Segmentez des clients d'un site e-commerce

Projet 5 du parcour
Data Scientist

Dernière MàJ
14 Avril 2023

Raquel Sanchez Pellicer

OPENCLASSROOMS

olist

Context

Data preprocessing

Modeling

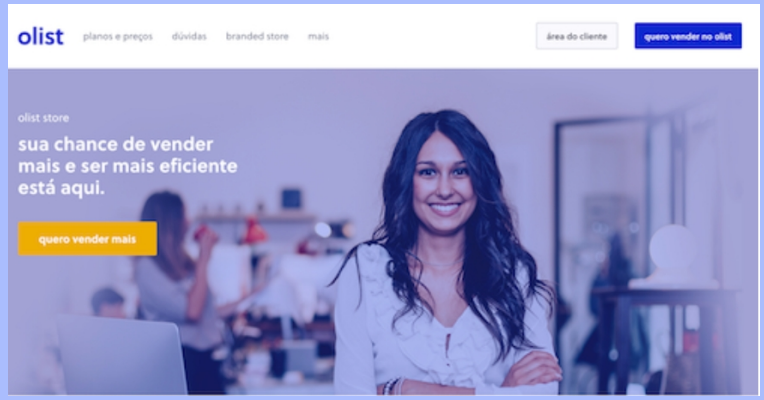
- RFM
- Unsupervised learning

Maintenance simulation

Conclusions

olist est une plateforme e-commerce du Brésil.

Met en lien les consomateurs et les vendeurs.



MISSION : fournir aux équipes d'e-commerce d'Olist une segmentation des clients qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.

OBJECTIF :

- **Comprendre** les différents **types de consommateurs** qui font des achats sur ***olist***, grâce à leur comportement et à leurs données personnelles.
    - Fournir à l'équipe marketing une **description actionnable de la segmentation**.
    - Fournir une **proposition** de **contrat de maintenance** basée sur une analyse de la stabilité des segments au cours du temps.

CONTEXT. APPROACH

olist

Données disponibles

- 9 jeux de données indépendants
- Variables communes permettant de les relier
- Nombre de lignes par client variable

EDA

- Réunir tous les données dans un seul dataset
- Axer le dataset au tour des clients : une ligne par client
- Transformation variables pour obtention d'indicateurs

Segmentation

- Analyse multidimensionnelle
- Identification de relation entre variables
- Création de groupes de clients homogènes

Interprétation

- Connaissance consommateurs améliorée
- Segmentation
- Description actionnable

Simulation maintenance

- Analyse évolution groupes de clients avec le temps
- Identification période actualisation

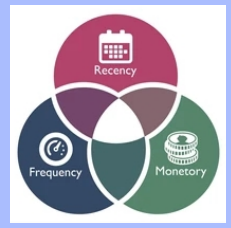
Fusion

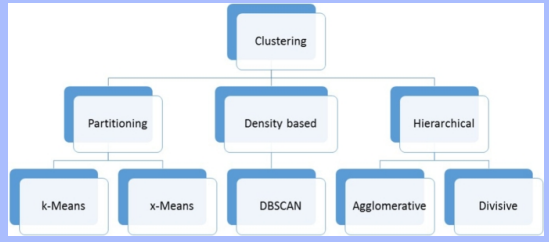
Transformation des variables

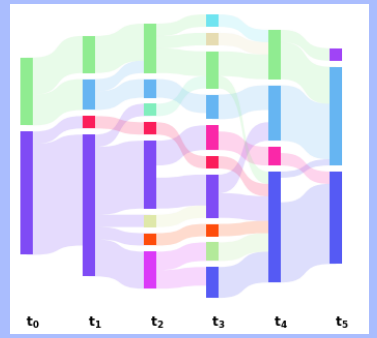
Encodage

RFM analyses

cluster analyses







olist

Context

Data preprocessing

Modeling

- RFM
- Unsupervised learning

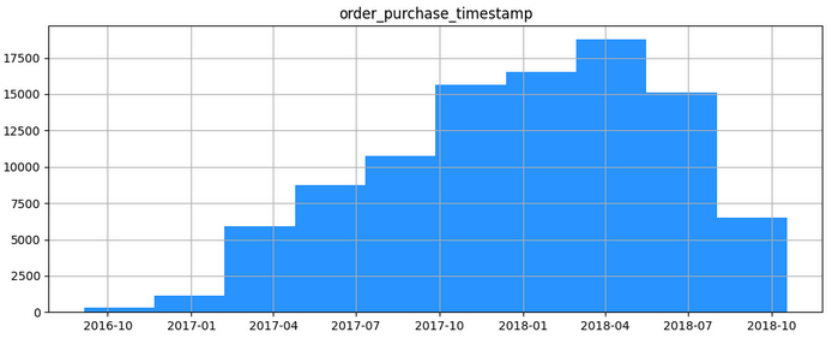
Maintenance simulation

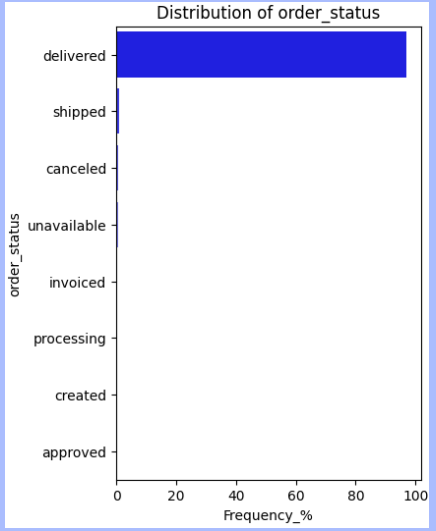
Conclusions

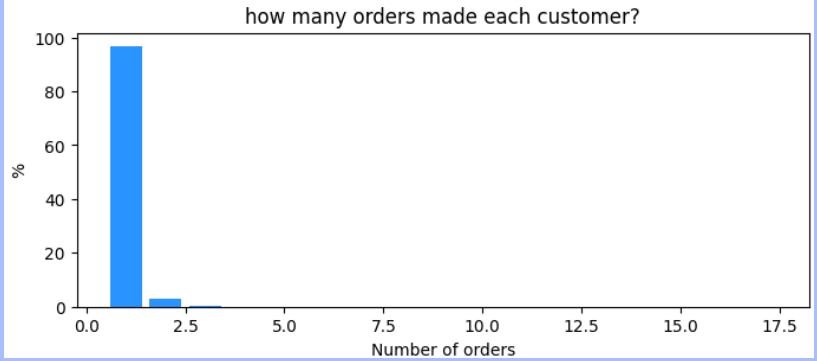
DATA PREPROCESSING. EDA

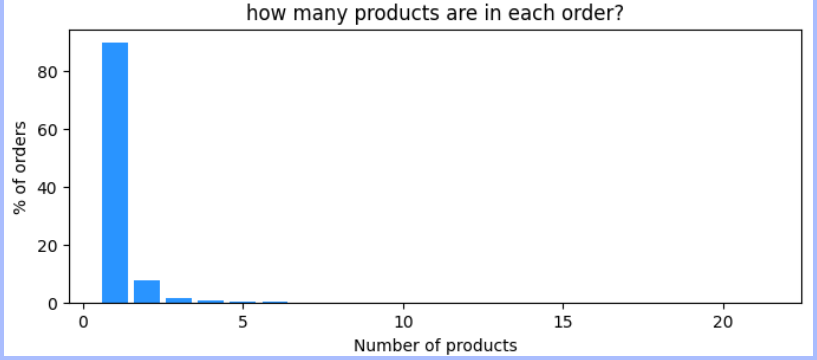
olist

Commandes

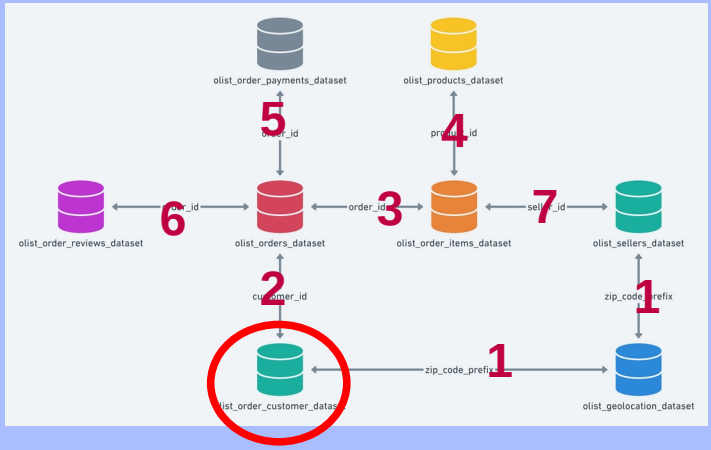






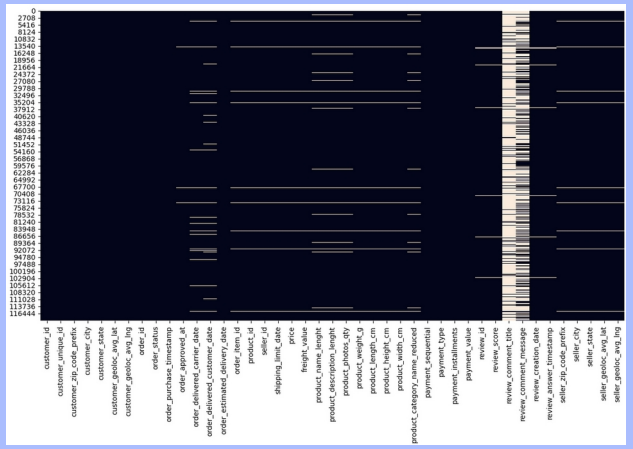


DATA PREPROCESSING. DATA CLEANING & FEATURE ENGINEERING



1. Fusion des jeux de données.

Préparation préalable de certaines colonnes (zip code, catégorie).
Identification et suppression outliers (geoloc)



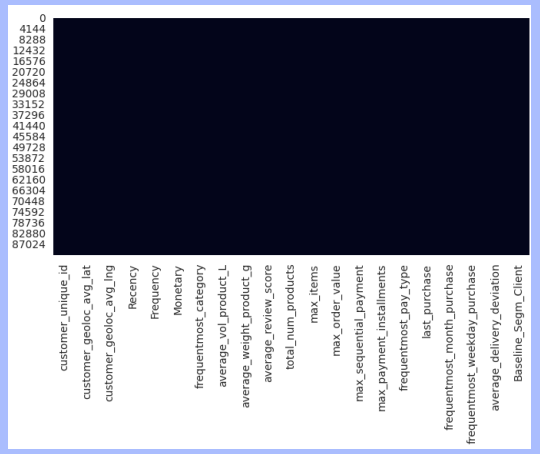
115529 lignes ; 42 colonnes

2. Nettoyage

Suppression variables non pertinentes (variables clés pour fusionner les datasets et information vendeurs).
Suppression **lignes avec données manquantes**.
Réduction dataset **commandes livrées** sans erreurs de date.

3. Feature engineering

Transformation variables pour axer le dataset sur les clients et obtenir des indicateurs.
Détails Annexe A.



91160lignes ; 19 colonnes

olist

Context

Data preprocessing

Modeling

- **RFM**
- **Unsupervised learning**

Maintenance simulation

Conclusions

MODELING. RFM APPROACH

olist

RANKING APPROACH

$$RFM_{score} = R_{(rank-norm)} \times 0.28 + F_{(rank-norm)} \times 0.15 + M_{(rank-norm)} \times 0.57$$

	customer_unique_id	Recency	Frequency	Monetary	R_rank_norm	F_rank_norm	M_rank_norm	RFM_Score	Customer_segment
1542	04544a89f79714c0abcc3ae49e884209	580	1	15.71	0.78	42.49	0.75	0.62	Lost Customers
22624	3f4f614c632af7fc7508462a7cb55ac2	695	1	18.62	0.01	42.49	0.99	0.62	Lost Customers
60784	aa811dd4c5b1ca28de6db905ec51ad5e	601	1	17.62	0.29	42.49	0.91	0.62	Lost Customers
10114	1c259d4575f154dccfefa0576b0ca987	598	1	16.62	0.31	42.49	0.82	0.62	Lost Customers
9699	1b10d58aef8d0d26651c02d40f765cae	574	1	15.29	1.07	42.49	0.72	0.62	Lost Customers

Rank → Normalize → Score → Rate

Rating Customer based upon the RFM score

- rfm score >4.5 : Top Customer
- 4.5 > rfm score > 4 : High Value Customer
- 4>rfm score >3 : Medium value customer
- 3>rfm score>1.6 : Low-value customer
- rfm score<1.6 :Lost Customer

✗ Les sous groupes sont trop héterogènes

✗ Seulement la variable « monetary » est bien captée

QUANTILE CODING APPROACH

	customer_unique_id	Recency	Frequency	Monetary	R_quartile	F_quartile	M_quartile	RFM_class
0	0000366f3b9a7992bf8c76cfdf3221e2	112	1	141.90	4	1	3	413
1	0000b849f77a49e4a4ce2b2a4ca5be3f	115	1	27.19	3	1	1	311
2	0000f46a3911fa3c0805444483337064	537	1	86.22	1	1	2	112
3	0000f6ccb0745a6a4b88665a16c9f078	321	1	43.62	2	1	1	211
4	0004aac84e0df4da2b147fca70cf8255	288	1	196.89	2	1	4	214

Quartile identification → Coding → Assembling

- Top/Best cusotmers = 444
- Lower/Lost cusotmers = 111

Sur le 64 classes possibles classes, 29 sont présentes. 13 des 29 classes inclus moins de 4 clients.

✗ Trop de classes différentes

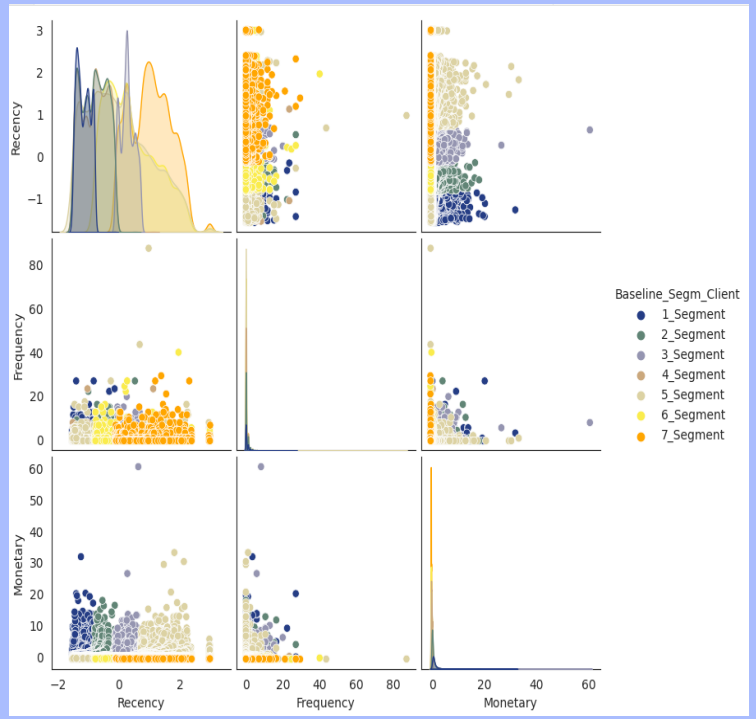
✗ Classes non différentiables

MODELING. RFM APPROACH

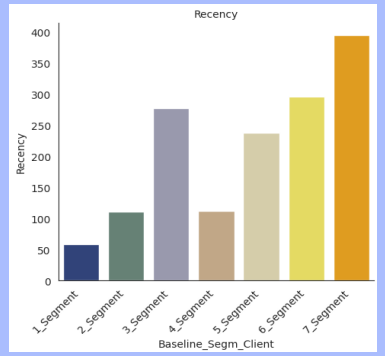
olist

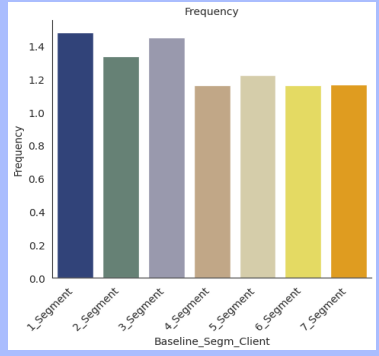
INTEGRATIVE APPROACH

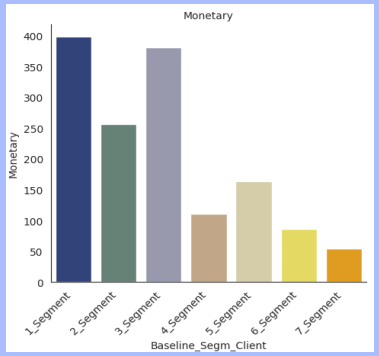
	customer_unique_id	Recency	Frequency	Monetary	RFM_Score	Customer_segment	R_quartile	F_quartile	M_quartile	RFM_class	RFM_class_score	Baseline_Segm_Client
27720	2255f2032f5376177c2b40499c881c84	322	1	87.84	1.90	Low Value Customers	2	1	2	212	5	6_Segment
35375	e9f1564a928d766c0e3060dc9347d84a	130	1	67.54	2.15	Low Value Customers	3	1	2	312	6	5_Segment
11325	ac1199e07a382c4b92e2bf1e0f179772	346	1	58.71	1.34	Lost Customers	2	1	1	211	4	7_Segment
90269	eca61ae704cc5ecddd7a6e558b6f04b1	152	3	1097.70	4.47	High value Customer	3	1	4	314	8	2_Segment
74627	694bde413be910bc563ab60079e097b4	436	2	336.28	3.46	Medium Value Customer	1	1	4	114	6	5_Segment
4844	e2af1e4f7b1b25b6b9edf561515cc304	295	1	35.72	1.00	Lost Customers	2	1	1	211	4	7_Segment
46188	4b787b82d4a7d50a53f918c12fc4827f	213	1	109.41	2.51	Low Value Customers	3	1	3	313	7	4_Segment



Customer Segment	RFM_class_score	Ranking classification
1	9-10	Anyone
2	8	Anyone
3	7	Top or High
4	7	Medium to Lost
5	6	Medium to Lost
6	5	Top to Lost
7	3-4	Top to Lost







MODELING. RFM APPROACH

olist

INTEGRATIVE APPROACH

Customer segment	Name	Description	Actions
Champions	Segment 1	Are the biggest spenders Show a high purchase frequency Have purchased recently.	Offer rewards Offer incentives
Loyal customers to protect	Segment 2	Have purchased high purchase frequency High monetary value orders Have placed an order in the last three months	Organize contests Offer a loyalty program
Need attention	Segment 3	The pruchased expensive items They have not purchased for a long They were frequent customers	Offer a free trial Make them feel special
New customers	Segment 4	Have placed an order in the last three months Low spenders Mostly single buyers	Provide welcome assistance Offer them discounts Build a relationship
Endangered	Segment 5	They placed low price orders Bad recency values They are not frequent customers	Offer credit Offer a wish list
Sleeping customer	Segment 6	Low spenders They have not purchased for a long Mostly single buyers	Understand them Do one last promotion
Lost customers	Segment 7	Worst recency values Worst monetary values Mostly single buyers	Forget them

MODELING. UNSUPERVISED LEARNING

olist

Data preprocessing

Sous-échantillons :
- RFM
- Quantitative
- Quanti&Quali (encodage)

Scaling

Réduction de dimensions
Optimisation hyperparamètres
- ACP
- Isomap
- t-SNE

Tuning cluster hypermaramètres

Sélection configurations Algorithme + hyperparamètres

Analyse coefficients

Silhouette
Calinsky-Harabasz
Temps calcul

Test hyperparamètres

Entraînement algorithmes

K-Means
DBScan
Hierarchical,
Wards method

Modeling

Entraînement configuration

Étude préliminaire

Homogénéité taille clusters Distribution discriminante

Stabilité d'initialisation

Sélection modèles

Évaluation de la robustesse

Sélection Données Scaler Algorithme Hyperparamètres

Analyse interprétabilité

Analyse homogénéité taille clusters

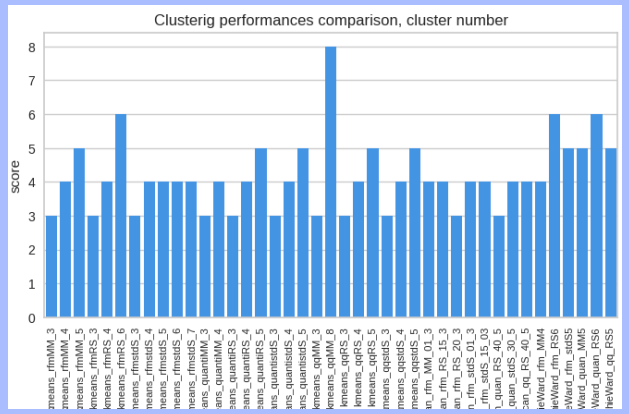
Analyse distribution discriminante

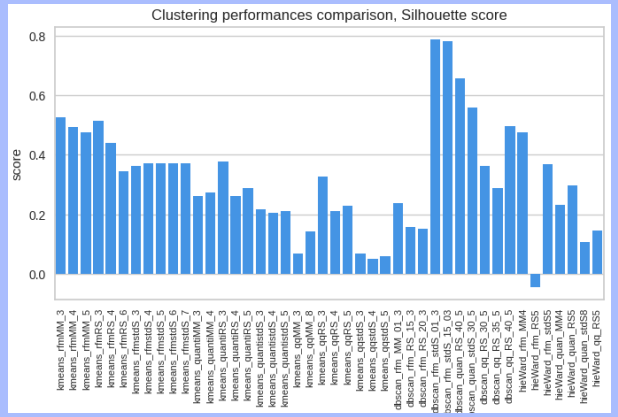
Étude approfondie

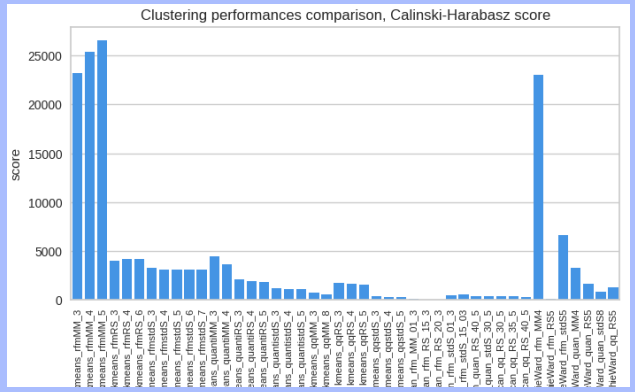
MODELING. UNSUPERVISED LEARNING

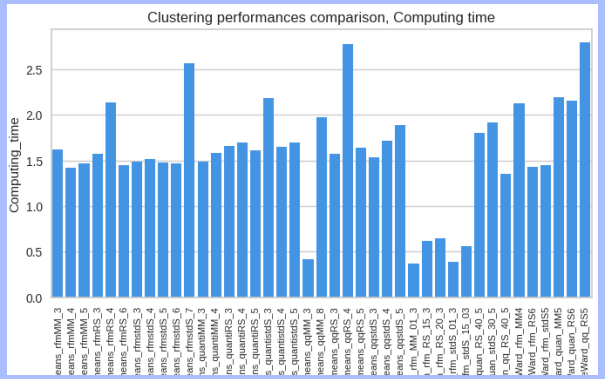
olist

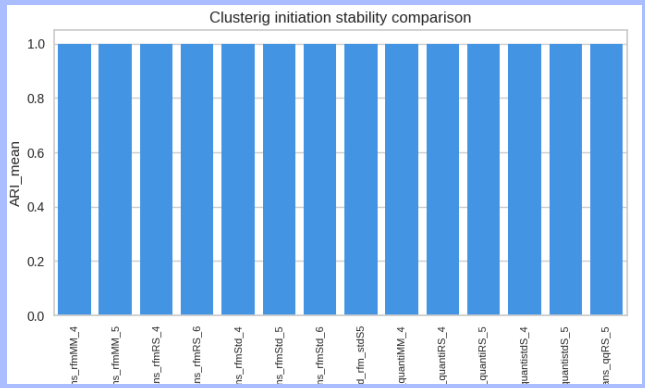
Comparaison performances











MODELING. UNSUPERVISED LEARNING

olist

Data preprocessing

Sous-échantillons :
- RFM
- Quantitative
- Quanti&Quali (encodage)

Scaling

Réduction de dimensions
Optimisation hyperparamètres
- ACP
- Isomap
- t-SNE

Tuning cluster hypermaramètres

Sélection configurations Algorithme + hyperparamètres

Analyse coefficients

Silhouette
Calinsky-Harabasz
Temps calcul

Test hyperparamètres

Entraînement algorithmes

K-Means
DBScan
Hierarchical,
Wards method

Modeling

Entraînement configuration

Étude préliminaire

Homogénéité taille clusters Distribution discriminante

Stabilité d'initialisation

Sélection modèles

Évaluation de la robustesse

Sélection Données Scaler Algorithme Hyperparamètres

Analyse interprétabilité

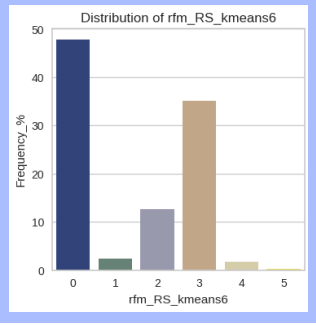
Analyse homogénéité taille clusters

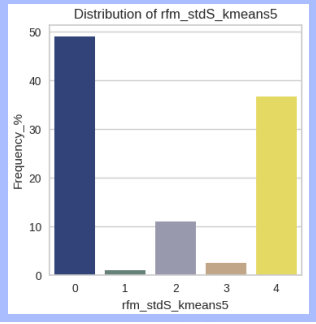
Analyse distribution discriminante

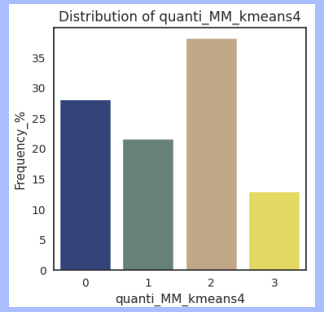
Étude approfondie

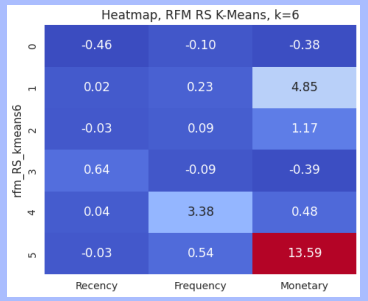
olist

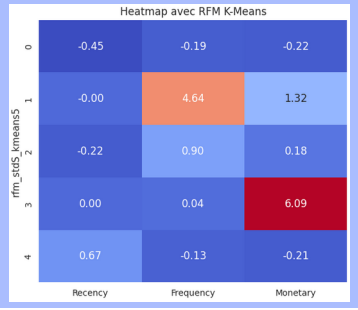
Évaluation robustesse

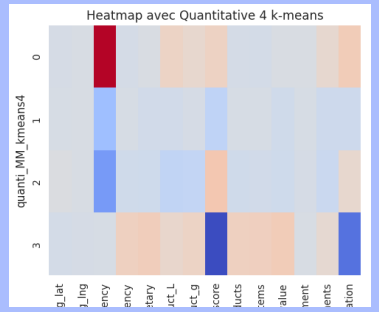


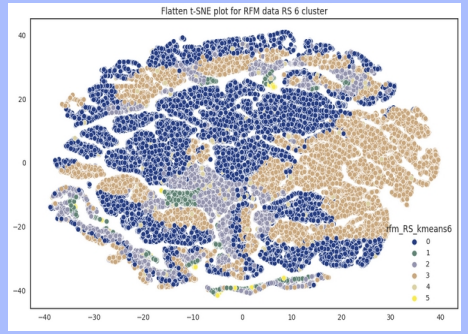


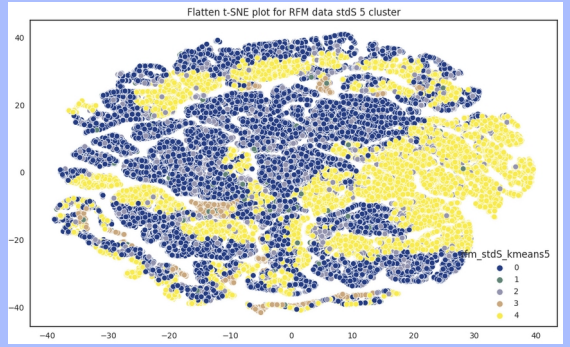


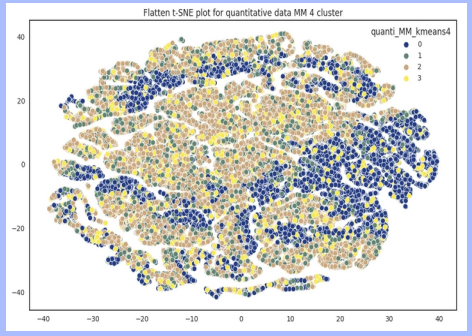








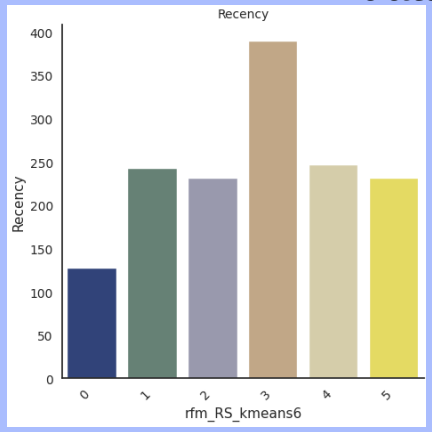


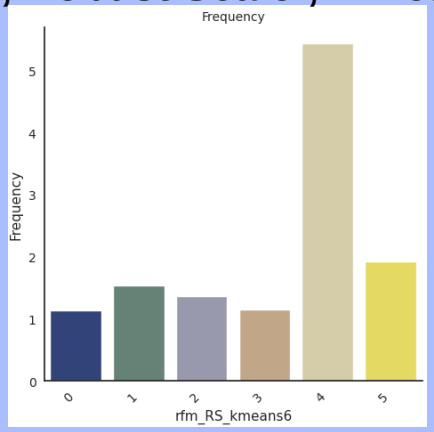


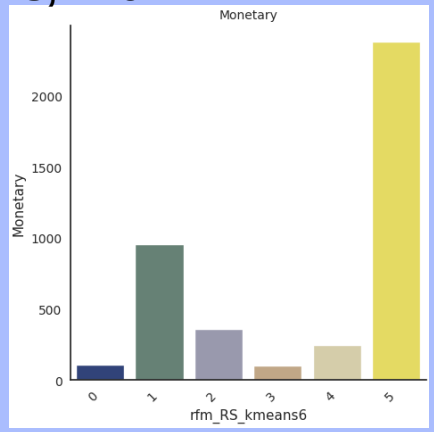
olist

Évaluation interprétabilité

RFM data, Robust Scaler, k-Means, k=6

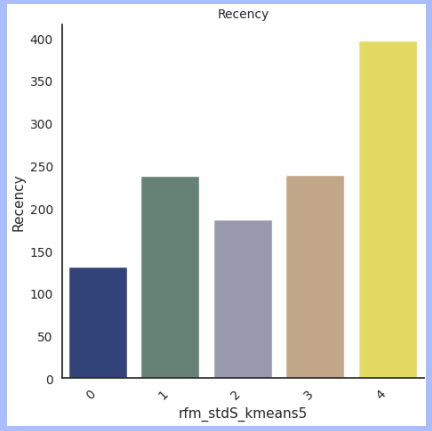


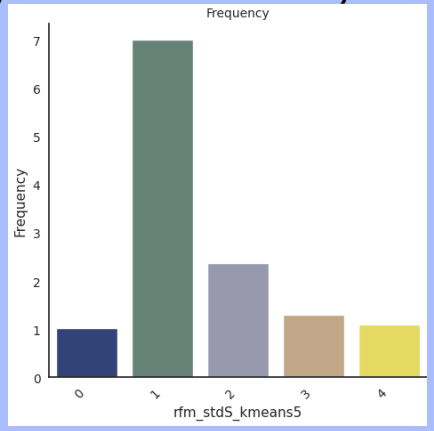


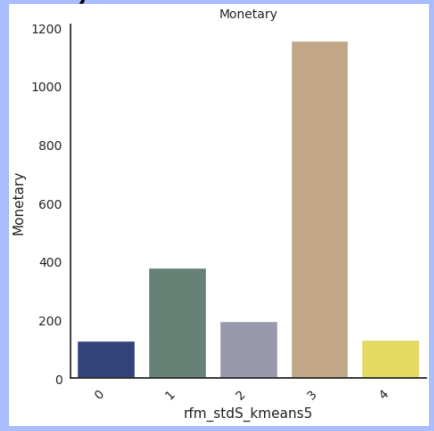


- The **size** of the clusters is **uneven**.
- **All** clusters can be **discriminated** according to the **RFM** variables.
- Visualization of the distribution of the elements of each cluster using **t-SNE** reveals that:
  - the elements in clusters **0 to 3** are **well** grouped and **discriminable**
  - the elements in clusters **4 and 5 are mixed** in the visualization

RFM data, Standard Scaler, k-Means, k=5







- The **size** if the clusters is **uneven**.
- The analysis of the **features mean values** by cluster reveals that **not all clusters** can be **discrimitated**.
- Visualization of the distribution of the elements of each cluster using **t-SNE** reveals that:
  - the elements in clusters **3 and 4** are well grouped and **discriminable**
  - the elements in clusters **0, 1 and 2** are **mixed** in the visualizations

MODELING. RFM Robust Scaled, k-Means, k=6

olist

Customer segment	Cluster number	Description	Actions
Champions	5	Highest expenses Frequency above the mean Mean values of recency Buy more than one product per order	***Offer incentives:*** *Offering incentives to loyal customers encourages them to buy more.* ***Offer rewards***: Give them one of your most popular products.
Potential loyalist	1	Second highest expenses Frequency above the mean Mean values of recency	***Organize contests***: Keep those customers in the loop by organizing contests with attractive prizes. ***Offer a loyalty program***: Introduce them to a loyalty program or perhaps an elite club membership so you don't lose them.
Promising customers	4	Highest frequency Mean to low expenses Mean values of recency	***Offer a free trial:*** e.g. premium features and let them decide whether to purchase after the expiration period. ***Make them feel special***: Connect with these customers by wishing them a birthday, etc.
New customers	0	These are the customers who have most recently purchased Lowest expenses Their frequency is low	***Provide welcome assistance*** ***Offer them discounts:*** Welcome your new customers by offering them discount points or coupons on their next purchase to keep them coming back. ***Build a relationship:*** Offer to help at every step.
Need attention	2	They haven't placed any orders for quite some time Frequency slightly above the mean Expenses slightly above the mean	***Offer credit :*** This type of offer makes the offer tangible and customers feel it is worthwhile. ***Offer a wish list***: Run a sale based on their wish list
Lost customers	3	These are the customers who have not purchased for the longest time Lowest expenses Their frequency is low	***Understand them*** : Research your lost customers to develop a strategy to bring them back. ***Do one last promotion :*** for example, "We miss you. Here's your gift: get 25% off your entire order."

olist

Context

Data preprocessing

Modeling

- RFM
- Unsupervised learning

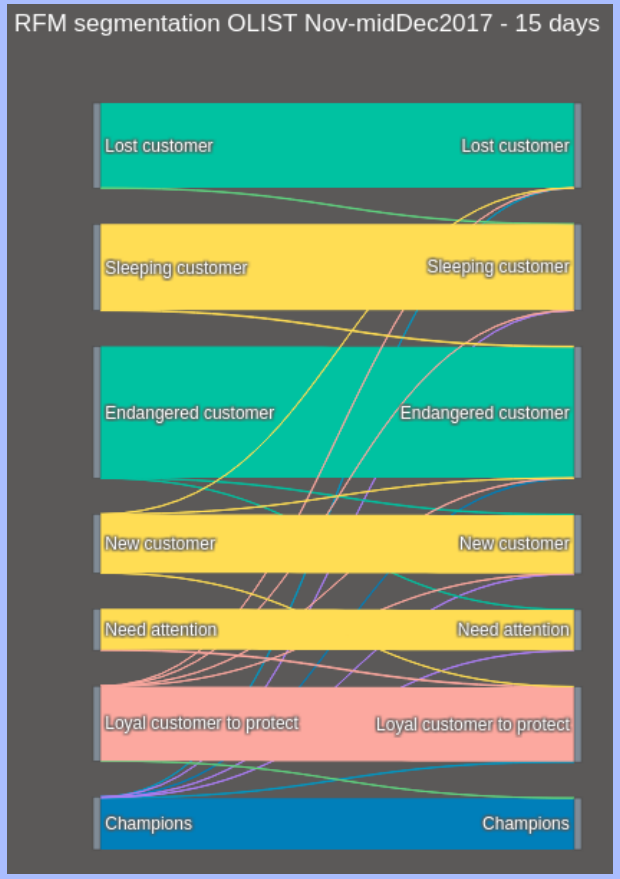
Maintenance simulation

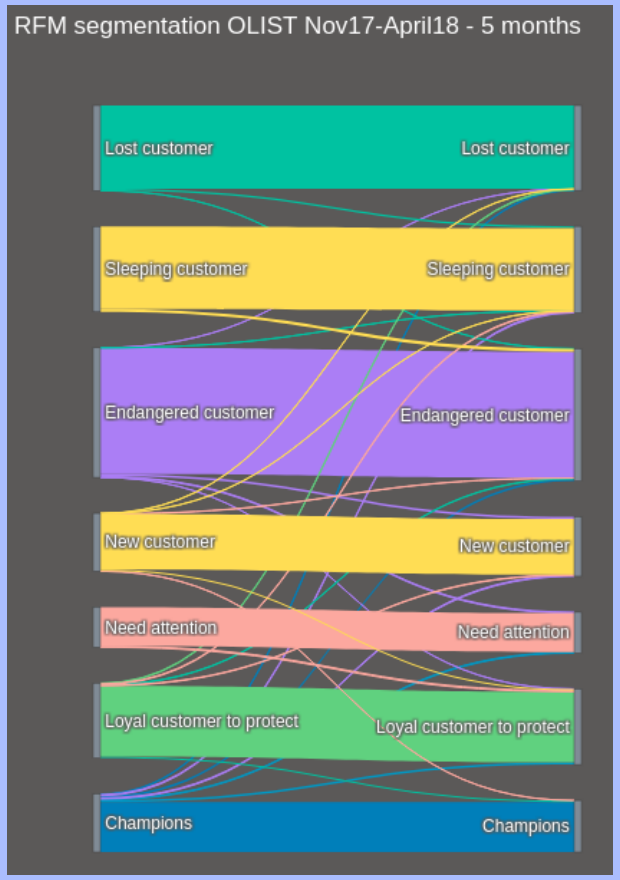
Conclusions

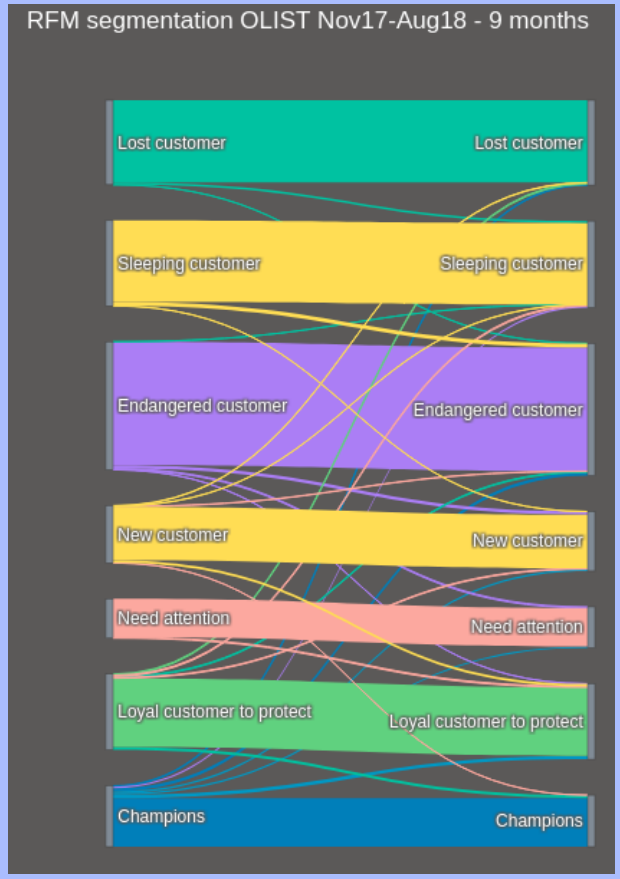
MAINTENANCE SIMULATION. RFM APPROACH

INTEGRATIVE APPROACH

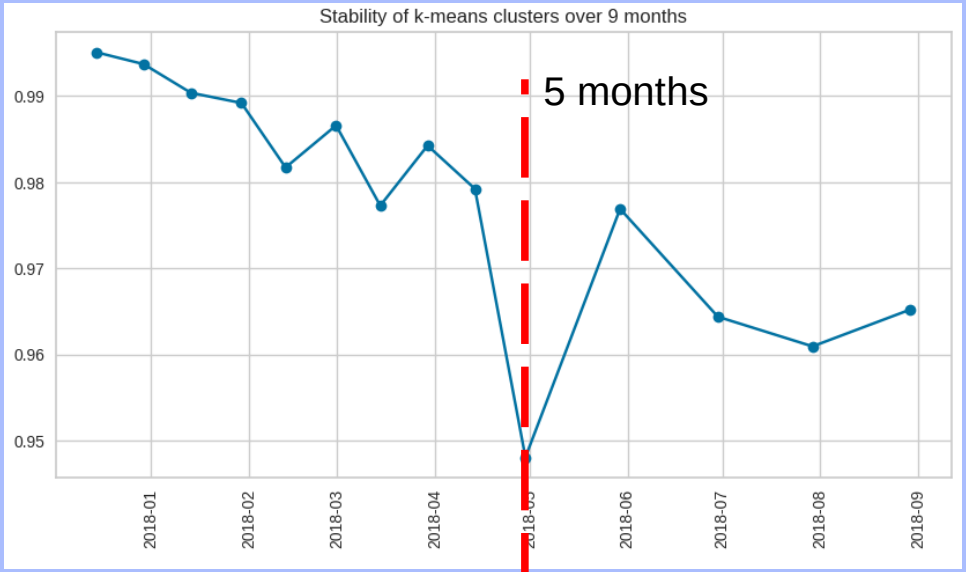
Period	ARI
15d_Nov_midDecember	0.985330
1_month_Nov_December	0.981552
2_month_Nov_January	0.960526
3_month_Nov_February	0.954867
5_month_Nov_April	0.923830
6_month_Nov_May	0.909313
9_month_Nov_Aug	0.893506







MAINTENANCE SIMULATION. RFM Robust Scaled, k-Means, k=6



Period	Date	ARI
15_days_Nov_midDecember	2017-12-14 15:00:37+00:00	0.995073
1_month_Nov_December	2017-12-29 15:00:37+00:00	0.993694
45_days_Nov_midJanuary	2018-01-13 15:00:37+00:00	0.990361
2_month_Nov_January	2018-01-29 15:00:37+00:00	0.989183
75_days_Nov_midFebruary	2018-02-12 15:00:37+00:00	0.981696
3_month_Nov_February	2018-02-28 15:00:37+00:00	0.986576
105_days_Nov_midMarch	2018-03-14 15:00:37+00:00	0.977264
4_month_Nov_March	2018-03-29 15:00:37+00:00	0.984239
135_days_Nov_midApril	2018-04-13 15:00:37+00:00	0.979217
5_month_Nov_April	2018-04-29 15:00:37+00:00	0.948048
6_month_Nov_May	2018-05-29 15:00:37+00:00	0.976891
7_month_Nov_June	2018-06-29 15:00:37+00:00	0.964366

- Analysis does not show important changes
- If we consider the analysis valid no maintenance is needed
- **BUT !** The dataset is heavily biased towards customers who have made only one purchase (frequency =1)
- It is necessary to **monitor the evolution of the database structure**, if changes are observed in the type of customers (for example, increase of customers who make more than one purchase) it will be necessary to make an update.
- After five months, the exchange rate curve becomes more unstable.
- An **update at five months is recommended**.

olist

Context

Data preprocessing

Modeling

- RFM
- Unsupervised learning

Maintenance simulation

Conclusions

CONCLUSIONS

olist

Model	Advantages	Inconvénients
RFM integrative	Simple à mettre en œuvre. Rapide. Marketing traditionnel.	3 variables seulement prise en compte. Tout refaire pour l'ajout de nouveaux clients.
k-Means	Toutes les variables numériques peuvent être prises en compte. Entraînement rapide et stable à l'initialisation. Segments assez homogènes, interprétables et actionnables. Nouveaux clients pris en compte facilement (.predict).	Lent. Paramétrage assez difficile. Paramètre k à initialiser.

Modèle final :

k-Means avec **6** clusters, données **RFM** mises à l'échelle avec **Robust Scaler**, est une bonne piste de départ pour démarrer la segmentation de clientèle en partant d'un contrat de maintenance avec mise à jour au but de **5 mois**.

olist

Possibles pistes d'amélioration :

Jeu de données

Biaisés :

- 96% des clients ne commandent qu'une seule fois.
- Notes toutes très positives.

Nécessite plus de données :

- démographiques (âge, profession, sexe, nombre d'enfants..)
- psychographiques (avis sur le produit, centre d'intérêt…)

Segmentation

Explorer des nouvelles transformations sur les variables (log)

Explorer des nouvelles variables (feature engineering)

Collaborer avec l'équipe Marketing métier :

- Valider les regroupements des catégories de produits.
- Valider le choix des variables ajoutées lors du feature engineering.
- Définir la finesse du nombre de segments souhaités par OLIST.
- Valider les premiers résultats (modifier le paramétrage/modèle si besoin).

Annexes

ANNEXE A. Transformation variables

olist

Nouvelle variable	Variable d 'origine	Action	Explication
Recency	order_purchase_timestamp	{"order_purchase_timestamp": lambda x: (date_last_purchase - x.max()).days}	Nombre de journées depuis le dernier achat (par rapport au dernier achat enregistré dans la base de données)
Frequency	order_id	'order_id' : 'count'	Fréquence d'achat sur l'historique
Monetary	payment_value	payment_value: 'sum'	Montant total des achats sur l'historique

ANNEXE A. Transformation variables

olist

Nouvelle variable	Variable d 'origine	Action	Explication
average_vol_product_L	"product_length_cm" "product_height_cm" "product_width_cm"	'reduced_data_olist["product_length_cm"] * reduced_data_olist["product_height_cm"] * reduced_data_olist["product_width_cm"] / 1000' : 'mean'	Volume moyen des produits achetés
frequentmost_category	product_category_name_reduced	'product_category_name_reduced': lambda x: x.mode()[0]	Catégorie de produits la plus acheté
average_weight_product_g	product_weight_g	'product_weight_g': 'mean'	Poids moyen des produits
average_review_score	review_score_avg	'review_score_avg' : 'mean'	Note moyenne des avis
total_num_products	product_id	'product_id' : 'count'	Nombre total de produits achetés
total_num_reviews	review_score	'review_score' : 'count'	Nombre total d'avis déposés

ANNEXE A. Transformation variables

Nouvelle variable	Variable d 'origine	Action	Explication
average_items	order_item_id	'.groupby("order_id") ["order_item_id"].max()' : 'mean'	Nombre moyen d'articles dans le panier
max_items	order_item_id	'.groupby("order_id") ["order_item_id"]' : 'max'	Nombre maximal d'articles dans le panier
average_order_value	payment_value	'.groupby("order_id") ["payment_value"].max()' : 'mean'	Prix moyen du panier
max_order_value	payment_value	'.groupby("order_id") ["payment_value"]' : 'max'	Prix maximale du panier
max_sequential_payment	payment_sequential	"payment_sequential" : "max"	Nombre maximale d'échéances
max_payment_installments	payment_installments	"payment_installments" : "max"	Nombre maximale de moyens de payement
frequentmost_pay_type	payment_type	"payment_type" : lambda x: x.mode()[0]	Moyen de payement le plus frequent

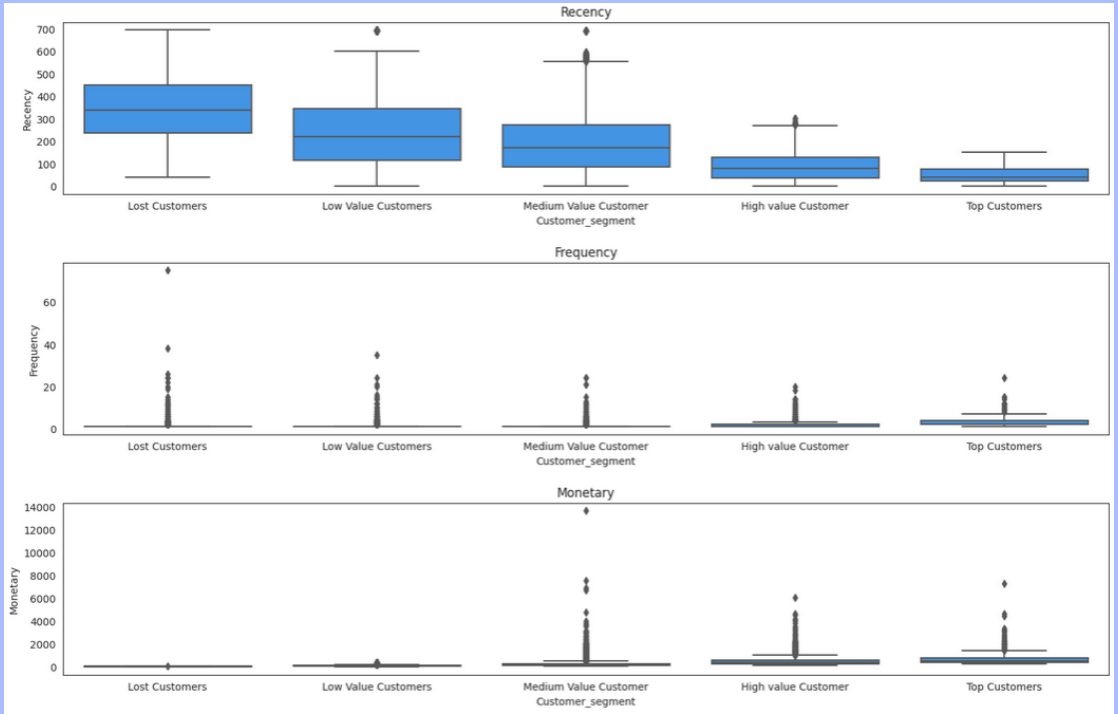
ANNEXE A. Transformation variables

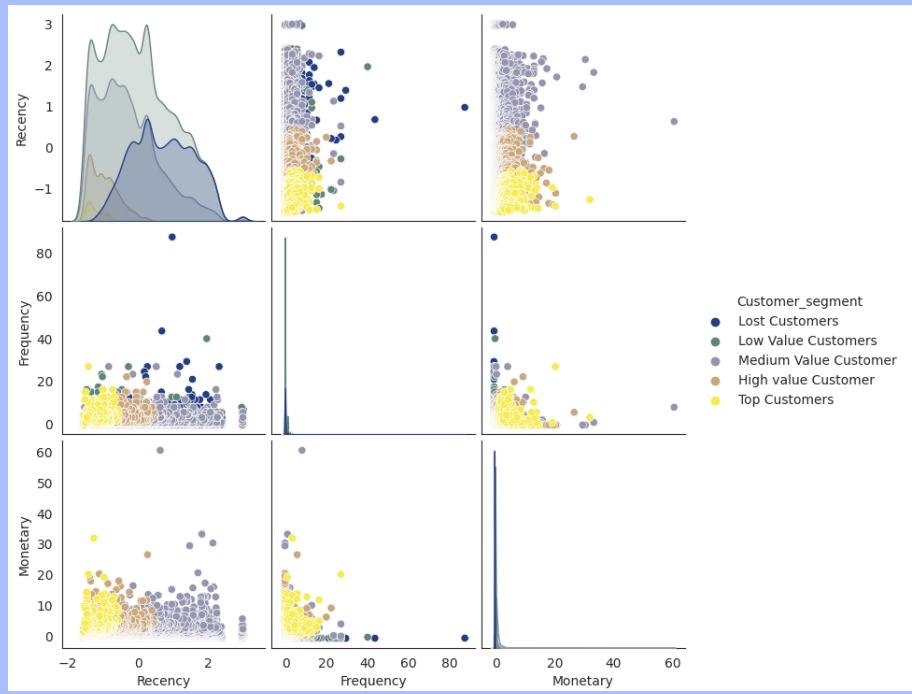
olist

Nouvelle variable	Variable d 'origine	Action	Explication
last_purchase	order_purchase_ timestamp	"last_purchase" : "max"	Date du dernier achat
frequentmost_month_ purchase	order_purchase_ timestamp	'[order_purchase_timestamp].dt. month' : lambda x: x.mode()[0]	Mois préféré pour acheter
frequentmost_weekday_ purchase	order_purchase_ timestamp	'[order_purchase_timestamp].dt. weekday: lambda x: x.mode()[0]	Jour de la semaine préféré pour acheter
average_delivery_ deviation	order_delivered_ customer_date, order_estimated_ delivery_date	'[order_delivered_customer_date] [order_estimated_delivery_date]. dt.days' : 'mean'	Déviation moyenne par rapport à la date de livraison prévue
max_delivery_ delay	order_delivered_ customer_date, order_estimated_ delivery_date	'[order_delivered_customer_date] [order_estimated_delivery_date]. dt.days' : 'max'	Retard maximum dans la livraison
max_delivery_ advance	order_delivered_ customer_date, order_estimated_ delivery_date	'[order_delivered_customer_date] [order_estimated_delivery_date]. dt.days' : 'min'	Anticipation maximale dans la livraison

MODELING. RFM RANKING APPROACH

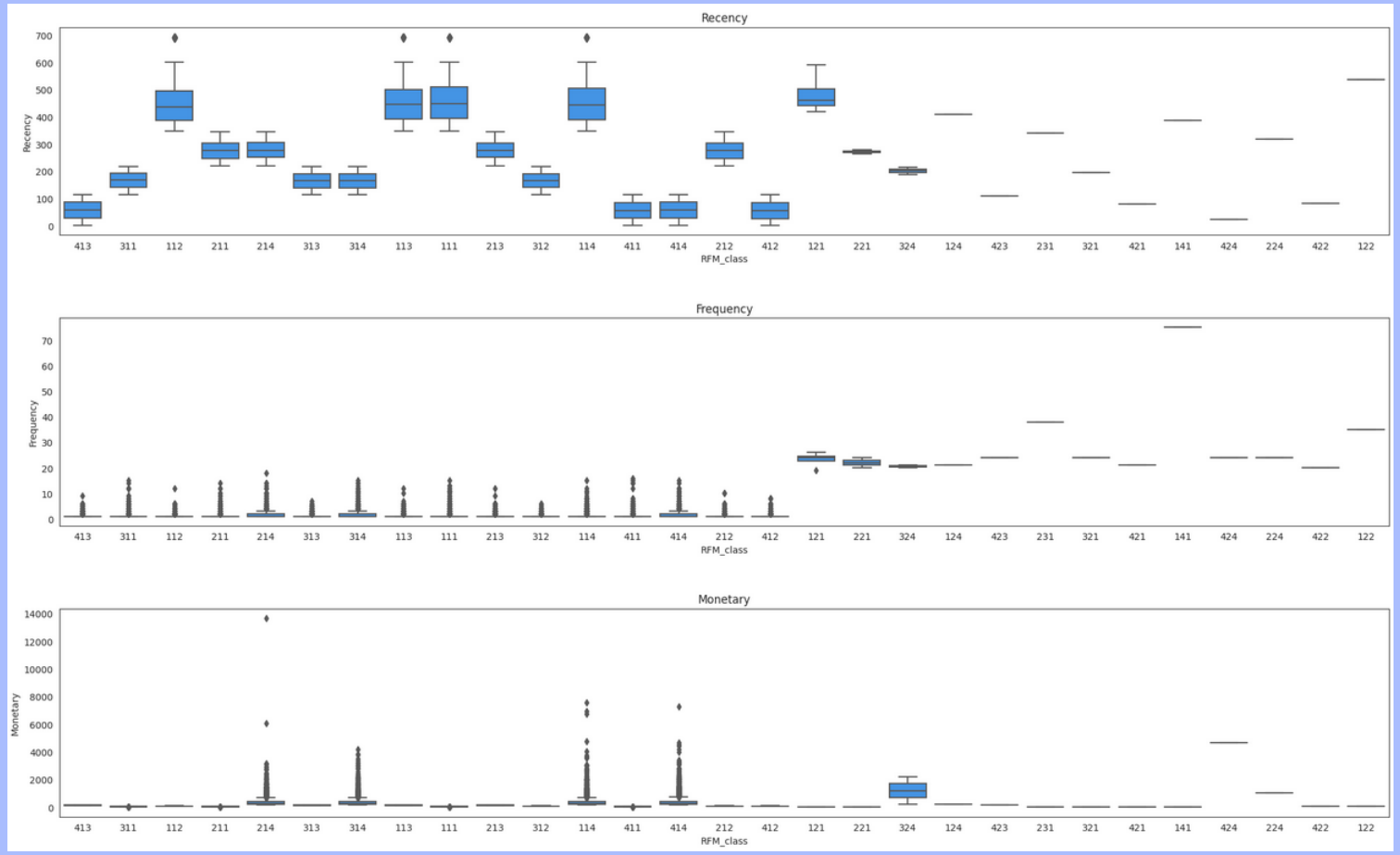
RANKING APPROACH

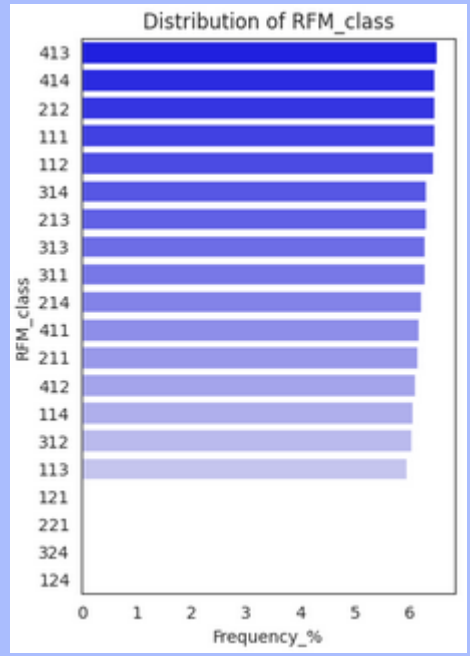




MODELING. RFM RANKING APPROACH

QUANTILE CODING APPROACH



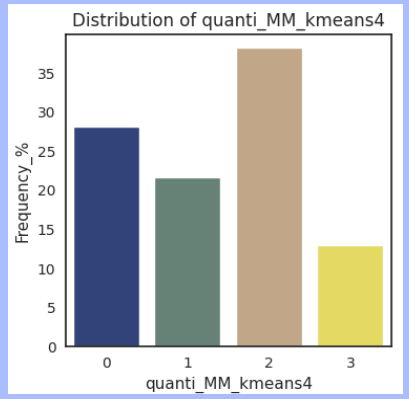


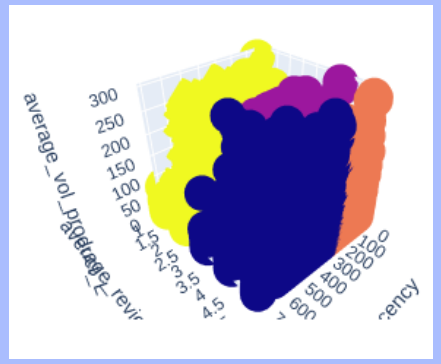
olist

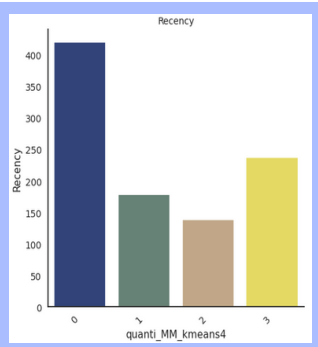
MAINTENANCE SIMULATION. TEST Quantitative data MinMax Scaled, k=4

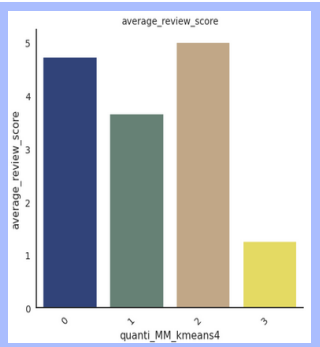
MAINTENANCE SIMULATION. TEST Quantitative data MinMax Scaled, k=4

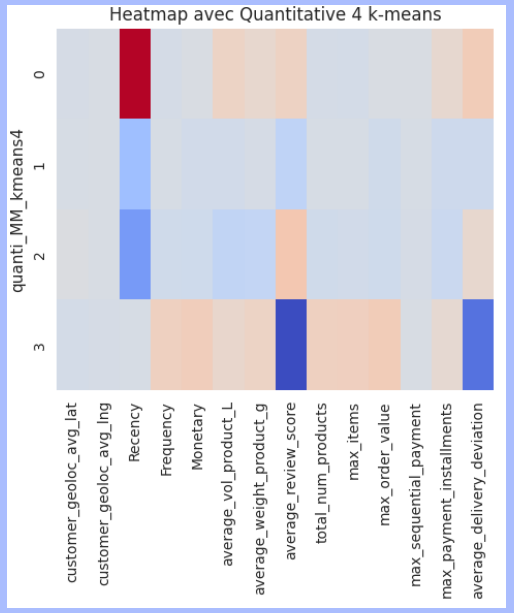
olist

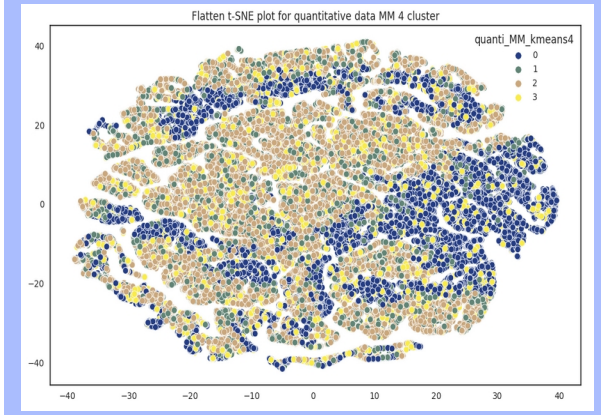


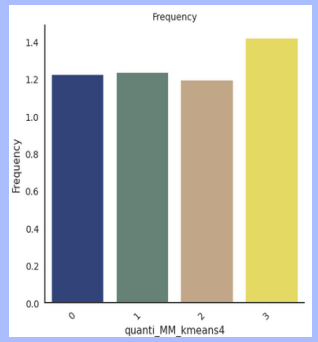


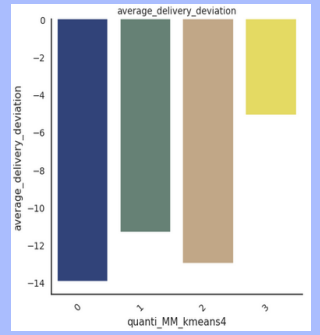


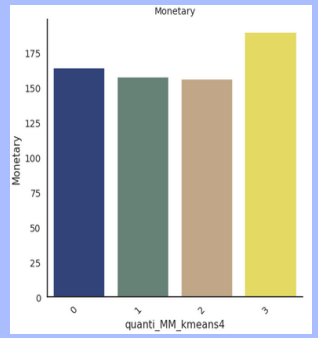




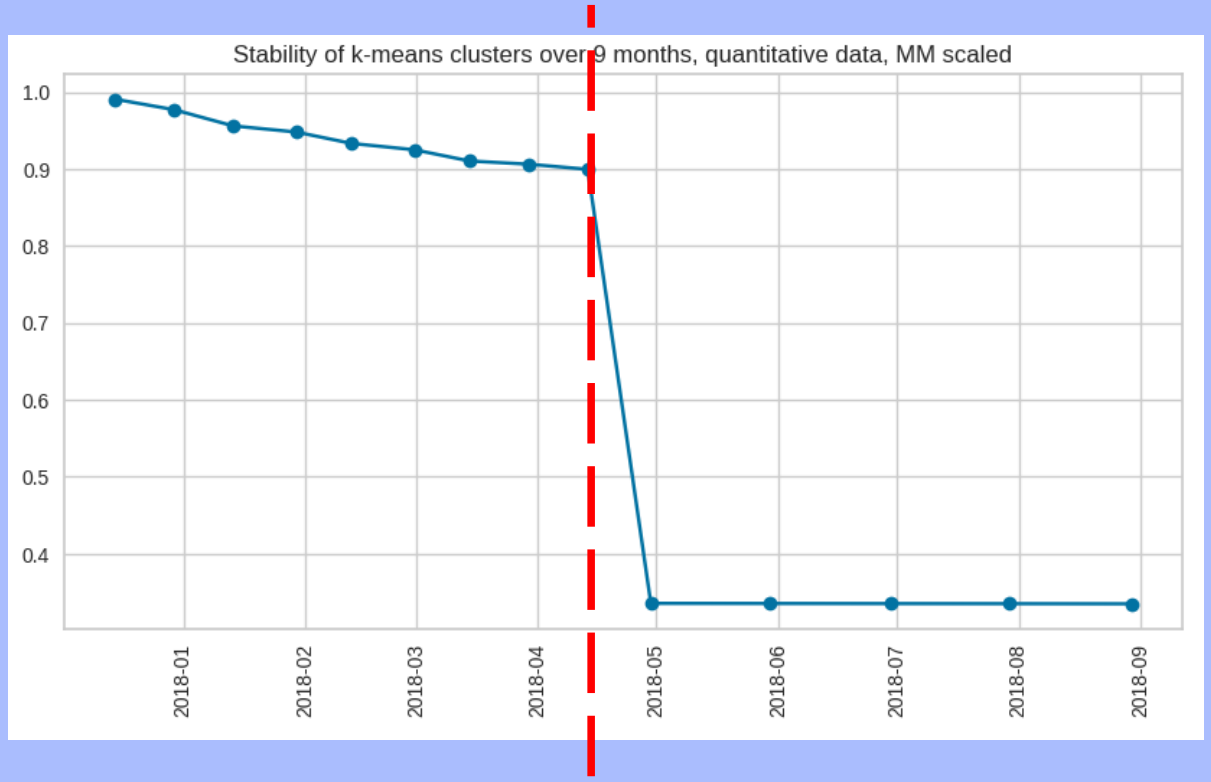








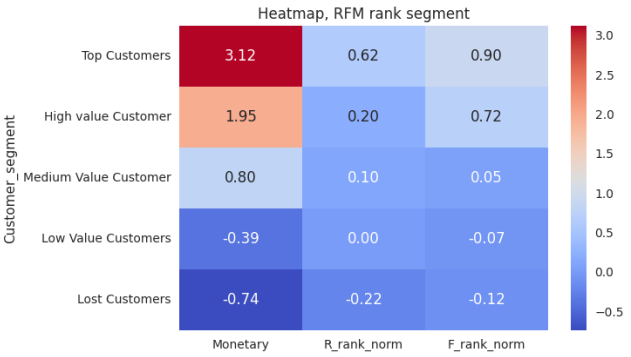
MAINTENANCE SIMULATION. TEST Quantitative data MinMax Scaled, k=4

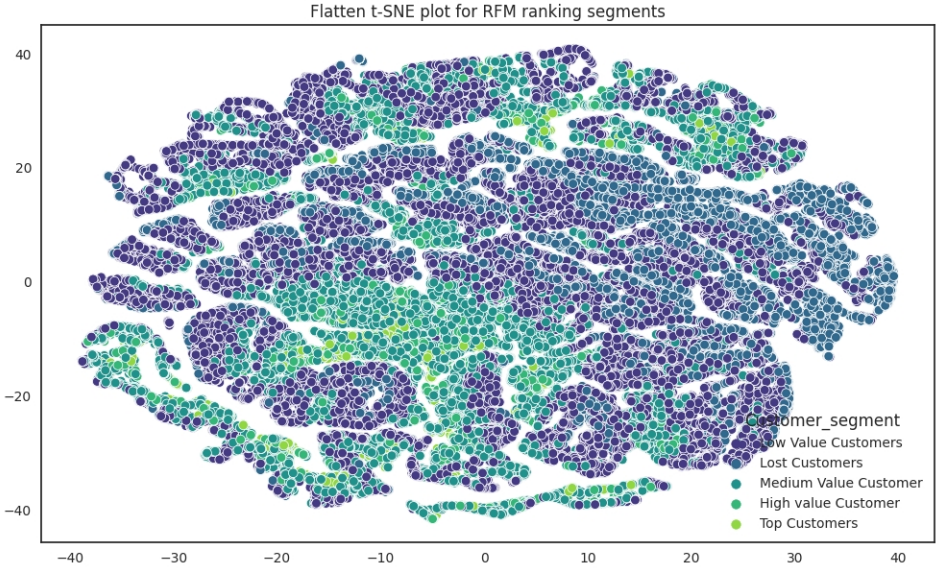


4,5 months

The evolution of the values of the variables studied in the different intervals compared has been analyzed and no significant changes have been observed.

Period	Date	ARI
15_days_Nov_midDecember	2017-12-14 15:00:37+00:00	0.990609
1_month_Nov_December	2017-12-29 15:00:37+00:00	0.976951
45_days_Nov_midJanuary	2018-01-13 15:00:37+00:00	0.955894
2_month_Nov_January	2018-01-29 15:00:37+00:00	0.947928
75_days_Nov_midFebruary	2018-02-12 15:00:37+00:00	0.933136
3_month_Nov_February	2018-02-28 15:00:37+00:00	0.924866
105_days_Nov_midMarch	2018-03-14 15:00:37+00:00	0.910285
4_month_Nov_March	2018-03-29 15:00:37+00:00	0.906143
135_days_Nov_midApril	2018-04-13 15:00:37+00:00	0.899451
5_month_Nov_April	2018-04-29 15:00:37+00:00	0.335329
6_month_Nov_May	2018-05-29 15:00:37+00:00	0.335196
7_month_Nov_June	2018-06-29 15:00:37+00:00	0.335054
8_month_Nov_July	2018-07-29 15:00:37+00:00	0.334946
9_month_Nov_August	2018-08-29 15:00:37+00:00	0.334775


Heatmap, RFM rank segment
Customer_segment
Top Customers
High value Customer
Medium Value Customer
Low Value Customers
Lost Customers
3.12
0.62
0.90
1.95
0.20
0.72
0.80
0.10
0.05
-0.39
0.00
-0.07
-0.74
-0.22
-0.12
Monetary
R_rank_norm
F_rank_norm
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5


Flatten t-SNE plot for RFM ranking segments
40
20
0
−20
−40
−40
−30
−20
−10
0
10
20
30
40
Customer_segment
Low Value Customers
Lost Customers
Medium Value Customer
High value Customer
Top Customers