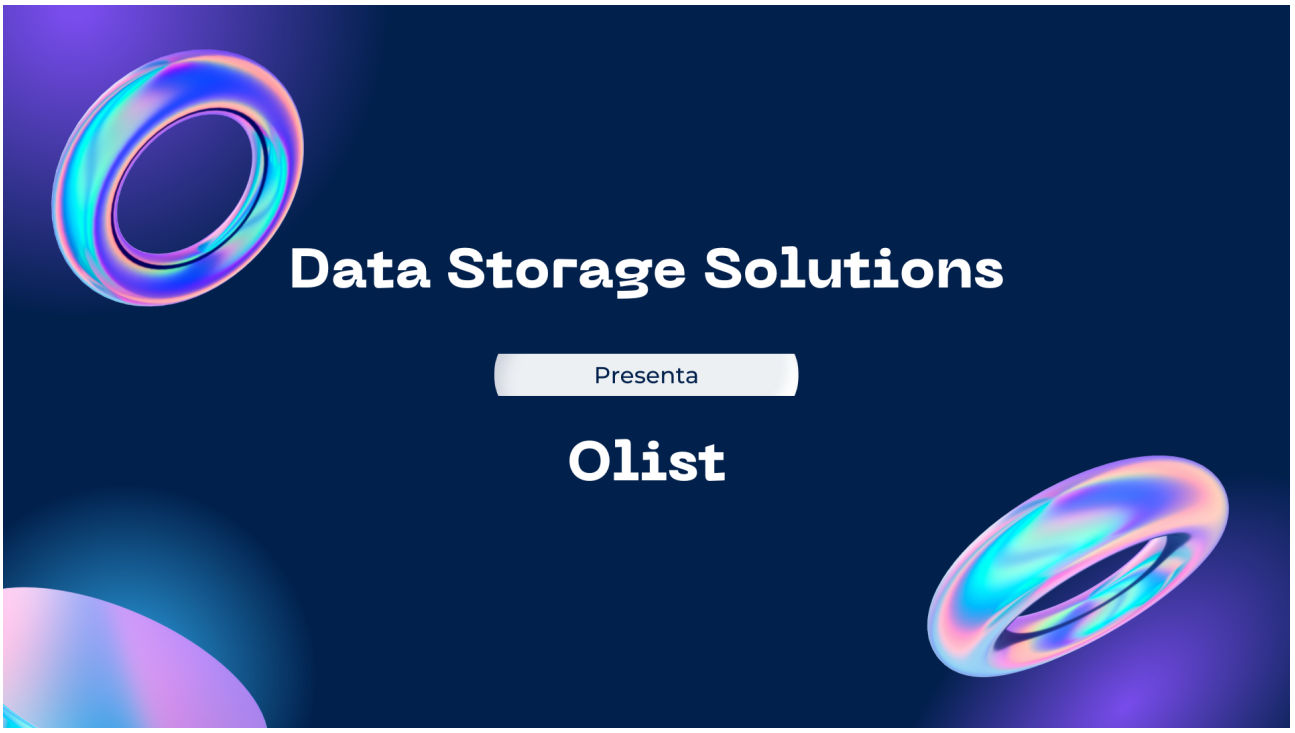


Project Proposal

Company: Olist

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

Index

Introduction

Our company has designed an action plan to connect SMEs with broader markets, improving the user experience. This scheme includes exhaustive research of the e-commerce market in Brazil with the aim of developing strategies that facilitate business intelligence and finding innovative solutions to help users sell their products to a broader customer base. We will use information obtained through an open access source known as "Olist" to carry out these operations.

Team formation: Roles and functions

To maximize theefficiencyof this project, tasks will be distributed among team members based on their individual skills and preferences. Greater flexibility in task assignment will be allowed, allowing team members to work in different areas depending on the complexity of the process. In addition, the progress will be monitored by all team members to ensure the progress of the project.

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

Work methodology

The project will be managed using the Scrum methodology. A daily team meeting will be held withoutlimit of time to discuss possible changes in the project and follow the progress of the established objectives. In addition, at the end of each week there will be a global review of the progress of the project. I also knowsoftware Regular meetings with the client to inform him of the progress of the work and receive his feedback. The team will have a schedule of tasks, but will be open to adaptations based on the needs and ideas that arise throughout the process, both from the team and from the client.

objectives

Main goal:

- Connect small businesses (SMEss) with larger markets and improve the user experience.

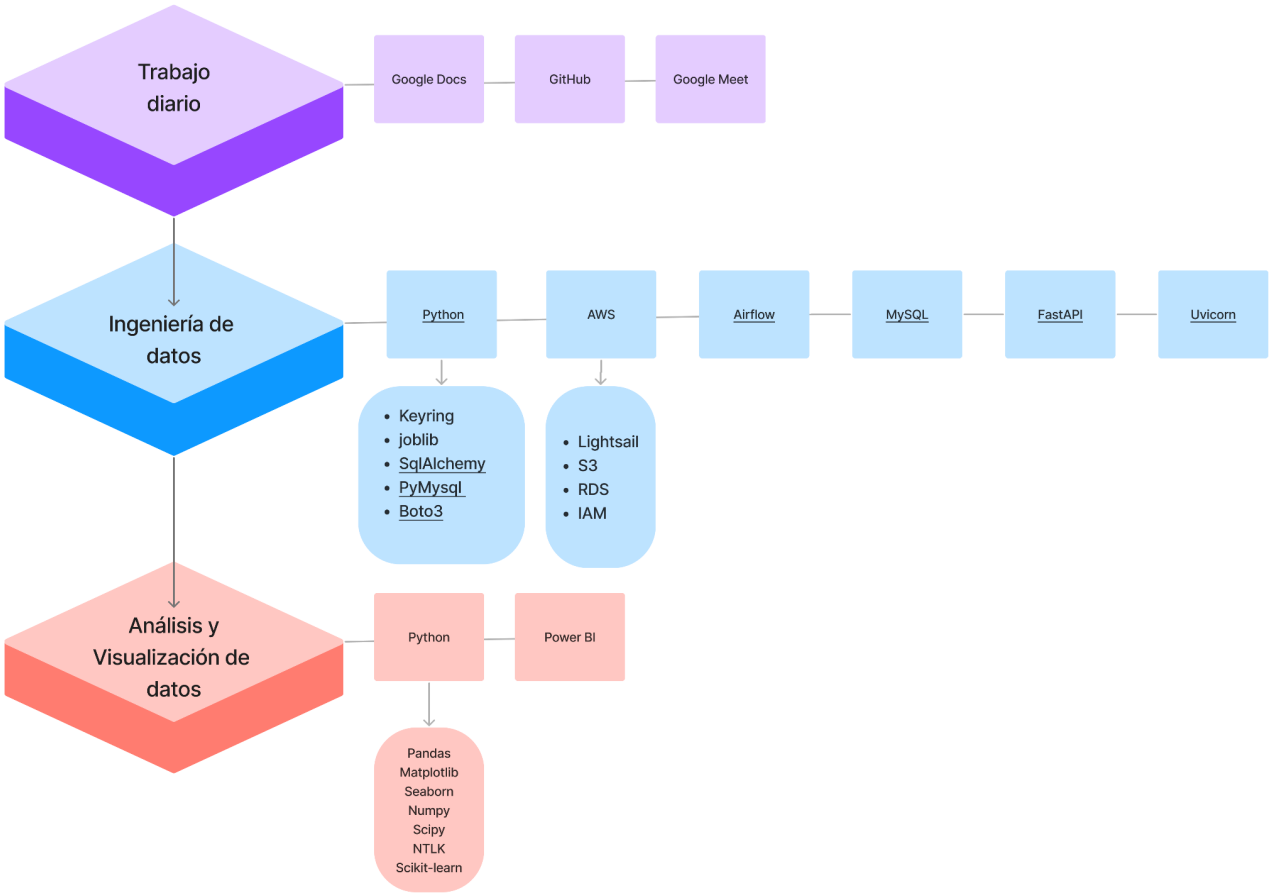
Specific objectives:

- Build an efficient and scalable Data Lake and Data Warehouse.
- Analyze vendors and their performance.
- Extract KPIs that encompass the performance of each salesperson.
- Offer recommendations to the company based on available data.
- Build a report for scalable, online viewing.

Scopes of the project

This project seeks to provide the client with different tools to work with their data and in turn also an analysis of the available data, with which they will seek to offer recommendations to the client for the improvement of their services.

Proposed technological stack:



work schedule

WEEK 1

During the first week of the project we will focus on establishing a detailed plan to carry out the work and on collecting and analyzing the data necessary to carry out the project. It is important to have a good understanding of the data before starting work on project development.

Monday and Tuesday:

- Analysis and understanding of the project
- Definition of objectives and scope

Wednesday and Thursday:

- Analysis exploratory data (EDA)

Friday:

- meeting with theproduct owner y pproject proposal

WEEK 2

Saturday and Sunday:

- Define Infrastructure andchoice cloud service

Monday, Tuesday andWednesday:

- creation from S3 bucket and database in amazon RDS
- Arming airflow in the amazon lightsail virtual machine

Thursday:

- Cloud infrastructure testing and latest changes

Friday:

- Presentation from the infrastructure running to the product owner

WEEK 3

Sonbado toWednesday:

- Building of a preliminary dashboard
- Building a machine learning model

Thursday:

- organizeof the storytelling

Friday:

- Presentation del dashboard preliminar al product owner

WEEK 4

Saturday and Sunday:

- Dashboard completion
- Analysis and search for improvements to offer recommendations
- Armed powerpoint for final presentation

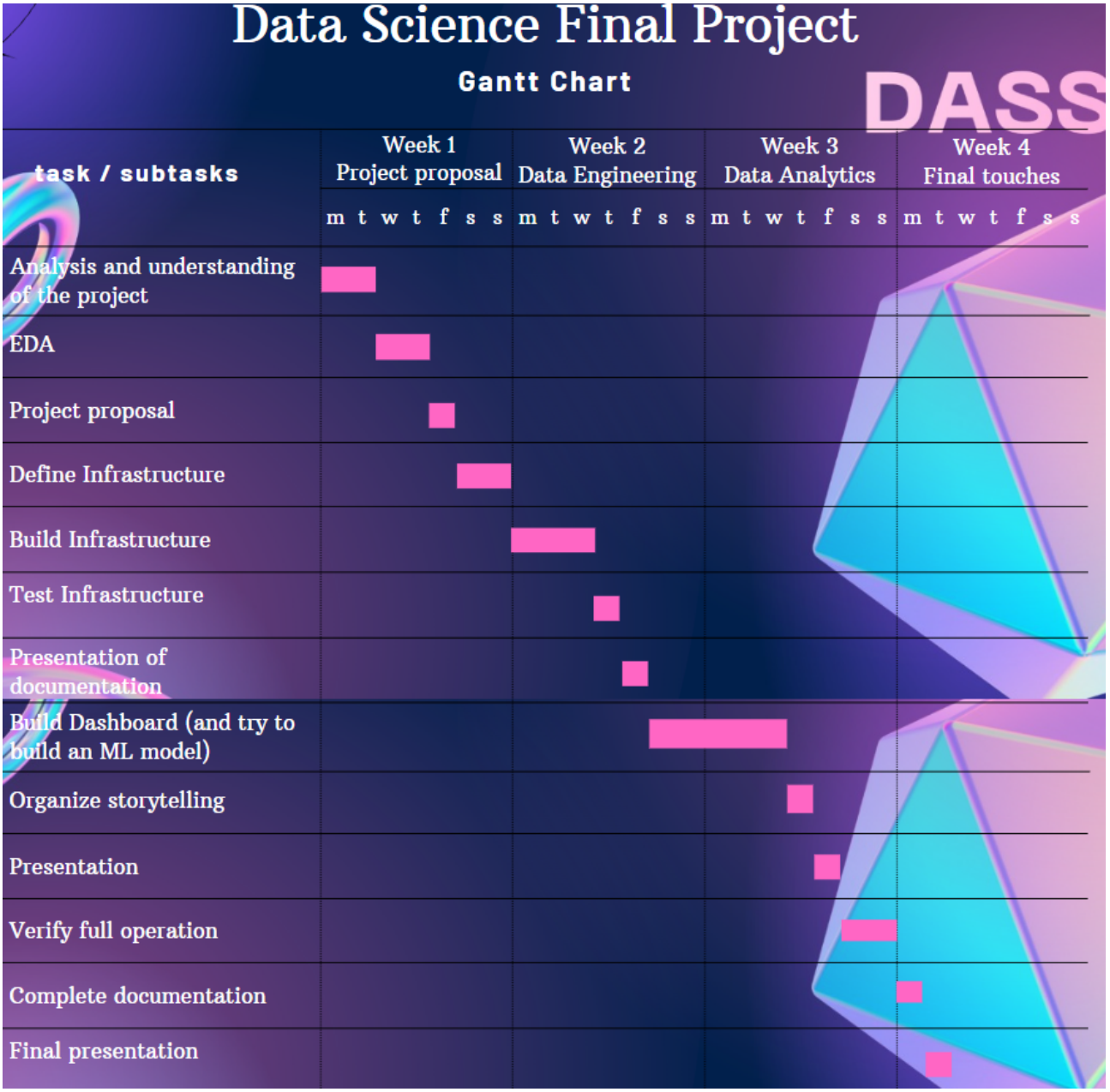
Monday:

- Final meeting to organize storytelling and finalize details

MArt:

- Project presentation

Effort estimation



Deliverables

- All code organized in a repository on GitHub
- Scalable Database
- automated ETL
    - Configured auto-incremental load
    - audit tables
    - New tables created for analysis
- Analysis of sellers and their performance
- business recommendations
- extracted KPIs
- Dashboard the Power BI
- ML model on delivery times and reviews
- API for vendor performance queries

Project development

Data dictionaries

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

Preliminary data quality report

The Olist datasets, which are publicly accessible, were imported. The data consists of 11 .csv files of which 10 were used. These files are located in the folder "Datasets/Datasets_original”

bugs found

The errors found in the exploratory analysis of the data of each dataset are detailed below.

• **olist_customers_dataset**

this tablecontains 5 columns, with 99441 values in each. Missing data fromcustomers_unique_id that there is no reference to how those were createdcodes. This table can be used to find out which cities the consumptions come from. Contains no null data.

• **olist_geolocation_dataset**

This table has 5 columns and 1000163 values in each. This table can be used to locate the cities correctly inanddashboard map. Contains no null data.

• **olist_order_items_dataset**

This table contains 7 columns with 112650 values in each. It can be used to find out which is the most sold product as well as how much it invoiced, which is the seller with the most orders and for how much. Does not contain null values. Presents outliers.

• olist_order_payments_dataset

This table has 5 columns and 103886 values in each. You could find out the behavior of payments. The utility will depend on the PO proposal. Does not contain null values.

• **olist_order_reviews_dataset**

This table has 7 columns and 99224 values in each one. The usefulness of this table will depend on the proposed objective. Contains null values in two columns.

• **olist_orders_dataset**

This table contains 8 columns and 99441 values in each. The usefulness of this table will depend on the objective. Contains few null values in 3 columns.

• **olist_products_dataset**

This table contains 9 columns and 32951 values in each. The usefulness of this table will depend on the objective. It has very few null values in most columns.

• **olist_sellers_dataset**

This table contains 4 columns with 3095 values in each. It does not present null data. It could be known which is the city with the highest number of orders as well as the highest amount in sales.

• **product_category_name_translation**

This table contains 2 columns with 71 values in each. It does not present errors.

• **olist_marketing_qualified_leads_dataset**

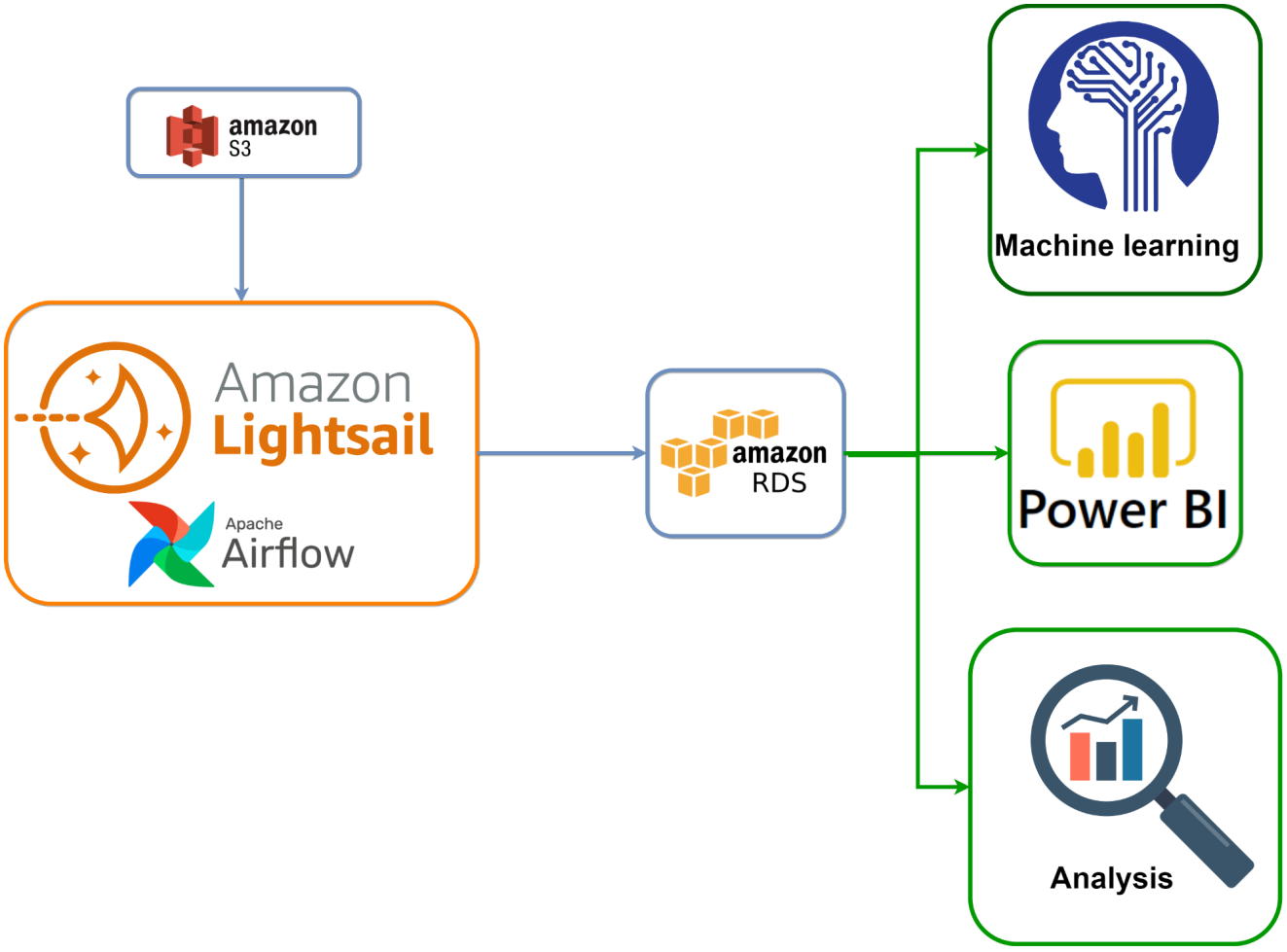
This table contains 4 columns, with a total of 8000 values in each of them. This table does not provide significant data for the analysis we want to perform. Contains few null values in a column.

• **olist_closed_deals_dataset**

This table consists of 14 data columns, with a total of 842 values in each.Lthe columnshas_company, has_got, average_stock and declared_product_catalog_size havemore 90% of your missing data. This table does not provide very relevant data. No duplicate records found. LThe names of the fields are in English.

Extraction, transformation and load ofthe data

Once the preliminary analysis of the data is done, we proceed to the transformation and loading of the data. First we loaded the datasets to work in the cloud using the Amazon S3 service, then we designed a python script that contains the load and transformation functions and then airflow was used in a virtual machine in the cloud, using the Amazon Lightsail service, to Build a pipeline with this python script. We chose to configure the pipeline to run daily from the cloud, which will take the data from the S3 bucket, transform it, and then upload it to the Amazon RDS database.



Data cleaning and normalization

Duplicate rows were removed from each dataset.

All dataset "Closed_deals” The data type of the columns has been changed 'has_got' Y 'has_company' to the float64 data type.

The Orders dataset added a new column called “delivery time” which is obtained by calculating the difference in days of the columns “order_approved_at” y “order_delivered_customer_date”.

The main transformation was carried out by combining the datasets “orders”, ”order_reviews”, ”order_payments”, ”order_items”, “sellers” and “products” into a new table calledcombined_datasets.

A new column calledavg_income_month which contains the calculation of the average monthly income of each seller.

Then a new table called Evaluation is created and the columns “seller_id”, “seller_city”, “product_id”, “product_category_name”, “delivery time”, “review_score”, “order_id”, “payment_value” y “avg_income_month” del dataframe “datasets_combinados”.

The respective calculations are performed and the following column names are changed:

'product_id'→'distinct_prod'

'Delivery time'→'delivery_avg'

'product_category_name'→'distinct_categories'

'review_score'→'review_avg'

'order_id'→'total_orders'

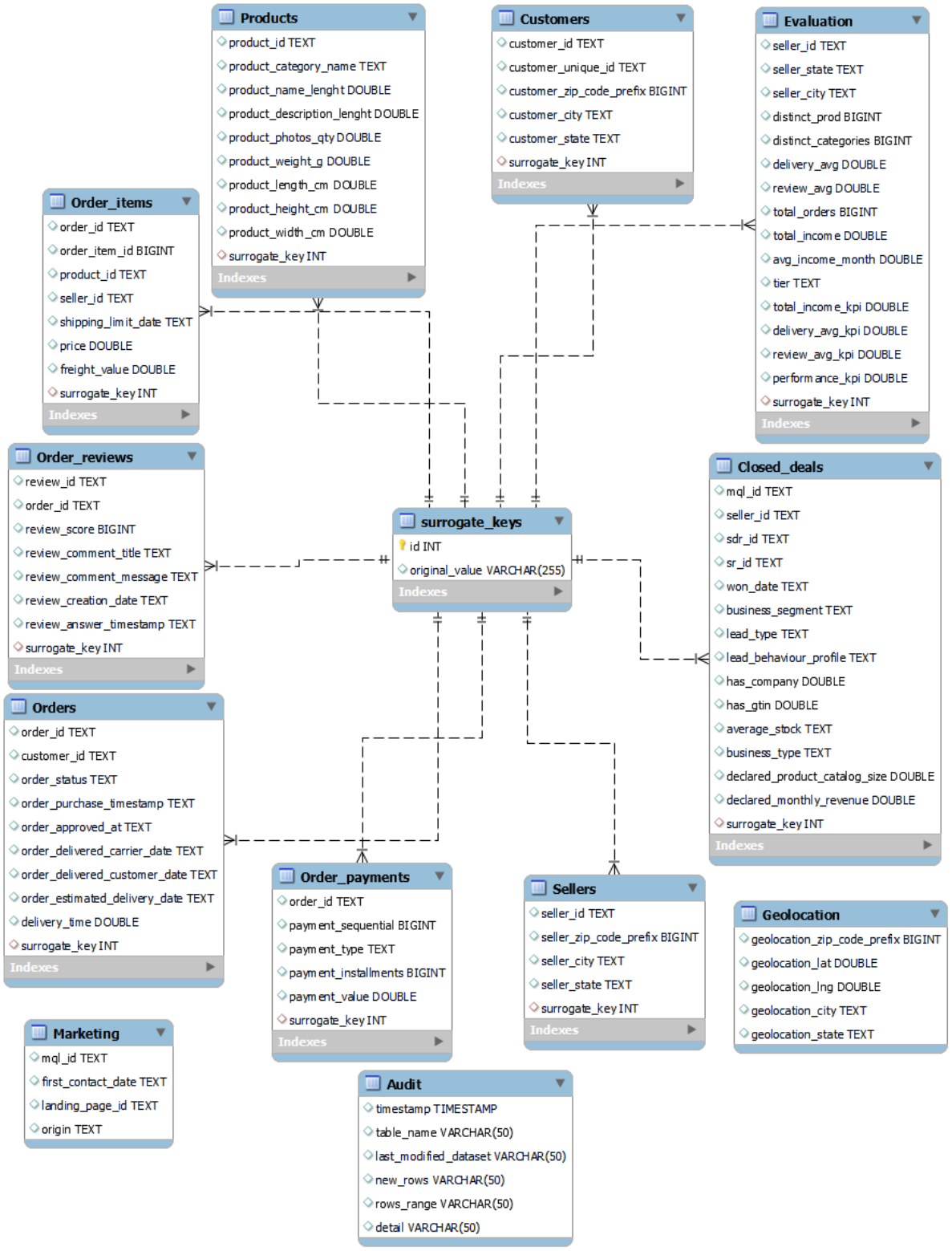
'payment_value'→'total_income'

The rows with values ofdelivery_avg older than 50 days were eliminated as there were few rows and were considered outliers. Then 125 rows containing null values in any column were eliminated.

Finally, functions were created that divide the Evaluation table into different categories and assign different scores to sellers based on different metrics. Added these values in 5 new columns:

- **tier** = This column assigns a category to the seller
- **total_income_kpi=** Calculated based on the column “total_income” that contains the total income of the seller and according to the category of the seller
- **delivery_avg_kpi=** Calculated according to the columndelivery_avg and according to the category assigned to the seller.
- **review_avg_kpi=** It is calculated with the columnreview_avg which contains the average ofreview_score of the seller and based on the category assigned to the seller.
- **performance_score=** It is an average of the three kpi previously calculated and is taken as the final score that is assigned to the salesperson to evaluate their performance

Entity-relationship model



For the relationship entity model, a table of surrogate keys was created, which assigns to the IDs of the different tables, which have alphanumeric values, a new integer as ID in a new column for each table.blah callsubrrogate_key.

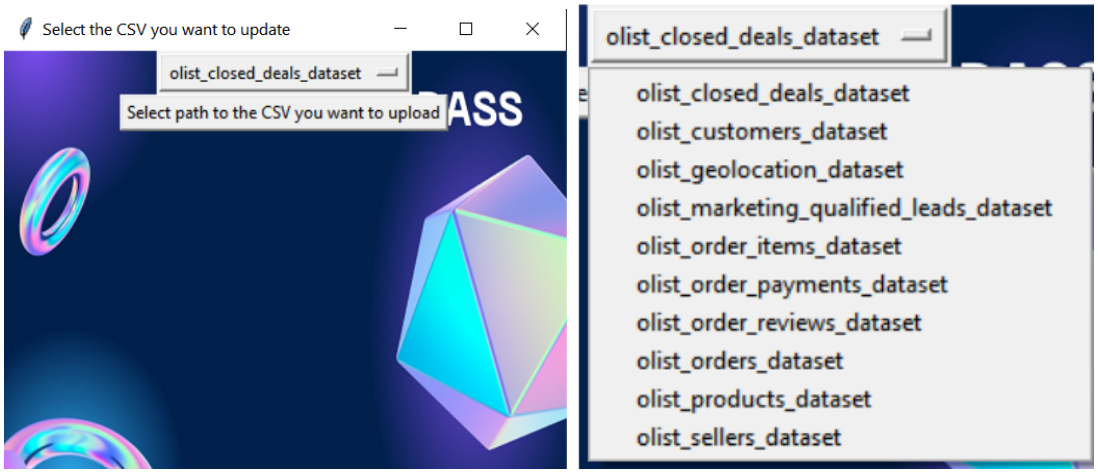
This allows us to connect all the tables and in turn keep the original IDs, which are encrypted, so that the client can access that data again if needed by decrypting the ID values.

At the same time, we designed the automatic loading system to be compatible with an incremental loading of data. In this way, when updating the datasets in the S3 bucket, the airflow pipeline will automatically upload only the new rows to the database, thus avoiding uploading duplicate rows. The pipeline in turn will check the last modification date of thedatasets uploaded in the cloud, so that it will only operate with those that present modifications from the last modification date for which there are records, for each specific dataset.

Finally, the airflow pipeline will load information into an audit table, which contains the following columns:

- **Timestamp=** date and time the operation was performed on the database.
- **table_name=** name of the table on which the operation is performed.
- **last_modified_dataset=** last modified date of the .csv file found in the S3 bucket, of the table being operated on.
- **new_rows=** number of new rows added to the table
- **rows_range=** index of the range of new rows added to the table.

We also designed a script to make it easy to update the datasets that are in the S3 bucket.



This script first asks the user to choose the dataset they want to update and then select the location of the .csv file to upload to the cloud. The script will throw an error message in case the columns of the chosen .csv file do not match the columns of the .csv file uploaded to the cloud, in order to avoidoverwrite wrong files.

Machine learning

For the machine learning model, the original datasets were loaded first “olist_order_items_dataset.csv” , “olist_order_reviews_dataset.csv” y “olist_orders_dataset.csv”

Then the delivery time column was added to the Orders dataframe and the 3 dataframes were combined into one called DF.

A new column was added to this combined dataset where a categorical variable is calculated based on the review_score column. This variable assigns a value of 0 to review scores less than or equal to 3 and 1 to review scores greater than 3. removed duplicate rows from the dataframe and rows with null values.

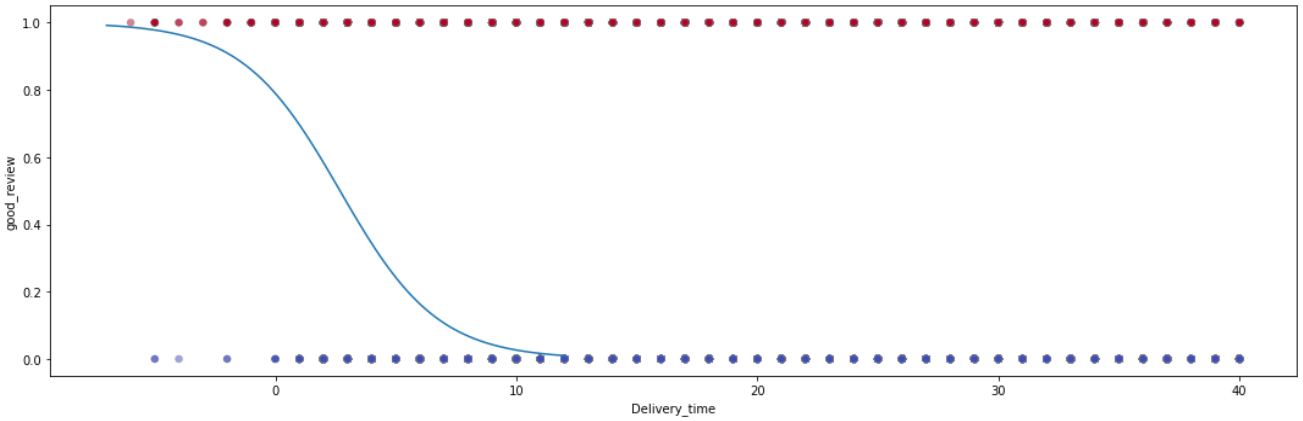
Outliers were removed using the scipy library and then scikit learn was used to do a split test of the data, where 25% of the data was taken for testing and the rest for training.

4 pipelines were set up to test different models and see which oneoffered best results. Each pipeline applies a standard scaler to the data and the 4 models tested along with their accuracy were:

- 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

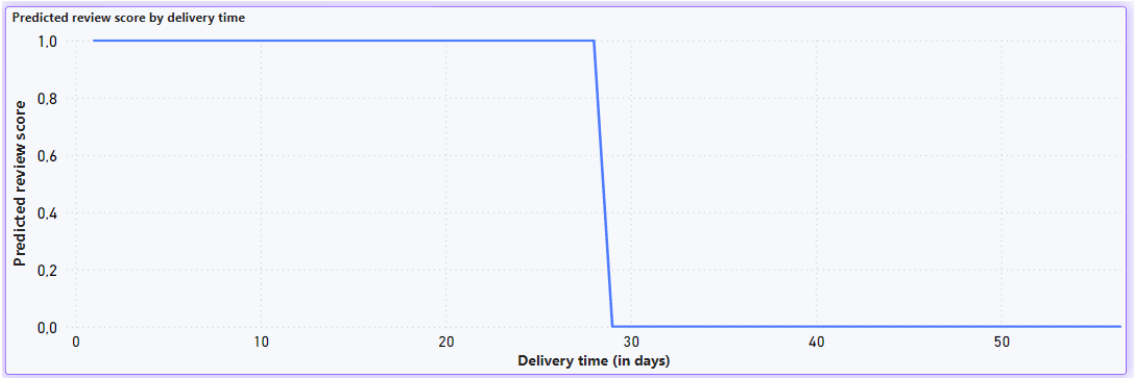
Therefore, the model chosen was the classification tree which has an accuracy of 0.79, a recall of 0.98 and an F1-score of 0.88.

In the following graph it can be seen that most of the data falls within the classification of the model, so it has enough precision to use it.



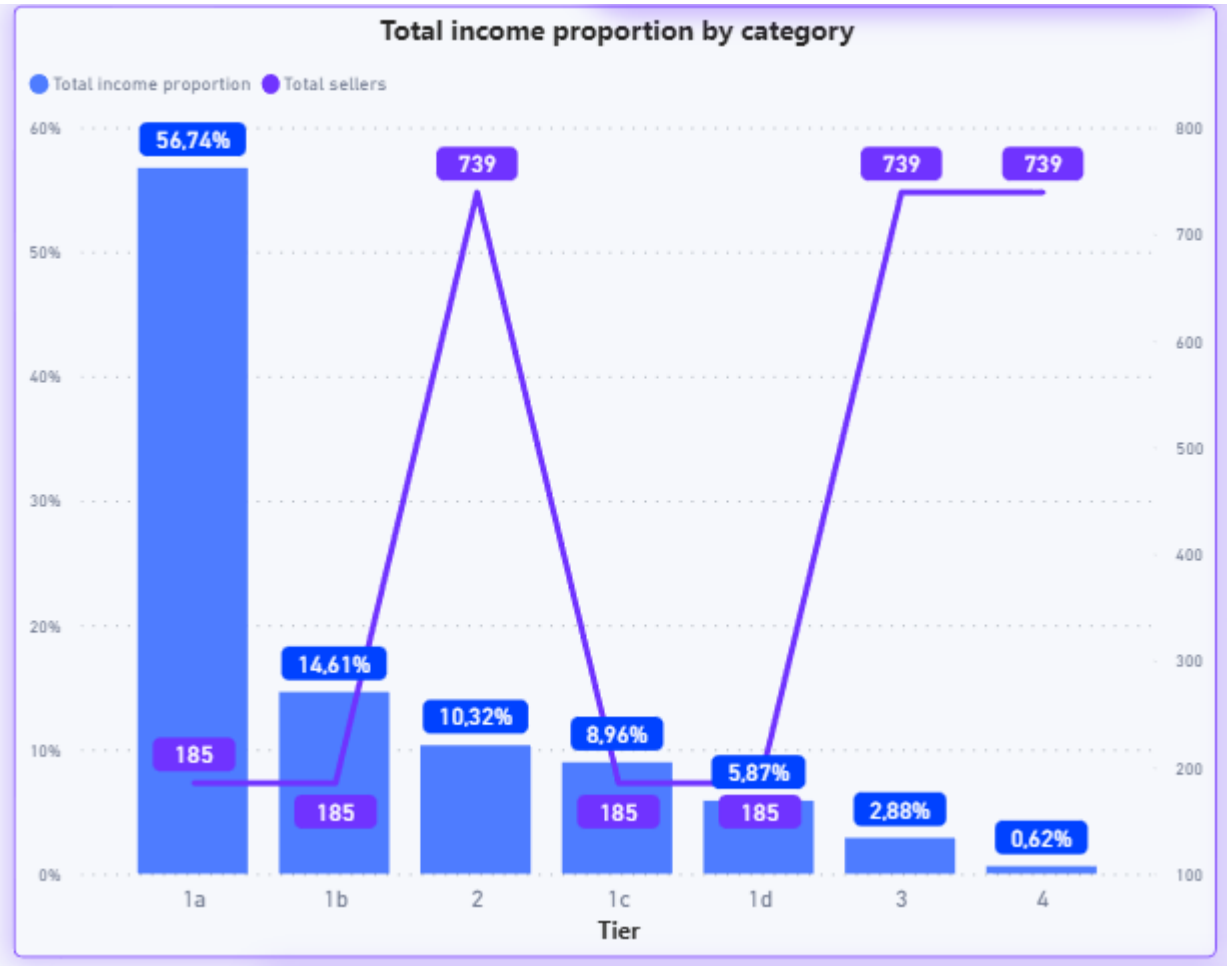
Finally, a prediction is made on the target column using the trained model, to predict the review score based on the average delivery time, which was used as a feature for the model.

With the prediction made, the information was obtained that after 28 days it is much more likely to obtain a score of 3 or less, so now we have a clearer limit on the change of review score according to delivery time, and this information can be given to sellers who aim to improve their review score.



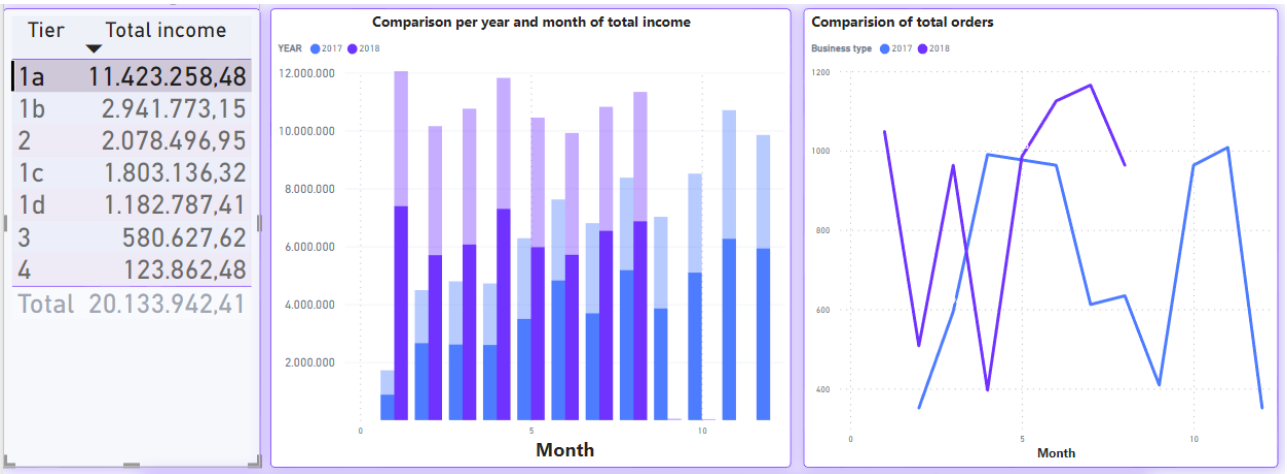
Analysis

In the Evaluation table, sellers are categorized based on their revenue. In the following graph, it can be seen that in category 1-A, which is assigned to the sellers with the highest income, 56.74% of the total income of the platform is found in the year 2017-2018 and despite the fact that only 6.25% of sellers are found.



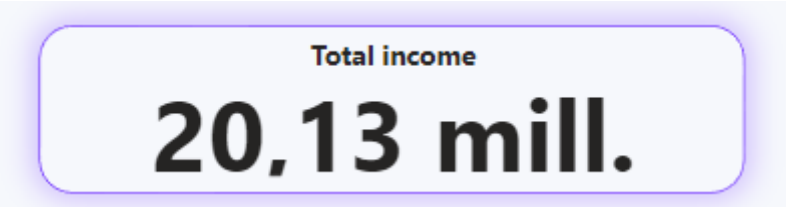
All these data are calculated based on a total of 117 thousand orders, most of which are provided by category 1-A.

Next, the amount of total income contributed by each category can be seen with greater precision in the table on the left. The central graph shows the total income of category 1-A compared to the total income of all categories and The graph on the right shows a comparison of category 1-A orders between the years 2017 and 2018.

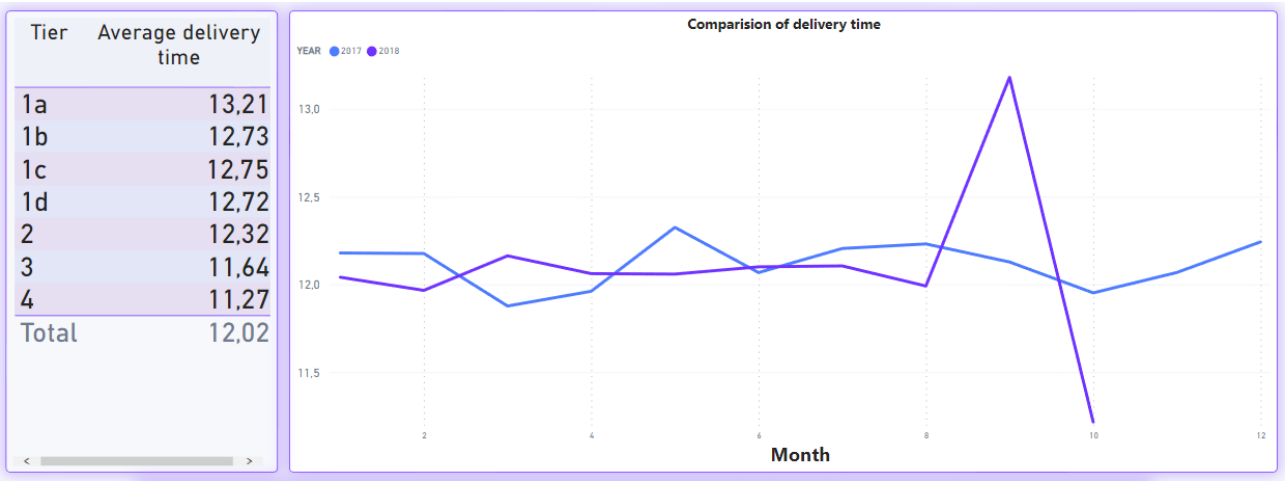


This tells us that the company should pay attention to sellers in this category to see what makes them stand out from the rest in terms of revenue and number of orders.

Finally, to finish with the analysis of total income, it can be seen that in the years studied the total income was 20,130,000 Reales.



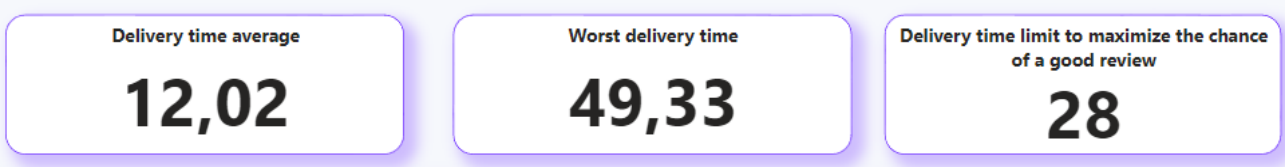
The next thing that was done was to analyze the evolution of delivery times in the years 2017 and 2018 (in days), in the table on the left you can see the comparison of the average delivery times by category, it is seen that category 1 -A has a higher average delivery time, which is related to the greater number of orders they handle.



Tier	Average delivery time
1a	13,21
1b	12,73
1c	12,75
1d	12,72
2	12,32
3	11,64
4	11,27
Total	12,02

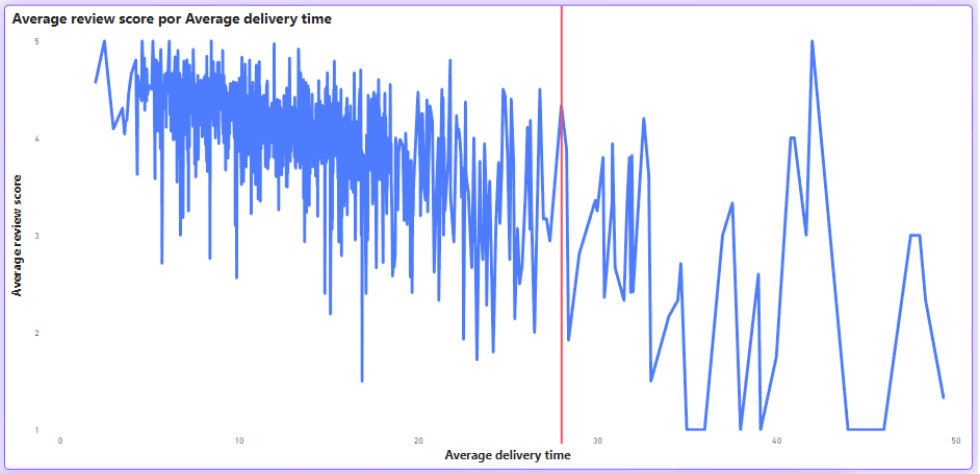
The average delivery time for all sellers was 12.02 days, the worst delivery time on record, having eliminated outliers, is 49.33 days.

28 days is the recommended delivery time limit according to the information obtained with the machine learning model, to reduce the probability that the customer will leave a review score less than or equal to 3.

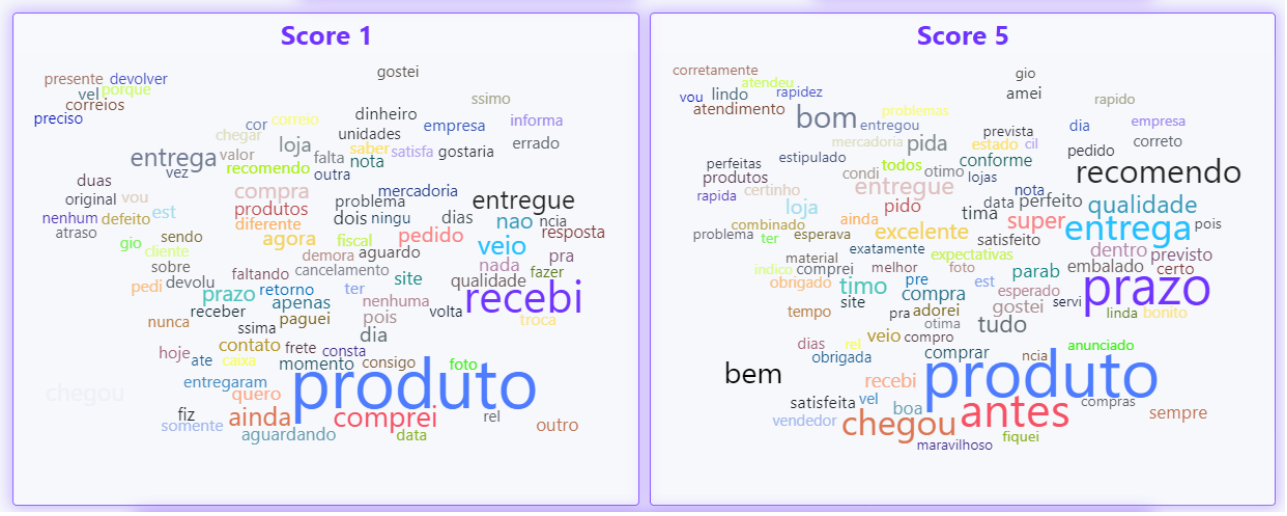


In the following graph you can see more clearly the 28-day limit (red vertical line) and how it affects the average score and in the table on the left we have an analysis of the average review score by category.

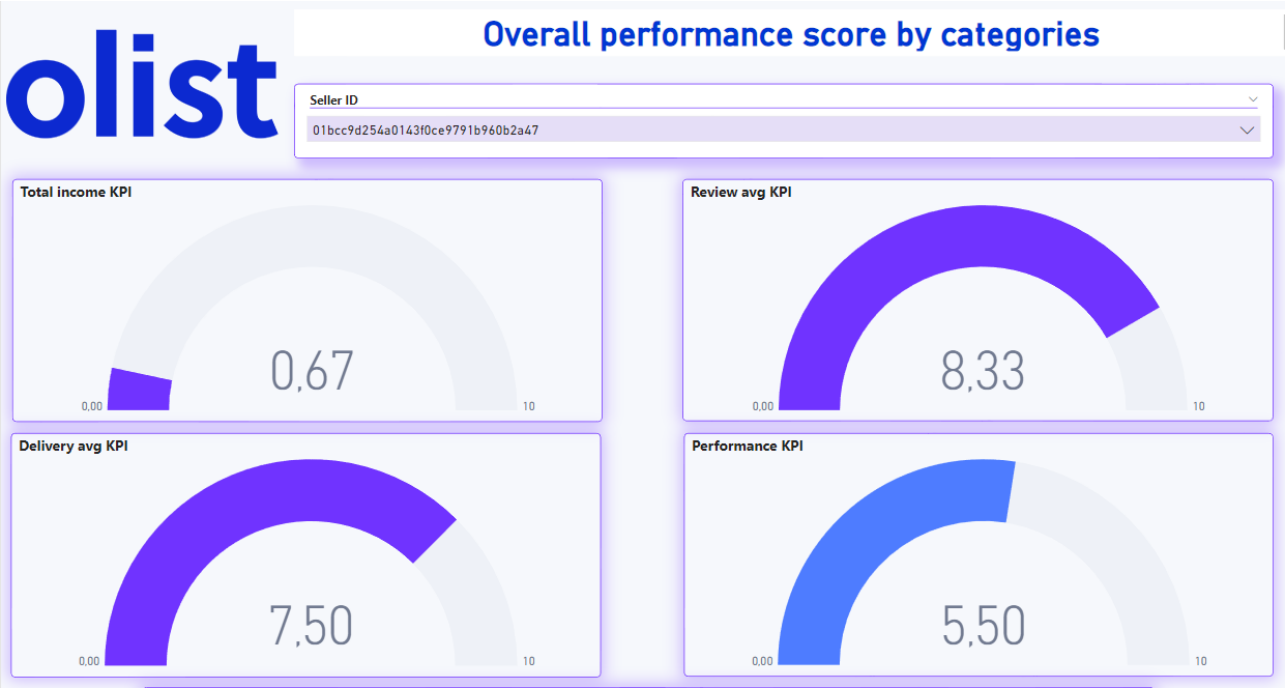
Tier	Average review score
1a	3,98
1b	4,00
1c	4,07
1d	4,02
2	4,04
3	4,10
4	4,18
Total	4,08



The next thing that was done was an analysis of the comments in the reviews and two word clouds were made with the most frequent keywords for score 1 and score 5. It can be seen that one of the things that most influences the score It is once again the delivery time, and that in reviews with a score of 5, customers recommend the seller or product, which translates into more sales.



Finally, a Seller ID filter was added where you can check the 4 KPIs of each seller regardless of their category.



conclusions

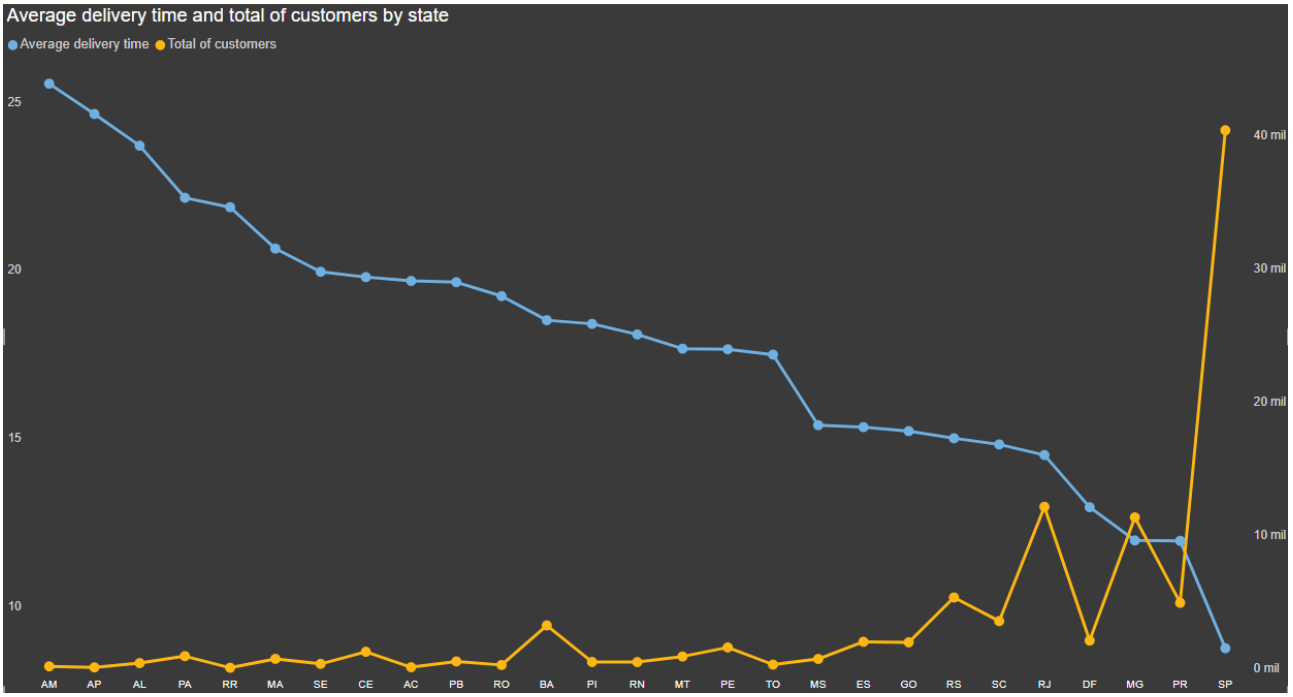
From the analysis of the information, we can conclude the following:

- Less than 6% of customers contribute more than 55% of total revenue, and also contribute more than 50% of orders.
- If we extend the number of customers up to 25%, the total income exceeds 85%, as do the orders.
- Sales do not increase during Carnival and Christmas times, which indicates a lack of promotion of the festivities.
- Deliveries with long delays are, in 80% of cases, the cause of bad evaluations.
- The city and state with the highest number of orders is São Paulo
- There are deliveries with delays well above the average, reaching extremes of more than 100 days.

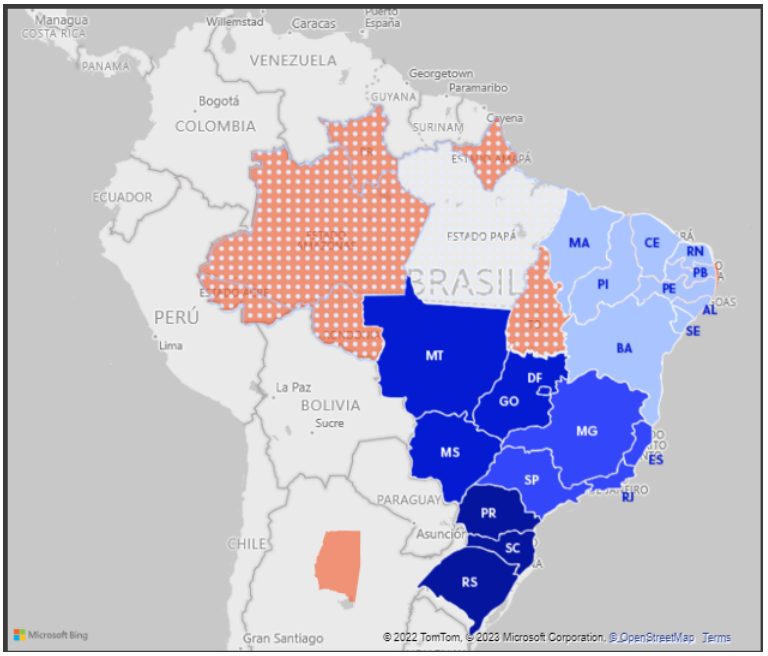
recommendations

Reduction of delivery times

When analyzing the number of customers and the relationship with delivery times in the different states of Brazil, it can be seen that by improving delivery times there are more customers, so reducing delivery times would not only improve the service for customers but will also increase the number of these.



In recent years, Olist acquired the logistics company PAX to improve the quality of the product delivery service. In the following map you can see in blue the area where the service currently works and in orange the areas where Olist PAX is not available and where the average delivery time is higher.

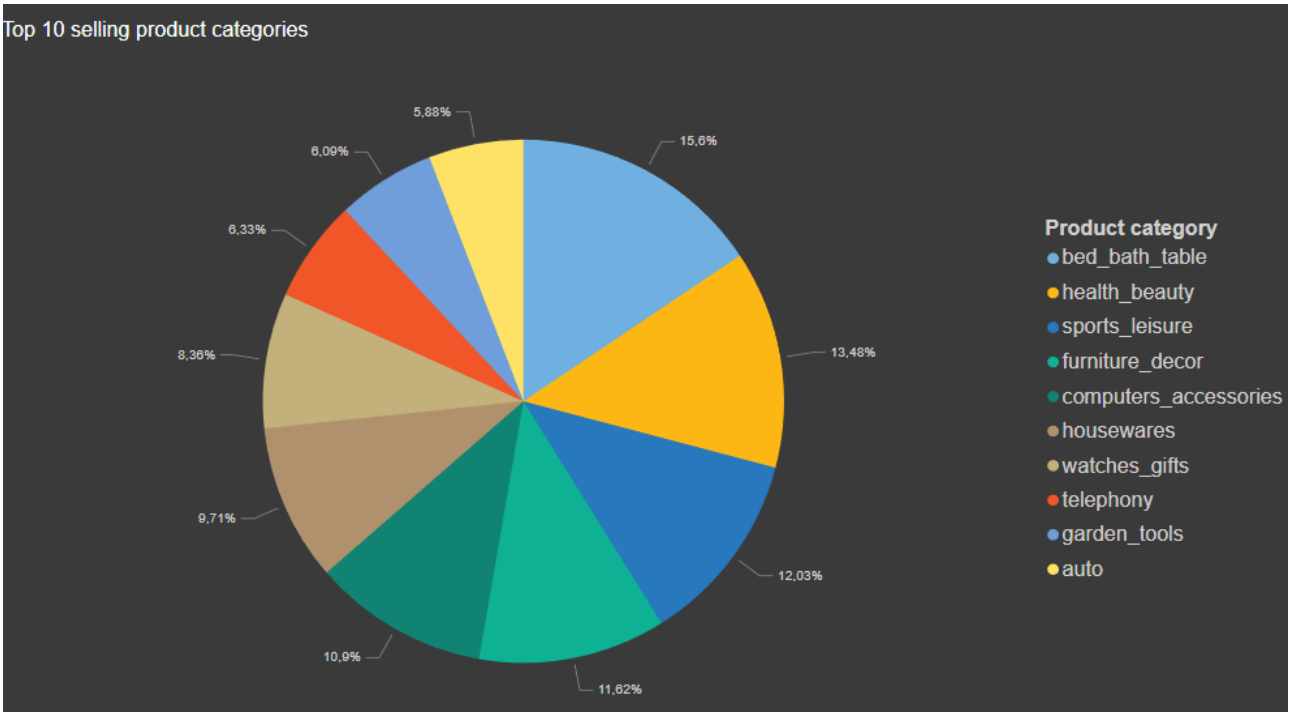


Based on the previous map and the information available on the average delivery time, the recommendation is to expand the service first in the states of Rondonia and Tocantis, since these areas border the current coverage area of the Olist PAX service.

For a more complete analysis of delivery times, you can see the Recommendations file in the dashboard folder.

Recommended product categories

In the following graph you can see the best-selling product categories according to the available data



In turn, the following are the categories with the highest sales growth in 2021 and the categories with the highest number of sales orders in 2022

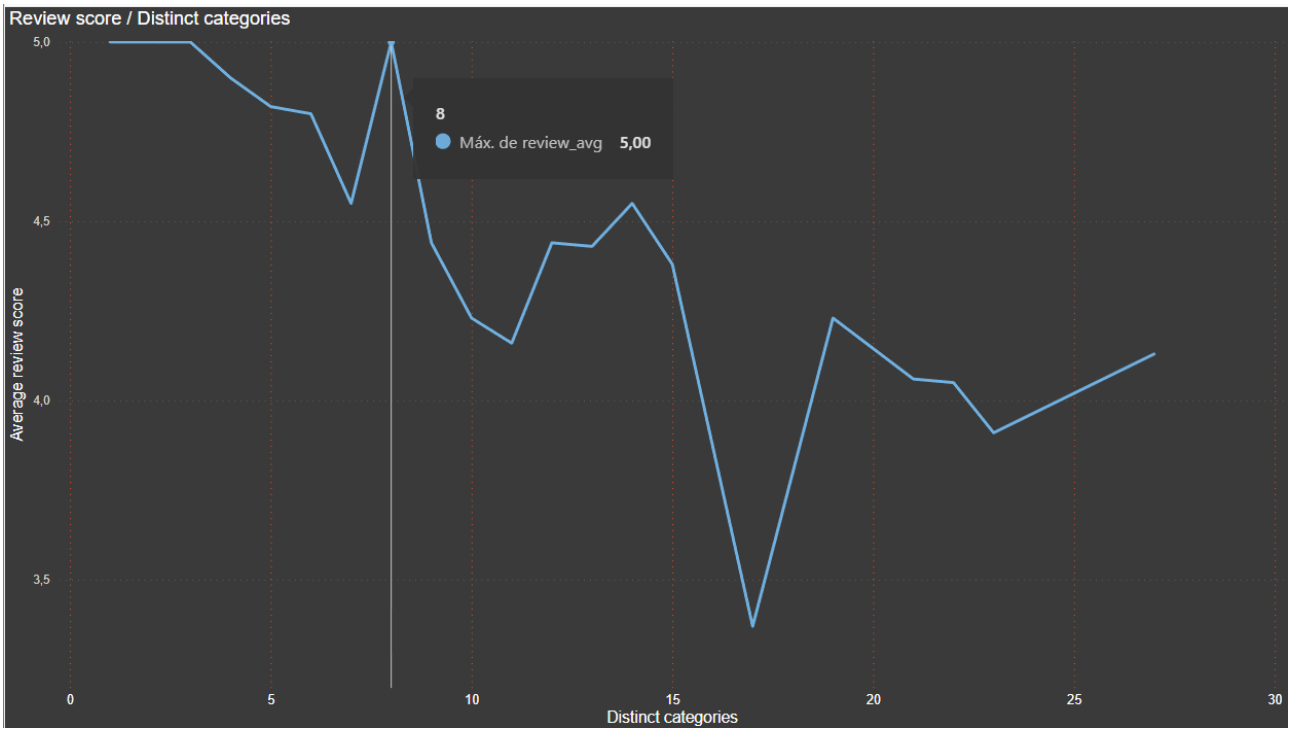
1. Home and decoration
2. Perfumery and cosmetics
3. Fashion and accessories
4. Household appliances
5. Foods and beverages

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

Based on this, it is recommended that the marketing sector give priority to these best-selling categories, mainly the cosmetics section, beauty and health accessories, and perfumery.

Number of different product categories

In the following graph you can see a comparison between the review scores and the different product categories of the sellers.

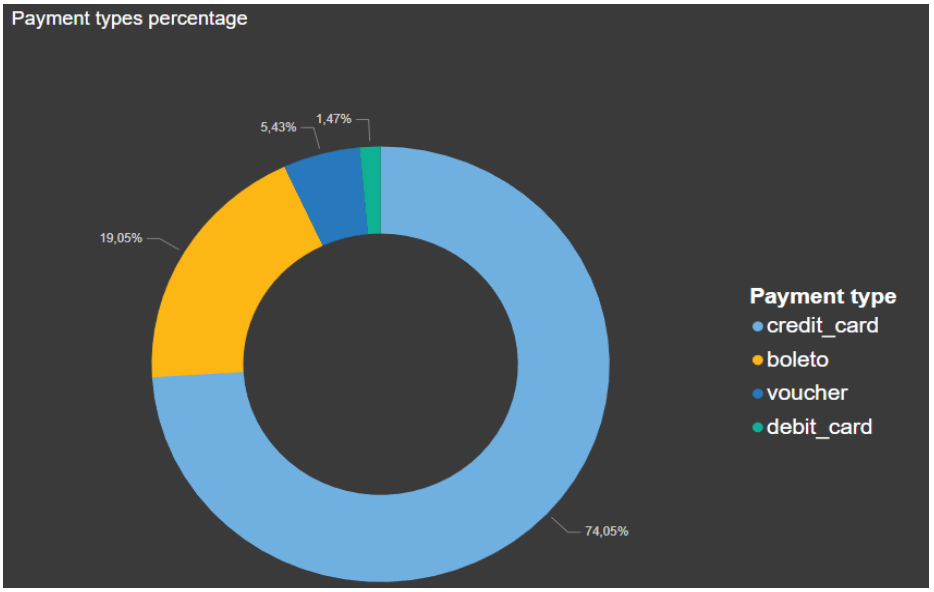


It is observed that sellers who specialize in some types of products and who do not handle a wide variety of product categories have a higher average review score.

It would be useful to advise sellers not to expand into selling many different types of products if they do not have the necessary logistics to maintain a higher review score and in turn make customers more satisfied with their products.

types of payments

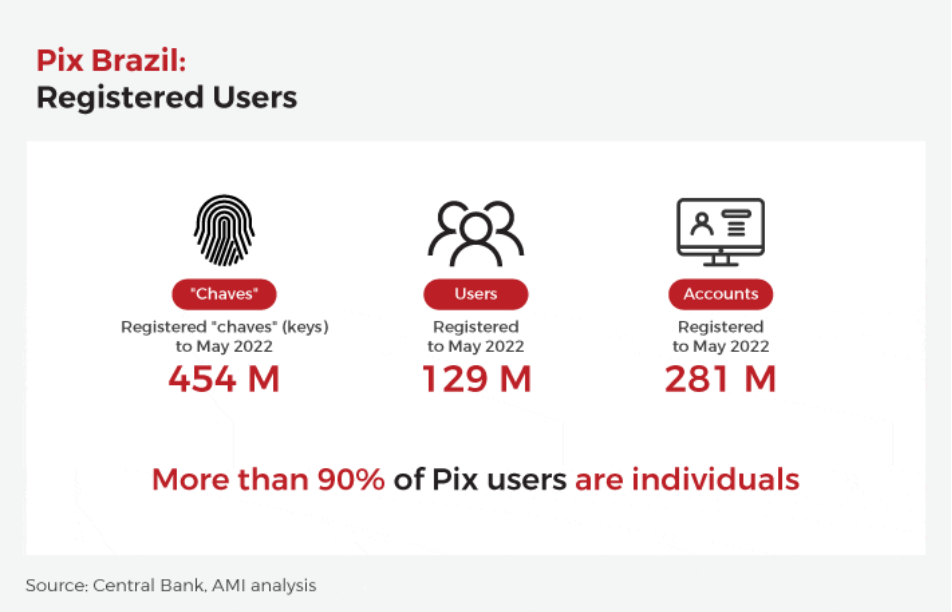
The following is a graph of the percentage of use of each type of payment



It can be seen that the use of credit cards far exceeds the rest of the types of payments, in turn, according to other statistics, it is known that credit cards represented 52% of payments in electronic commerce.

However, it is estimated that by 2025 the use of credit cards will drop 5% and a high growth of the PIX payment platform is estimated.

Pix is a new instant payment platform that has experienced high growth in recent years and it is estimated that this will increase, so the recommendation is to support this type of payment on Olist and in turn encourage its use by vendors and customers. to improve the quality of service and sales.



Sources:

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