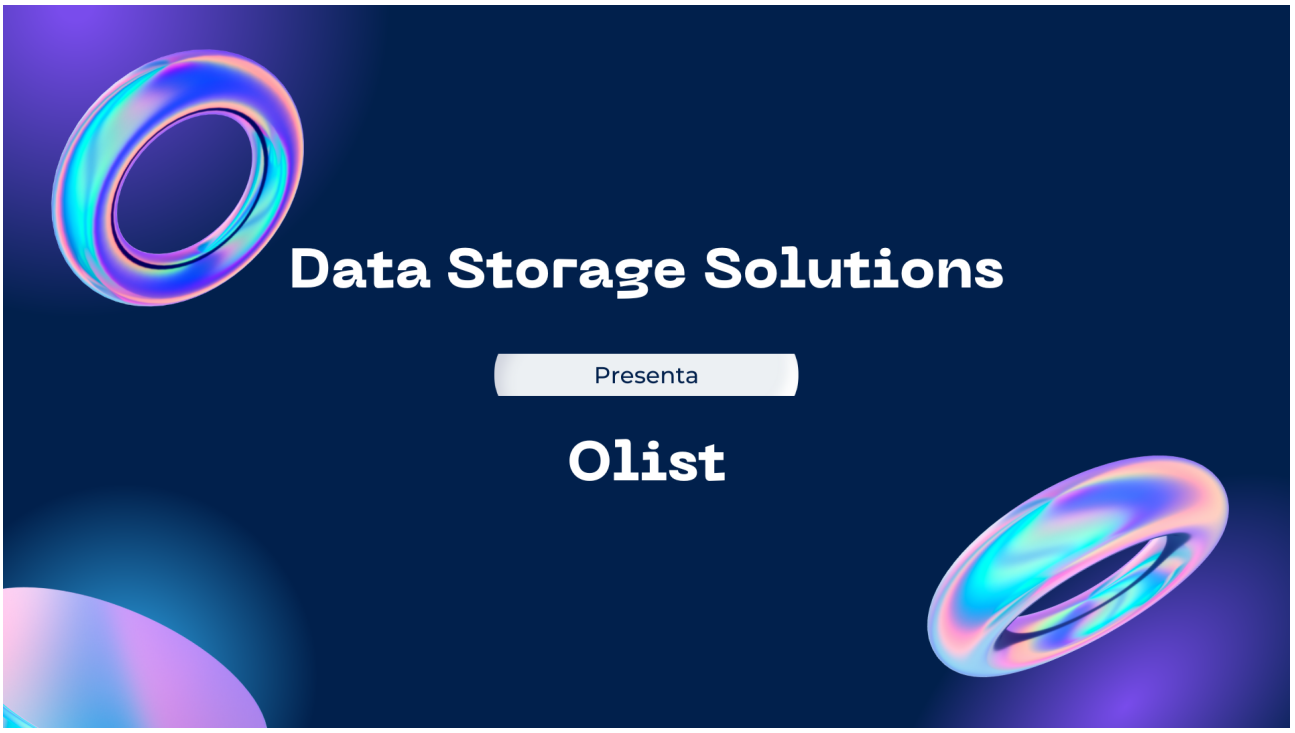


Propuesta de trabajo

Empresa: Olist

Castro P. - Fernandez G. - Pierotti L. - Tacchella F.

Índice

Índice 2

Introducción 3

Conformación del equipo: Roles y funciones 3

Metodología de trabajo 4

Objetivos 4

Alcances del proyecto 4

Stack tecnológico propuesto: 5

Cronograma de trabajo 6

SEMANA 1 6

SEMANA 2 6

SEMANA 3 7

SEMANA 4 7

Estimación de esfuerzos 8

Entregables 9

Desarrollo del proyecto 10

Diccionarios de datos 10

Informe preliminar de calidad del dato 12

Errores encontrados 12

Extracción, transformación y carga de los datos 15

Limpieza y normalización de los datos 16

Modelo entidad-relación 18

Machine learning 20

Análisis 22

Conclusiones 27

Recomendaciones 28

Reducción de los tiempos de entrega 28

Categorías de productos recomendables 29

Cantidad de distintas categorías de productos 31

Tipos de pagos 31

Fuentes: 34

Introducción

Nuestra empresa ha diseñado un plan de acción para conectar a las PYMEs con mercados más amplios, mejorando la experiencia del usuario. Este esquema incluye una investigación exhaustiva del mercado de comercio electrónico en Brasil con el objetivo de desarrollar estrategias que faciliten la inteligencia del negocio y encontrar soluciones innovadoras para ayudar a los usuarios a vender sus productos a una base de clientes más amplia. Utilizaremos información obtenida a través de una fuente de acceso libre conocida como "Olist" para llevar a cabo estas operaciones.

Conformación del equipo: Roles y funciones

Para maximizar la eficiencia de este proyecto, las tareas serán distribuidas entre los miembros del equipo en función de sus habilidades y preferencias individuales. Se permitirá una mayor flexibilidad en la asignación de tareas, permitiendo a los miembros del equipo trabajar en diferentes áreas según la complejidad del proceso. Además, los avances serán supervisados por todos los miembros del equipo para garantizar el progreso del proyecto.

Project Manager: Franco Tacchella
Data Engineers: Lautaro Pierotti y Guillermo Fernandez
Data Analytics: Franco Tacchella y Pablo Castro
Machine Learning: Guillermo Fernandez
DAG & Cloud Engineer: Lautaro Pierotti

Metodología de trabajo

El proyecto se gestionará utilizando la metodología Scrum. Se llevará a cabo una reunión diaria de equipo sin límite de tiempo para discutir posibles cambios en el proyecto y seguir el progreso de los objetivos establecidos. Además, al final de cada semana se realizará una revisión global de los avances del proyecto. También se programaron reuniones regulares con el cliente para informarle del progreso del trabajo y recibir su retroalimentación. El equipo tendrá un cronograma de tareas, pero estará abierto a adaptaciones en función de las necesidades y las ideas que surjan a lo largo del proceso, tanto del equipo como del cliente.

Objetivos

Objetivo principal:

- Conectar a pequeñas empresas (PYMEs) con mercados más grandes y mejorar la experiencia del usuario.

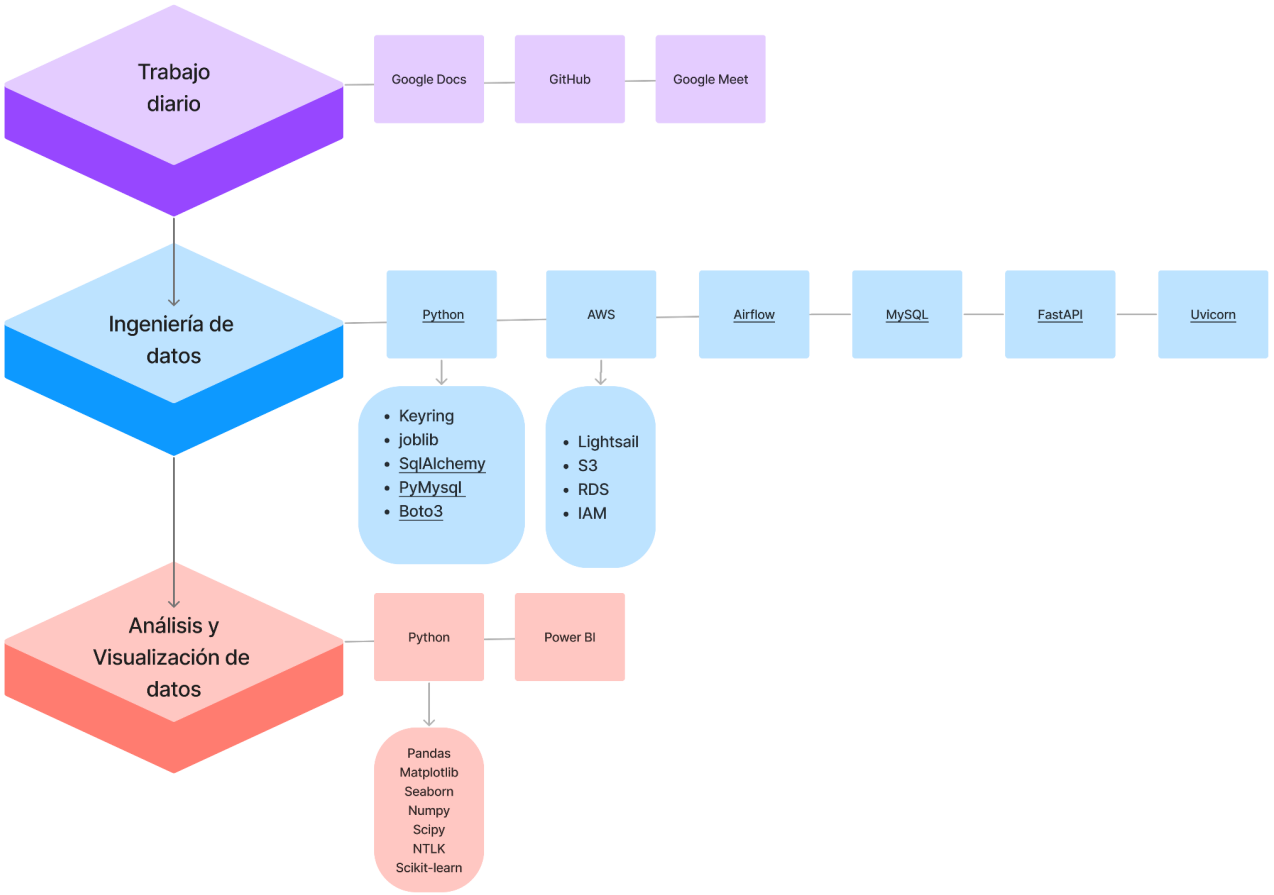
Objetivos específicos:

- Construir un Data Lake y Data Warehouse eficiente y escalable.
- Analizar a vendedores y su performance.
- Extraer KPIs que engloban el performance de cada vendedor.
- Ofrecer recomendaciones a la empresa en base a los datos disponibles.
- Armar un reporte para visualización escalable y en línea.

Alcances del proyecto

En este proyecto se busca proporcionar al cliente distintas herramientas para trabajar con sus datos y a su vez también un análisis de los datos disponibles, con los cuales se buscará ofrecer recomendaciones al cliente para la mejora de sus servicios.

Stack tecnológico propuesto:



Cronograma de trabajo

SEMANA 1

Durante la primera semana del proyecto nos enfocaremos en establecer un plan detallado para llevar a cabo el trabajo y en recopilar y analizar los datos necesarios para llevar a cabo el proyecto. Es importante tener una buena comprensión de los datos antes de comenzar a trabajar en el desarrollo del proyecto.

Lunes y Martes:

- Análisis y entendimiento del proyecto
- Definición de objetivos y alcance

Miércoles y Jueves:

- Análisis exploratorio de los datos (EDA)

Viernes:

- Reunión con el product owner y propuesta de proyecto

SEMANA 2

Sábado y Domingo:

- Definir Infraestructura y elección del servicio en la nube

Lunes, Martes y Miércoles:

- Creación del bucket de S3 y la base de datos en amazon RDS
- Armado de airflow en la máquina virtual de amazon lightsail

Jueves:

- Testeo de la infraestructura en la nube y últimos cambios

Viernes:

- Presentación de la infraestructura funcionando al product owner

SEMANA 3

Sábado a Miércoles:

- Construcción de un dashboard preliminar
- Armado de modelo de machine learning

Jueves:

- Organizado de storytelling

Viernes:

- Presentación del dashboard preliminar al product owner

SEMANA 4

Sábado y domingo:

- Finalización del dashboard
- Análisis y búsqueda de mejoras para ofrecer recomendaciones
- Armado de powerpoint para presentacion final

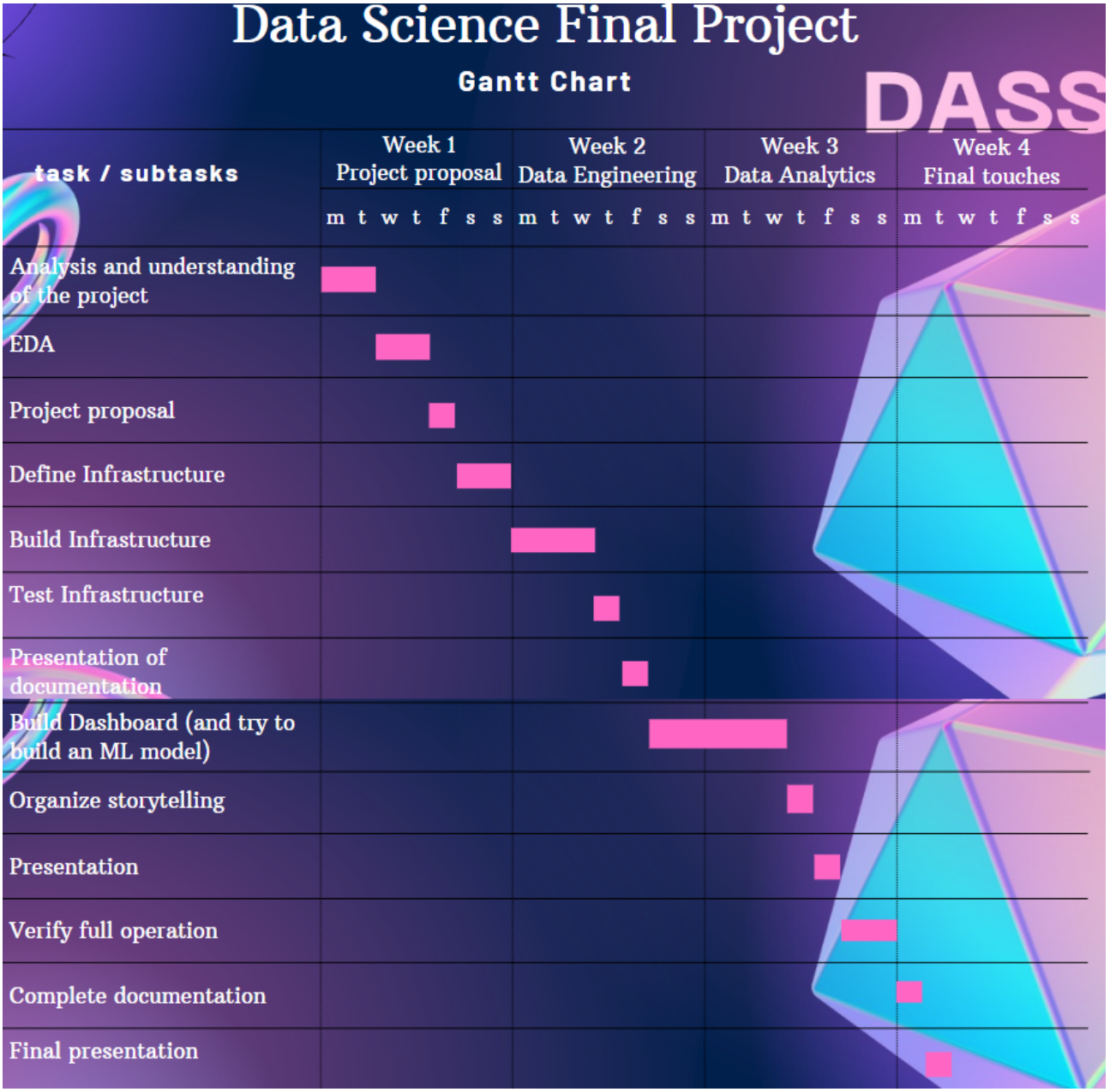
Lunes:

- Reunión final para organizar storytelling y finalizar detalles

Martes:

- Presentación del proyecto

Estimación de esfuerzos



Entregables

- Todo el código organizado en un repositorio en GitHub
- Base de Datos escalable
- ETL automatizado
    - Carga auto-incremental configurada
    - Tablas de auditoría
    - Tablas nuevas creadas para el análisis
- Análisis de vendedores y su desempeño
- Recomendaciones de negocio
- KPIs extraídos
- Dashboard de Power BI
- Modelo de ML sobre tiempos de entrega y reviews
- API para consultas de performance de vendedores

Desarrollo del proyecto

Diccionarios de datos

Dataset	Closed_deals		
Columna	Detalle	Tipo de dato	Ejemplo
mql_id	ID cliente potencial	Alfanumerico	a0604c9d9ef23fbf7cb7be5091201041
seller_id	ID del vendedor	Alfanumerico	b7140ce94c4514bf136a2c3f98e0476c
sdr_id	informacion incompleta	Alfanumerico	b90f87164b5f8c2cfa5c8572834dbe3f
sr_id	informacion incompleta	Alfanumerico	d3d1e91a157ea7f90548eef82f1955e3
won_date	informacion incompleta	Fecha y hora	2018-07-31 20:01:32
business_segment	Segmento del negocio	Texto	audio_video_electronics
lead_type	Tipo de cliente	Texto	online_small
lead_behaviour_profile	informacion incompleta	Texto	cat
business_type	Tipo de negocio	Texto	reseller
declared_monthly_revenue	informacion incompleta	Numerico decimal	0.0
Dataset	Customers		
Columna	Detalle	Tipo de dato	Ejemplo
customer_id	ID cliente	Alfanumerico	bc3c9c45fe3fd83f49adbcbf50daa3da
customer_unique_id	ID unico del cliente	Alfanumerico	e72bbc364013bd2f23ace1b4e3c43be6
customer_zip_code_prefix	Codigo postal del cliente	Numerico entero	89252
customer_city	Ciudad del cliente	Texto	jaragua do sul
customer_state	Estado del cliente	Texto	SC
Dataset	Geolocation		
Columna	Detalle	Tipo de dato	Ejemplo
geolocation_zip_code_prefix	Codigo postal	Numerico entero	31035
geolocation_lat	Latitud	Numerico decimal	-19.898.344
geolocation_lng	Longitud	Numerico decimal	-43.921.363
geolocation_city	Ciudad	Texto	belo horizonte
geolocation_state	Estado	Texto	MG
Dataset	Marketing		
Columna	Detalle	Tipo de dato	Ejemplo
mql_id	ID cliente potencial	Alfanumerico	434c2eb8627ed4e1a0a4f0ee5d6022aa
first_contact_date	Fecha primer contacto	Fecha	2018-03-22
landing_page_id	informacion incompleta	Alfanumerico	0d6bc3c00e4e64927cae2e8d9c6a0b9b
origin	Origen contacto	Texto	paid_search
Dataset	Order_items		
Columna	Detalle	Tipo de dato	Ejemplo
order_id	ID orden de compra	Alfanumerico	ec1ecce6ed2f4a351a045a8de255e3af
order_item_id	ID identificacion articulo	Numerico entero	1
product_id	ID producto	Alfanumerico	35afc973633aaeb6b877ff57b2793310
seller_id	ID vendedor	Alfanumerico	4a3ca9315b744ce9f8e9374361493884
shipping_limit_date	Fecha limite de envio	Fecha y hora	2017-09-21 10:10:20
price	Precio	Numerico decimal	89.90
freight_value	Precio flete	Numerico decimal	14.95

Dataset	Order_payments		
Columna	Detalle	Tipo de dato	Ejemplo
order_id	ID orden de compra	Alfanumerico	647255bbedcdf748e7496180374b0dfe
payment_sequential	secuencia unificacion de medios de pago	Numerico entero	1
payment_type	TIpo de medio de pago	Texto	credit_card
payment_installments	numero de cuotas	Numerico entero	1
payment_value	valor del pago	Numerico decimal	28.09
Dataset	Order_reviews		
Columna	Detalle	Tipo de dato	Ejemplo
review_id	ID reseña	Alfanumerico	a51b17cc0ae35deec0466cbf057b6700
order_id	ID orden de compra	Alfanumerico	d3d8fd64df1cf1428ea9f88aaabb4713
review_score	Calificacion reseña	Numerico entero	3
review_comment_title	Titulo reseña	Texto	Talvez recomendaria
review_comment_message	Mensaje reseña	Texto	A empresa deveria responder aos e-mail dos cli...
review_creation_date	Fecha envio reseña	Fecha y hora	2018-05-17 0:00:00
review_answer_timestamp	Fecha respuesta reseña	Fecha y hora	2018-05-21 17:40:50

Dataset	Orders		
Columna	Detalle	Tipo de dato	Ejemplo
order_id	ID orden de compra	Alfanumerico	bb01789fc0271409b394c7b283e9dd29
customer_id	ID cliente	Alfanumerico	ac3e01509ab5a9e7bb1bb344048ef81d
order_status	Estado pedido	Texto	delivered
order_purchase_timestamp	fecha compra	Fecha y hora	2018-08-07 12:08:49
order_approved_at	fecha aprobacion compra	Fecha y hora	2018-08-07 12:24:47
order_delivered_carrier_date	fecha entrega socio logistico	Fecha y hora	2018-08-09 14:22:00
order_delivered_customer_date	fecha entrega pedido	Fecha y hora	2018-08-14 22:13:42
order_estimated_delivery_date	fecha estimada de entrega informada	Fecha y hora	2018-08-28 0:00:00
Dataset	Products		
Columna	Detalle	Tipo de dato	Ejemplo
product_id	ID producto	Alfanumerico	093f7389fa2eccda5e86add8da4aa19e
product_category_name	nombre categoria producto	Texto	cama_mesa_banho
product_name_lenght	longitud del nombre del producto	Numerico decimal	40.0
product_description_lenght	longitud de la descripcion del producto	Numerico decimal	718.0
product_photos_qty	cantidad de fotos del producto	Numerico decimal	1.0
product_weight_g	peso del producto en gramos	Numerico decimal	2000.0
product_length_cm	longitud del producto en cm	Numerico decimal	38.0
product_height_cm	altura del producto en cm	Numerico decimal	20.0
product_width_cm	ancho del producto en cm	Numerico decimal	25.0
Dataset	Sellers		
Columna	Detalle	Tipo de dato	Ejemplo
seller_id	ID vendedor	Alfanumerico	282c7480173bb9c01dd41cc739fec010
seller_zip_code_prefix	codigo postal vendedor	Numerico entero	4795
seller_city	ciudad del vendedor	Texto	sao paulo
seller_state	estado del vendedor	Texto	SP
Dataset	Product_category_name_translation		
Columna	Detalle	Tipo de dato	Ejemplo
product_category_name	nombre categoria producto	Texto	moveis_quarto
product_category_name_english	nombre categoria producto en ingles	Texto	furniture_bedroom
Dataset	Valoracion		
Columna	Detalle	Tipo de dato	Ejemplo
seller_id	ID vendedor	Numerico entero	629
seller_city	Ciudad vendedor	Texto	maua
seller_state	Estado vendedor	Texto	SP
distinct_prod	ID producto	Numerico entero	58
delivery_avg	promedio de envios	Numerico decimal	-6.981.012
review_avg	promedio de reseñas	Numerico decimal	4.431.034
total_orders	ID orden	Numerico entero	58
total_income	precio total	Numerico decimal	2548.80

Informe preliminar de calidad del dato

Se importaron los datasets de Olist los cuales son de acceso público. Los datos constan de 11 archivos .csv de los cuales se utilizaron 10. Estos archivos se encuentra en la carpeta "Datasets/Datasets_original”

Errores encontrados

A continuación se detallan los errores encontrados en el análisis exploratorio de los datos de cada dataset

• **olist_customers_dataset**

Esta tabla contiene 5 columnas, con 99441 valores en cada una. Faltan datos de customers_unique_id que no hay referencia de cómo fueron creados esos codigos. Esta tabla se puede utilizar para saber de qué ciudades provienen los consumos. No contiene datos nulos.

• **olist_geolocation_dataset**

Esta tabla cuenta con 5 columnas y 1000163 valores en cada una. Esta tabla puede servir para ubicar las ciudades correctamente en un mapa del Dashboard. No contiene datos nulos.

• **olist_order_items_dataset**

Esta tabla contiene 7 columnas con 112650 valores en cada una. Se puede utilizar para conocer cuál es el producto más vendido como así también cuánto facturó, cuál es el vendedor con más órdenes y por cuanto. No contiene valores nulos. Presenta outliers.

• olist_order_payments_dataset

Esta tabla cuenta con 5 columnas y 103886 valores en cada una. Se podría averiguar el comportamiento de los pagos. La utilidad dependerá de la propuesta del PO. No contiene valores nulos.

• **olist_order_reviews_dataset**

Esta tabla cuenta con 7 columnas y 99224 valores en cada una. La utilidad de esta tabla dependerá del objetivo propuesto. Contiene valores nulos en dos columnas.

• **olist_orders_dataset**

Esta tabla contiene 8 columnas y 99441 valores en cada una. La utilidad de esta tabla dependerá del objetivo. Contiene pocos valores nulos en 3 columnas.

• **olist_products_dataset**

Esta tabla contiene 9 columnas y 32951 valores en cada una. La utilidad de esta tabla dependerá del objetivo. Presenta muy pocos valores nulos en la mayoría de las columnas.

• **olist_sellers_dataset**

Esta tabla contiene 4 columnas con 3095 valores en cada una. No presenta datos nulos. Se podría conocer cuál es la ciudad con mayor cantidad de órdenes como así también el monto mayor en ventas.

• **product_category_name_translation**

Esta tabla contiene 2 columnas con 71 valores en cada una. No presenta errores.

• **olist_marketing_qualified_leads_dataset**

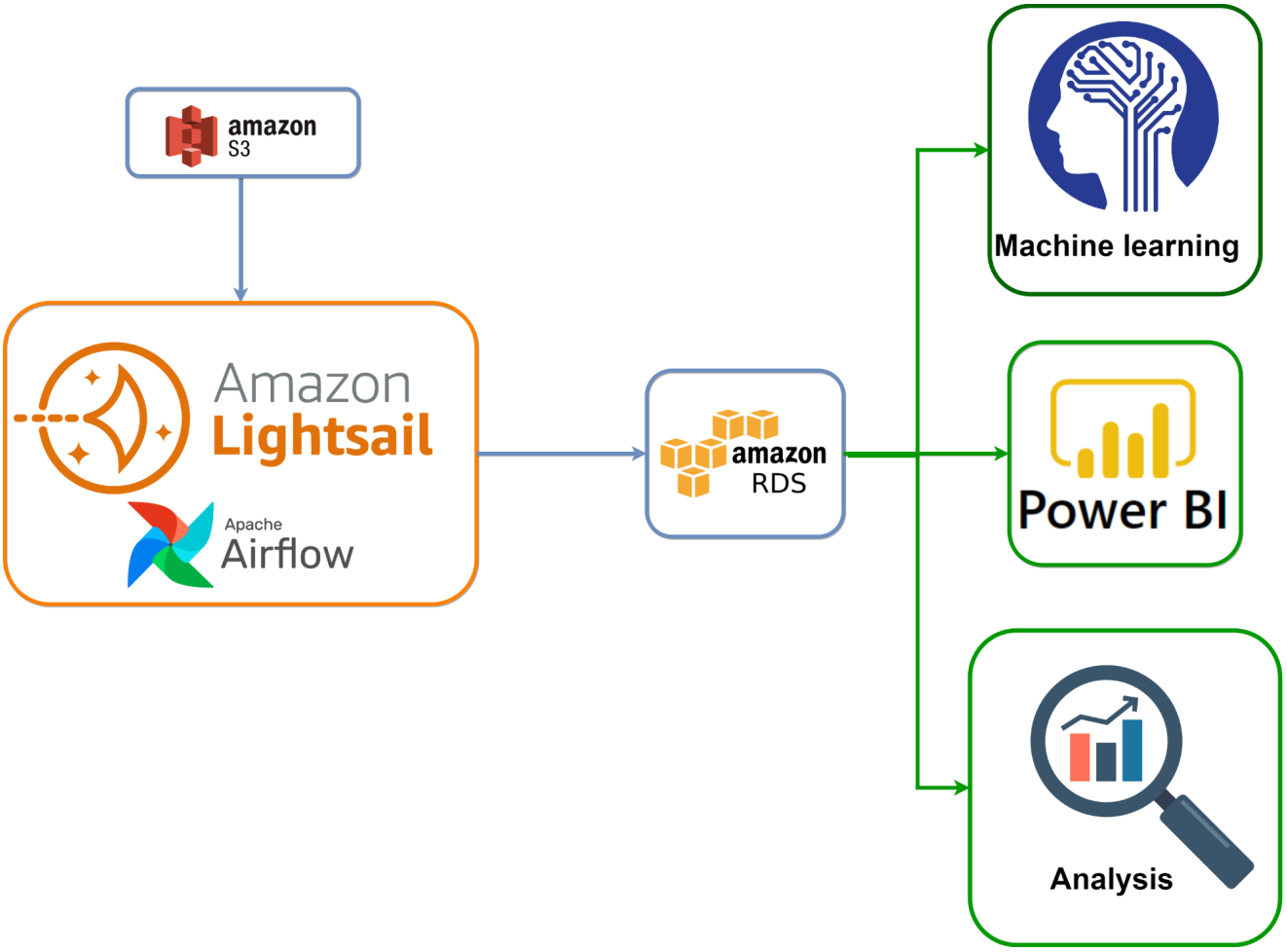
Esta tabla contiene 4 columnas, con un total de 8000 valores en cada una de ellas. Esta tabla no aporta datos significativos para el análisis que queremos realizar. Contiene pocos valores nulos en una columna.

• **olist_closed_deals_dataset**

Esta tabla consta de 14 columnas de datos, con un total de 842 valores en cada una de ellas. Las columnas has_company, has_gtin, average_stock y declared_product_catalog_size tienen más del 90 % de sus datos faltantes. Esta tabla no aporta datos muy relevantes. No se encontraron registros duplicados. Los nombres de los campos se encuentran en inglés.

Extracción, transformación y carga de los datos

Una vez hecho el análisis preliminar de los datos, se procede a la transformación y carga de los mismos. Primero cargamos los datasets a trabajar en la nube usando el servicio de Amazon S3, luego diseñamos un script de python que contiene las funciones de carga y transformación y luego se utilizó airflow en una máquina virtual en la nube, utilizando el servicio Amazon Lightsail, para armar un pipeline con este script de python. Elegimos configurar el pipeline para que se ejecute diariamente desde la nube, el cual va a tomar los datos del bucket de S3, los va a transformar y luego los va a cargar a la base de datos en amazon RDS



Limpieza y normalización de los datos

A cada dataset se le eliminaron las filas duplicadas.

Al dataset “Closed_deals” se le cambió el tipo de datos en las columnas 'has_gtin' y 'has_company' al tipo de dato float64.

Al dataset Orders se le agregó un nueva columna llamada “tiempo_entrega” la cual se obtiene calculando la diferencia en días de las columnas “order_approved_at” y “order_delivered_customer_date”.

La transformación principal se llevó a cabo al combinar los datasets “orders”, ”order_reviews”, ”order_payments”, ”order_items”, “sellers” y “products” en una nueva tabla llamada datasets_combinados.

A esta tabla combinada se le creó una nueva columna llamada avg_income_month la cual contiene el cálculo del promedio de ingresos mensuales de cada vendedor.

Luego se crea una nueva tabla llamada Evaluation y se toman las columnas “seller_id”, “seller_city”, “product_id”, “product_category_name”, “tiempo_entrega”, “review_score”, “order_id”, “payment_value” y “avg_income_month” del dataframe “datasets_combinados”.

Se realizan los respectivos cálculos y se cambian los siguiente nombres de columnas:

'product_id'→'distinct_prod'

'Tiempo_entrega'→'delivery_avg'

'product_category_name'→'distinct_categories'

'review_score'→'review_avg'

'order_id'→'total_orders'

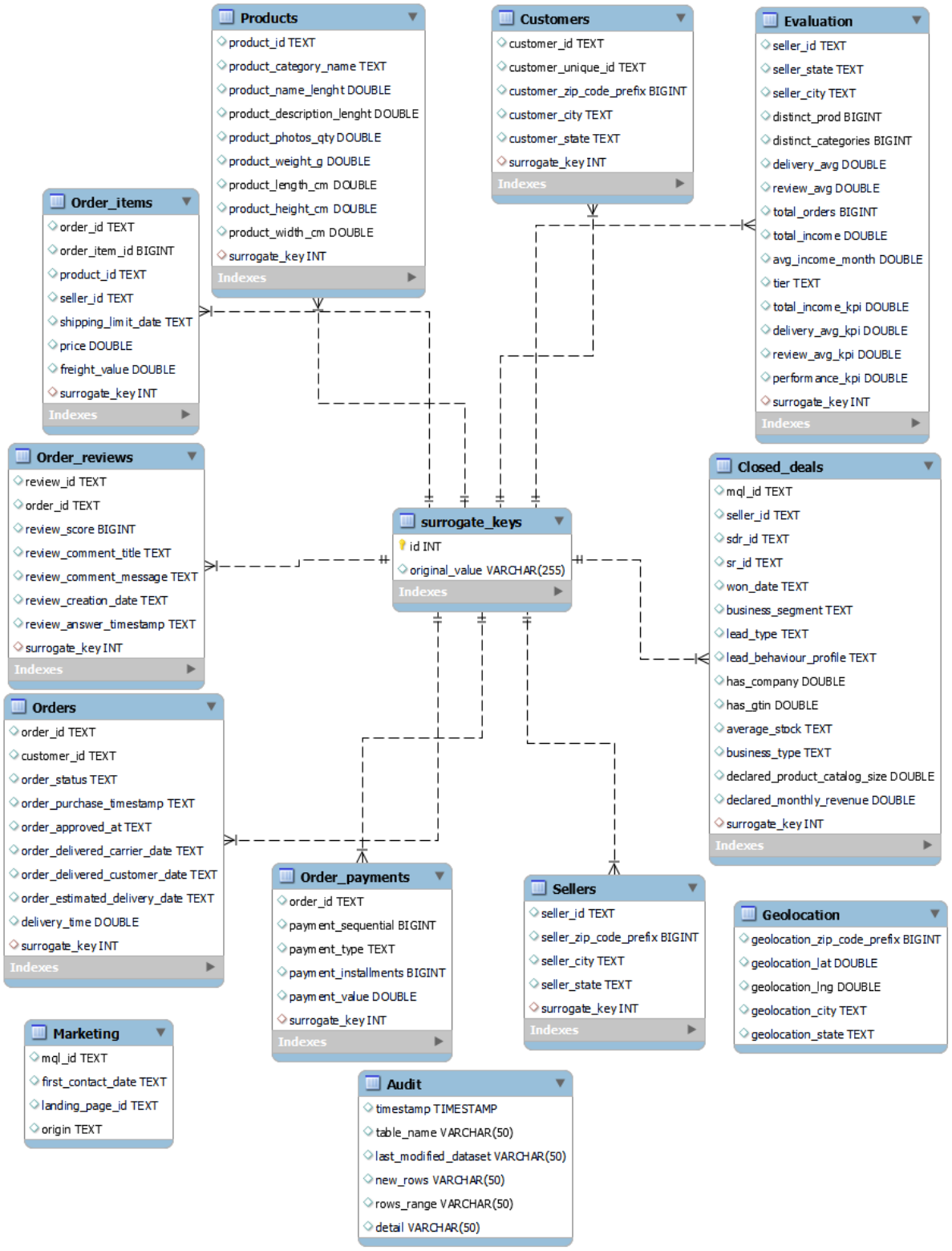
'payment_value'→'total_income'

Las filas con valores de delivery_avg mayores a 50 días se eliminaron al ser pocas filas y considerarse outliers, Luego se eliminaron 125 filas que contengan valores nulos en alguna columna.

Por último se crearon funciones que dividen la tabla Evaluation en distintas categorías y asignan distintos score a los vendedores en base a distintas métricas. Se agregaron estos valores en 5 nuevas columnas:

- **tier** = Esta columna asigna una categoría al vendedor
- **total_income_kpi=** Calculada en base a la columna “total_income” que contiene los ingresos totales del vendedor y segun la categoria del vendedor
- **delivery_avg_kpi=** Calculada según la columna delivery_avg y según la categoría asignada al vendedor.
- **review_avg_kpi=** Se calcula con la columna review_avg la cual contiene el promedio de review_score del vendedor y en base a la categoría asignada al vendedor.
- **performance_score=** Es un promedio de los tres kpi calculados anteriormente y se toma como el score final que se le asigna al vendedor para evaluar su rendimiento

Modelo entidad-relación



Para el modelo de entidad relación se creó una tabla de claves subrogadas, la cual asigna a los ID de las distintas tablas, los cuales tienen valores alfanuméricos, un nuevo número entero como ID en una nueva columna para cada tabla llamada subrrogate_key.

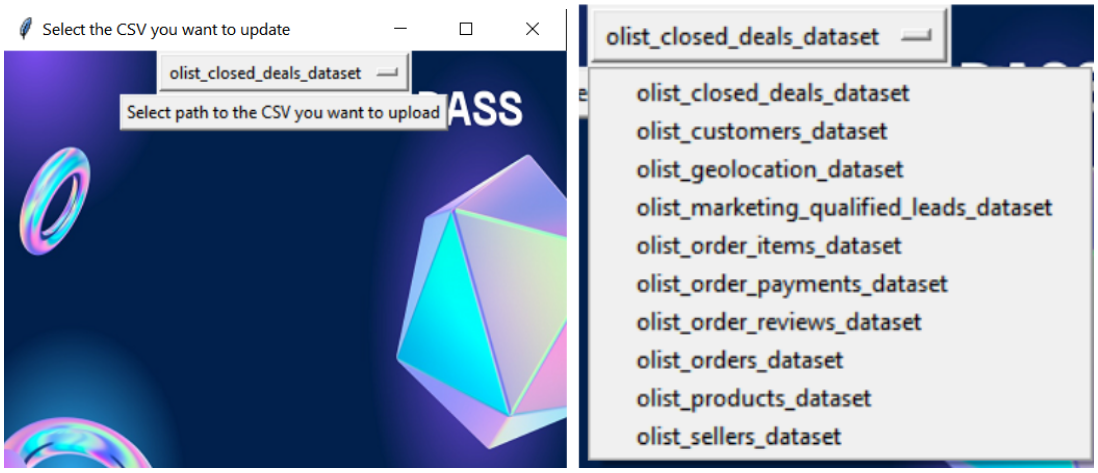
Esto nos permite conectar todas las tablas y a su vez mantener los ID originales, los cuales están encriptados, de forma que el cliente puede volver a acceder a esos datos en caso de necesitarlo desencriptando los valores del ID.

A su vez diseñamos el sistema de carga automática para ser compatible con una carga incremental de datos. De esta forma, al actualizar los datasets del bucket de S3 el pipeline de airflow automáticamente cargará solo las nuevas filas a la base de datos, evitando así cargar filas repetidas. El pipeline a su vez comprobará la última fecha de modificación de los datasets cargados en la nube, de manera que solo va a operar con los que presenten modificaciones de la última fecha de modificación de la que se tenga registros, para cada dataset en específico.

Por último el pipeline de airflow va a cargar informacion en una tabla de auditoria, la cual contienen las siguientes columnas:

- **Timestamp=** fecha y hora en la que se realizó la operación en la base de datos.
- **table_name=** nombre de la tabla en la que se realiza la operación.
- **last_modified_dataset=** última fecha de modificación del archivo .csv que se encuentra en el bucket de S3, de la tabla con la que se está operando.
- **new_rows=** número de nuevas filas agregadas a la tabla
- **rows_range=** índice del rango de nuevas filas agregadas a la tabla.

También diseñamos un script para facilitar la actualización de los datasets que se encuentran en el bucket de S3.



Este script solicita primero que el usuario elija el dataset que quiere actualizar y luego seleccione la ubicación del archivo .csv para subir a la nube. El script arrojará un mensaje de error en caso de que las columnas del archivo .csv elegido no coincidan con las columnas del archivo .csv subido en la nube, para evitar así que se sobreescriban archivos equivocados.

Machine learning

Para el modelo de machine learning primero se cargaron los datasets originales “olist_order_items_dataset.csv” , “olist_order_reviews_dataset.csv” y “olist_orders_dataset.csv”

Luego se agregó la columna de delivery time al dataframe Orders y se combinaron los 3 dataframes en uno llamado DF.

A este dataset combinado se le agregó una nueva columna donde se calcula una variable categórica en base a la columna review_score, esta variable asigna un valor 0 a los review score menores o iguales a 3 y 1 a los review score mayores a 3. Luego se eliminaron filas duplicadas del dataframe y filas con valores nulos.

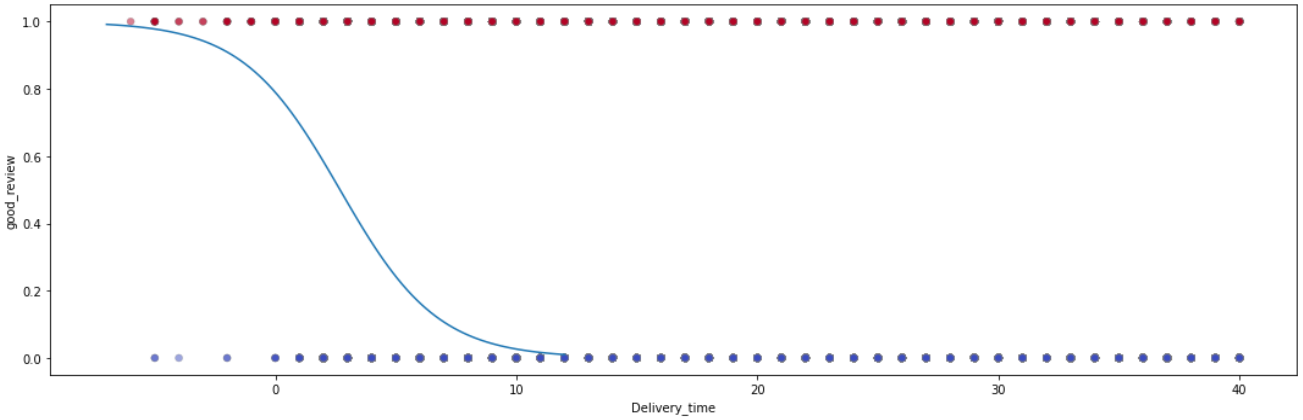
Se eliminaron outliers usando la librería scipy y luego se usó scikit learn para hacer un test split de los datos, donde se tomó el 25% de los datos para test y el resto para entrenamiento.

Se armaron 4 pipelines para probar distintos modelos y ver cuál ofrecia mejores resultados. Cada pipeline aplica standar scaler a los datos y los 4 modelos testeados junto con su accuracy fueron:

- Linear Regression accuracy en test: 0.04970180902756638
- Logistic Regression accuracy en test: 0.7847160603371783
- KNeighbors Classifier accuracy en test: 0.7495563442768411
- Decision Tree Classifier accuracy en test: 0.7866385684708667

El modelo elegido por lo tanto fue el árbol de clasificación el cual tiene un accuracy de 0.79 un recall de 0.98 y un F1-score de 0.88.

En el siguiente gráfico se puede observar que la mayoría de los datos entran dentro de la clasificación del modelo por lo que tiene suficiente precisión para darle uso.

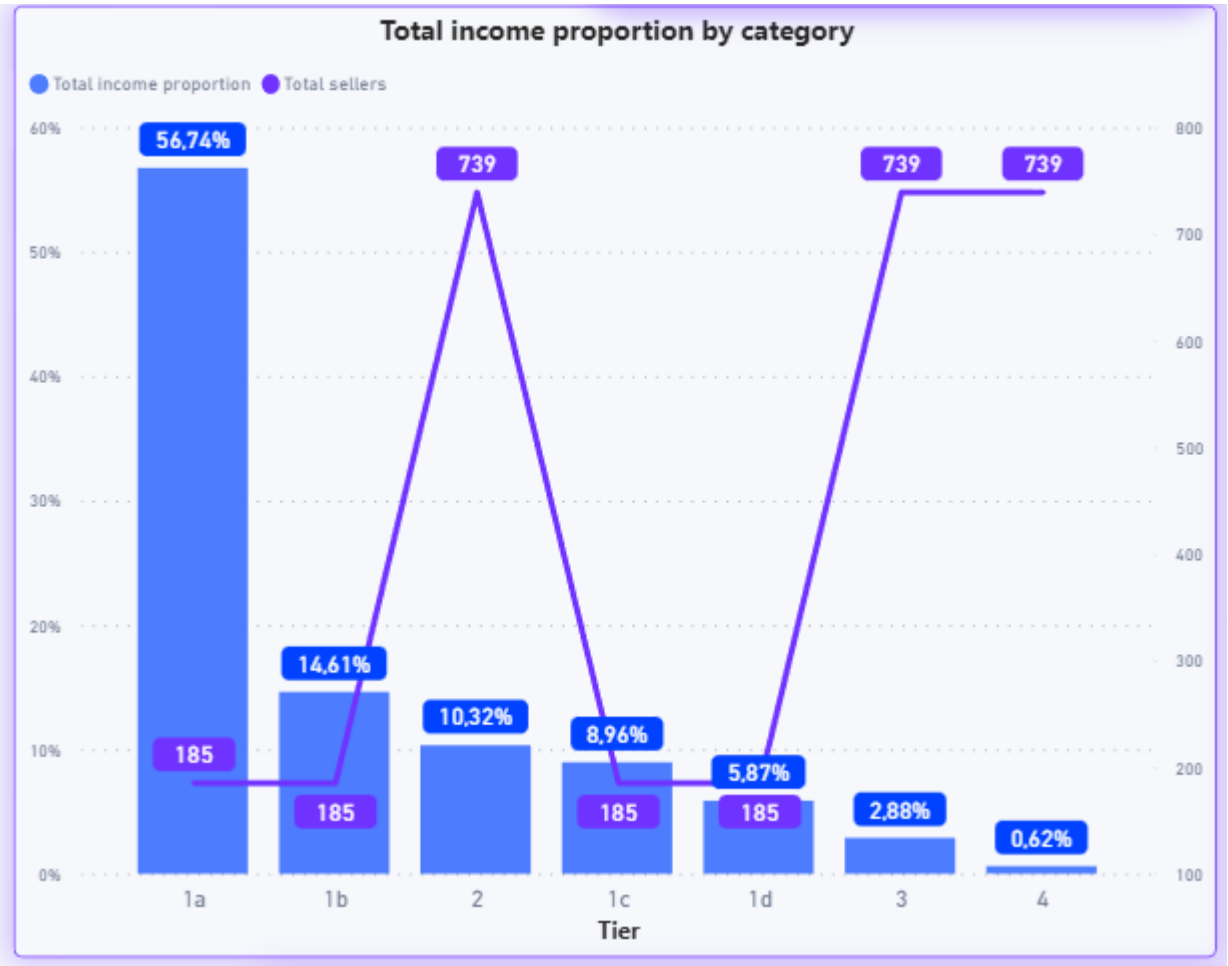


Por último se hace una predicción sobre la columna objetivo usando el modelo entrenado, para predecir el review score en base al tiempo de entrega promedio, el cual se usó como feature para el modelo.

Con la predicción hecha, se obtuvo la información de que a partir de los 28 días es mucho más probable obtener un score de 3 o menos, por lo que ahora tenemos un límite más claro sobre el cambio de review score según el tiempo de entrega, y se puede dar esta información a los vendedores que tengan como objetivo mejorar sus review score.

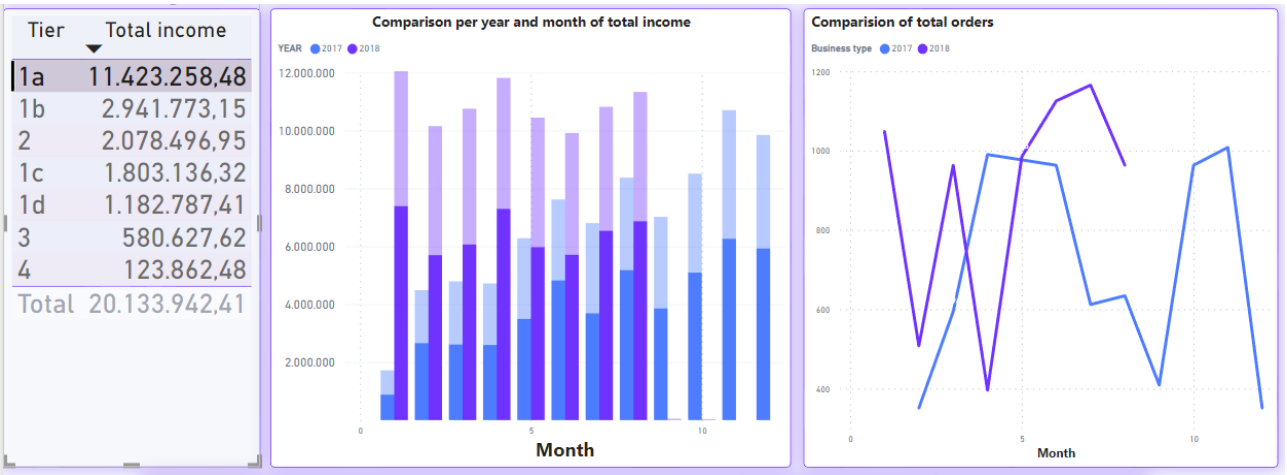
Análisis

En la tabla Evaluation, los vendedores están divididos en categorías según sus ingresos. En el siguiente gráfico se puede ver que en la categoría 1-A, la cual se asigna a los vendedores con mayores ingresos, se encuentra el 56,74% de los ingresos totales de la plataforma en el año 2017-2018 y a pesar de que solo se encuentra el 6,25% de los vendedores.



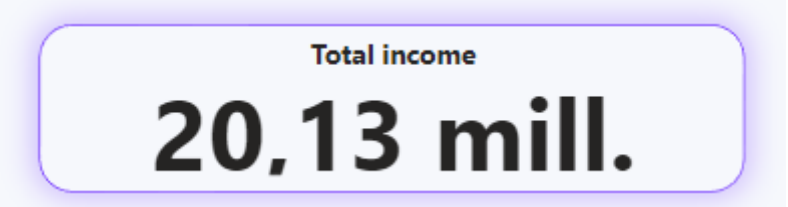
Todos estos datos son calculados en base a un total de 117 mil órdenes, las cuales su mayor parte las aporta la categoría 1-A.

A continuación se observa con mejor precisión la cantidad de ingresos totales que aporta cada categoría en la tabla de la izquierda, en el gráfico central se ve el total de ingresos de la categoría 1-A comparada sobre el total de ingresos de todas las categorías y en el gráfico de la derecha se ve una comparación de las órdenes de la categoría 1-A entre los años 2017 y 2018.

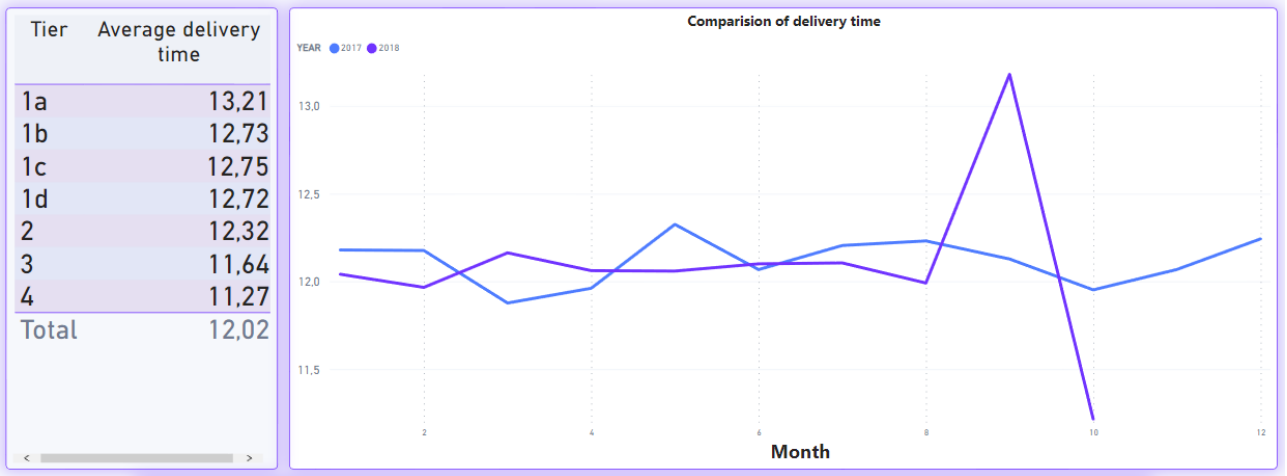


Esto nos indica que la empresa debería poner atención en los vendedores de esta categoría para ver qué es lo que los destaca por sobre los demás en términos de ingresos y cantidad de órdenes.

Por último, para finalizar con el análisis de ingresos totales, se ve que en los años estudiados el total de ingresos fue de 20.130.000 Reales.

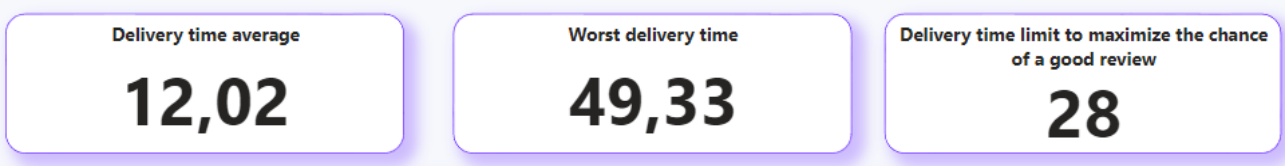


Lo siguiente que se hizo fue analizar la evolución de los tiempos de entrega en los años 2017 y 2018 (en días), en la tabla de la izquierda se ve la comparación del promedio de tiempos de entrega por categoría, se ve que la categoría 1-A tiene mayor promedio de tiempo de entrega, lo cual se relaciona con la mayor cantidad de órdenes que manejan.



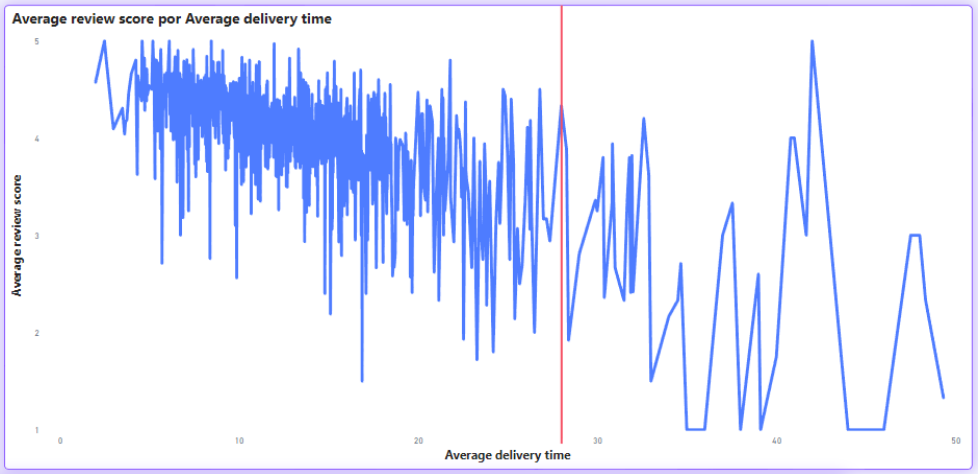
El tiempo de entrega promedio de todos los vendedores fue de 12,02 días, el peor tiempo de entrega registrado, habiendo eliminado outliers es de 49,33 días.

28 días es el límite de tiempo de entrega recomendado según la información obtenida con el modelo de machine learning, para disminuir la probabilidad de que el cliente deje un review score menor o igual a 3.

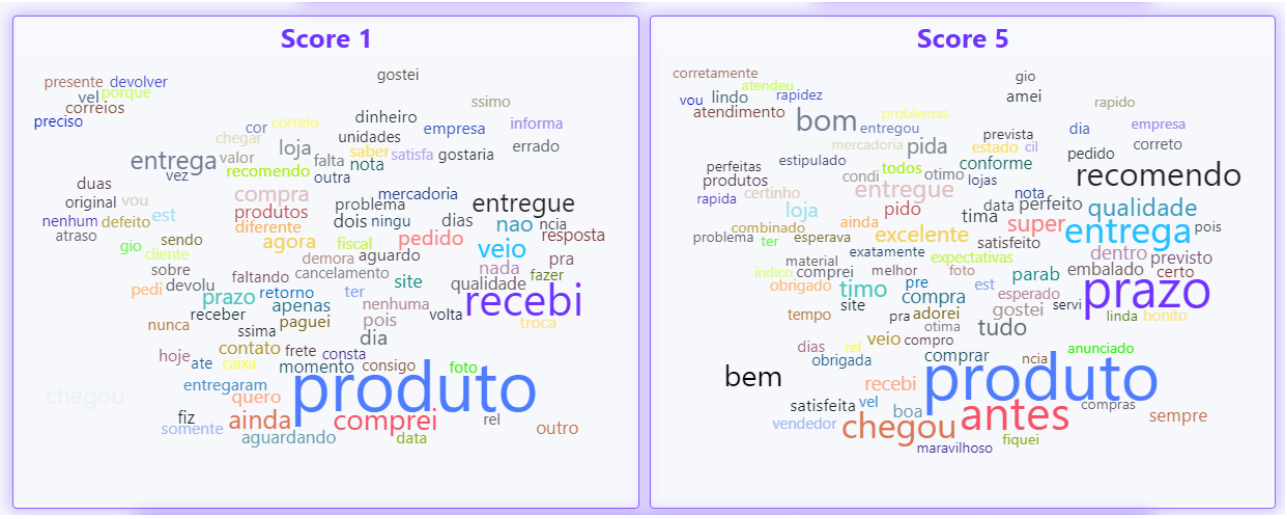


En el siguiente gráfico se puede ver con mejor claridad el límite de los 28 días (línea vertical roja) y como afecta al promedio de score y en la tabla de la izquierda tenemos un análisis del promedio de review score por categoría.

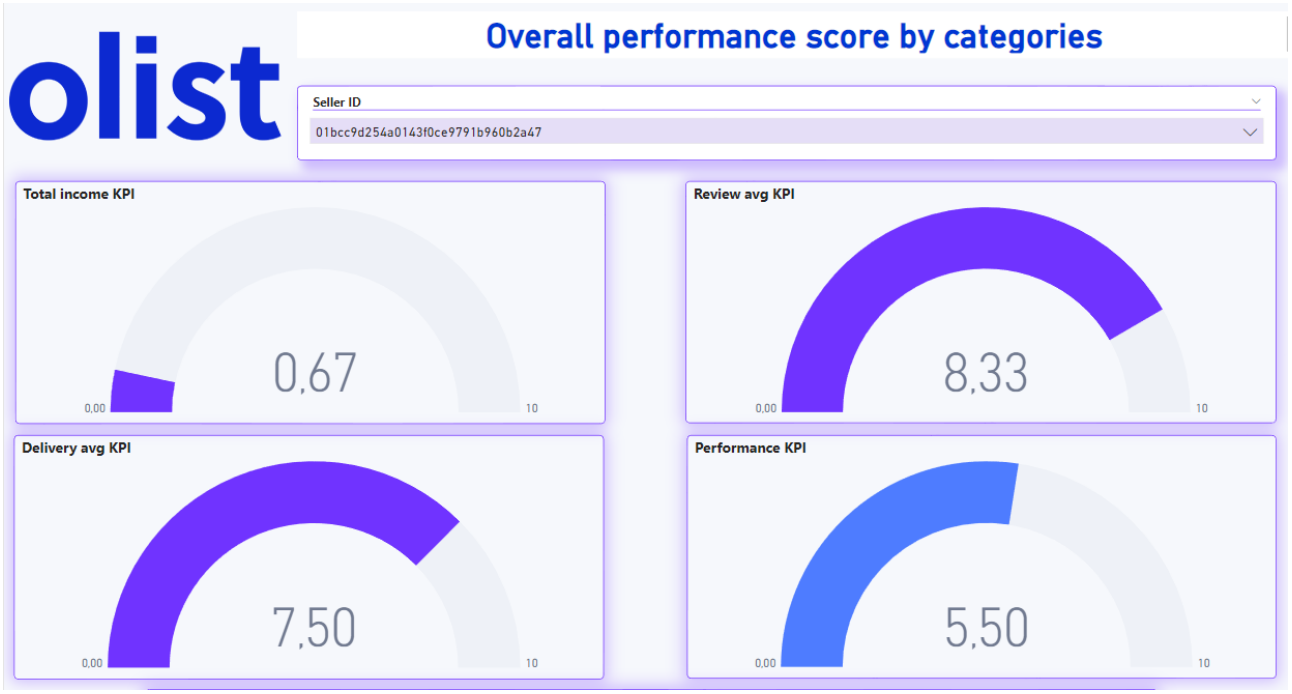
Tier	Average review score
1a	3,98
1b	4,00
1c	4,07
1d	4,02
2	4,04
3	4,10
4	4,18
Total	4,08



Lo siguiente que se hizo fue un análisis de los comentarios en los review y se hicieron dos nubes de palabras con las palabras claves más frecuentes para los score 1 y los score 5. Se puede ver que una de las cosas que más influye en los score es nuevamente el tiempo de entrega, y que en las review con score 5 los clientes recomiendan al vendedor o producto, lo que se traduce en más ventas.



Por último se agregó un filtro de Seller ID donde se pueden consultar los 4 KPIs de cada vendedor independientemente de su categoría.



Conclusiones

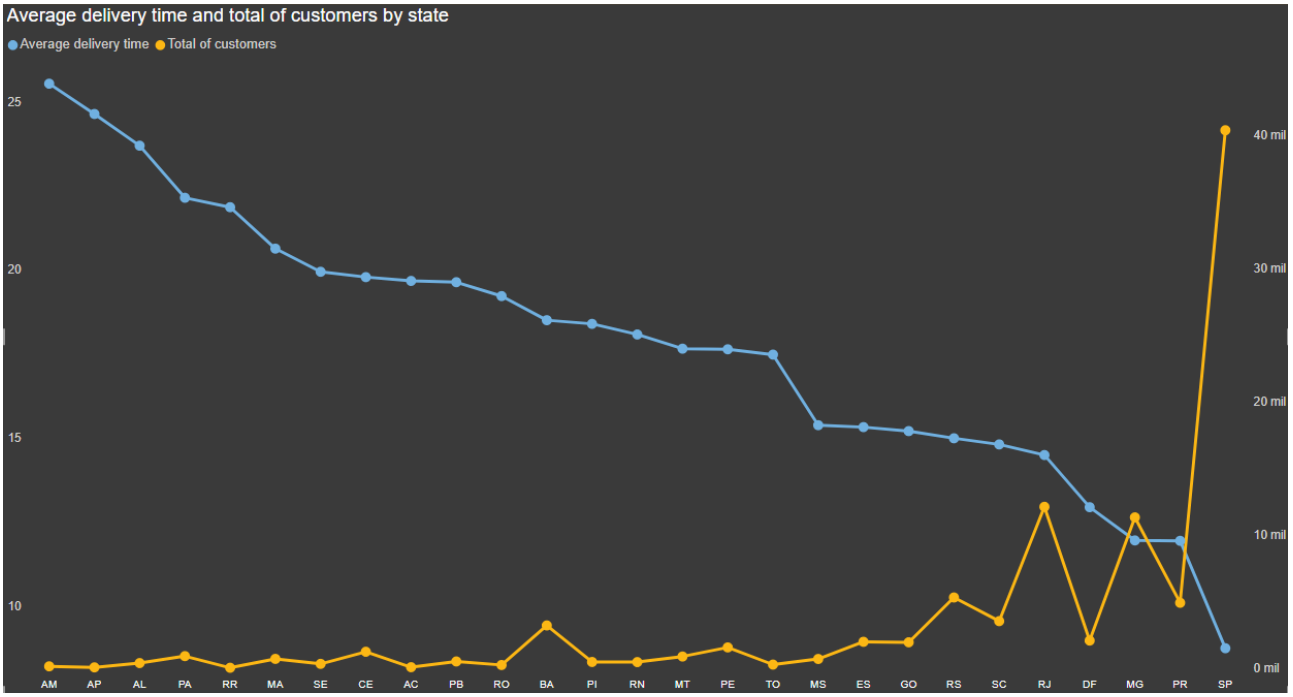
A partir del análisis de la información, podemos concluir lo siguiente:

- Menos del 6% de los clientes aportan más del 55% de los ingresos totales, y también aporta más del 50% de las órdenes.
- Si extendemos la cantidad de clientes hasta un 25%, los ingresos totales superan el 85%, al igual que las órdenes.
- Las ventas no aumentan en época de Carnaval y Navidad, lo que indica falta de promoción de las festividades.
- Las entregas con demoras grandes son, en el 80% de los casos, causantes de malas valoraciones.
- La ciudad y estado con mayor cantidad de órdenes es São Paulo
- Hay entregas con demoras muy por encima de la media, llegando a extremos de más de 100 días.

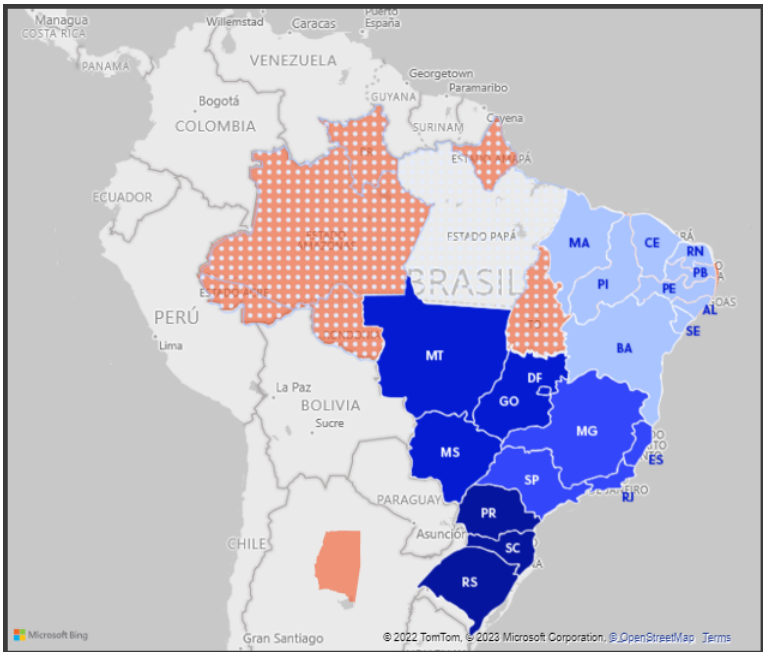
Recomendaciones

Reducción de los tiempos de entrega

Al analizar la cantidad de clientes y la relación con los tiempos de entrega en los distintos estados de brasil se puede ver que al mejorar los tiempos de entrega hay mayor cantidad de clientes, por lo que reducir los tiempos de entrega no solo mejoraría el servicio para los clientes sino que también aumentará la cantidad de estos.



En los últimos años Olist adquirió la empresa de logística PAX para mejorar la calidad del servicio de entrega de productos. En el siguiente mapa se puede ver en azul la zona en donde actualmente funciona el servicio y en naranja las zonas donde no está disponible Olist PAX y donde el promedio de tiempo de entrega es más alto.

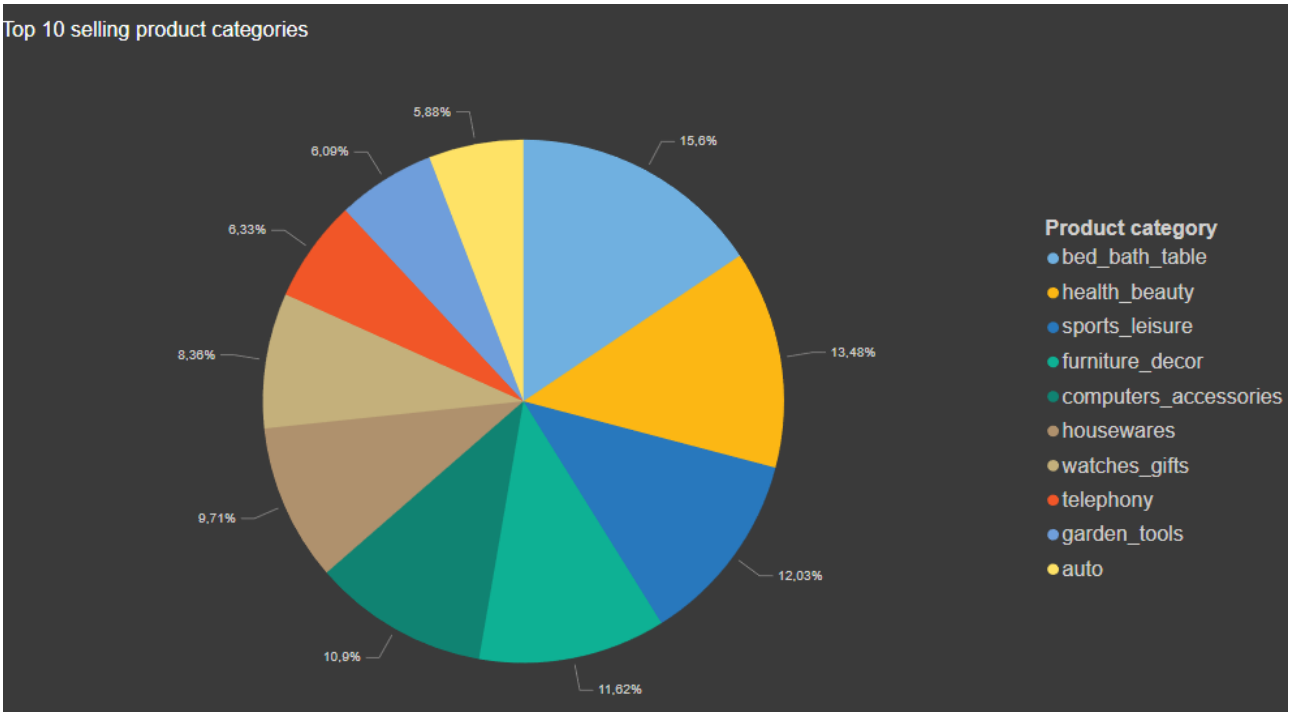


Basándose en el mapa anterior y la información que se tiene sobre el promedio de tiempo de entrega, la recomendación es expandir el servicio primero en los estados de Rondonia y Tocantis ya que estas zonas limitan con la zona de cobertura actual del servicio Olist PAX.

Para un análisis más completo sobre los tiempos de entrega se puede ver el archivo Recomendations de la carpeta dashboard.

Categorías de productos recomendables

En el siguiente gráfico se pueden ver las categorías de productos más vendidas según los datos disponibles



A su vez las siguientes son las categorías con mayor crecimiento de ventas en el año 2021 y las categorías con mayor cantidad de órdenes de venta en 2022

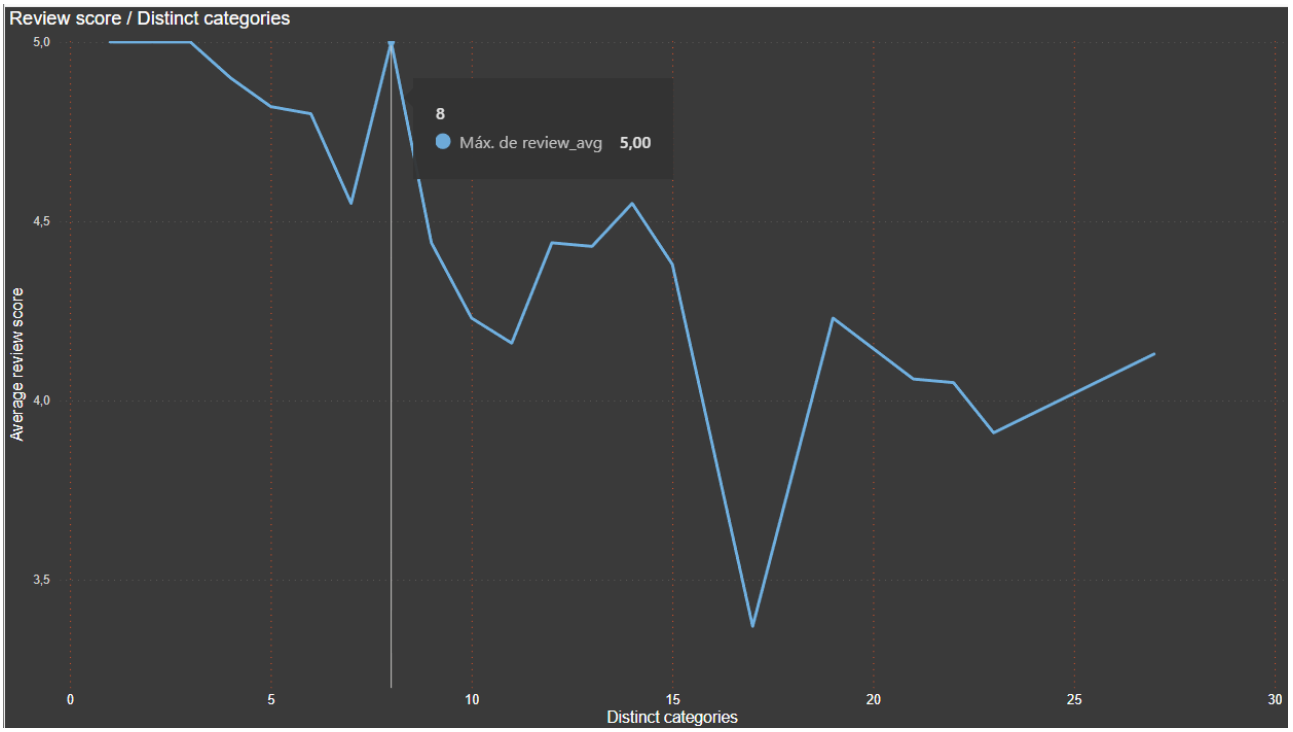
1. Home and decoration
2. Perfumery and cosmetics
3. Fashion and accessories
4. Household appliances
5. Foods and beverages

1. Fashion and accessories
2. Beauty, and Perfumery
3. Health
4. Foods and beverages
5. Household utilities

En base a esto se recomienda darle prioridad desde el sector de marketing a estas categorías más vendidas, principalmente a la sección de cosméticos, accesorios de belleza y salud, y perfumería.

Cantidad de distintas categorías de productos

En el siguiente gráfico se puede ver una comparación entre los review score y las distintas categorías de productos de los vendedores.

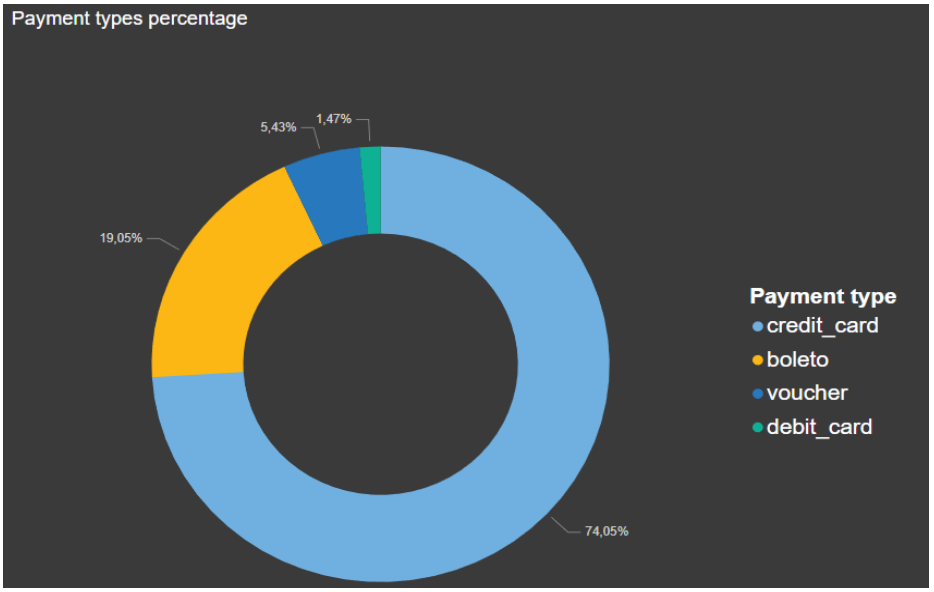


Se observa que los vendedores que se especializan en algunos tipos de productos y que no manejan mucha variedad de categorías de productos tienen un promedio de review score mayor.

Sería útil recomendarle a los vendedores que no se expandan a vender muchos tipos de productos distintos si no cuentan con la logística necesaria para que mantengan un review score más alto y a su vez los clientes estén más satisfechos con sus productos.

Tipos de pagos

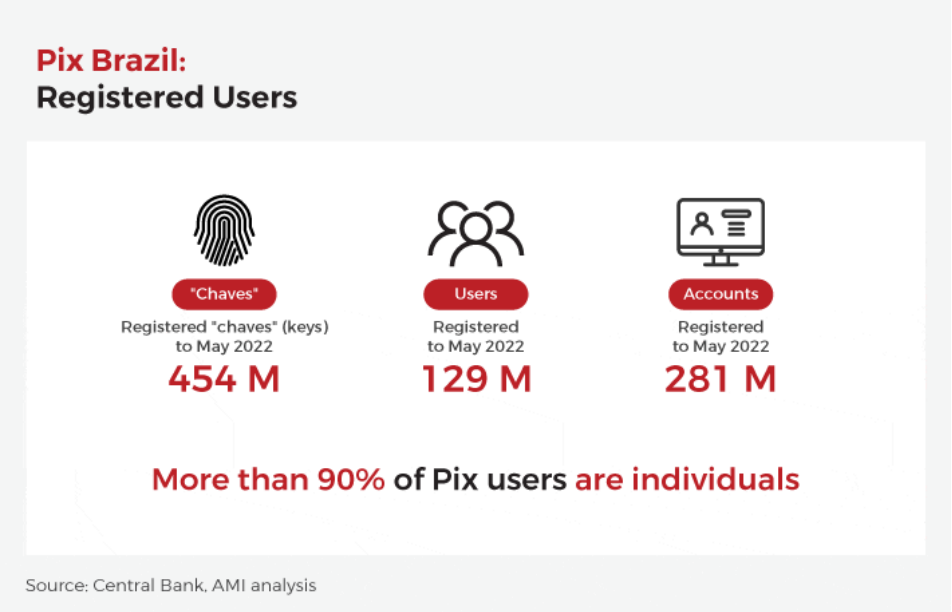
El siguiente es un gráfico del porcentaje de uso de cada tipo de pago



Se puede observar que el uso de tarjeta de crédito supera por amplia diferencia al resto de tipo de pagos, a su vez según otras estadísticas se sabe que la tarjeta de crédito representó el 52% de los pagos en comercio electrónico.

Sin embargo se estima que para 2025 el uso de tarjeta de crédito baje 5% y se estima un alto crecimiento de la plataforma de pago PIX.

Pix es una nueva plataforma de pago instantaneo que en los últimos años tuvo un alto crecimiento y se estima que esto aumente, por lo que la recomendación es dar soporte a este tipo de pago en Olist y a su vez fomentar su uso en los vendedores y clientes para mejorar la calidad del servicio y las ventas.



Fuentes:

- https://americasmi.com/insights/brazil-ecommerce-market-data/
- https://olistpax.com.br/