Grupo 3. Avances del proyecto - semana 3

YOSELIN NIETO GIL, FELIPE DIEGO FELIPE DIEGO LOBATO DA SILVA, NICOLAS ALEJANDRO YEPES JOVEN y ANA MARIA SOTO OROZCO.

1. Definición del problema.

El proyecto se enfoca en abordar un desafío común en la gestión de productos de consumo masivo: la predicción de la demanda estacional. El problema específico es la dificultad que enfrentan las empresas para anticipar los cambios en el comportamiento de compra de los consumidores, especialmente durante las diferentes temporadas del año (por ejemplo, vacaciones, festividades, cambios climáticos, etc.). Estos cambios afectan las ventas y, si no son correctamente previstos, pueden resultar en problemas como exceso o falta de inventario, pérdidas de oportunidades de venta y costos innecesarios.

El objetivo del proyecto es utilizar datos históricos de ventas para identificar patrones estacionales y prever la demanda futura, lo que permitirá a las empresas tomar decisiones más informadas sobre la gestión de inventarios y estrategias comerciales.

2. Definición de la pregunta de negocio y alcance del proyecto.

La pregunta del negocio principal que guiara este proyecto es:

¿Cómo podemos predecir la demanda estacional de productos de consumo masivo, teniendo en cuenta factores históricos de ventas y comportamiento de compra, para optimizar la gestión de inventario y mejorar la eficiencia operativa?

Este proyecto tiene un alcance limitado a productos de consumo masivo vendidos a través de plataformas de e-commerce (en este caso, Olist Store), durante un período determinado (3 años inicialmente, pero ajustable). El modelo debe ser capaz de anticipar cuáles productos tendrán mayor demanda en diferentes épocas del año, considerando patrones estacionales y de comportamiento del consumidor. La solución está diseñada para empresas que buscan mejorar su gestión de inventarios y maximizar la disponibilidad de productos según las necesidades de sus clientes.

3. Definición de los conjuntos de datos a emplear.

Para el desarrollo del proyecto de predicción de demanda estacional, se utilizarán datos históricos de la plataforma Olist Store. Estos datos proporcionan información detallada sobre los productos vendidos, las transacciones realizadas, los clientes, los vendedores y las reseñas. A continuación, se describen las principales tablas que serán empleadas en el análisis:

Tabla geo

Esta tabla contiene información geográfica relevante sobre los códigos postales de los clientes y vendedores.

Nombre de la Columna	Descripción	Tipo de Dato	Ejemplo del Dato	Clave
`cep_prefix`	Prefijo del CEP (Código Postal)	string	"64091"	Clave primaria
`city`	Nombre de la ciudad	string	"teresina"	
`uf`	Unidad Federativa (Estado)	string	"PI"	
`lat`	Latitud de la ubicación	doble	-5.0874608	
`lon`	Longitud de la ubicación	doble	-42.8049571	

Tabla customers

Información detallada sobre los clientes que realizaron compras en Olist Store.

Nombre de la Columna	Descripción	Tipo de Dato	Ejemplo del Dato	Clave
customer_id	Identificador único del cliente	string	"f2a1d75b74d9ec748"	Clave primaria
customer_unique_id	Identificador único y persistente del cliente	string	"15ee900ec703c9a10"	
customer_zip_code_prefix	Prefijo del código postal del cliente	string	"68590"	Clave foránea (geo.cep_prefix)
customer_city	Ciudad del cliente	string	"jacunda"	
customer_state	Estado del cliente	string	"PA"	

Tabla order_items

Esta tabla contiene los detalles de los ítems incluidos en cada pedido realizado en Olist Store.

Nombre de la Columna	Descripción	Tipo de Dato	Ejemplo del Dato	Clave
order_id	Identificador único del pedido	string	"8ac26cb70a17887c..."	Clave foránea (orders.order_id)
order_item_id	Identificador del ítem dentro del pedido	integer	1	Clave primaria compuesta junto a order_id
product_id	Identificador único del producto	string	"4lebbrb7a41c44632..."	Clave foránea (products.product_id)
seller_id	Identificador único del vendedor	string	"7a67c85e85bbc2e85..."	Clave foránea (sellers.seller_id)
shipping_limit_date	Fecha límite para el envío	timestamp	"2017-05-22 16:05:44"	
price	Precio del producto	doble	109.99	
freight_value	Valor del envío	doble	18.02	

Tabla order_payments

Registra los detalles de los pagos realizados en los pedidos.

Nombre de la Columna	Descripción	Tipo de Dato	Ejemplo del Dato	Clave
order_id	Identificador único del pedido	string	"b81ef226f3fe1789b..."	Clave foránea (orders.order_id)
payment_sequential	Secuencia del pago dentro del pedido	integer	1	Clave primaria compuesta junto a order_id
payment_type	Tipo de pago utilizado	string	"credit_card"	
payment_installments	Número de cuotas del pago	integer	8	
payment_value	Valor del pago	doble	99.33	

Tabla order_reviews

Incluye información sobre las reseñas de los clientes en Olist Store.

Nombre de la Columna	Descripción	Tipo de Dato	Ejemplo del Dato	Clave
review_id	Identificador único de la reseña	string	"f21e5dff1c6efe..."	Clave primaria
order_id	Identificador único del pedido	string	"d6027c4fb846f61d7..."	Clave foránea (orders.order_id)
review_score	Calificación de la reseña (1 a 5)	string	4	
review_comment_title	Título de la reseña (si existe)	string	NULL	
review_comment_message	Mensaje o comentario de la reseña	string	"Adorei a mercadoria!"	
review_creation_date	Fecha en que se creó la reseña	timestamp	"2017-09-21 00:00:00"	
review_answer_timestamp	Fecha y hora en que se respondió la reseña	timestamp	"2017-09-23 11:06:26"	

Tabla orders

Contiene información sobre los pedidos realizados en la plataforma.

Nombre de la Columna	Descripción	Tipo de Dato	Ejemplo del Dato	Clave
order_id	Identificador único del pedido	string	"995392413cee616c1..."	Clave primaria
customer_id	Identificador único del cliente	string	"4bf2490c4245cdb25a..."	Clave foránea (customers.customer_id)
order_status	Estado del pedido (ej. delivered, canceled)	string	"delivered"	
order_purchase_timestamp	Fecha y hora en que se realizó la compra	timestamp	"2017-09-04 22:24:05"	
order_approved_at	Fecha y hora en que se aprobó el pedido	timestamp	"2017-09-04 22:43:54"	
order_delivered_carrier_date	Fecha y hora en que el transportista entregó el pedido	timestamp	"2017-09-13 17:20:04"	
order_delivered_customer_date	Fecha y hora en que el cliente recibió el pedido	timestamp	"2017-09-22 21:09:32"	
order_estimated_delivery_date	Fecha estimada de	timestamp	"2017-09-27 00:00:00"	

	entrega del pedido			

Tabla products

Incluye detalles sobre los productos vendidos en la plataforma.

Nombre de la Columna	Descripción	Tipo de Dato	Ejemplo del Dato	Clave
product_id	Identificador único del producto	string	"1e9e8ef04dbcff454..."	Clave primaria
product_category_name	Categoría del producto	string	"perfumaria"	
product_name_lenght	Longitud del nombre del producto	integer	40	
product_description_lenght	Longitud de la descripción del producto	integer	287	
product_photos_qty	Cantidad de fotos del producto	integer	1	
product_weight_g	Peso del producto en gramos	integer	225	
product_length_cm	Longitud del producto en centímetros	integer	16	
product_height_cm	Altura del producto en centímetros	integer	10	
product_width_cm	Ancho del producto en centímetros	integer	14	

Tabla sellers

Información sobre los vendedores que ofrecen productos en Olist Store.

Nombre de la Columna	Descripción	Tipo de Dato	Ejemplo del Dato	Clave
seller_id	Identificador único del vendedor	string	"1f50f92002b6aeb04..."	Clave primaria
seller_zip_code_prefix	Prefijo del código postal del vendedor	string	"13023"	Clave foránea (geo.cep_prefix)
seller_city	Ciudad del vendedor	string	"campinas"	
seller_state	Estado del vendedor	string	"SP"	

Tabla datos product_category

Esta tabla contiene información sobre las categorías de productos disponibles en la plataforma, tanto en portugués como en inglés.

Nombre de la Columna	Descripción	Tipo de Dato	Ejemplo del Dato	Clave
product_category_name	Nombre de la categoría de productos en portugués	string	"beleza_saude"	Clave primaria
product_category_name_english	Nombre de la categoría de productos en inglés	string	"health_beauty"	

A continuación, se adjunta la imagen que muestra el esquema de datos de las tablas descritas anteriormente y se identifica las relaciones que existen entre ellas.

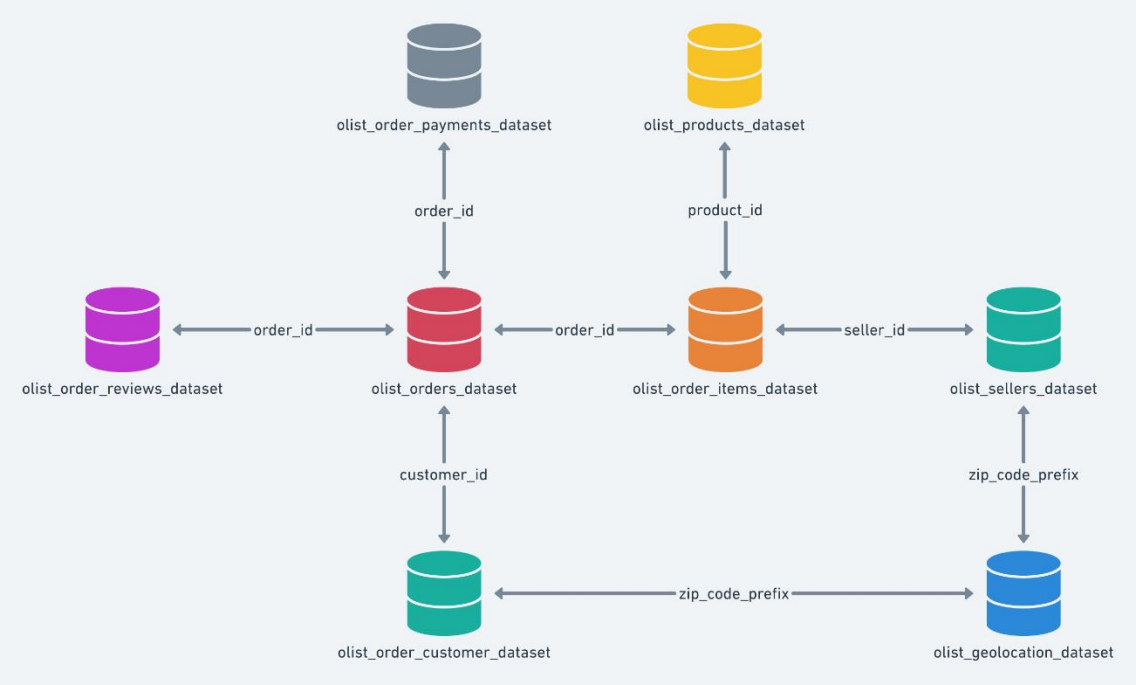


Imagen "Esquema de datos de insumos"

3.1 Tratamiento inicial para conexión de datos.

El primer paso en el proceso fue garantizar que los datos estuvieran preparados y organizados para un uso eficiente en los análisis. Todas las bases de datos se obtuvieron del enlace https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce. Además, utilizamos el archivo `log.csv`, que contenía información actualizada de los códigos postales (CEP), adquirido a través de un vendedor por 9 dólares que realiza raspado de datos.

Para preparar los datos, utilizamos cuatro scripts, cada uno desempeñando un papel esencial en las etapas de estandarización, transformación y enriquecimiento de los datos de geolocalización. La conversión de los archivos originales en formato CSV al formato Parquet, que ofrece una estructura más eficiente para el almacenamiento y consulta de grandes volúmenes de datos, fue una de estas etapas fundamentales.

Script 1: Comenzamos identificando y tratando los archivos que necesitaban estandarización. Por ejemplo, los prefijos de los códigos postales (CEP) recibieron ceros a la izquierda para mantener la consistencia en todos los registros. El archivo `log.csv` se incorporó en este proceso, asegurando que su información estuviera alineada con las demás bases. Después de esta estandarización, los archivos se convirtieron y guardaron en una nueva carpeta en formato Parquet, optimizando el acceso y la manipulación de los datos para las etapas posteriores.

Script 2: A continuación, nos centramos en la transformación y unión de los datos de geolocalización. Combinamos diferentes fuentes de datos, incluyendo `log.csv`, estandarizando y ajustando las coordenadas de latitud y longitud para garantizar la conformidad. Este proceso fue esencial para asegurar que los datos geográficos estuvieran alineados, permitiendo el cruce con otras tablas y la realización de análisis más complejos, como análisis espaciales.

Script 3: Con los datos estandarizados, enriquecimos la información utilizando la API de Nominatim para obtener coordenadas geográficas precisas. A partir de información parcial, como los prefijos de los CEP, fue posible obtener las coordenadas exactas de cada registro, incluyendo los datos presentes en `log.csv`. Esto posibilitó un análisis más detallado de las ubicaciones de clientes y vendedores.

Script 4: Por último, refinamos y unificamos los datos. Leímos los archivos `geolocation_correios.parquet` y `geo_coords.parquet`, realizando un *inner join* basado en la columna `cep_prefix` para asegurar la correspondencia entre las ubicaciones y las coordenadas geográficas. Después de la unión, las columnas de latitud y longitud se convirtieron al tipo de dato *double*, garantizando una mayor precisión en el tratamiento de las coordenadas geográficas.

Los datos enriquecidos se consolidaron y guardaron en el archivo `geolocation_correios_coords.parquet`. Con las columnas de latitud y longitud ajustadas al tipo de dato adecuado (*double*), aseguramos la precisión necesaria para análisis posteriores. Los datos están ahora listos para análisis espaciales, con precisión garantizada para consultas y cruces eficientes.

4. Análisis Exploratorio de los datos.

a. Primero se procedió con una exploración de las dimensiones de cada una de las bases de datos con las que se contaba. Para determinar sus dimensiones, el nombre de las variables las posibles llaves que permitirían hacer un join para formar una base de datos principal a explorar.
b. Se hace una exploración de la base de datos combinada y se procede ha hacer limpieza de la misma de la siguiente manera:
    - El análisis del DataFrame revela que contiene un total de 119,143 filas y 39 columnas, lo que sugiere una base de datos considerable para el estudio de órdenes de compra.
    - Las estadísticas descriptivas muestran que las fechas de las órdenes de compra varían desde el 5 de septiembre de 2016 hasta el 17 de octubre de 2018. Las columnas relacionadas presentan valores nulos, indicando que algunos pedidos no fueron entregados según lo previsto.
    - En cuanto a las estadísticas numéricas, el precio medio de los productos es de 120.65 reales, con un valor de flete promedio de 20.03 reales. Sin embargo, se presentan discrepancias notables en los datos, ya que hay valores nulos en varias columnas, como order_approved_at, order_delivered_carrier_date, y order_delivered_customer_date, lo que puede complicar un análisis enfocado en de la eficiencia en la entrega (para este ejercicio el enfoque está en predicción de precios y demanda, no eficiencias de entrega).
    - Además, el análisis muestra que las variables de longitud, peso y dimensiones de los productos también tienen una cantidad significativa de valores nulos, lo que podría limitar las conclusiones sobre la relación entre estas características y otros aspectos, como el precio o el tiempo de entrega.
    - Los tipos de variables en el DataFrame incluyen una mezcla de datos categóricos y numéricos, así como datos temporales. En particular, se observa que la mayoría de las columnas son de tipo object, seguidas de float64 para los datos numéricos y datetime64 para las fechas. Por último, las primeras cinco filas del DataFrame proporcionan una visión inicial de la estructura de los datos, mostrando pedidos con información sobre el estado, las fechas de compra y entrega, así como los detalles del producto y del vendedor. Esto establece una base sólida para un análisis más profundo de los patrones de compra y entrega en la plataforma.
    - Dado el considerable tamaño de valores nulos en las variables relacionadas a la calificación del producto y experiencia de usuario se determina descartar el uso de esta información para el análisis, pues eliminar los valores nulos o tener en cuenta los pocos que se han dispuesto puede afectar la efectividad del modelo.

Exploración de valores nulos por variable

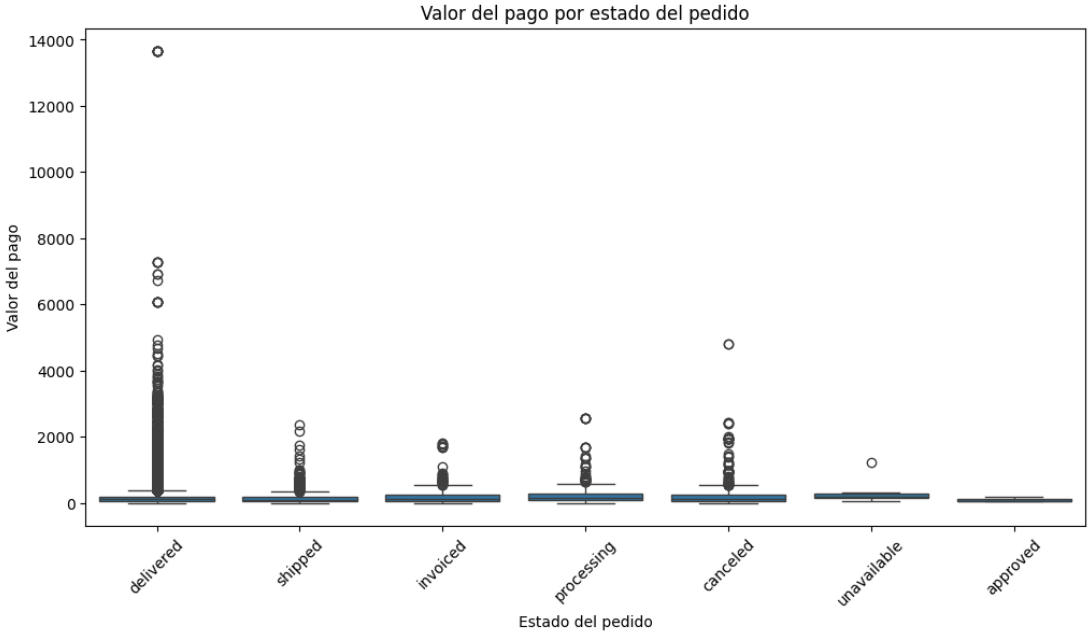
```
order_status                        0
order_purchase_timestamp            0
order_approved_at                 177
order_estimated_delivery_date       0
order_item_id                     833
product_id                        833
seller_id                         833
shipping_limit_date               833
price                             833
freight_value                     833
payment_sequential                  3
payment_type                        3
payment_installments                3
payment_value                       3
customer_unique_id                  0
customer_zip_code_prefix            0
customer_city                       0
customer_state                      0
product_category_name            2542
product_name_lenght              2542
product_description_lenght       2542
product_photos_qty               2542
product_weight_g                  853
product_length_cm                 853
product_height_cm                 853
product_width_cm                  853
seller_zip_code_prefix            833
seller_city                       833
seller_state                      833
dtype: int64
```

c. Después de la depuración de variables con alto grado de valores nulos, se procede a eliminar los valores nulos de variables. Si bien existen grupos de tamaño considerable (como product_category_name con una cantidad de nulos de 2542), estos son se estiman relevantes para el entrenamiento del modelo, de manera que se mantienen las variables a costa de eliminar sus registros inválidos.

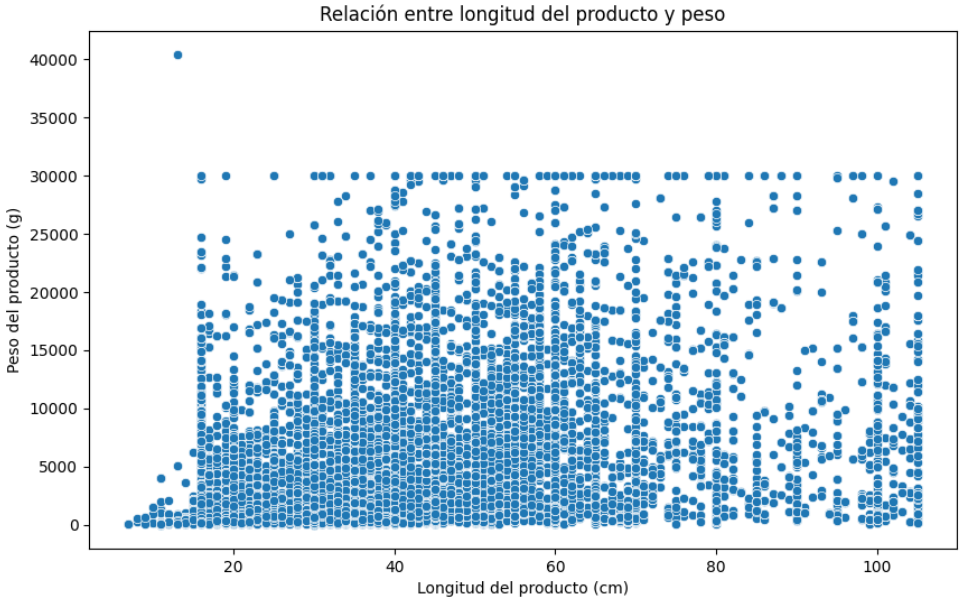
d. Sobre la base de datos depurada, sin nulos, se procedió a hacer una revisión de las estadísticas cuantitativas disponibles y una revisión en la distribución entre algunas de las variables de interés, así:

- Las estadísticas descriptivas de las columnas numéricas del DataFrame revelan información clave sobre los datos. En cuanto a los precios, la media es de 120.93, con un mínimo de 0.85 y un máximo de 6735.00; la desviación estándar es de 184.19. En lo que respecta al valor de flete, la media es de 20.08, el mínimo es 0.00 y el máximo alcanza 409.68, con una desviación estándar de 15.87. Las fechas de compra muestran una fecha mínima de 2016-09-05 y una máxima de 2018-09-03, con un promedio de 2017-12-31.
- Al visualizar las primeras filas del DataFrame, se pueden observar ejemplos concretos de los datos, confirmando que las columnas están correctamente pobladas. A partir de esta información, se pueden considerar diversas áreas de análisis, como el estudio de tendencias de ventas a lo largo del tiempo utilizando las fechas de compra para identificar patrones estacionales, el análisis de los productos más vendidos mediante product_id y order_item_id, y la evaluación del impacto del flete en las ventas comparando price y freight_value.
- Grafica 1 / Valor del pago por estado del pedido: Esta gráfica presenta la distribución del valor del pago según diferentes estados del pedido, como "delivered", "shipped", "invoiced", "processing", etc. La

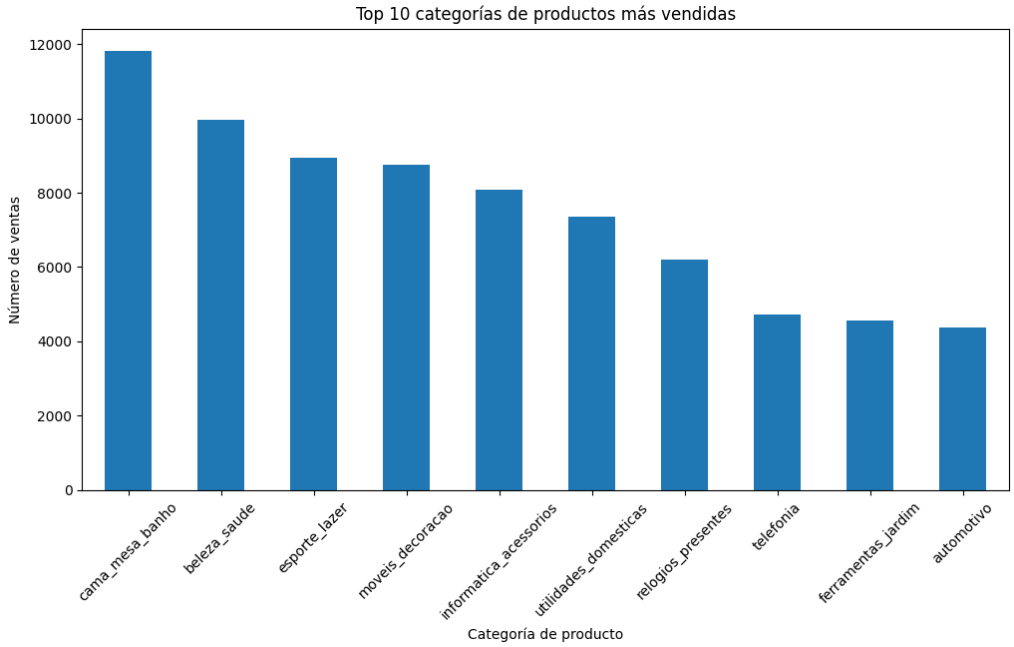
mayoría de los pagos están concentrados en valores bajos, pero existen algunos valores atípicos (outliers) que alcanzan hasta los 14,000. En particular, el estado "delivered" parece tener una mayor variabilidad y más valores atípicos que el resto de los estados, lo que sugiere que hay una amplia gama de pagos en pedidos entregados, posiblemente debido a diferencias en los tipos de productos o en el tamaño de los pedidos.



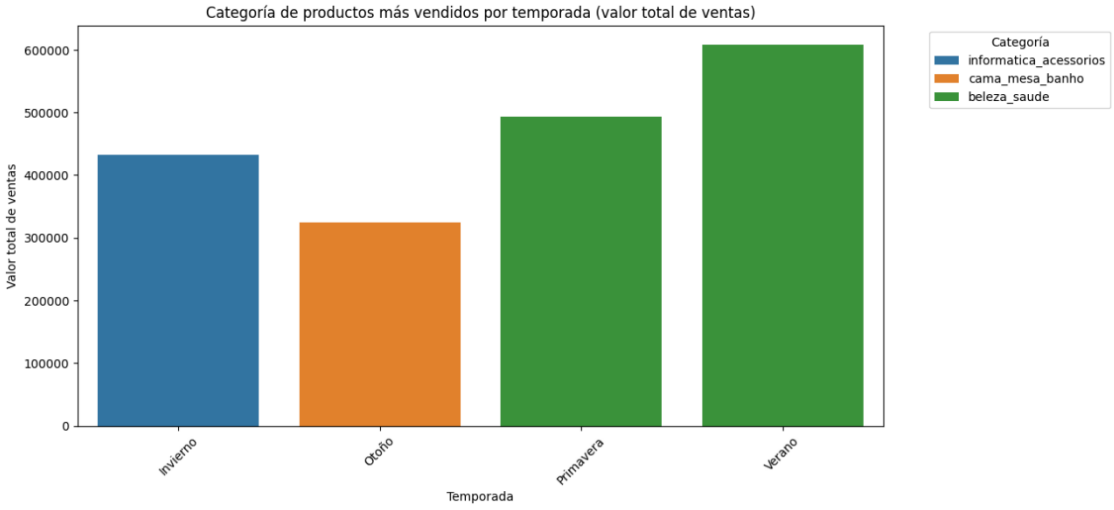
- Grafica 2 / Relación entre longitud del producto y peso: Esta gráfica muestra la relación entre la longitud (en cm) y el peso (en gramos) de los productos. A simple vista, no parece haber una relación clara o lineal entre la longitud y el peso; los puntos están distribuidos de manera dispersa. Sin embargo, se observa que a medida que la longitud aumenta, el peso de los productos tiende a ser más variable, y algunos productos de longitud media (alrededor de 50-80 cm) alcanzan pesos altos, lo cual puede indicar que los productos más largos no siempre son más pesados.



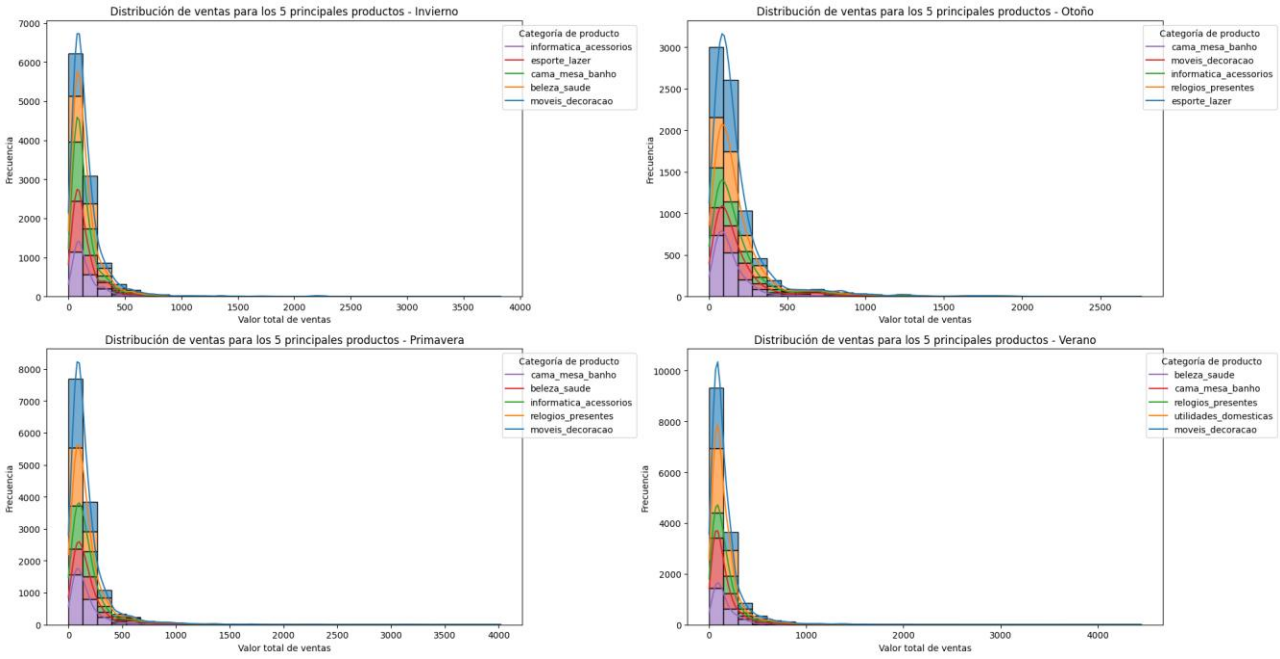
- Grafica 3 / Top 10 productos más vendidos según su categoria: Se observa que las categorías de hogaar, belleza, bienestar y tecnología liderán las ventas del market place Olist.



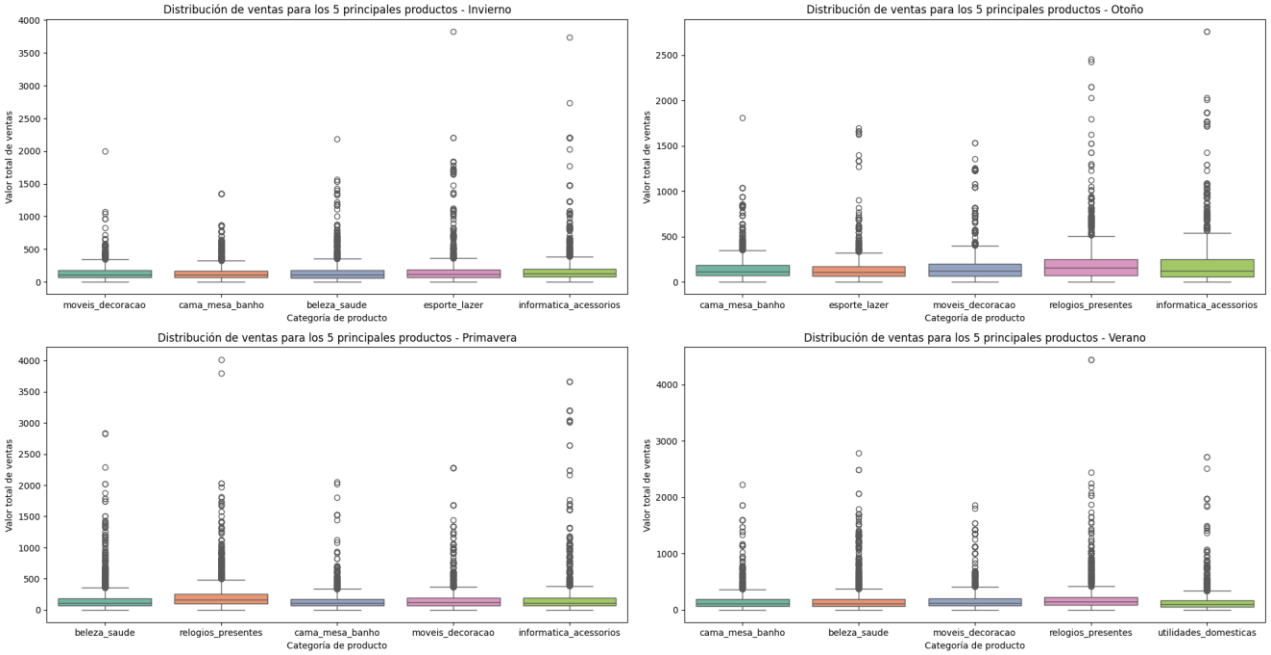
- La grafica de categoria de producto vendido por temporeda sugiere que "beleza_saude" tiene una gran demanda tanto en primavera como en verano, mientras que en invierno se observa un aumento en la venta de productos tecnológicos. Esto puede reflejar una estacionalidad en las preferencias de los clientes, probablemente impulsada por factores como el clima o eventos de temporada.



- A continuación, se exploraron los 5 productos más vendidos por temporada para determinar la distribución de las ventas, en todas las temporadas se pueden ver los productos de las categorías “cama_mesa_banho” y “moveis_decoracao” y en otoño es la época de donde son mas frecuentes ventas por mayores valores de manera constante para todas las categorías, aunque la mayor frecuencia de ventas se obtiene en verano, en la categoría de muebles y decoración.

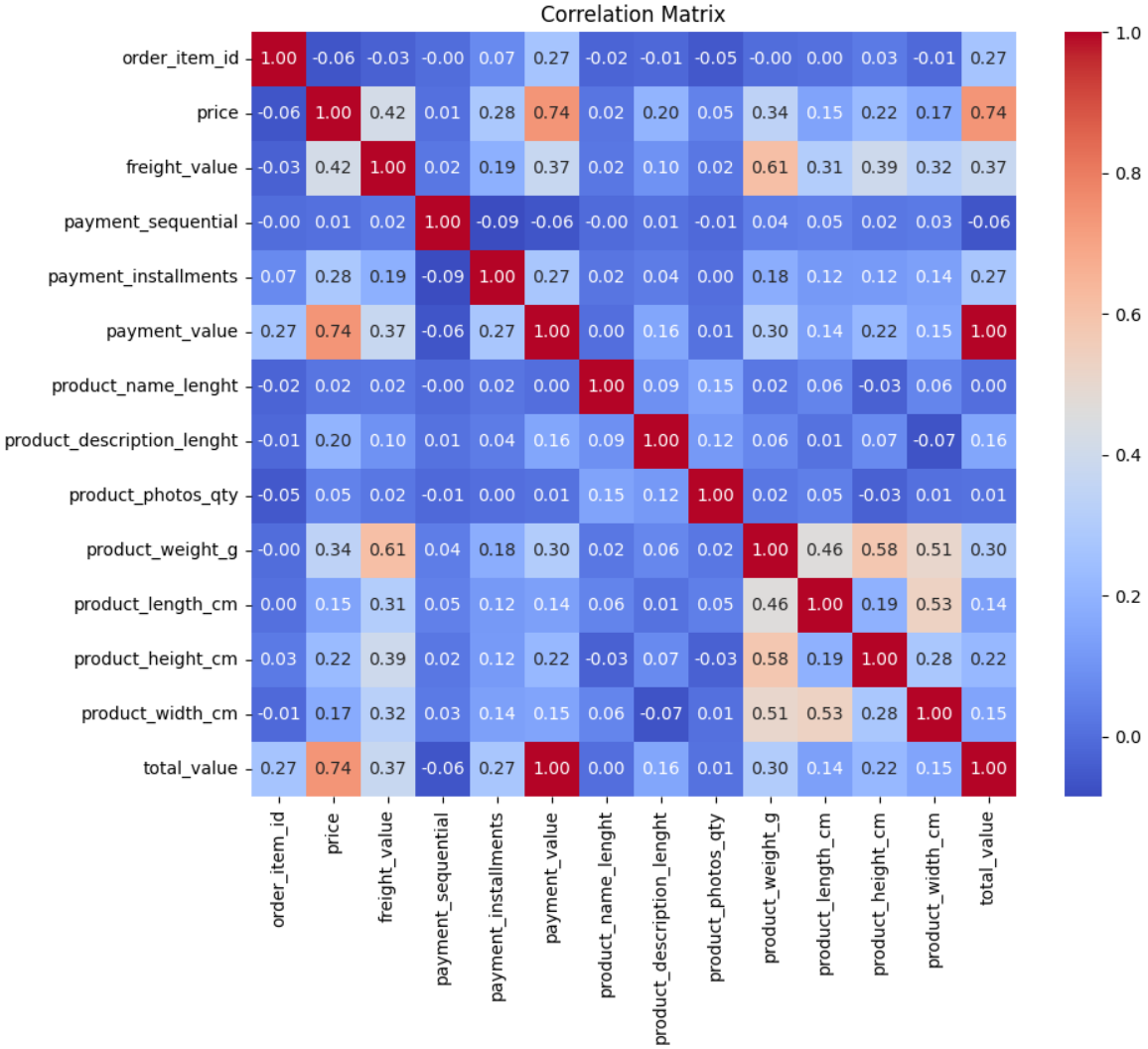


- Para complementar este análisis se realizaron boxplot que permiten identificar la variación en los valores totales de ventas para cada temporada en las 5 categorías de muestra, se encontró que en todos los casos los valores atípicos están presentes y corresponden al rango alto. En invierno y primavera informática y accesorios tiene una media mayor del total de ventas, mientras en otoño y verano predominan los presentes religiosos.



e. Finalmente, se hizo un ejercicio de revisión de correlación de las variables cuantitativas alrededor de la variable Price (ignorando las variables de payment value y total value). Se definió lo siguiente:

- El valor de flete presenta una correlación positiva moderada de 0.416653 con el precio. Esto sugiere que los productos más caros tienden a tener mayores costos de envío, lo que podría deberse al tamaño, peso o distancia de entrega de estos.
- Otro aspecto destacado es la correlación positiva con el número de cuotas de pago, que tiene un valor de 0.279394. Esto indica que, a medida que aumenta el número de cuotas, el precio de los artículos también tiende a ser más alto. Esta relación sugiere que los productos de mayor precio a menudo son adquiridos a través de planes de pago a plazos.
- La longitud de las descripciones de los productos también muestra una correlación positiva, aunque más débil, de 0.201853. Esto implica que los productos con descripciones más extensas suelen tener precios ligeramente más altos, lo que podría estar relacionado con la presencia de más detalles o características que destacan su valor.
- Asimismo, el peso del producto tiene una correlación positiva moderada de 0.340408 con el precio. Este hallazgo indica que los productos más pesados tienden a ser más caros, lo cual es coherente con la idea de que los artículos de mayor tamaño o peso generalmente tienen un costo más elevado. Las dimensiones del producto, como la altura (0.224383), el ancho (0.172371) y la longitud (0.145765), también muestran correlaciones positivas, aunque más débiles. Esto sugiere que los productos de mayor tamaño tienden a tener precios más altos, alineándose con la noción de que el tamaño es un factor importante en la determinación del costo.
- Por otro lado, el identificador del artículo del pedido presenta una correlación negativa débil de -0.060448 con el precio, lo que indica que no existe una relación significativa entre estas dos variables. Esto sugiere que el número de identificación del artículo no influye en el precio del producto.



5. Maqueta del dashboard.

A continuación, se presenta la maqueta inicial del proyecto, La idea es que el usuario selecciona la temporada para obtener los resultados del modelo. Así, puede obtener un precio estimado de ventas en la categoría seleccionada para esa temporada, también obtendrá información sobre cómo es la distribución espacial de los clientes en los estados y ciudades de Brasil para identificar donde está ubicada la mayor demanda. En el tablero también se mostrará la a proyección de ventas en el top 10 de categorías y una tabla donde se encuentra clasificada la demanda en: alta, media y baja según el rango de ventas por categoría en la temporada seleccionada.

6. Repositorios del proyecto.

https://github.com/felipelobatodasilva/despliegue_analytica.git
en el anexo se encuentran detalles sobre el repositorio.

7. Reporte de trabajo en equipo

Sobre los principales responsables de las principales actividades realizadas: Felipe trabajó en la creación de repositorio y cargue de las bases de datos, Ana en la definición del problema de negocio, la pregunta, alcance y apoyo en la unión de bases de datos, Nicolás en unión de Bases de datos y exploratoria de datos, Yoselin en elaboración de la maqueta y apoyo en la exploratoria de datos.

Anexo. Capturas creación de repositorio

Git : Configuré el repositorio remoto en GitHub y subí todos los archivos pequeños, incluyendo el código y otros datos ligeros.

DVC : Los archivos grandes de las carpetas files_csv y files_parquet los gestioné con DVC y los almacené en un bucket de S3.

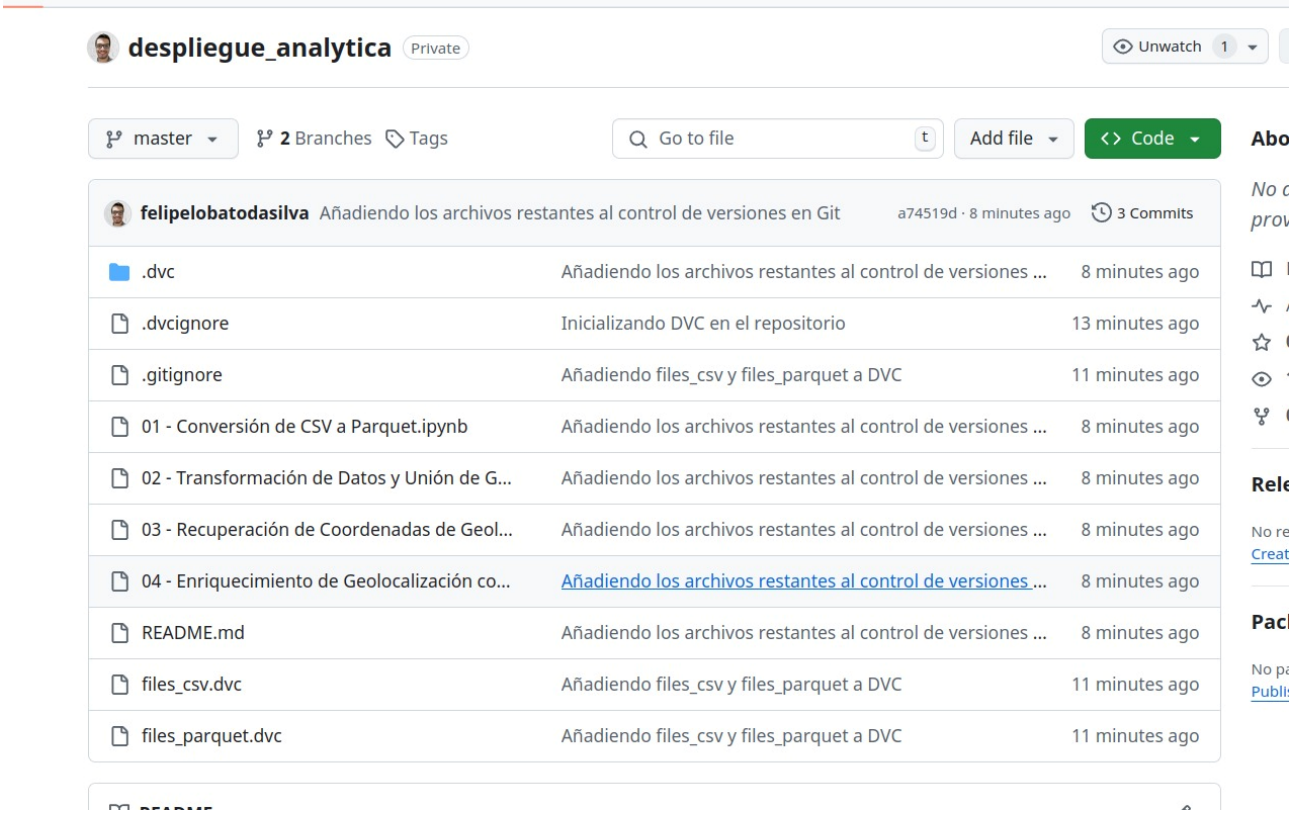


Figura 01 – Vistazo del repo en que están todos los archivos pequeños

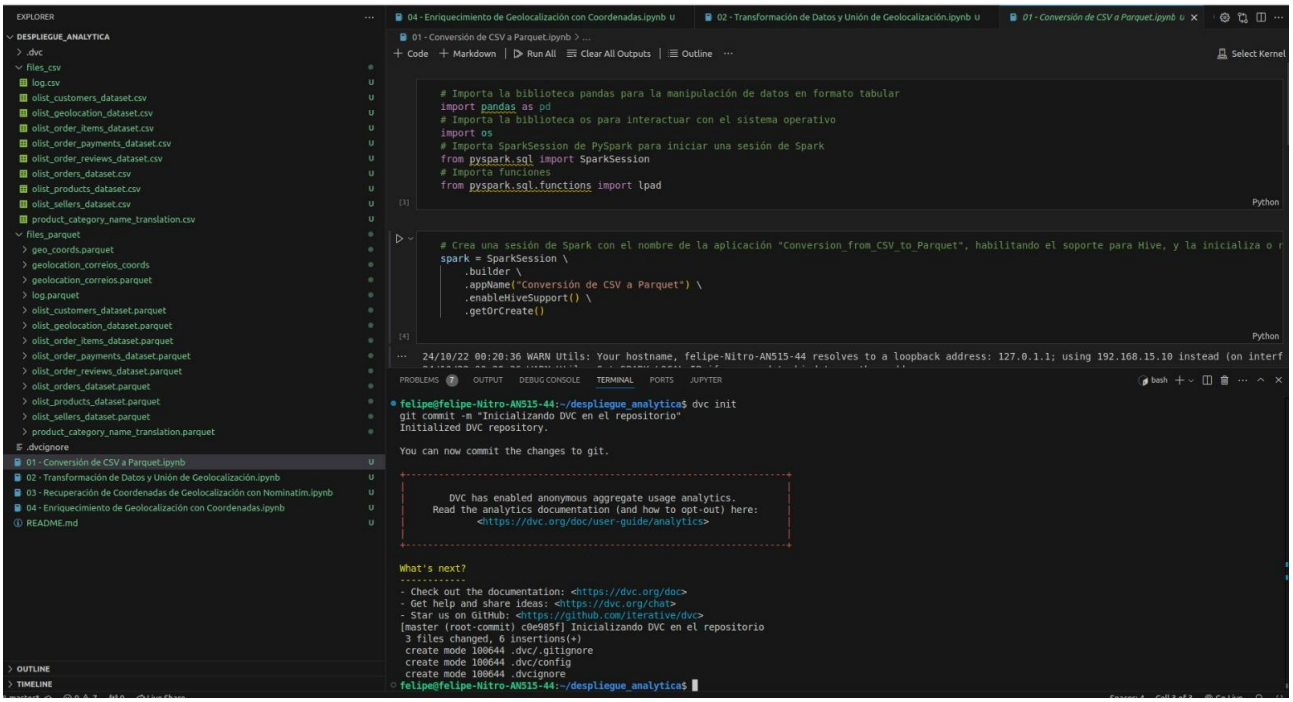


Figura 02 – Preparando el DVC en el mismo repositorio em donde están los archivos pequeños

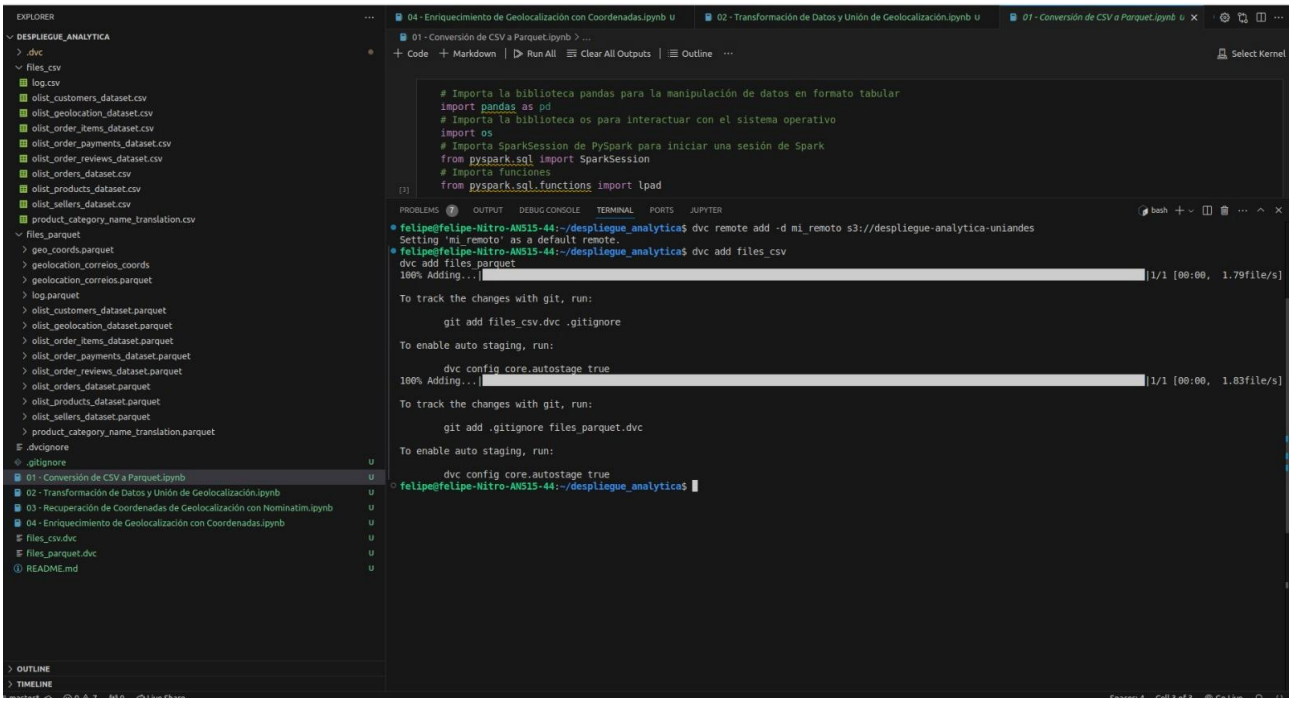


Figura 03 – Agregando el bucket de s3 remoto a dvc para que se guarde los archivos largos adentro de él. Además, se carga las carpetas con los archivos pesados a ese bucket remoto

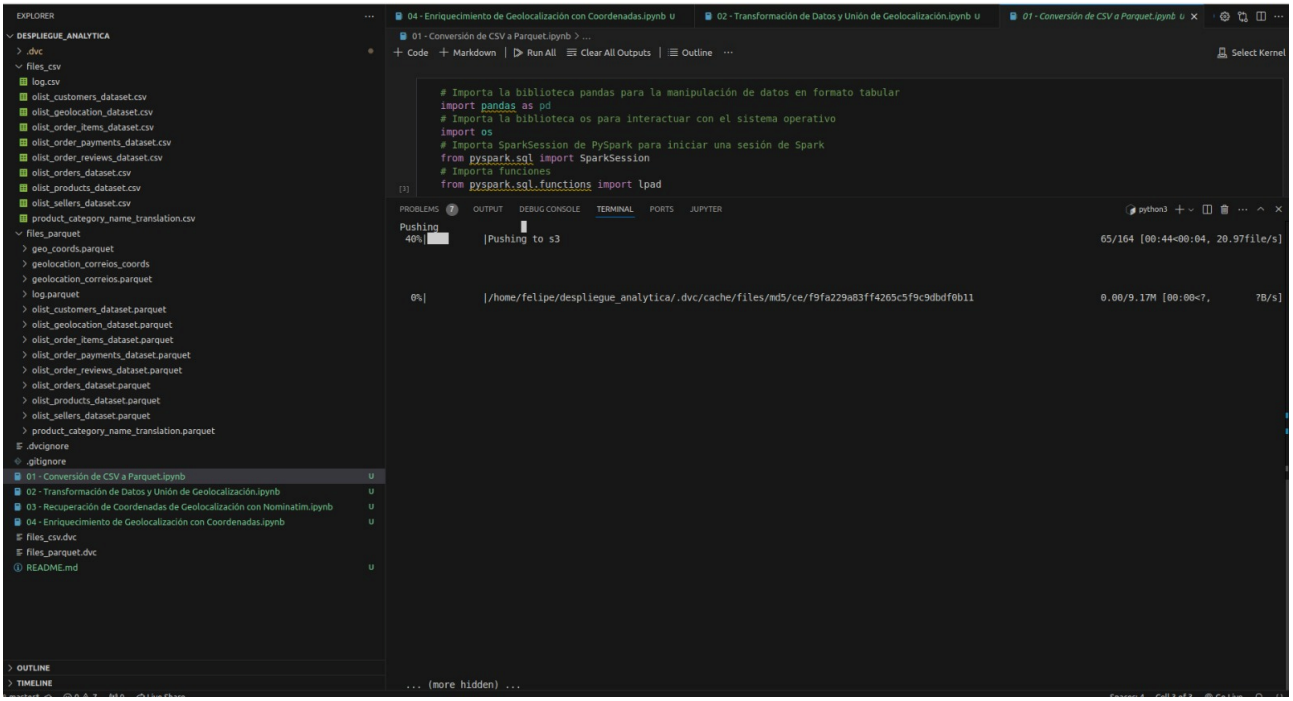


Figura 04 – Empujando los archivos locales para el s3

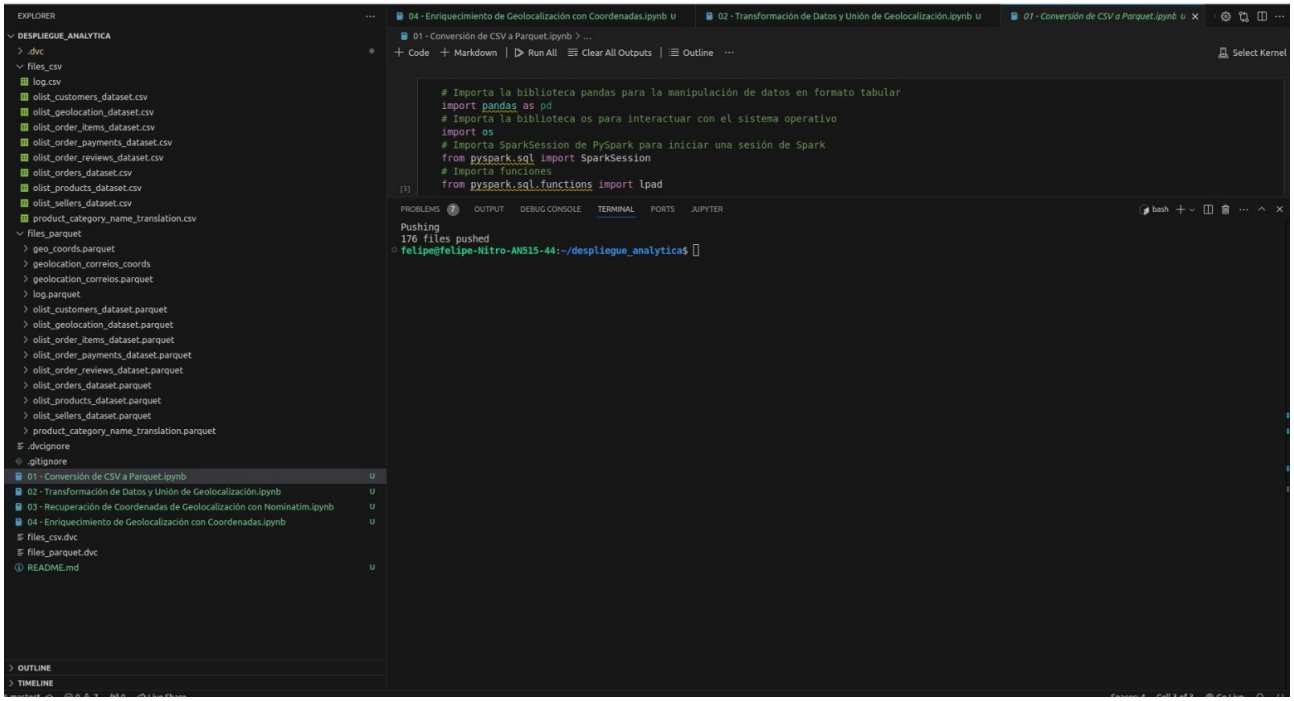


Figura 05 – Archivos empujados con éxito

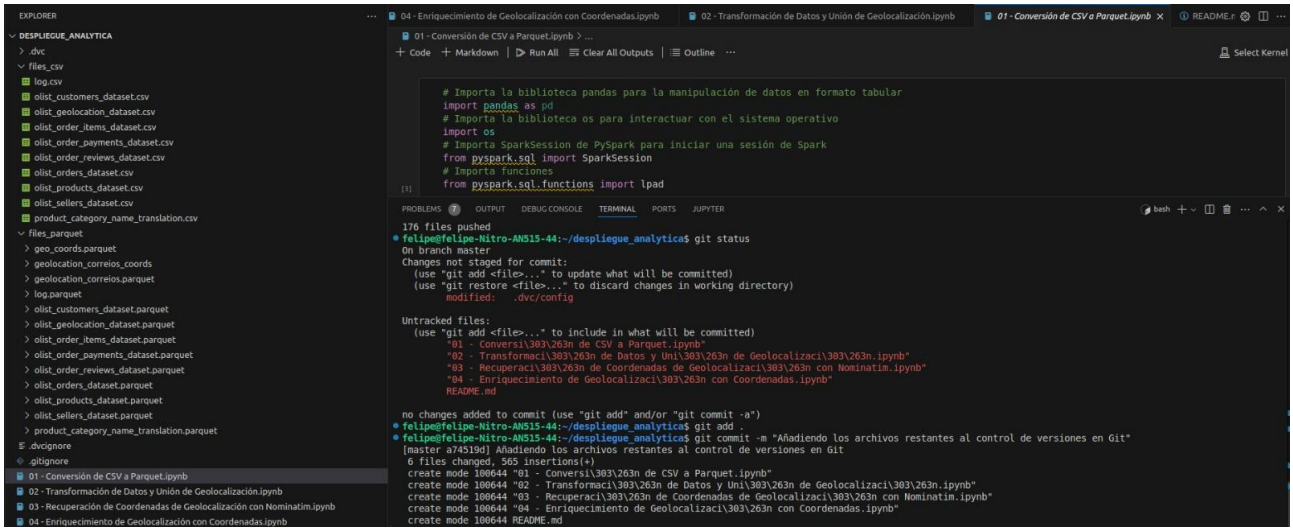


Figura 06 – Commiteando todas las modificaciones en el repo

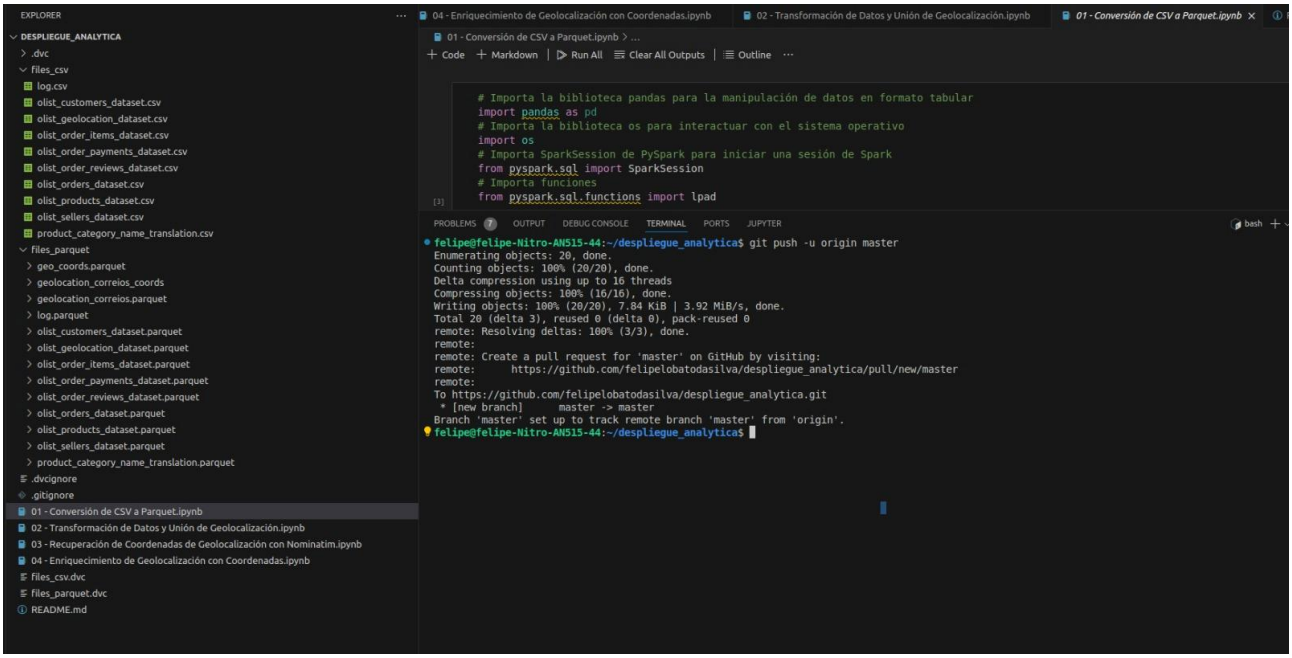


Figura 07 – Realizando el git push al repo

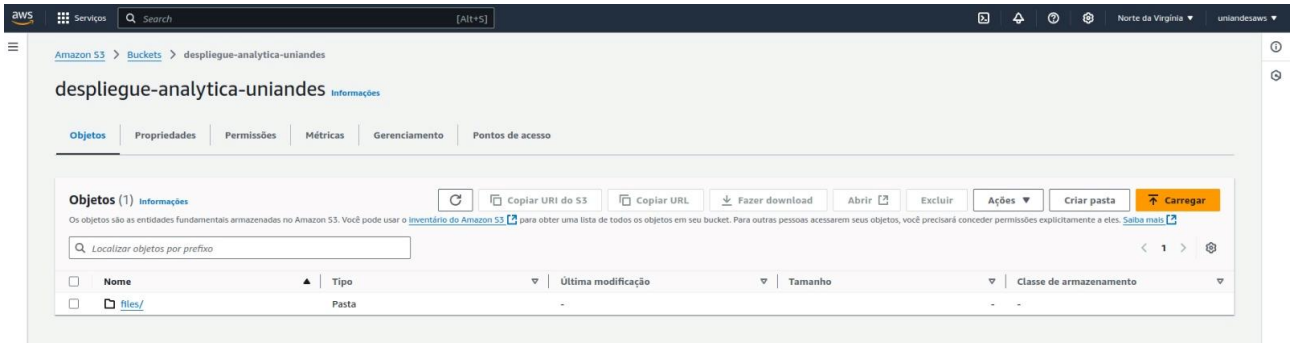


Figura 08 – Captura de los archivos que han sido empujados al s3 en la nube

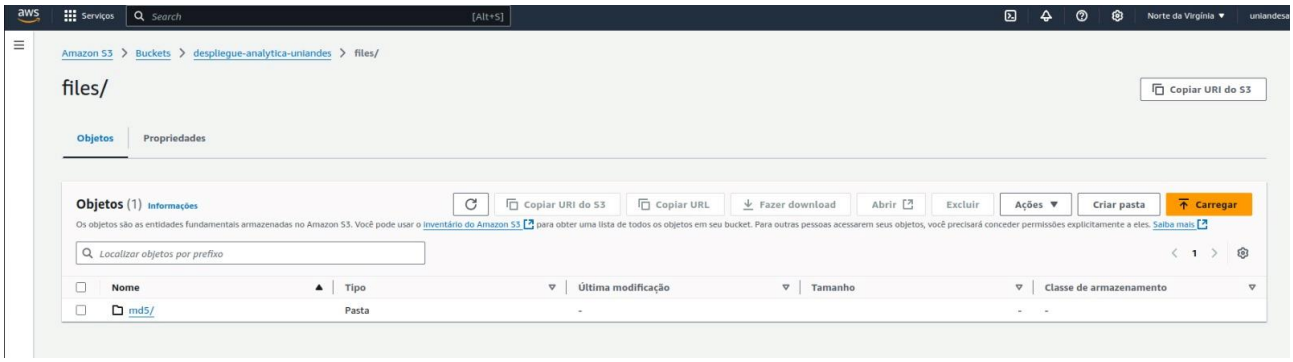


Figura 9 – Captura del hash md5 adentro de la carpeta files

Esos fueron los comandos para subir y aplicar el DVC:

dvc init

dvc remote add -d mi_remoto s3://despliegue-analytica-uniandes

dvc add files_csv

dvc add files_parquet

dvc push

git status

git add .

git commit -m "Añadiendo los archivos restantes al control de versiones en Git"

git push -u origin master