Clustering des clients du site de e-commerce olist

Open Classroom

Benjamin Bouchard

1- Clustering des clients du site olist: travail préparatoire

2 - Clustering des clients du site olist: modélisation

3- Clustering des clients du site olist: interprétation des clusters

4- Clustering des clients du site olist: plan de maintenance

1- Clustering des clients du site olist: travail préparatoire

- Introduction
- Données olist
- Feature engineering
- Exploration

Introduction

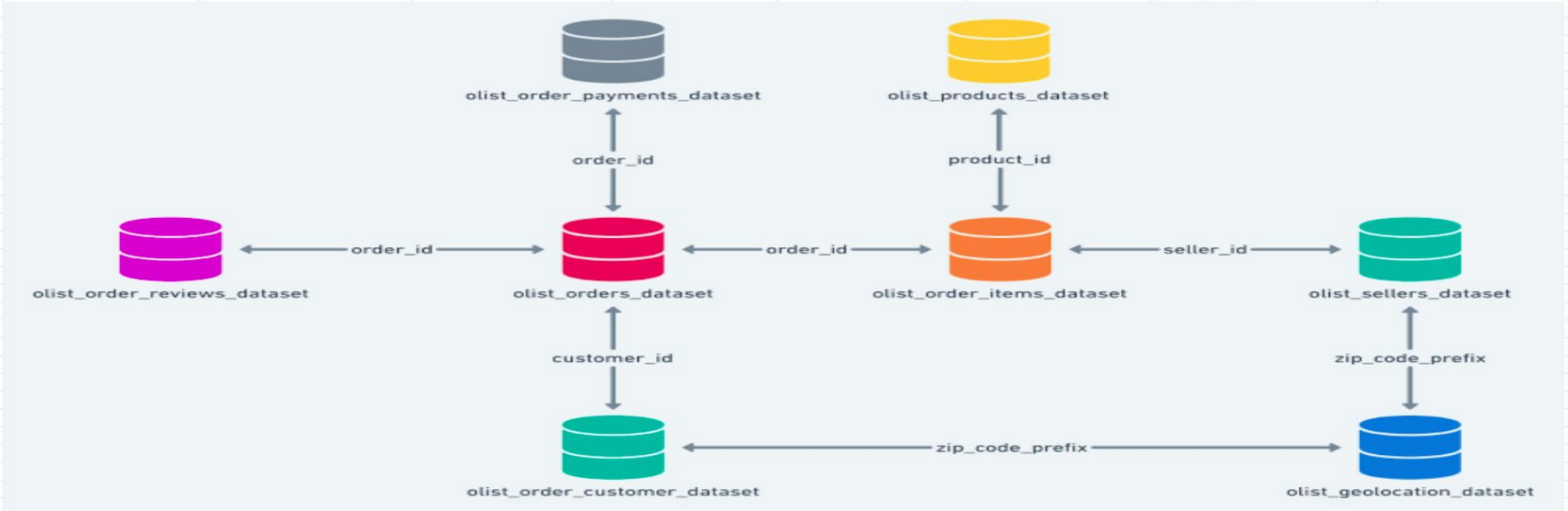
- olist site de e-commerce brésilien
- clustering des clients du site olist
    - définir des groupes de clients homogènes et cohérents :
        - sur la base de critères ***inconnus***
        - en un nombre de groupes ***inconnus***
        - si possible de tailles comparables
    - exhiber les caractéristiques métier de chacun de ces groupes

Données olist

- Données fournies sous la forme d'un dump de base de données:
  - 2 ans d'historique (de sept 2016 à oct 2018)
  - 96 096 clients (93 349 clients actifs)
  - 99 441 commandes

Données olist

customers	orders	geolocation	order_items	order_payments	order_reviews	products	sellers
customer_id	order_id	geolocation_zip_code_prefix	order_id	order_id	review_id	product_id	seller_id
customer_unique_id	customer_id	geolocation_lat	order_item_id	payment_sequential	order_id	product_category_name	seller_zip_code_prefix
customer_zip_code_prefix	order_status	geolocation_lng	product_id	payment_type	review_score	product_name_lenght	seller_city
customer_city	order_purchase_timestamp	geolocation_city	seller_id	payment_installments	review_comment_title	product_description_lenght	seller_state
customer_state	order_approved_at	geolocation_state	shipping_limit_date	payment_value	review_comment_message	product_photos_qty	
	order_delivered_carrier_date		price		review_creation_date	product_weight_g	
	order_delivered_customer_date		freight_value		review_answer_timestamp	product_length_cm	
	order_estimated_delivery_date					product_height_cm	
						product_width_cm	



Feature Engineering

- Méthode retenue pour manipuler les données olist:
  - Chargement des données dans une base de données (postgres)
  - Création d'une vue (SQL) pour la performance
  - Traitement des données effectués en SQL
  - Définition d'un "moteur" pour définir les caractéristiques client (*facilement modifiable*) en utilisant la vue précédente (ci-contre)
  - Création de fichiers plats (dataframe) pour les différentes phases de l'étude (clustering, maintenance)
  - Utilisation de l'identifiant client comme index des dataframes

```
(
lambda t:
f"""
    select distinct  c_customer_unique_id,max(r_review_score), avg(r_review_score)::numeric(10,2) as avg,min(r_review_score)
    from olist_view
    where o_order_delivered_customer_date <= '{t[0]}' and o_order_status = 'delivered' and r_review_creation_date <= '{t[0]}'
    group by c_customer_unique_id
""",
    [
        lambda t: (
            f"""select {t[2]}.min"""
            ,
            "review_score_min"
            ,
            f"left outer join {t[0]} as {t[2]} on {t[1]}.c_customer_unique_id = {t[2]}.c_customer_unique_id"
        ),
        lambda t: (
            f"""select {t[2]}.avg"""
            ,
            "review_score_avg"
            ,
            f"left outer join {t[0]} as {t[2]} on {t[1]}.c_customer_unique_id = {t[2]}.c_customer_unique_id"
        ),
        lambda t: (
            f"""select {t[2]}.max"""
            ,
            "review_score_max"
            ,
            f"left outer join {t[0]} as {t[2]} on {t[1]}.c_customer_unique_id = {t[2]}.c_customer_unique_id"
        )
    ]
),
```

Exploration

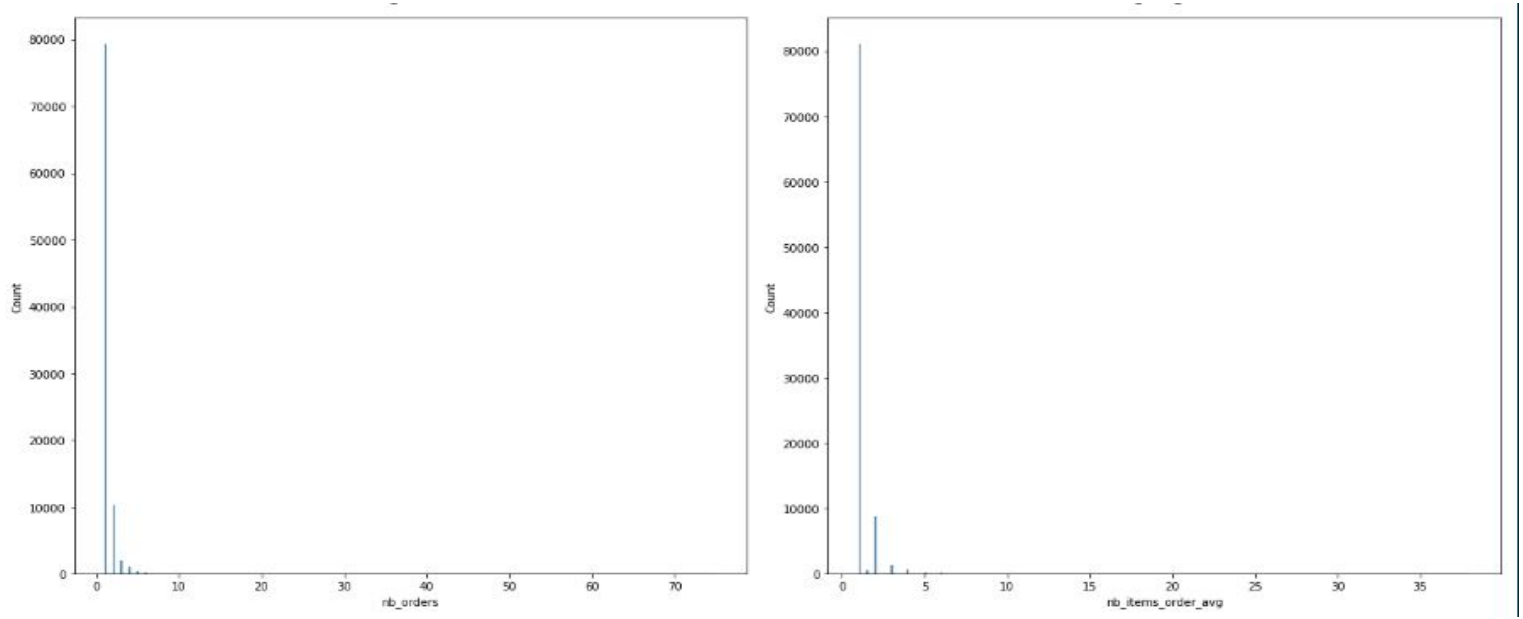
- *c_customer_unique_id* : identifiant client.
- **Commande**:
  - *order_value_max* : montant de la commande la plus chère
  - *order_value_min* : montant de la commande la moins chère
  - *order_value_avg* : montant moyen de commande
  - *order_value_stddev* : écart-type du montant de commande
  - *nb_orders_canceled* : nombre de commandes annulés
  - *nb_orders* : nombre de commandes (F)
  - *nb_orders_canceled* : nombre de commandes annulés
  - *nb_items_order_avg* : nombre moyen d'articles par commande
  - *amount_last_order* : montant de la dernière commande (M)
  - total_order : montant total des commandes
  - *nb_orders_canceled* : nombre de commandes annulés
  - *max_delay_delivery* *: délai de livraison (par rapport à la date estimée)*
  - *nb_days_since_last_order* : nombre de jours depuis la dernière commande (R)
- **Paiement de la commande**:
  - *nb_boleto* : nombre de paiements effectués en liquide.
  - *nb_debit_card* : nombre de paiements effectués avec une carte à débit.
  - *nb_voucher* : nombre de paiements effectués avec des bons.
  - *nb_credit_card* : nombre de paiements effectués avec une carte de crédit.
  - *payment_type_max* : type de paiement utilisé pour le plus gros montant.
  - *payment_installments_max* : nombre maximum de versements effectués pour un seul paiement.
  - *payment_sequential_max* :nombre maximum de types de paiement utilisés pour un seul paiement.
- ***Article:***
  - *r_freight_price_max* *: ratio prix article/frais de livraison maximum*
  - *product_cat_total_max* *: catégorie de produit*
  - *high_cat_total_max* *: catégorie de produit (regroupement)*
  - *weight_max* *: poids maximum*
  - *volume_max* *: volume maximum*
- ***Avis:***
  - *review_score_min* *: note la plus basse*
  - *review_score_avg* *: note moyenne*
  - *review_score_max* *: note la plus haute*
- ***Géolocalisation:***
  - *city* *: ville*
  - *state* *: état*
  - *distance* *: distance depuis le siège social d'olist*

2 - Clustering des clients du site olist: modélisation

- Exploration des features client
- Pistes de modélisation
- Grille de recherche
- Critères de sélection
- Modèle sélectionné

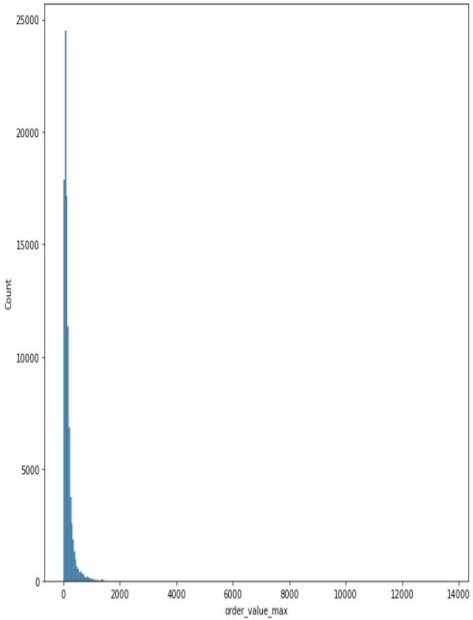
Exploration des features client

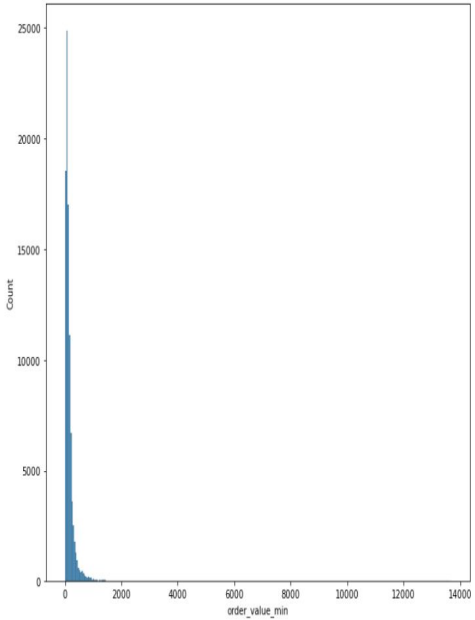
- Jeu de données très uniforme
    - La très grande majorité des clients n'a passé qu'une seule et unique commande avec un seul article

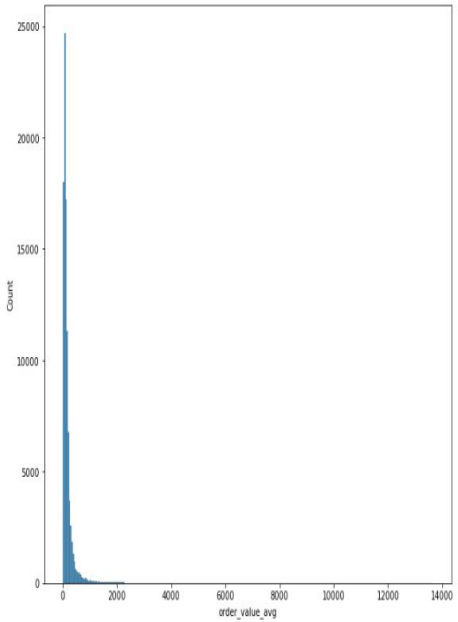


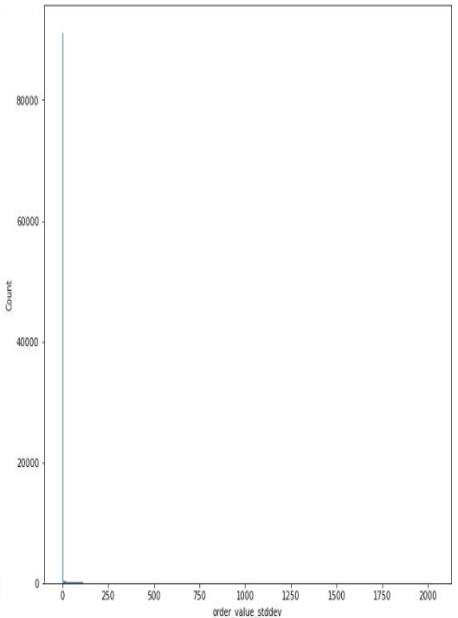
Exploration des features client

- Jeu de données très uniforme
    - Montant de commande min, max et moyen égaux et un écart type nul.



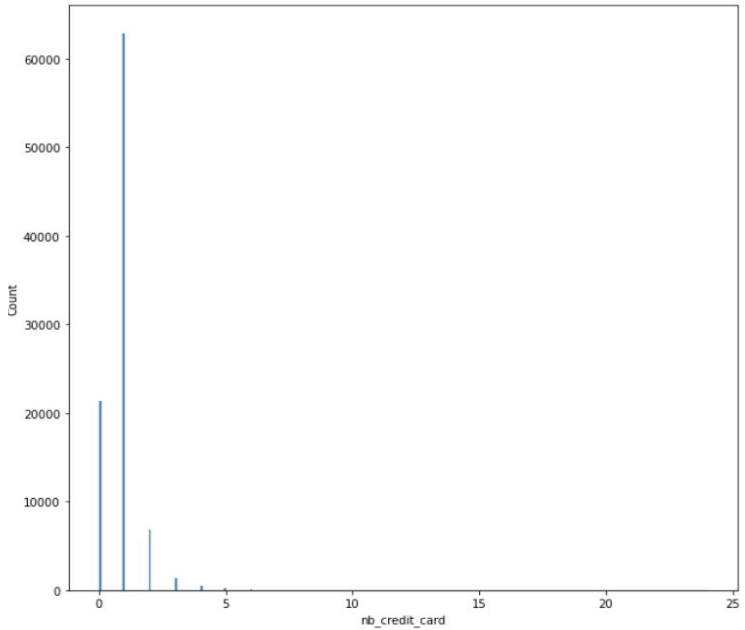


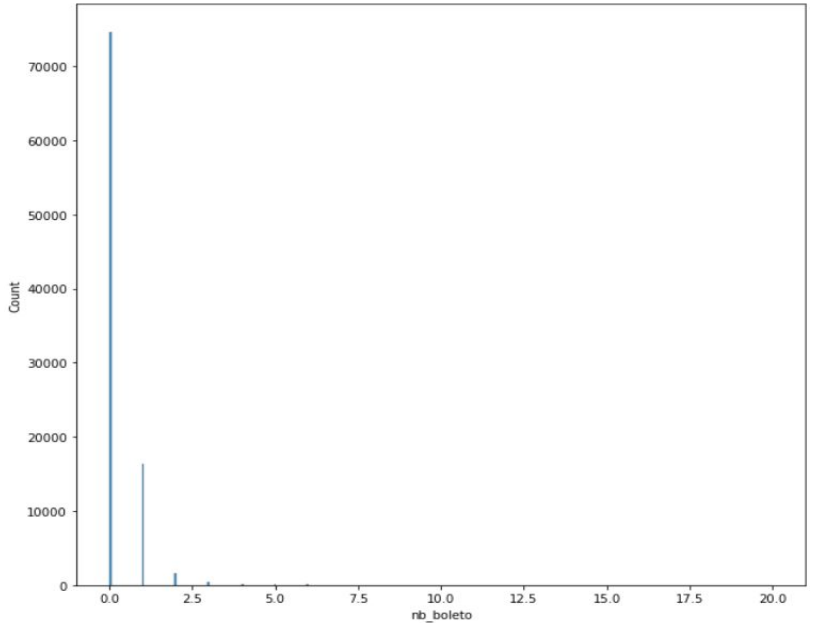




Exploration des features client

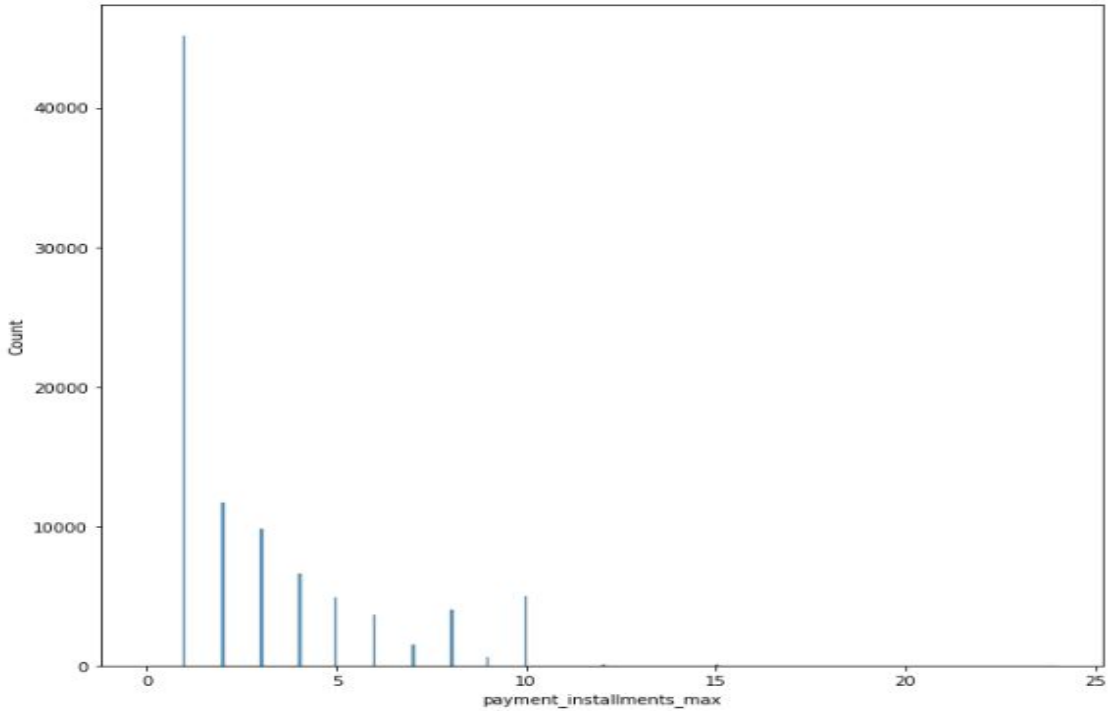
- Jeu de données très uniforme
    - L'essentiel des clients paie avec une carte de crédit, une minorité en espèce.





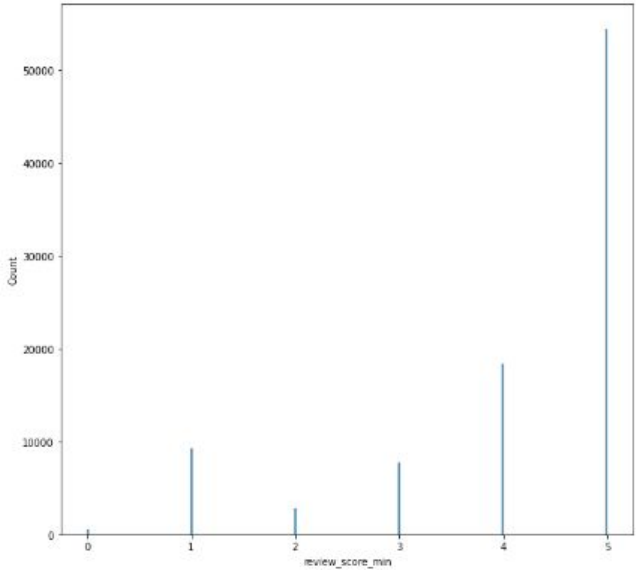
Exploration des features client

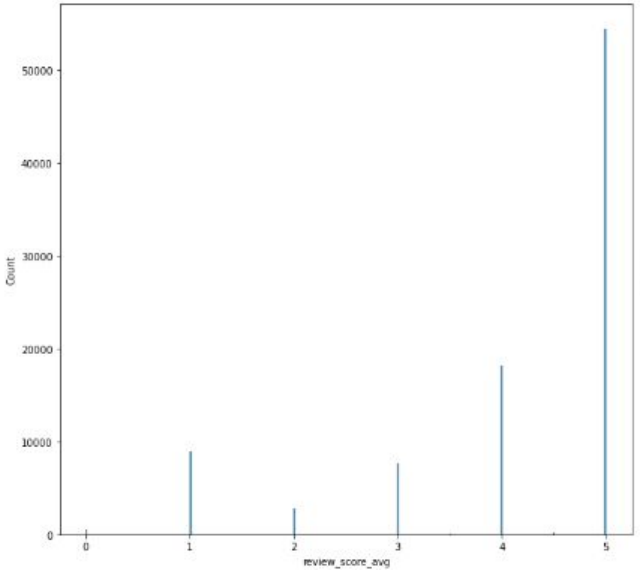
- Le nombre de versements présente une distribution moins uniforme

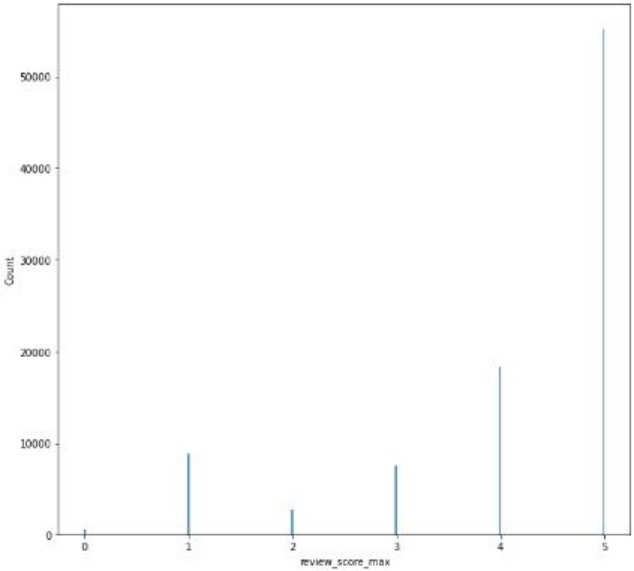


Exploration des features client

- L'essentiel des clients fournissent un commentaire sur leur commande.

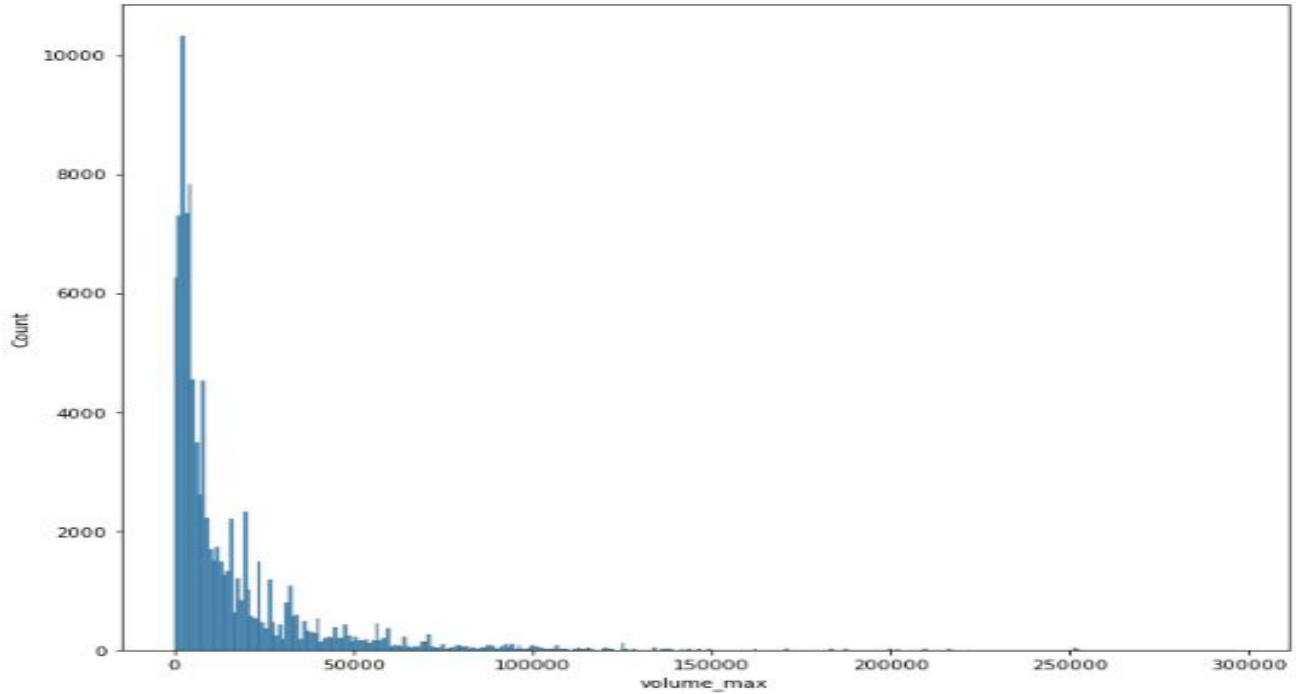






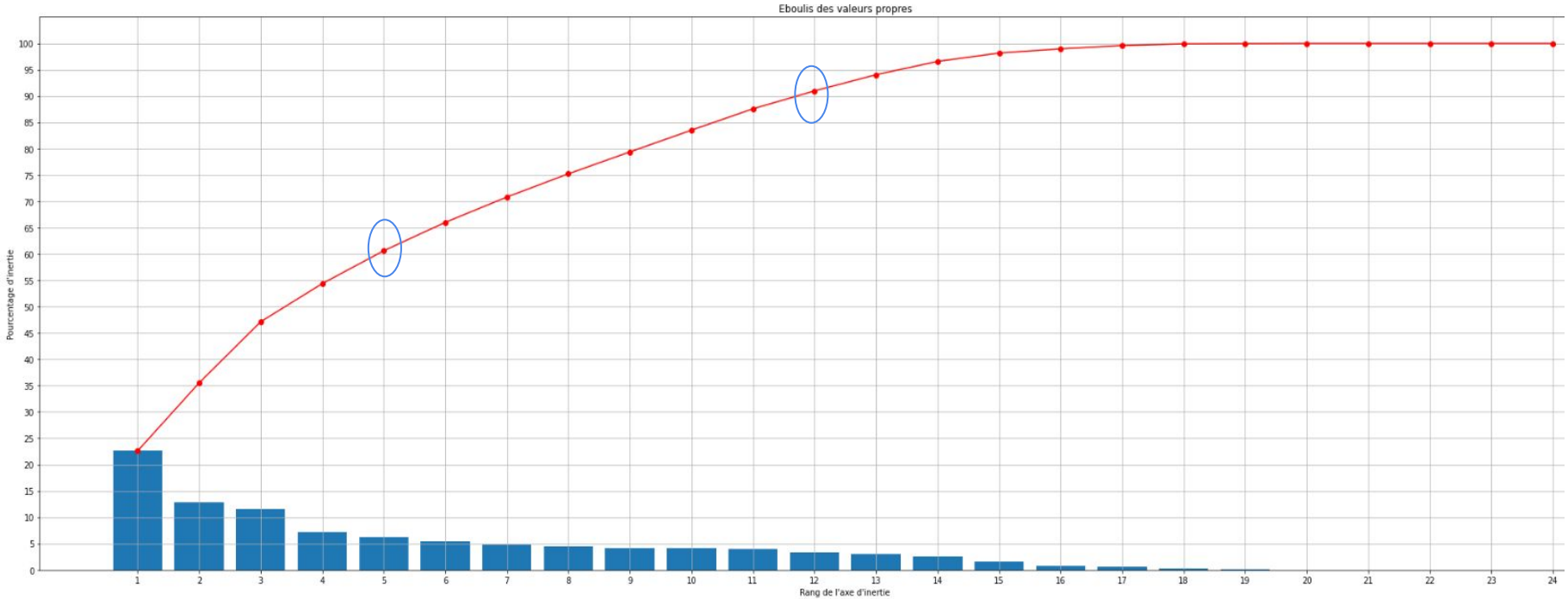
Exploration des features client

- Le volume des articles présente la distribution la moins uniforme.



Exploration des features client

- Analyse en composantes principales.



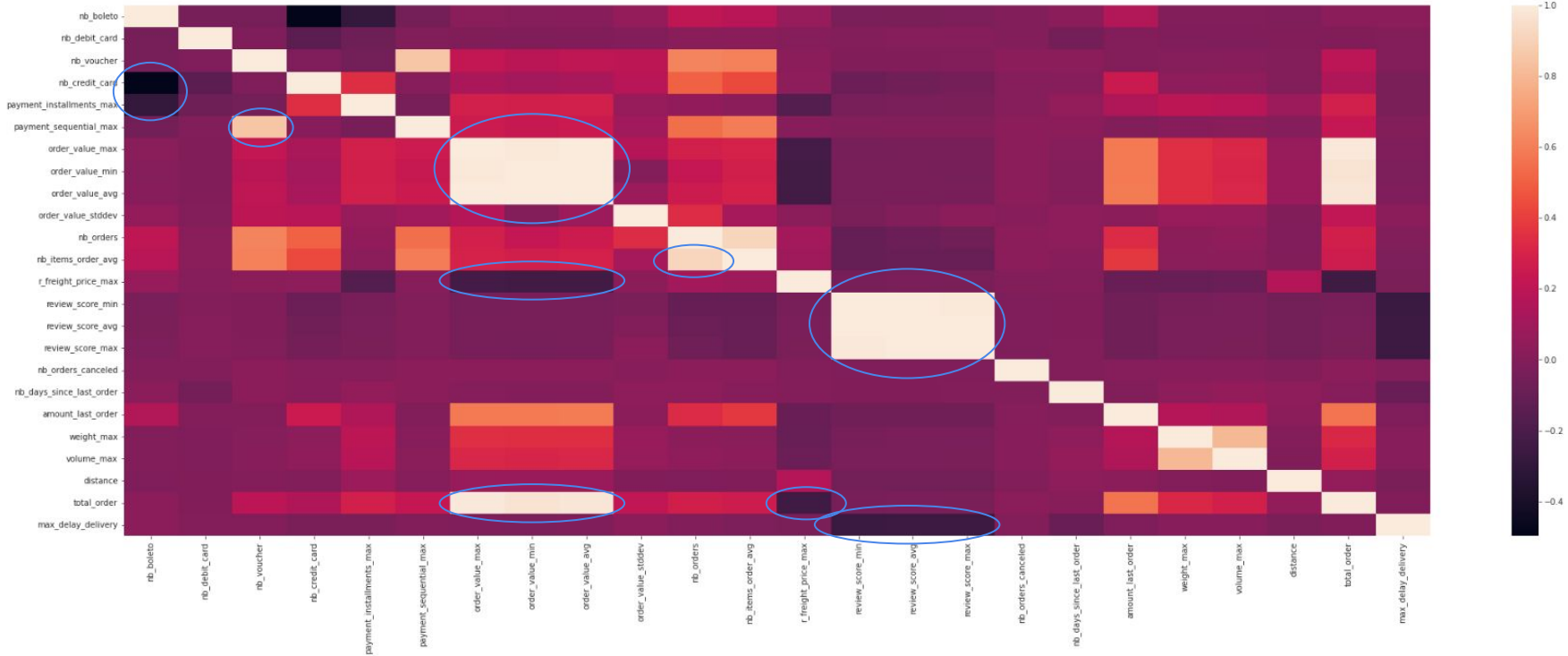
Exploration des features client

- Analyse en composantes principales.
    - Features prédominantes dans les premières composantes:

Composantes	Features
C1	order_value_min order_value_max order_value_avg total_order
C2	review_score_min review_score_max review_score_avg
C3	nb_voucher payment_sequential_max

Exploration des features client

- Matrice de corrélation



Pistes de modélisation

- Algorithmes de clustering considérés
  - K-means
  - DBScan
- Sélection des features
  - RFM
  - Sur la base d'un niveau de variance expliquée
  - Sélection manuelle de feature
- Différents prétraitements:
  - Analyse en composantes principales
  - Application de log + 1
- Différents scalers
  - Standard
  - RobustScaler
  - MinMaxScaler

Grille de recherche

- Les algorithmes de clustering ont des hyper-paramètres qu'il faut déterminer:
  - le nombre de cluster dans le kmeans

- Certains des paramètres des éléments de modélisation précédents peuvent également être considérés comme des "hyper-paramètres"
  - les features sélectionnés, le niveau de variance expliqué

- Mise en place d'une grille de recherche:
  - qui permet de définir des pipes sur la base des éléments de modélisation précédents
  - les pipes sont ensuite exécutés avec les différents hyper-paramètres spécifiés dans la configuration

Pistes de modélisation

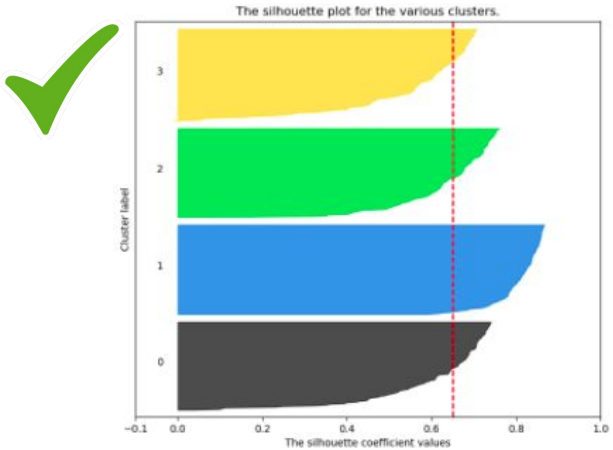
```
high_correlated = ['nb_voucher','nb_credit_card','nb_orders','nb_items_order_avg', 'nb_boleto','payment_installments_max','payment_sequential_max','amount_last_order']
low_correlated  = ['distance','volume_max','weight_max','nb_orders_canceled','nb_days_since_last_order','review_score_avg','r_freight_price_max','order_value_stddev','nb_debit_card']

config = {
        'RFM-kmeans':{
            'pipe':  Pipeline([("featuresSelection",FeaturesSelection(['nb_days_since_last_order','nb_orders','amount_last_order'])),("k-means",KMeans())]),
            'params':{'k-means__n_clusters':range(2,11),'k-means__algorithm':['full'],'k-means__max_iter':[500],'k-means__n_init':[30]},
            'fitted_params':['k-means__n_clusters']
        },
        'RFM-DBSCAN':{
            'pipe':  Pipeline([("featuresSelection",FeaturesSelection(['nb_days_since_last_order','nb_orders','amount_last_order'])),("stdScaler",StandardScaler()),("dbscan",DBSCAN())]),
            'params':{'dbscan__eps':[0.1,0.2,0.5],'dbscan__min_samples':[5,10,50],'dbscan__n_jobs':[-1]},
            'fitted_params':['dbscan__eps','dbscan__min_samples']
        },
        'kmeans':{
            'pipe':  Pipeline([("FS",RuleSelection(rule=lambda X: list(X.select_dtypes('number').columns.values))),("SC",FeaturesScaler(scaler=StandardScaler())),
                               ("pcaFeaturesSelection",PCAFeaturesSelection()),("k-means",KMeans())]),
            'params':{'k-means__n_clusters':range(2,7),'k-means__algorithm':['full'],'k-means__max_iter':[500],'k-means__n_init':[30],'pcaFeaturesSelection__level':[0.75,0.8,0.85,0.9] },
            'fitted_params':['k-means__n_clusters','pcaFeaturesSelection__level']
        },
        'kmeans1':{
            'pipe':  Pipeline([ ("FS",FeaturesSelection(['nb_days_since_last_order','nb_orders','amount_last_order'])) ,("k-means",KMeans())]),
            'params':{'k-means__n_clusters':[3,4,5] },
            'fitted_params':['k-means__n_clusters']
        },
        'kmeans2':{
            'pipe':  Pipeline([("FS",FeaturesSelection(['nb_boleto','nb_credit_card','order_value_avg','nb_items_order_avg','review_score_avg'])),("k-means",KMeans())]),
            'params':{'k-means__n_clusters':[3,4,5] },
            'fitted_params':['k-means__n_clusters']
        },
        'kmeans3':{
            'pipe':  Pipeline([("FS",FeaturesSelection(['nb_boleto','nb_credit_card','order_value_avg','nb_items_order_avg','review_score_avg','payment_installments_max'])),("k-means",KMeans())]),
            'params':{'k-means__n_clusters':[3,4,5] },
            'fitted_params':['k-means__n_clusters']
        },
        'kmeans4':{
            'pipe':  Pipeline([("FS",FeaturesSelection(['nb_boleto','nb_credit_card','order_value_avg','nb_items_order_avg','review_score_avg','r_freight_price_max'])),("k-means",KMeans())]),
            'params':{'k-means__n_clusters':[3,4,5] },
            'fitted_params':['k-means__n_clusters']
        },
```

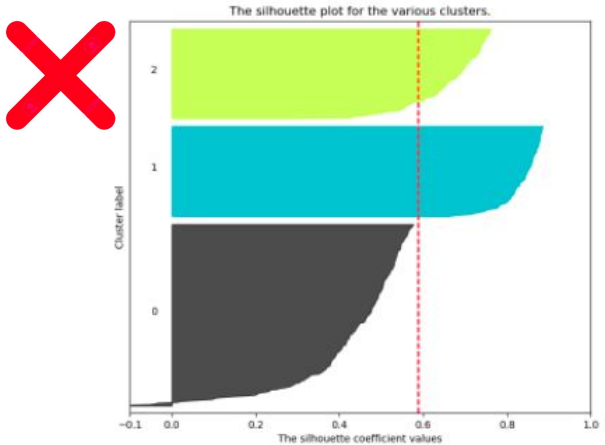
Pistes de modélisation

```
'highc-kmeans':{
    'pipe':  Pipeline([("featuresSelection",FeaturesSelection(high_correlated )),("k-means",KMeans())]),
    'params':{
        'k-means__n_clusters':range(3,5),'k-means__algorithm':['full'],'k-means__max_iter':[300],'k-means__n_init':[10]

    },
    'fitted_params':['k-means__n_clusters']
},
'lowc-kmeans':{
    'pipe':  Pipeline([("featuresSelection",FeaturesSelection(low_correlated )),("k-means",KMeans())]),
    'params':{
        'k-means__n_clusters':range(3,5),'k-means__algorithm':['full'],'k-means__max_iter':[300],'k-means__n_init':[10]

    },
    'fitted_params':['k-means__n_clusters']
},
'kmeans5':{
    'pipe':  Pipeline([
        ("FS",FeaturesSelection()),
        ("L",LambdaFeatures(f=Log_plus1(),features=['order_value_avg','volume_max','weight_max','r_freight_price_max'])),
        ("SC",FeaturesScaler()),
        ("kM",KMeans())
    ]),
    'params':{
        'kM__n_clusters':[3,4,5],'kM__algorithm':['full'],'kM__max_iter':[500],'kM__n_init':[30],
        'FS__features':[
            ['order_value_avg','review_score_avg','volume_max'],
            ['order_value_avg','review_score_avg','volume_max','r_freight_price_max'],
            ['order_value_avg','review_score_avg','weight_max','r_freight_price_max'],
            ['order_value_avg','review_score_avg','volume_max','weight_max','r_freight_price_max'],
            ['order_value_avg','review_score_avg','volume_max','weight_max','distance'],
            ['order_value_avg','review_score_avg','volume_max','payment_installments_max','nb_credit_card'],
            ['order_value_avg','review_score_avg','volume_max','weight_max','r_freight_price_max','payment_installments_max','nb_credit_card','distance']
        ],
        "SC__scaler":[RobustScaler()],
    },
    'fitted_params':['kM__n_clusters','FS__features']
},
'kmeans-wPCA':{
    'pipe':  Pipeline([("FS",RuleSelection(rule=lambda X: list(X.select_dtypes('number').columns.values)),("SC",FeaturesScaler(scaler=StandardScaler())),
                       ("pcaFeaturesSelection",PCAFeaturesSelection()),("k-means",KMeans())]),
    'params':{'k-means__n_clusters':range(2,7),'k-means__algorithm':['full'],'k-means__max_iter':[500],'k-means__n_init':[30],'pcaFeaturesSelection__level':[0.20,0.35,0.50,0.6,0.65] },
    'fitted_params':['k-means__n_clusters','pcaFeaturesSelection__level']
},
```

Critères de sélection

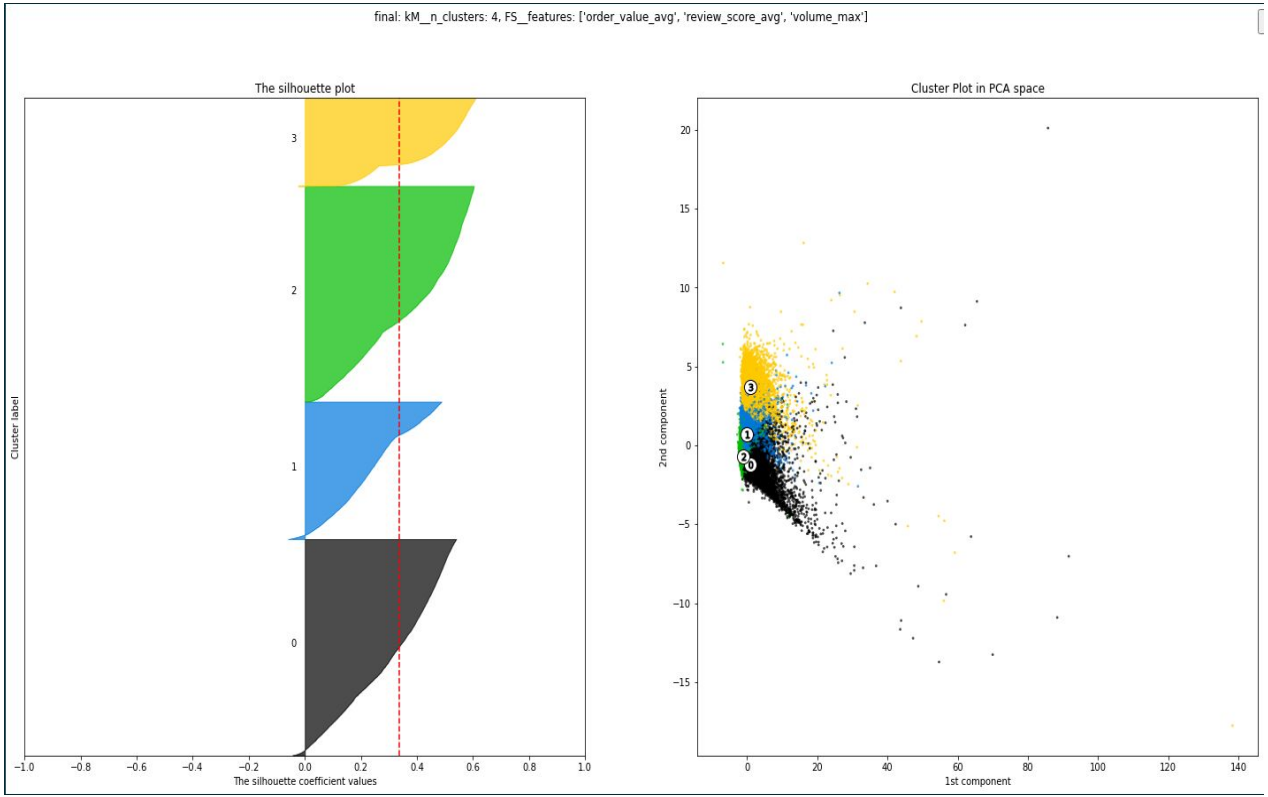
- Inspiré par cet article: Selecting the number of clusters with silhouette analysis on KMeans clustering
- La sélection est basée sur le score silhouette (par valeur et moyen)
- On s'appuie sur la représentation du score silhouette par valeur:
  - les formes les plus homogènes possibles
  - toutes les formes "dépassent" la valeur moyenne





Modèle sélectionné

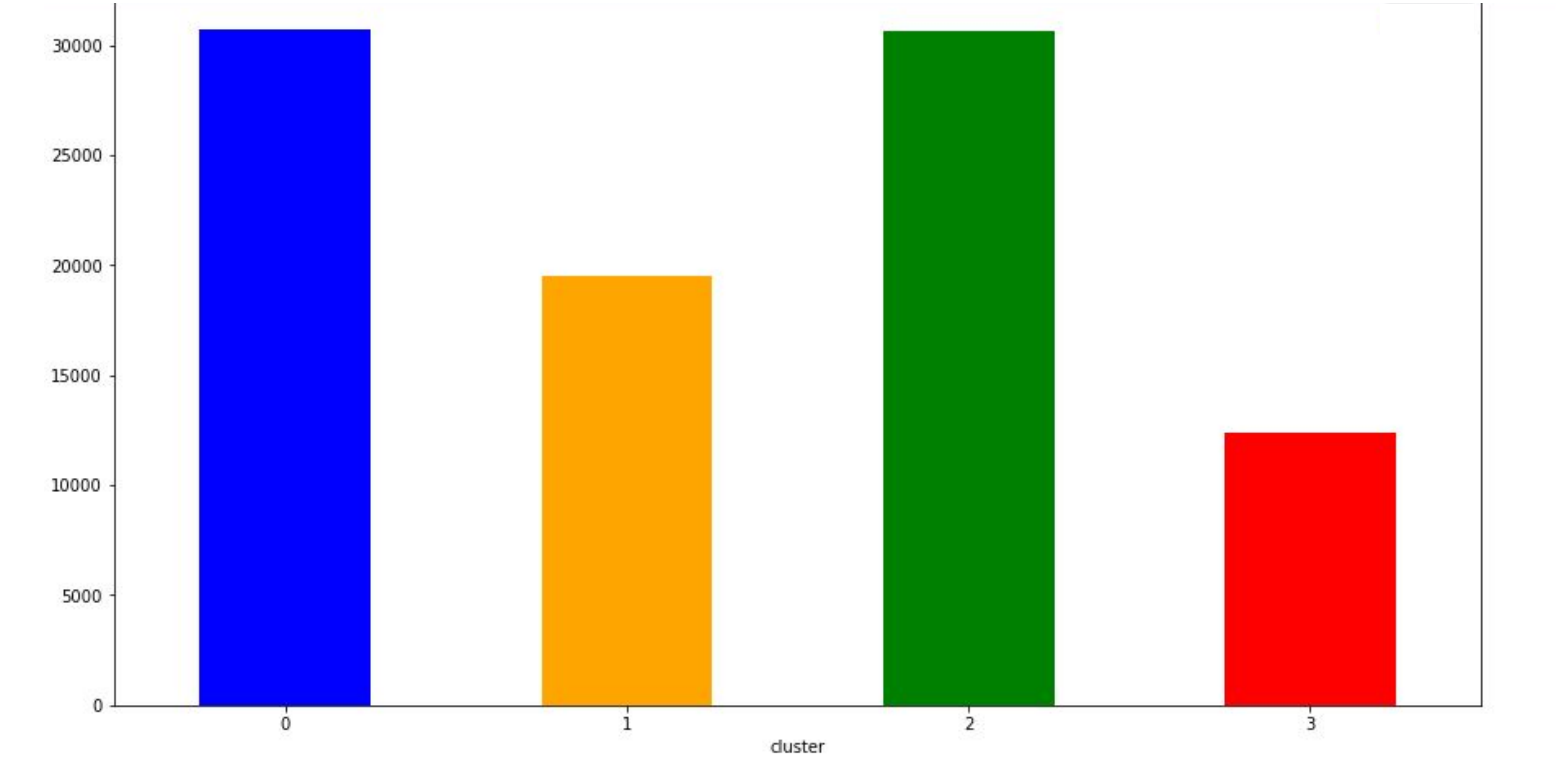
- Algorithme: K-means
- Nombre de cluster: 4
- Features utilisés:
  - order_value_avg
  - review_score_avg
  - volume_max



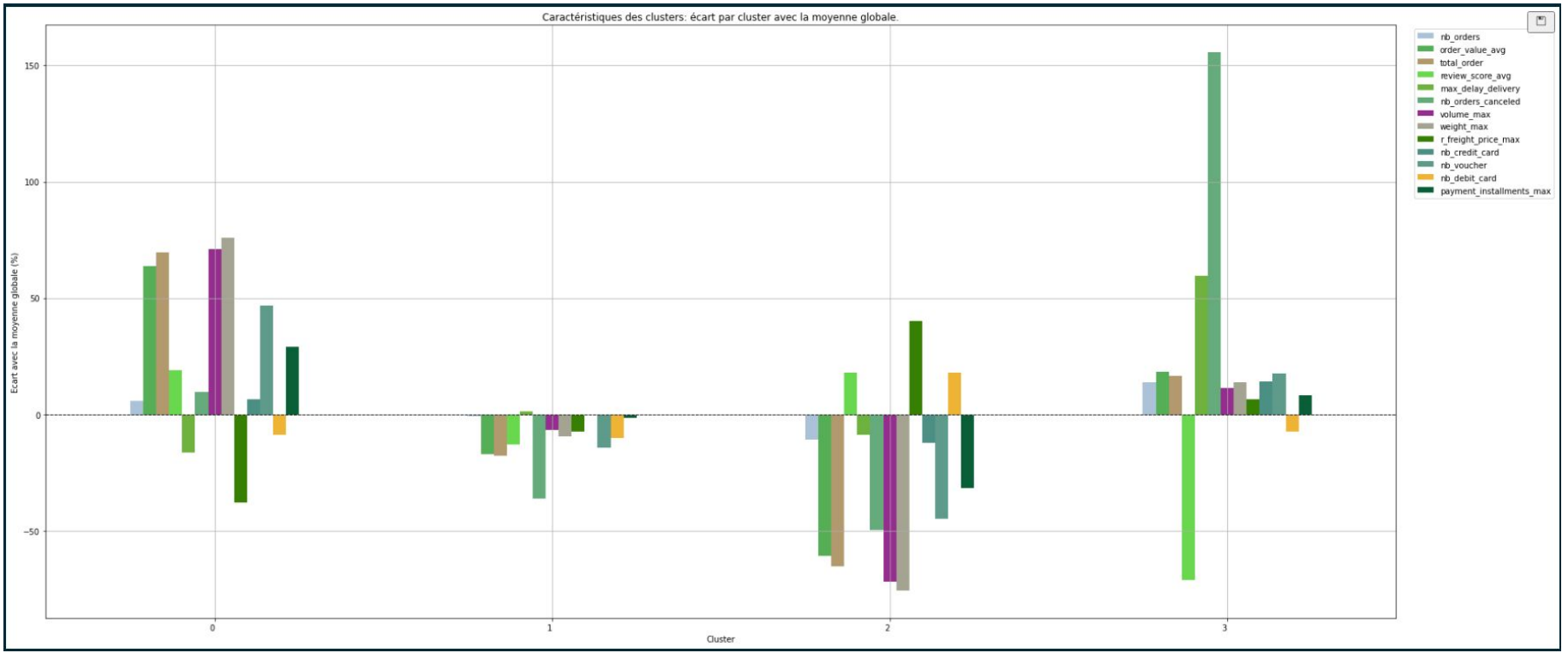
3- Clustering des clients du site olist: Interprétation des clusters

- Population des clusters
- Variation par rapport à la moyenne
- Graphe Radar
- Répartition du CA et des commandes par cluster

Population des clusters



Variation par rapport à la moyenne



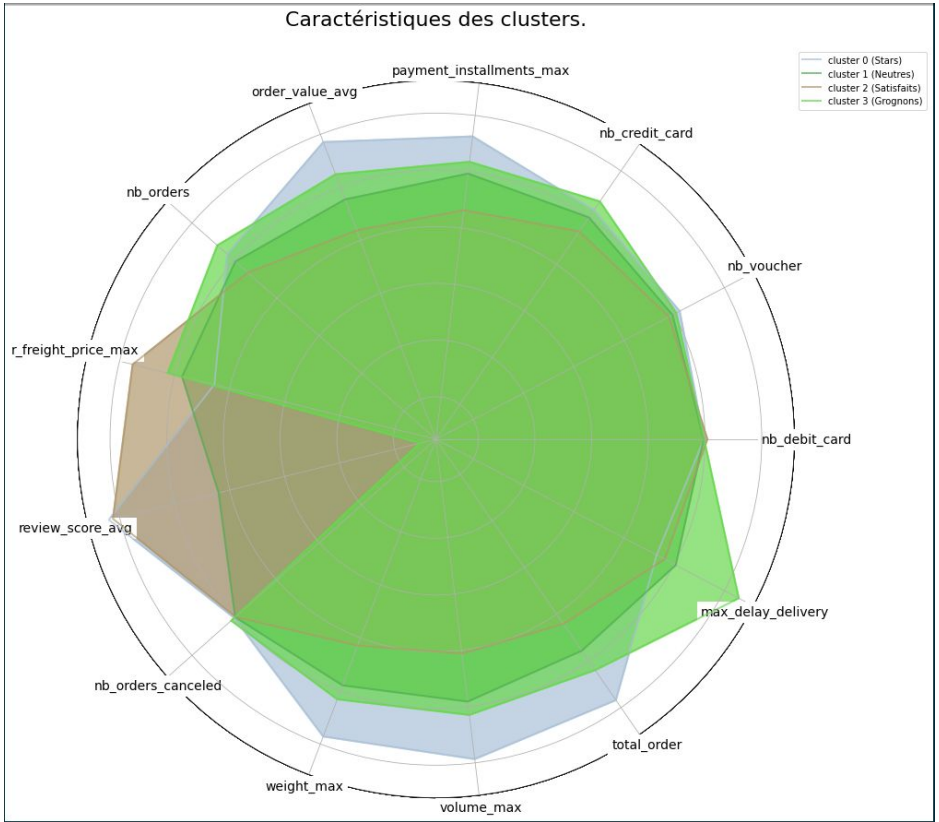
Stars

Neutres

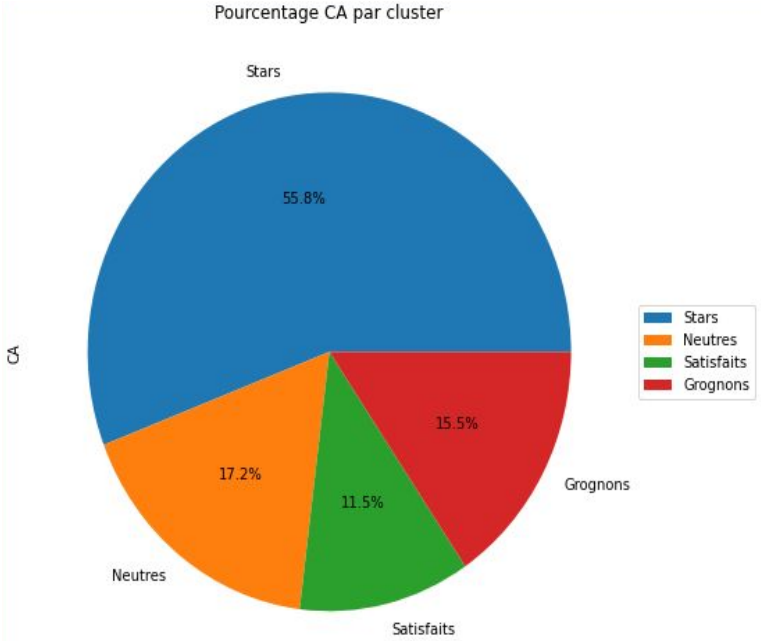
Satisfaits

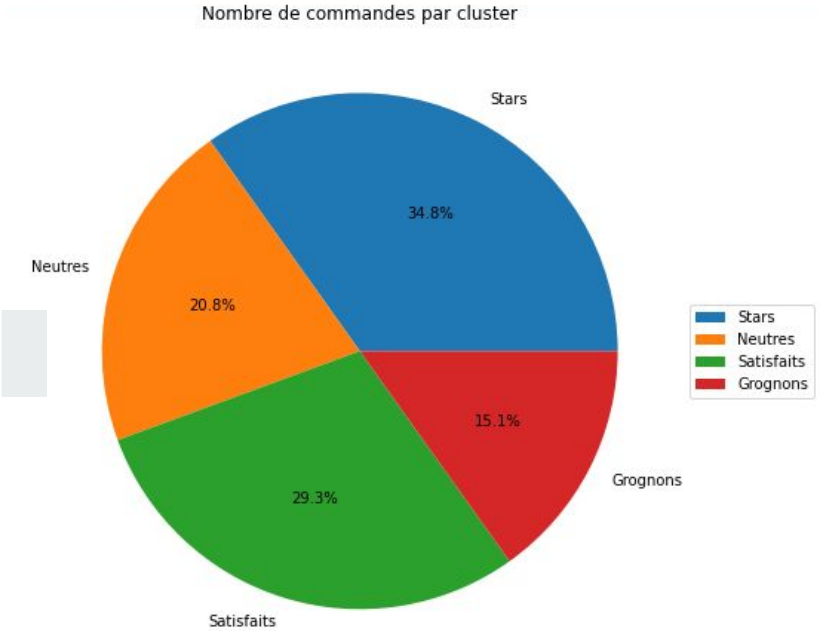
Grognons

Graphe Radar



Répartition du CA et des commandes par cluster





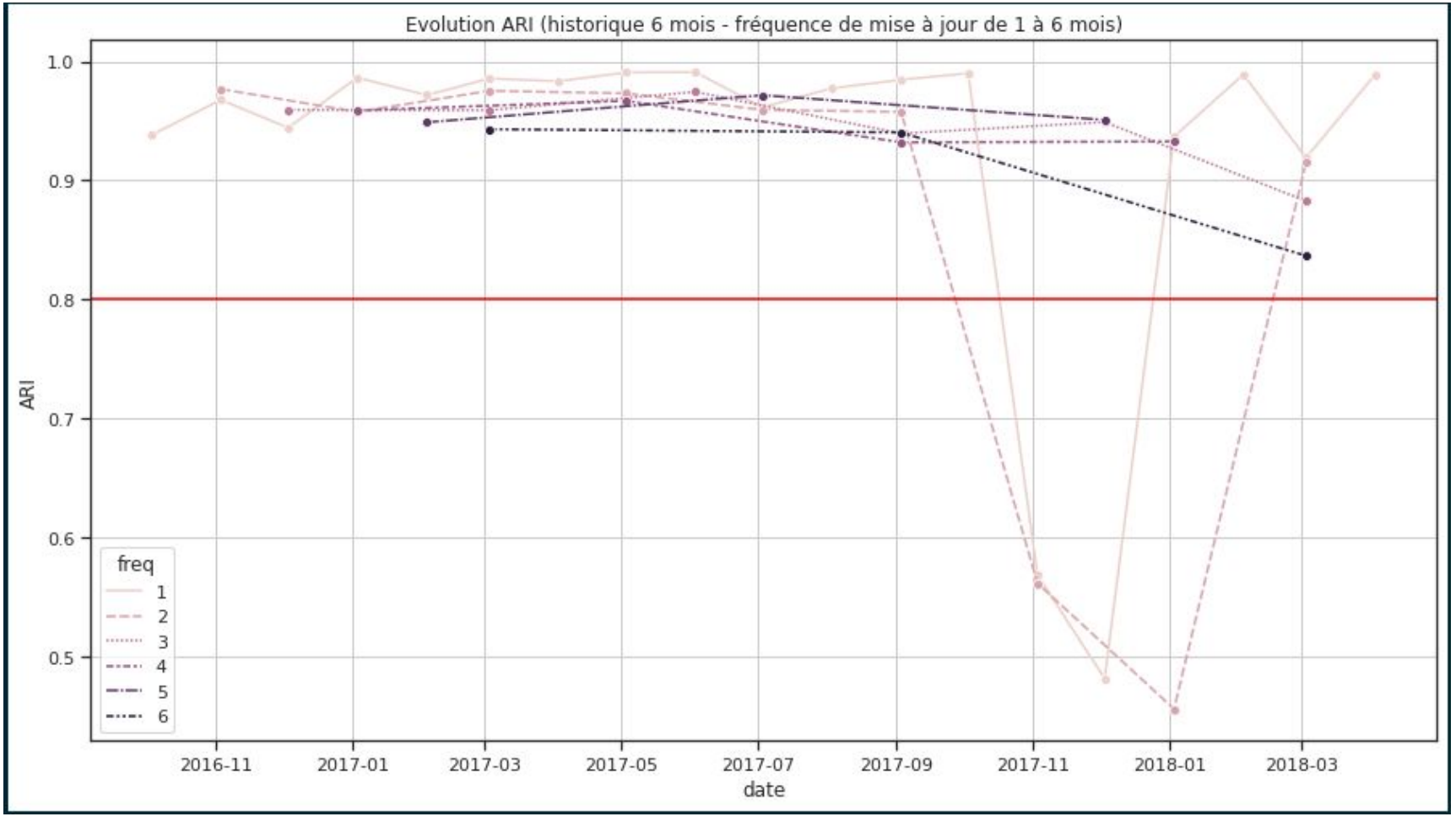
Caractéristiques des clusters

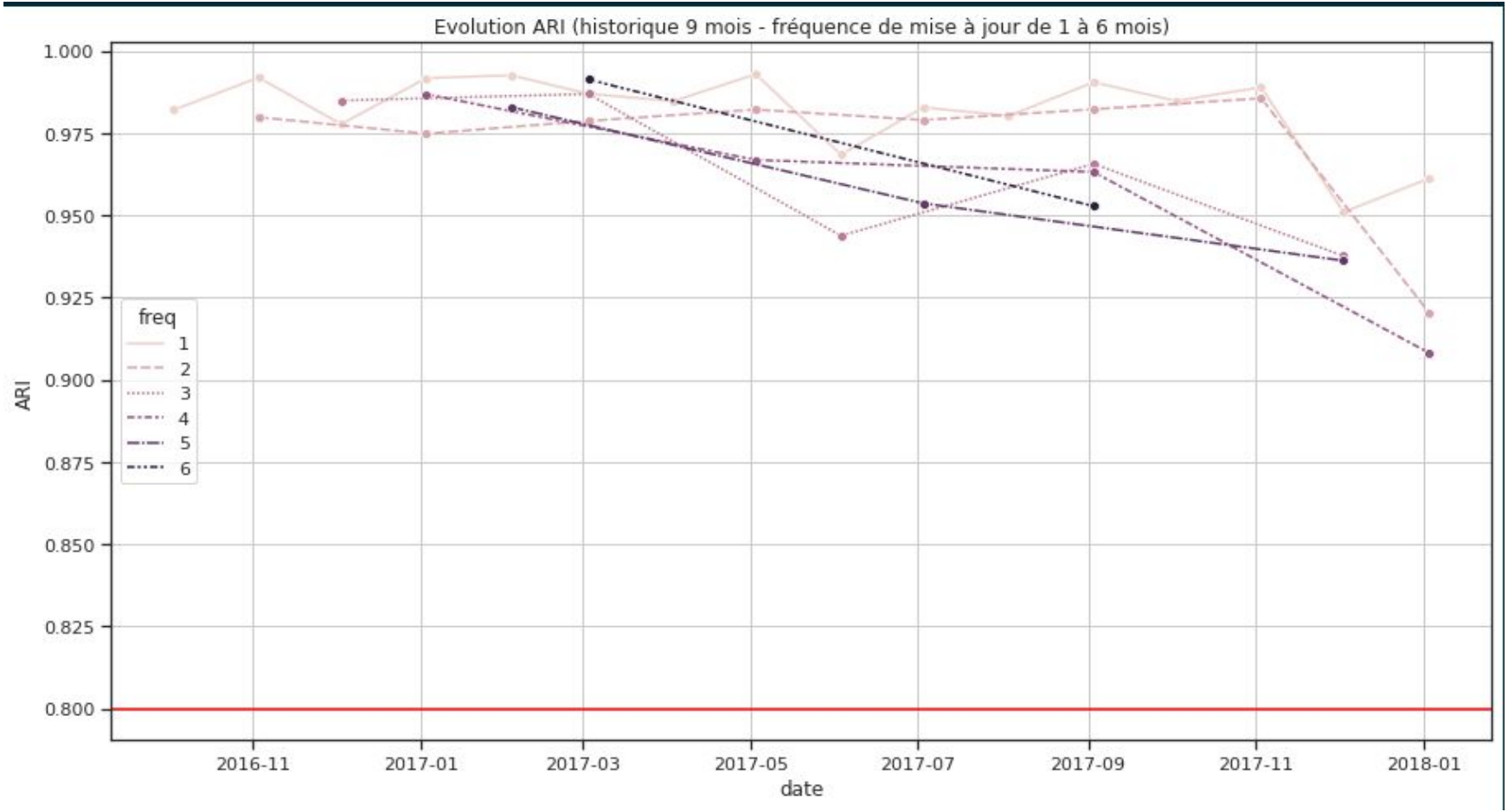
Caractéristiques des clusters- Recommandations

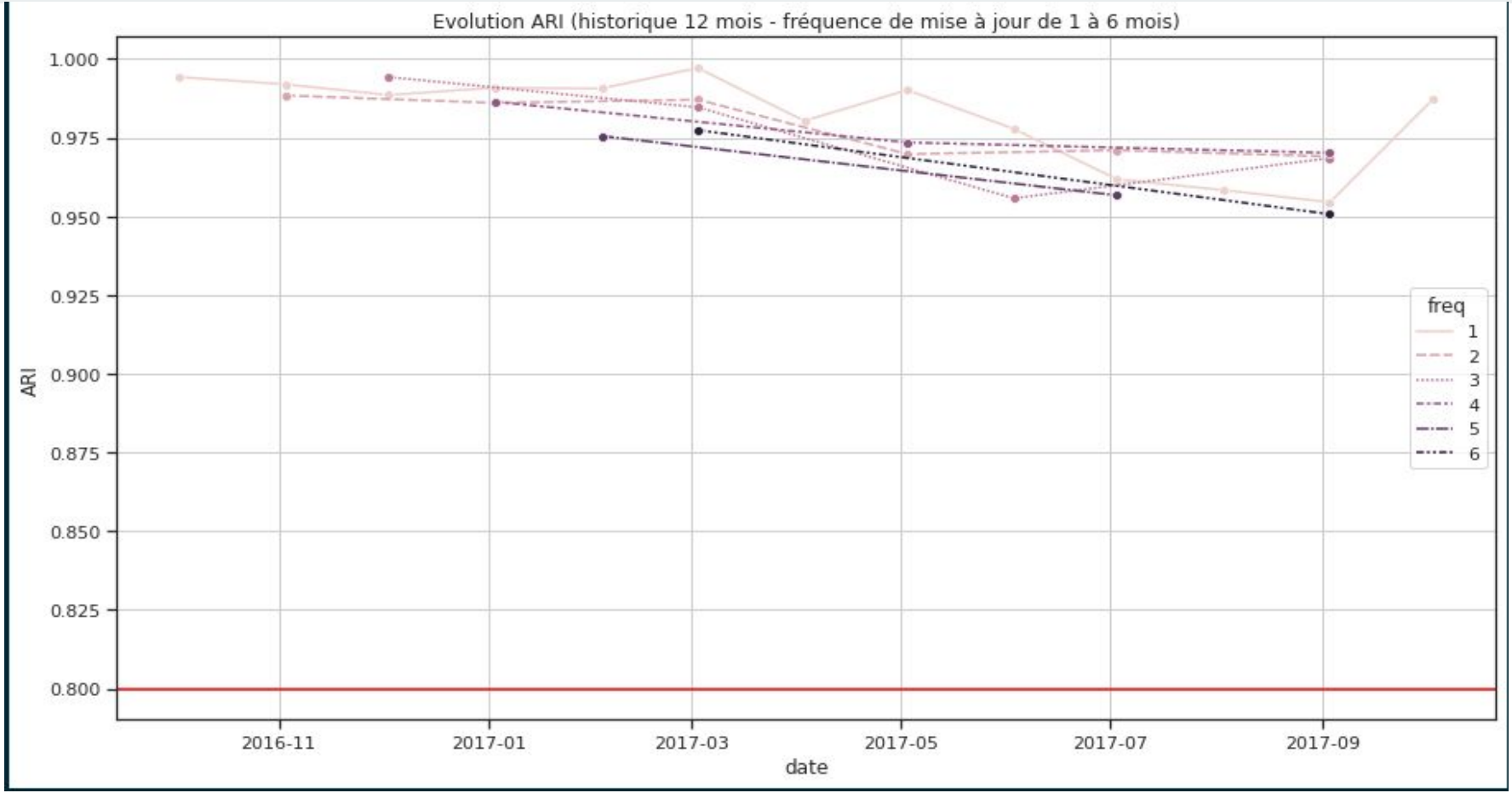
- Ne pas modifier la politique des frais de livraison.
- Ne pas modifier la politique d’annulation.
- Ne pas modifier la possibilité de paiement en plusieurs versements.
- Revoir l’attribution/la politique des vouchers.

4- Clustering des clients du site olist: Plan de maintenance

- Nécessité de re-calibrer le modèle à interval de temps régulier
- Indicateur **ARI** (Ajusted Rand Index) permet de quantifier la qualité du clustering:
  - ARI < 0.8 ⇒ nouvelle calibration nécessaire.
- Critères à définir:
  - profondeur d'historique (à minimiser - moins de consommation de ressource)
  - interval de mise à jour (à maximiser - faire la calibration le moins souvent possible)
  - paramètres étudiés :
    - historiques: 6,9 et 12 mois.
    - intervalles: de 1 à 6 mois.


Evolution ARI (historique 6 mois - fréquence de mise à jour de 1 à 6 mois)
ARI
1.0
0.9
0.8
0.7
0.6
0.5
freq
1
2
3
4
5
6
2016-11
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
date


Evolution ARI (historique 9 mois - fréquence de mise à jour de 1 à 6 mois)
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
ARI
freq
1
2
3
4
5
6
2016-11
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
date


Evolution ARI (historique 12 mois - fréquence de mise à jour de 1 à 6 mois)
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
ARI
2016-11
2017-01
2017-03
2017-05
2017-07
2017-09
date
freq
1
2
3
4
5
6

Plan de maintenance - Conclusion

- historique d'un an avec mise à jour des données tous les trimestres.