CUSTOMER SEGMENTATION

Project

olist

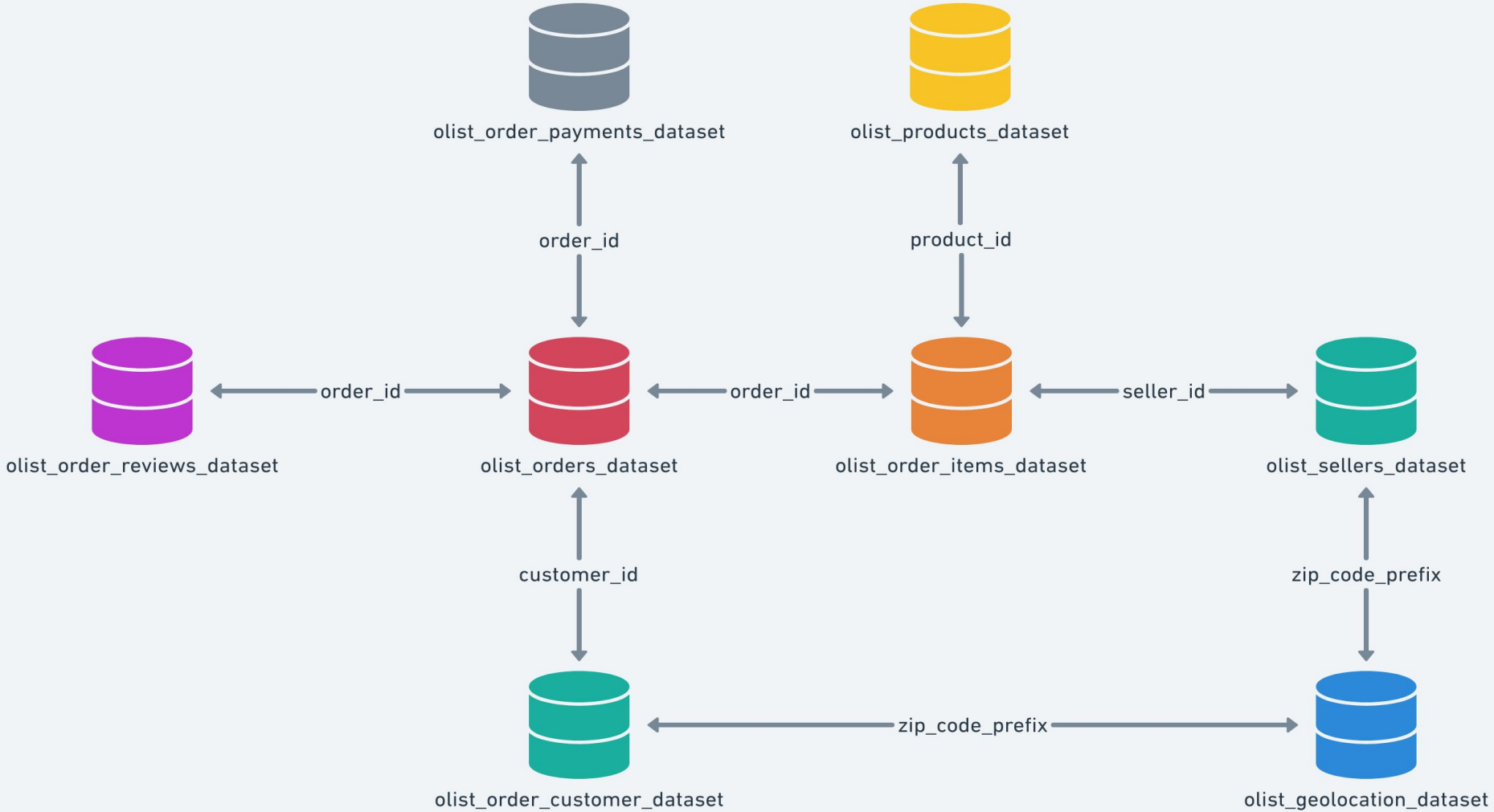
Provide a customer segmentation for the e-commerce teams to optimize marketing campaigns

MISSION

Determine customer profiles

Projection for the update frequency

DATA


olist_order_payments_dataset
olist_products_dataset
order_id
product_id
olist_order_reviews_dataset
order_id
olist_orders_dataset
order_id
olist_order_items_dataset
seller_id
olist_sellers_dataset
customer_id
zip_code_prefix
zip_code_prefix
olist_order_customer_dataset
olist_geolocation_dataset

Quick overview

- 8 tables
- 99441 orders
- 96096 unique customers
- sept 2016 - oct 2018 order dates range

Data processing

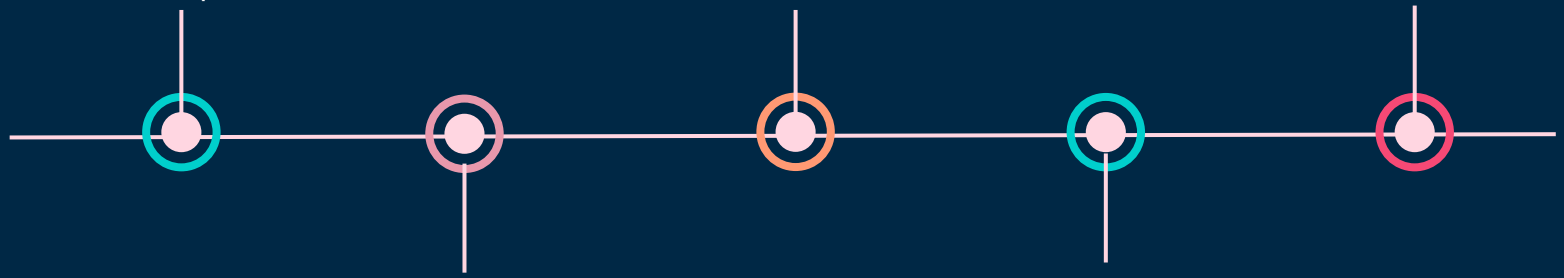
Merge datasets
(orders, payments, customers, items, reviews, products)

Fill missing values
(delivery date = estimated delivery date, review score = mean score)

Select most relevant features

Perform exploratory analysis

Create RFM dataset
(grouped by customer)

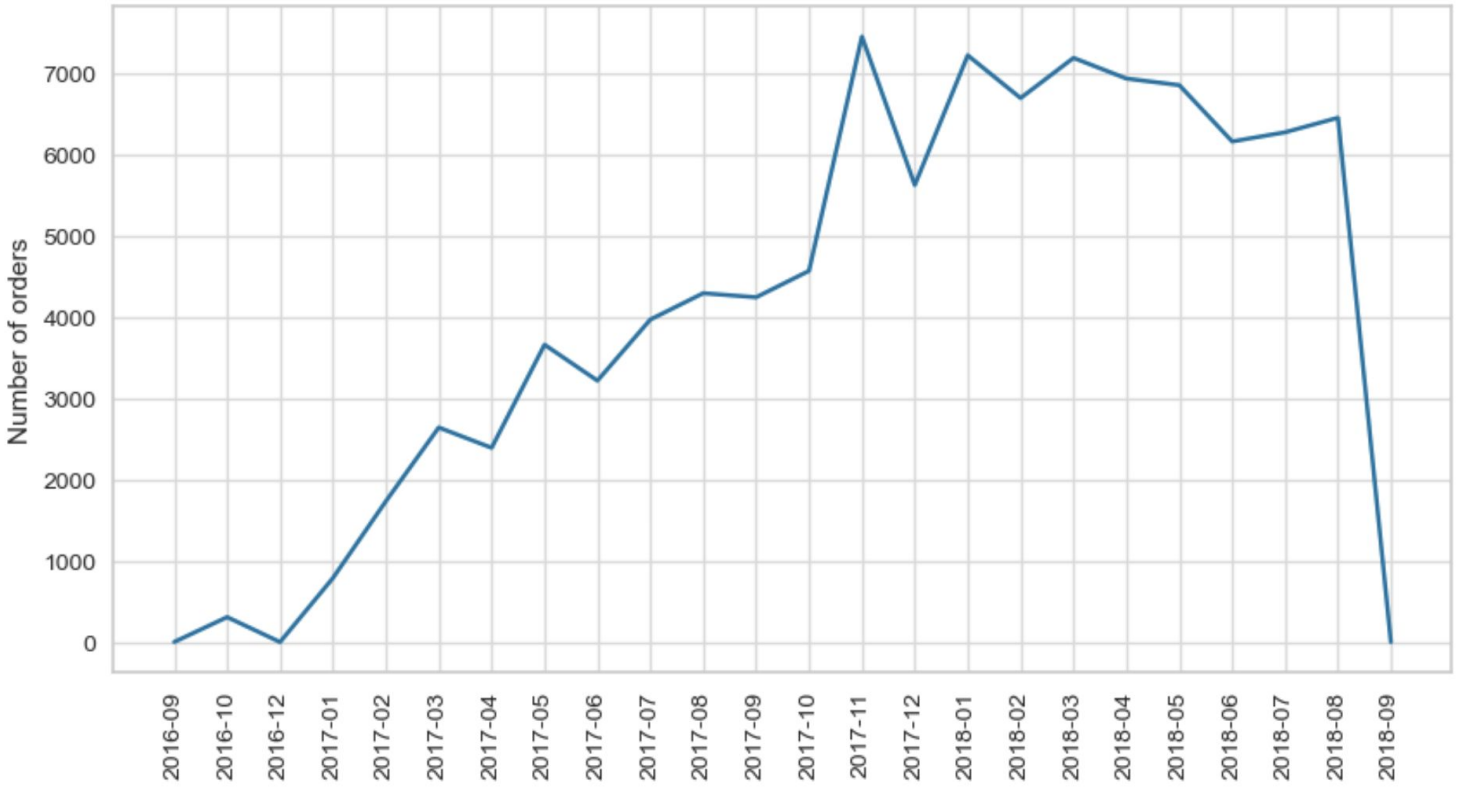


RFM feature transformation

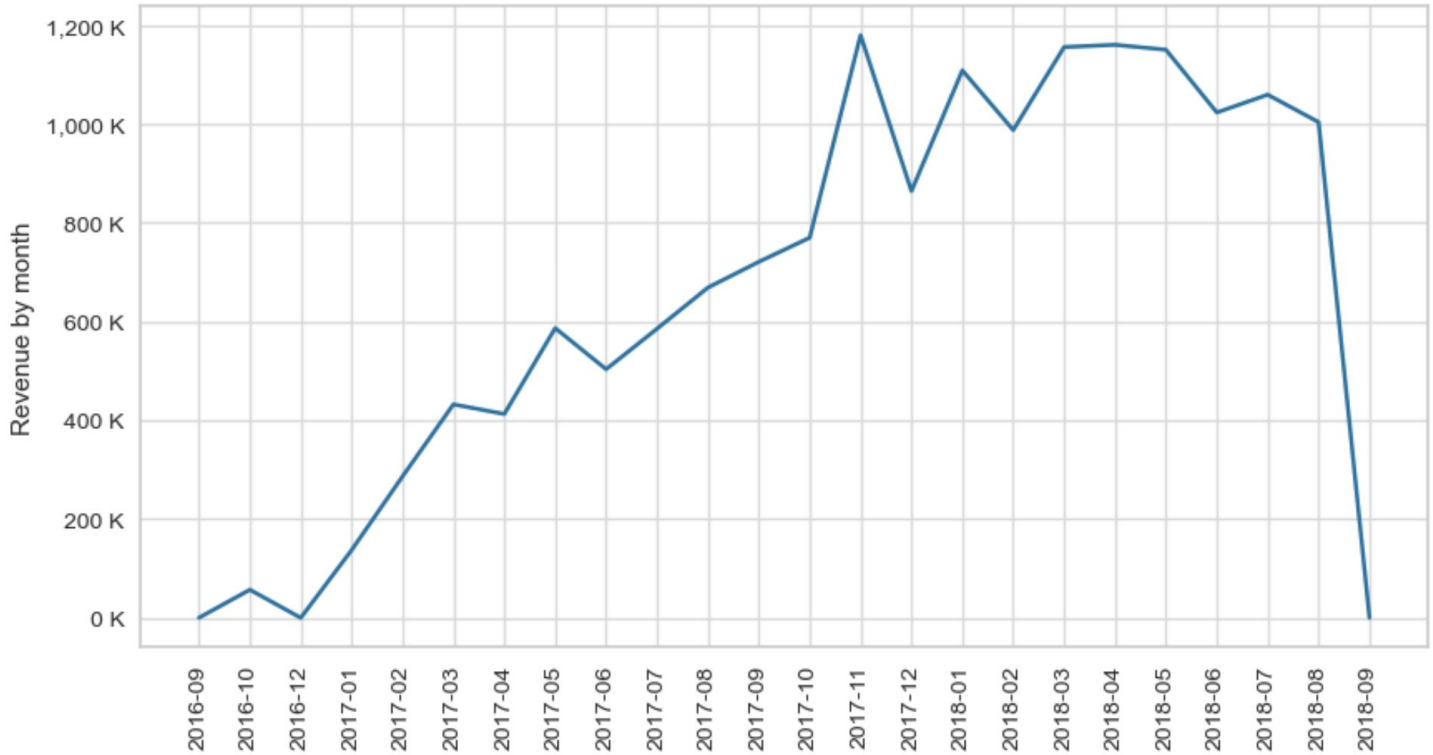
R	RECENCY	Difference in days between the last order and the customer's last order
F	FREQUENCY	Number of orders per customer
M	MONETARY	Total payment value per customer
S	SATISFACTION	Average review score per customer

EXPLORATORY ANALYSIS

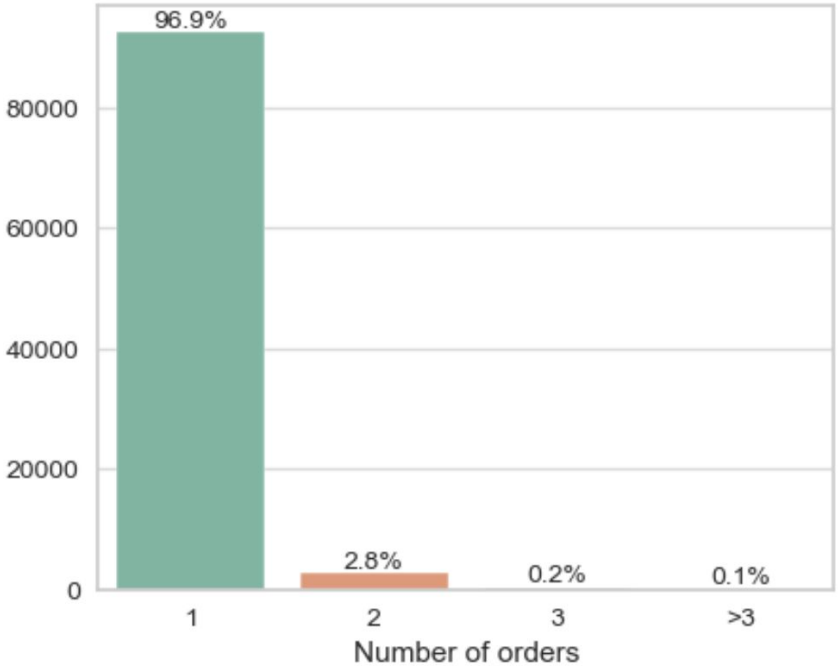
ORDERS PER MONTH



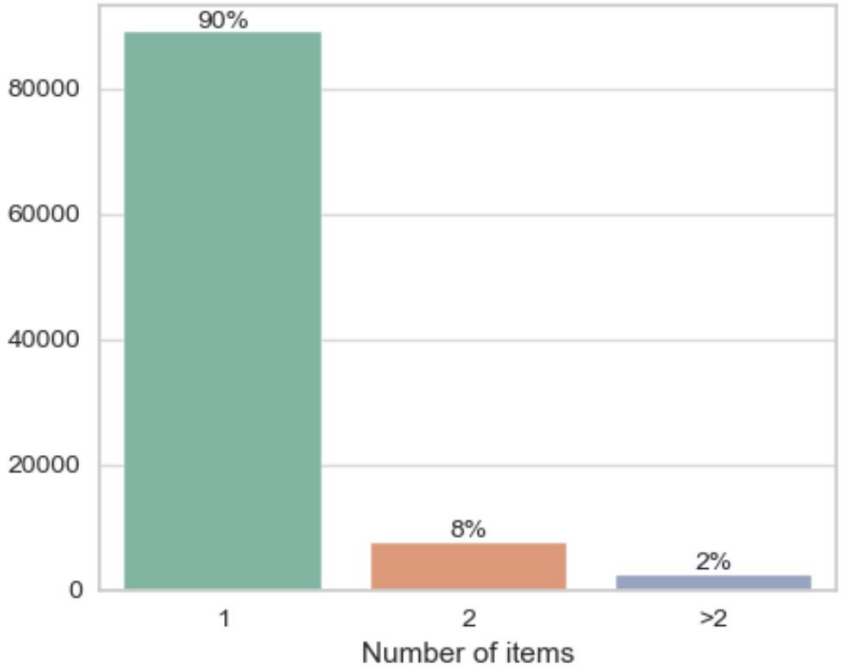
REVENUE PER MONTH



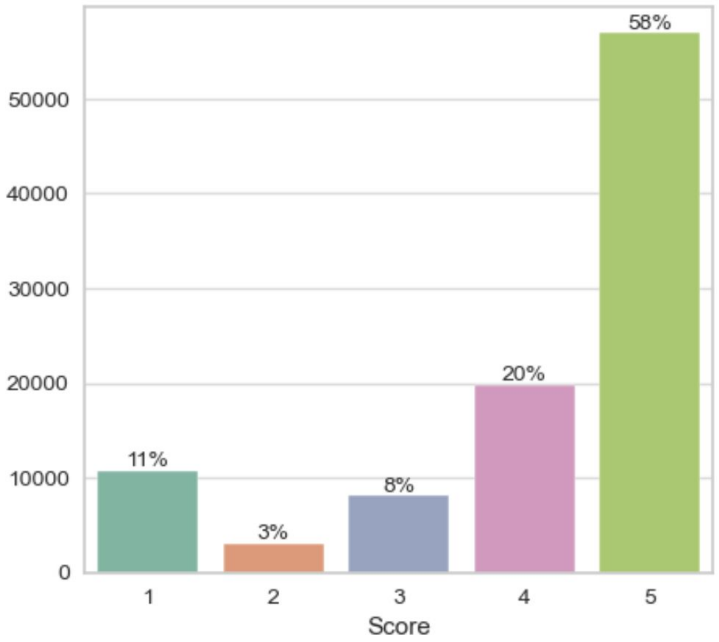
ORDERS PER CUSTOMER

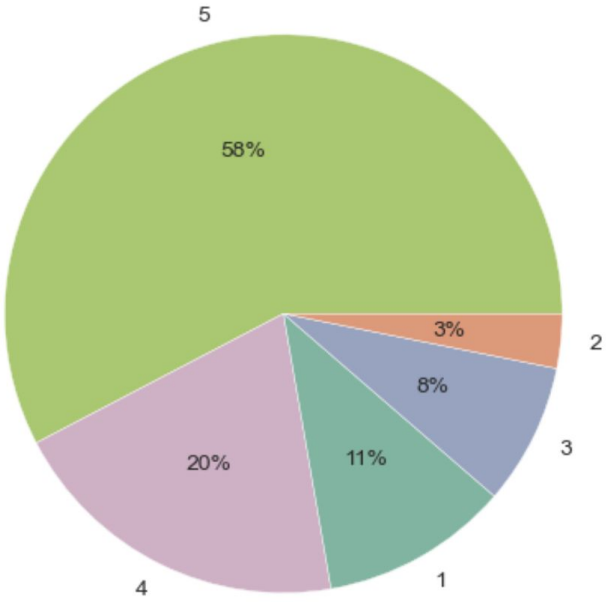


ITEMS PER ORDER



REVIEW SCORES PER ORDER





Key figures

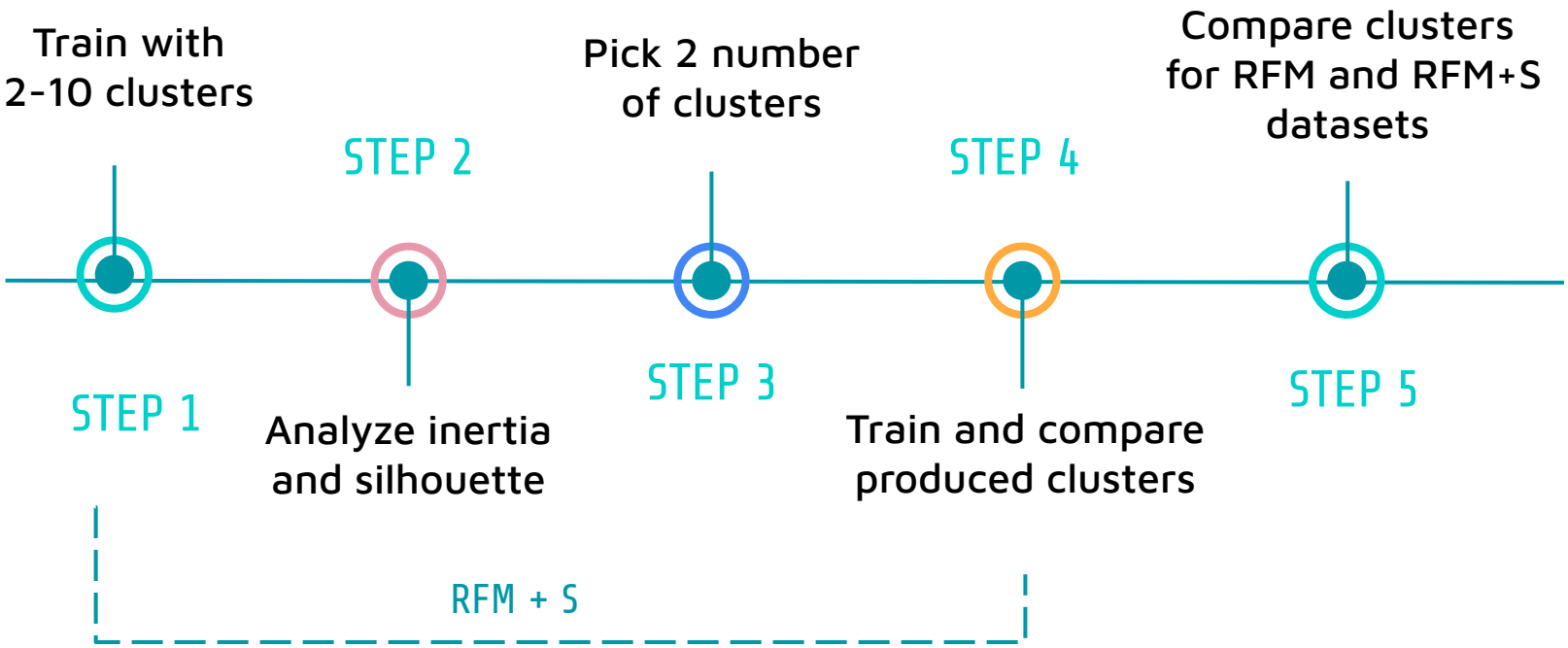
- **November 2017** is the most profitable month
- **97%** of customers with only 1 order
- **90%** of orders with only 1 item
- **78%** of satisfied customers

CLUSTERING

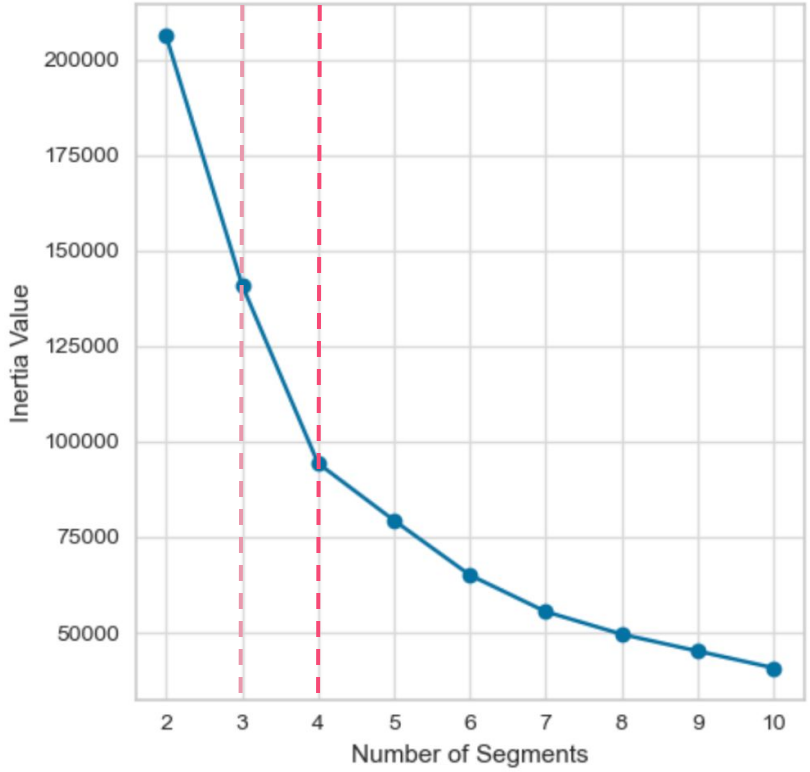
Tested algorithms

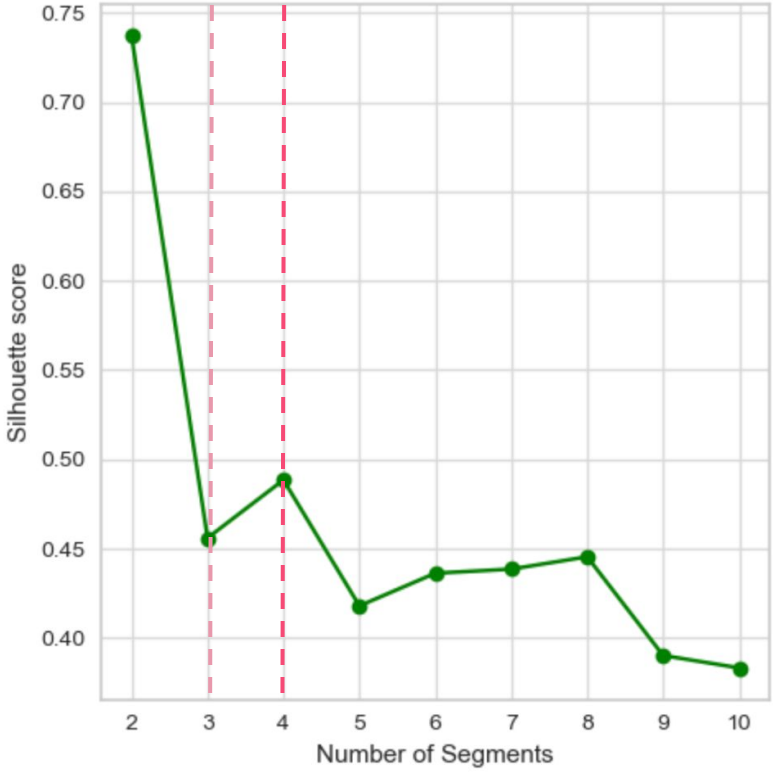
- K-Means
- Hierarchical Clustering
- DBSCAN

K-MEANS strategy

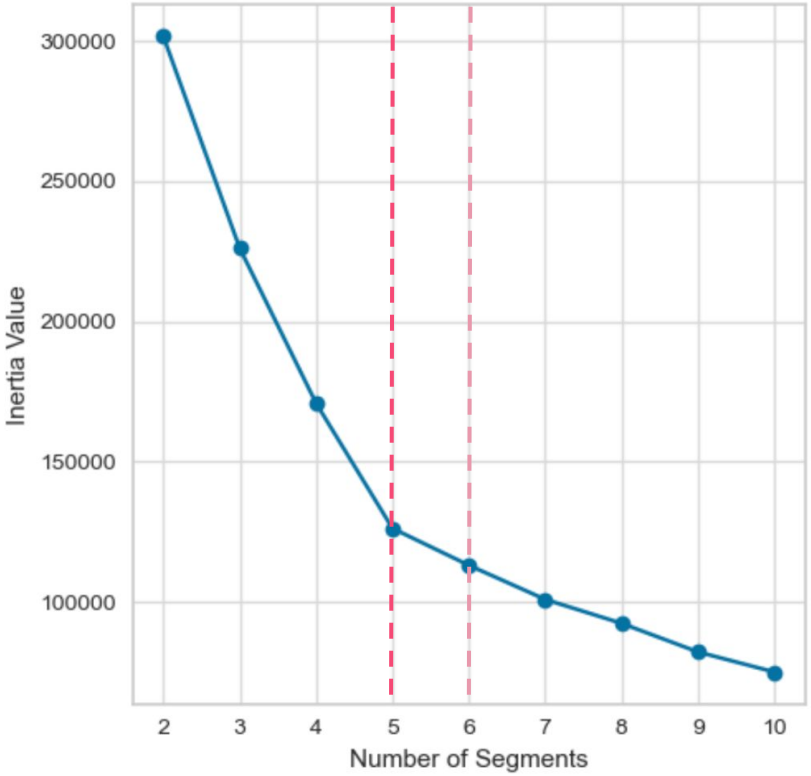


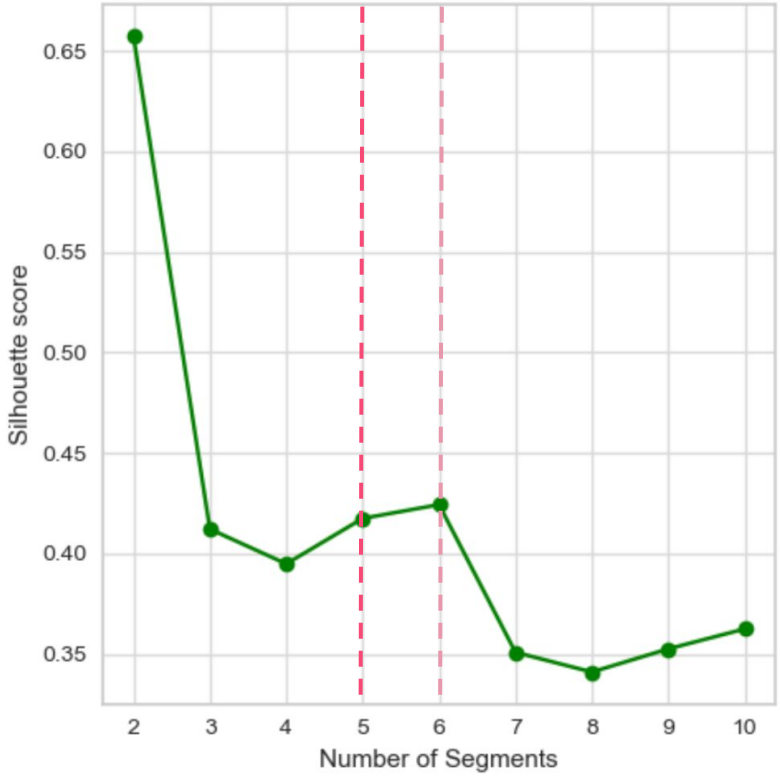
K-MEANS RFM


Inertia Value
200000
175000
150000
125000
100000
75000
50000
2
3
4
5
6
7
8
9
10
Number of Segments


Silhouette score
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
2
3
4
5
6
7
8
9
10
Number of Segments

K-MEANS RFM+S clusters

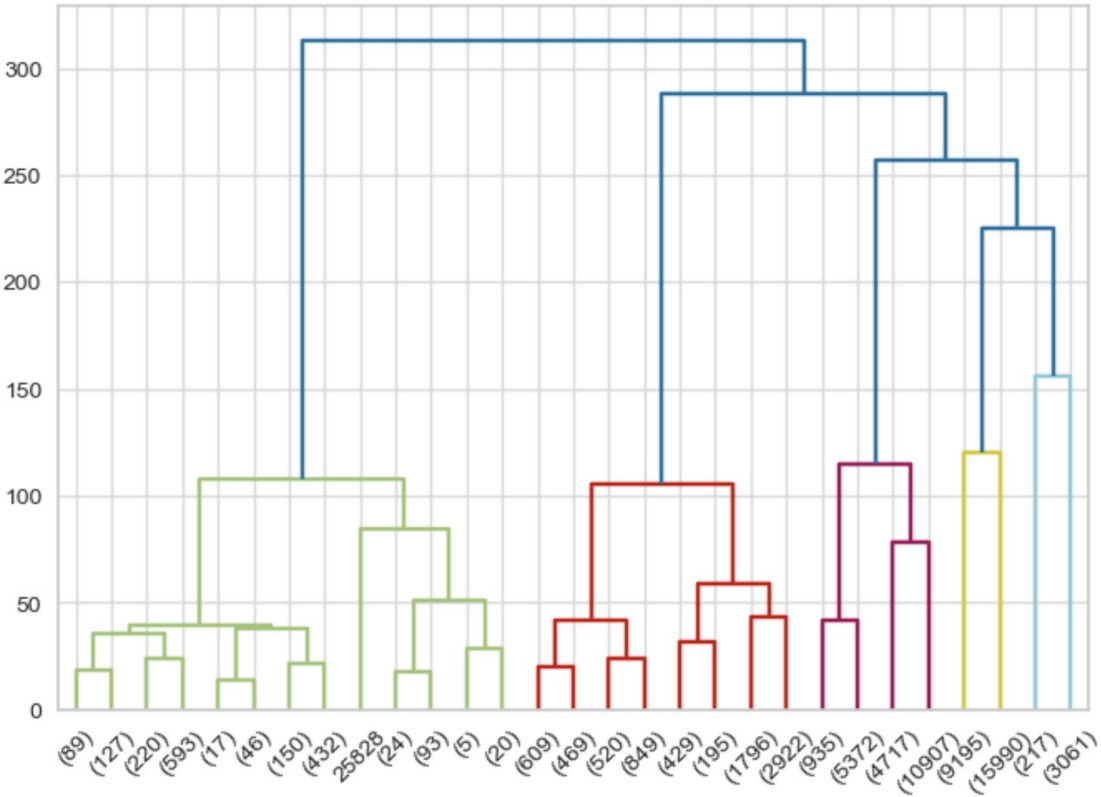




Hierarchical Clustering Linkage

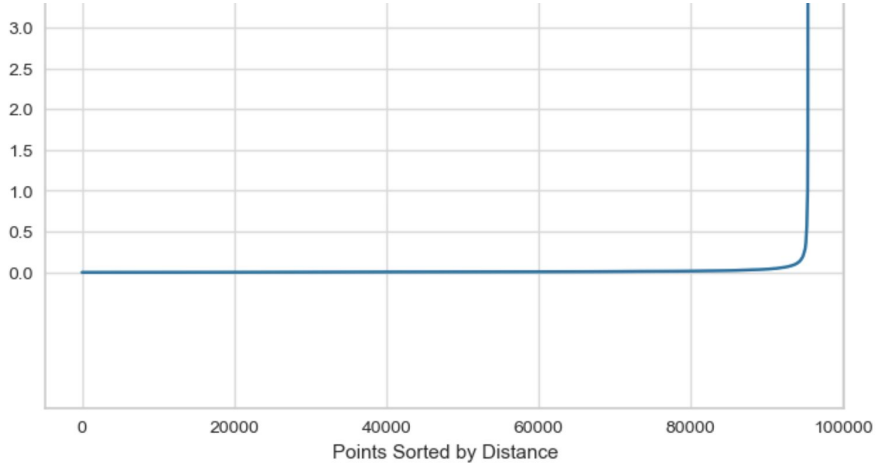
Methods

- Single
- Complete
- Average
- **Ward**

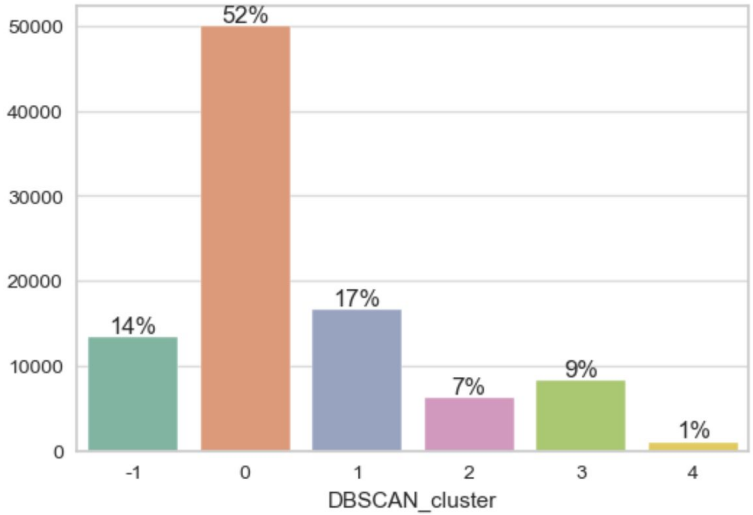


DBSCAN

Selecting hyperparameters with KNN

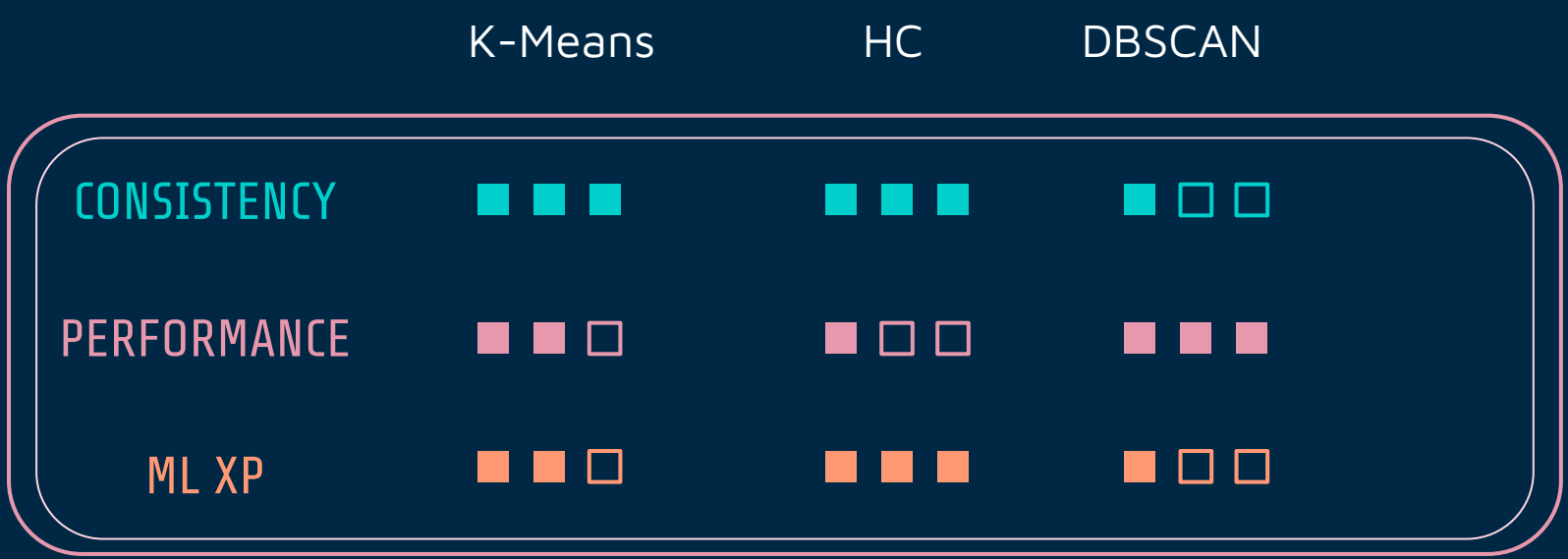


5 clusters + noise



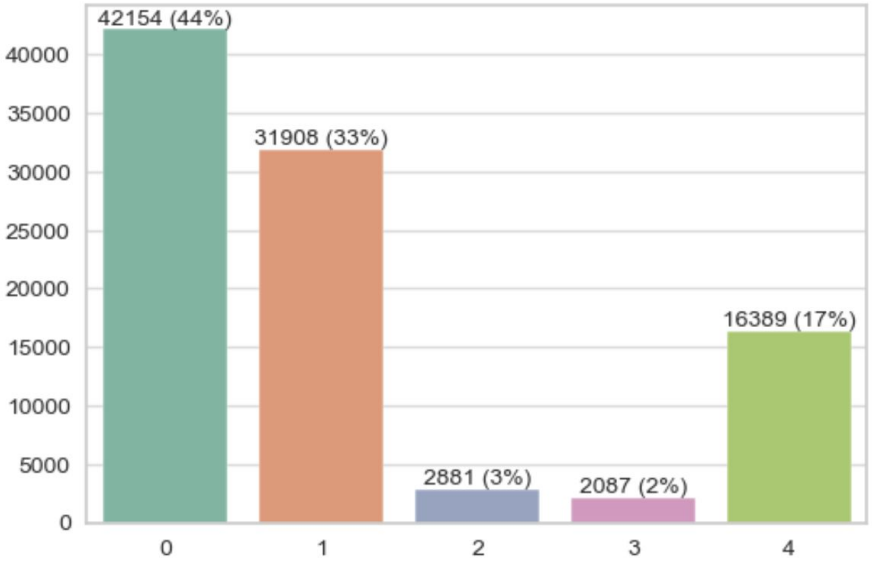
min_samples = 150
eps = 0.2

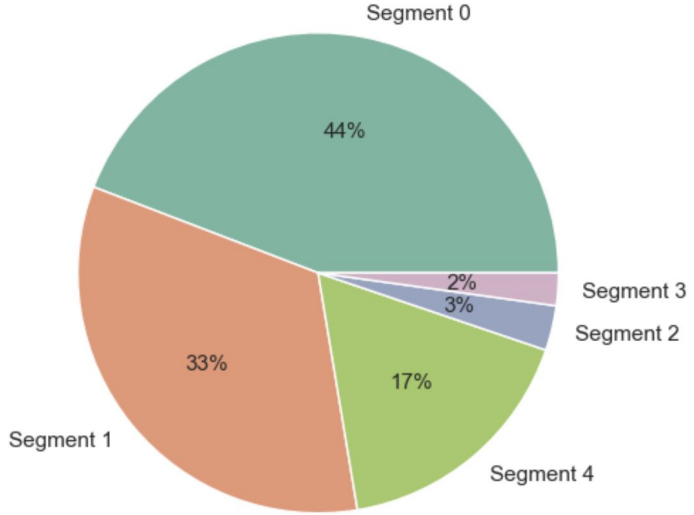
Benchmark



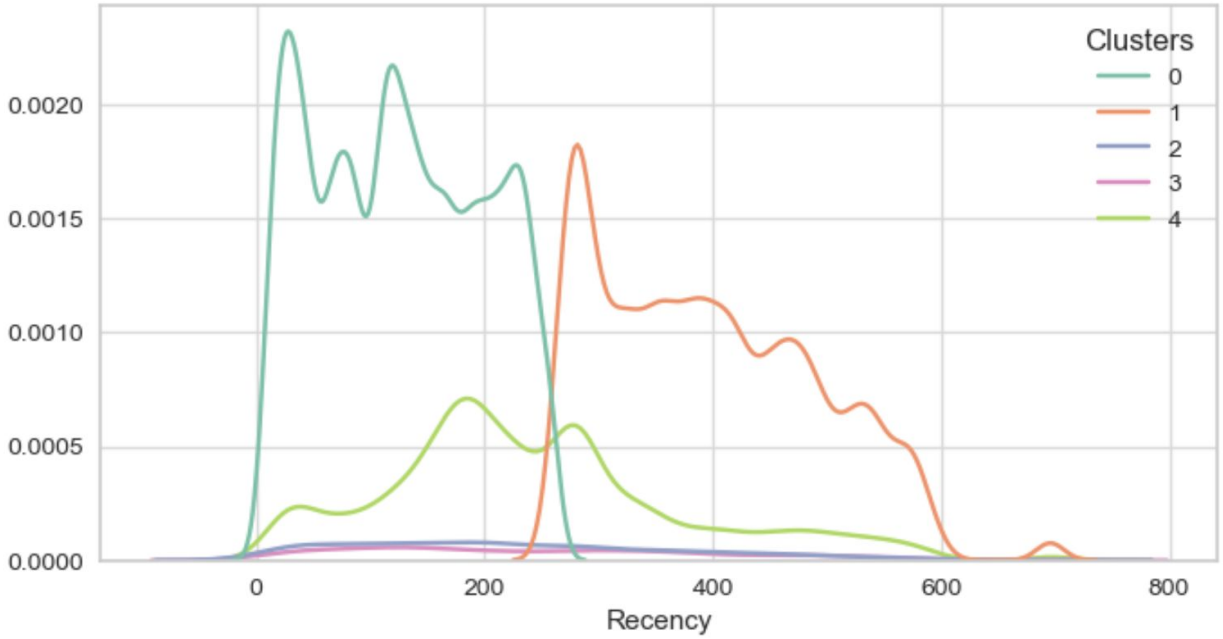
RESULTS

FINAL MODEL K-Means with RFM+S





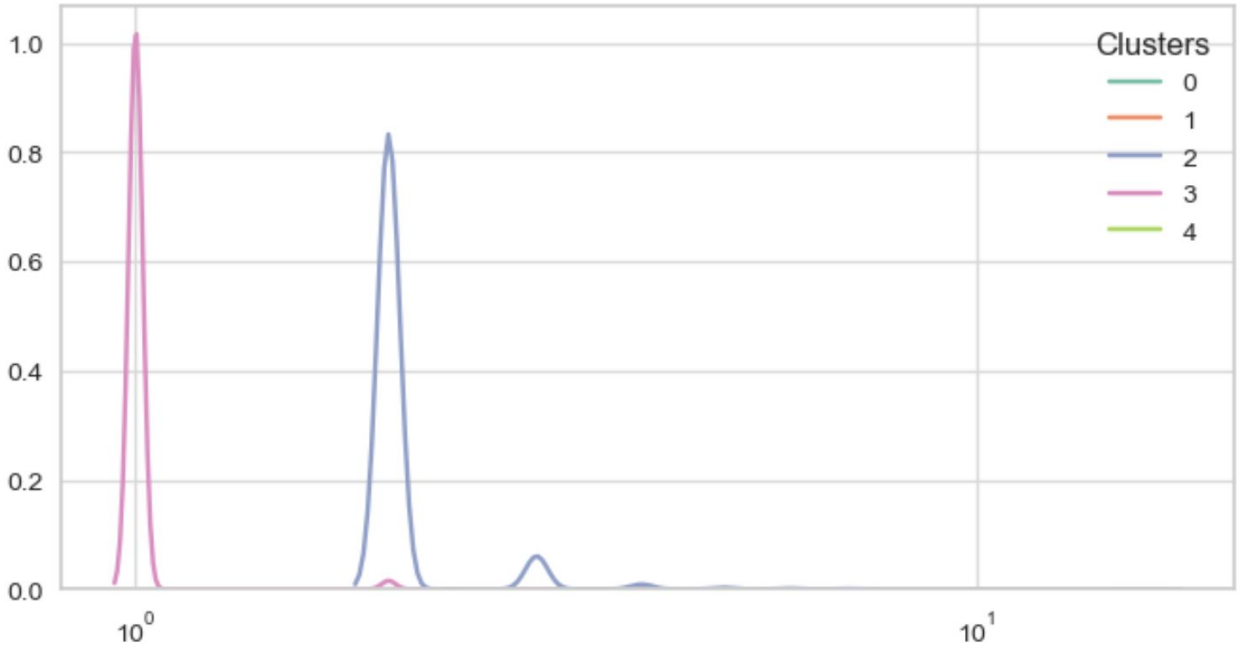
CLUSTER ANALYSIS Recency



RANKING in days

Cluster 0	126
Cluster 2	225
Cluster 3	241
Cluster 4	243
Cluster 1	397

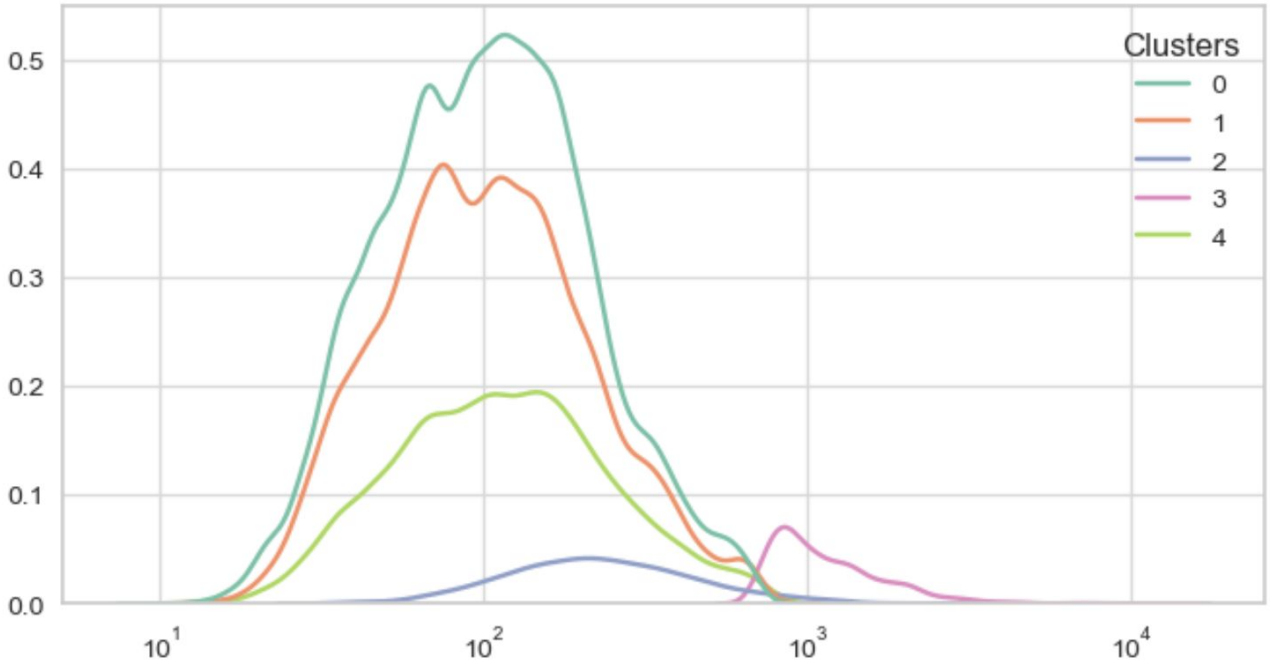
CLUSTER ANALYSIS Frequency



RANKING

Cluster 2	2
Cluster 0	1
Cluster 1	1
Cluster 3	1
Cluster 4	1

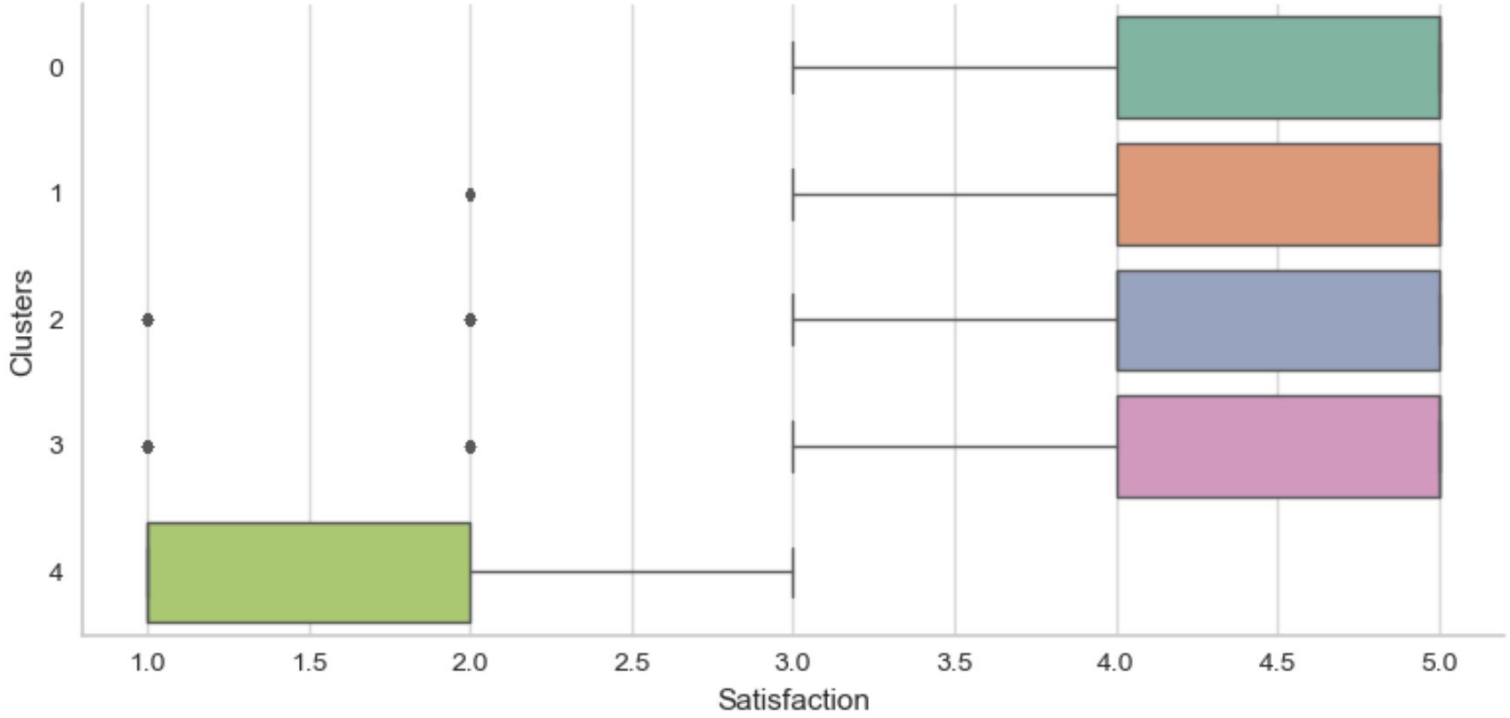
CLUSTER ANALYSIS Monetary



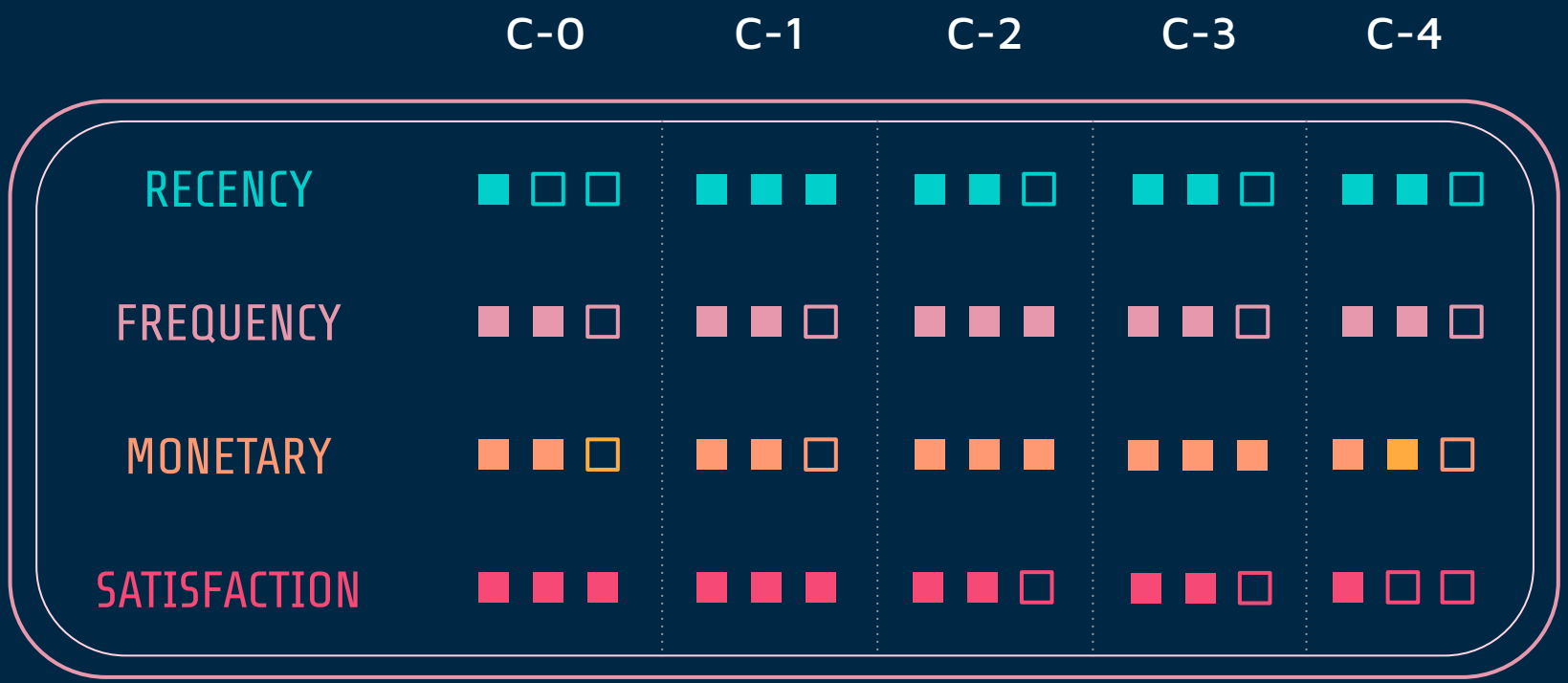
RANKING in BRL

Cluster 3	1258
Cluster 2	290
Cluster 4	152
Cluster 1	134
Cluster 0	133

CLUSTER ANALYSIS Satisfaction



Clusters overview

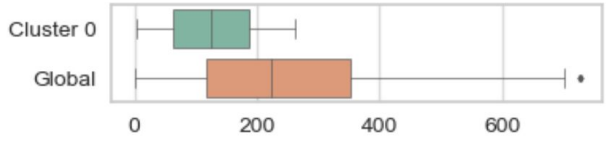


CLUSTER 0 New clients

Profile

- the largest group (44%)
- **the most recent**
- 1 order
- average spenders
- **happy**

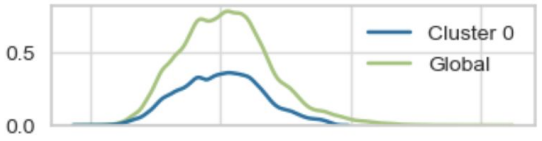
RECENCY
Below the average
(4 - 262 days)



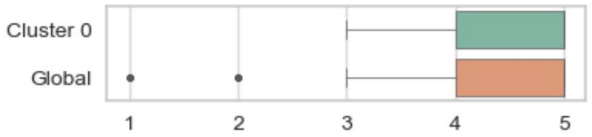
FREQUENCY
In the average

1 order

MONETARY
In the average
(133 BRL)



SATISFACTION
Above average
(4.7)

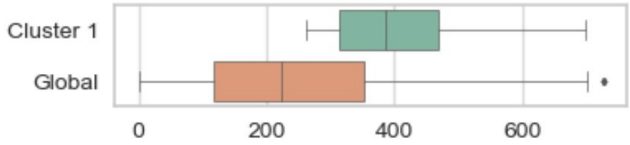


CLUSTER 1 Disengaged clients

Profile

- the 2nd largest group (33%)
- **the oldest**
- 1 order
- average spenders
- **happy**

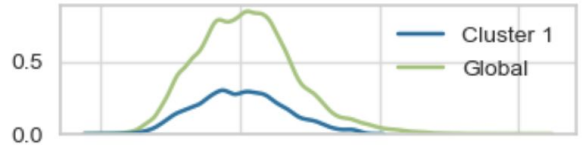
RECENCY
Above the average
(266 - 699 days)



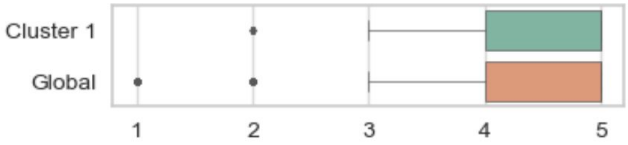
FREQUENCY
In the average

1 order

MONETARY
In the average
(134 BRL)



SATISFACTION
Above average
(4.6)

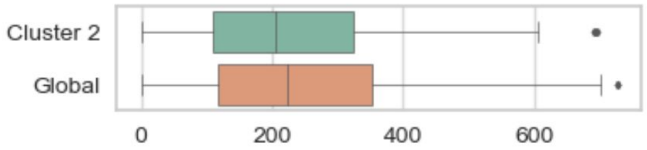


CLUSTER 2 Loyal clients

Profile

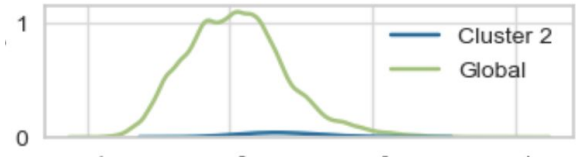
- 3% of all clients
- average recency
- **2 order and more**
- **good spenders**
- satisfied

RECENCY
In the average
(0 - 696 days)



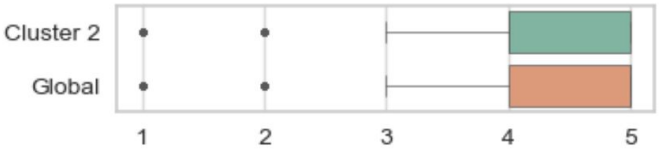
FREQUENCY
Above the average
(2 - 16)

2 orders in average



MONETARY
Above the average
(290 BRL)

SATISFACTION
In the average
(4.1)

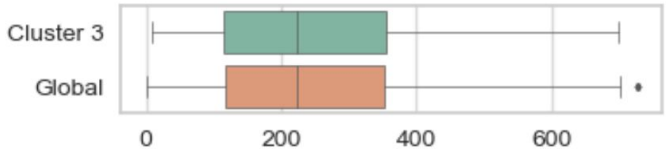


CLUSTER 3 Best spending clients

Profile

- The smallest group (2% of all clients)
- average recency
- 1 order
- **the best spenders**
- satisfied

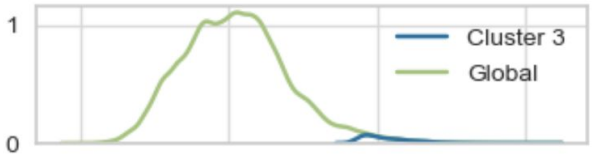
RECENCY
In the average
(8 - 698 days)



FREQUENCY
In the average

1 order

MONETARY
Above the average
(1258 BRL)



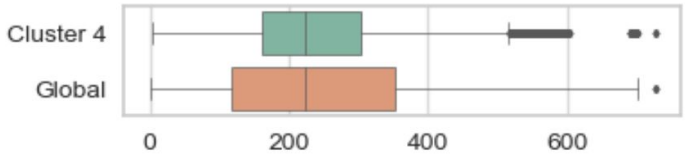
SATISFACTION
In the average
(4)

CLUSTER 4 Unhappy clients

Profile

- **17% of all clients**
- average recency
- 1 order
- average spenders
- **very unhappy**

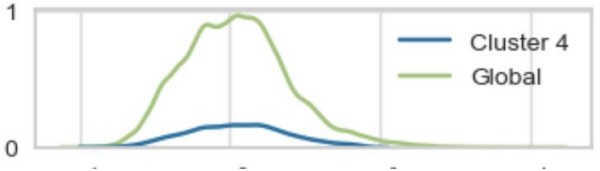
RECENCY
In the average
(4 - 728 days)



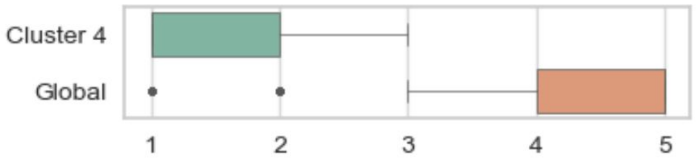
FREQUENCY
In the average

1 order

MONETARY
In the average
(152 BRL)

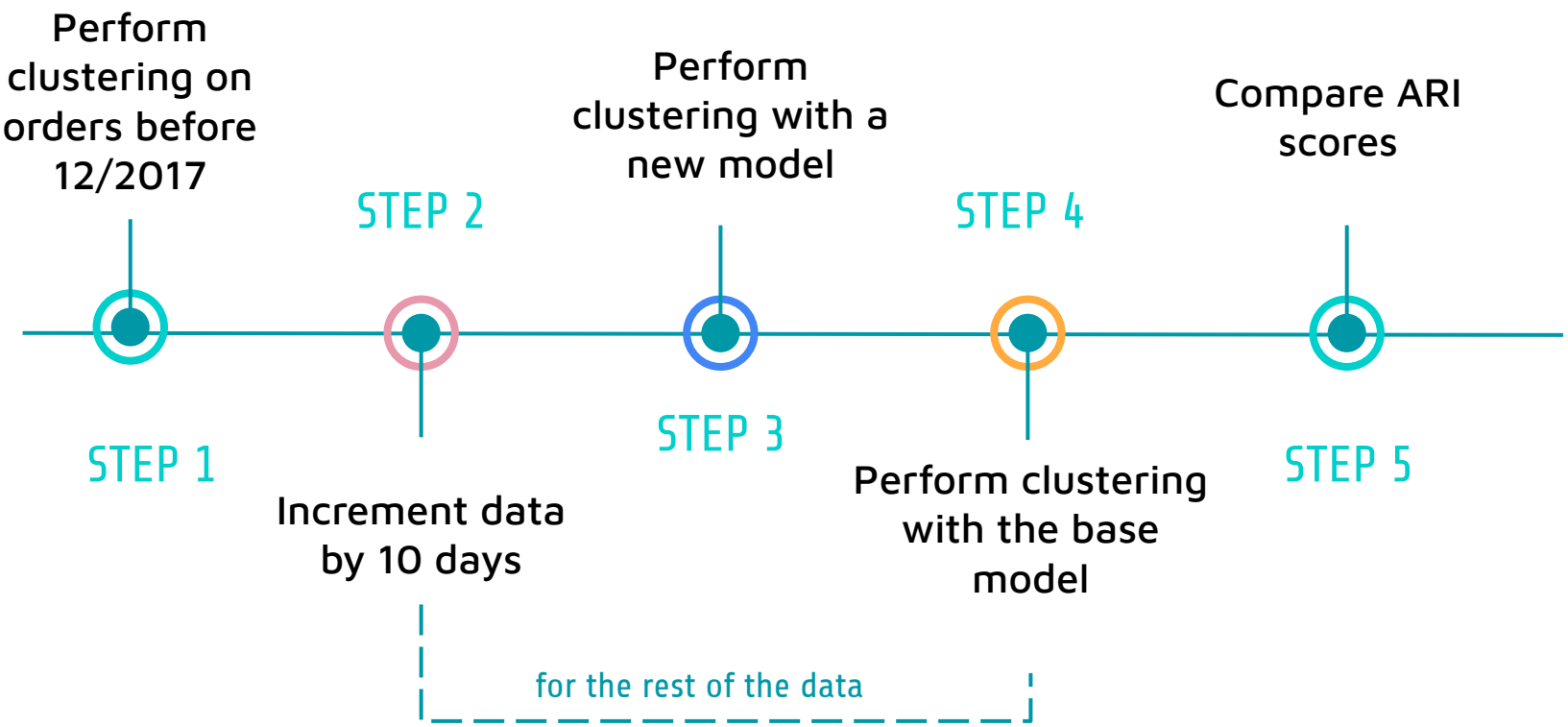


SATISFACTION
Below the average
(1.6)



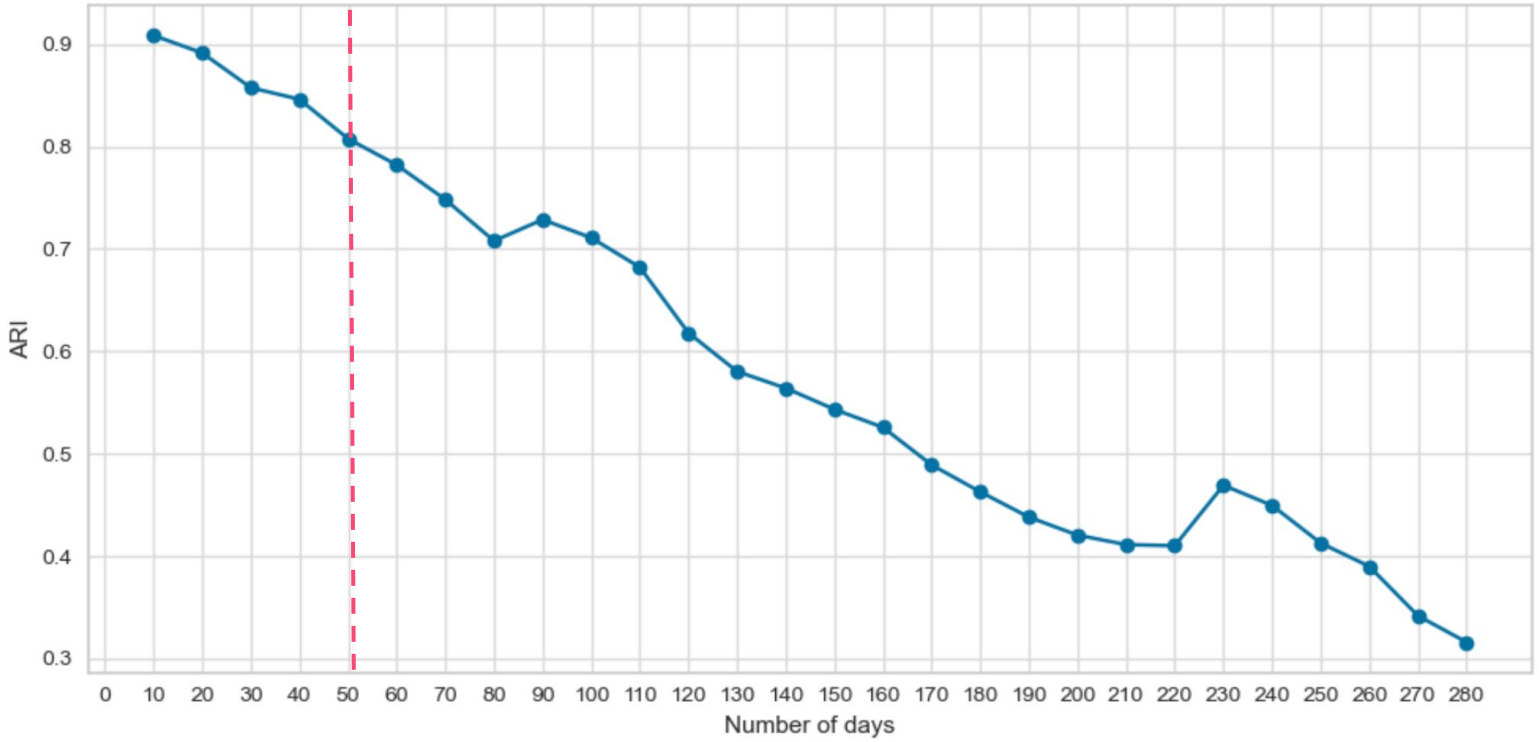
UPDATE FREQUENCY

SIMULATION PROCESS



Projection of database updates

every 50 days



KEY OUTCOMES

- The best algorithm: **K-Means**
- Number of clusters: **5**
- Update frequency: **50 days**

TO GO FURTHER

- Analyze order and revenue seasonality (with data covering a higher time frame, more than 2 years)
- Analyze customer duplicates (same customers creating new accounts to use welcome vouchers) and aggregate data
- Add additional features to the RFM dataset (average basket, items per order, geolocation: big city/small city, etc) with more rich data
- Try other versions of clustering algorithms (Bisecting K-Means, HDBSCAN, etc)

THANKS!