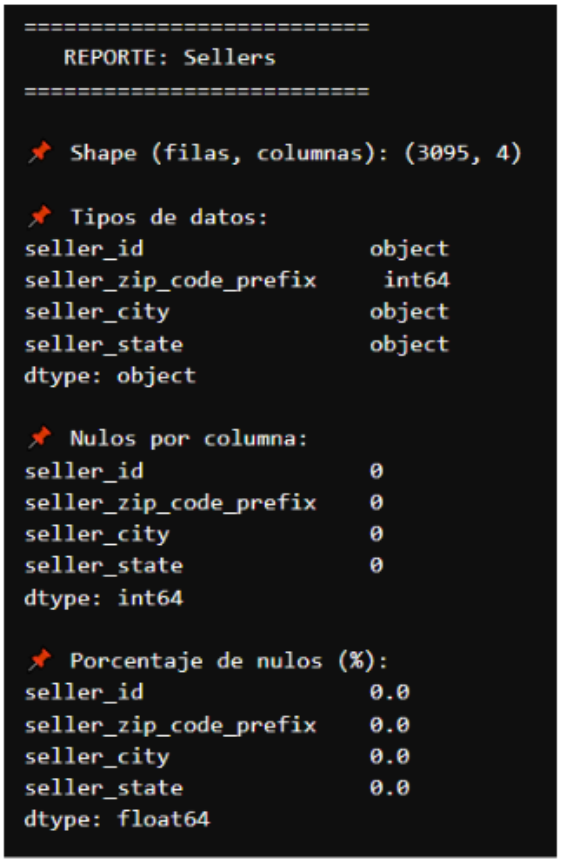
- ingresos por vendedor,
- número de productos vendidos por región,
- desempeño logístico por ubicación del vendedor.

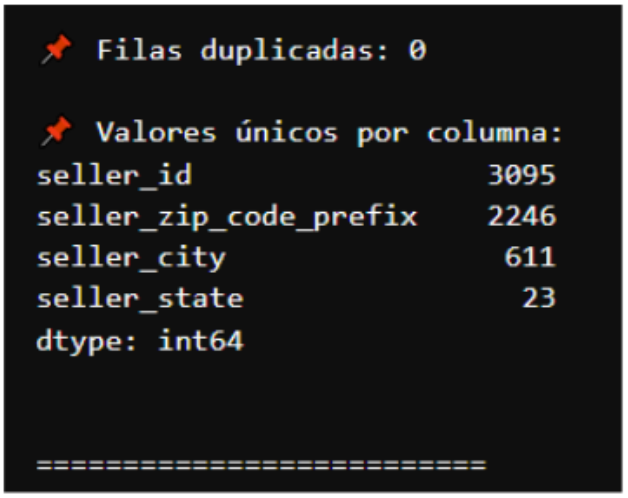
Conclusión

La tabla **Sellers** presenta:

- datos completos,
- estructura consistente,
- correcta definición de claves.

Resultado:
Se incorpora al modelo analítico **sin modificaciones**, funcionando como una dimensión estable y confiable.





Análisis de la tabla Payments

Descripción general

La tabla **Payments** contiene la información relacionada con los **pagos realizados por los pedidos** en la plataforma de e-commerce.
Cuenta con **103.886 registros** y **5 columnas**, donde cada fila representa **un pago asociado a un pedido**.

Es importante destacar que **un pedido puede tener más de un pago**, lo que explica que el número de filas sea mayor al de la tabla *Orders*.

Estructura de la tabla

- **Filas:** 103.886
- **Columnas:** 5

La relación entre pedidos y pagos es **uno a muchos**.

Tipos de datos

Decisión:
Los tipos de datos son correctos para el análisis financiero y no requirieron conversiones.
Análisis de valores nulos

Todas las columnas presentan **0 valores nulos**.

Interpretación:
La información de pagos está completamente informada, lo cual es esencial para análisis monetarios confiables.

Decisión de limpieza:
No fue necesario realizar imputaciones ni eliminaciones.

Análisis de duplicados

- **Filas duplicadas:** 0

Decisión:
No se detectaron duplicados.

Análisis de valores únicos

Interpretación clave:
La diversidad en métodos de pago y planes de cuotas permite analizar:

- preferencias de pago de los clientes,
- impacto de las cuotas en el valor de compra.

Rol en el modelo analítico

La tabla **Payments** puede utilizarse de dos formas:

Opción 1 – Tabla de hechos auxiliar

- Para análisis específicos de pagos.
- Métricas: monto total pagado, cuotas promedio, distribución por método.

Opción 2 – Fuente para enriquecer la tabla de hechos principal

- Agregando el monto total pagado por pedido a **fact_order_items**.

En este proyecto, se utiliza principalmente como **fuente de métricas financieras complementarias**.

Decisiones para el modelo estrella

- Se mantiene la tabla completa.
  Se preserva **order_id** como clave de la relación.
- No se agregan ni eliminan registros.

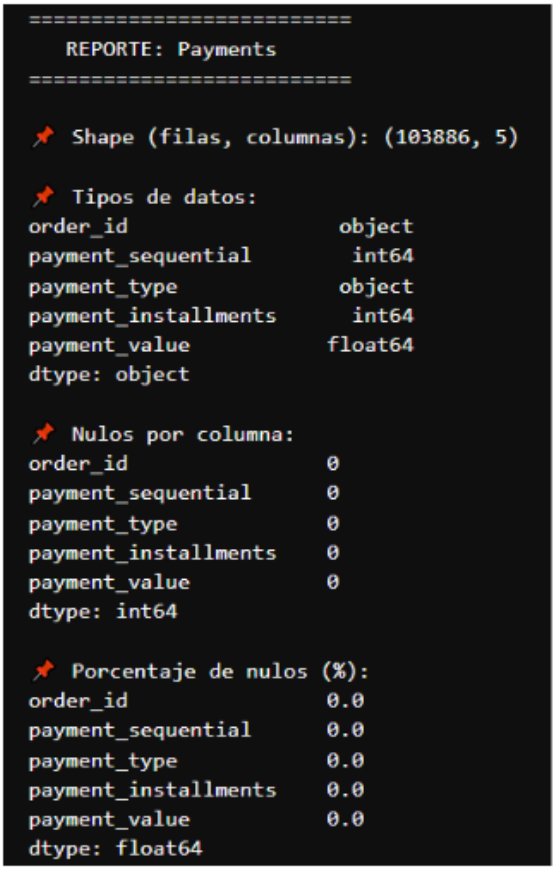
Resultado: La tabla se integra al modelo analítico sin modificaciones estructurales.

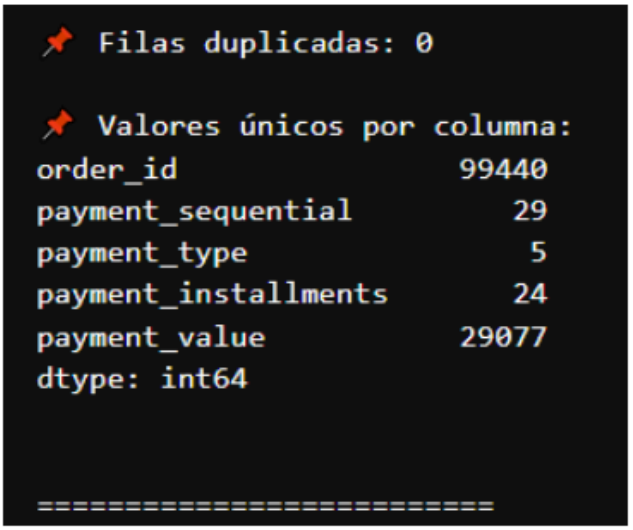
Conclusión

La tabla **Payments** presenta:

- alta calidad de datos,
- métricas financieras completas,
- estructura adecuada para análisis monetarios.

Por lo tanto, **no requirió limpieza** y se utiliza directamente para el análisis financiero.





Análisis de la tabla Reviews

Descripción general

La tabla **Reviews** contiene la información relacionada con las **reseñas y calificaciones realizadas por los clientes** luego de completar un pedido.
Cuenta con **99.224 registros** y **7 columnas**, donde cada fila representa una reseña asociada a un pedido.

Estructura de la tabla

- **Filas:** 99.224
- **Columnas:** 7

Cada registro se identifica mediante **review_id** y se relaciona con un pedido a través de **order_id**.

Tipos de datos

Decisión de limpieza:
Las columnas de fecha fueron convertidas a tipo **datetime** para permitir análisis temporales y de tiempos de respuesta.

Análisis de valores nulos

Interpretación clave:
Una gran proporción de clientes **califica con puntaje sin dejar comentario escrito**, lo cual es un comportamiento típico en plataformas de e-commerce.

Los valores nulos **no representan errores**, sino **ausencia voluntaria de texto**.

Decisiones de limpieza sobre nulos

Campos de texto (review_comment_title, review_comment_message)

- Los valores nulos se mantuvieron como `NULL`.
- No se imputaron valores artificiales (por ejemplo, “sin comentario”), ya que:
    - no aportan valor analítico,
    - podrían distorsionar análisis de texto o sentimiento.

Campos críticos (review_score, order_id)

- No presentan nulos, por lo tanto se conservaron sin cambios.

Análisis de duplicados

- **Filas duplicadas:** 0

Decisión:
No se detectaron duplicados, por lo tanto no se realizaron ajustes.

Análisis de valores únicos

Interpretación:
La mayoría de los pedidos tiene una sola reseña asociada, lo cual permite analizar la satisfacción por pedido sin ambigüedad.
Rol en el modelo analítico

La tabla **Reviews** se utiliza como:

Tabla auxiliar de análisis de satisfacción, no como tabla de hechos principal.

Permite:

- calcular promedios de calificación,
- analizar satisfacción por producto, vendedor o región,
- evaluar tiempos de respuesta del soporte.

Se relaciona con:

- **orders** mediante **order_id**,
- indirectamente con productos y vendedores.

Decisiones para el modelo estrella

- Se preserva la tabla completa.
- Se mantienen los valores nulos en campos de texto.
- Se utiliza principalmente para **KPIs de calidad y experiencia del cliente**.

Resultado:
La tabla se integra al modelo analítico sin eliminar registros ni alterar su significado original.

Conclusión

La tabla **Reviews** presenta:

- alta proporción de nulos en campos no críticos,
- datos completos en métricas clave,
- información valiosa para análisis cualitativos.

Por lo tanto, **no se realizó limpieza destructiva**, respetando el comportamiento real de los usuarios.

```
==========================
   REPORTE: Reviews
==========================

📌 Shape (filas, columnas): (99224, 7)

📌 Tipos de datos:
review_id                 object
order_id                  object
review_score               int64
review_comment_title      object
review_comment_message    object
review_creation_date      object
review_answer_timestamp   object
dtype: object

📌 Nulos por columna:
review_id                     0
order_id                      0
review_score                  0
review_comment_title      87656
review_comment_message    58247
review_creation_date          0
review_answer_timestamp       0
dtype: int64
```

```
📌 Porcentaje de nulos (%):
review_id                  0.00
order_id                   0.00
review_score               0.00
review_comment_title      88.34
review_comment_message    58.70
review_creation_date       0.00
review_answer_timestamp    0.00
dtype: float64

📌 Filas duplicadas: 0

📌 Valores únicos por columna:
review_id                 98410
order_id                  98673
review_score                  5
review_comment_title       4527
review_comment_message    36159
review_creation_date        636
review_answer_timestamp   98248
dtype: int64


==========================
```

Análisis de la tabla Geolocation

Descripción general

La tabla **Geolocation** contiene información geográfica asociada a los **códigos postales** de Brasil, incluyendo coordenadas y ubicación administrativa.
Cuenta con **1.000.163 registros** y **5 columnas**, siendo la tabla con **mayor volumen de datos** del dataset.

Estructura de la tabla

- **Filas:** 1.000.163
- **Columnas:** 5

A diferencia de otras tablas, esta **no representa entidades únicas**, sino múltiples registros geográficos para un mismo código postal.

Tipos de datos

Decisión:
Los tipos de datos son correctos y no requirieron transformaciones.

Análisis de valores nulos

Todas las columnas presentan **0 valores nulos**.

Interpretación:
La información geográfica está completamente informada en términos de campos obligatorios.

Decisión:
No se realizaron imputaciones ni eliminaciones por nulos.

Análisis de duplicados

- **Filas duplicadas:** 261.831

Interpretación clave:
La presencia de una gran cantidad de duplicados se debe a que:

- un mismo código postal puede tener múltiples coordenadas,
- existen registros repetidos de ciudad y estado para un mismo prefijo postal.

Esto **no es un error**, sino una característica del dataset original.

Análisis de valores únicos

Interpretación:
La tabla presenta **alta redundancia**, especialmente a nivel de coordenadas, lo que incrementa el volumen sin aportar valor analítico adicional.
Decisiones de limpieza

Problema identificado

- Exceso de registros duplicados.
- Gran volumen que afecta performance.
- Múltiples coordenadas para un mismo código postal.

Decisión tomada
Se realizó una **normalización y agregación** de la tabla, conservando **un solo registro por código postal**.

Criterio aplicado

- Agrupación por **geolocation_zip_code_prefix**.
- Selección de:
    - una ciudad y estado representativos,
    - el promedio de latitud y longitud.

Motivo
Este enfoque:

- reduce drásticamente el volumen de datos,
- mantiene la información geográfica relevante,
- facilita su uso en análisis regionales y mapas.

Rol en el modelo analítico

La tabla **Geolocation** se utiliza como:

Dimensión geográfica (dim_geolocation), relacionada con:

- clientes (**customer_zip_code_prefix**),
- vendedores (**seller_zip_code_prefi**x).

Permite análisis como:

- ventas por región,
- distribución geográfica de clientes y vendedores,
- visualizaciones en mapas.

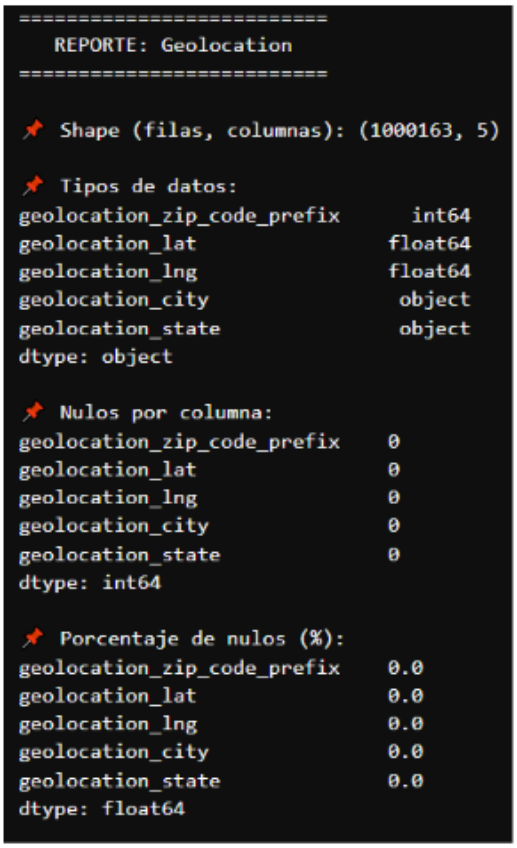
Conclusión

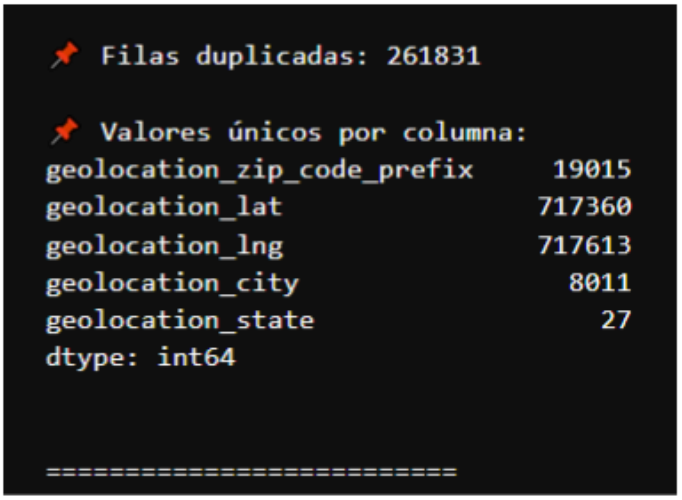
La tabla **Geolocation** presenta:

- datos completos,
- alta redundancia,

- gran volumen innecesario para análisis analítico.

Resultado:
Se realizó una **limpieza estructural no destructiva**, reduciendo duplicados y optimizando la tabla para su uso en el modelo estrella.





Análisis de la tabla Translation

Descripción general

La tabla **Translation** contiene la **traducción de las categorías de productos** del idioma original (portugués) al inglés.
Cuenta con **71 registros** y **2 columnas**, donde cada fila representa una categoría de producto y su equivalente en inglés.

Estructura de la tabla

- **Filas:** 71
- **Columnas:** 2

La tabla tiene una granularidad **uno a uno**: cada categoría en portugués tiene exactamente una traducción al inglés.

Tipos de datos

Decisión:
Los tipos de datos son correctos y no requirieron transformaciones.

Análisis de valores nulos: Todas las columnas presentan **0 valores nulos**.

Interpretación:
La tabla está completa y no presenta información faltante.

Decisión de limpieza:
No fue necesario aplicar imputaciones ni eliminaciones.

Análisis de duplicados

- **Filas duplicadas:** 0

Decisión:
No se detectaron duplicados, por lo tanto no se realizaron ajustes.

Análisis de valores únicos

Interpretación:
Existe una correspondencia exacta entre categorías y traducciones, lo que garantiza consistencia semántica.
Rol en el modelo analítico

La tabla **Translation** no funciona como una dimensión independiente, sino como una **tabla de enriquecimiento semántico**.

Se utiliza para:

- traducir las categorías de productos en reportes,
- mejorar la legibilidad de dashboards,
- facilitar la interpretación por usuarios no hispanohablantes/portugueses.

Se integra mediante:

- **product_category_name → dim_product**

Decisiones de uso

- No se eliminan ni modifican registros.
- Se utiliza como **lookup table** para enriquecer la dimensión de productos.
- Permite mantener los datos originales sin perder trazabilidad.

Conclusión

La tabla **Translation** presenta:

- estructura simple y completa,
- ausencia de nulos y duplicados,

- alto valor para la interpretación del negocio.

Resultado:
Se utiliza como tabla auxiliar para mejorar la calidad semántica del modelo analítico, sin necesidad de limpieza ni transformaciones.

```
========================
   REPORTE: Translation
========================

📌 Shape (filas, columnas): (71, 2)

📌 Tipos de datos:
product_category_name            object
product_category_name_english    object
dtype: object

📌 Nulos por columna:
product_category_name            0
product_category_name_english    0
dtype: int64

📌 Porcentaje de nulos (%):
product_category_name            0.0
product_category_name_english    0.0
dtype: float64

📌 Filas duplicadas: 0

📌 Valores únicos por columna:
product_category_name            71
product_category_name_english    71
dtype: int64


========================
```

3. Etapa de staging del dataset

Antes de construir la base de datos analítica y el modelo estrella, se implementó una **etapa de staging**.
Esta etapa funciona como una **zona intermedia** entre los datos crudos (CSV originales) y la base de datos analítica final.

El objetivo principal del staging es **preparar, estandarizar y validar los datos**, sin alterar aún su estructura de negocio.

¿Qué es la etapa de Staging?

La etapa de *staging* consiste en cargar los datos originales en una base de datos relacional (**PostgreSQL**) luego de realizar una **limpieza mínima y controlada**, conservando la granularidad y el significado original de cada tabla.

En esta fase:

- No se agregan métricas
- No se desnormalizan relaciones
- No se crean tablas de hechos ni dimensiones

¿Por qué se realizó una etapa de Staging?

Separación de responsabilidades

Dividir el proceso en etapas permite:

- aislar la limpieza de datos,
- evitar errores en la base analítica,
- facilitar la depuración y el control de calidad.

Tareas realizadas en la etapa de Staging

Durante esta etapa se realizaron las siguientes acciones:

Carga de datos desde archivos CSV

- Se importaron todas las tablas del dataset original.
- Cada tabla se cargó en el esquema staging de PostgreSQL.
- Se mantuvo una correspondencia uno a uno con los archivos originales.

Estandarización de tipos de datos

- Las columnas de fecha fueron convertidas a tipo datetime.
- Se utilizó errors="coerce" para manejar valores inválidos sin perder registros.
- Los valores NaT fueron almacenados como NULL en PostgreSQL.

Esta decisión permitió respetar la lógica del negocio (pedidos no aprobados o no entregados).

Control de valores nulos

- No se eliminaron filas por presencia de nulos.
- Los valores nulos se conservaron cuando representaban ausencia real de información.
- No se realizaron imputaciones artificiales en esta etapa.

Validación de duplicados

- Se verificó la existencia de filas duplicadas.
- Solo se trató el caso de la tabla **Geolocation**, donde la duplicación era estructural.
- En el resto de las tablas no se detectaron duplicados.

Creación de tablas de staging

Cada tabla limpia fue cargada en PostgreSQL utilizando una función genérica de carga (to_staging), lo que permitió:

- trazabilidad del proceso,
- consistencia en la carga,
- facilidad de mantenimiento.

Relación entre Staging y Base de Datos Analítica

La etapa de staging funciona como la **fuente directa** para la construcción de la base analítica.

Desde el staging:

- se seleccionan las columnas relevantes,
- se crean dimensiones y hechos,
- se realizan desnormalizaciones controladas,
- se construye el modelo estrella.

De esta manera, la base analítica se apoya sobre datos:

- limpios,
- consistentes,
- validados.

Este enfoque refleja una **buena práctica de ingeniería de datos**, asegurando que las decisiones analíticas posteriores se basen en datos confiables y correctamente estructurados.

4. Creación de la base de datos analítica

Luego de completar la etapa de *staging*, se procedió a la construcción de la **base de datos analítica**, cuyo objetivo es facilitar el análisis de la información, la generación de métricas y la visualización de indicadores clave del negocio.

A diferencia de la base transaccional y del staging, la base analítica está diseñada para:

- optimizar consultas,
- reducir la complejidad de los joins,
- facilitar el uso por herramientas de BI.

Para ello, se implementó un **modelo estrella (Star Schema)** utilizando Python y PostgreSQL.

¿Qué es una base de datos analítica?

Una base de datos analítica se caracteriza por:

- separar **hechos** y **dimensiones**,
- almacenar métricas numéricas en tablas de hechos,
- organizar los atributos descriptivos en tablas dimensionales.

Este tipo de diseño permite realizar análisis históricos, comparativos y agregados de forma eficiente.

Fuente de datos para la analítica

La base analítica se construyó **exclusivamente a partir del esquema staging**, garantizando que:

- los datos ya estén limpios y validados,
- no se dependan de archivos CSV,
- el proceso sea reproducible y escalable.

Las tablas de staging fueron leídas desde PostgreSQL utilizando Python y SQLAlchemy.

Diseño del modelo estrella

Tabla de hechos

Se definió como tabla de hechos principal:

fact_order_items

Motivo de la elección:

- representa el mayor nivel de granularidad del negocio (producto por pedido),
- permite calcular métricas económicas reales,
- evita problemas de duplicación de montos presentes en la tabla orders.

Métricas incluidas:

- precio del producto,
- valor del flete,
- total por ítem.

Dimensiones creadas

Dimensión Cliente (dim_customer)

Contiene información descriptiva del cliente:

- ubicación,
- identificadores únicos.

Se utiliza para análisis geográficos y segmentación de clientes.

Dimensión Producto (dim_product)

Incluye:

- categoría del producto,
- atributos físicos,
- traducción de categorías (enriquecida con la tabla translation).

Permite análisis por tipo de producto y categoría.

Dimensión Vendedor (dim_seller)

Describe a los vendedores:

- ubicación geográfica,
- identificación única.

Facilita análisis de desempeño por vendedor o región.

Dimensión Tiempo (dim_date)

Se creó a partir de las fechas de compra:

- año,
- mes,
- día,
- día de la semana.

Esta dimensión es clave para análisis temporales y tendencias.

Proceso de desnormalización

Durante esta etapa se realizaron:

- joins controlados entre tablas de staging,
- selección de columnas relevantes,
- eliminación de dependencias innecesarias.

La desnormalización se realizó **solo en la base analítica**, nunca en staging, para preservar la trazabilidad del dato.

Manejo de valores nulos en la analítica

Los valores nulos provenientes del staging:

- se mantuvieron como NULL,
- no fueron imputados artificialmente,
- se respetaron como eventos reales del negocio (por ejemplo, pedidos no entregados).

Esto asegura que las métricas calculadas no estén distorsionadas.

Carga de datos en PostgreSQL

Las tablas analíticas fueron cargadas en un esquema separado (analytics) utilizando funciones genéricas de carga desde Python.

Este enfoque permitió:

- reutilizar el mismo engine de conexión,
- asegurar consistencia entre tablas,
- automatizar el proceso de carga.

Beneficios del modelo analítico implementado

Consultas más simples
Mejor performance
Claridad conceptual
Compatibilidad con herramientas BI (Metabase, Power BI, etc.)
Separación clara entre datos operacionales y analíticos

La creación de la base de datos analítica permitió transformar datos transaccionales complejos en una estructura optimizada para el análisis.

Este enfoque replica prácticas utilizadas en entornos profesionales de ingeniería y análisis de datos.

5. Visualización de datos y Dashboards

Una vez construida la base de datos analítica y definido el modelo estrella, se procedió a la etapa final del proyecto: la visualización de métricas y el análisis exploratorio a través de dashboards.

El objetivo de esta etapa es transformar los indicadores calculados en información visual clara, permitiendo interpretar rápidamente el comportamiento del negocio, detectar patrones y comunicar resultados de forma efectiva.

Para ello, se utilizó **Metabase** como herramienta de Business Intelligence, conectada directamente al esquema `analytics` de PostgreSQL. Esta elección permitió aprovechar un entorno visual intuitivo, ideal para análisis interactivos, filtros dinámicos y visualizaciones orientadas a usuarios de negocio.

Tendencias de Ventas

El sector de **Tendencias** del dashboard tiene como objetivo analizar la evolución temporal de las ventas, permitiendo identificar patrones de crecimiento, estacionalidad y variaciones en la demanda a lo largo del tiempo. Este análisis resulta clave para comprender el comportamiento del negocio, detectar períodos de alto y bajo rendimiento y apoyar la toma de decisiones estratégicas.

Ventas por Categoría de Producto

El gráfico presenta la **evolución de las ventas totales por categoría de producto** a lo largo del tiempo, considerando un período específico seleccionado mediante filtros dinámicos de fecha.

En la visualización se observa un **crecimiento sostenido de las ventas durante el año 2017**, alcanzando un pico significativo hacia el último trimestre del año. Posteriormente, se registra una **caída abrupta**, seguida de una recuperación moderada durante 2018, con fluctuaciones que indican un comportamiento estacional de la demanda.

Este análisis permite identificar:

- Picos de ventas en períodos específicos
- Momentos de contracción o desaceleración
- Tendencias generales de crecimiento o descenso por categoría

La métrica resulta fundamental para anticipar comportamientos futuros, planificar campañas comerciales y optimizar la gestión del catálogo de productos.

Ventas por mes en un determinado año

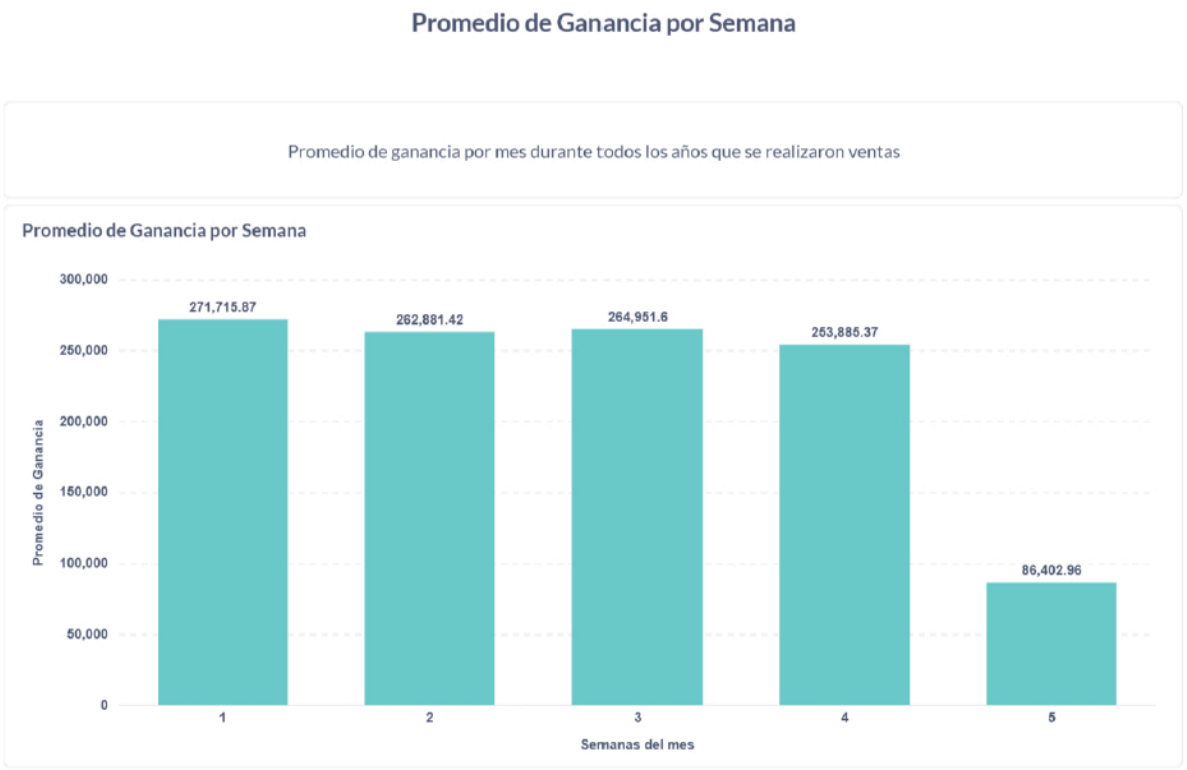
Este gráfico presenta la **evolución mensual de las ventas** para el año seleccionado (2018), permitiendo analizar el comportamiento del negocio a lo largo del tiempo e identificar patrones de crecimiento, estabilidad o estacionalidad.

Durante los primeros meses del año se observa un **nivel de ventas elevado y relativamente estable**, con valores cercanos al **millón de unidades monetarias mensuales**, alcanzando picos en los meses de **marzo, abril y mayo**. A partir de junio se registra una **leve desaceleración**, aunque las ventas se mantienen en niveles similares hasta agosto.

En el mes de **septiembre** se evidencia una **caída abrupta del volumen de ventas**, lo que sugiere un período con datos parciales o un cierre anticipado del registro para ese año. Este comportamiento refuerza la importancia de considerar el contexto temporal y la completitud de los datos al interpretar tendencias.

En conjunto, el gráfico permite concluir que el negocio presenta una **fuerte consistencia mensual**, con picos bien definidos en el primer semestre y una dinámica que facilita la identificación de períodos clave para la planificación comercial y estratégica.

Promedio de Ganancia por Semana



El gráfico muestra el **promedio de ganancia por semana del mes**, calculado a lo largo de todos los años en los que se registraron ventas. Esta visualización permite analizar la distribución temporal de la rentabilidad dentro de cada mes y detectar patrones recurrentes en el comportamiento del negocio.

Se observa que las **primeras cuatro semanas del mes presentan niveles de ganancia promedio relativamente estables**, con valores que oscilan entre aproximadamente **253.000 y 272.000**. En particular, la **primera semana** se destaca como la de mayor ganancia promedio, seguida muy de cerca por la segunda y la tercera semana, lo que indica una concentración de la rentabilidad en la primera parte del mes.

En contraste, la **quinta semana** muestra una caída significativa en el promedio de ganancia, con un valor considerablemente menor. Este comportamiento puede explicarse por el hecho de que no todos los meses cuentan con una quinta semana completa, lo que reduce la cantidad de días y transacciones consideradas en el cálculo.

Este análisis permite identificar una **tendencia clara de mayor estabilidad y rendimiento durante las primeras semanas del mes**, información relevante para la planificación

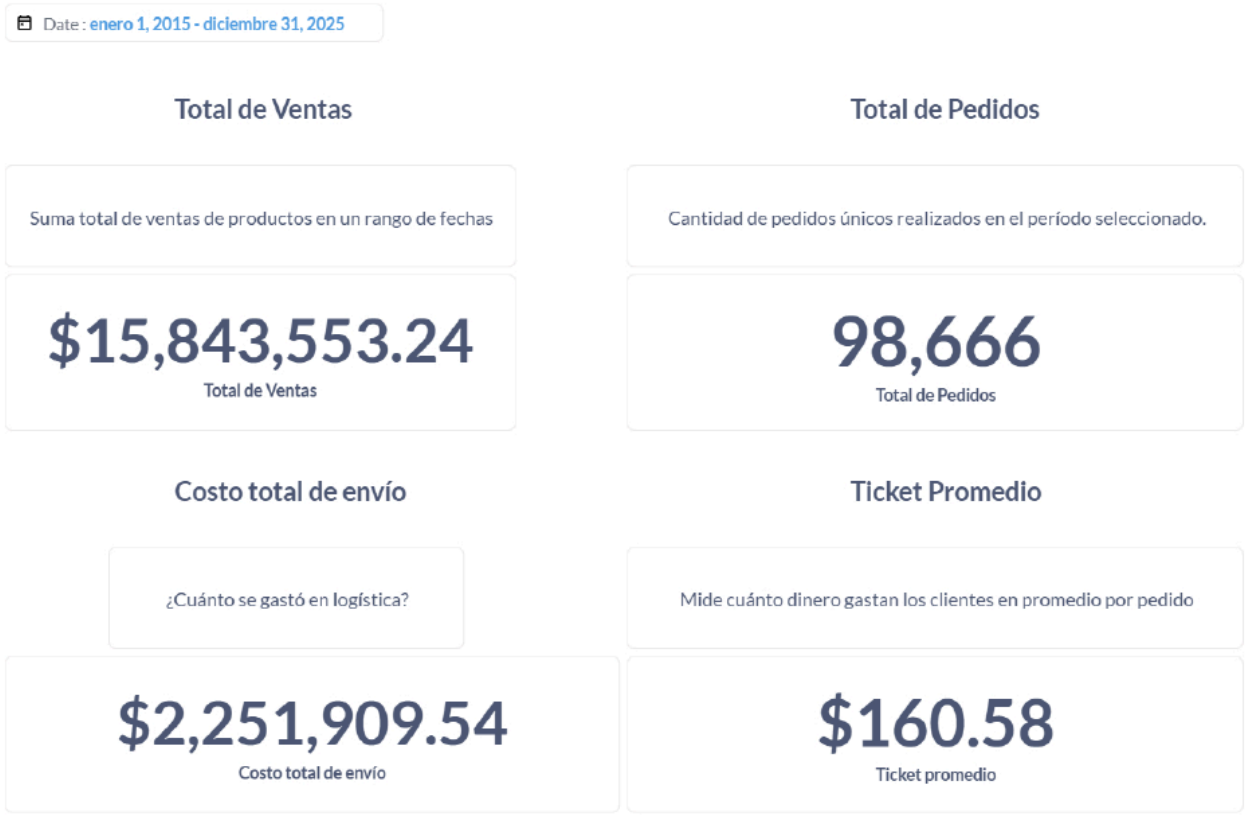
operativa, la asignación de campañas comerciales y la optimización de recursos en función de los períodos con mayor impacto económico.

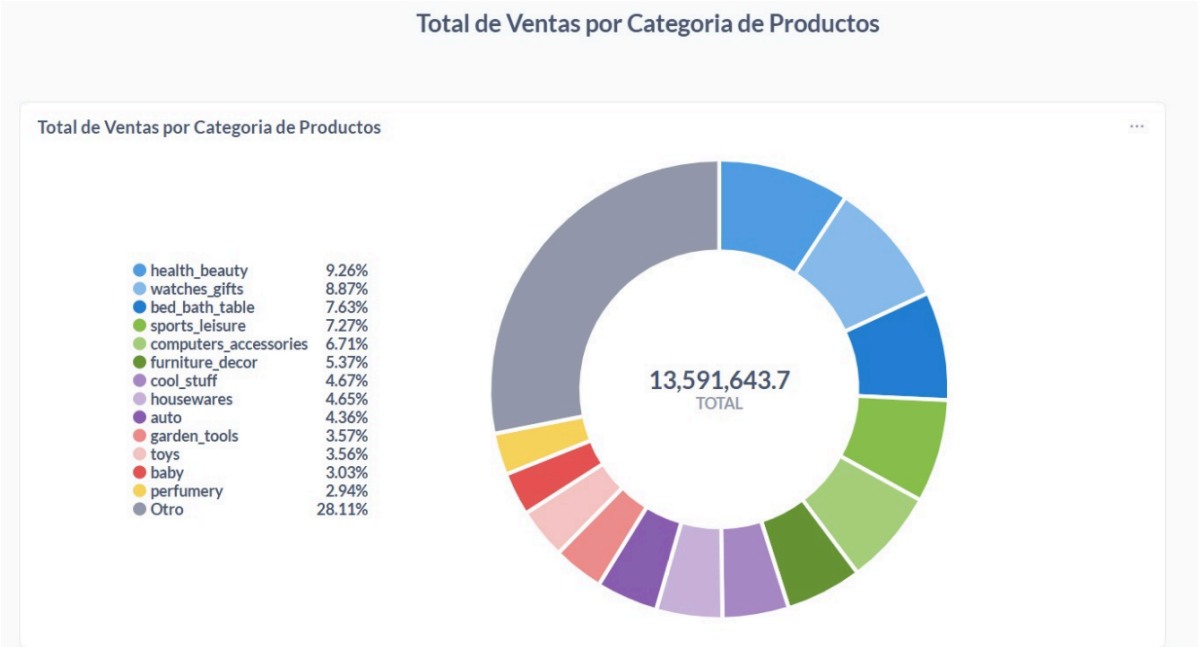
KPIs (Indicadores Clave de Desempeño)

El sector de **KPIs** del dashboard presenta una visión general y sintética del desempeño del negocio de e-commerce. Su objetivo es brindar, de forma inmediata, los indicadores más relevantes para evaluar la **magnitud de la operación**, el **volumen de actividad comercial** y la **evolución global de las ventas** dentro del período de tiempo seleccionado.

Estos indicadores permiten comprender rápidamente el estado del negocio y funcionan como punto de partida para análisis más detallados en los sectores de ranking, tendencias y segmentación.

Descripción de los indicadores





Total de Ventas

Representa la **suma total del valor monetario de las ventas de productos** realizadas dentro del rango de fechas seleccionado. Este indicador refleja el ingreso bruto generado por la plataforma de e-commerce y permite evaluar el crecimiento o contracción del negocio a lo largo del tiempo.

Total de Pedidos

Corresponde a la **cantidad de pedidos únicos** efectuados en el período analizado. Este KPI mide el nivel de actividad comercial y la demanda de los clientes, siendo fundamental para analizar el volumen operativo independientemente del monto facturado.

Costo Total de Envío

Este indicador muestra el **monto total invertido en logística y envíos** durante el período analizado. El valor obtenido refleja el impacto directo de los costos logísticos sobre la operación del negocio, siendo un factor clave a considerar en la rentabilidad general. Su análisis permite evaluar oportunidades de optimización en procesos de envío, negociación con transportistas o estrategias de reducción de costos.

Ticket Promedio

El **ticket promedio** representa el **gasto medio realizado por los clientes en cada pedido**. Este KPI es fundamental para comprender el comportamiento de compra y el valor promedio de las transacciones. Un ticket promedio elevado suele asociarse a estrategias efectivas de venta cruzada, bundles o productos de mayor valor, mientras que su evolución en el tiempo permite detectar cambios en los hábitos de consumo.

Total de Ventas por Categoría de Productos

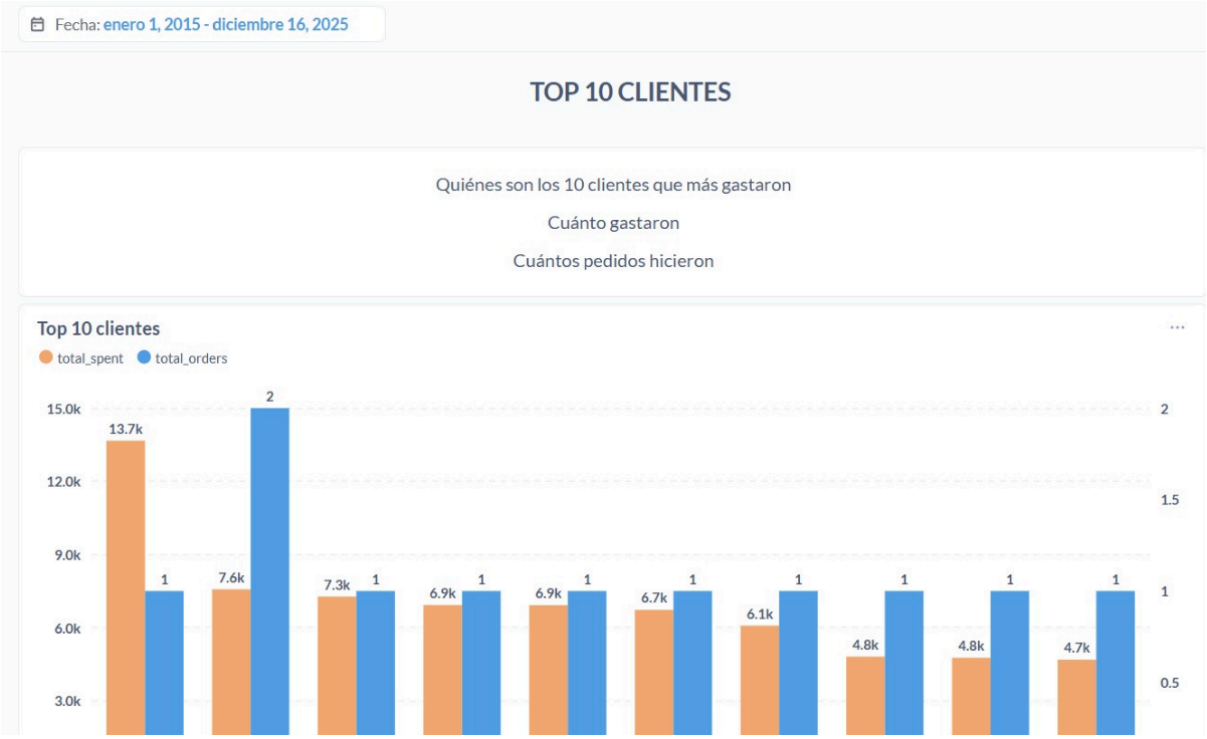
El gráfico de distribución muestra cómo se reparte el **total de ventas entre las distintas categorías de productos**. Se observa que algunas categorías concentran una mayor proporción del revenue, destacándose como los principales motores del negocio. Este análisis permite identificar categorías estratégicas, evaluar su peso relativo dentro del total de ventas y orientar decisiones comerciales, promocionales y de gestión del catálogo.

En conjunto, estos KPIs ofrecen una **visión ejecutiva del rendimiento general**, permitiendo identificar rápidamente períodos de alto desempeño, comparar resultados entre distintos rangos temporales y servir como base para el análisis de métricas más específicas presentadas en el resto del dashboard.

Ranking

El sector **Ranking** tiene como objetivo identificar a los **principales actores del negocio**, tanto desde el punto de vista de los clientes como de los productos. A través de estos análisis se busca reconocer patrones de concentración, detectar clientes de alto valor y comprender qué productos impulsan la mayor parte de las ventas. Esta información resulta clave para la toma de decisiones comerciales, estrategias de fidelización y optimización del catálogo.

Top 10 Clientes



El gráfico presenta el **Top 10 de clientes con mayor gasto total** durante el período seleccionado. Para cada cliente se muestran dos métricas clave:

- **Monto total gastado**, que refleja su aporte directo al revenue.
- **Cantidad de pedidos realizados**, que permite evaluar su nivel de recurrencia.

El análisis evidencia que algunos clientes combinan **alto gasto con una elevada cantidad de pedidos**, lo que los posiciona como clientes estratégicos y recurrentes. En otros casos, se observan clientes con **menor frecuencia de compra pero con tickets más altos**, lo que sugiere distintos perfiles de consumo. Este ranking permite identificar oportunidades para acciones de fidelización, segmentación y programas de beneficios orientados a los clientes de mayor valor.

Top 10 Productos

El segundo gráfico muestra el **Top 10 de productos más vendidos**, medidos por la **cantidad de unidades vendidas**. Se observa una clara concentración de ventas en determinadas categorías, destacándose productos del rubro **hogar, bienestar y tecnología**, que lideran el volumen total de ventas.

Este ranking permite identificar los **productos más demandados** del marketplace, facilitando decisiones relacionadas con stock, promociones y estrategias de pricing. Asimismo, ayuda a detectar categorías clave que impulsan el negocio y que pueden ser potenciadas para maximizar el rendimiento comercial.

Meses con Mayor Ganancia

El gráfico presenta el **Top 2 de los meses con mayor ganancia total** para el año 2017, permitiendo identificar los períodos de mayor rentabilidad dentro del negocio. En la visualización se observa que **noviembre** se posiciona como el mes con mayor ganancia, alcanzando un valor aproximado de **1.010.271**, lo que lo convierte en el período más rentable del año.

En segundo lugar se encuentra **diciembre**, con una ganancia cercana a **743.914**, manteniéndose también como uno de los meses con mejor desempeño económico, aunque con un valor inferior respecto a noviembre.

Este ranking evidencia una **fuerte concentración de ganancias en los últimos meses del año**, lo que sugiere un comportamiento estacional del negocio, asociado a eventos comerciales y festividades como el Black Friday y las celebraciones de fin de año. La información resulta clave para comprender qué períodos impulsan la rentabilidad total y para apoyar decisiones estratégicas relacionadas con planificación comercial, promociones y asignación de recursos en los momentos de mayor impacto económico.

Ventas por Categoría – Caso: Christmas Supplies

El gráfico muestra la evolución mensual de las ventas correspondientes a la categoría *Christmas Supplies* durante el año 2017. A lo largo de los primeros meses analizados (abril a agosto), se observa un volumen de ventas bajo y relativamente estable, lo que indica una demanda limitada fuera de la temporada festiva.

A partir del mes de septiembre se registra un crecimiento progresivo, con un aumento significativo en octubre y un pico marcado en noviembre, mes en el cual las ventas alcanzan su valor máximo. Este comportamiento refleja claramente un patrón de estacionalidad asociado a las festividades de fin de año, donde los consumidores incrementan sus compras de productos vinculados a la Navidad.

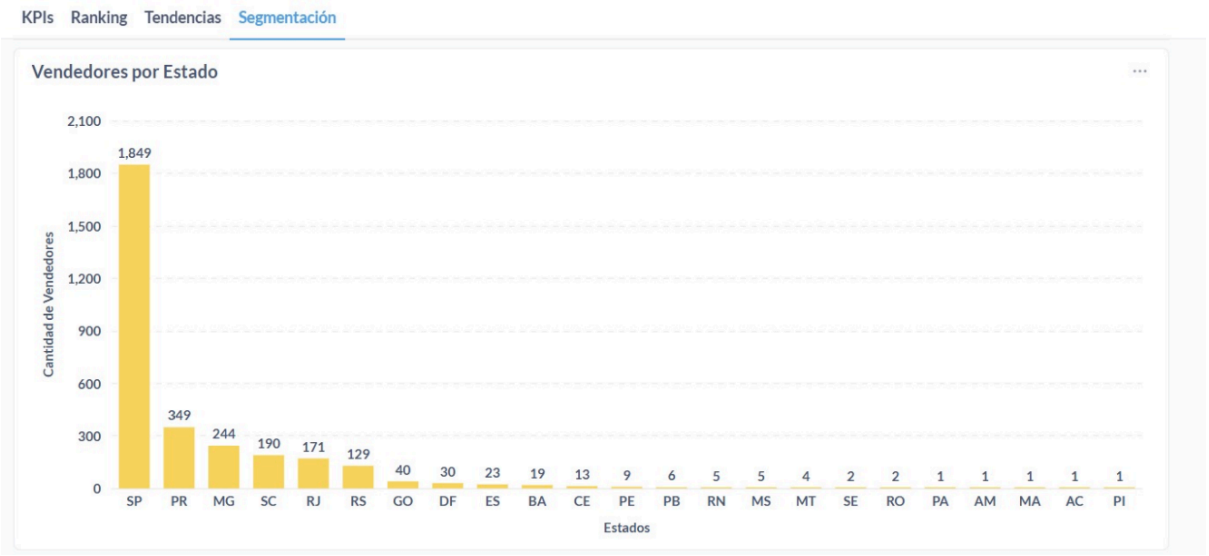
En diciembre, si bien las ventas descienden respecto al pico de noviembre, se mantienen en un nivel considerablemente superior al resto del año, lo que confirma la alta concentración de compras en el período previo a las fiestas.

Segmentación

El sector de **Segmentación** tiene como objetivo analizar la distribución del negocio a partir de diferentes dimensiones geográficas y comerciales. Este tipo de análisis permite identificar **concentraciones de oferta**, desequilibrios regionales y oportunidades de expansión en zonas con menor presencia de vendedores.

A través de esta segmentación, se obtiene una visión más profunda de cómo está estructurado el ecosistema de vendedores dentro del e-commerce.

Vendedores por Estado

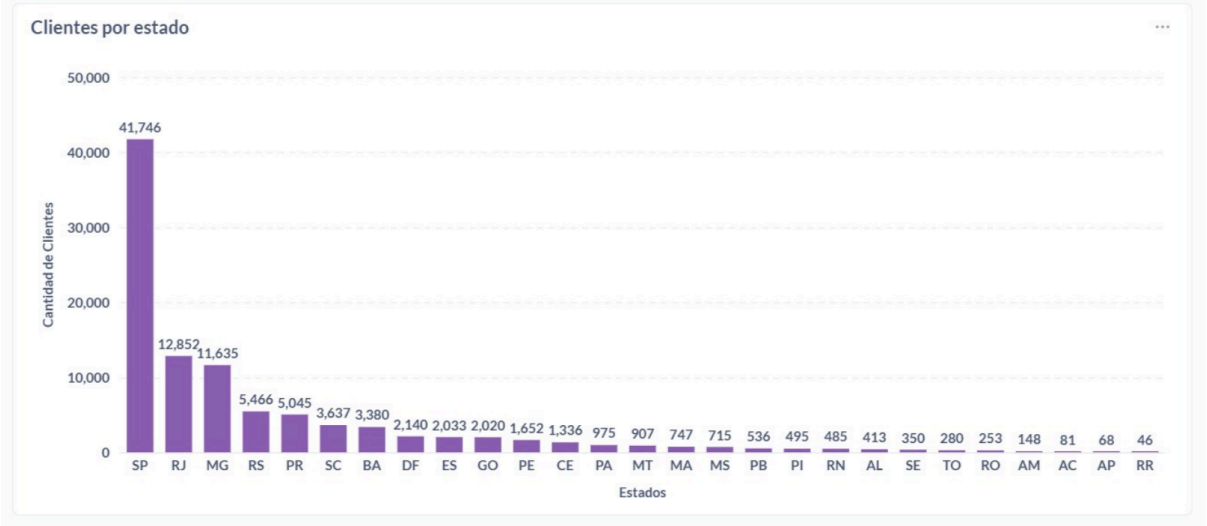


El gráfico presentado muestra la **cantidad de vendedores registrados por estado**, permitiendo observar la distribución geográfica de la oferta comercial.

Se destaca una **alta concentración de vendedores en el estado de São Paulo (SP)**, que supera ampliamente al resto de las regiones, consolidándose como el principal polo comercial del e-commerce analizado. En un segundo nivel se encuentran estados como **Paraná (PR), Minas Gerais (MG), Santa Catarina (SC) y Río de Janeiro (RJ)**, con una presencia significativa pero considerablemente menor.

En contraste, varios estados presentan una **baja cantidad de vendedores**, lo que evidencia una distribución desigual de la actividad comercial a nivel nacional. Este comportamiento sugiere posibles oportunidades estratégicas para fomentar la incorporación de nuevos vendedores en regiones menos representadas y lograr una mayor diversificación geográfica del negocio.

Clientes por Estado



El gráfico muestra la **distribución de clientes por estado**, evidenciando una fuerte concentración en determinadas regiones del país. El estado de **São Paulo (SP)** se destaca ampliamente como el principal núcleo de clientes, seguido por **Rio de Janeiro (RJ)** y **Minas Gerais (MG)**, lo que refleja una mayor penetración del e-commerce en los estados más poblados y con mayor desarrollo económico.

A medida que se avanza hacia otros estados, la cantidad de clientes disminuye progresivamente, lo que permite identificar zonas con menor participación pero con potencial de crecimiento. Este análisis resulta fundamental para orientar estrategias de marketing regional, optimizar la logística y evaluar oportunidades de expansión hacia mercados menos explotados.

6. Conclusión General

A lo largo de este proyecto se llevó a cabo un análisis integral del dataset **Brazilian E-Commerce**, abordando todas las etapas fundamentales de un proceso de análisis de datos: comprensión del negocio, exploración y limpieza de los datos, modelado analítico y visualización de la información.

Partiendo del dataset original, se realizó un proceso de **preparación y transformación de los datos**, que permitió estructurar la información en un **esquema estrella** orientado al análisis, facilitando la obtención de métricas clave y mejorando el rendimiento de las consultas analíticas. Este modelo resultó fundamental para responder de manera eficiente a preguntas de negocio relacionadas con ventas, clientes, productos, vendedores y comportamiento temporal.

A partir de esta base analítica se construyeron **dashboards interactivos** utilizando **Metabase**, organizados en secciones de **KPIs, Ranking, Tendencias y Segmentación**. Estas visualizaciones permitieron analizar el desempeño general del e-commerce, identificar los productos y clientes más relevantes, comprender la evolución de las ventas a lo largo del tiempo y detectar patrones geográficos de concentración de clientes y vendedores.

El proyecto demuestra cómo una correcta **modelización de datos**, combinada con herramientas de **visualización**, permite transformar grandes volúmenes de información en **insights claros y accionables**, aportando valor tanto a la toma de decisiones estratégicas como operativas.