OLIST
ECOMMERCE

FINAL MARKET PLACE REPORT

REPORT

AGENDA

1. Objective of the Work
2. Contextualization of the Problem
3. Database
    i. Original bases
    ii. Base transformations
4. Exploratory Data Analysis
5. Modeling with Traditional Statistics
6. Modeling with Artificial Intelligence
7. Conclusions

DATA ANALYSIS METHODOLOGY

Planning → **Exploratory Analysis** → **Statistical Modelling** → **ML Algorithms** → **Conclusion**

Problem definition

- Goals
- Concepts
- Criteria
- Data history
- Variables

Primary Analysis

- Position measurements
- Frequency analysis Graphics
- Outlier analysis
- Missing analysis
- Validation on the
- Consistency of information

Evaluation of techniques

- Native K-means using Spark

Evaluation of techniques

• biSecting K-means Spark native

• Gaussian Mixture Native model in Spark

• Scikit Learn: DBSCAN, MeanShift, K- means Clustering agorithms

Key Actionable Insights & takeaways

• Definition of the technique

• Validation of results

• Choice of technique what better if suitable for use and strategies

DATA ANALYSIS METHODOLOGY

Problem definition

- Goals
- Concepts
- Criteria
- Data history
- Variables

Primary Analysis

- Position measurements
- Frequency analysis Graphics
- Outlier analysis
- Missing analysis
- Validation on the
- Consistency of information

Evaluation of techniques

- Native K-means using Spark

Evaluation of techniques

- biSecting K-means Spark native
- Gaussian Mixture Native model in Spark
- Scikit Learn: DBSCAN, MeanShift, K- means Clustering agorithms

Key Actionable Insights & takeaways

- Definition of the technique
- Validation of results
- Choice of technique what better if suitable for use and strategies

1. PURPOSE OF THE WORK

Olist Customer Segmentation for definition of commercial policy and actions marketing

Main objective: Find the best way to grouping of Olist customers (sellers of marketplaces) to *establish commercial policy* and *differentiated marketing for each group*, aiming to increase the revenue and improve the company's results.

2. CONTEXTUALIZATION OF THE PROBLEM

Olist is a company that helps sellers advertise their products on the main internet marketplaces (Amazon, Mercado Livre, Americanas, Carrefour, Submarino, Via Varejo, Casas Bahia, B2W Digital, Extra, Shoptime, Ponto Frio, Madeira Madeira and Zoom), providing centralized way tools for registration and management of products and stocks; sales, finance and strategy management; logistics management of deliveries; and improving positions on highly reputable sales sites from Olist.

Currently, the company has 3 commercial plans:

- Olist lite: revenue of up to R$3,000.00, no monthly fee, commission 21% per order.
- Olist pro: revenue from R$3,000.00 to R$20,000.00, monthly fee R$79.90, 19% commission per order.
- Olist Premium: revenue above R$20,000.00, conditions specific to each client.

2. CONTEXTUALIZATION OF THE PROBLEM

The company would like to know whether this division of customers into billing is the most appropriate to establish your commercial policy,
In addition to obtaining insights for personalized marketing actions according to
with customer profiles, to increase your revenue and improve your result.

3. DATABASE

Brazilian E-Commerce Public Dataset by Olist

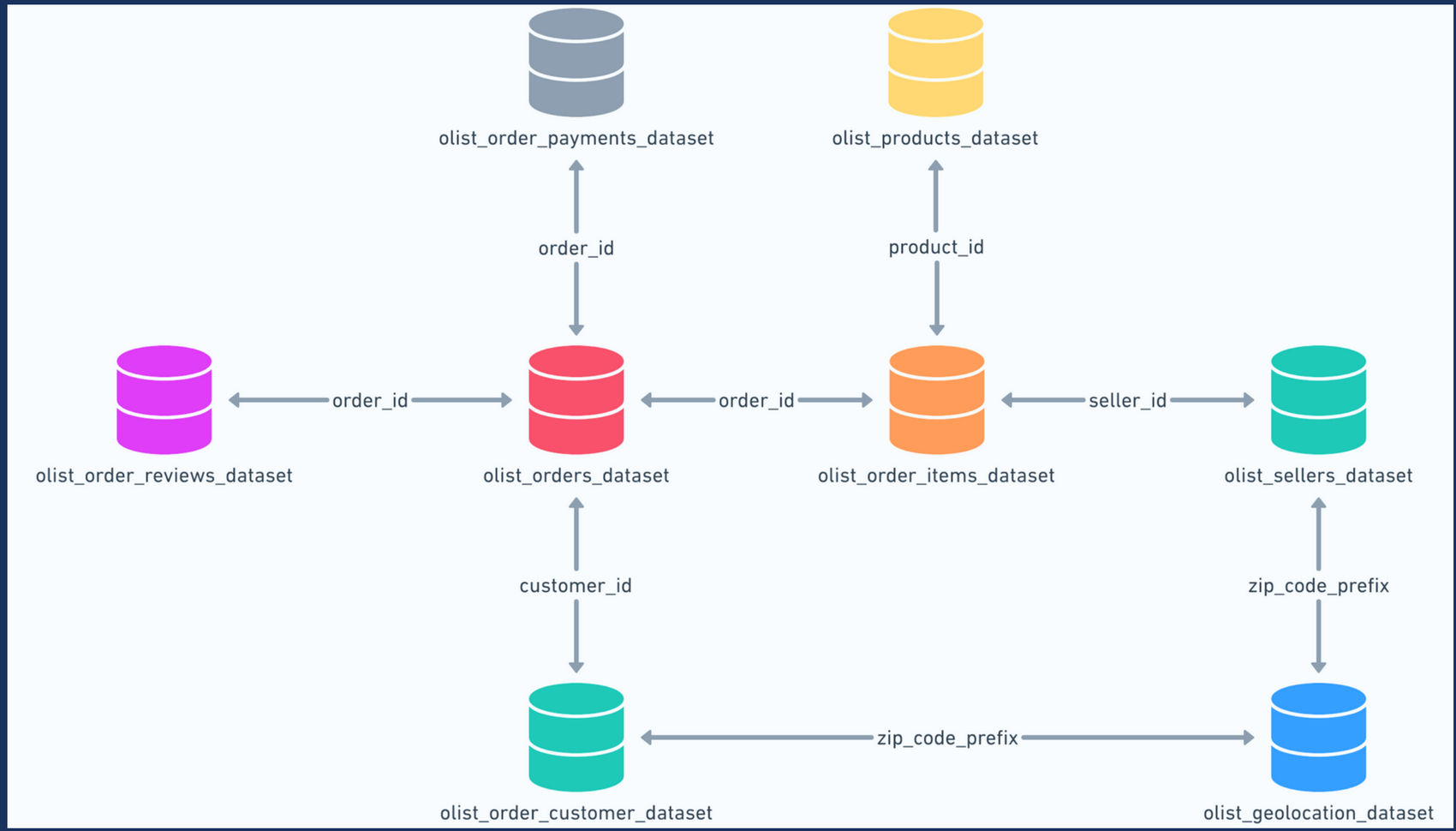
100,000 Orders with product, customer and reviews info

k kaggle.com

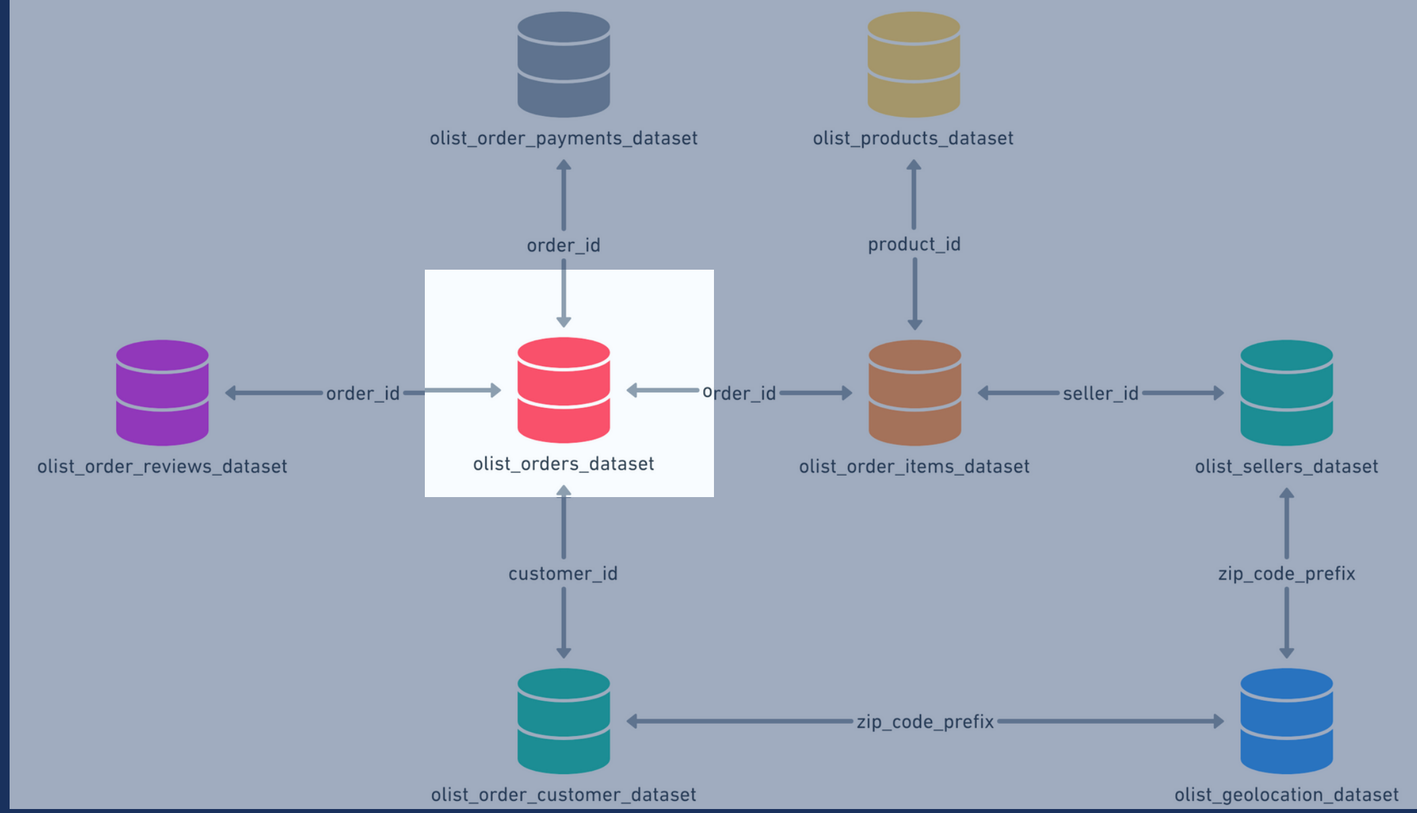
Olist has a public dataset available on kaggle.
Brazilian E Commerce Public Dataset by Olist dataset consists of customer transactions in the most diverse Brazilian marketplaces.
The main base is made up of a historical sales series with 99,441 orders issued between September 2016 and October 2018.

In addition to this, there are auxiliary bases with details about the products sold and
the products of each order, about the sellers, about the buyers and reviews from buyers, payments, and customer service data geolocation. The last two will not be used in our study.

3.1 DATABASE SCHEMA (SALES)



3.2. ORIGINAL BASES



Description of the **olist_orders_dataset:**

Number of records

- **99,441**

Period analyzed:

- **09/04/2016 to 10/16/2018**

Number of variables

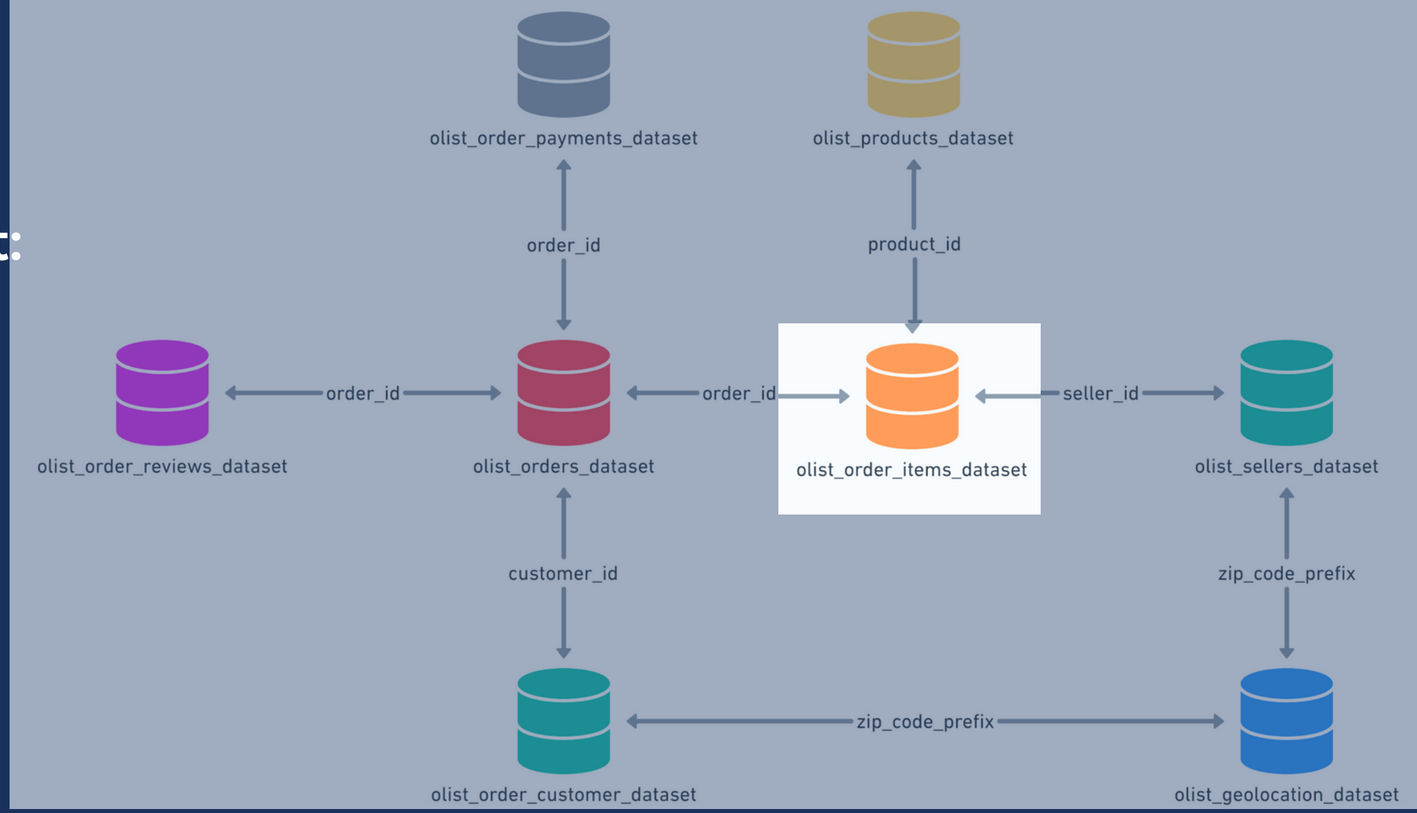
- **8**

Variable names:

1. **order_id:** Unique purchase ID
2. **customer_id:** Key to the customer dataset, each purchase has a unique customer_id
3. **order_status:** Purchase status
4. **order_purchase_timestamp:** Time of purchase
5. **order_approved_at:** Purchase approval time
6. **order_delivered_carrier_date:** Purchase posting time
7. **order_delivered_customer_date:** Purchase delivery time
8. o**rder_estimated_delivery_date:** Estimated delivery time informed to the buyer at the time of purchase

3.2. ORIGINAL BASES

Description of the **olist_order_items_dataset:**



Number of records

- **112,650**

Number of variables

- **7**

Variable names:

1. **order_id:** Unique ID of the purchase
2. **order_item_id:** Sequential number that identifies items included in the same purchase
3. **product_id:** Unique product ID
4. **seller_id:** Seller's unique ID
5. **shipping_limit_date: Deadline for delivery of the product to the delivery service**
6. **price:** Price of the item upon purchase
7. **freight_value: Shipping price corresponding to the item**

3.2. ORIGINAL BASES

Description of the **olist_sellers_dataset:**

Number of records

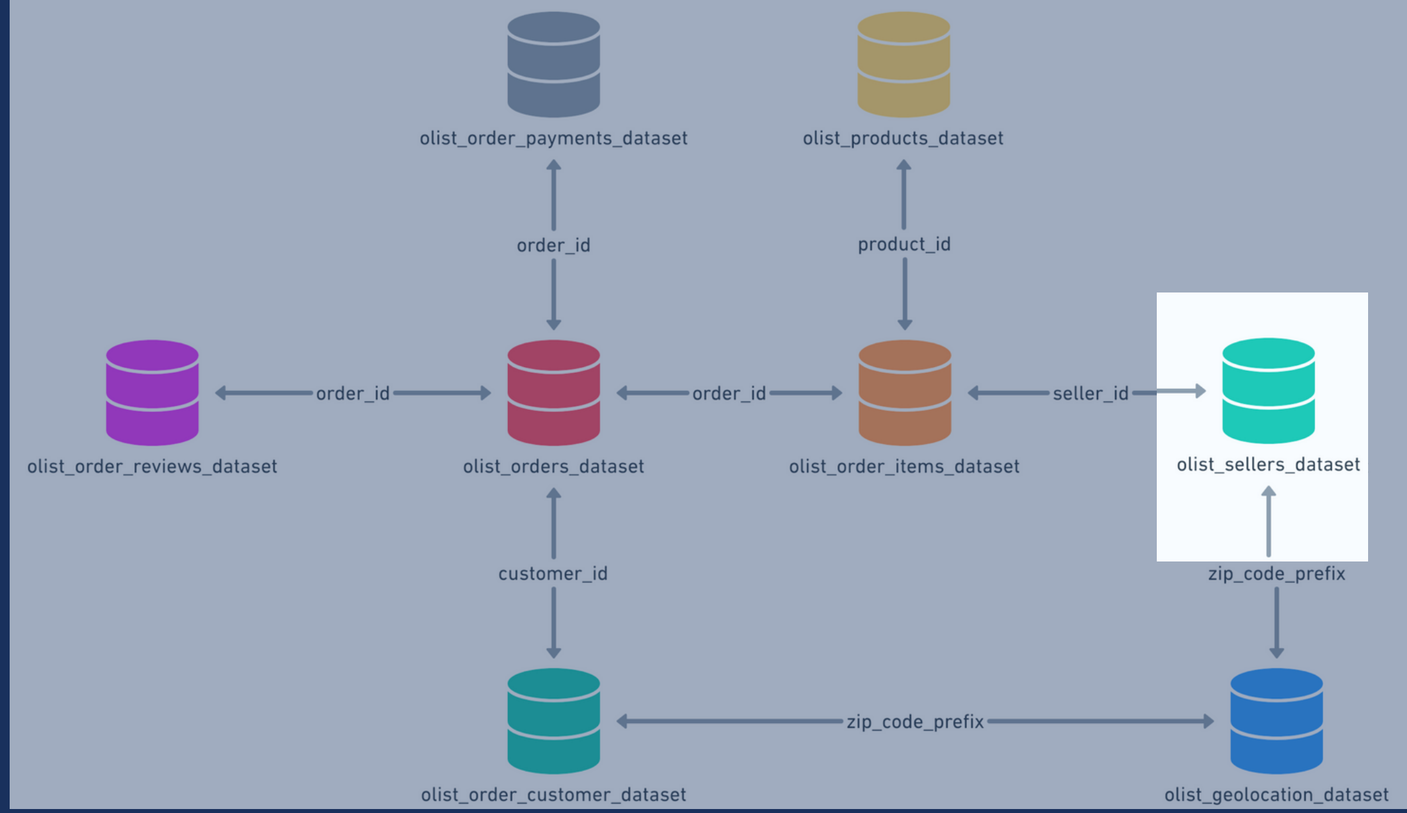
- **3,095**

Number of variables

- **4**

Variable names:

1. **Seller_id**: Unique seller ID
2. **Seller_zip_code_prefix**: First 5 digits of the seller's zip code
3. **Seller_city**: Name of the seller's city
4. S**eller_state**: State of the seller



3.2. ORIGINAL BASES

Description of the **olist_products_dataset:**

Number of records

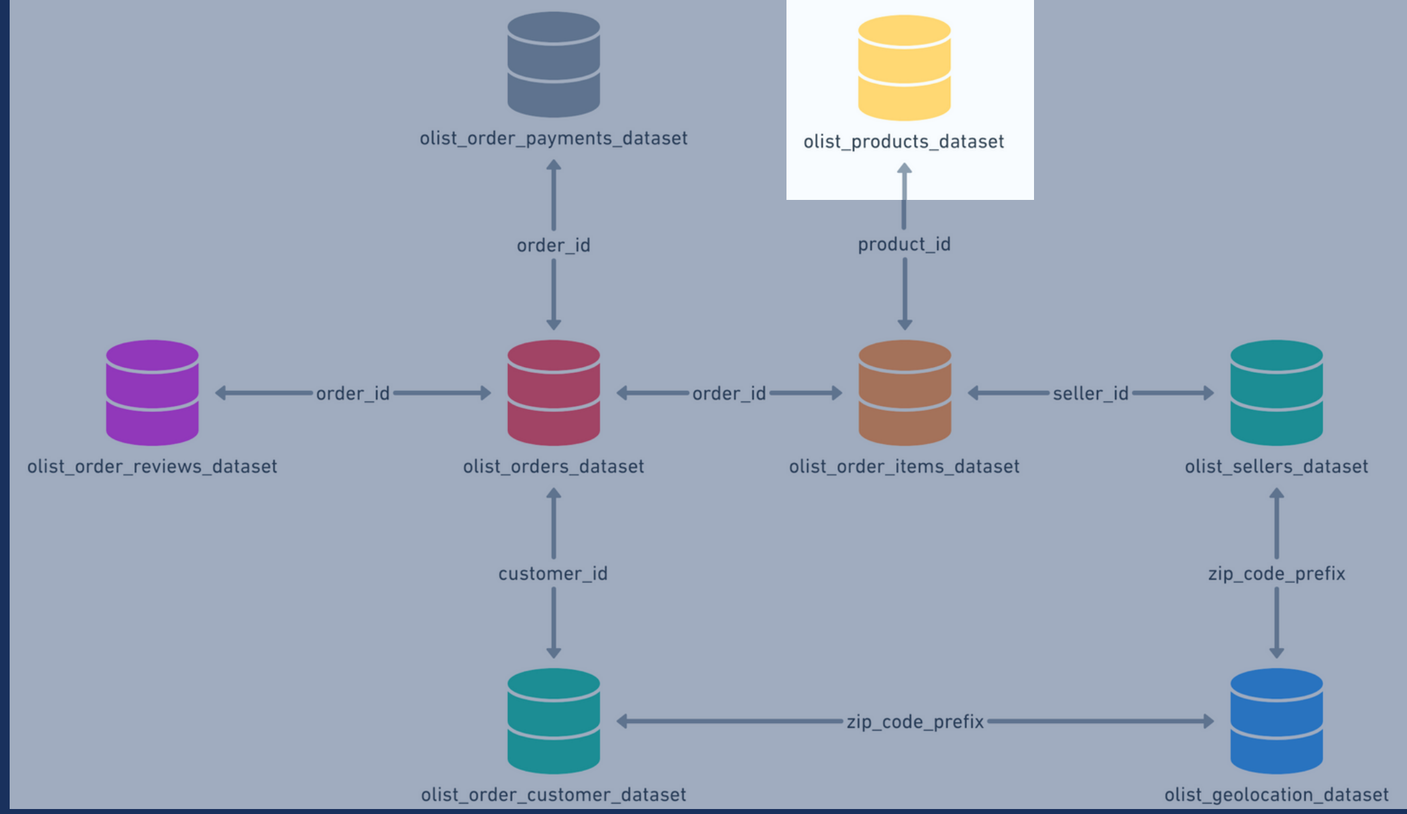
- **32,951**

Number of variables

- **9**

Variable names:

1. **Product_id**: Unique product ID
2. **product_category_name**: Root of the product category
3. **Product_name_lenght**: Length of the product name
4. **Product_description_lenght**: Length of the product description
5. **Product_photos_qty**: Number of product photos
6. **Product_weight_g**: Product weight in grams
7. **Product_lenght_cm**: Product length in centimeters
8. **Product_height_cm**: Product height in centimeters
9. **Product_width_cm**: Product width in centimeters



3.2. ORIGINAL BASES

Description of the **olist_customers_dataset:**

Number of records

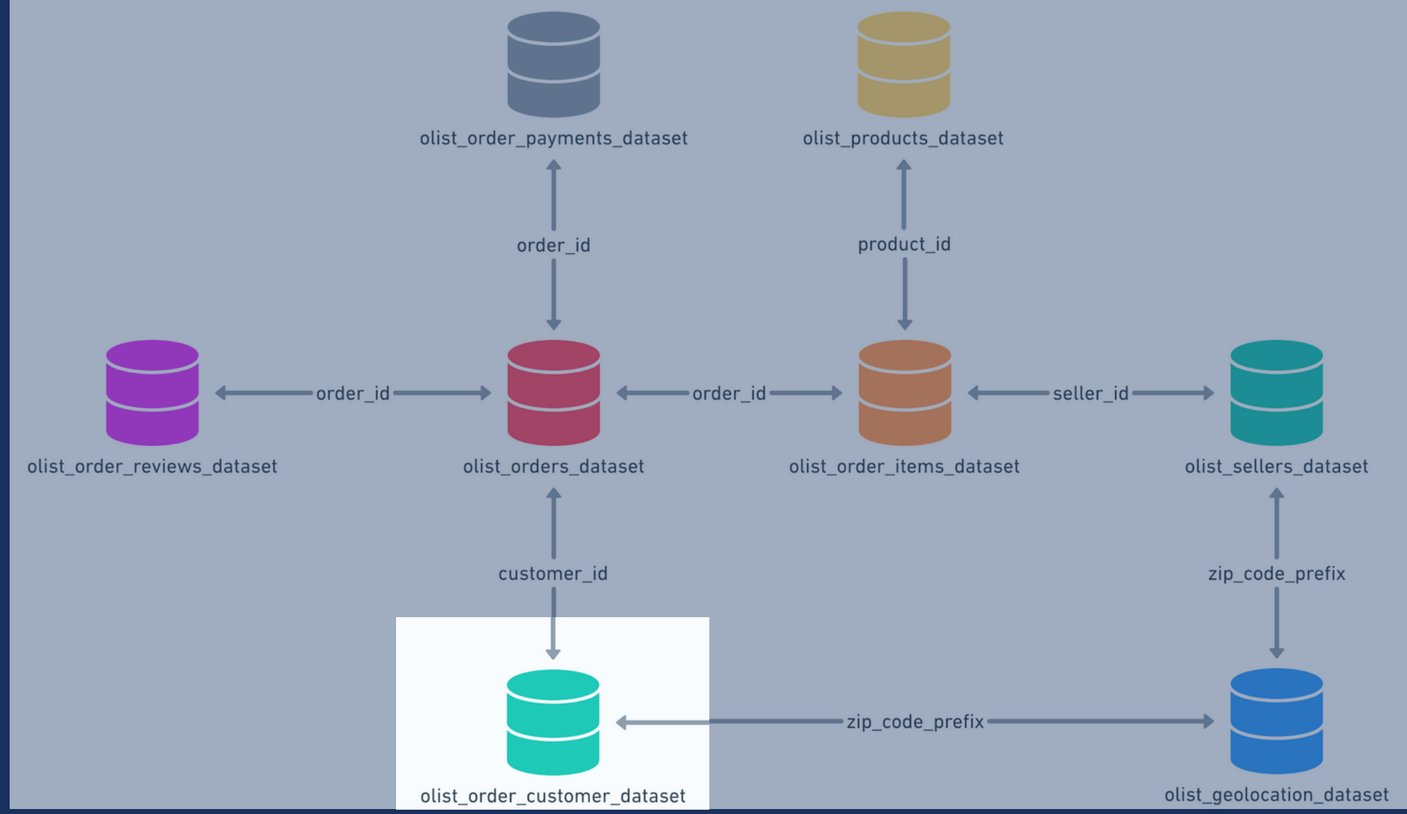
- **99.441**

Number of variables

- **5**

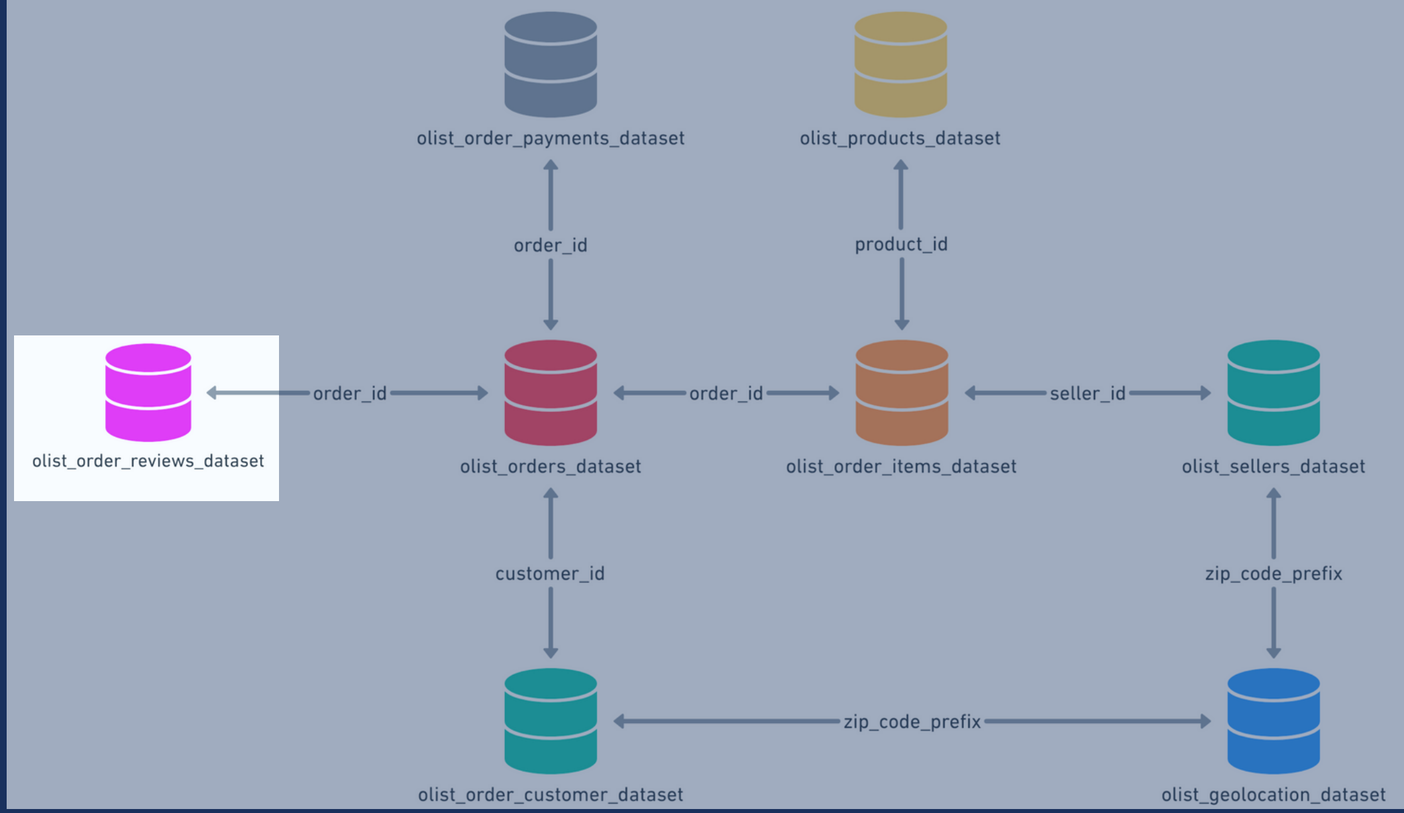
Variable names:

1. **Customer_id**: Key to the order dataset, each order has a unique customer_id
2. **Customer_unique_id**: Buyer's unique ID
3. **Customer_zip_code_prefix**: First five digits of the buyer's zip code
4. **Customer_city**: City of the buyer
5. **Customer_state**: State of the buyer



3.2. ORIGINAL BASES

Description of the **order_reviews_dataset:**



Number of records

- **100,000**

Number of variables

- **7**

Variable names:

1. **Review_id:** Unique review id
2. **Order_id:** Unique order id
3. **Review_score:** Rating from 1 to 5 given by the buyer
4. **Review_comment_title:** Title of the review
5. **Review_comment_message:** Review comment
6. **Review_creation_date:** Date the satisfaction survey was sent to the buyer
7. **Review_answer_timestamp:** Date the buyer responded to the satisfaction survey

3.2. ORIGINAL BASES

Description of the **order_payments_dataset:**

Number of records

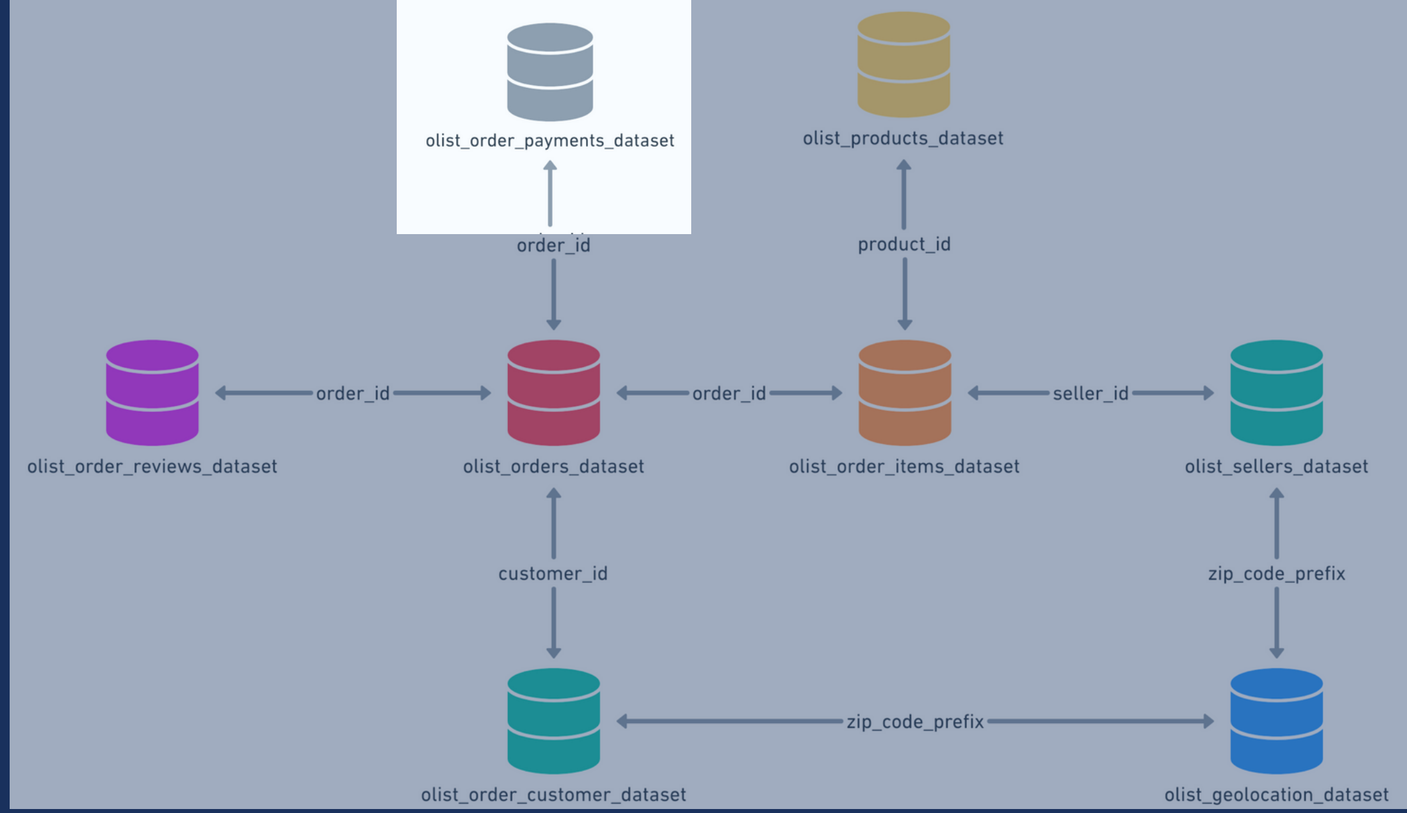
- **99,440**

Number of variables

- **5**

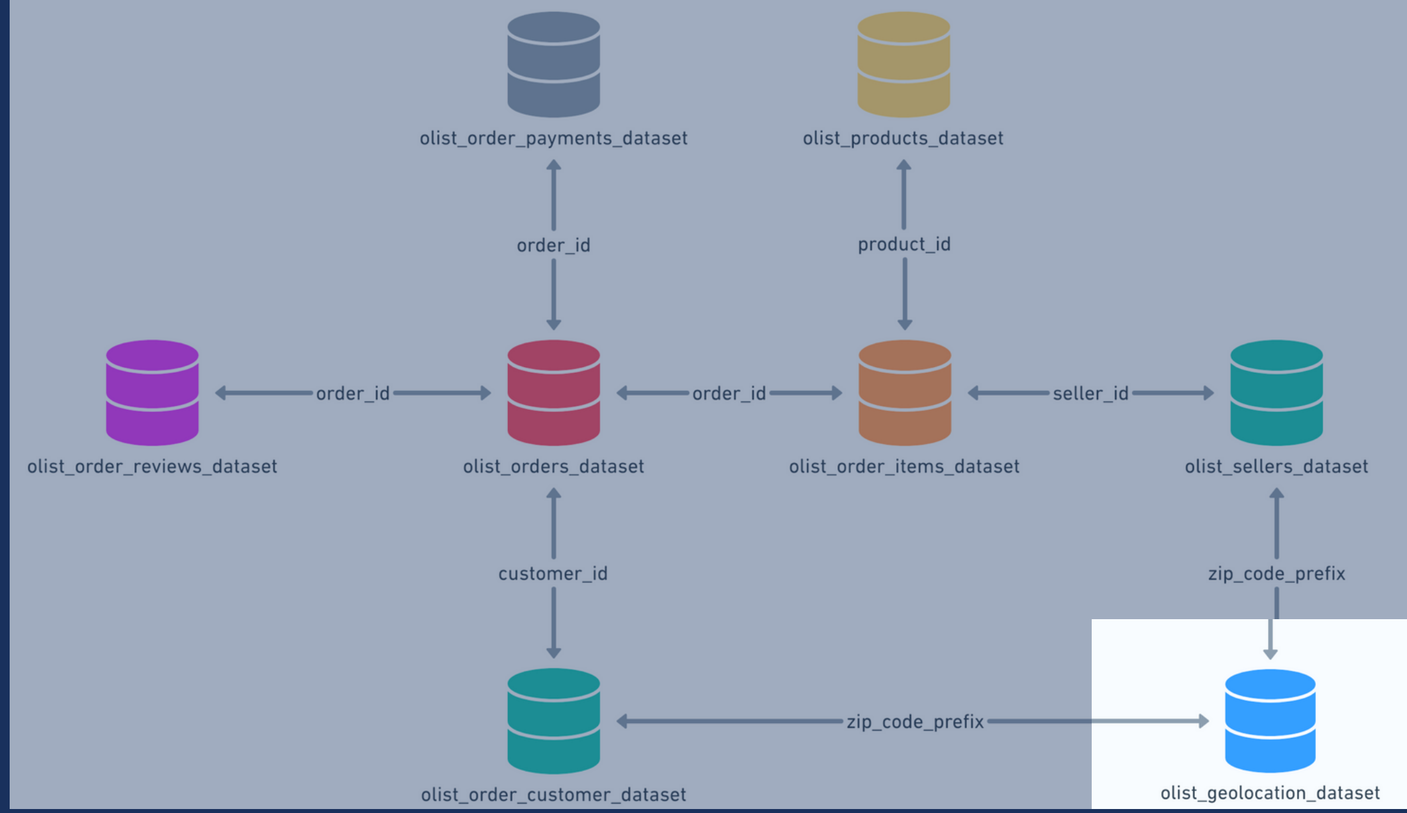
Variable names:

1. **order_id:** Unique identifier of an order.
2. **payment_sequential:** A customer may pay an order with more than one payment method. If he does so, a sequence will be created to accommodate all payments.
3. **payment_type:** Method of payment chosen by the customer.
4. **payment_installments:**number of installments chosen by the customer.
5. **payment_value:** transaction value.



3.2. ORIGINAL BASES

Description of the **order_reviews_dataset:**



Number of records

- **1,000,000**

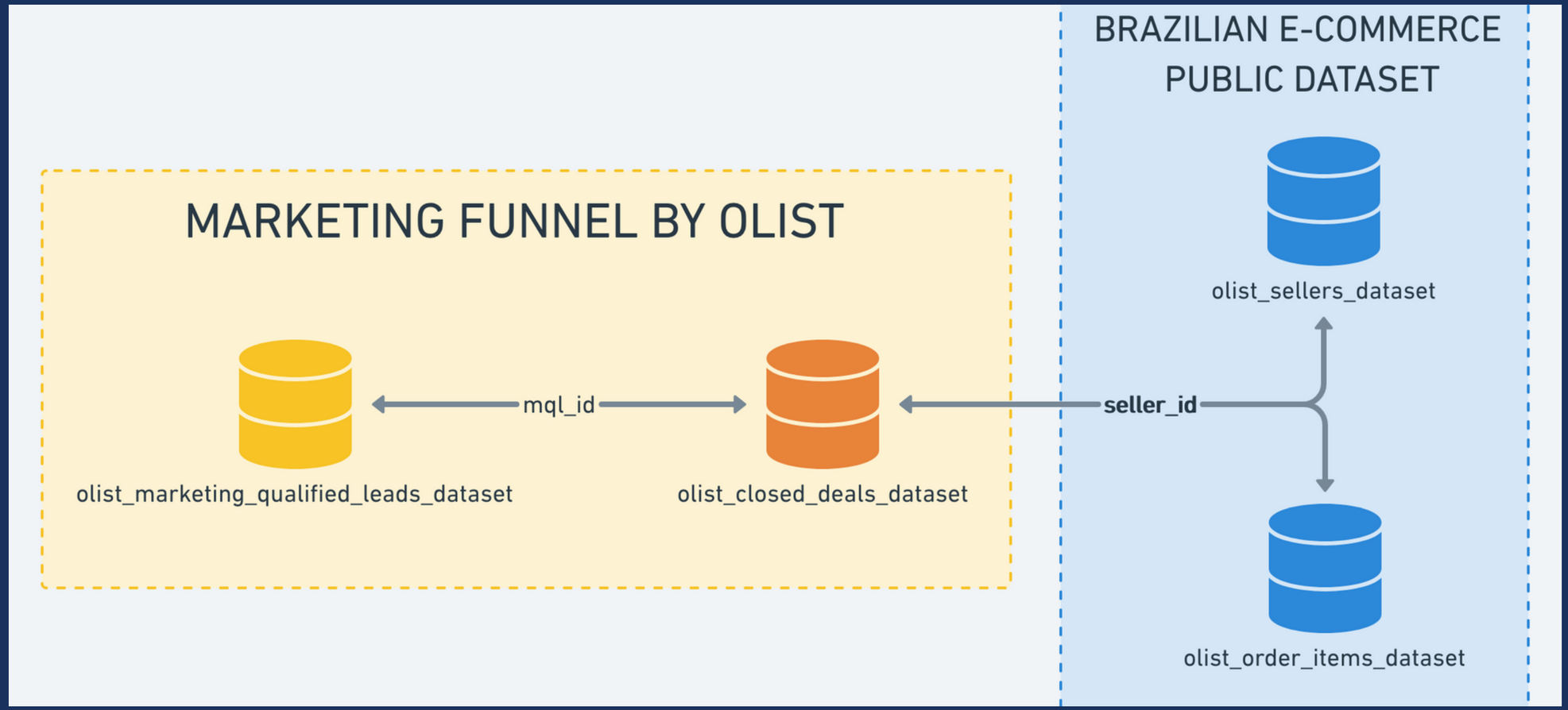
Number of variables

- **5**

Variable names:

1. **geolocation_zip_code_prefix:** first 5 digits of zip code
2. **geolocation_lat:** latitude
3. **geolocation_lng:** longitude
4. **geolocation_city:** city name
5. **geolocation_state:** state

3.3 DATABASE SCHEMA (MARKETING)



3.4. ORIGINAL BASES

Description of the **olist_marketing_qualified_leads_dataset:**

Number of records

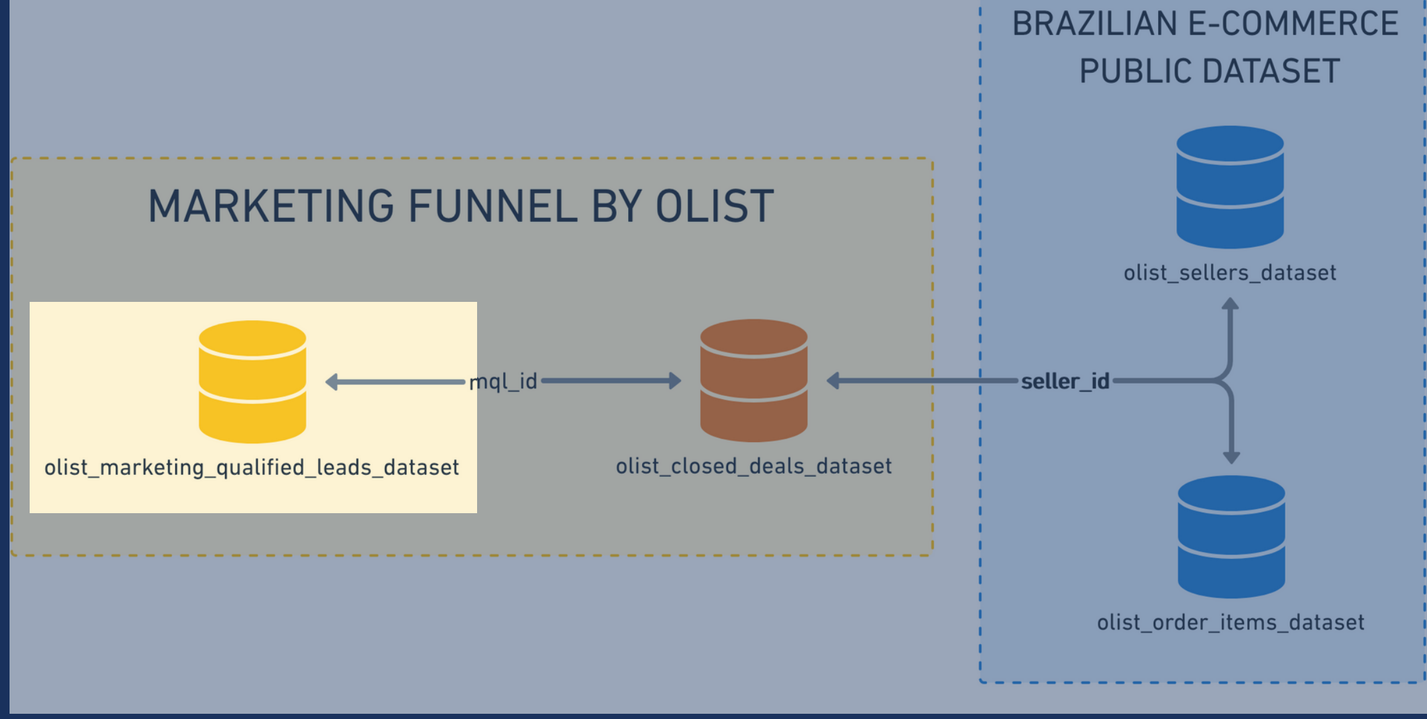
- **8,000**

Number of variables

- **4**

Variable names:

1. **mql_id:**Marketing Qualified Lead id
2. **first_contact_date:** Date of the first contact solicitation.
3. **landing_page_id:** Landing page id where the lead was acquired
4. **Type of media:** where the lead was acquired



3.4. ORIGINAL BASES

Description of the **olist_closed_deals_dataset:**

Number of records

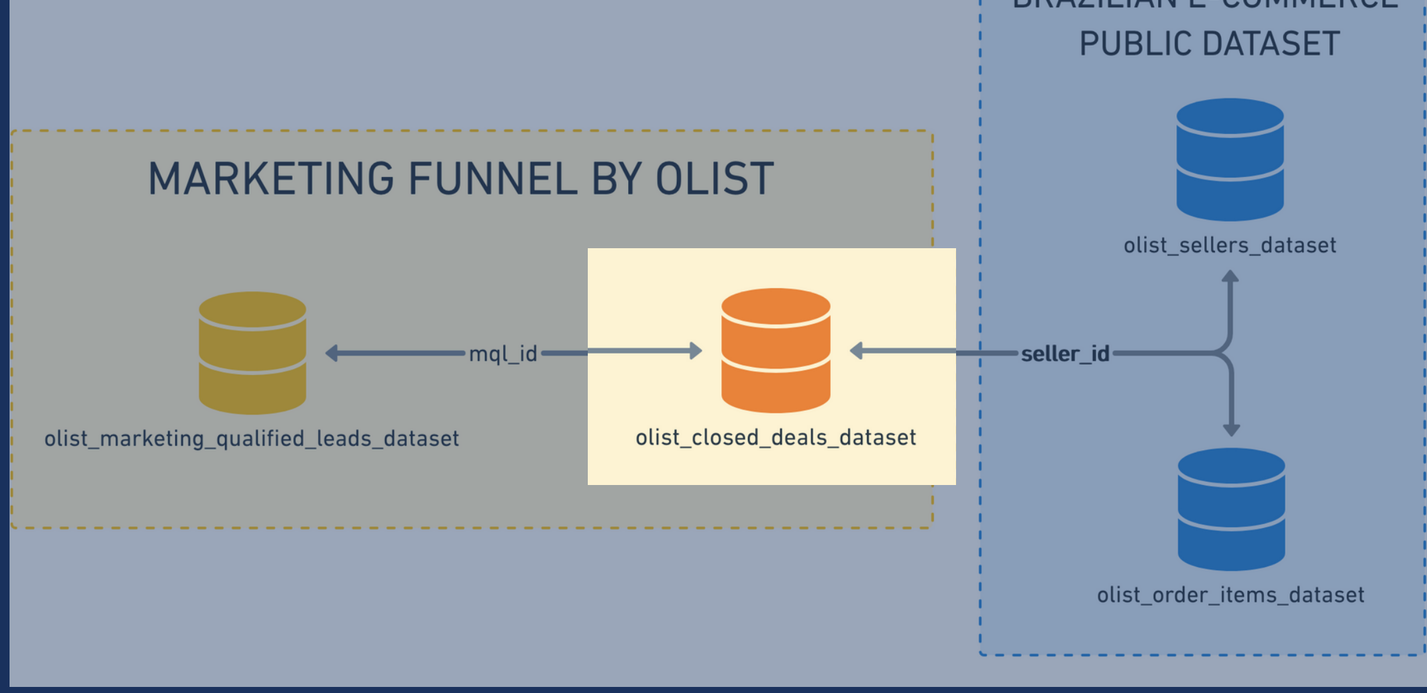
- **842**

Number of variables

- **14**

Variable names:

1. **mql_id:** Marketing Qualified Lead id
2. **seller_id:** Seller id
3. **sdr_id:** Sales Development Representative id
4. **sr_id:**Sales Representative
5. **won_date:**Date the deal was closed.
6. **business_segment:**Lead business segment. Informed on contact.
7. **lead_type:**Lead type. Informed on contact.
8. **lead_behaviour_profile:**Lead behaviour profile. SDR identify it on contact.
9. **has_company:**Does the lead have a company (formal documentation)?
10. **has_gtin:** Does the lead have Global Trade Item Number (barcode) for his products?
11. **average_stock:** Lead declared average stock. Informed on contact.
12. **business_type:** Type of business (reseller/manufacturer etc.)
13. **declared_product_catalog_size:** Lead declared catalog size. Informed on contact.
14. **declared_monthly_revenue:** Lead declared estimated monthly revenue. Informed on contact.



4. DATA TRANSFORMATION

Original Bases

- 10 different tables
- 68 variables

Transformations

- Table joins
- Removal of 12 unused variables
- Imputation of missing values
- Development of calculated columns, measures and tables

Final Base

- Analytical Base Table (ABT) for customer segmentation through unsupervised learning models.
- As Olist's customers are sellers, we need a table with the characteristics of each seller:
- Count of unique order_ids (qty_orders);
- Count of product_id sold by the seller (qty_products);
- Division of the product_id count by the count of unique order_ids (products_per_order);
- Sum of the seller's price and freight_value (sales_value);
- Division of sales_value by count of unique order_id (ticket_medio);
- Count of unique customer_unique_ids of the seller (qty_buyers);
- Count of unique order_status, to quantify sellers' delivery quality;
- Recency = last date of the dataset minus the date of the seller's last order;
- Frequency = average number of sales per month;
- Average of the differences between the seller's order_estimated_delivery_date and order_delivered_customer_date (average_delay_days);
- Average seller review_score;
- Product categories sold by the seller;
- seller_state column.

DATA ANALYSIS METHODOLOGY

Planning → **Exploratory Analysis** → Statistical Modelling → ML Algorithms → Conclusion

Problem definition

- Goals
- Concepts
- Criteria
- Data history
- Variables

Primary Analysis

- Position measurements
- Frequency analysis Graphics
- Outlier analysis
- Missing analysis
- Validation on the
- Consistency of information

Evaluation of techniques

- Native K-means using Spark

Evaluation of techniques

• biSecting K-means Spark native

• Gaussian Mixture Native model in Spark

• Scikit Learn: DBSCAN, MeanShift, K- means Clustering agorithms

Key Actionable Insights & takeaways

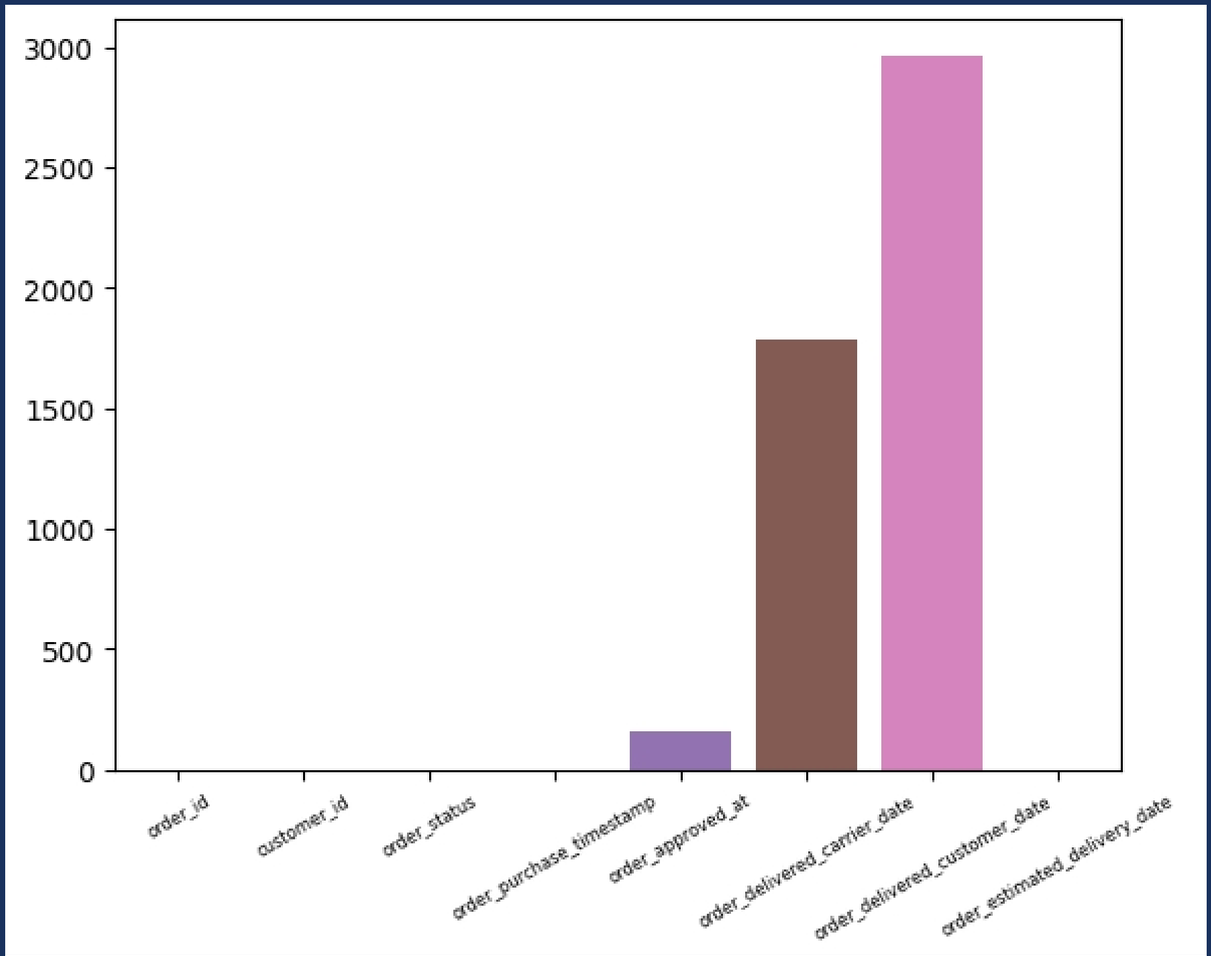
• Definition of the technique

• Validation of results

• Choice of technique what better if suitable for use and strategies

4. EXPLORATORY DATA ANALYSIS IN POWER BI

Through data exploration, we defined what direction we would give to creation of the models and necessary treatments for the databases.



As null values represent a small volume in the bases used for the creation of **Analytical Base Table**(ABT), we decided to just remove them. The **olist_orders_table** base null value graph exemplifies this

4. EXPLORATORY DATA ANALYSIS IN POWER BI

- **Number of orders Over time**

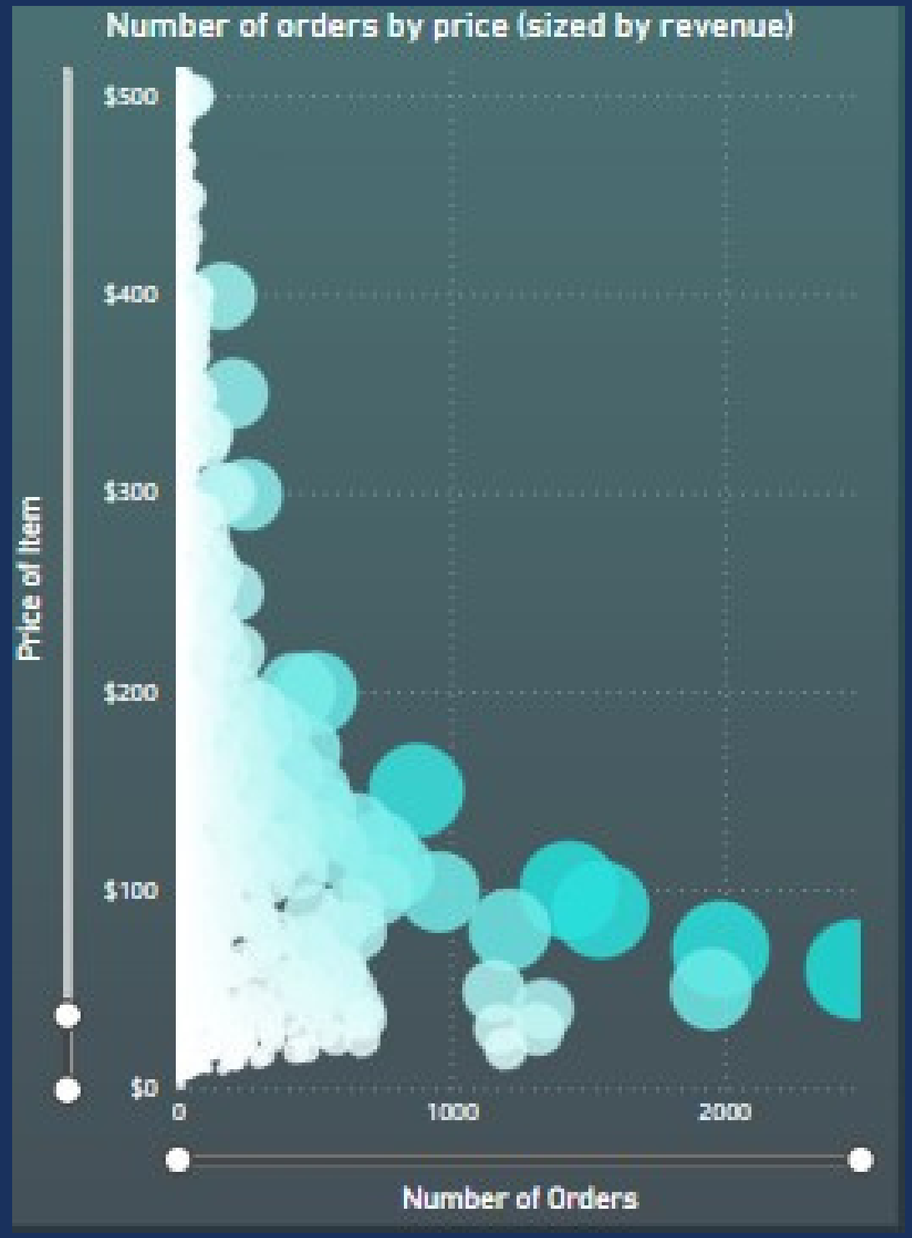
We noticed a spike in the number of registrations for the day November 24, 2017, this is because Black Friday occurred on this day. Over all there is a noticeable upward trend in monthly orders.

4. EXPLORATORY DATA ANALYSIS IN POWER BI

- **Number of orders by Pricing & sized by revenue generated**

We can see that its a right skewed distribution, where the number of orders are densely located on the lower end of price variable.

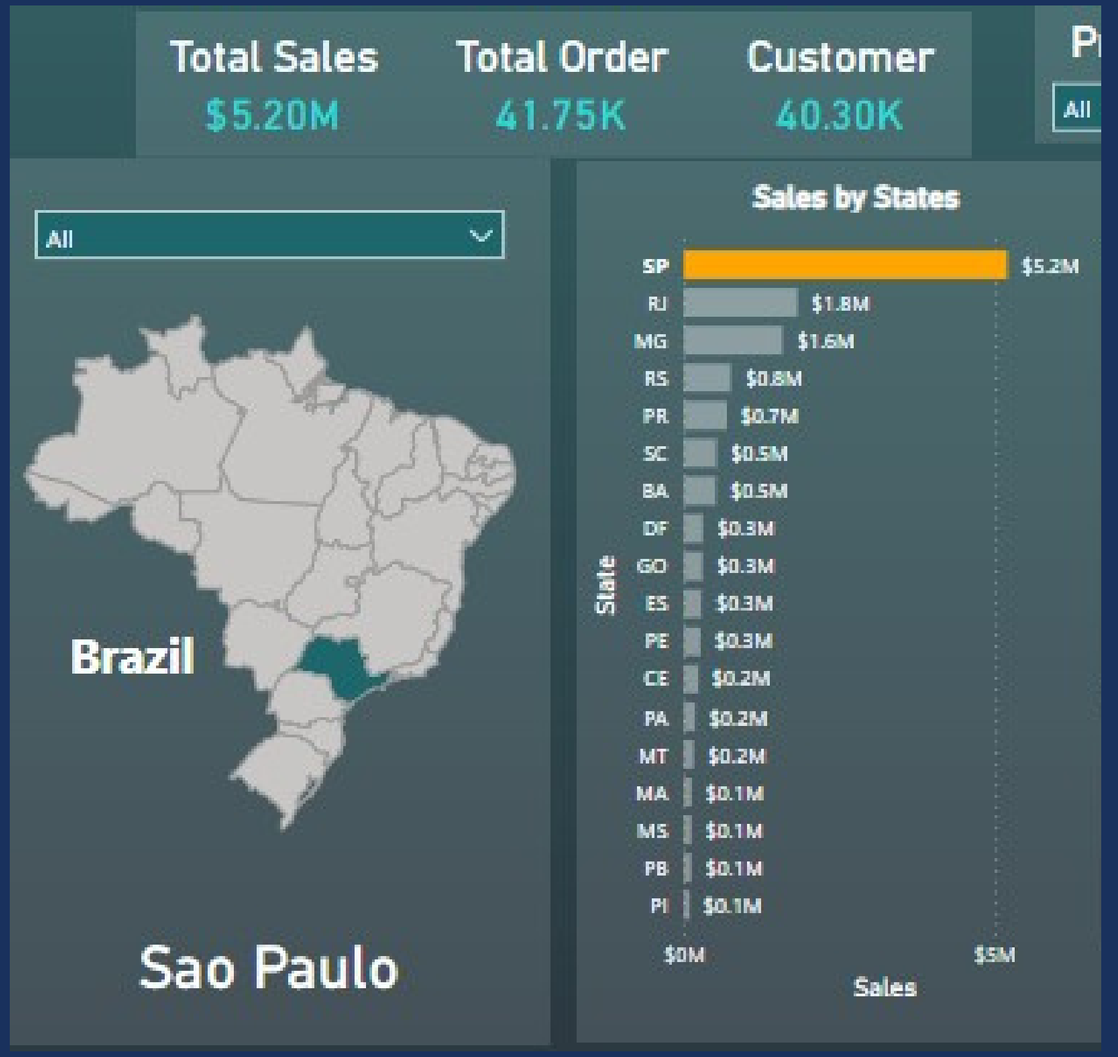
- 25% of the products sold are in the range up to R$40
- 75% of the products sold are in the range up to R$130.
- Even if the sales volume is high in the under $130 range, the profit margin on each individual item sold might be lower compared to higher-priced categories.



4. EXPLORATORY DATA ANALYSIS IN POWER BI

- **Which state has the highest sales?**

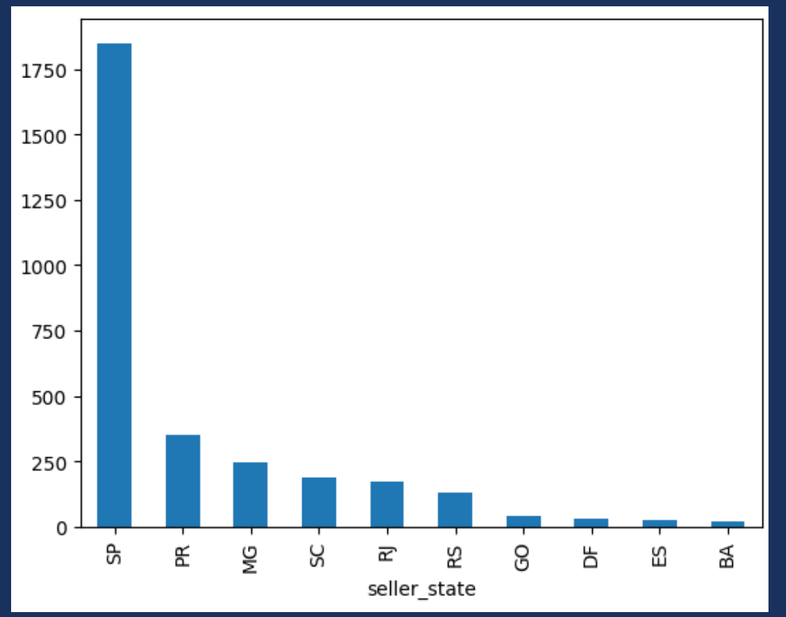
The state with the highest sales is in **São Paulo(SP) at $5.2M,** followed by Rio de Janero (RJ) at $1.8M and Minas Gerais(MG) at $1.6M

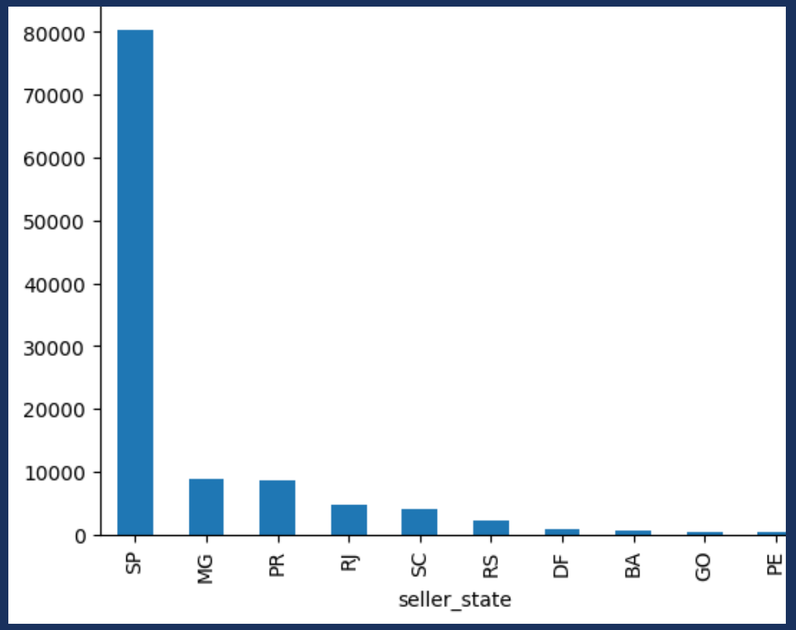


4. EXPLORATORY DATA ANALYSIS IN POWER BI

- **Which state has the highest number of registered sellers?**

The vast majority of sellers registered with sales base is in São Paulo. Followed by Paraná and Minas Gerais.





- **Which state has the highest number of orders sold by sellers?**

After putting the bases together and comparing the number of sales by seller status, we can see that sellers in São Paulo make much more sales than those in others States.

4. EXPLORATORY DATA ANALYSIS IN POWER BI

- **How many orders are we expecting in the near future?**

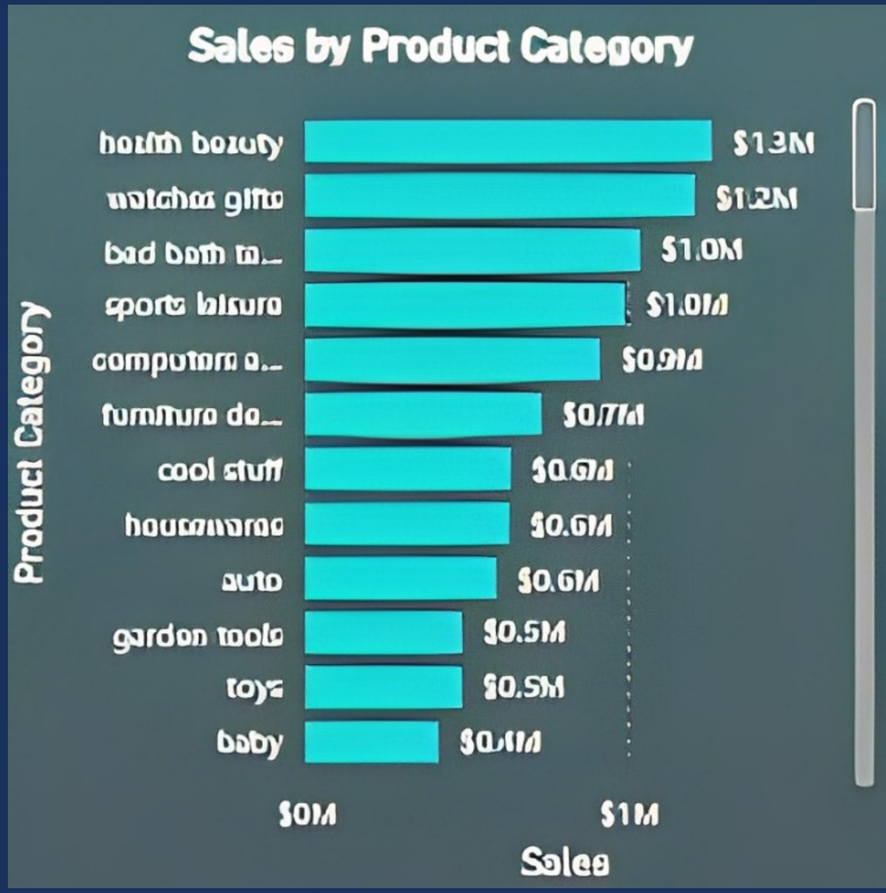
As per forecasting feature from power BI, In the next 6 months, sales will continue to grow at this rate, there is going to be a spike in demand in upcoming Oct-Dec 2018. The number of orders are going to cross 10k in the next year, so inventory planning, warehousing and logistics

4. EXPLORATORY DATA ANALYSIS IN POWER BI

- **Which product category generates highest revenue?**

Health & beauty is the highest revenue generating product at **$1.3M,** followed by watches gifts at $1.2M and then bed bath table at $1.0M

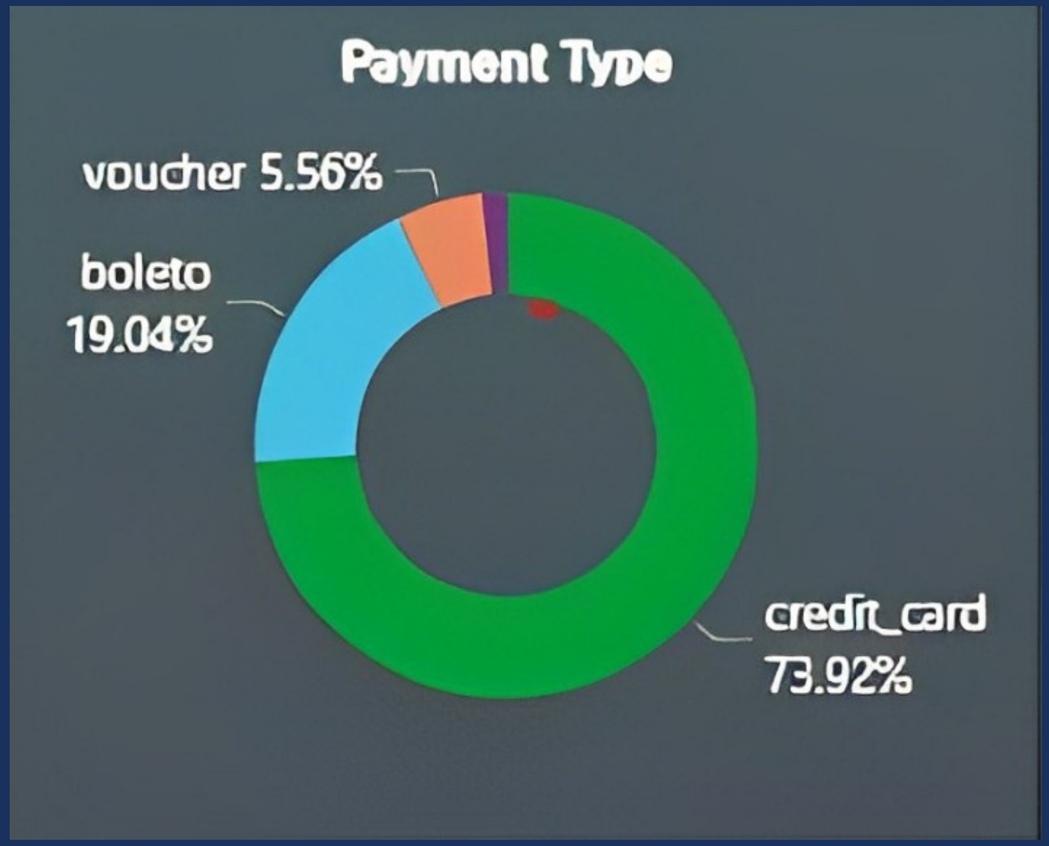
Category	Revenue	Profit ratio	Average price	Order
bed bath table	$1,036,900.60	7.60%	$93.30	
health beauty	$1,258,601.34	9.26%	$130.16	
sports leisure	$988,048.97	7.27%	$114.34	
furniture decor	$729,762.49	5.37%	$87.56	
computers accessories	$911,954.32	6.71%	$116.51	
housewares	$632,248.66	4.65%	$90.79	
watches gifts	$1,205,005.68	8.87%	$201.14	
telephony	$323,667.53	2.38%	$71.21	
garden tools	$485,256.46	3.57%	$111.69	
auto	$592,720.11	4.36%	$139.96	



4. EXPLORATORY DATA ANALYSIS IN POWER BI

- **What is the preferred mode of payment among customers?**

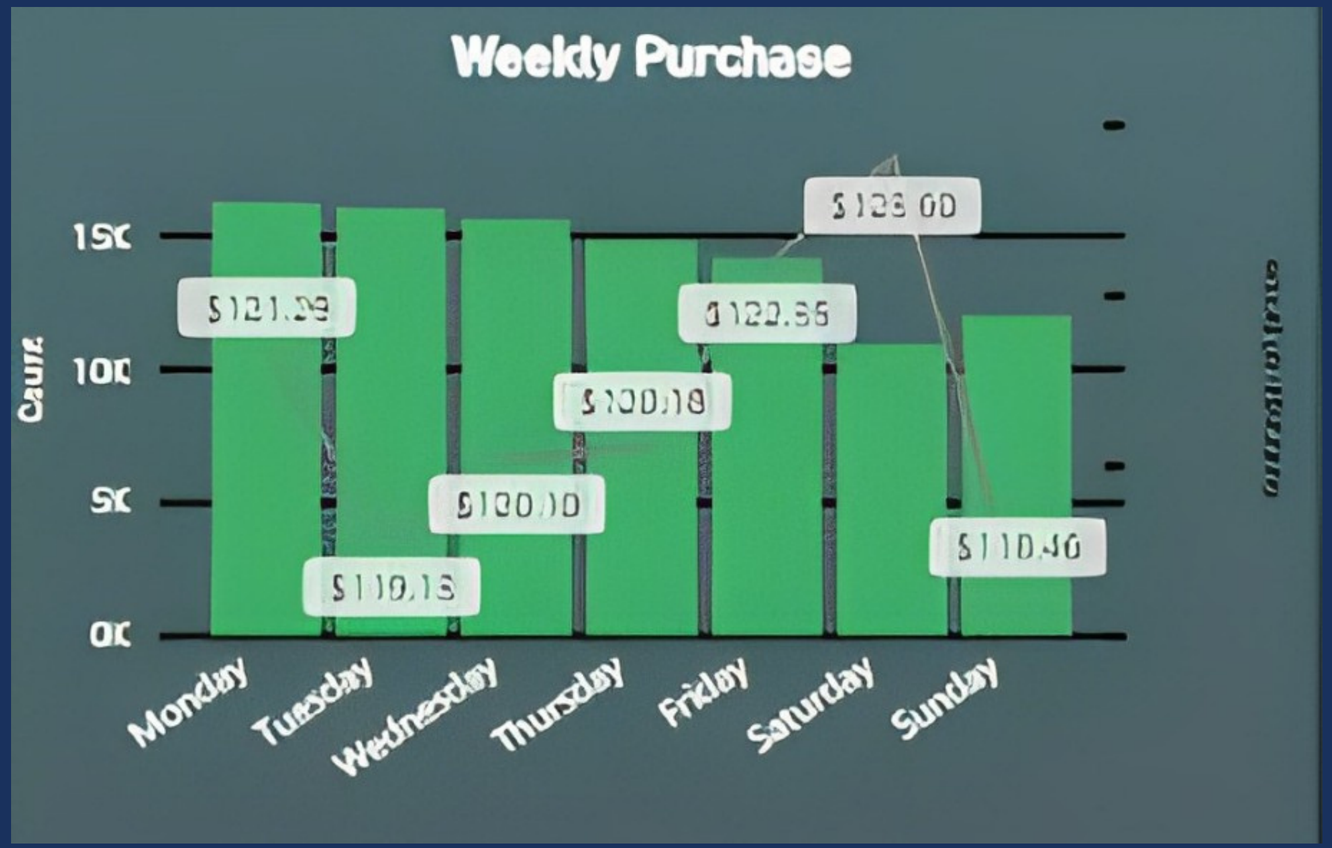
Credit Card is the most prefered payment method with a majority of 73.9% of users, followed by boleto and vouchers



4. EXPLORATORY DATA ANALYSIS IN POWER BI

- **What time of the week purchases are made the most?**

An interesting insight here is that eventhough Monday has the most number of orders purchases (15k+ orders), the Average Order Value (AOV) on Saturday is the highest ($123.60)

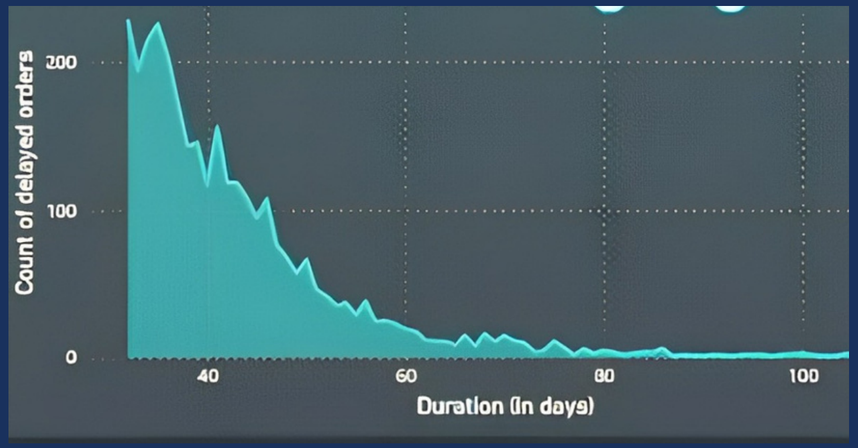


4. EXPLORATORY DATA ANALYSIS IN POWER BI

- **How many orders were delivered- on time vs delayed?**

A total of 96,480 orders were delivered out of which 93% were on time and 7% were delayed.

- **What was the delivery delay period?**



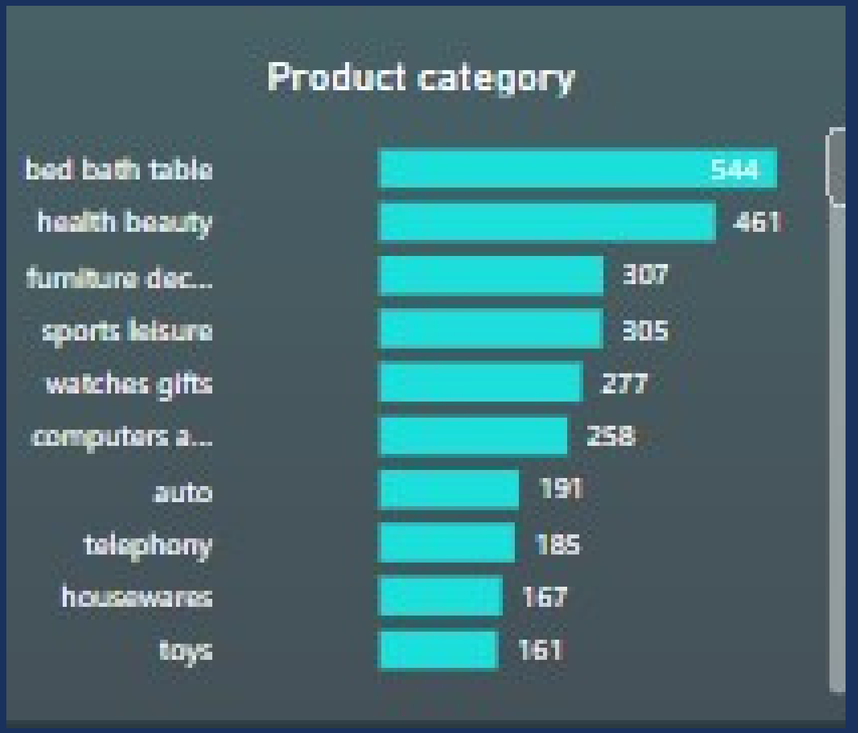
The delay duration in late delivery ranges from a min of 4 days to max of 210 days with An average delivery delay of 34 days.

4. EXPLORATORY DATA ANALYSIS IN POWER BI

- **Which product category accounts for highest delayed deliveries?**

A shocking insight: our top 2 highest revenue generating product categories(bed bath table and health & beauty) have seen the most number of delayed deliveries.

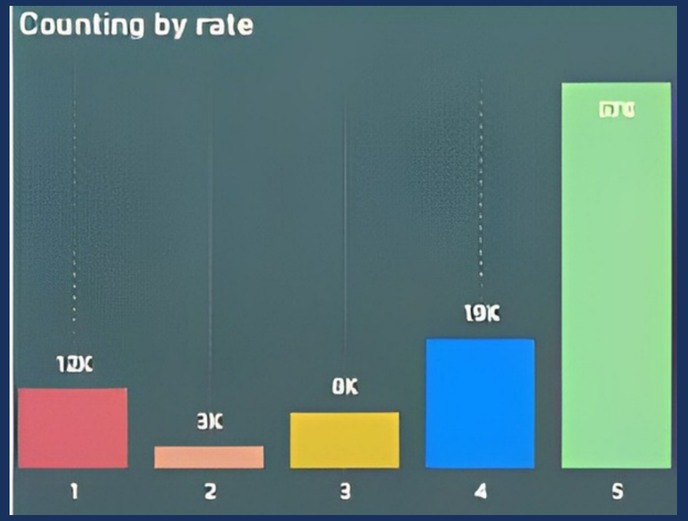
Around 27% (1k/3.6k=0.27) of total late deliveries are coming from these two categories.

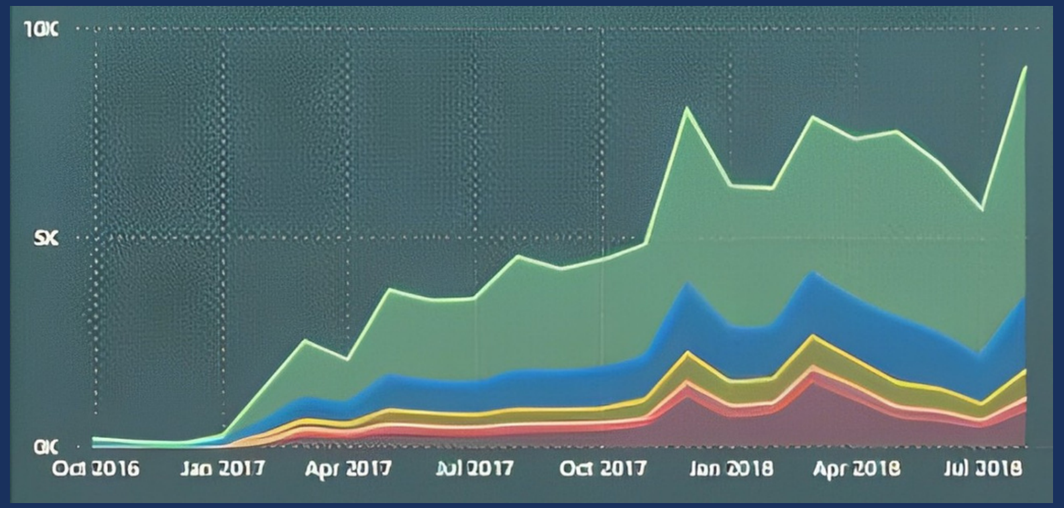


4. EXPLORATORY DATA ANALYSIS IN POWER BI

- **What is the count of various rating?**

A total of 57K review of 5 star rating out of 114k reviews, deriving from this we have 76% positive, 6% neutral and 18% negative sentiment





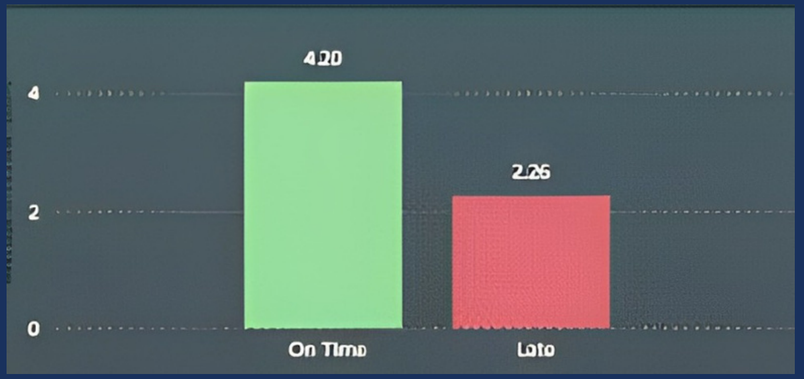
- **How has the rating changed overtime?**

The ratings have picked up since Jan 2017, and it is at an all time high in July 2018. **5 stars** rating observing the most drastic spike out of the whole group

4. EXPLORATORY DATA ANALYSIS IN POWER BI

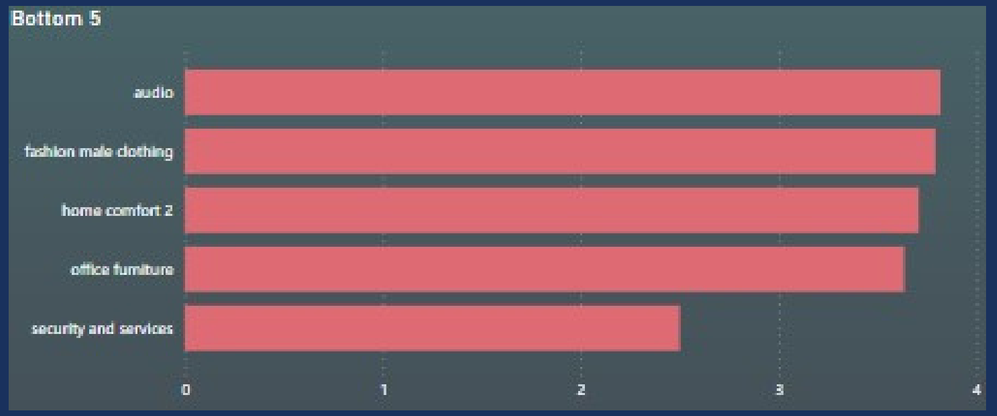
- **How many orders were delivered- on time vs delayed?**

We can observe that there is a direct correlation between delivery time and rating score. On-time deliveries have an average rating of 4.2 while late deliveries have a score of 2.26

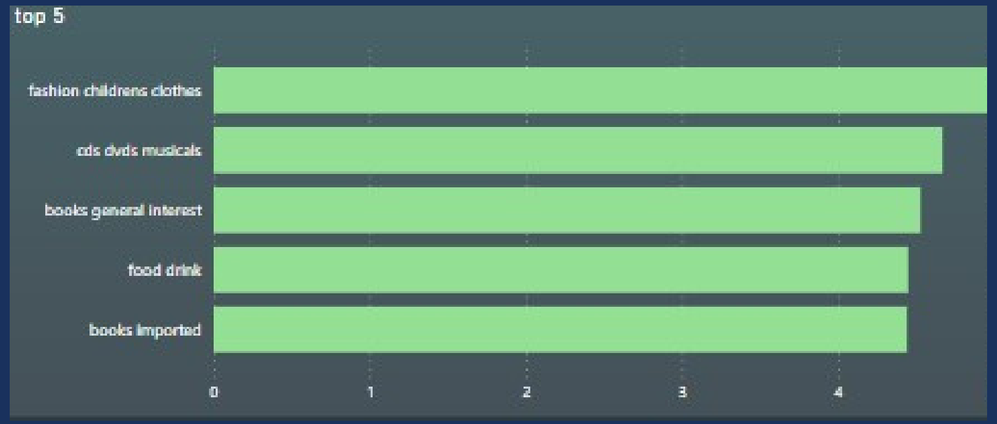


- **Which are the best vs worst rated product catogories?**

Worst rated products have an average rating of 3.6. Consting of Audio, mens fashion, home comfort & furniture

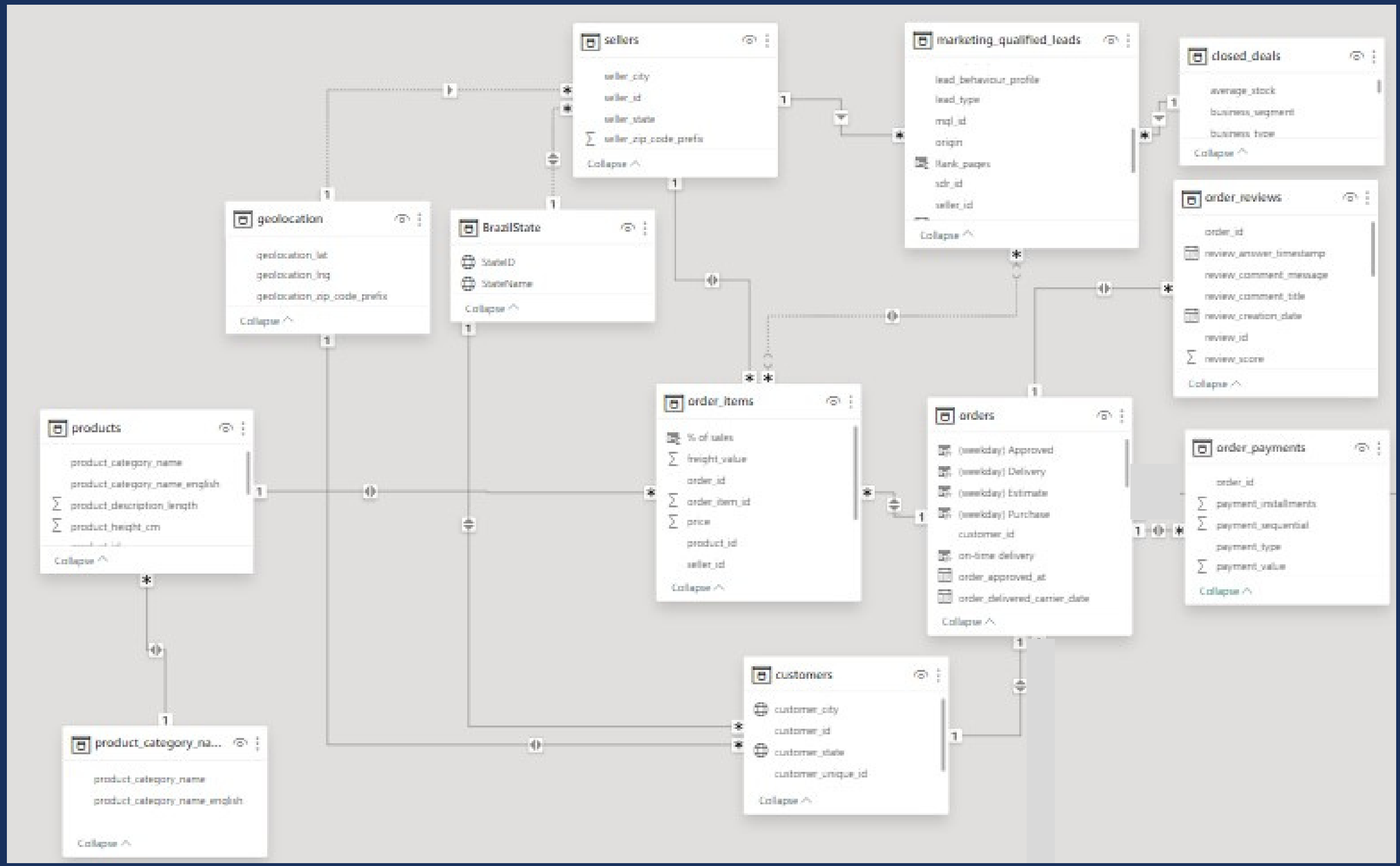


Best rated products have an average rating of 4.5. Consting of kids fashion, books & dvds/cds



Power BI

5. DATA MODELING IN POWER BI



DATA ANALYSIS METHODOLOGY

Planning → Exploratory Analysis → **Statistical Modelling** → ML Algorithms → Conclusion

Problem definition

- Goals
- Concepts
- Criteria
- Data history
- Variables

Primary Analysis

- Position measurements
- Frequency analysis Graphics
- Outlier analysis
- Missing analysis
- Validation on the
- Consistency of information

Evaluation of techniques

- Native K-means using Spark

Evaluation of techniques

- biSecting K-means Spark native
- Gaussian Mixture Native model in Spark
- Scikit Learn: DBSCAN, MeanShift, K- means Clustering agorithms

Key Actionable Insights & takeaways

- Definition of the technique
- Validation of results
- Choice of technique what better if suitable for use and strategies

STATISTICAL MODELING

ANALYTICAL BASE TABLE | VARIABLE SELECTION

- # Requests
- # Products
- # Products per Order
- $ Sales Value
- $ Average Ticket
- # Buyers
- # Orders by Status
- Recency (days since last sale)
- Frequency (sales per month)
- Average days late
- Average review score
- Categories of products sold by the seller
- Seller's UF

STATISTICAL MODELING

ANALYTICAL BASE TABLE | FEATURE ENGINEERING

Due to the application of clustering models, some treatments were applied to the base so that the training was permitted.

Most grouping models do not support categorical variables.

One Hot Encoding

Just like most classification models and regression, clustering models do not support null values.

Removal of Nulls

Because they work with distances, the variables need to be on the same scale in the grouping.

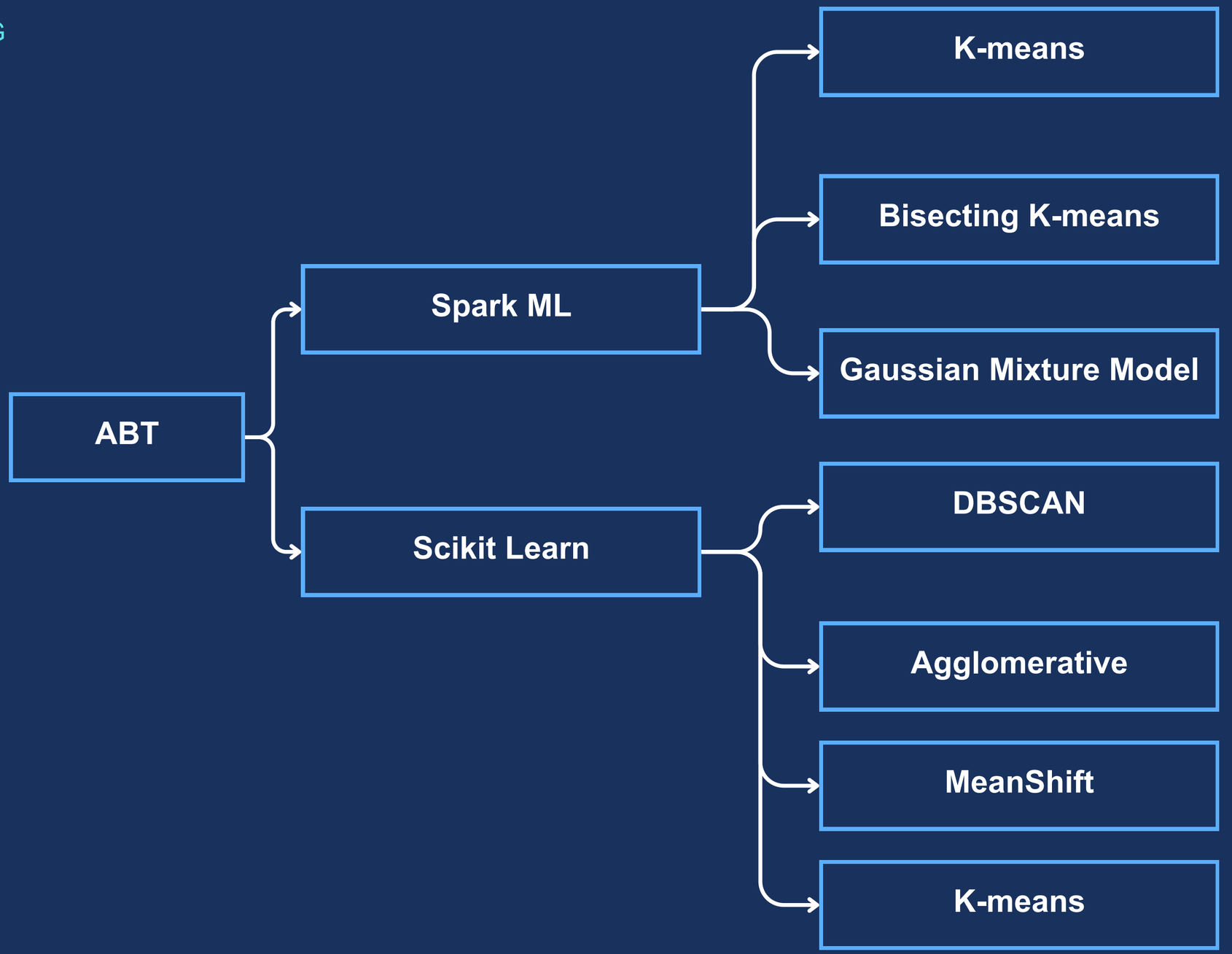
Standardization

STATISTICAL MODELING

SEGMENTATION | MACHINE LEARNING

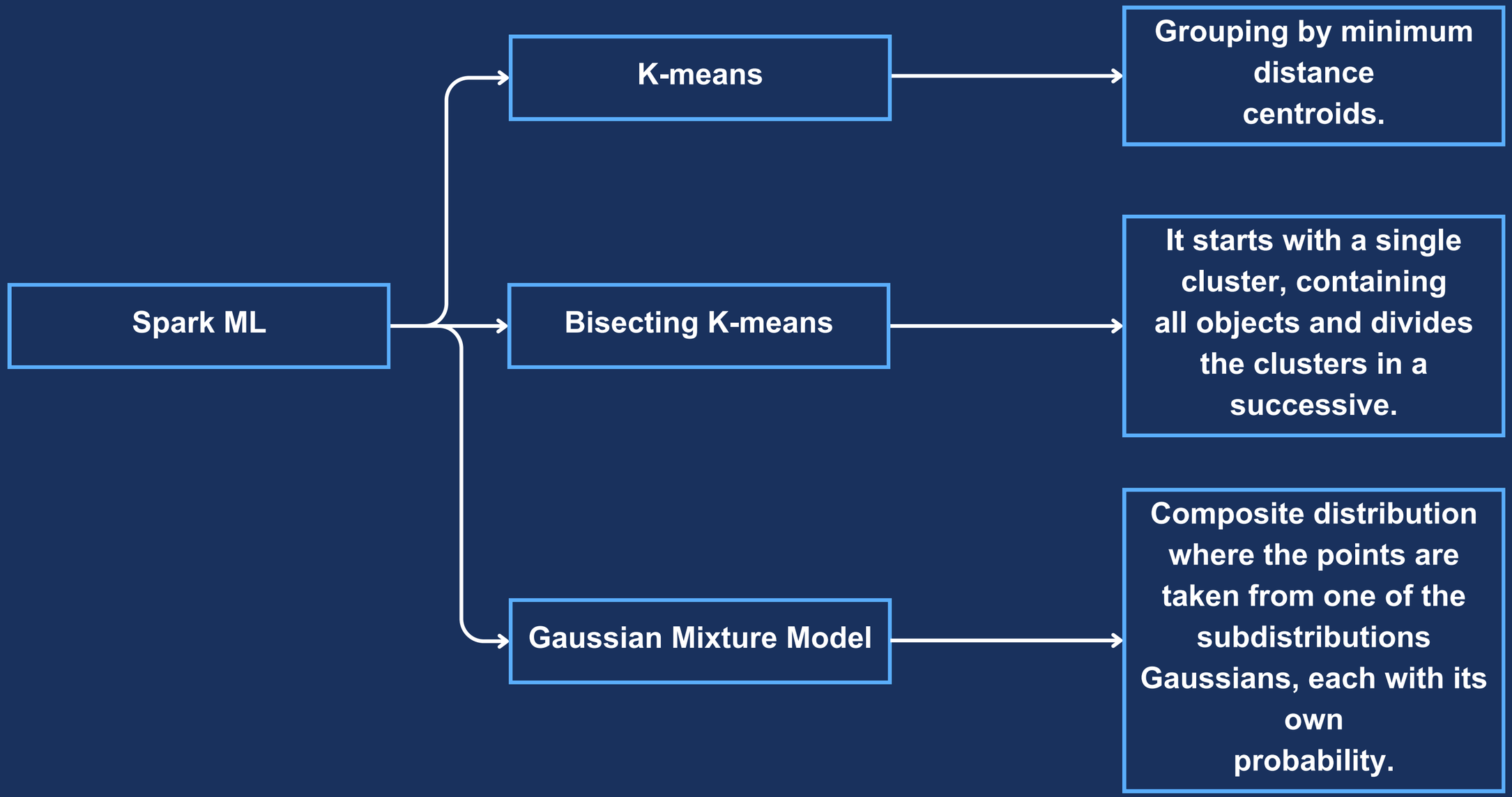
Next, we segmented
of sellers through sales models
machine learning no
supervised: when we do not have
the target variable, we develop
complex mathematical models that
test the correlation between
variables
explanations until you find patterns
that enable grouping.

To define the best way to
grouping, we use techniques of
native modeling of Spark and
Scikit Learn, with different degrees
of success



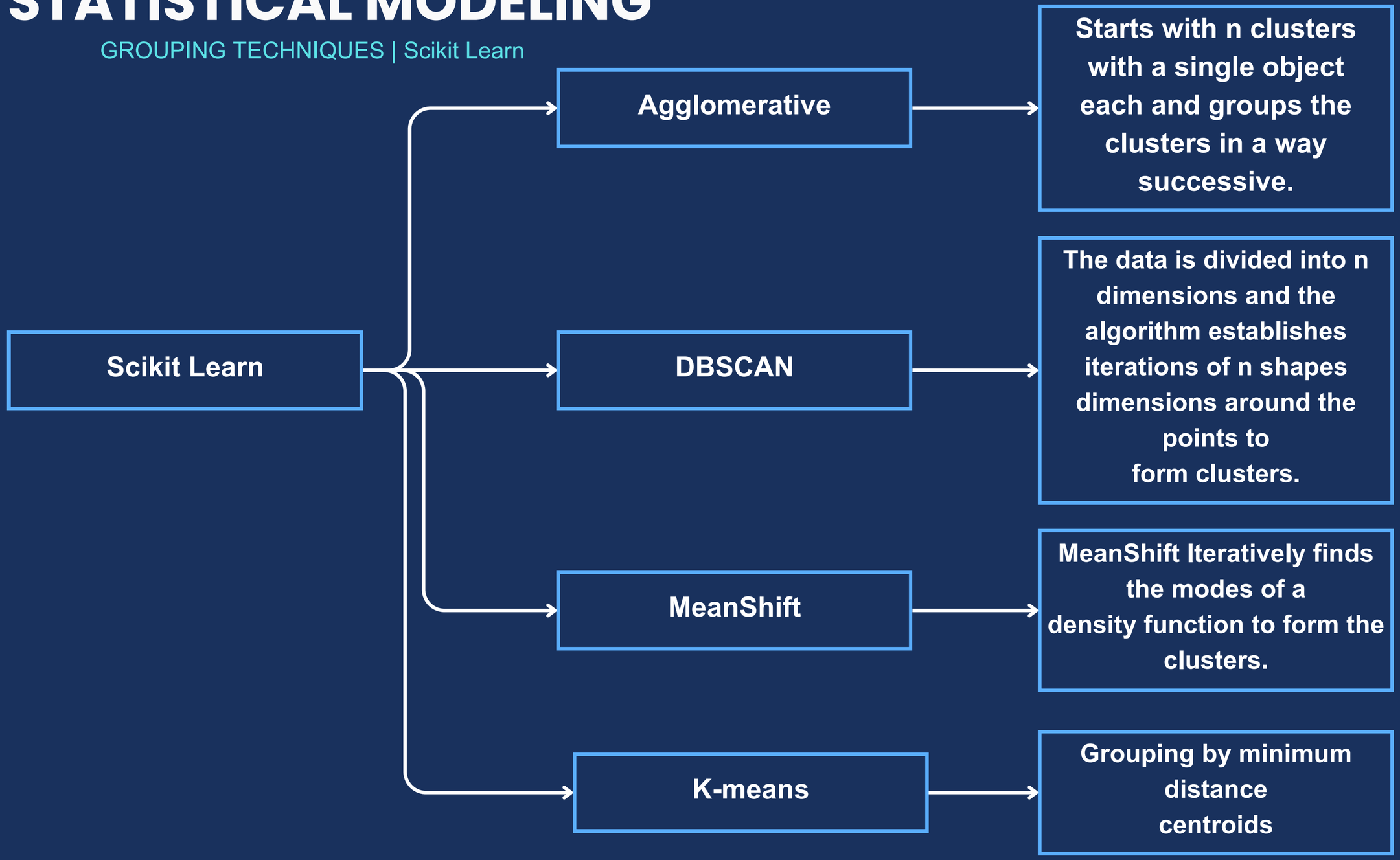
STATISTICAL MODELING

GROUPING TECHNIQUES | SPARK ML



STATISTICAL MODELING

GROUPING TECHNIQUES | Scikit Learn

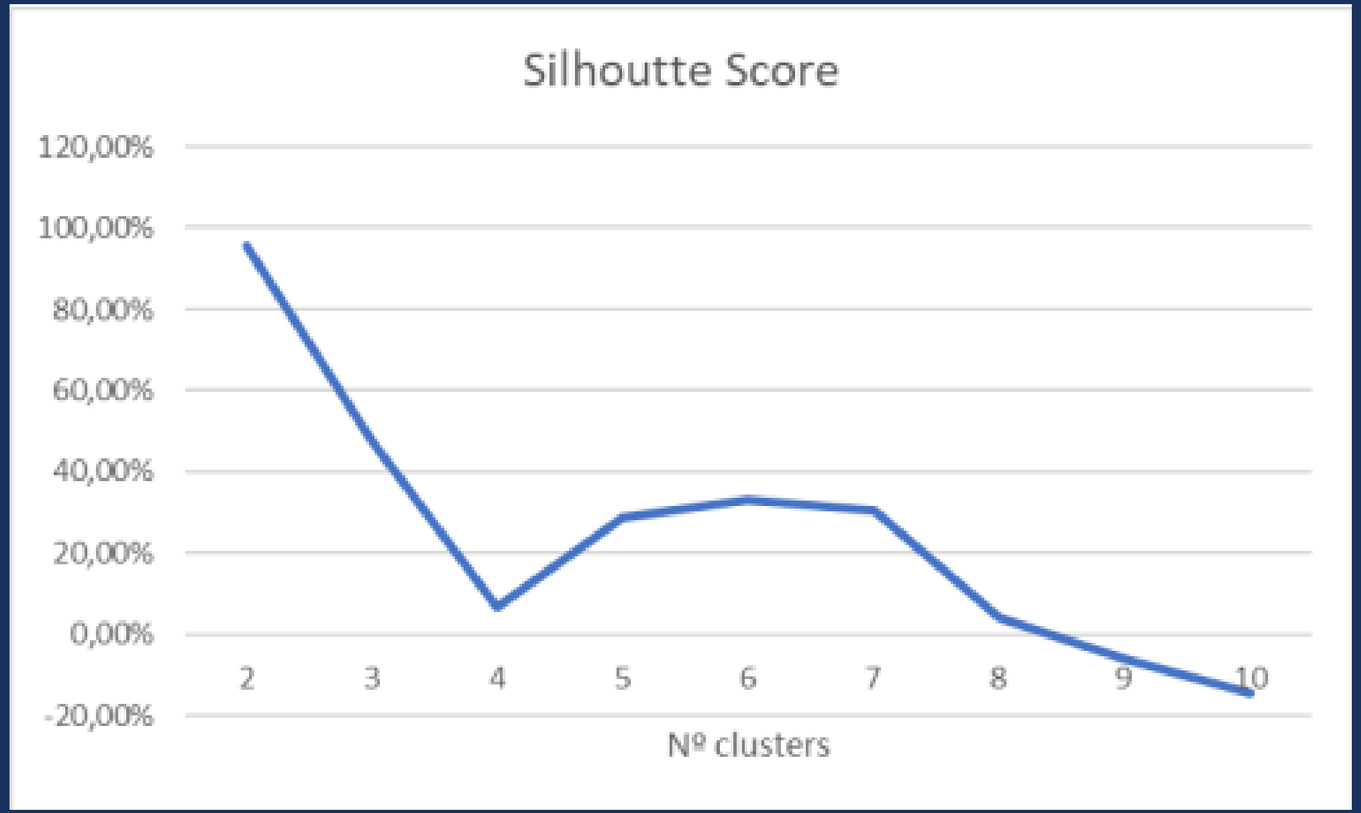


STATISTICAL MODELING

RESULTS | SPARK ML K-MEANS

SparkML
We arrive at the possibility of 3 different groupings: models with 2, 3 or 6 groups.

In all of them the most relevant factor appears to be the number of sales, Then we can infer that the Segmentation occurs mainly among large sellers (probably large stores and retailers) and smaller sellers.



STATISTICAL MODELING

RESULTS | SPARK ML K-MEANS

SparkML
We arrive at the possibility of 3 different groupings: models with 2, 3 or 6 groups.

In all of them the most relevant factor appears to be the number of sales, Then we can infer that the Segmentation occurs mainly among large sellers (probably large stores and retailers) and smaller sellers.

Model separated the sellers among the big sellers (minority) and small sellers (majority). Do not do a lot of sense from the point of view of business because the division is very radical: discarded model

K=2	K=3	K=6
Silhoutte score: 95% Group 0: 3.023 vendors Group 1: 72 vendors	Silhoutte score: 47% Group 0: 2,999 vendors Group 1: 33 vendors Group 2: 72 vendors	Silhoutte score: 33% Group 0: 2,732 vendors Group 1: 2 vendors Group 2: 23 vendors Grupo 3: 331 vendors Group 4: 6 vendors Group 5: 1 vendors

STATISTICAL MODELING

RESULTS | SPARK ML K-MEANS

SparkML

We arrive at the possibility of 3 different groupings: models with 2, 3 or 6 groups.

In all of them the most relevant factor appears to be the number of sales, Then we can infer that the Segmentation occurs mainly among large sellers (probably large stores and retailers) and smaller sellers.

Group 0 makes few sales and is only interesting to to Olist in volume.
Group 1 makes more sales than most sellers, but still at a level well below that of premium sellers. For To compensate, it sells more expensive products.
Group 2 does many sales and guarantees good revenue. They are profiles distinct enough to apply in commercial policy: model selected for final analysis.

K=2	K=3	K=6
Silhoutte score: 95% Group 0: 3.023 vendors Group 1: 72 vendors	Silhoutte score: 47% Group 0: 2,999 vendors Group 1: 33 vendors Group 2: 72 vendors	Silhoutte score: 33% Group 0: 2,732 vendors Group 1: 2 vendors Group 2: 23 vendors Grupo 3: 331 vendors Group 4: 6 vendors Group 5: 1 vendors

STATISTICAL MODELING

RESULTS | SPARK ML K-MEANS

SparkML

We arrive at the possibility of 3 different groupings: models with 2, 3 or 6 groups.

In all of them the most relevant factor appears to be the number of sales, Then we can infer that the Segmentation occurs mainly among large sellers (probably large stores and retailers) and smaller sellers.

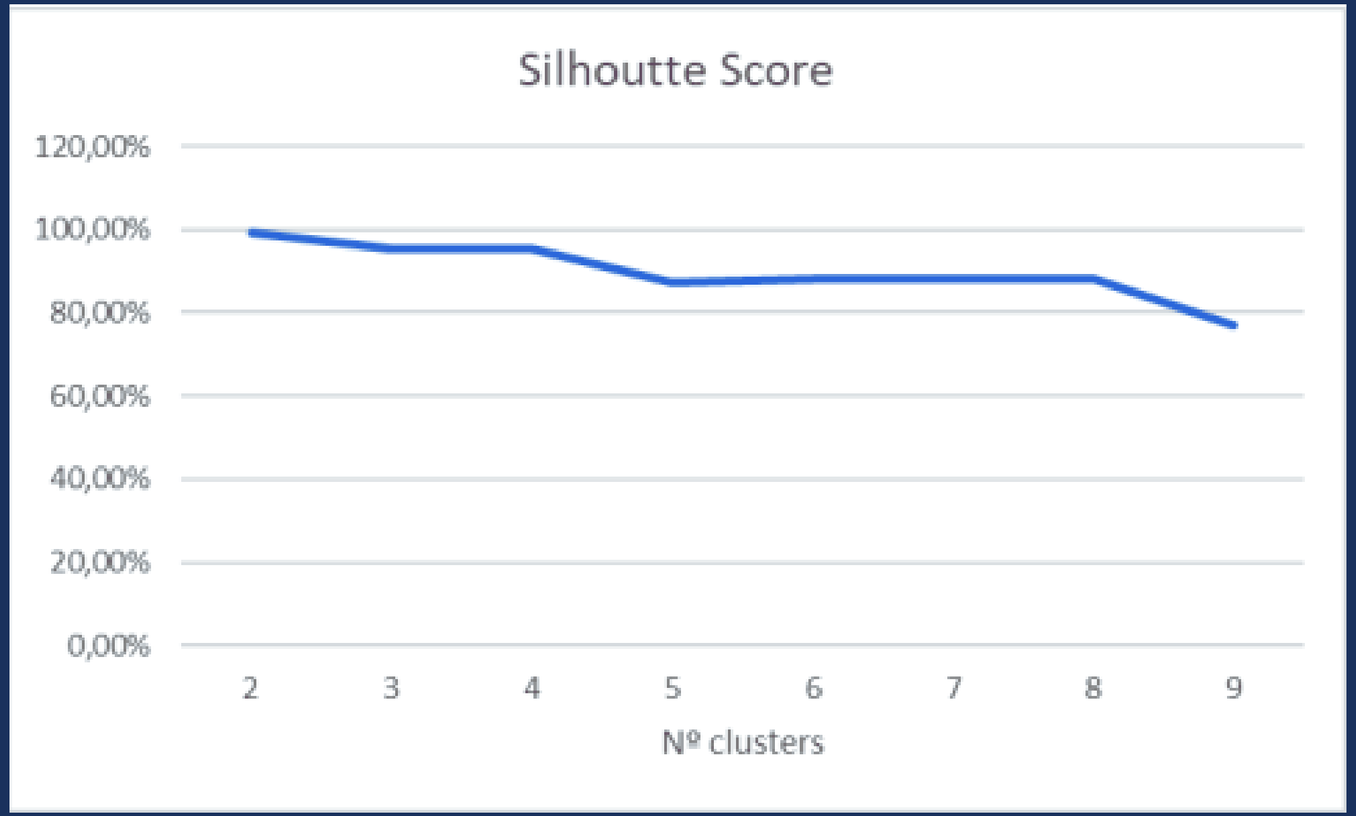
Group 1 has only 2 sellers, with completely distinct: one has 3 orders, the other 396; one sold 4 products, the another sold 430; one made 945 and the other 18,470. Group 5 has only 1 seller, and Group 4 has only 6 sellers.
We concluded that it does not make sense from a business point of view, as any decision to be made to affect less than 1% of base sellers, even if it brings benefits, it will be very costly considering hours spent on analysis and implementation

K=2	K=3	K=6
Silhoutte score: 95% Group 0: 3.023 vendors Group 1: 72 vendors	Silhoutte score: 47% Group 0: 2,999 vendors Group 1: 33 vendors Group 2: 72 vendors	Silhoutte score: 33% Group 0: 2,732 vendors Group 1: 2 vendors Group 2: 23 vendors Group 3: 331 vendors Group 4: 6 vendors Group 5: 1 vendors

STATISTICAL MODELING

RESULTS | SPARK ML BISECTING K-MEAN

Spark ML Bisecting K-means
We arrive at the possibility of 3 different groupings: models with 2, 3 or 4 groups.
It is possible to see that the variables quantity of orders and value of sales have a great influence on customer segmentation, sorting them into sellers small, medium and large.



STATISTICAL MODELING

RESULTS | SPARK ML BISECTING K-MEANS

Spark ML Bisecting K-means

We arrive at the possibility of 3 different groupings: models with 2, 3 or 4 groups.

It is possible to see that the variables quantity of orders and value of sales have a great influence on customer segmentation, sorting them into sellers small, medium and large.

The result of biSec K-means with 2 clusters became similar to K-means native to Spark, but with performance bottom: there, the second group was left with 72 sellers. Therefore, let's discard the profile analysis of this technique

K=2	K=3	K=4
Silhoutte score: 99% Group 0: 3.076 vendors Group 1: 19 vendors	Silhoutte score: 95% Group 0: 2,957 vendors Group 1: 119 vendors Group 2: 19 vendors	Silhoutte score: 95% Group 0: 2,957 vendors Group 1: 119 vendors Group 2: 12 vendors Group 3: 17 vendors

STATISTICAL MODELING

RESULTS | SPARK ML BISECTING K-MEANS

Spark ML Bisecting K-means
We arrive at the possibility of 3 different groupings: models with 2, 3 or 4 groups.
It is possible to see that the variables quantity of orders and value of sales have a great influence on customer segmentation, sorting them into sellers small, medium and large.

The biSec K-means result with 3 clusters was as interesting as Spark's native K-means: there, the 1st group had 2990 sellers, the 2nd group got 33 and the 3rd got 72.
Let's carry out the profile analysis of this technique and compare with the profiles generated through Spark's native K-means with 3 clusters to define the best model.

K=2	K=3	K=4
Silhoutte score: 99% Group 0: 3.076 vendors Group 1: 19 vendors	Silhoutte score: 95% Group 0: 2,957 vendors Group 1: 119 vendors Group 2: 19 vendors	Silhoutte score: 95% Group 0: 2,957 vendors Group 1: 119 vendors Group 2: 12 vendors Group 3: 7 vendors

STATISTICAL MODELING

RESULTS | SPARK ML BISECTING K-MEANS

Spark ML Bisecting K-means

We arrive at the possibility of 3 different groupings: models with 2, 3 or 4 groups.

It is possible to see that the variables quantity of orders and value of sales have a great influence on customer segmentation, sorting them into sellers small, medium and large.

biSec K-means result with 4 clusters does not seem to add much in compared to the grouping with 3 clusters seen above. There was no difference in the number of salespeople in the 1st and 2nd groups, and the 3rd group was divided in 2, with little differentiation between them. We conclude that it makes no sense business point of view, as any decision to be made to affect less than 0.1% of base sellers, even if it brings benefit, will be very costly considering hours spent on analysis, solution design and Implementation. Therefore, we will exclude this cluster from the final analysis.

K=2	K=3	K=4
Silhoutte score: 99% Group 0: 3.076 vendors Group 1: 19 vendors	Silhoutte score: 95% Group 0: 2,957 vendors Group 1: 119 vendors Group 2: 19 vendors	Silhoutte score: 95% Group 0: 2,957 vendors Group 1: 119 vendors Group 2: 12 vendors Group 3: 7 vendors

STATISTICAL MODELING

RESULTS | SPARK ML GAUSSIAN MIXTURE MODEL

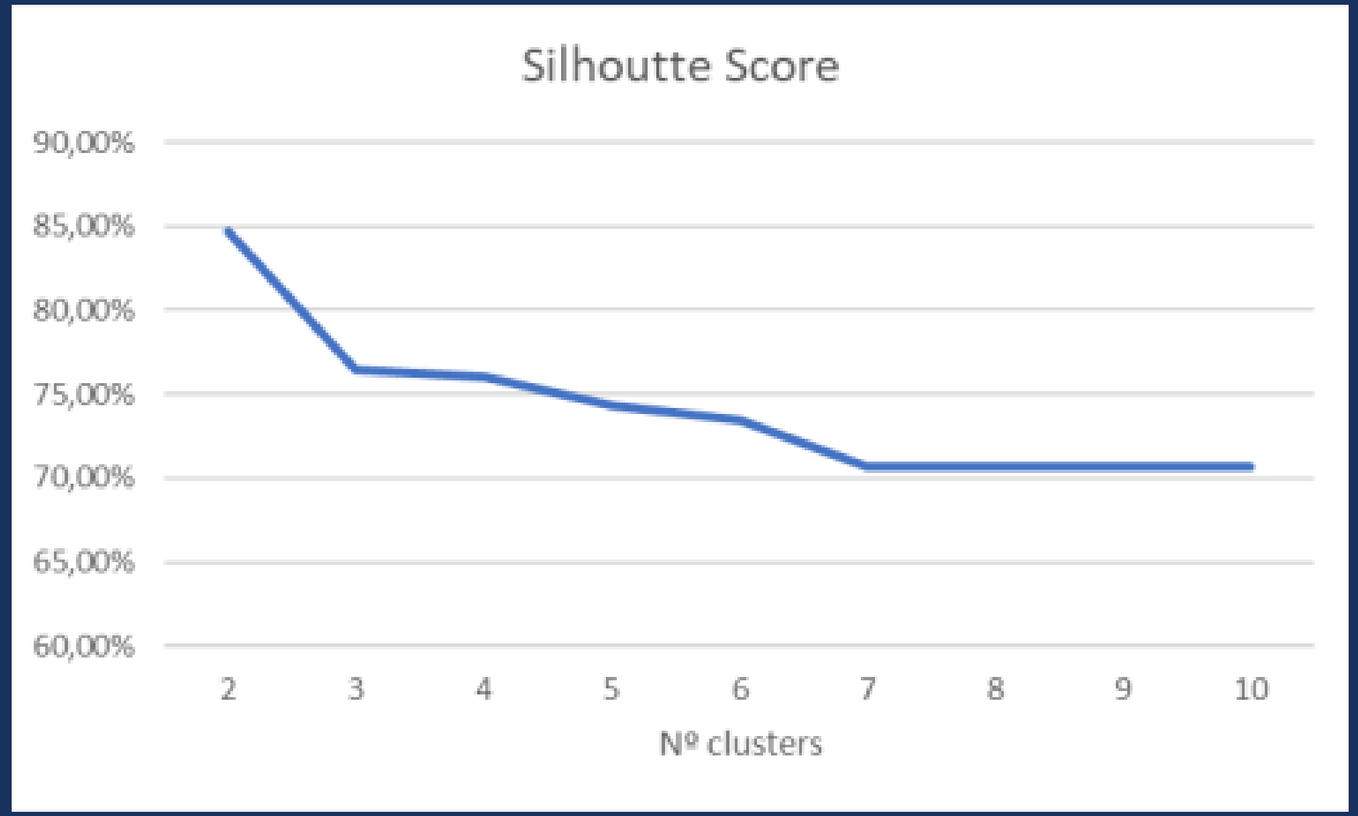
Spark ML Gaussian Mixture Model

We cannot understand the how the Gaussian Mixture works Model due to the lack of well documentation. We run with 2 and with 3 clusters and the result was same: prediction = 0 for all Appetizer. Therefore, we chose address the problem using the Pandas and Scikit Learn, where we will apply 4 techniques unsupervised clustering: Agglomeration, DBSCAN, MeanShift and K-means.

STATISTICAL MODELING

RESULTS | SCIKIT LEARN AGGLOMERATIVE

Scikit Learn Agglomerative
We arrive at the possibility of 3 different groupings: models with 2, 3 or 8 groups.
Unfortunately, despite the high silhoutte
score, the divisions didn't make sense from a business point of view and we decided to discard this technique



STATISTICAL MODELING

RESULTS | SCIKIT LEARN AGGLOMERATIVE

Scikit Learn Agglomerative

We arrive at the possibility of 3 different groupings: models with 2, 3 or 8 groups.

Unfortunately, despite the high silhoutte score, the divisions didn't make sense from a business point of view and we decided to discard this technique

The division by agglomeration with 2 clusters does not make sense from the business point of view, as any decision to be taken to affect only 1 seller in the base, even if bring benefit, it will be very costly considering hours spent on analysis, solution design and implementation. As we already have good results with division by K-means native to Spark, let's discard this division.

K=2	K=3	K=8
Silhouette score: 84% Group 0: 3094 vendors Group 1: 1 salesperson	Silhoutte score: 76% Group 0: 3089 vendors Group 1: 1 salesperson Group 2: 5 vendors	Silhoutte score: 70% Group 0: 5 vendors Group 1: 2 vendors Group 2: 3080 vendors Group 3: 1 salesperson Group 4: 1 salesperson Group 5: 2 vendors Group 6: 3 vendors Group 7: 1 salesperson

STATISTICAL MODELING

RESULTS | SCIKIT LEARN AGGLOMERATIVE

Scikit Learn Agglomerative

We arrive at the possibility of 3 different groupings: models with 2, 3 or 8 groups.

Unfortunately, despite the high silhoutte score, the divisions didn't make sense from a business point of view and we decided to discard this technique

Division by agglomeration with 3 clusters does not sense from a business point of view, as any decision to be taken to affect less than 0.5% of the sales base, even if it brings benefits, it will be very costly considering hours spent on analysis, solution design and implementation. As we already have good results with the native K-means division Spark, let's discard this division.

K=2	K=3	K=8
Silhouette score: 84% Group 0: 3094 vendors Group 1: 1 salesperson	Silhoutte score: 76% Group 0: 3089 vendors Group 1: 1 salesperson Group 2: 5 vendors	Silhoutte score: 70% Group 0: 5 vendors Group 1: 2 vendors Group 2: 3080 vendors Group 3: 1 salesperson Group 4: 1 salesperson Group 5: 2 vendors Group 6: 3 vendors Group 7: 1 salesperson

STATISTICAL MODELING

RESULTS | SCIKIT LEARN AGGLOMERATIVE

Scikit Learn Agglomerative

We arrive at the possibility of 3 different groupings: models with 2, 3 or 8 groups.

Unfortunately, despite the high silhoutte score, the divisions didn't make sense from a business point of view and we decided to discard this technique

Division by agglomeration with 8 clusters does not make sense from the point of view of business, as any decision to be made to affect less than 0.5% of the seller base and with most groups having 5 or fewer sellers, Even if it brings benefits, it will be very costly considering the hours spent for analysis, solution design and implementation. As we already have good results with Spark's native K-means division, we will discard this division and, consequently, the agglomeration technique as a whole.

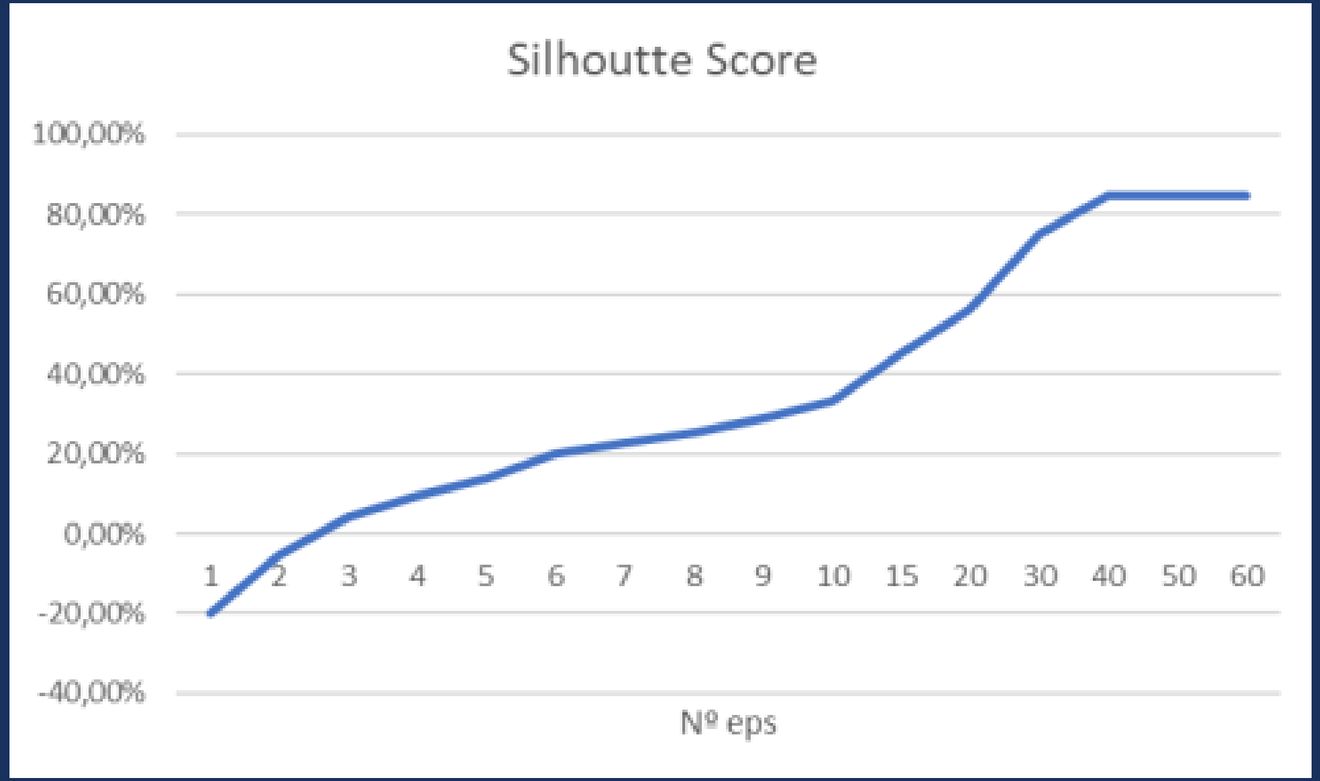
K=2	K=3	K=8
Silhouette score: 84% Group 0: 3094 vendors Group 1: 1 salesperson	Silhoutte score: 76% Group 0: 3089 vendors Group 1: 1 salesperson Group 2: 5 vendors	Silhoutte score: 70% Group 0: 5 vendors Group 1: 2 vendors Group 2: 3080 vendors Group 3: 1 salesperson Group 4: 1 salesperson Group 5: 2 vendors Group 6: 3 vendors Group 7: 1 salesperson

STATISTICAL MODELING

RESULTS | SCIKIT LEARN DBSCAN

Scikit Learn DBSCAN

We select DBSCAN with eps = 40, 20 and 18, equivalent to k = 2, 3 and 7. Unfortunately, despite the high silhoutte score, the divisions didn't make sense from a business point of view and we decided to discard this technique.



STATISTICAL MODELING

RESULTS | SCIKIT LEARN DBSCAN

Scikit Learn DBSCAN

We select DBSCAN with eps = 40, 20 and 18, equivalent to k = 2, 3 and 7. Unfortunately, despite the high silhoutte score, the divisions didn't make sense from a business point of view and we decided to discard this technique.

division by DBSCAN with 2 clusters does not make sense business point of view, as any decision to be taken to affect only 1 seller in the base, even if bring benefit, it will be very costly considering hours spent on analysis, solution design and implementation. As we already have good results with division by K-means native to Spark, let's discard this division.

K=2	K=3	K=7
Silhouette score: 84% Group 0: 3094 vendors Group 1: 1 salesperson	Silhoutte score: 56% Group 0: 3037 vendors Group 1: 51 salesperson Group 2: 7 vendors	Silhoutte score: 45% Group 0: 2099 vendors Group 1: 7 vendors Group 2: 6 vendors Group 3: 6 vendors Group 4: 8 vendors Group 5: 5 vendors Group -1: 64 vendors

STATISTICAL MODELING

RESULTS | SCIKIT LEARN DBSCAN

Scikit Learn DBSCAN

We select DBSCAN with eps = 40, 20 and 18, equivalent to k = 2, 3 and 7. Unfortunately, despite the high silhoutte score, the divisions didn't make sense from a business point of view and we decided to discard this technique.

The division by DBSCAN with 3 clusters does not make sense from the business point of view, as any decision to be taken to affect less than 2% of the base sellers, even if it brings benefit, it will be very costly considering hours spent for analysis, solution design and implementation. As we already have good results with the native K-means division Spark, let's discard this division.

K=2	K=3	K=7
Silhouette score: 84% Group 0: 3094 vendors Group 1: 1 salesperson	Silhoutte score: 56% Group 0: 3037 vendors Group 1: 51 salesperson Group 2: 7 vendors	Silhoutte score: 45% Group 0: 2099 vendors Group 1: 7 vendors Group 2: 6 vendors Group 3: 6 vendors Group 4: 8 vendors Group 5: 5 vendors Group -1: 64 vendors

STATISTICAL MODELING

RESULTS | SCIKIT LEARN DBSCAN

Division by DBSCAN with 7 clusters does not make sense from the point of view of
business, as any decision to be made to affect only 3% of the base of sellers and with most groups having 8 or fewer sellers,
Even if it brings benefits, it will be very costly considering the hours spent
for analysis, solution design and implementation. As we already have good
results with Spark's native K-means division, we will discard this division and, consequently, the DBSCAN technique as a whole.

Scikit Learn DBSCAN

We select DBSCAN with eps = 40, 20 and 18, equivalent to k = 2, 3 and 7. Unfortunately, despite the high silhoutte score, the divisions didn't make sense from a business point of view and we decided to discard this technique.

K=2	K=3	K=7
Silhouette score: 84% Group 0: 3094 vendors Group 1: 1 salesperson	Silhoutte score: 56% Group 0: 3037 vendors Group 1: 51 salesperson Group 2: 7 vendors	Silhoutte score: 45% Group 0: 2099 vendors Group 1: 7 vendors Group 2: 6 vendors Group 3: 6 vendors Group 4: 8 vendors Group 5: 5 vendors Group -1: 64 vendors

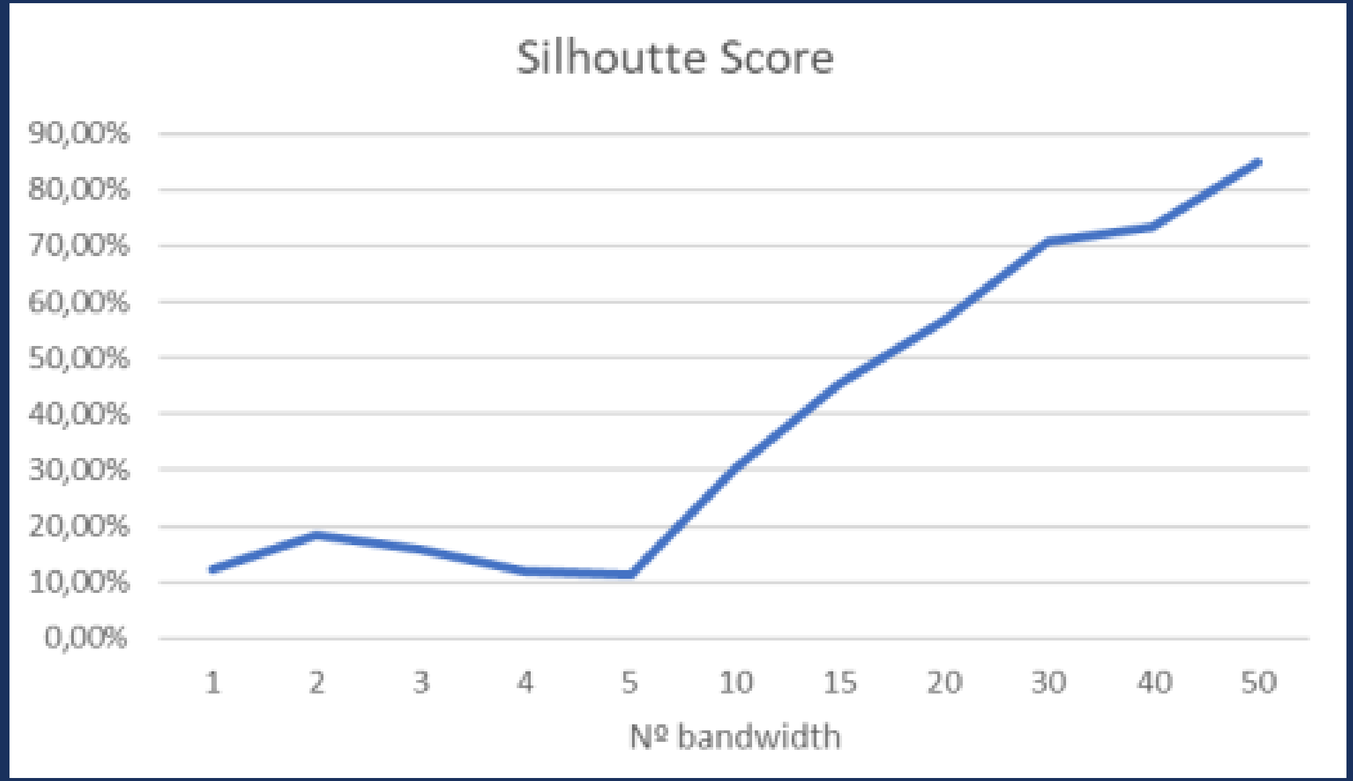
STATISTICAL MODELING

RESULTS | SCIKIT LEARN MEANSHIFT

Scikit Learn MeanShift

We select MeanShift with bandwidth = 50, 40 and 35, equivalent at k = 2, 4 and 8.

Unfortunately, despite the high silhoutte score, the divisions didn't make sense from a business point of view and we decided to discard this technique.



STATISTICAL MODELING

RESULTS | SCIKIT LEARN MEANSHIFT

Scikit Learn MeanShift

We select MeanShift with bandwidth = 50, 40 and 35, equivalent at k = 2, 4 and 8.

Unfortunately, despite the high silhoutte score, the divisions didn't make sense from a business point of view and we decided to discard this technique.

Dividing by MeanShift with 2 clusters does not make sense from the business point of view, as any decision to be taken to affect only 1 seller in the base, even if bring benefit, it will be very costly considering hours spent on analysis, solution design and implementation. As we already have good results with division by K-means native to Spark, let's discard this division..

K=2	K=4	K=8
Silhouette score: 84% Group 0: 3094 vendors Group 1: 1 salesperson	Silhoutte score: 73% Group 0: 3090 vendors Group 1: 2 vendors Group 2: 2 vendors Group 3: 1 salesperson	Silhouette score: 70% Group 0: 3083 vendors Group 1: 2 vendors Group 2: 2 vendors Group 3: 2 vendors Group 4: 3 vendors Group 5: 1 salesperson Group 6: 1 salesperson Group 7: 1 salesperson

STATISTICAL MODELING

RESULTS | SCIKIT LEARN MEANSHIFT

Scikit Learn MeanShift

We select MeanShift with bandwidth = 50, 40 and 35, equivalent at k = 2, 4 and 8.

Unfortunately, despite the high silhoutte score, the divisions didn't make sense from a business point of view and we decided to discard this technique.

Division by MeanShift with 4 clusters does not make sense from a business point of view, as any decision to be taken to affect only 0.2% of the base sellers, even if it brings benefit, it will be very costly considering hours spent for analysis, solution design and implementation. As we already have good results with the native K-means division Spark, let's discard this division.

K=2	K=4	K=8
Silhouette score: 84% Group 0: 3094 vendors Group 1: 1 salesperson	Silhoutte score: 73% Group 0: 3090 vendors Group 1: 2 vendors Group 2: 2 vendors Group 3: 1 salesperson	Silhouette score: 70% Group 0: 3083 vendors Group 1: 2 vendors Group 2: 2 vendors Group 3: 2 vendors Group 4: 3 vendors Group 5: 1 salesperson Group 6: 1 salesperson Group 7: 1 salesperson

STATISTICAL MODELING

RESULTS | SCIKIT LEARN MEANSHIFT

Scikit Learn MeanShift

We select MeanShift with bandwidth = 50, 40 and 35, equivalent at k = 2, 4 and 8.

Unfortunately, despite the high silhoutte score, the divisions didn't make sense from a business point of view and we decided to discard this technique.

The division by MeanShift with 8 clusters does not make sense from the point of view of business, as any decision to be made to affect only 0.4% of the base of sellers and with most groups having 3 or fewer sellers, Even if it brings benefits, it will be very costly considering the hours spent for analysis, solution design and implementation. As we already have good results with Spark's native K-means division, we will discard this division and, consequently, the MeanShift technique as a whole.

K=2	K=4	K=8
Silhouette score: 84% Group 0: 3094 vendors Group 1: 1 salesperson	Silhoutte score: 73% Group 0: 3090 vendors Group 1: 2 vendors Group 2: 2 vendors Group 3: 1 salesperson	Silhouette score: 70% Group 0: 3083 vendors Group 1: 2 vendors Group 2: 2 vendors Group 3: 2 vendors Group 4: 3 vendors Group 5: 1 salesperson Group 6: 1 salesperson Group 7: 1 salesperson

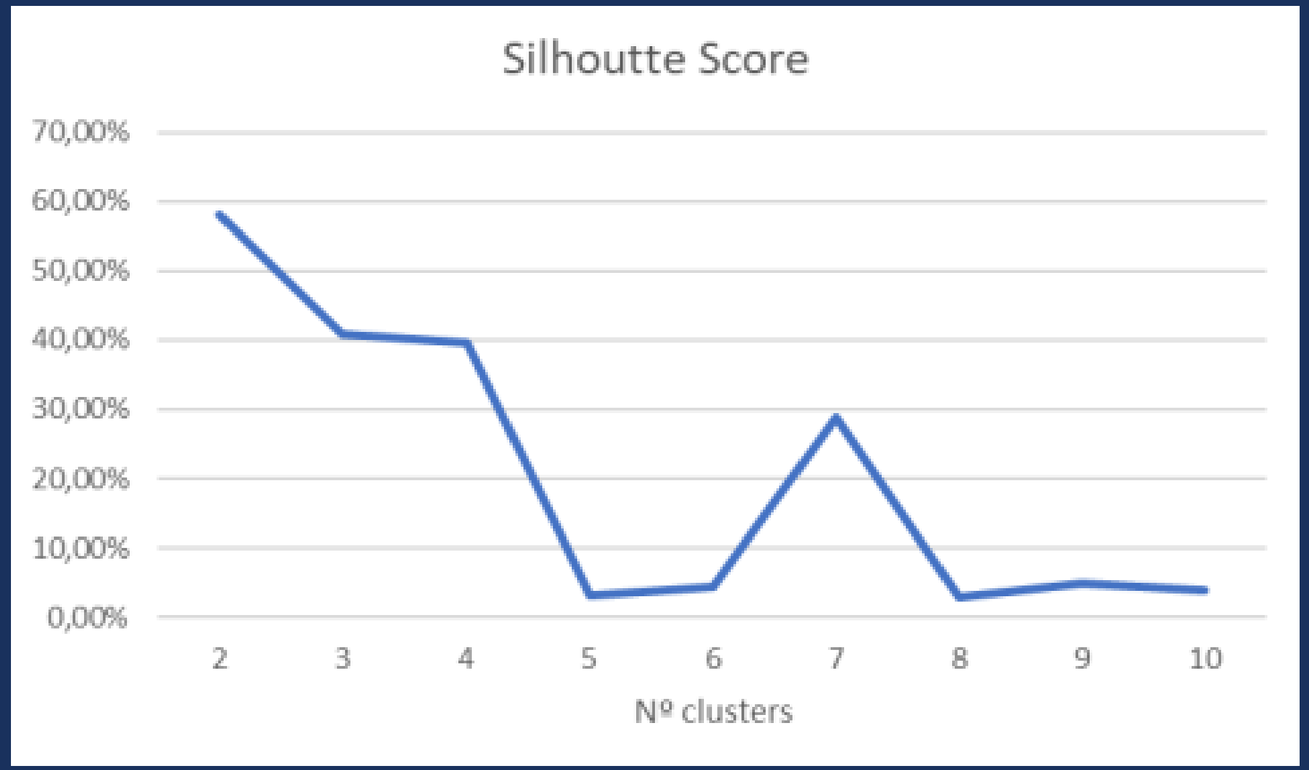
STATISTICAL MODELING

RESULTS | SCIKIT LEARN K-MEANS

Scikit Learn K-means

We selected models with 2, 3 and 4 clusters, as they have better silhouette score.

Of these, the grouping with 3 clusters turned out to be quite promising.



STATISTICAL MODELING

RESULTS | SCIKIT LEARN K-MEANS

Scikit Learn K-means

We selected models with 2, 3 and 4 clusters, as they have better silhouette score.

Of these, the grouping with 3 clusters turned out to be quite promising.

The K-means result with 2 Scikit Learn clusters was very similar to Spark's native K-means, where the second group of premium sellers got 72 sellers. Therefore, we will discard the analysis of profile of this technique.

K=2	K=3	K=4
Silhouette score: 58% Group 0: 3034 vendors Group 1: 61 vendors	Silhoutte score: 40% Group 0: 2814 vendors Group 1: 263 vendors Group 2: 18 vendors	Silhouette score: 39% Group 0: 2806 vendors Group 1: 18 vendors Group 2: 261 vendors Group 3: 10 vendors

STATISTICAL MODELING

RESULTS | SCIKIT LEARN K-MEANS

Scikit Learn K-means

We selected models with 2, 3 and 4 clusters, as they have better silhouette score.

Of these, the grouping with 3 clusters turned out to be quite promising.

The K-means result with 3 Scikit Learn clusters It was apparently better than K-means native to Spark, where the 1st group had 2999 sellers, the second group with 33 and the third with 72. It's worth looking into this grouping and generating profiles in the final analysis.

K=2	K=3	K=4
Silhouette score: 58% Group 0: 3034 vendors Group 1: 61 vendors	Silhoutte score: 40% Group 0: 2814 vendors Group 1: 263 vendors Group 2: 18 vendors	Silhouette score: 39% Group 0: 2806 vendors Group 1: 18 vendors Group 2: 261 vendors Group 3: 10 vendors

STATISTICAL MODELING

RESULTS | SCIKIT LEARN K-MEANS

K-means result with 4 Scikit Learn clusters does not seem to aggregate a lot compared to the 3-cluster grouping seen above. The 3rd group continues with 18 sellers and the 4th group was formed with just 10 sellers, taken from the 1st and 2nd groups. We concluded that it doesn't make sense from a business point of view, as any decision to be made to affect less than 0.3% of the sales base, even if it brings benefits, will be very costly considering hours spent on analysis, solution design and Implementation. Therefore, we will exclude this cluster from the final analysis.

Scikit Learn K-means

We selected models with 2, 3 and 4 clusters, as they have better silhouette score.

Of these, the grouping with 3 clusters turned out to be quite promising.

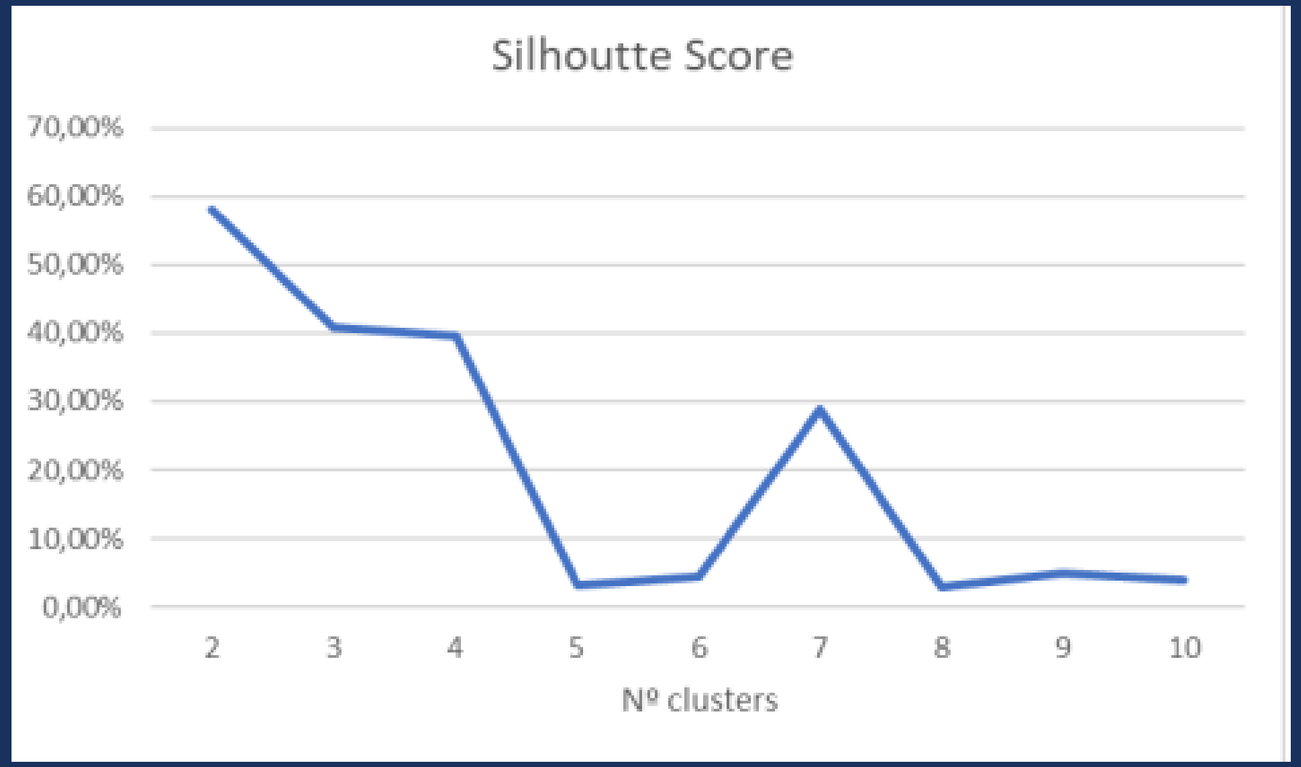
K=2	K=3	K=4
Silhouette score: 58% Group 0: 3034 vendors Group 1: 61 vendors	Silhoutte score: 40% Group 0: 2814 vendors Group 1: 263 vendors Group 2: 18 vendors	Silhouette score: 39% Group 0: 2806 vendors Group 1: 18 vendors Group 2: 261 vendors Group 3: 10 vendors

STATISTICAL MODELING

RESULTS | SCIKIT LEARN K-MEANS

Scikit Learn K-means

Still using this technique, we decided to test
the model with 6 clusters (even with very low silhouette score), to
be able to compare
it with the result of
K Spark's native -means, and the model with 7 clusters, which has silhoutte reasonable score.
Interestingly, the grouping with 6 clusters, despite having metrics
low stat, looks interesting
from a business perspective.



STATISTICAL MODELING

RESULTS | SCIKIT LEARN K-MEANS

Scikit Learn K-means

Still using this technique, we decided to test
the model with 6 clusters (even with very low silhouette score), to
be able to compare
it with the result of
K Spark's native -means, and the model with 7 clusters, which has silhoutte reasonable score.
Interestingly, the grouping with 6 clusters, despite having metrics
low stat, looks interesting
from a business perspective.

The K-means result with 2 Scikit Learn clusters was very similar to Spark's native K-means, where the second group of premium sellers got 72 sellers. Therefore, we will discard the analysis of profile of this technique.

K=6

Silhouette score: 4%
Group 0: 194 vendors
Group 1: 542 vendors
Group 2: 156 vendors
Group 3: 14 vendors
Group 4: 2,170 vendors
Group 5: 19 vendors

K=7

Silhoutte score: 28%
Group 0: 34 vendors
Group 1: 270 vendors
Group 2: 1 salesperson
Group 3: 35 vendors
Group 4: 19 vendors
Group 5: 55 vendors
Group 6: 2,681 vendors

STATISTICAL MODELING

RESULTS | SCIKIT LEARN K-MEANS

Scikit Learn K-means

Still using this technique, we decided to test
the model with 6 clusters (even with very low silhouette score), to
be able to compare
it with the result of
K Spark's native -means, and the model with 7 clusters, which has silhoutte reasonable score.
Interestingly, the grouping with 6 clusters, despite having metrics
low stat, looks interesting
from a business perspective.

Scikit Learn's K-means result with 7 clusters doesn't seem to add much in terms of
compared to the grouping with 6 clusters seen above. There are the 2 largest groups and the
another 5 groups correspond to less than 5% of the sales base, with 1 of the groups
with just 1 individual. We concluded that it doesn't make sense from a business perspective,
because any decision to be made to affect such small groups, even if it brings benefit, it will be very costly considering hours spent on analysis, solution design and implementation. Therefore, we will exclude this cluster from the final analysis.

K=6	K=7
Silhouette score: 4% Group 0: 194 vendors Group 1: 542 vendors Group 2: 156 vendors Group 3: 14 vendors Group 4: 2,170 vendors Group 5: 19 vendors	Silhoutte score: 28% Group 0: 34 vendors Group 1: 270 vendors Group 2: 1 salesperson Group 3: 35 vendors Group 4: 19 vendors Group 5: 55 vendors Group 6: 2,681 vendors

STATISTICAL MODELING

RESULTS | SELECTION OF MODELS

Among all the techniques and numbers of groups used, we chose those that appear to add the most value to the business.

KMeans MLSpark

- K = 3

Kmeans SKlearn

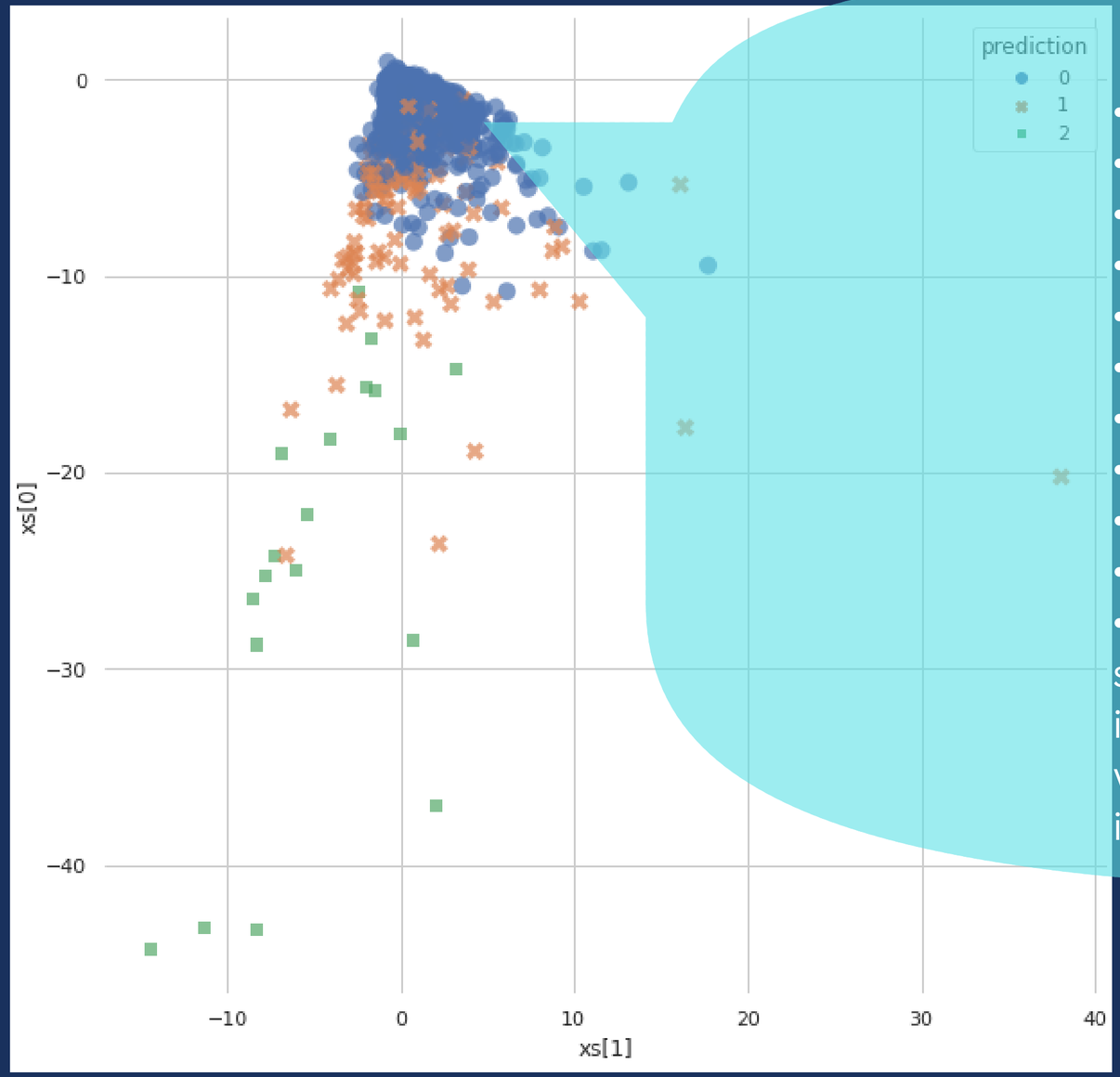
- K = 3
- K = 6

Bissecting Kmeans MLSpark

- K = 3

STATISTICAL MODELING

KMeans MLSpark K = 3 | GROUPS PROFILE

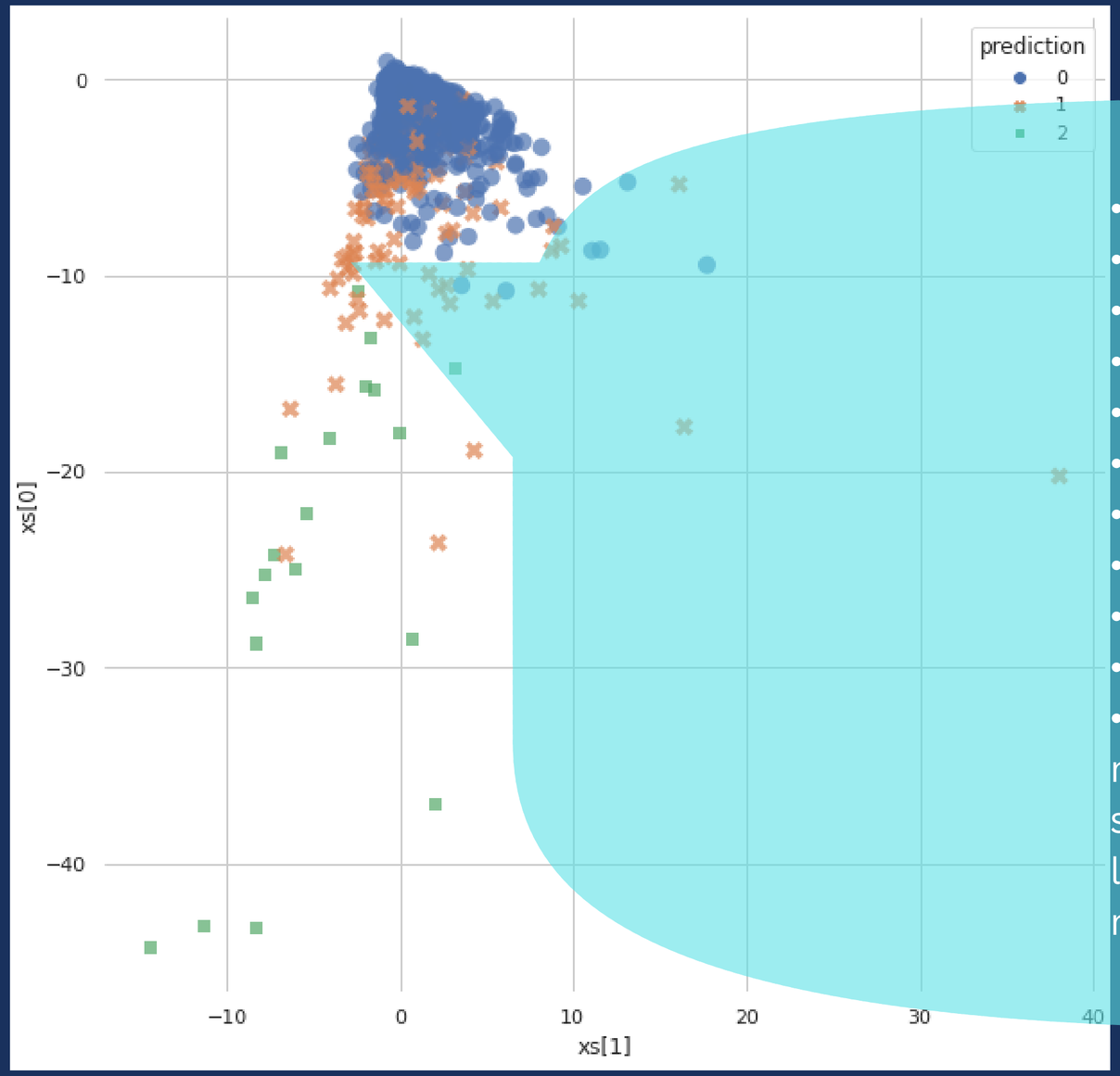


Prediction=0:

- Number of sellers: 2,990
- Average order: 20
- Average number of different buyers: 20
- Average number of products sold: 23
- Average revenue: just under R$3,500.00
- Average ticket: R$171.44
- Average price: R$151.03
- Recency: almost 5 months since the last sale
- Frequency: 3 sales every 2 months
- Delay in delivery: some sellers tend to be late
- This is the profile of most Olist salespeople: they make few sales and
it is only interesting for Olist in volume. Your products are of value
intermediate, compared to other groups.

STATISTICAL MODELING

KMeans MLSpark K = 3 | GROUPS PROFILE

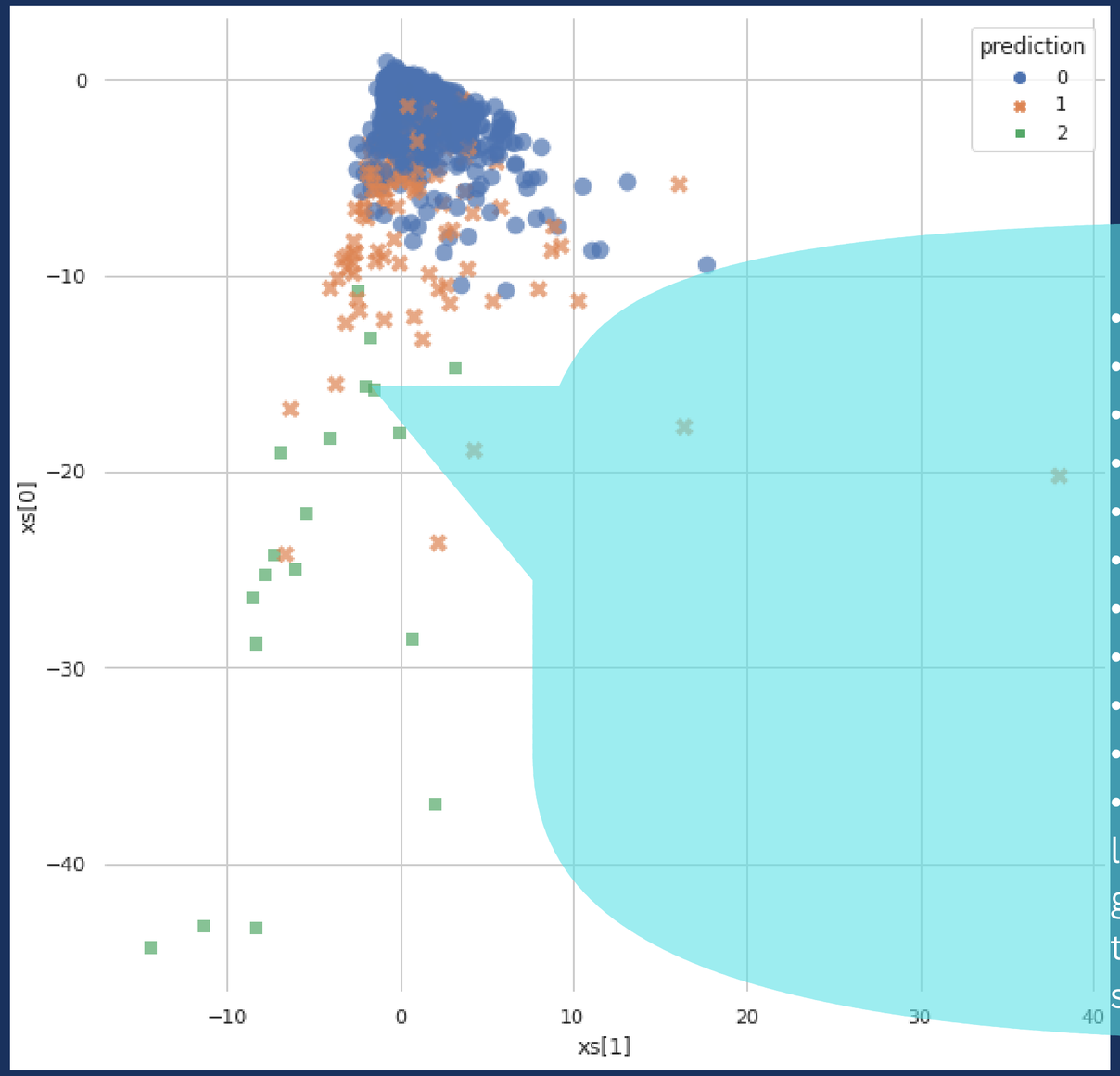


Prediction=1:

- Qty. Sellers: 33
- Average orders: 26
- Average number of different buyers: 25
- Average number of products sold: 31
- Average revenue: just over R$6,000.00
- Average ticket: R$228.41
- Average price: R$188.64
- Recency: just over 1 month since the last sale
- Frequency: almost 5 sales every 2 months
- Delay in delivery: no seller in this group is usually late
- This is the profile of aspiring premium Olist sellers: do more sales than most sellers, but still at a very high level. lower than premium sellers. To compensate, sell more expensive products.

STATISTICAL MODELING

KMeans MLSpark K = 3 | GROUPS PROFILE

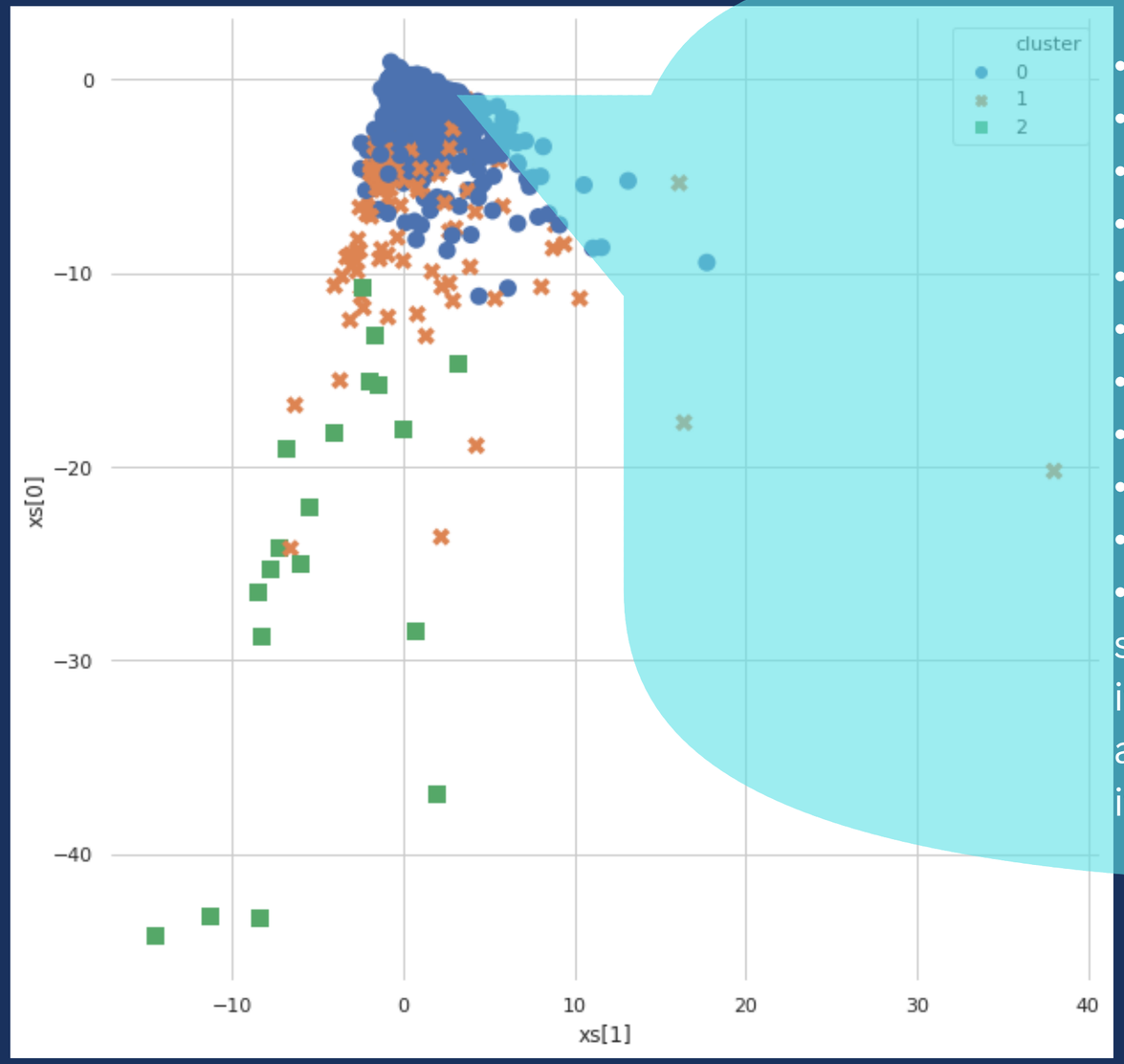


Prediction=2:

- Qty. Sellers: 72
- Average orders: 531
- Average number of different buyers: 525
- Average number of products sold: 599
- Average revenue: just over R$73,300
- Average ticket: R$138.05
- Average price: R$122.37
- Recency: just over 16 days since the last sale
- Frequency: almost 30 sales per month
- Delay in delivery: no seller in this group is usually late
- This is the profile of Olist's premium sellers: they make a lot of sales and guarantees good revenue, even selling products cheaper than sellers from other groups

STATISTICAL MODELING

Bissecting KMeans MLSpark K = 3| GROUPS PROFILE

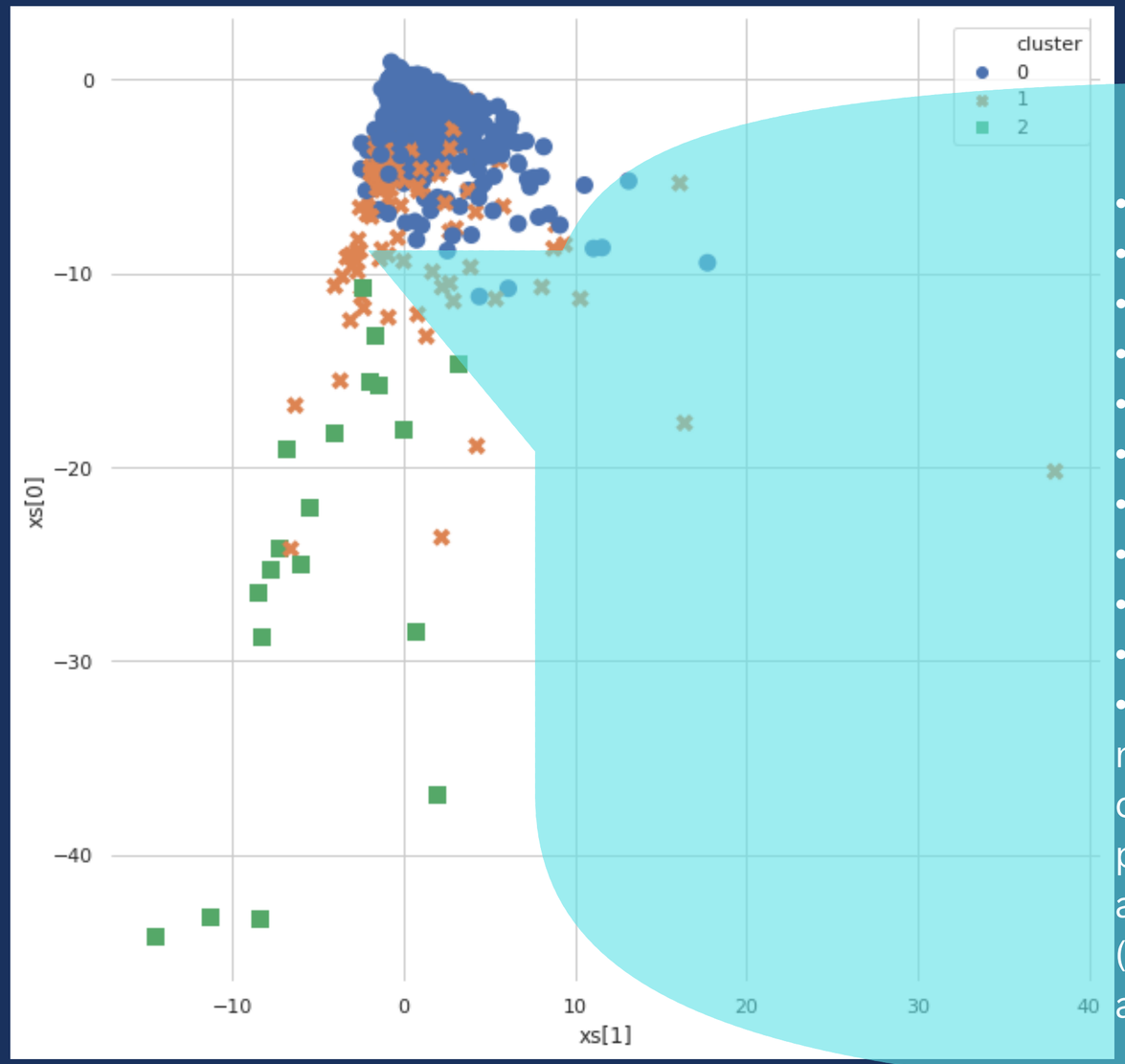


Prediction=0:

•Qty. Sellers: 2,957
• Average order: 18
• Average number of different buyers: 18
• Average number of products sold: 20
• Average revenue: just under R$2,700
• Average ticket: R$145.36
• Average price: R$127.90
• Recency: almost 5 months since the last sale
• Frequency: 3 sales every 2 months
• Delay in delivery: some sellers tend to be late
• This is the profile of most Olist salespeople: they make few sales,
its products have low added value in relation to other groups and
it is only interesting for Olist in volume.

STATISTICAL MODELING

Bissecting KMeans MLSpark K = 3| GROUPS PROFILE

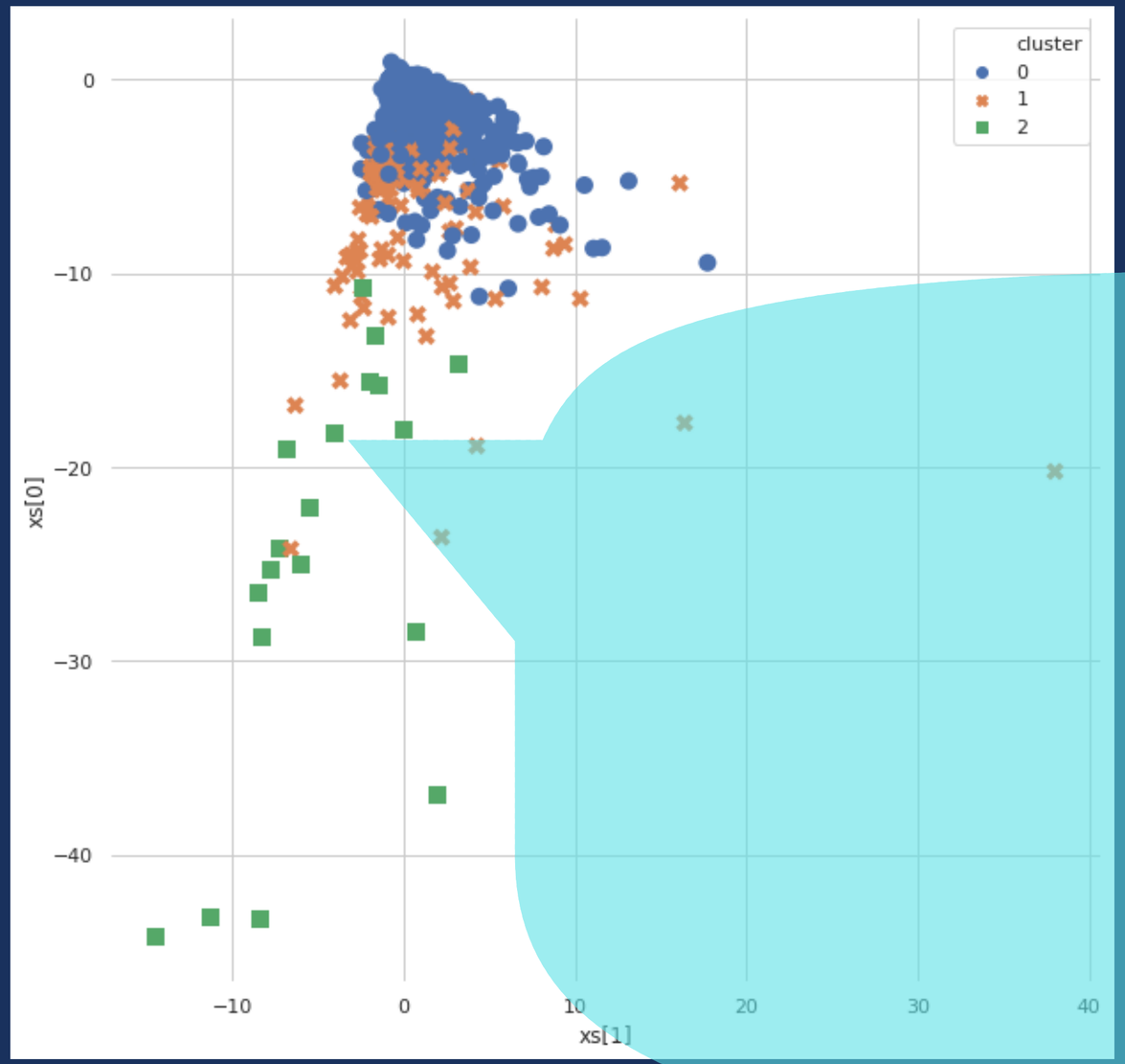


Prediction=1:

•Number of sellers: 119
• Average orders: 226
• Average number of different buyers: 223
• Average number of products sold: 252
• Average revenue: just over R$40,200
• Average ticket: R$387.74
• Average price: R$177.91
• Recency: just over 1 month since the last sale
• Frequency: almost 14 sales per month
• Delay in delivery: few sellers in this group tend to be late
• This is the profile of Olist's intermediary sellers: they have a number
considerable amount of more qualified sales, as it sells products of greater
added value (high average price) and more products per sale (higher
average ticket)

STATISTICAL MODELING

Bissecting KMeans MLSpark K = 3| GROUPS PROFILE

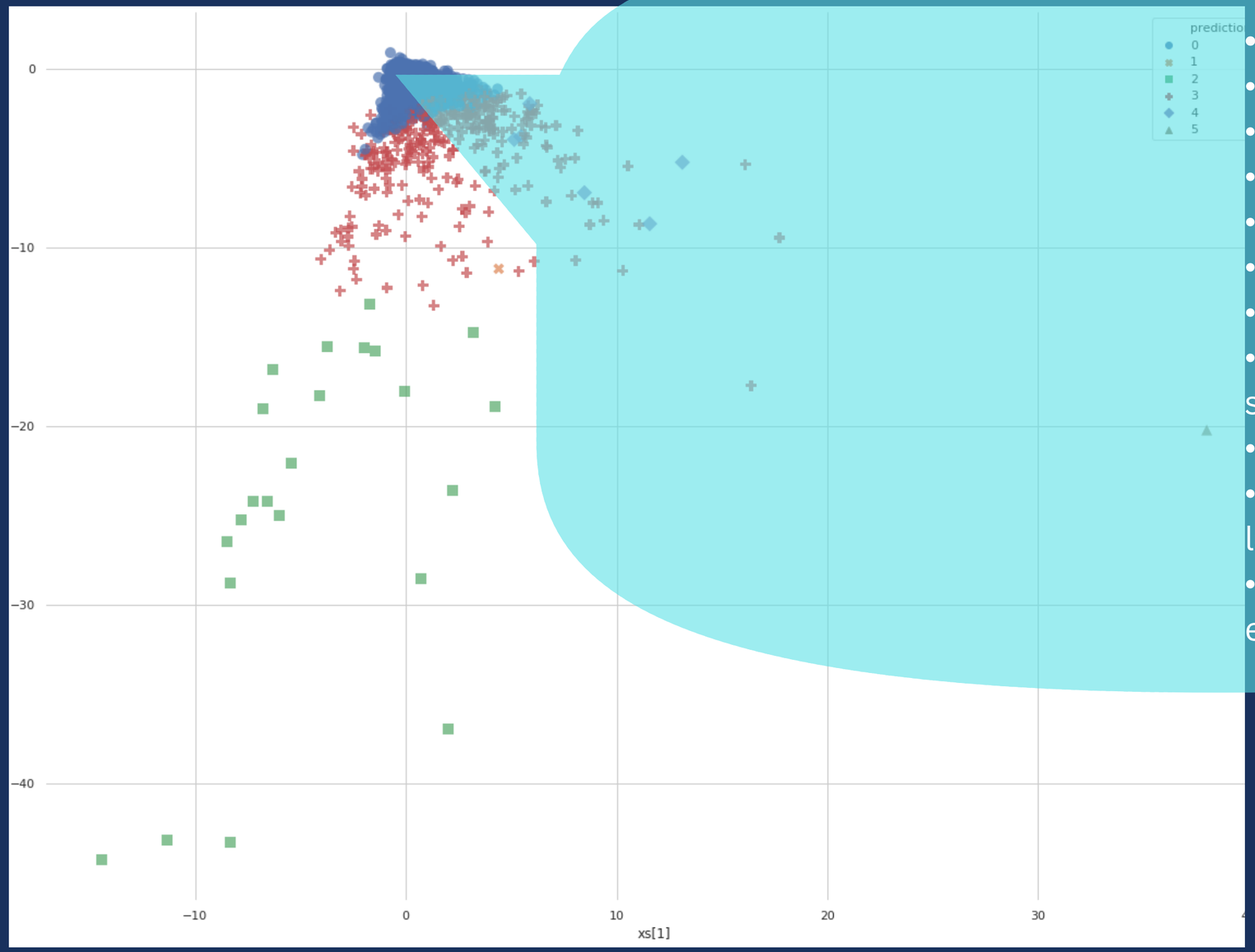


Prediction=2:

- Number of sellers: 19
- Average orders: 988
- Average number of different buyers: 976
- Average number of products sold: 1134
- Average revenue: just over R$170,200
- Average ticket: R$172.17
- Average price: R$150.07
- Recency: 15 days since the last sale
- Frequency: 52 sales per month
- Delay in delivery: no seller in this group is usually late
- This is the profile of Olist's elite salespeople: they make a lot of sales and guarantee good revenue. Your products are of intermediate value in relation to the other 2 groups.

STATISTICAL MODELING

Scikit Learn KMeans K = 6| GROUPS PROFILE

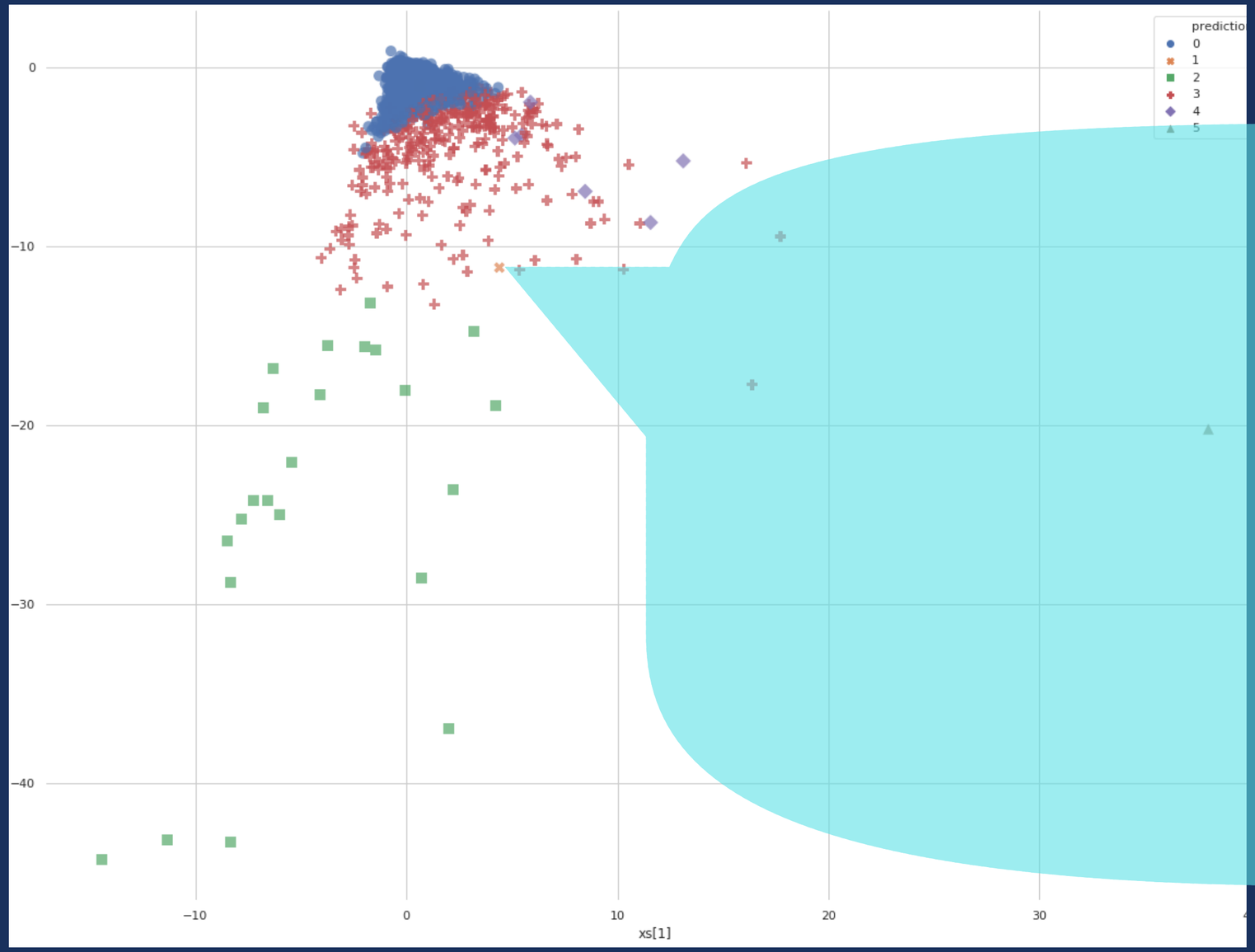


Prediction=0:

- **QTY**. Sellers: 194
- Average orders: 26
- Average number of different buyers: 26
- Average number of products sold: 29
- Average revenue: just over R$7,100
- Average ticket: R$271.18
- Average price: R$246.18
- Recency: more than 3 months since the last sale
- Frequency: 2 sales per month
- Delay in delivery: some sellers tend to be late
- This is the profile of sellers of more expensive products

STATISTICAL MODELING

Scikit Learn KMeans K = 6| GROUPS PROFILE

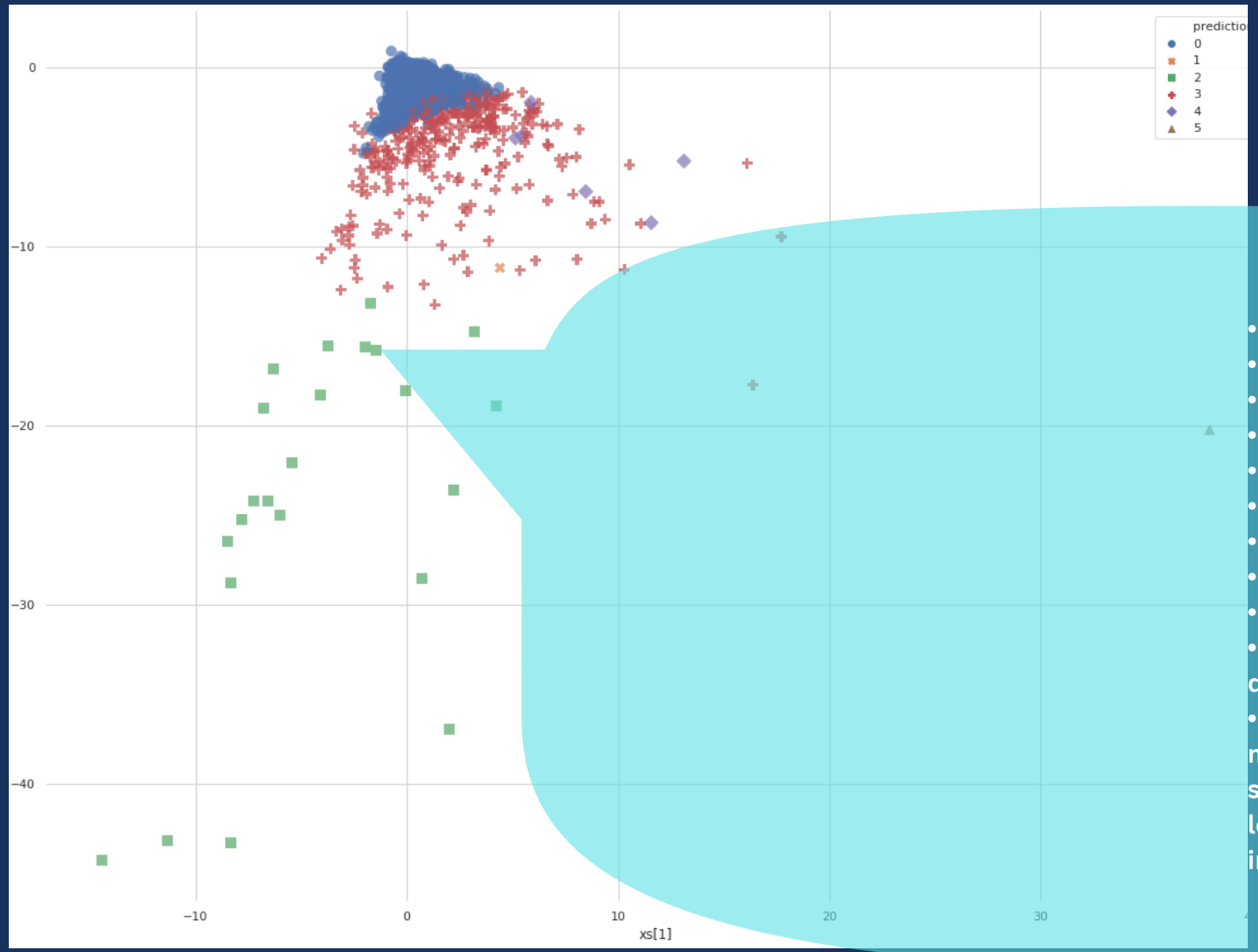


Prediction=1:

- Qty. Sellers: 542
- Average order: 5
- Average number of different buyers: 5
- Average number of products sold: 6
- Average revenue: just over R$1,100
- Average ticket: R$212.36
- Average price: R$175.57
- Recency: more than 11 months since the last sale
- Frequency: 1 sale every 3 months
- Delay in delivery: some sellers in this group tend to be late
- This is the profile of salespeople who didn't work out and gave up on the platform, with an average review score of 2.5.

STATISTICAL MODELING

Scikit Learn KMeans K = 6| GROUPS PROFILE



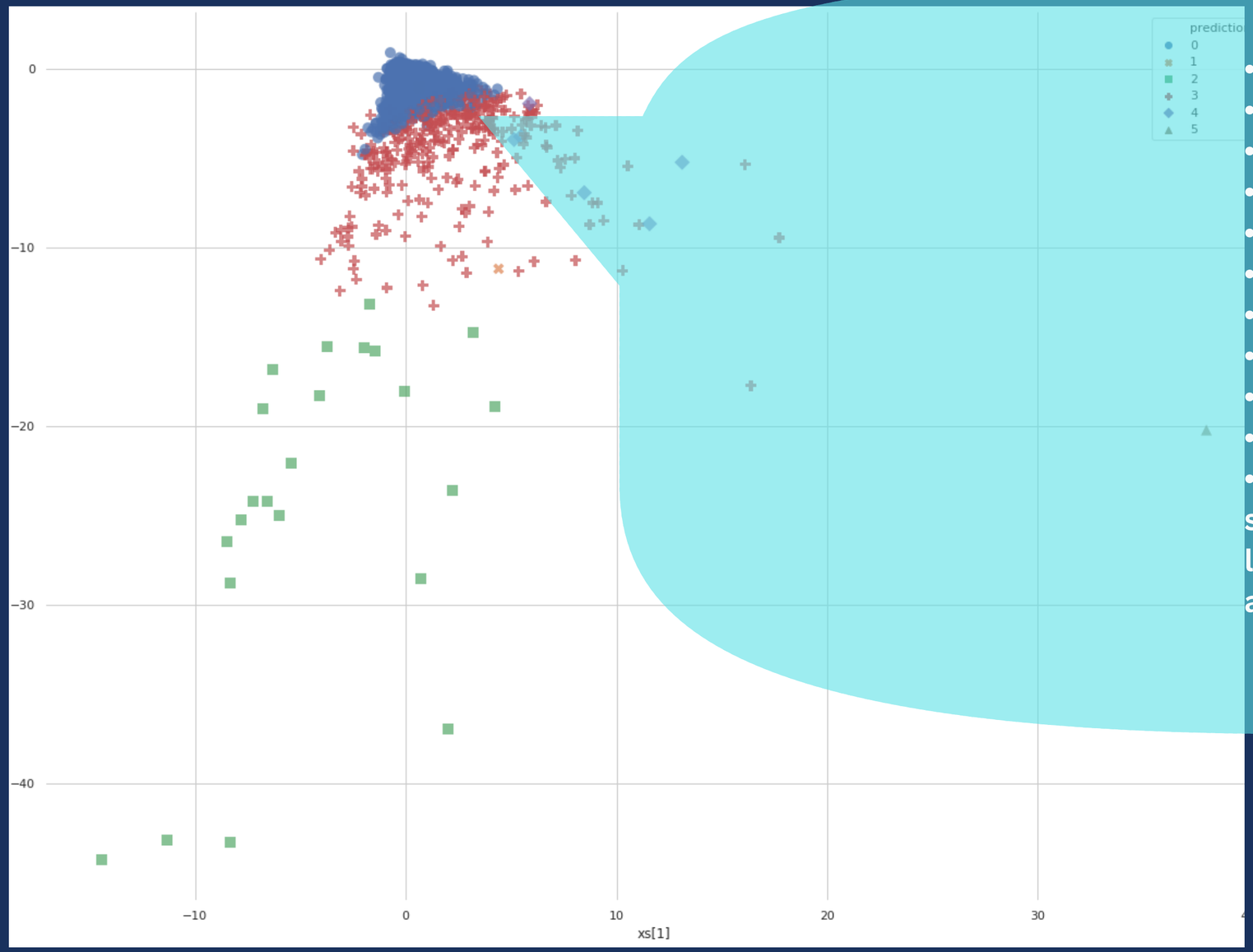
Prediction=2:

- Number of sellers: 156
- Average orders: 188
- Average number of different buyers: 186
- Average number of products sold: 208
- Average revenue: around R$26,600
- Average ticket: R$141.71
- Average price: R$127.78
- Recency: almost 1 and a half months since the last sale
- Frequency: 11 sales per month
- Delay in delivery: very few sellers in this group usually delay
- This is the profile of aspiring premium Olist sellers: do more

sales than most sellers, but still at a very high level. lower than premium sellers. Your products are of value intermediate in relation to other groups.

STATISTICAL MODELING

Scikit Learn KMeans K = 6| GROUPS PROFILE



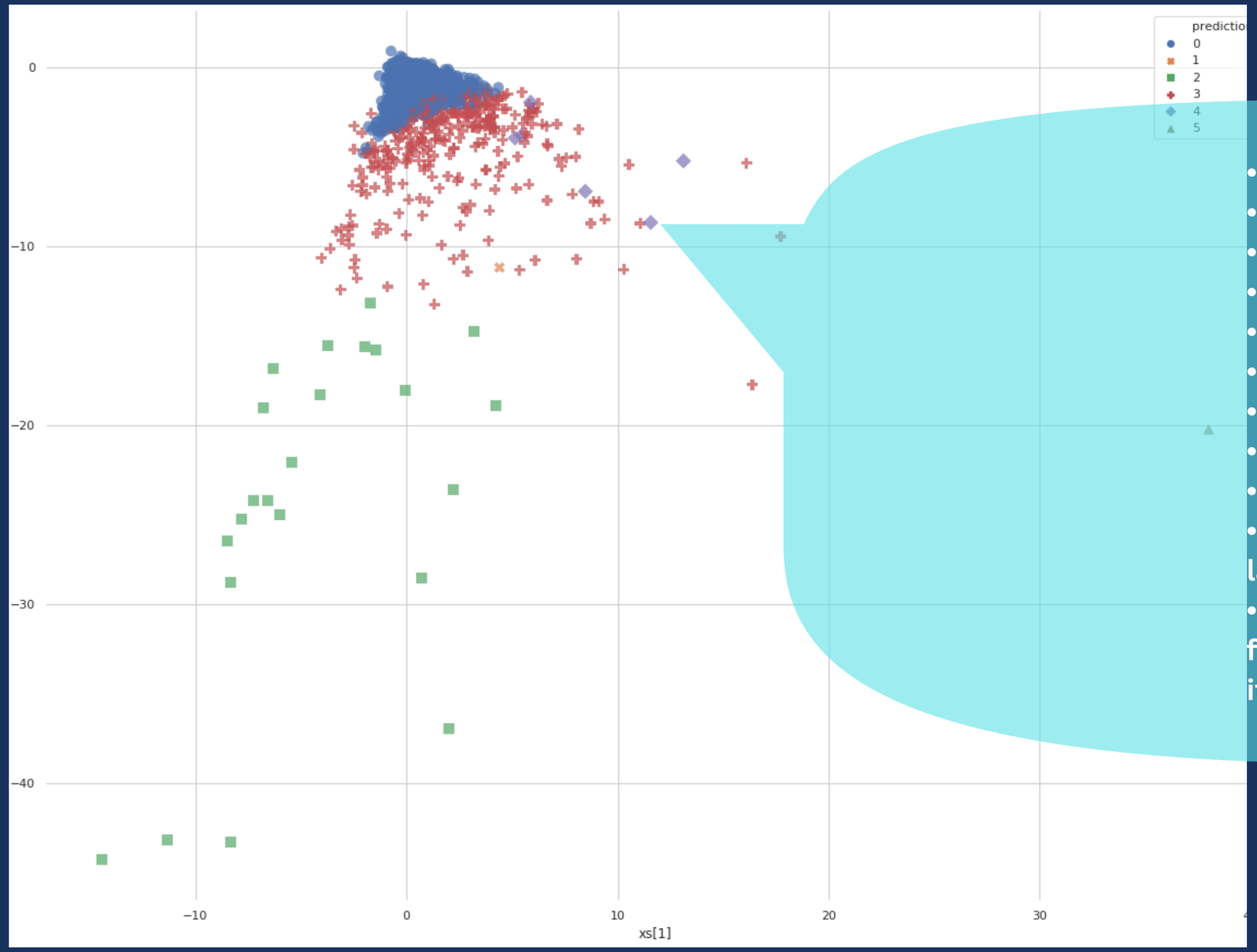
Prediction=3:

- Number of sellers: 14
- Average orders: 45
- Average number of different buyers: 45
- Average number of products sold: 50
- Average revenue: around R$4,500
- Average ticket: R$100.04
- Average price: R$90.26
- Recency: more than 1 month since the last sale
- Frequency: almost 5 sales per month
- Delay in delivery: no seller in this group is usually late
- This is the profile of a subgroup of lower clergy salespeople: number

low sales, works with cheaper products, but it pays off a little in volume.

STATISTICAL MODELING

Scikit Learn KMeans K = 6| GROUPS PROFILE



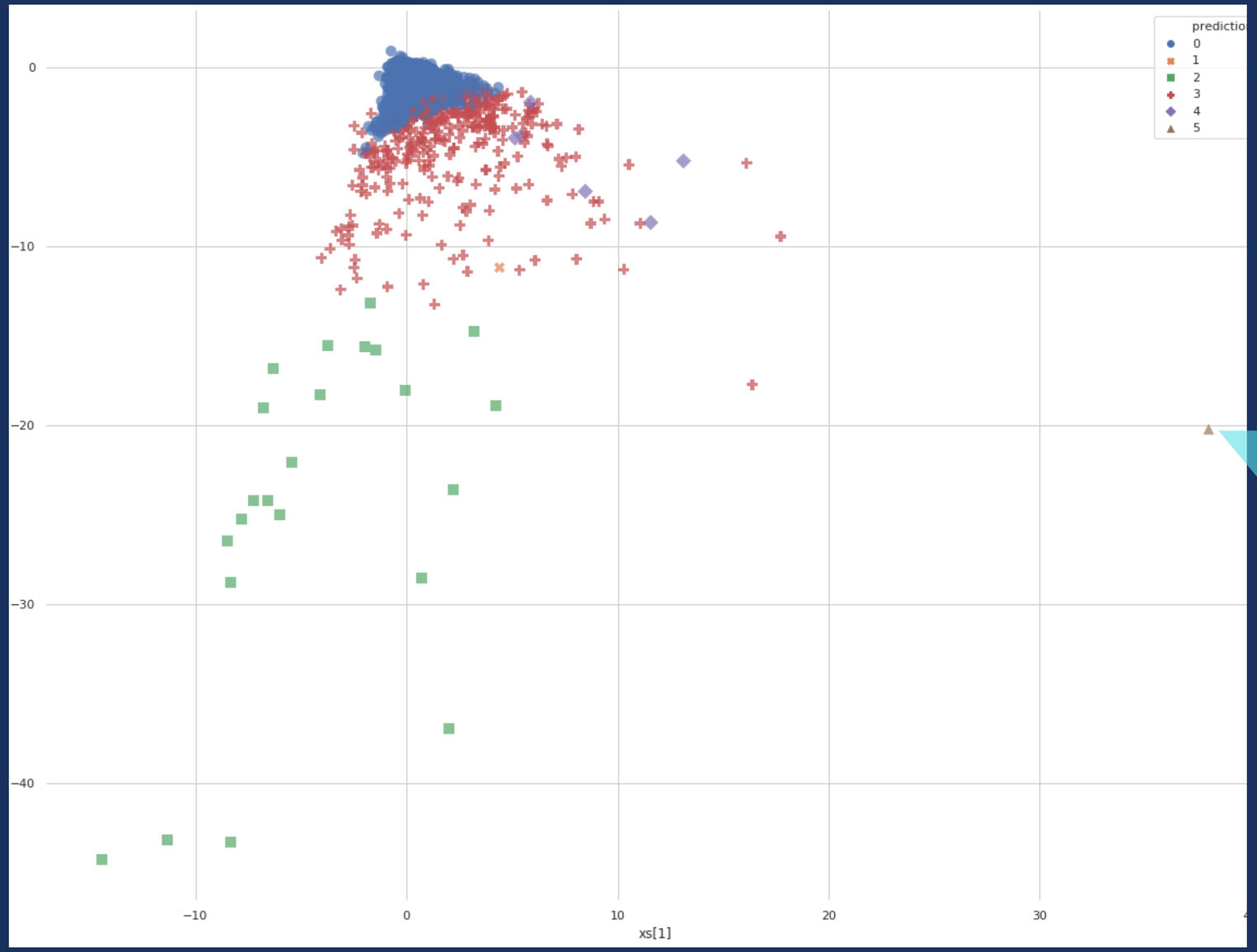
Prediction=4:

- **Number of sellers: 2,170**
- **Average orders: 19**
- **Average number of different buyers: 18**
- **Average number of products sold: 21**
- **Average revenue: almost R$3,200**
- **Average ticket: R$167.43**
- **Average price: R$146.40**
- **Recency: just over 3 months since the last sale**
- **Frequency: 5 sales every 3 months**
- **Delay in delivery: few sellers in this group tend to be late**
- **This is the profile of most Olist salespeople: they make few sales and**

it is only interesting for Olist in volume.

STATISTICAL MODELING

Scikit Learn KMeans K = 6| GROUPS PROFILE



Prediction=5:

- Qty. Sellers: 19
- Average orders: 1,090
- Average number of different buyers: 1,076
- Average number of products sold: 1,248
- Average revenue: around R$146,300
- Average ticket: R$134.22
- Average price: R$117.25
- Recency: almost 10 days since the last sale
- Frequency: almost 60 sales per month
- Delay in delivery: no seller in this group is usually late
- This is the profile of Olist's premium sellers: they make a lot of sales and guarantees good revenue, even selling valuable products lowest aggregate.

DATA ANALYSIS METHODOLOGY

Planning → Exploratory Analysis → Statistical Modelling → ML Algorithms → Conclusion

Problem definition

- Goals
- Concepts
- Criteria
- Data history
- Variables

Primary Analysis

- Position measurements
- Frequency analysis Graphics
- Outlier analysis
- Missing analysis
- Validation on the
- Consistency of information

Evaluation of techniques

- Native K-means using Spark

Evaluation of techniques

- biSecting K-means Spark native
- Gaussian Mixture Native model in Spark
- Scikit Learn: DBSCAN, MeanShift, K- means Clustering agorithms

Key Actionable Insights & takeaways

- Definition of the technique
- Validation of results
- Choice of technique what better if suitable for use and strategies

6. FINAL ASSESSMENT OF CLUSTERING TECHNIQUES

Comparing the grouping results, we have the following highlights:

Groups

•Groups with 3 clusters make a lot of sense from a business point of view

Metrics

Groups with 3 clusters have better silhoutte scores compared to Scikit Learn's k-means grouping of 6 clusters
- Spark native K-means with 3 clusters = 47.5%;
- K-means Scikit Learn with 3 clusters = 40.89%;
- Spark native biSec K-means with 3 clusters = 95.5%;
- K-means Scikit Learn with 6 clusters = 4.29%.

Model chosen

As the silhoutte score of the biSec K-means technique suggests, this was the best technique for separating the groups, which can be
evidenced when analyzing the distribution of the sales value: there are 3 very distinct ranges, without overlapping values (prediction= 0 varies from
R$12.22 to R$20,777.19; prediction=1 varies from R$21,591.68 to R$83,189.65; prediction=2 varies from R$117,409.50 to R$249,640.70). Not that
This happens in other forms of grouping, with bands overlapping.

6. FINAL ASSESSMENT OF CLUSTERING TECHNIQUES

Comparing the grouping results, we have the following highlights:

The separation into 6 groups using Scikit Learn's K-means, despite having the lowest silhoutte score among the selected techniques, brings 6 very interesting profiles from a business point of view:

- Sellers of more expensive products (194 sellers, 6.3% of the base);
- Sellers who did not work out and gave up on the platform (542, 17.5%);
- Aspiring Olist premium sellers: makes more sales than most sellers, but still at a level well below what

premium sellers (156,5%);

- Subgroup of lower clergy sellers: low number of sales, works with cheaper products, but compensates a little in volume(14, 0.5%);
- Majority of Olist sellers: make few sales and are only interesting to Olist in volume (2170, 70.1%);
- Olist premium sellers: makes a lot of sales and guarantees good revenue (19, 0.6%).

7. CONCLUSION

Recommendations for Olist:

1) Use the Bisecting K-means clustering technique native to Spark to classify sellers on an ABC curve enhanced by artificial intelligence, in order to define a commercial policy differentiated for each of the 3 groups of sellers, taking into account
counts the value that each group brings to the company

2) Use the K-means clustering technique with 6 Scikit clusters Learn to classify sellers by platform usage profile
Olist and, thus, create differentiated direct marketing actions to
each profile (e.g. rescue of abandoning sellers, bonus for aspiring premium sellers strive to reach the profile premium)