Segmenting customers of an e-commerce

APRIL 2022

Presentation Outline

1. Objectives
2. Dataset Preparation
3. Modeling options
4. Final model overview and associated maintenance time

Objectives

Context

- Olist: Brazilian e-commerce site
- Desire for customer segmentation, for the use of the marketing team

Business Problem

- Understanding the different types of users
- Targeting communication campaigns

Mission

- Provide Olist a segmentation of customers
- Provide an actionable description for each segment
- Analyse the stability of segments over time

Objetives

Approach

Methodology

1- Extract data from the database to characterise customers

2- Use unsupervised machine learning tools to partition clients based on these characteristics

3- Interpret the resulting segments from a business perspective

4- Analyse the stability to evaluate a maintenance frequency

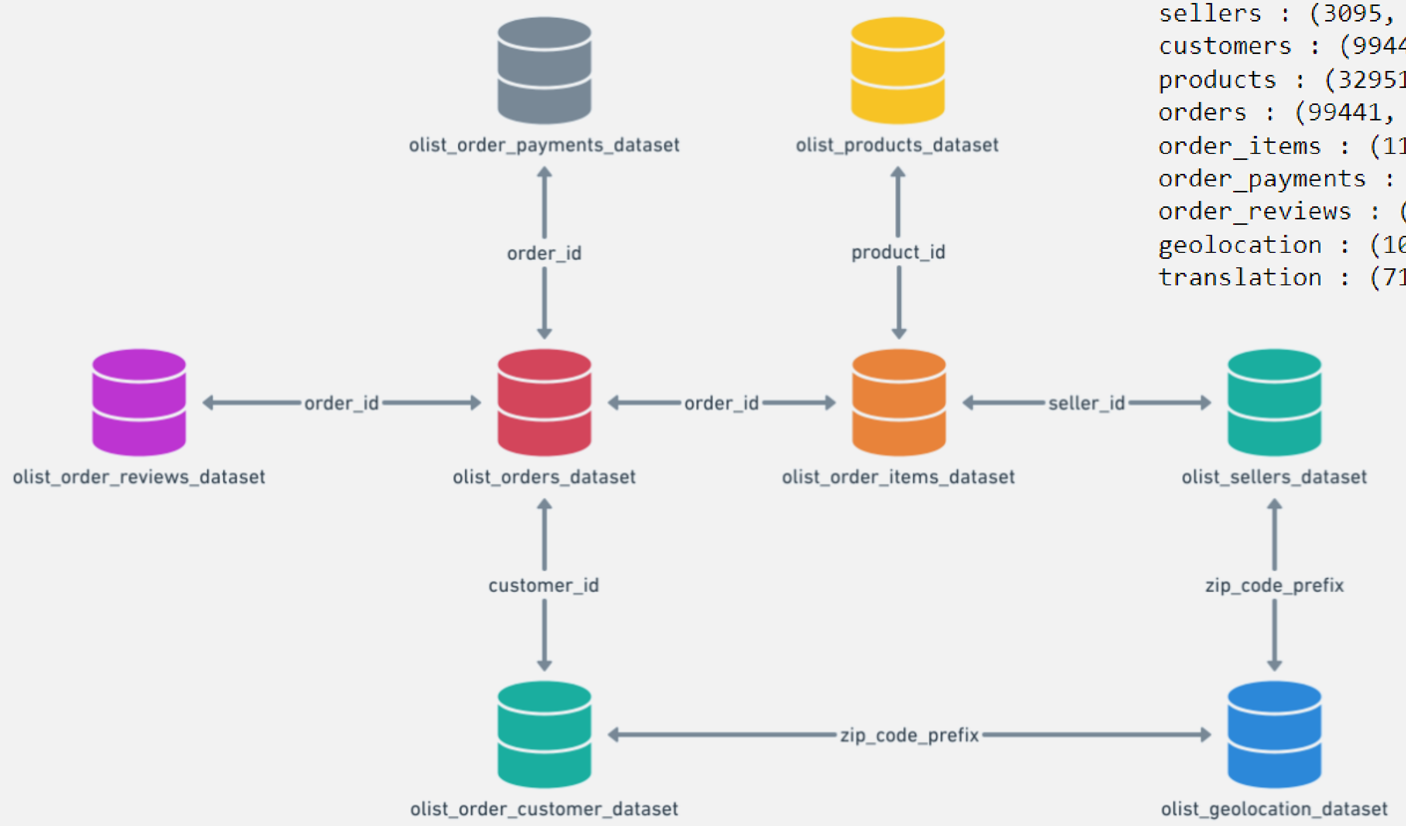
Dataset Preparation

Dataset to model

Dataset

Observations

- A dataset of 9 files detailing customers, orders, products, sellers from end 2016 to end 2018
- Customer, seller, order, product identified by a unique ID
- A well-filled dataset – all files < 1% NaN except Order_review with 21% NaN




```
sellers : (3095, 4)
customers : (99441, 5)
products : (32951, 9)
orders : (99441, 8)
order_items : (112650, 7)
order_payments : (103886, 5)
order_reviews : (99224, 7)
geolocation : (1000163, 5)
translation : (71, 2)
```

Dataset Preparation

Methodology

Process followed in 4 steps

Cleaning files
- Correction of types (date)
- Removal of duplicates (geolocation, ...)
- Dealing with missing values (category Unknown, ...)
- Dealing with outliers (payment_installment = 0)

File aggregation
- Merging files to 'order_id' or 'custumer_id'
- Selection of orders with status delivered

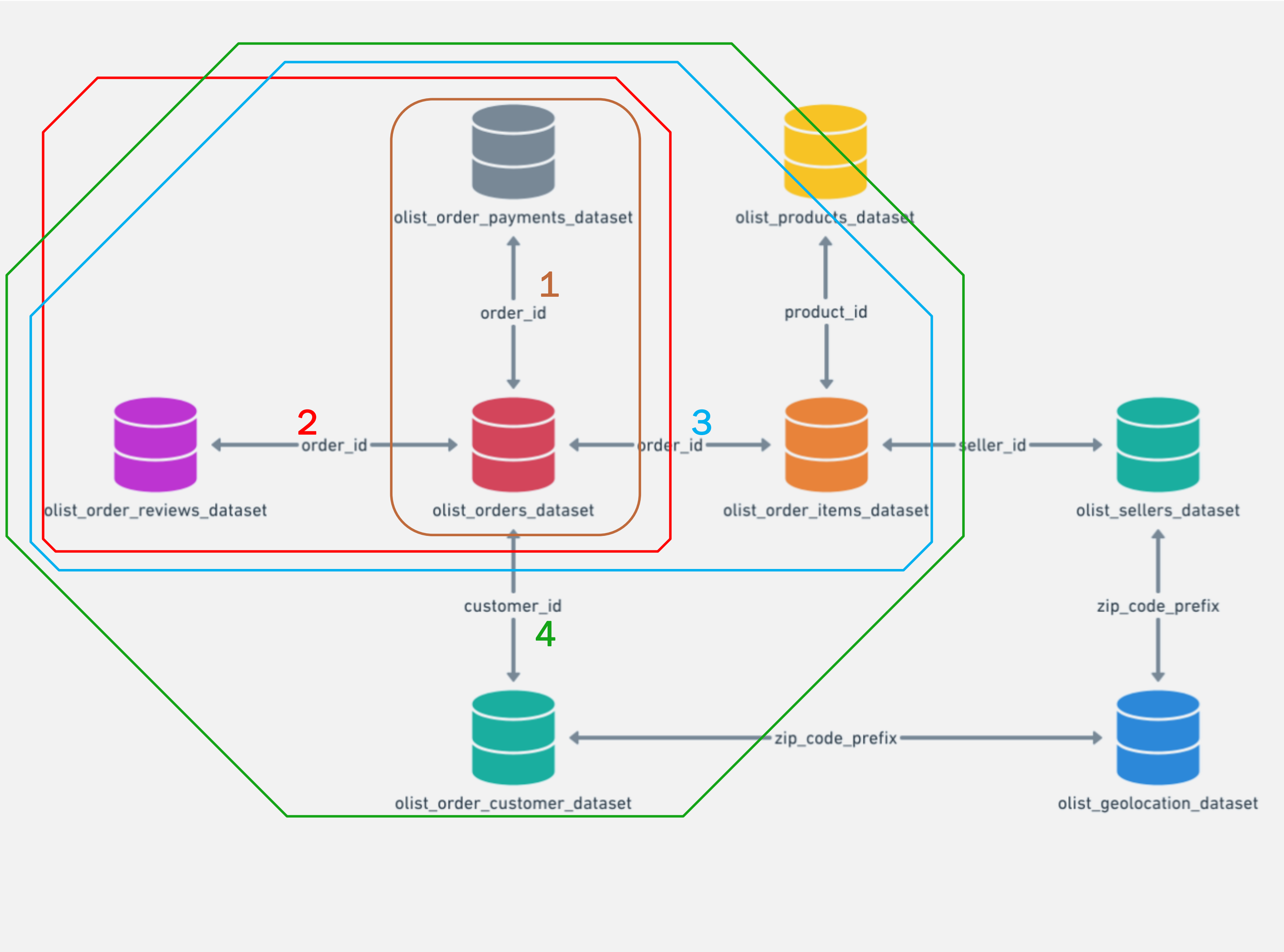
Feature Engineering
- Creating variables
- Transforming variables
- Selecting variables

Exploratory Analysis
- Distribution of variables
- Correlation between variables
- Selecting variables

Dataset Preparation

Files aggregation

Aggregation steps



Dataset Preparation

Feature Engineering

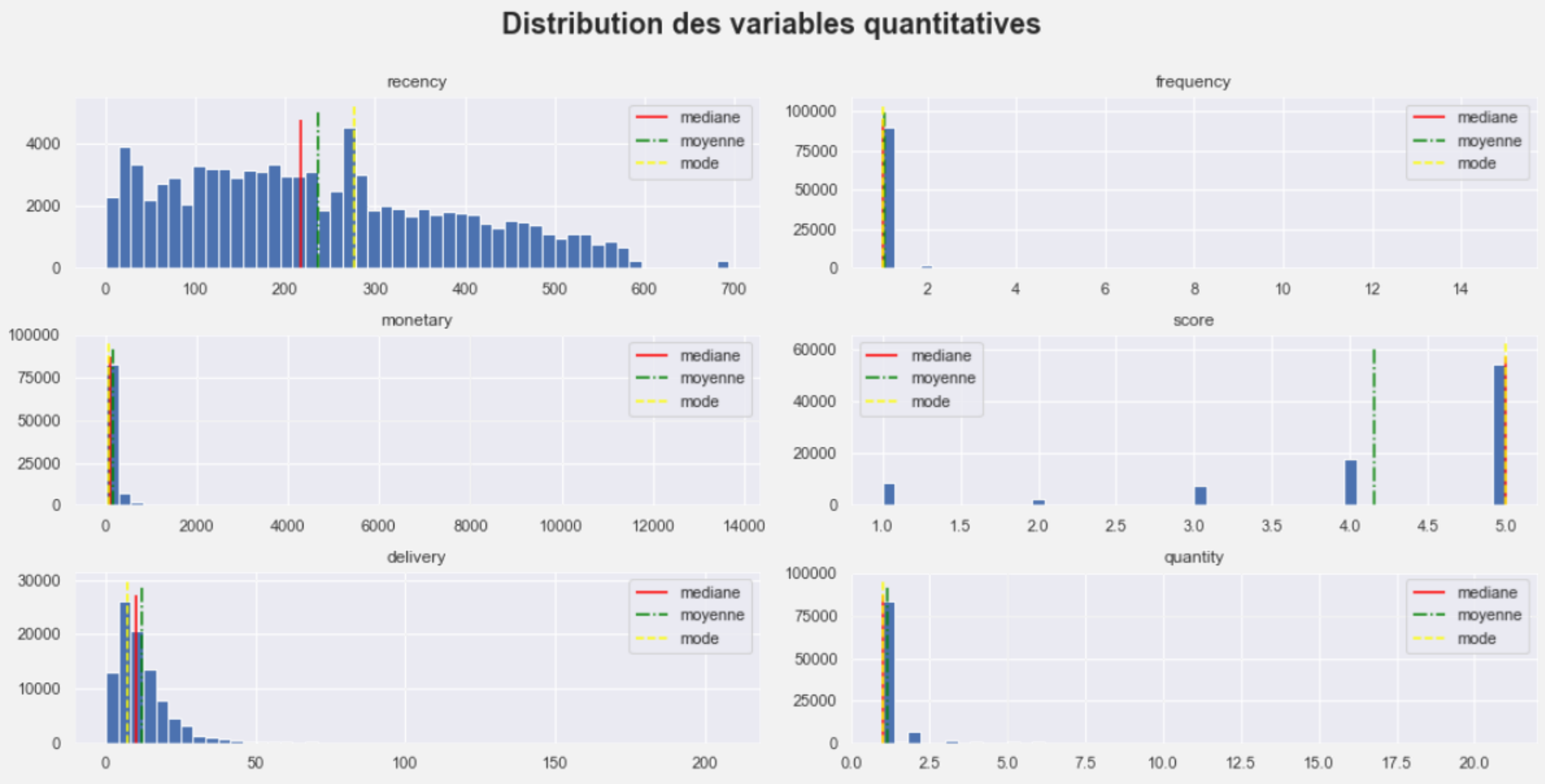
Creating variables per client

Variable	Meaning
Recency	Number of days between the customer's last order on the site and the last order on the site
Frequency	Number of orders
Monetary	Average amount per order spent
Score Score	Average score
Delivery	Average delivery time
Quantity	Average number of products purchased
Order_value	Average sum of products per order spent
Freight_value	Average sum of deliveries per order spent
Freight_per	Average % of freight value on order
Delivery_acc	Average number of days ahead of delivery compared to the estimated date
Delivery_per	Average % of delivery advance compared to effective delivery time of the order
Reaction	Number of days elapsed between receipt of the order and post of the review
Pay_inst	Average number of payment installs
Pay_type	Preferred payment type

Dataset Preparation

Exploratory Analysis

Distribution of variables



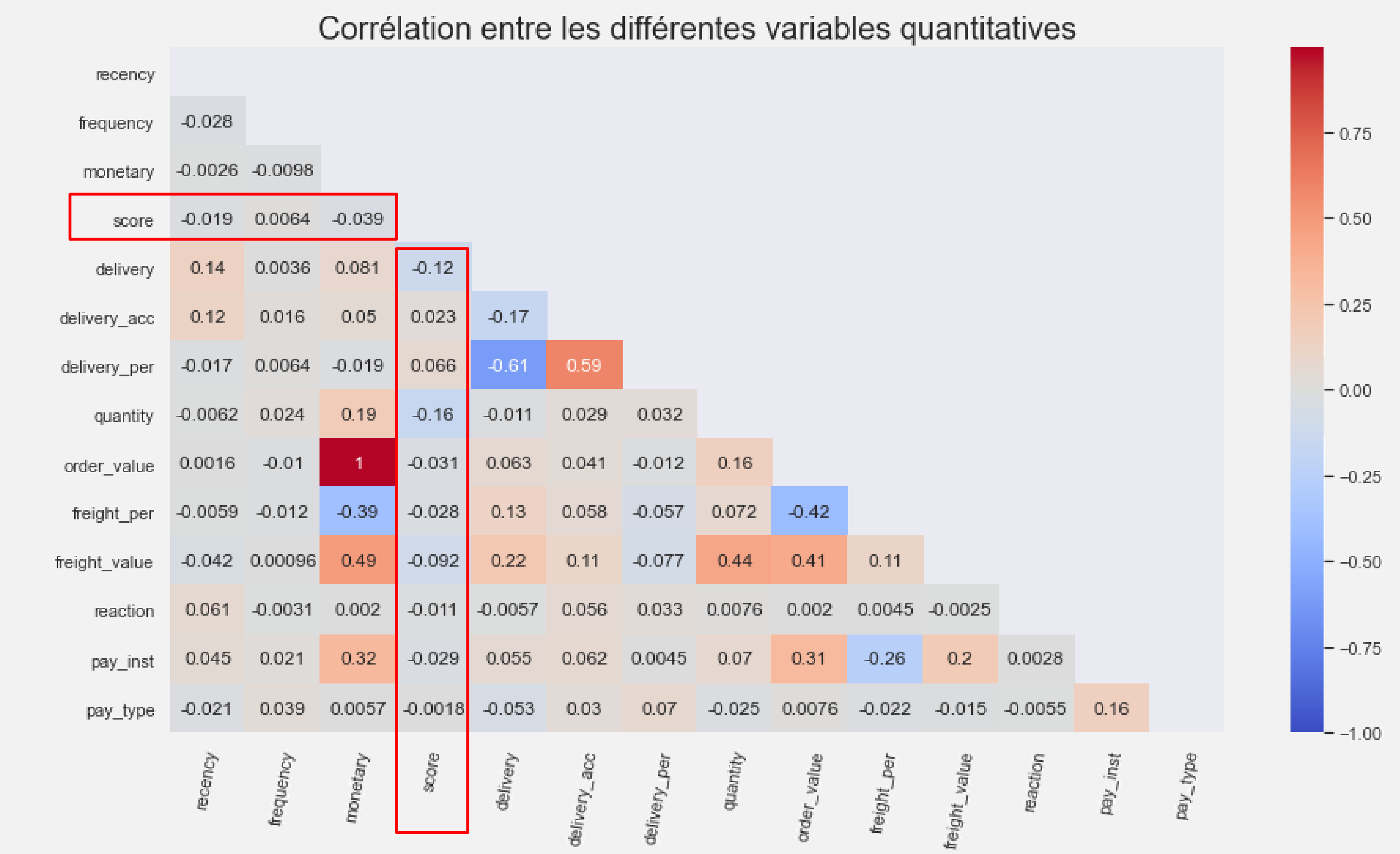
Observations

- Many distributions with a strong skewness on the right, hence the consideration of a standardisation of features for better clustering
- Some outliers

Dataset Preparation

Exploratory Analysis

Correlation between quantitative variables



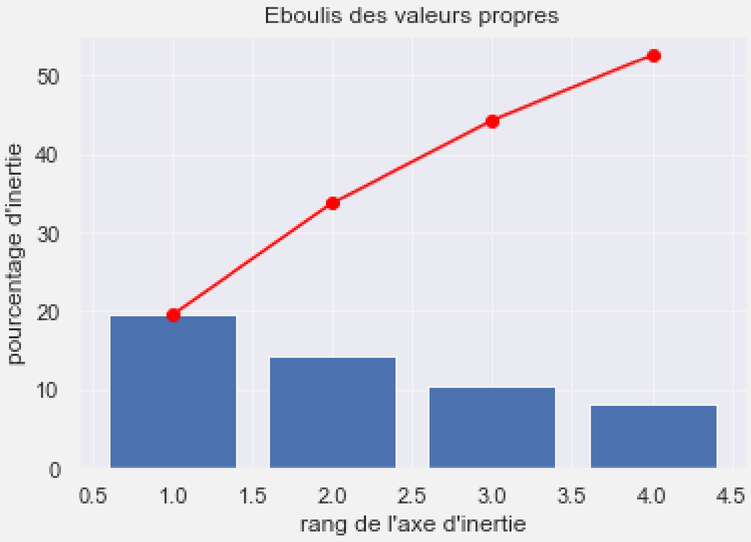
Observations

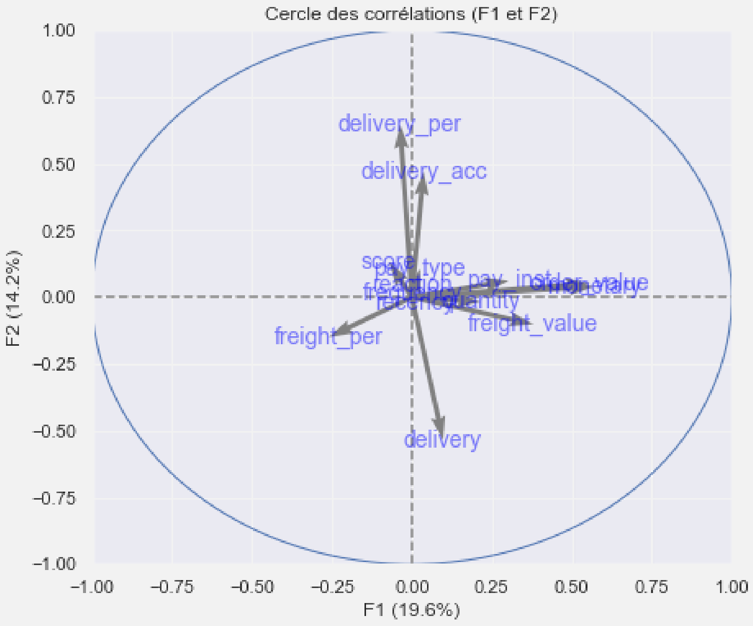
- No noticeable correlation between score and other variables
- Strong correlation between order_value and monetary

Dataset Preparation

Exploratory Analysis

Principal Component Analysis





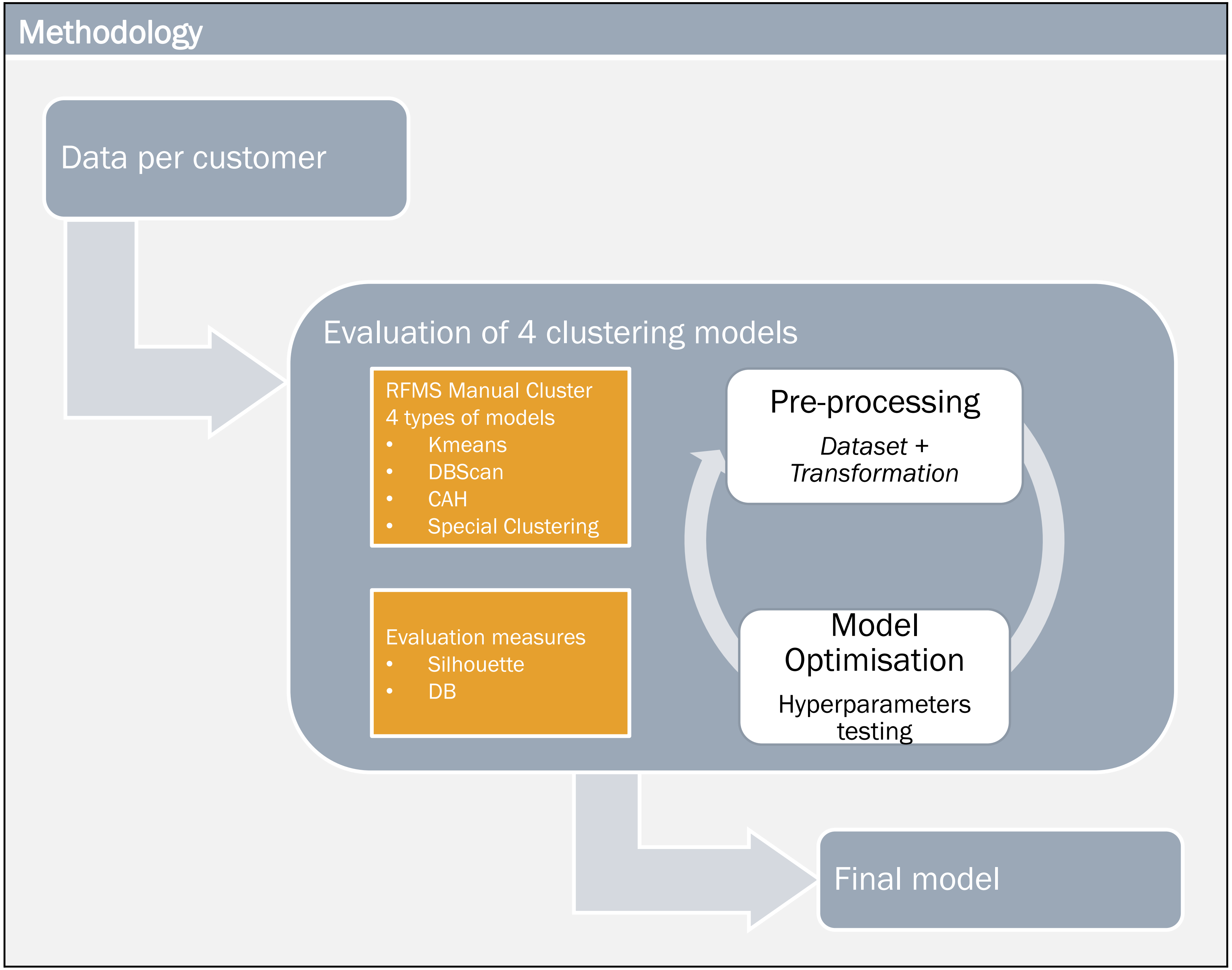
	PC1	PC2	PC3	PC4
recency	0.0099	-0.025	0.056	0.75
frequency	0.0027	0.017	0.012	-0.096
monetary	0.57	0.039	-0.095	-0.03
score	-0.071	0.13	-0.25	-0.052
delivery	0.098	-0.54	0.085	0.3
delivery_acc	0.037	0.47	0.31	0.3
delivery_per	-0.037	0.65	0.14	-0.0074
quantity	0.22	-0.017	0.51	-0.22
order_value	0.56	0.052	-0.15	-0.021
freight_per	-0.26	-0.15	0.53	-0.019
freight_value	0.38	-0.1	0.45	-0.11
reaction	0.0024	0.047	0.069	0.38
pay_inst	0.3	0.061	-0.14	0.19
pay_type	0.025	0.11	-0.086	0.057

Observations

- The first 4 components contain 55% of the variance

Modeling options

Methodology



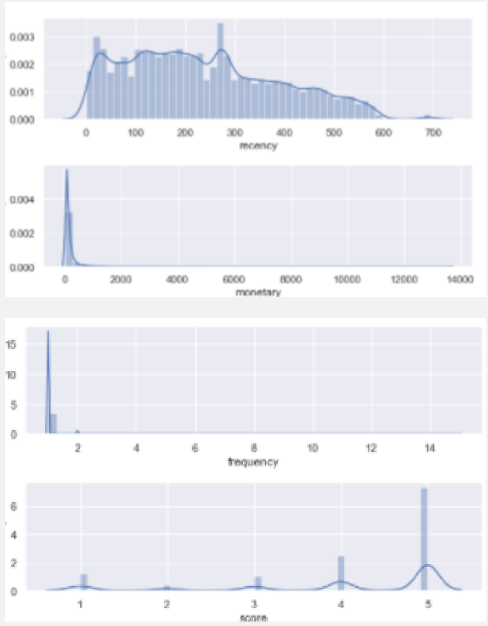
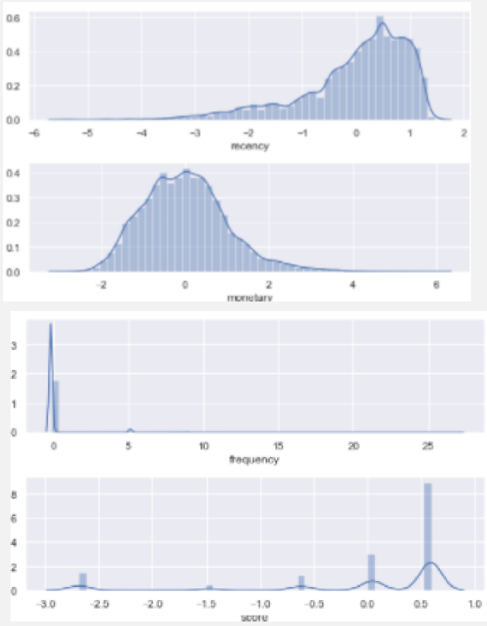
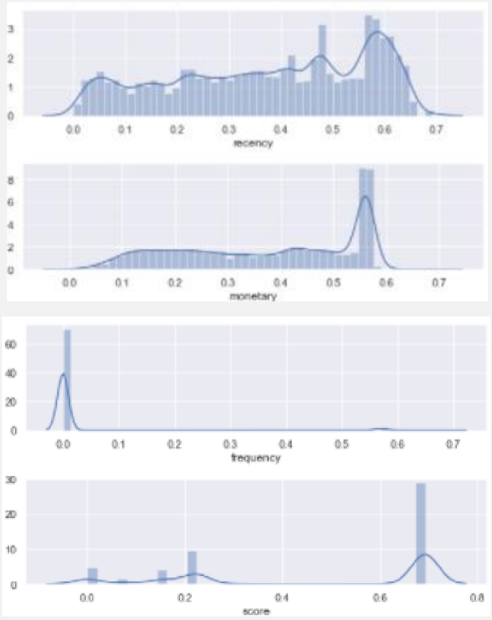
Modeling options

Pre-processing

Considered pre-processing options

Dataset	Features
RFMS	Recency, Frequency, Monetary, Score
Autre	Other combination of several features

Features transformation

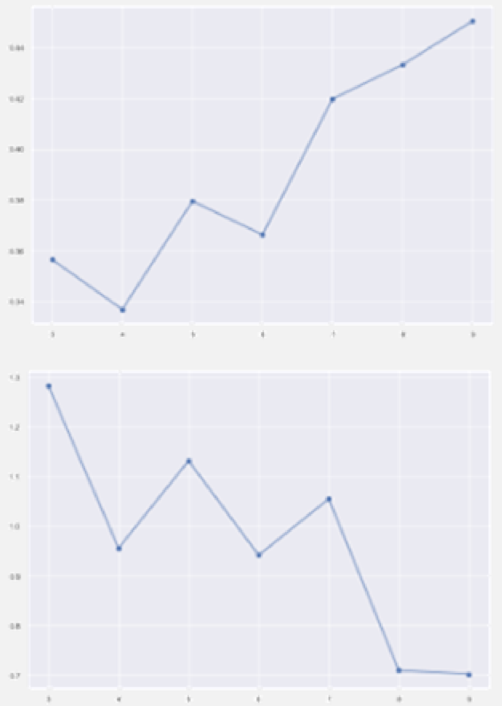
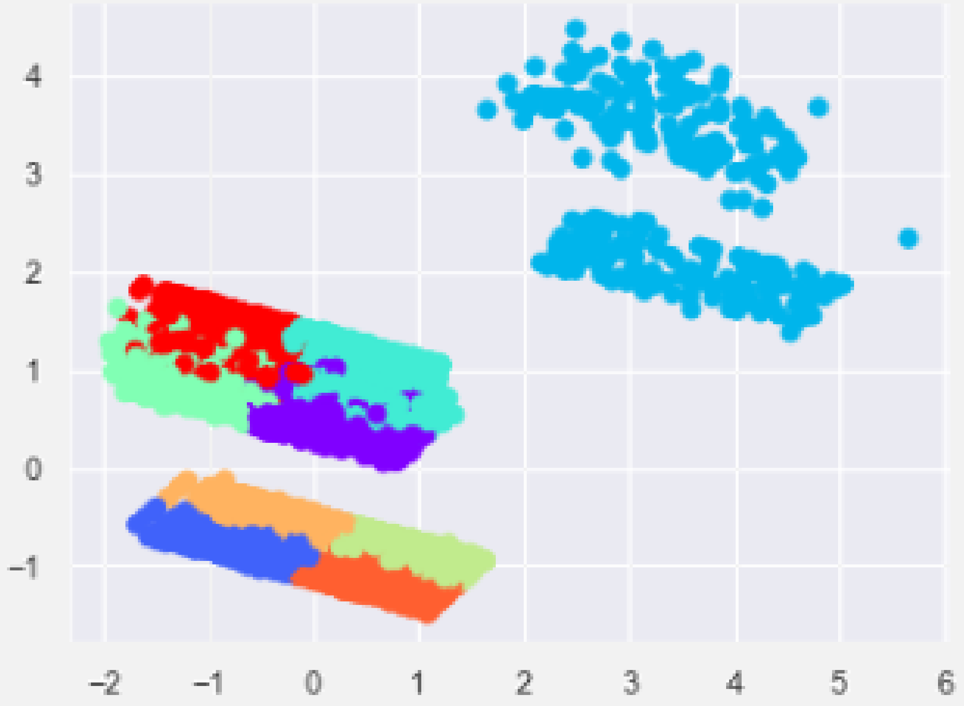
No transformation	Log + StandardScaler	QuantileTransformer
		

Modeling options

Model optimisation

KMeans

Dataset	RFMS (frac 10%)
Transfo features	QuantileTransformer
Algo clustering	KMeans
Silhouette – nb clusters	Best cluster = 9
DB – nb clusters	Best cluster = 9

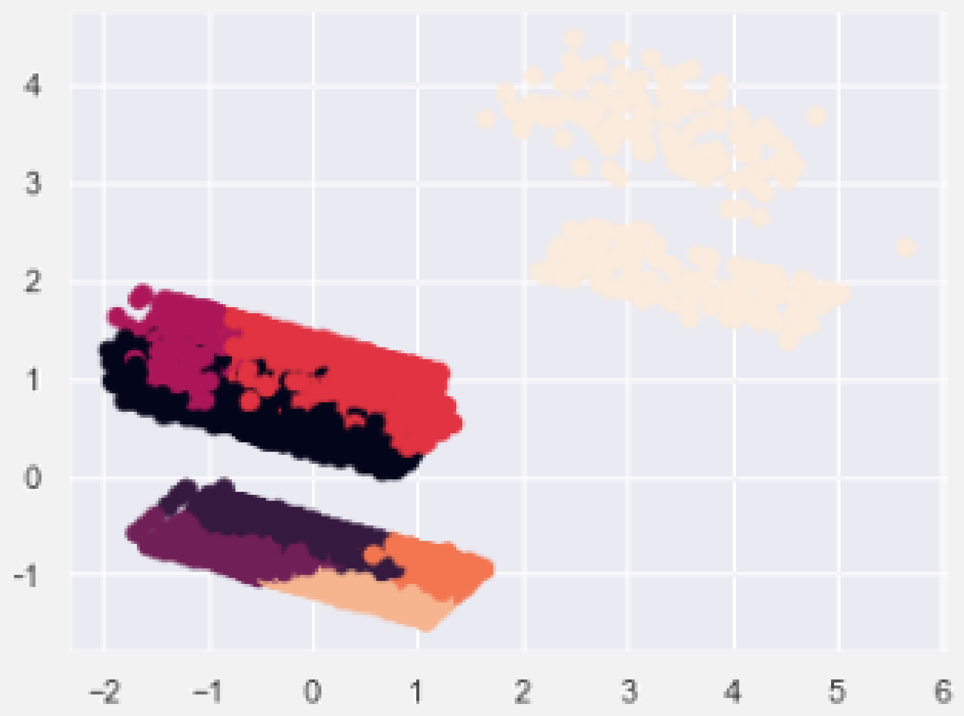
Silhouette (top) vs.DB (bottom)	Visualisation of clusters on PCA PC1 and PC2
	

Modeling options

Model optimisation

ACH

Dataset	RFMS (frac 10%)
Transfo features	QuantileTransformer
Algo clustering	Agglomerative Clustering
Silhouette – nb clusters	Best cluster = 8
DB – nb clusters	Best cluster = 9

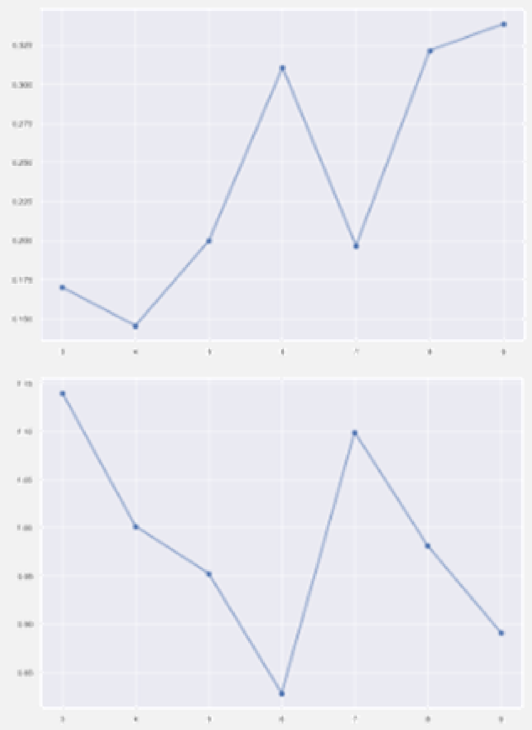
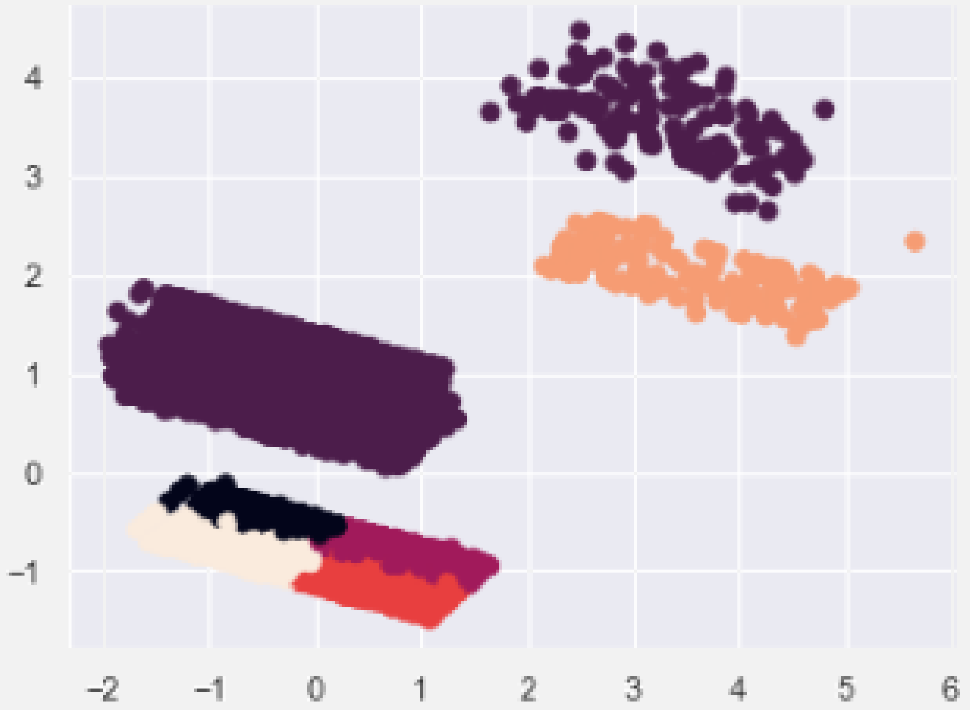
Silhouette (top) vs.DB (bottom)	Visualisation of clusters on PCA PC1 and PC2
	

Modeling options

Model optimisation

Spectral Clustering

Dataset	RFMS (frac 10%)
Transfo features	QuantileTransformer
Algo clustering	Spectral Clustering
Silhouette – nb clusters	Best cluster = 9
DB – nb clusters	Best cluster = 6

Silhouette (top) vs.DB (bottom)	Visualisation of clusters on PCA PC1 and PC2
	

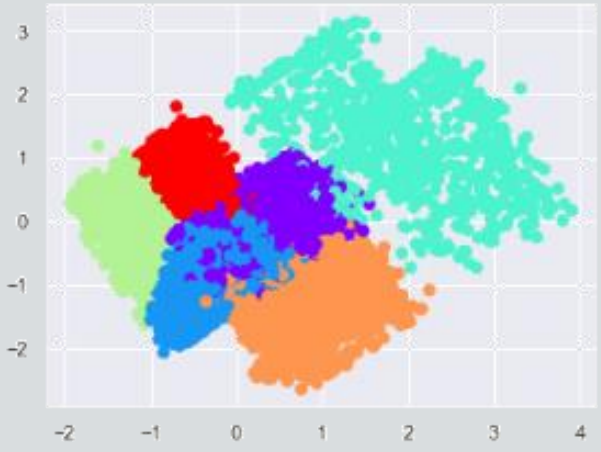
Final Model

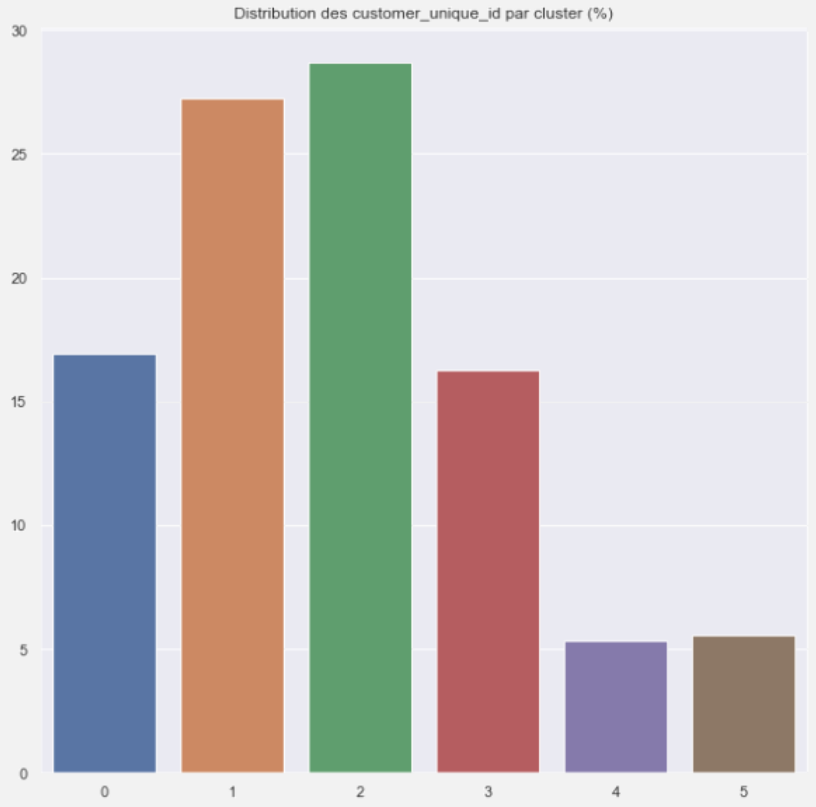
Final Model

Final model

Dataset	Score, Delivery_per, Freight_per, Quantity
Transfo features	QuantileTransformer
Algo clustering	Kmeans
Nb clusters	Silh 9 , DB 8, choice of 6 for better interpretation

Observations

- Each Customer cluster represents between 5 and 28% of customers
- Visualisation of clusters on PCA PC1 and PC2
- 

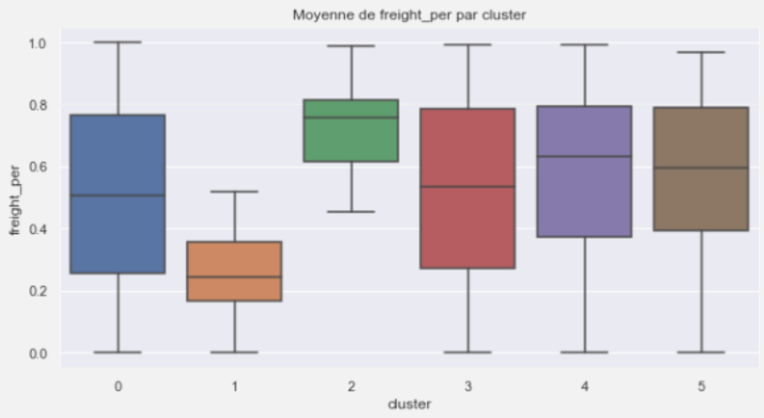


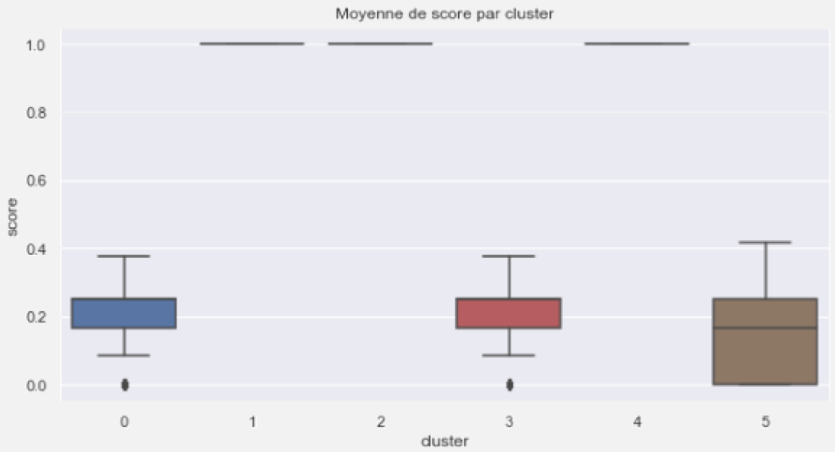
Final Model

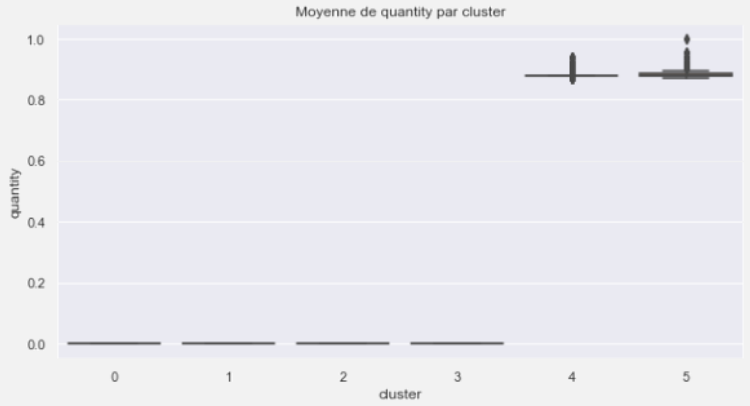
Customer segmentation

Customer segmentation

#	Segment
0	Dissatisfied customers – late delivery
1	Satisfied customer – small spender
2	Satisfied customer – big spender
3	Dissatisfied customers – due to the product?
4	Satisfied customer – large consumer
5	Dissatisfied customers – large consumer





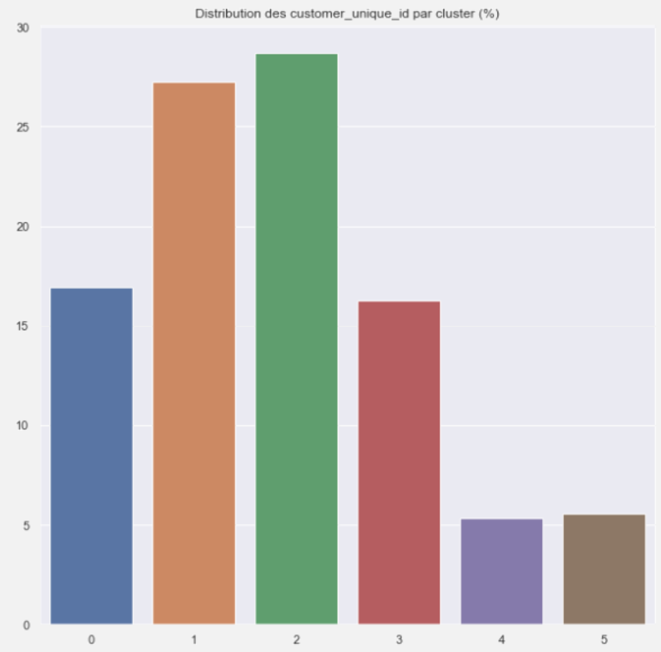


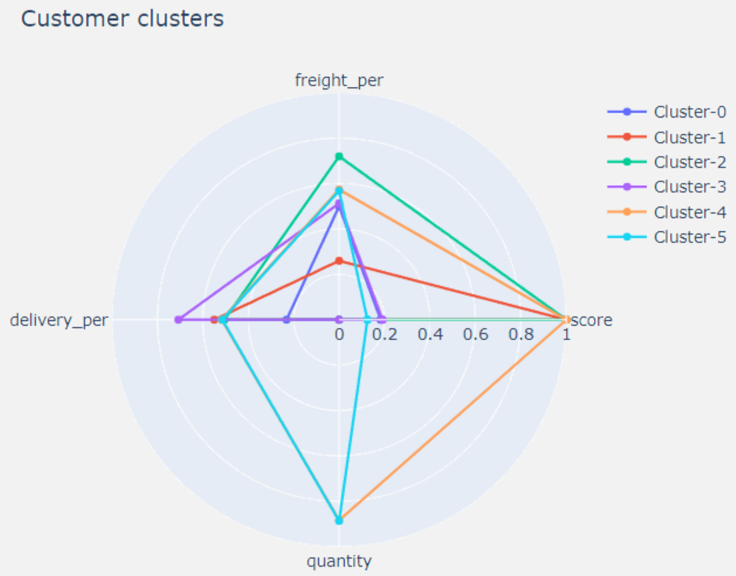
Final Model

Segment analysis

Marketing actions

C#	Segment	Marketing action
0	Dissatisfied customers – late delivery	Offer discounts
1	Satisfied customer – small spender	Offer other cheap products
2	Satisfied customer – big spender	Offer other expensive products
3	Dissatisfied customers – due to the product?	Dissatisfaction survey
4	Satisfied customer – large consumer	No action. Represents a minority
5	Dissatisfied customers – large consumer	No action. Represents a minority

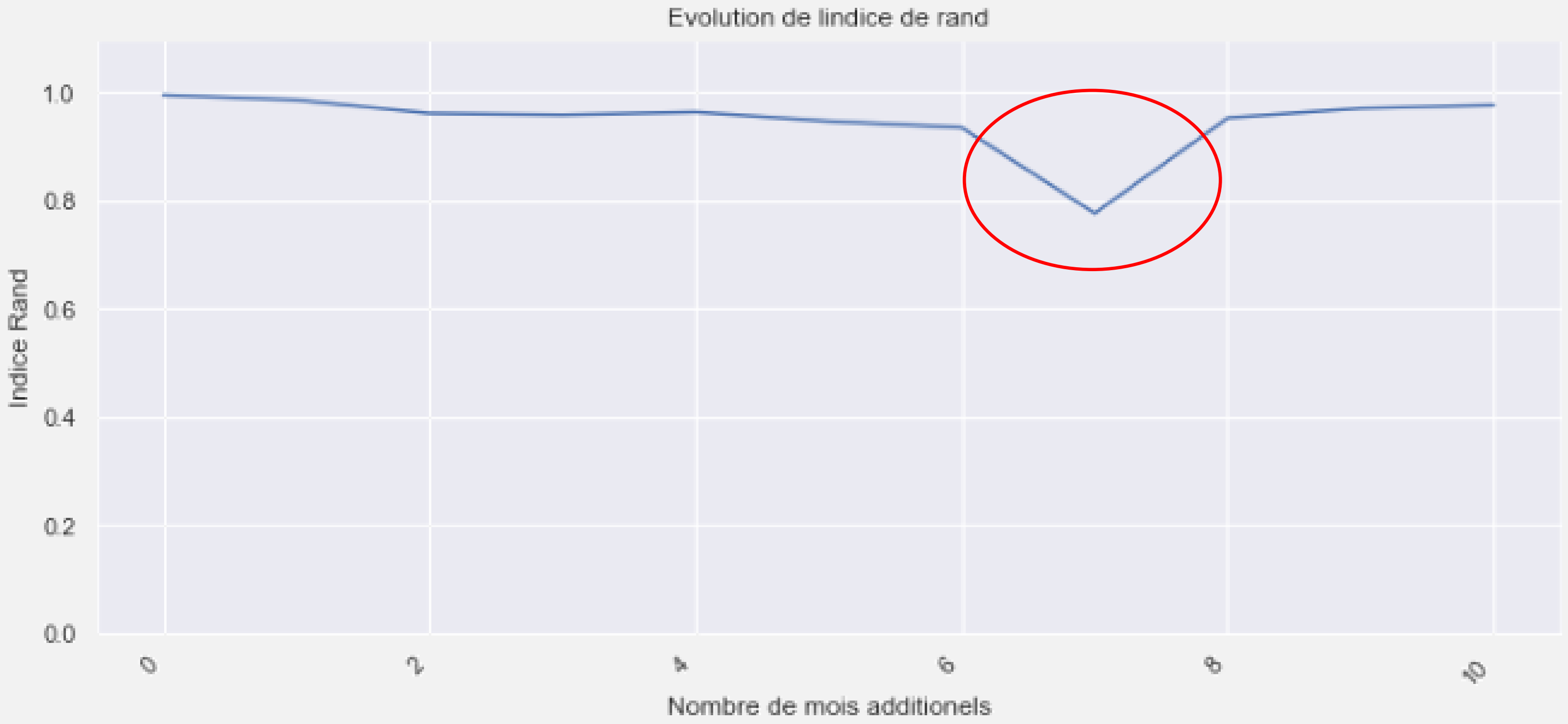




Final Model

Maintenance time

Contract maintenance



Observations

- Calculation of the Rand score Adjusted for the first 12 months (baseline), then 13 months, … up to 24 months corresponding to the whole dataset.
- Proposal to revise the clustering model after 7 months (index < 0.8)

Final Model

Conclusions

Relevance and axis of improvement

Relevance of clustering

- The final unsupervised model chosen is acceptable
- It makes it possible to identify a correct segmentation of customers and to define marketing actions
- However, some visible limitations to the proposed clustering

Areas for improving clustering

- Dataset with more than one order per customer
- More data knowledge of the customer: age, gender, interests
- Further identification of the most optimal hyperparameters for each model, excluding the number of clusters
- Consideration of other variables for modeling (purchase season, seller-customer distance, rental, purchase category, ...)

Thank you for your attention!