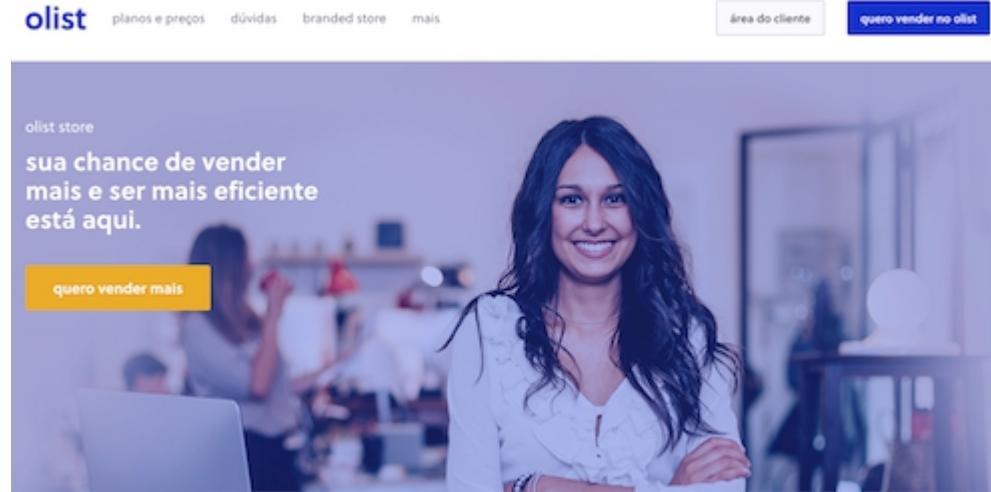


Soutenance Projet 5 : "Segmentez des clients d'un site e-commerce"



Sommaire

I. Problématique

II. Nettoyage des données / Analyse Exploratoire / Feature Engineering

III. Etude RFM

IV. Modèles de clustering (RFM +)

V. Modèle retenu

VI. Contrat de maintenance

VII. Conclusion



I. Problématique

- Olist : Solution de vente sur les marketplaces en ligne
- Besoin d'une segmentation client pour les équipes e-marketing afin de mettre en place une stratégie de campagne commerciale (actionnable)
- Base de données anonymisée disponible (à partir de Janvier 2017)
- Définir un modèle de segmentation afin d'identifier et de cibler les comportements clients (compréhensible)
- Proposer un contrat de maintenance sur ce modèle.
- Appliquer la convention Pep8 au code.



I. Problématique

Démarche de travail :

- Analyse exploratoire des données disponibles.
- Filtrer le jeu de données, évaluer les variables pertinentes pour notre problématique et appliquer du feature engineering pour en sortir des comportements intéressants.
- Effectuer une première analyse de segmentation classique de type RFM.
- Tester différents modèles d'apprentissage non-supervisé afin de proposer une solution (clustering) plus élaboré que la RFM.
- Sélectionner le meilleur modèle pour notre problématique et en sortir les caractéristiques des clusters (segments).
- Évaluer la stabilité dans le temps du modèle sélectionné et proposer un contrat de maintenance.
- Conclusion sur la solution proposée et les possibilités pour aller plus loin.



II. Nettoyage des données / Analyse exploratoire / Feature engineering

→ Neufs fichiers de données CSV disponibles:

- Fichier clients
- Géo-localisations
- Produits par commandes
- Paiement par commande
- Review par commande (note, commentaire)
- Commandes
- Produits
- Vendeurs
- Traduction produits

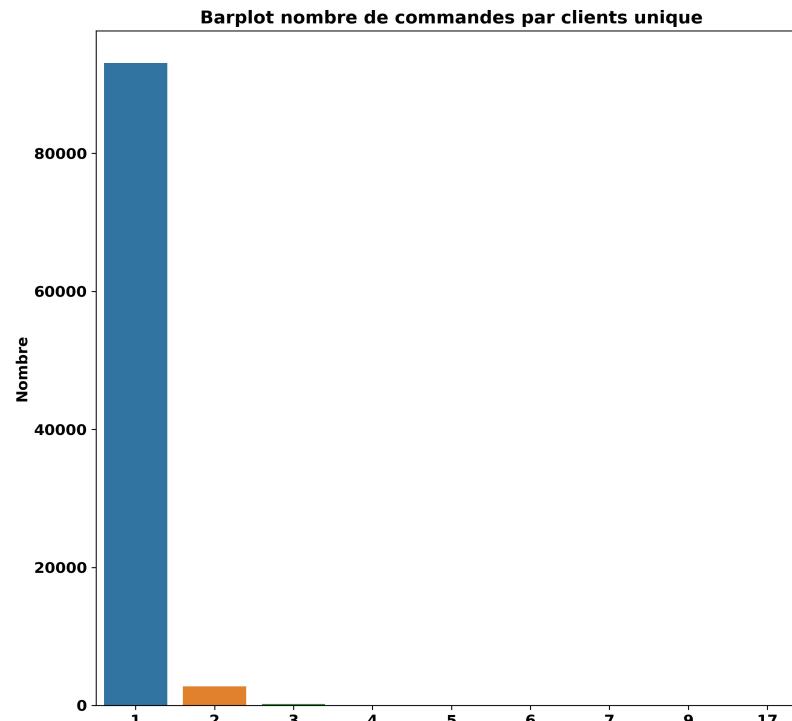
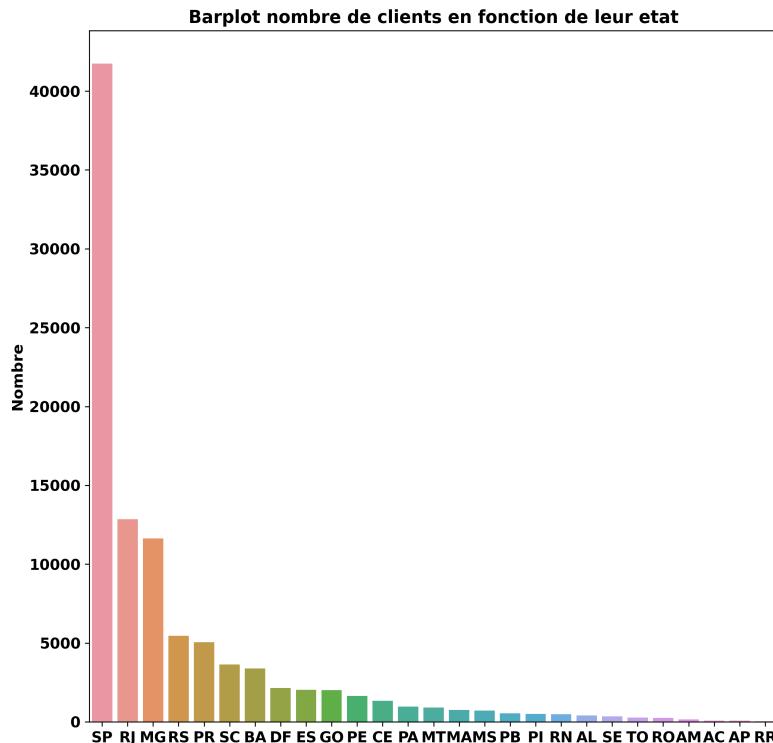
→ Exploration sommaire de chaque fichiers

→ Regroupement des différents fichiers



II. Nettoyage des données / Analyse exploratoire / Feature engineering

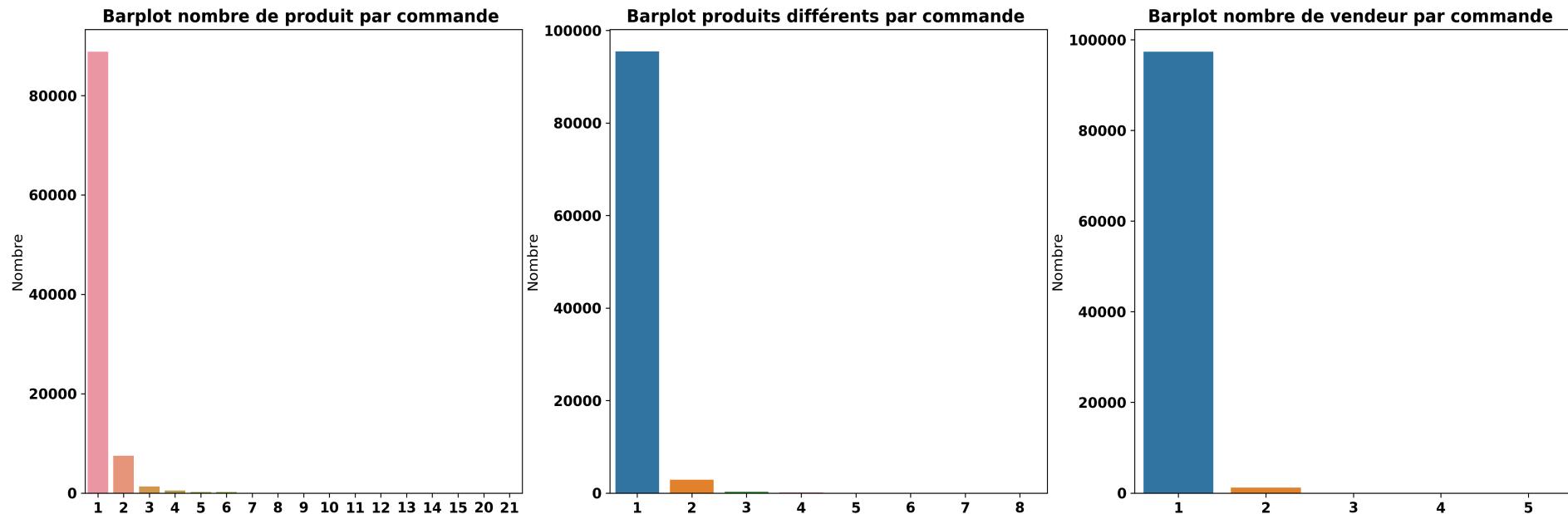
→ Informations clients :



- Concentration de clients dans l'état de São Paulo.
- Beaucoup de clients à une seule commande.

II. Nettoyage des données / Analyse exploratoire / Feature engineering

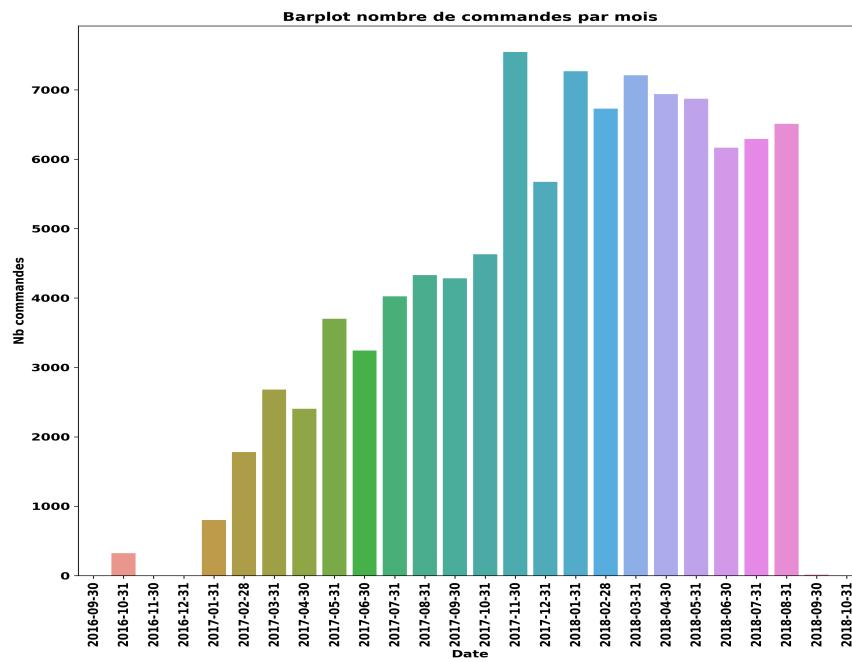
→ Informations commandes :



→ Beaucoup de commande à 1 seul produit / 1 seul vendeur.

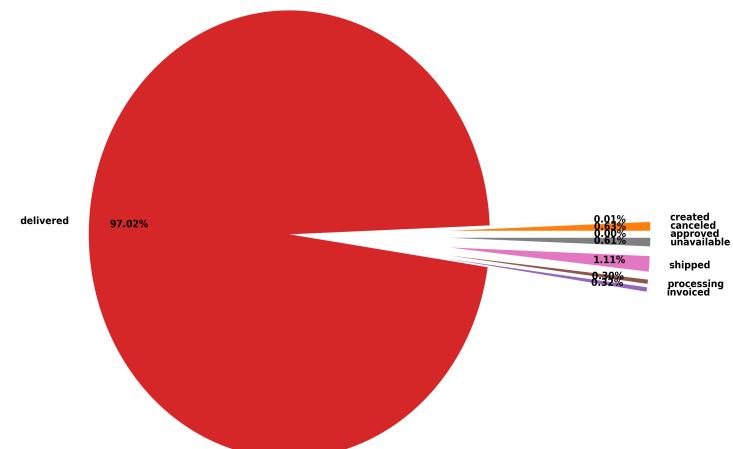
II. Nettoyage des données / Analyse exploratoire / Feature engineering

→ Informations commandes :



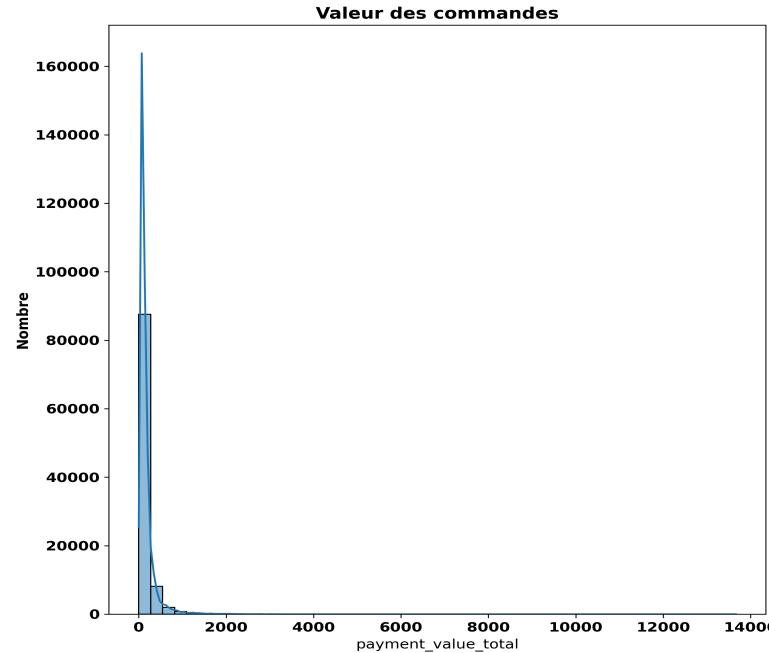
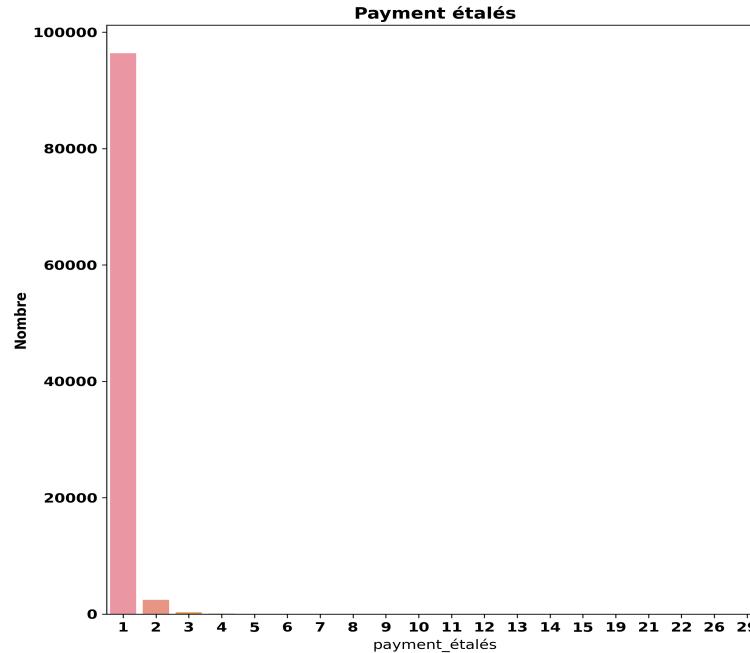
- Peu de commande avant 2017.
- Certaines commandes sont annulées.

Répartition des commandes suivant leur status (99441 observations)

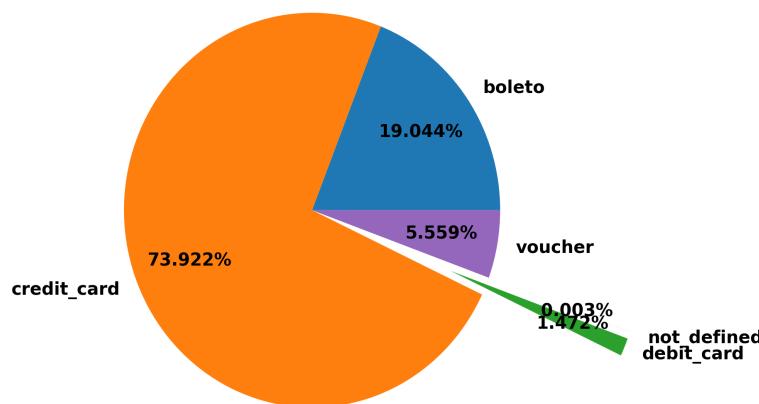


II. Nettoyage des données / Analyse exploratoire / Feature engineering

→ Informations payments :



Répartition des types de paiements(103886 observations)

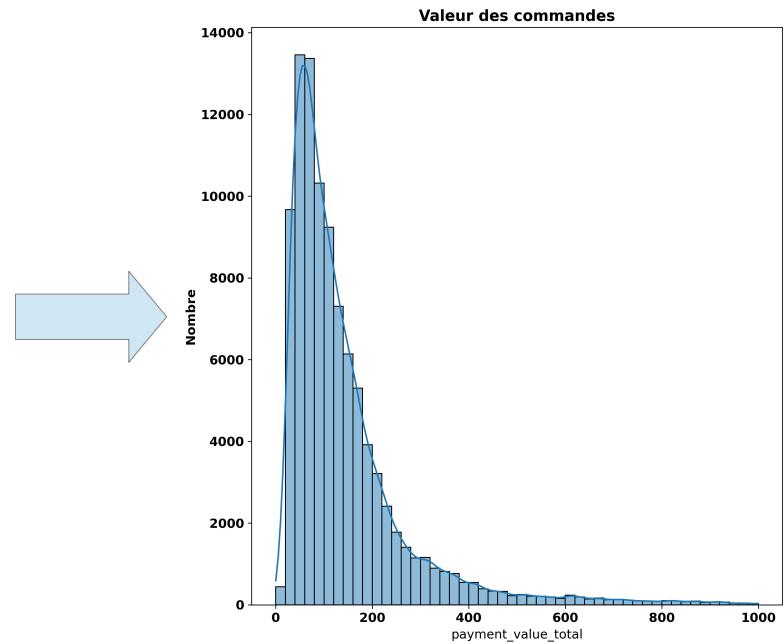
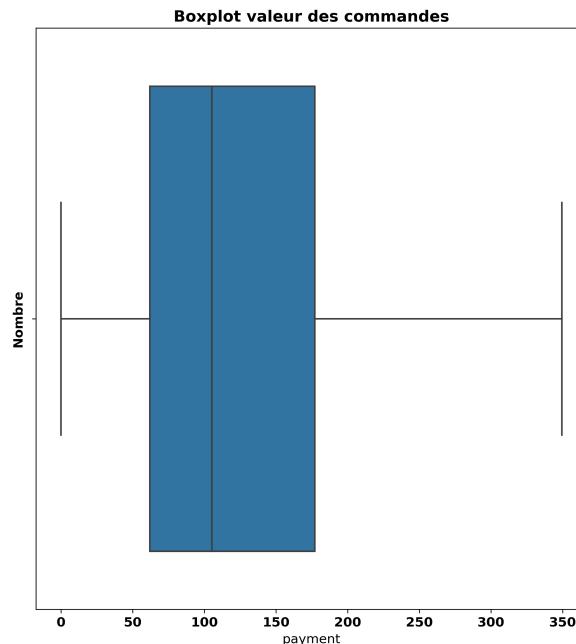
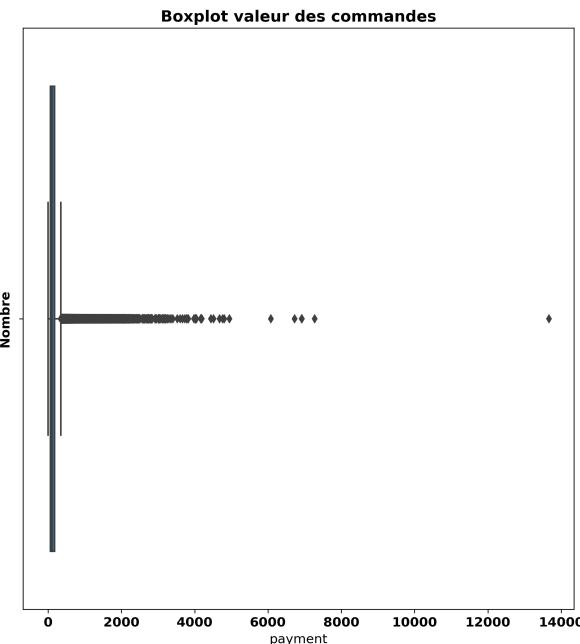


- Peu de paiement en plusieurs fois.
- Certaines commandes sont à montants élevés.
- L'essentiel des règlements se font en cartes bancaires.



II. Nettoyage des données / Analyse exploratoire / Feature engineering

→ Informations payements :

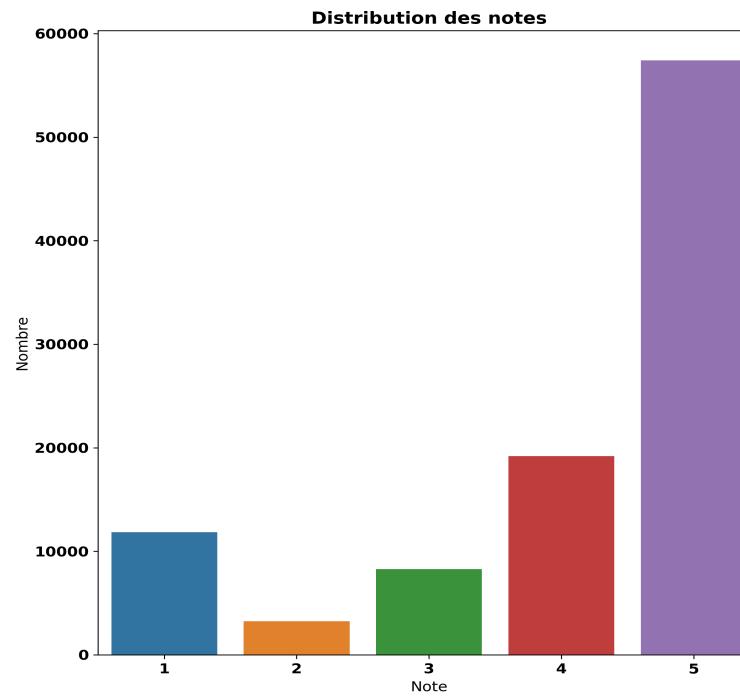
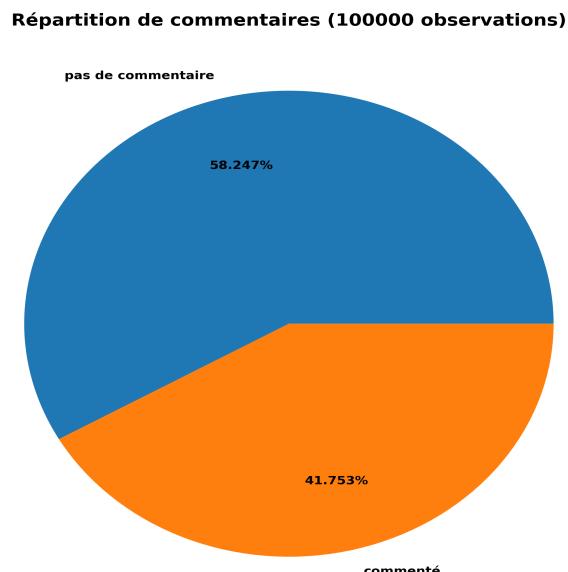


- Supprime les payements supérieurs à 1000 réal.
- Ces clients seront considérés comme 'à part' (à ne pas oublier).



II. Nettoyage des données / Analyse exploratoire / Feature engineering

→ Informations reviews :



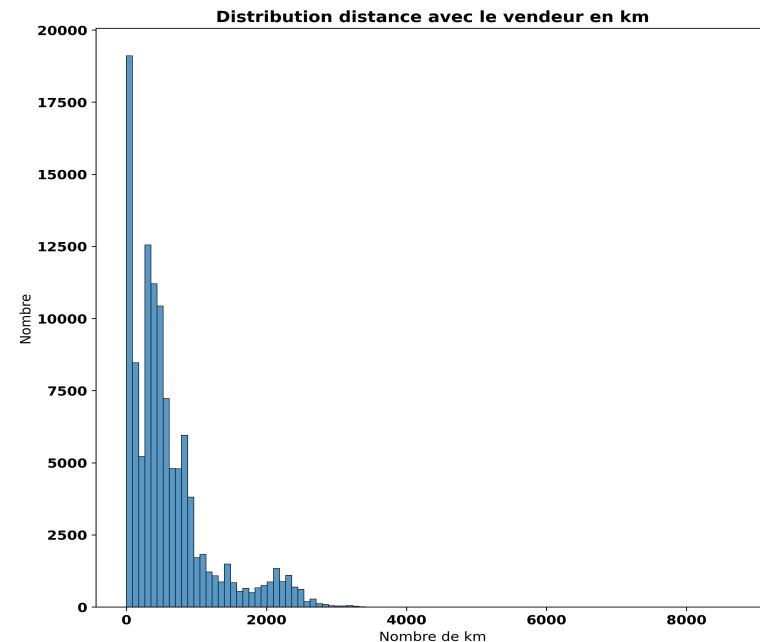
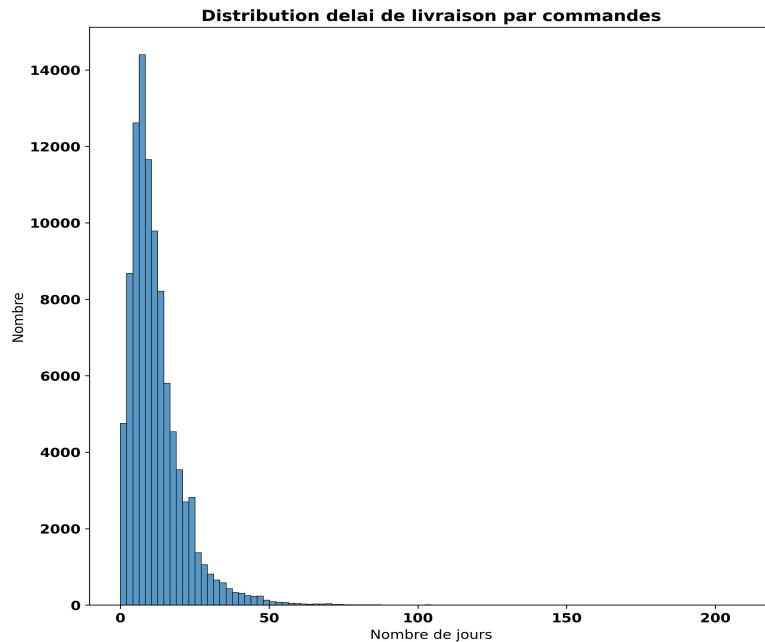
- L'essentiel des 'reviews' ne sont pas commentées.
- Les notes pourraient être intéressantes à évaluer dans notre analyse (satisfaction client).



II. Nettoyage des données / Analyse exploratoire / Feature engineering

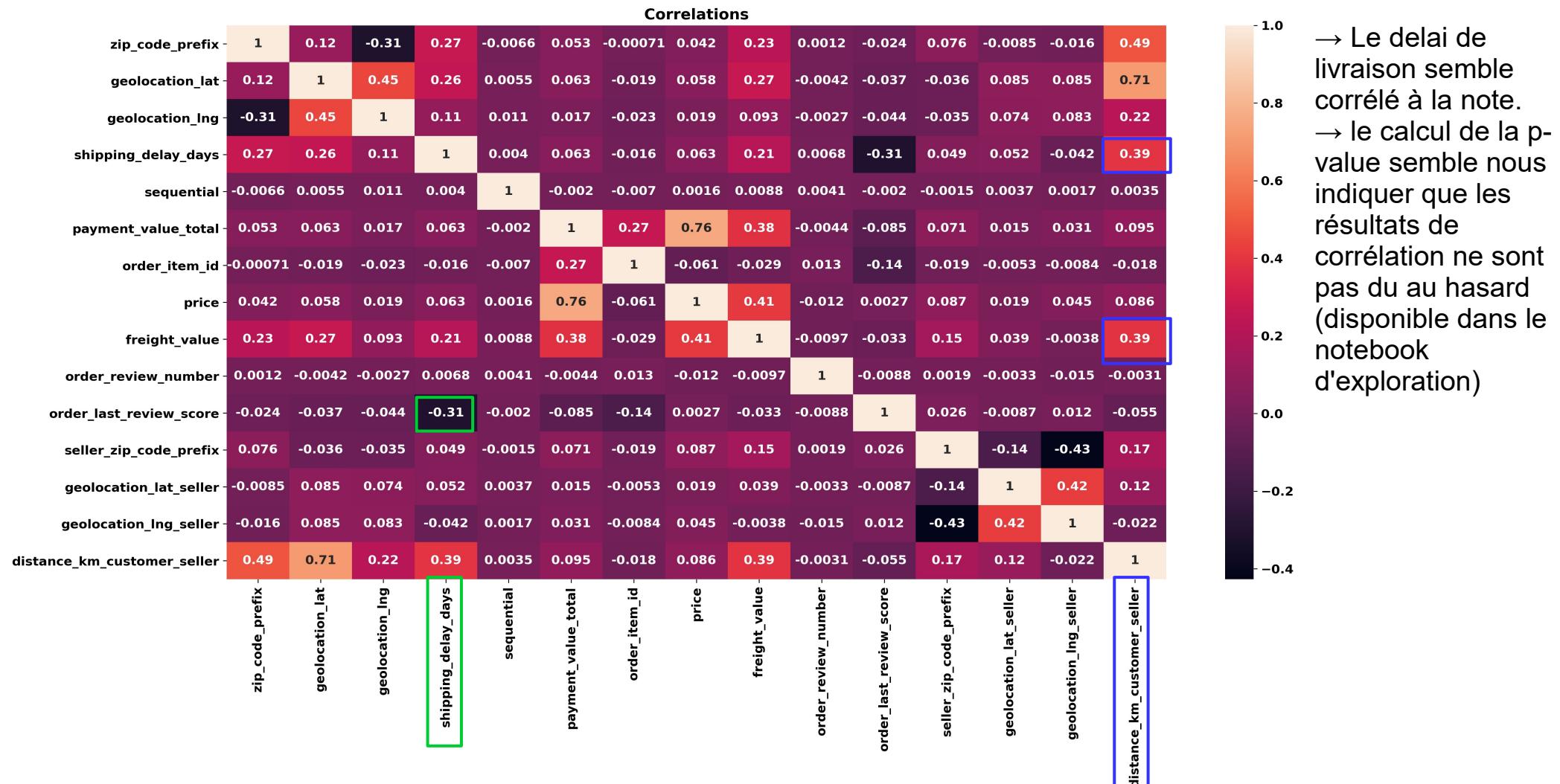
Feature engineering :

- Création de deux variables supplémentaires :
 - '*Shipping_delay_days*' : délai de livraison ('*order_purchase_timestamp*' - '*order_delivered_customer_date*')
 - '*distance_km_customer_seller*' : distance entre le vendeur et le client pour chaque produit (fonction avec latitude et longitude client / vendeur)



II. Nettoyage des données / Analyse exploratoire / Feature engineering

Analyse des corrélations linéaires :

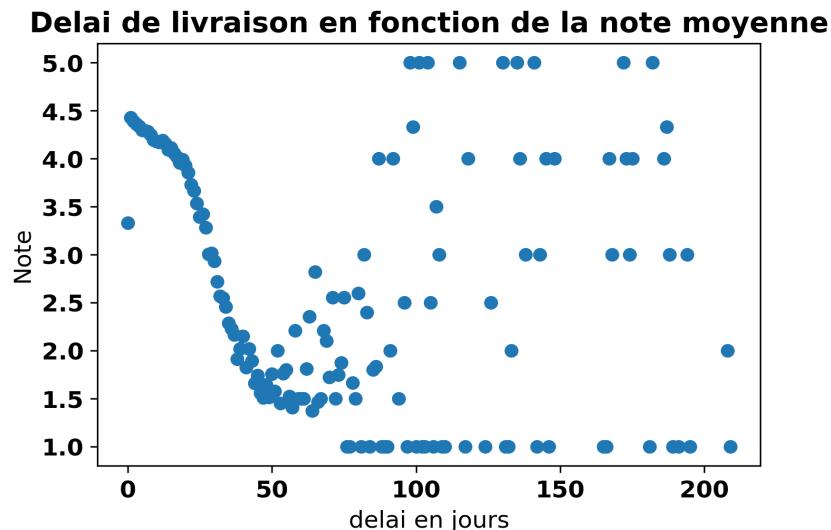


→ Le délai de livraison semble corrélé à la note.
→ le calcul de la p-value semble nous indiquer que les résultats de corrélation ne sont pas du au hasard (disponible dans le notebook d'exploration)

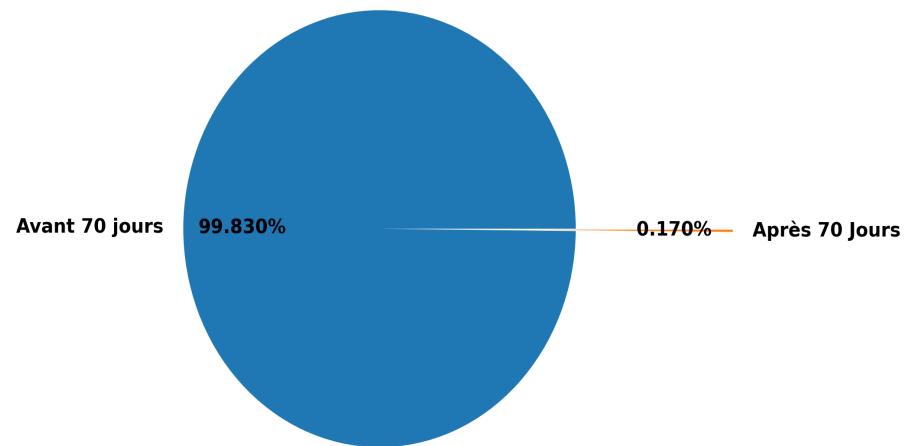


II. Nettoyage des données / Analyse exploratoire / Feature engineering

Delai de livraison en fonction de la note :



Répartition nombre de points utilisés(107936 observations)



- On peut raisonnablement confirmer notre corrélation.
- La tendance avant 70 jours représente 99% des points utilisés.



II. Nettoyage des données / Analyse exploratoire / Feature engineering

- Mise en place de l'analyse : fonction pour créer un fichier à une ligne par client et par période temporelle (pour nous servir aux différentes analyses) avec comme variables :

- Variables quantitatives :

R / F / M / nb_prod / nb_prod_by_order / mean_seller_dist_km / mean_payment_by_order / mean_sequential mean_review_score

- Variables qualitatives :

main_prod_cat / top_seller / main_payment_type

- Exemple :

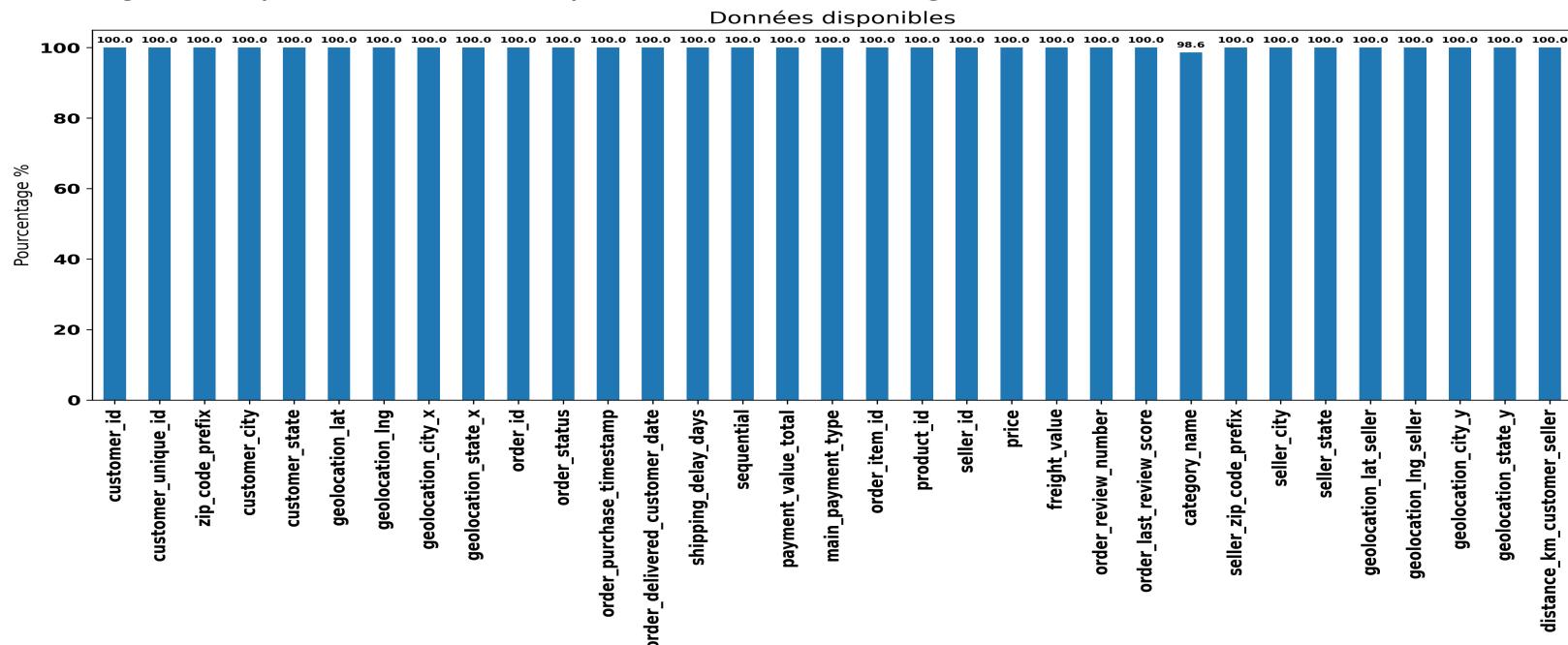
customer_unique_id	R	F	M	main_prod_cat	nb_prod	nb_prod_by_order	top_seller	mean_seller_dist_km	mean_payment_by_order	mean_sequential
0000366f3b9a7992bf8c76cfdf3221e2	111	1	141.90	bed_bath_table	1.0	1.0	da8622b14eb17ae2831f4ac5b9dab84a	109.307581	141.90	1.0
0000b849f77a49e4a4ce2b2a4ca5be3f	114	1	27.19	health_beauty	1.0	1.0	138dbe45fc62f1e244378131a6801526	22.853354	27.19	1.0
0000f46a3911fa3c0805444483337064	536	1	86.22	stationery	1.0	1.0	3d871de0142ce09b7081e2b9d1733cb1	517.883143	86.22	1.0
0000f6ccb0745a6a4b88665a16c9f078	320	1	43.62	telephony	1.0	1.0	ef506c96320abeedfb894c34db06f478	2484.002335	43.62	1.0
0004aac84e0df4da2b147fca70cf8255	287	1	196.89	telephony	1.0	1.0	70a12e78e608ac31179aea7f8422044b	154.073565	196.89	1.0



II. Nettoyage des données / Analyse exploratoire / Feature engineering

Au final:

- Suppression des outliers sur le montant des commandes(> 1000, 'client exceptionnels')
- Suppression des clients qui n'ont pas de localisation géographique / des commandes annulées et sans délai de livraison.
- Un fichiers global (base de travail) : 'df_olist_data_global.csv'



- 91830 clients et 94890 commandes (sur fin 2016 à 2018)



III. Etude RFM

A) Rappel :

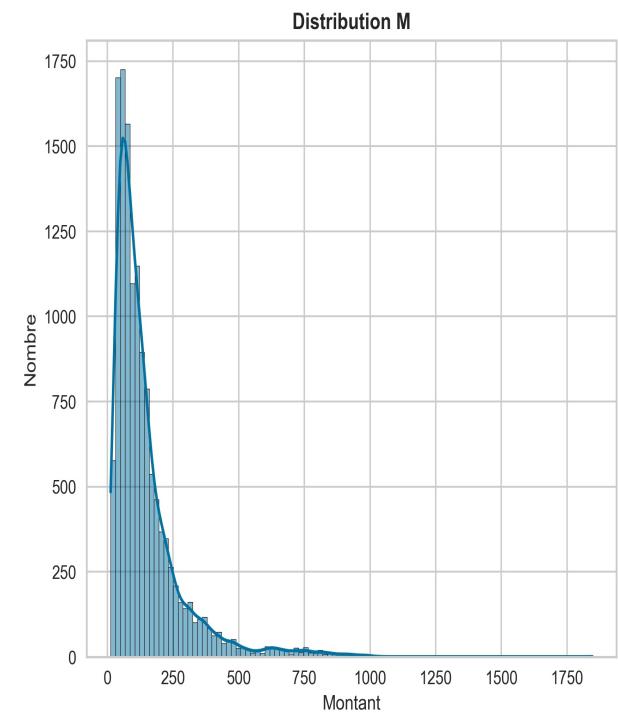
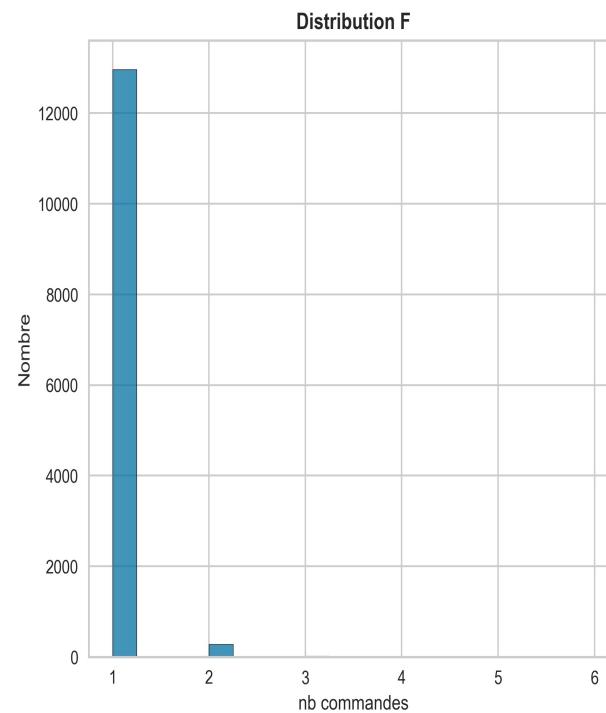
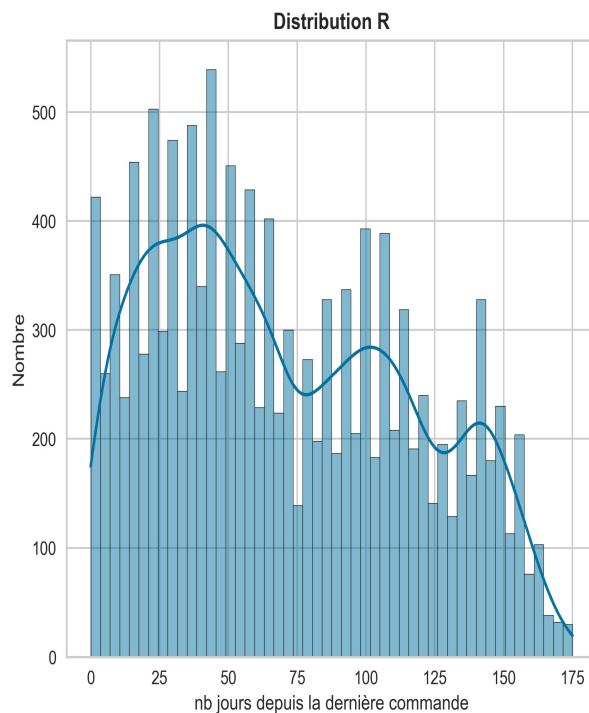
- L'analyse se base sur trois variables :
 - Recency (la date du dernier achat).
 - Frequency (la fréquence d'achat, le nombre de commandes)
 - Monetary (le montant d'achats, total ici)
- On définit une période temporelle d'analyse de nos clients.
- On construit un score avec les trois variables qui permet de segmenter nos clients :
 - Dans notre cas on a séparé chaque variable en 3 (qcut() et cut())
 - On a construit un score en assemblant les trois notes (ex : '121')



III. Etude RFM

B) Période d'analyse :

- On a choisi le premier semestre 2017 (Janvier à Juin / 6 mois) :



- 13266 clients

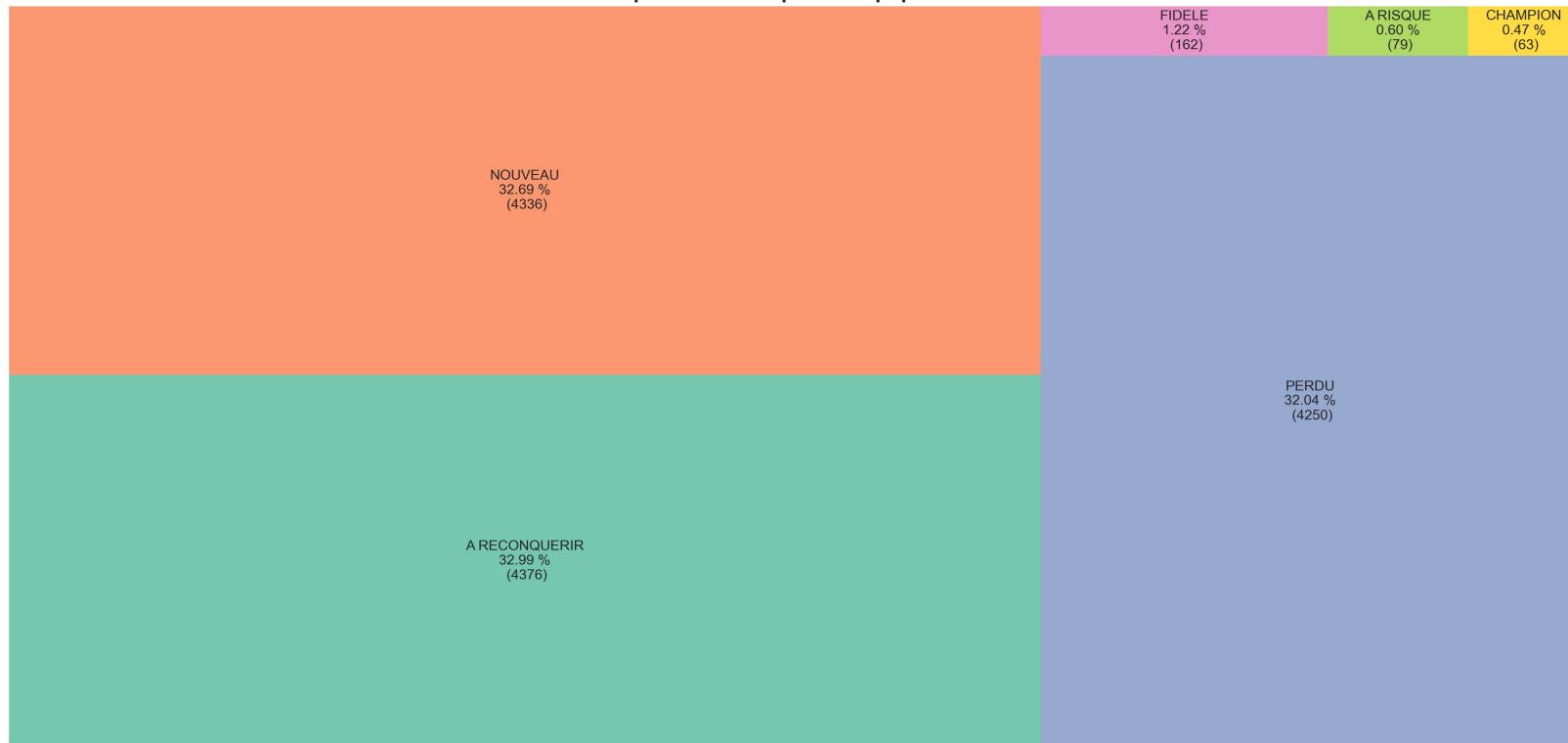


III. Etude RFM

C) Segmentation (6 groupes) :

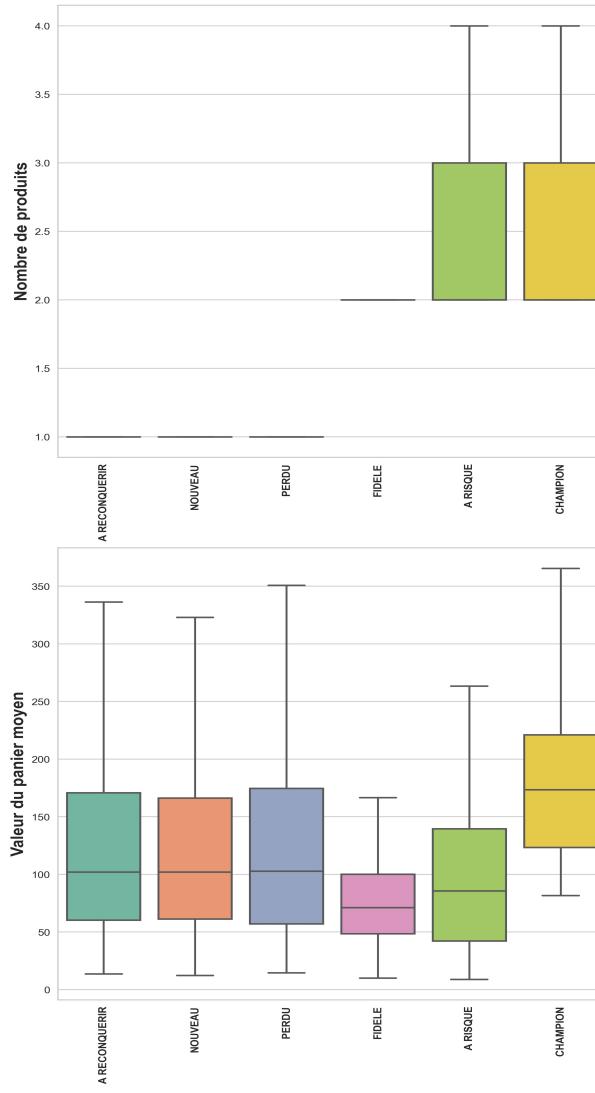
Champion → ['333', '323']
Fidèle → ['222', '223', '321', '322', '221']
Nouveau → ['311', '312', '313']
A risque → ['121', '122', '123', '132']
Perdu → ['111', '112', '113']
A reconquérir → ['211', '212', '213']

Répartition: Groupes RFM population

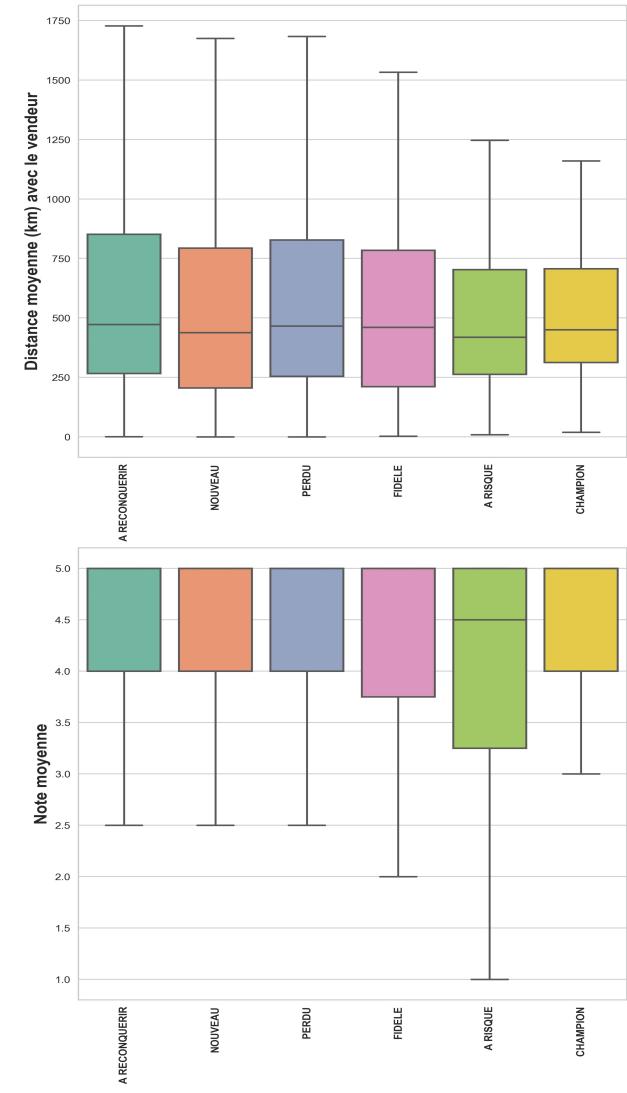
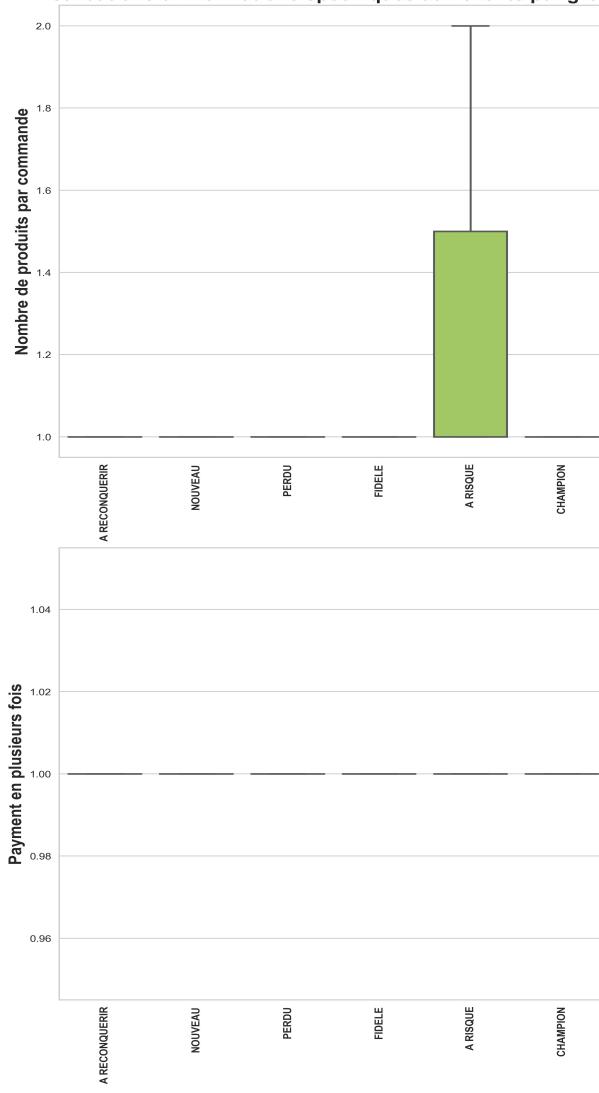


III. Etude RFM

C) Segmentation (6 groupes) :



Distributions d'informations spécifiques aux clients par groupe



IV. Modèles de clustering (RFM +)

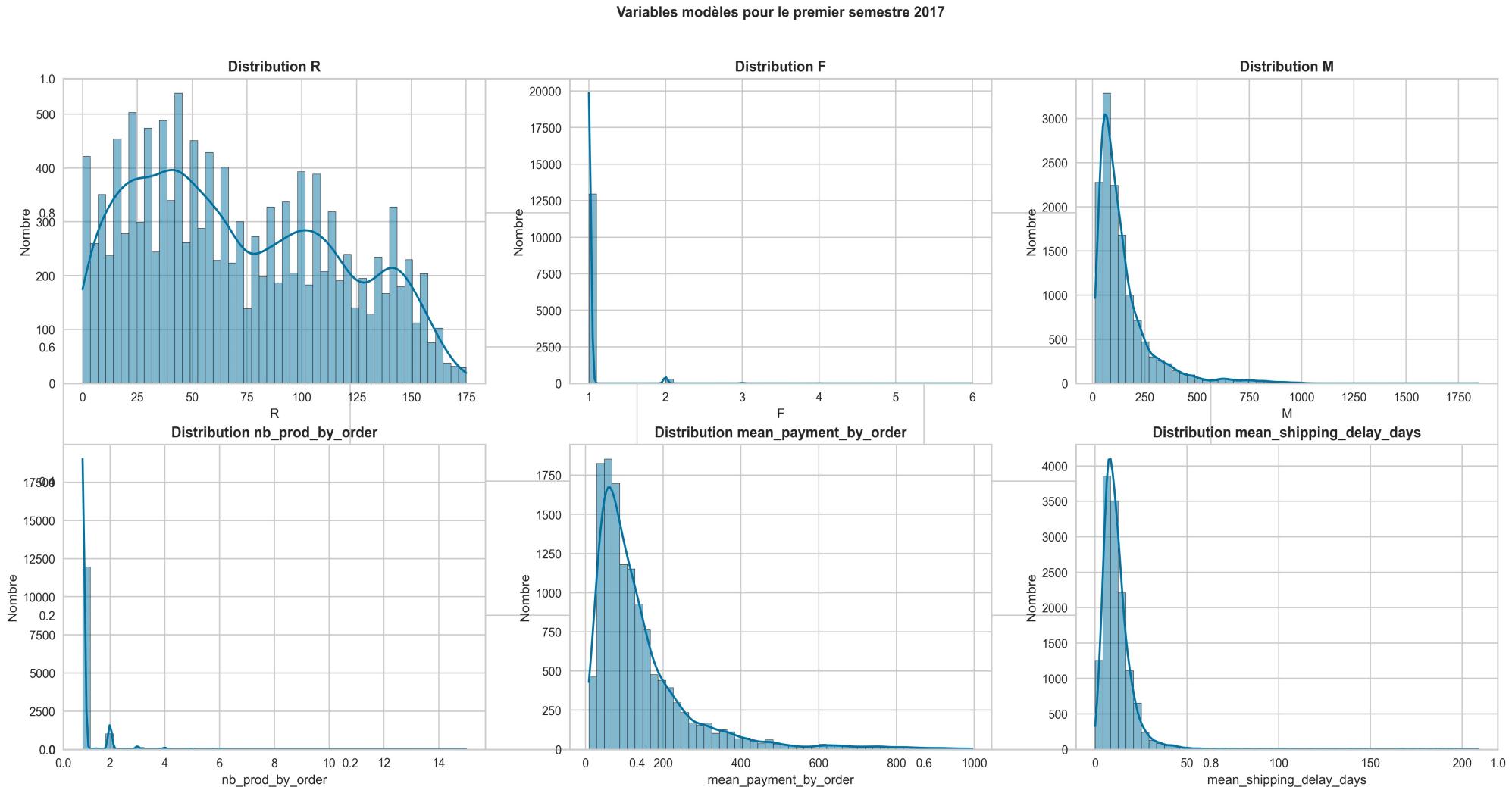
A) Les variables utilisées :

- Ajout de trois variables (en plus des RFM) :
 - *'nb_prod_by_order'*
 - *'mean_payment_by_order'*
 - *'mean_shipping_delay'*
- Preprocessing :
 - Log+1 sur les variables décentrées (décalées à gauche)
 - Standard_scaler



IV. Modèles de clustering (RFM +)

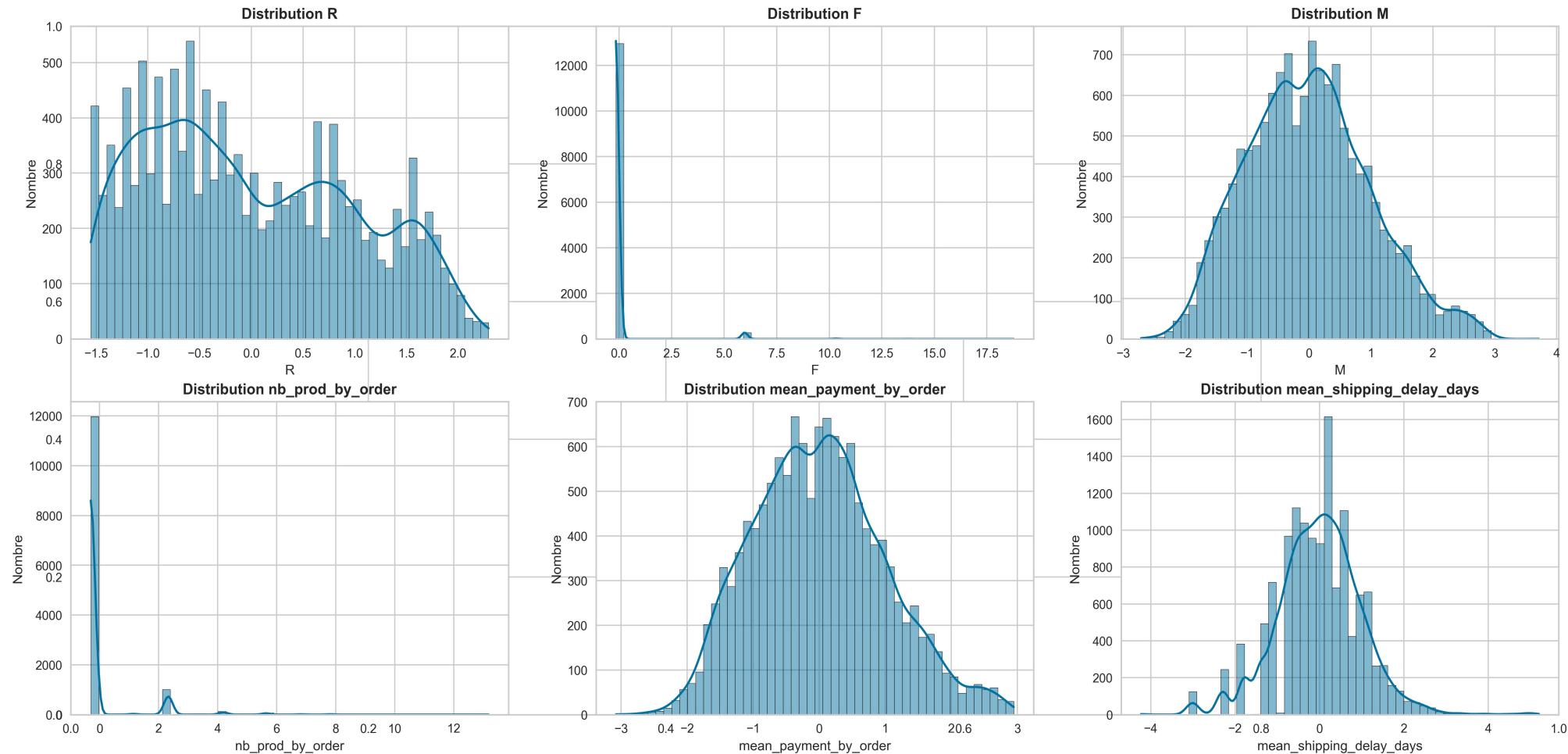
A) Les variables utilisées :



IV. Modèles de clustering (RFM +)

A) Les variables utilisées :

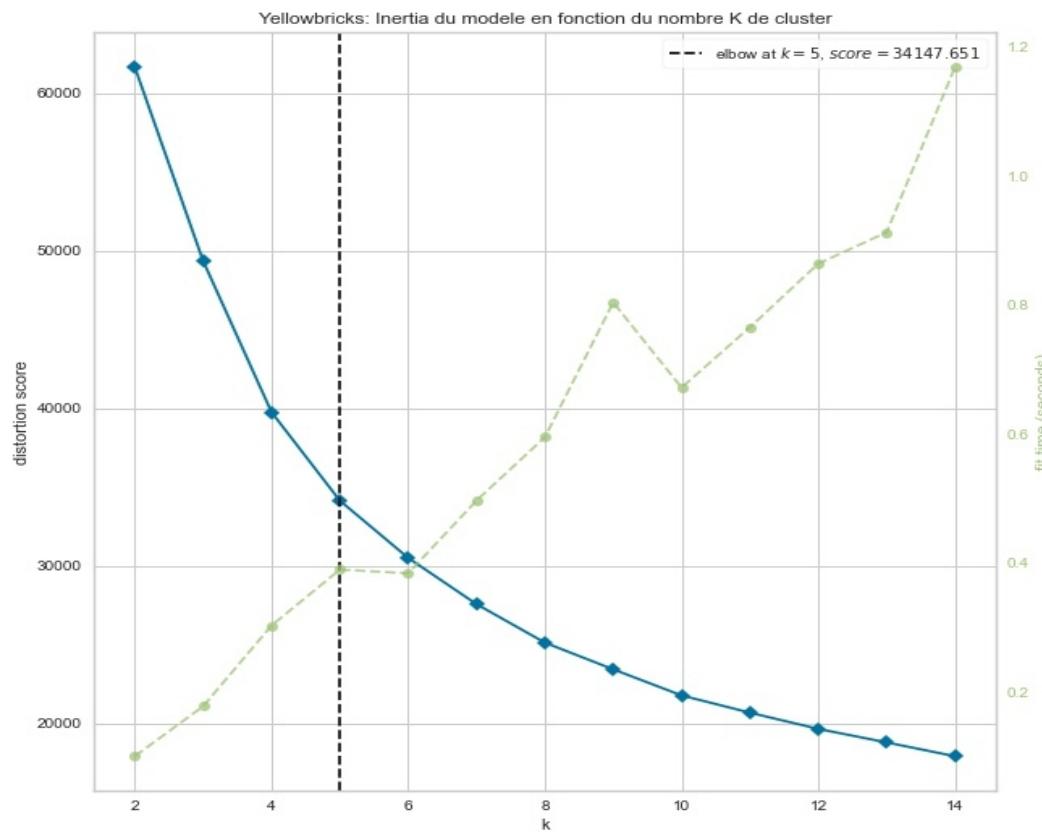
Variables après preprocessing pour le premier semestre 2017



IV. Modèles de clustering (RFM +)

B) Modèle K_Means :

- Optimisation :



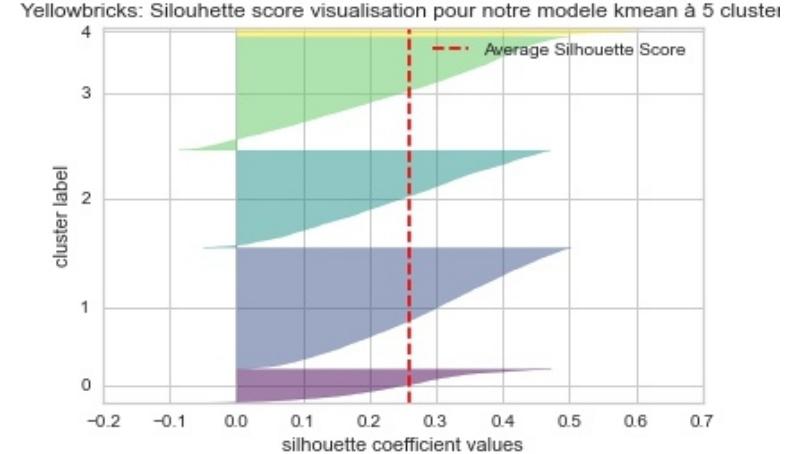
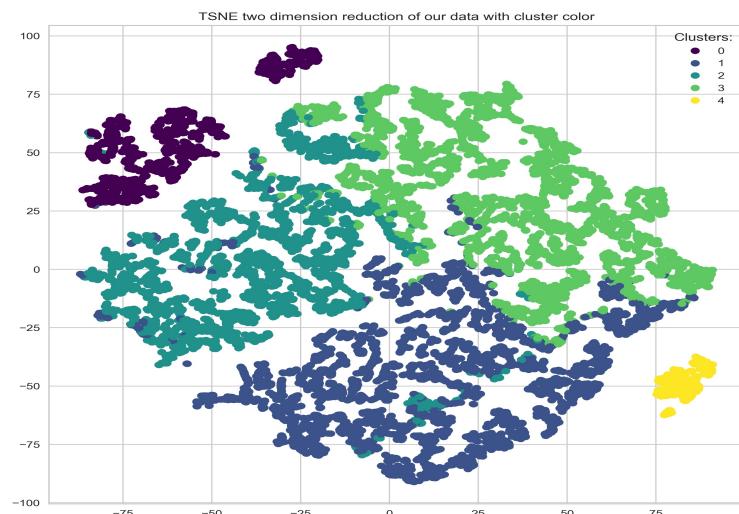
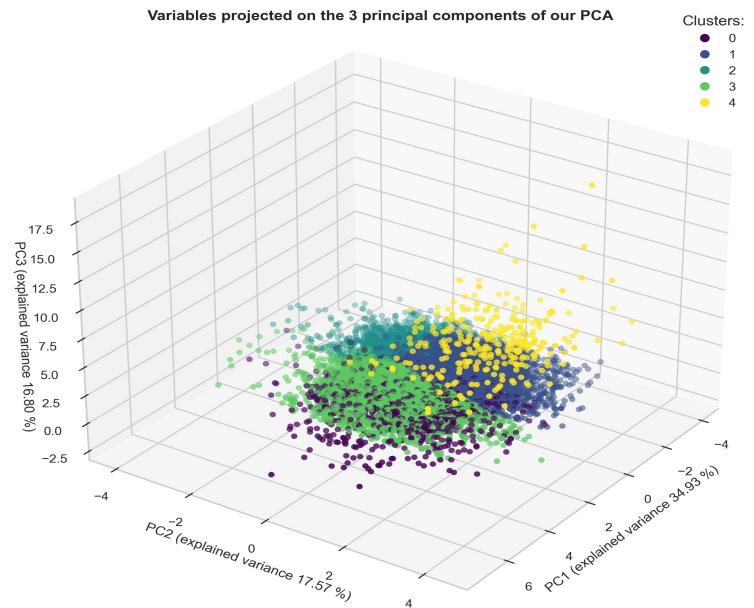
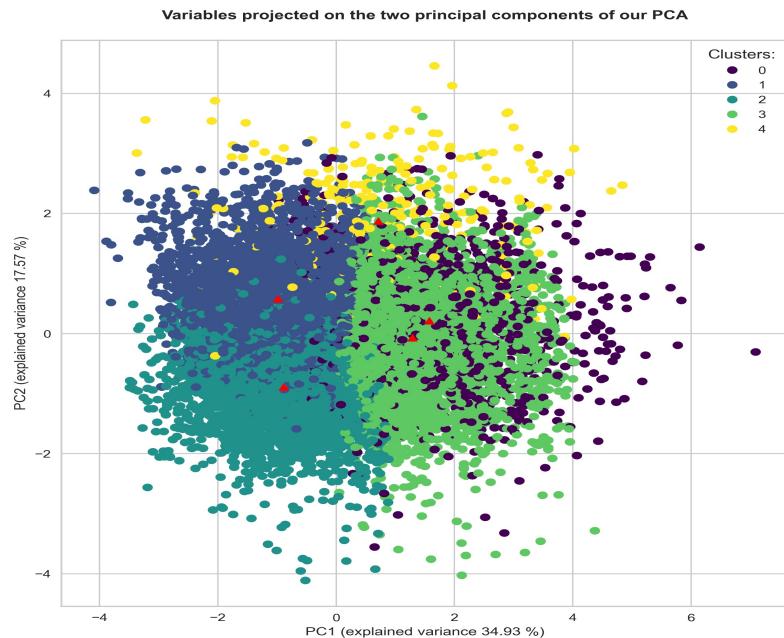
→ On choisit un modèle avec 5 clusters.



IV. Modèles de clustering (RFM +)

B) Modèle K_Means :

- Visualisations :



SENECHAL Yannick

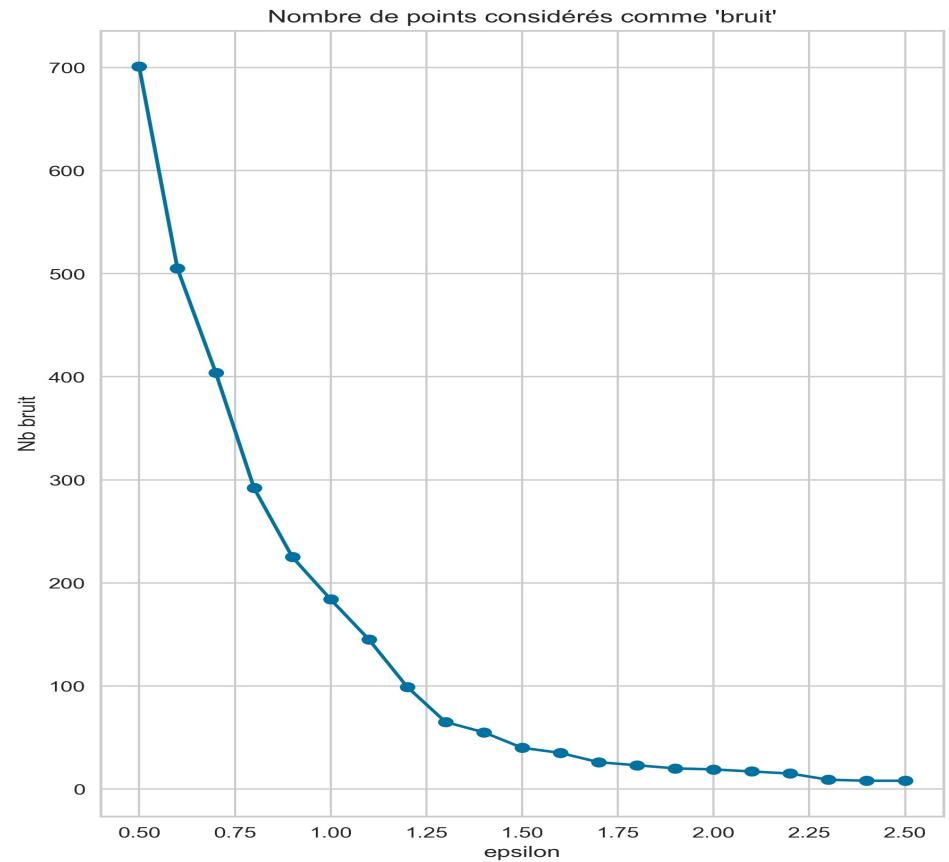
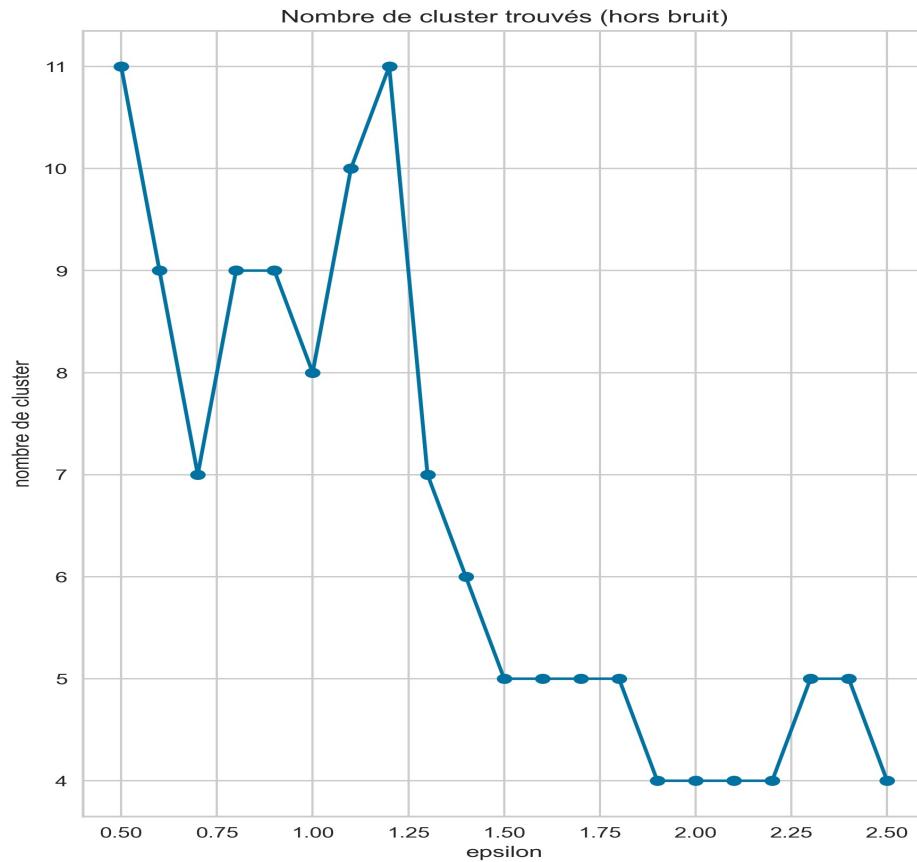


IV. Modèles de clustering (RFM +)

C) Modèle DBSCAN :

- Optimisation :

Optimisation 'epsilon' pour min_sample=6



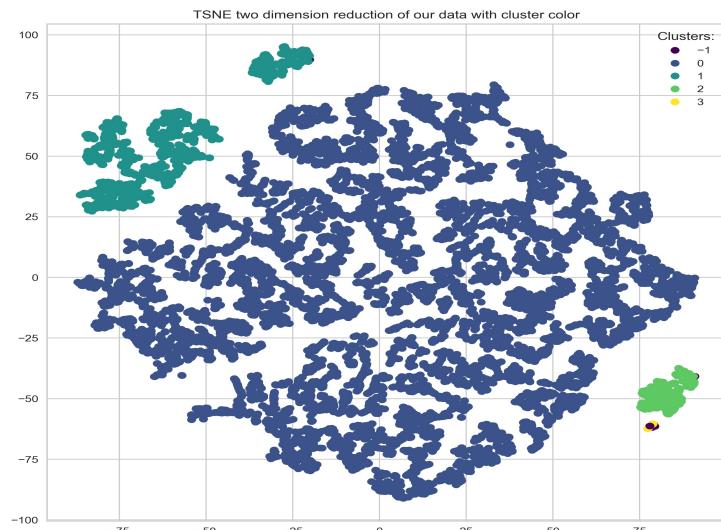
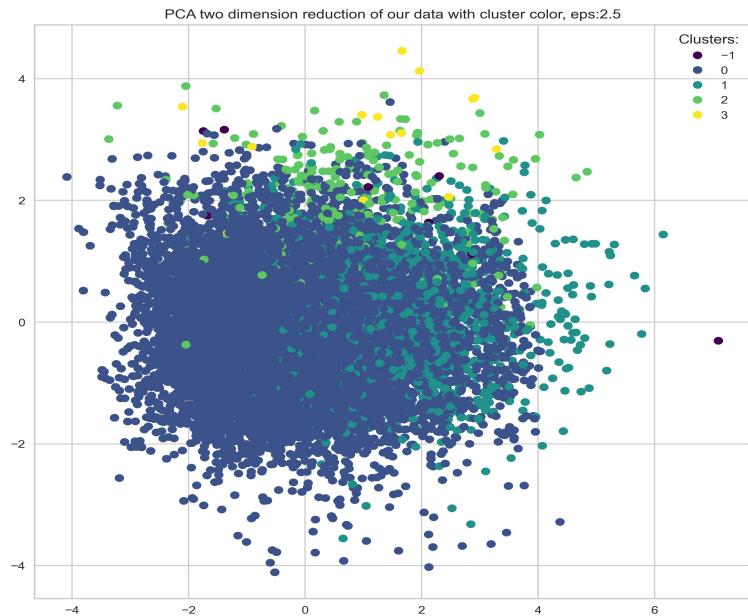
→ On choisit un modèle avec 4 clusters.



IV. Modèles de clustering (RFM +)

C) Modèle DBSCAN :

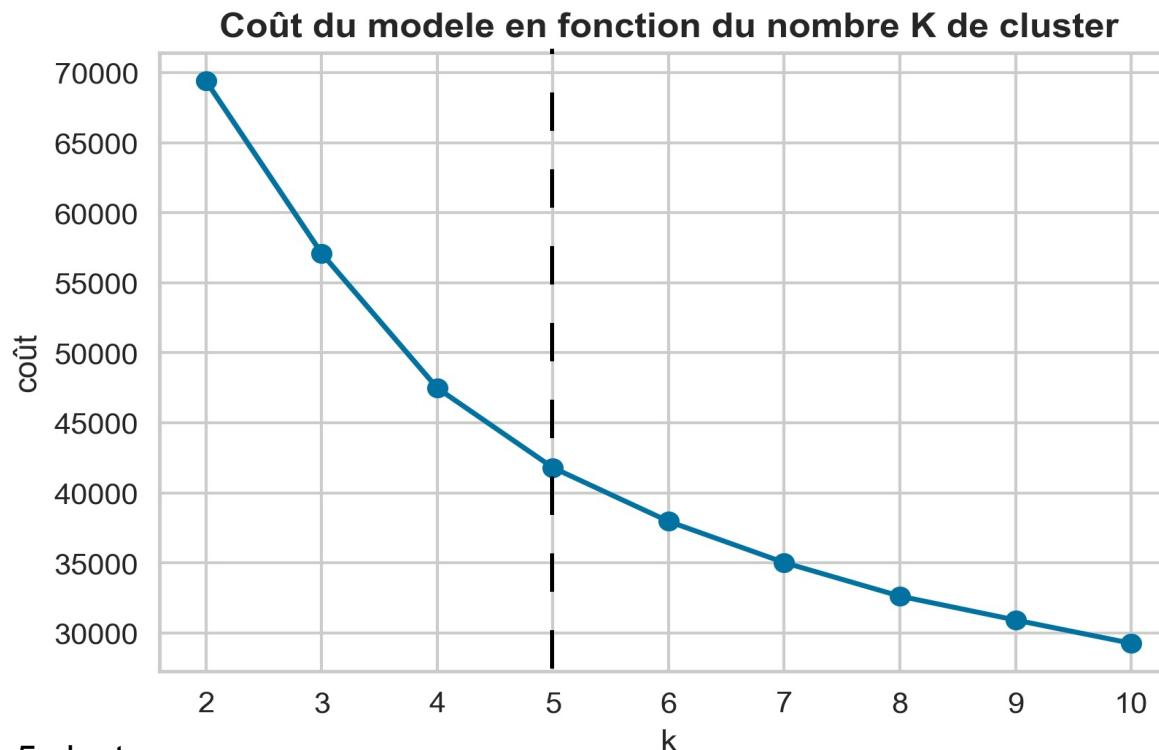
- Visualisations :



IV. Modèles de clustering (RFM +)

D) Modèle K_Prootypes :

- Ajout de deux variables catégorielles : '*main_prod_cat*', '*main_payment_type*'
- Optimisation :



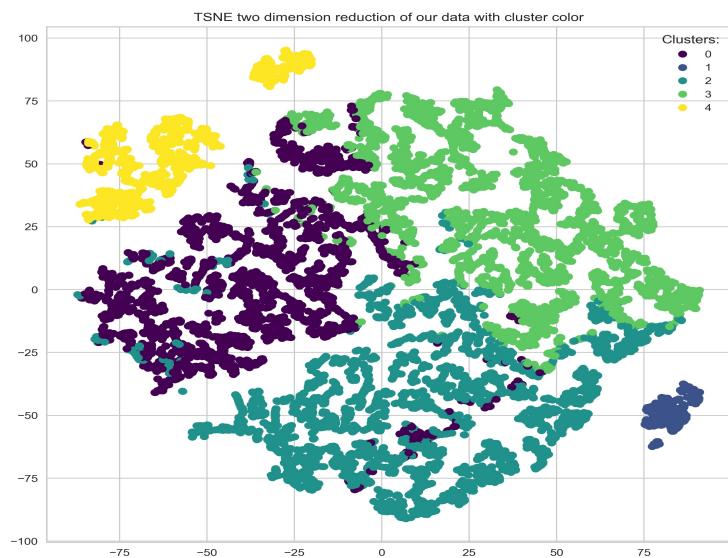
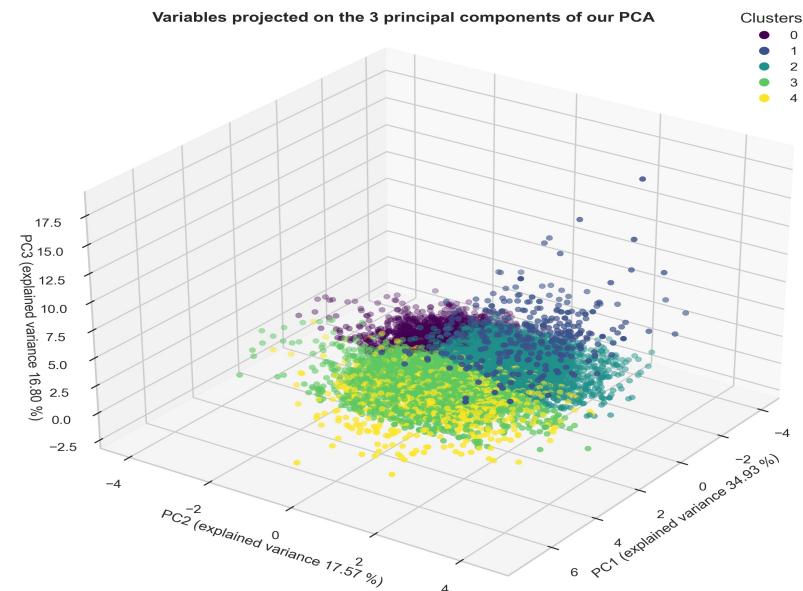
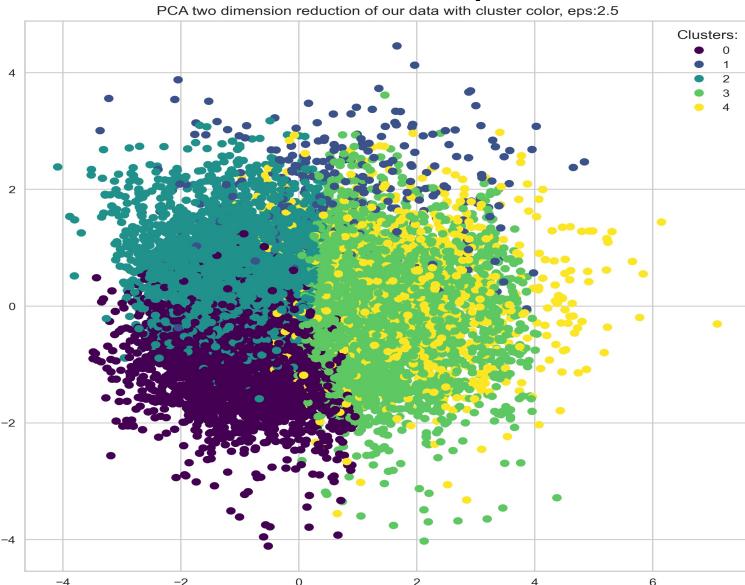
→ On choisit un modèle avec 5 clusters.



IV. Modèles de clustering (RFM +)

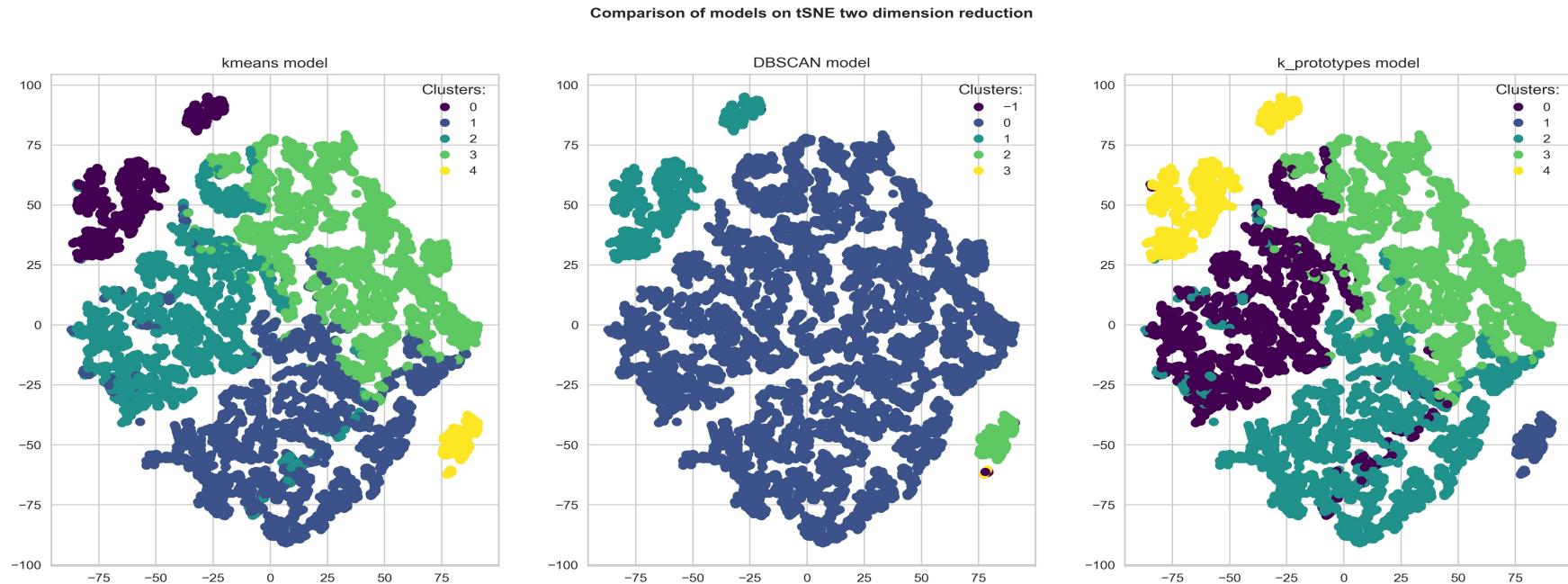
D) Modèle K_Protoypes :

- Visualisations (mêmes dimensions que les autres modèles) :



IV. Modèles de clustering (RFM +)

E) Comparaison des modèles :



	Model	Parameter_nb_cluster	Davies_boulin_score	Silhouette_score
0	Kmeans model	5	1.121932	0.259692
1	DBSCAN model	4	1.703873	0.390338
2	K_prototypes model	5	1.124184	0.258775



V. Modèle retenu (K_means)

A) Identification des clusters (5 groupes) :

Acheteur compulsif → 9.1%

Panier moyen : 210.8 réal

Fréquence : 1

Délai de livraison moyen : 12.4

Nombre de produit par commande : 2.3

Note moyenne : 3.8

Nouveau (curieux) → 32.4%

Panier moyen : 70.6 réal

Fréquence : 1

Délai de livraison moyen : 10.3

Nombre de produit par commande : 1

Note moyenne : 4.3

Occasionnel petit achats (perdu) → 26%

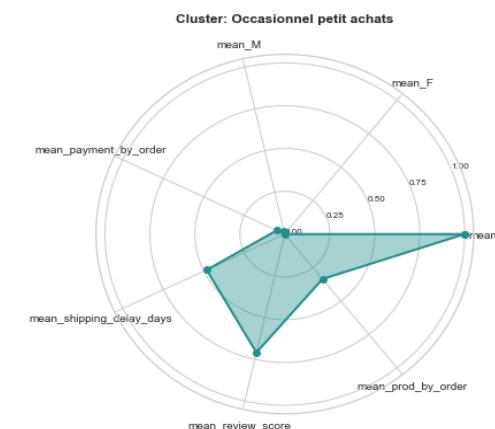
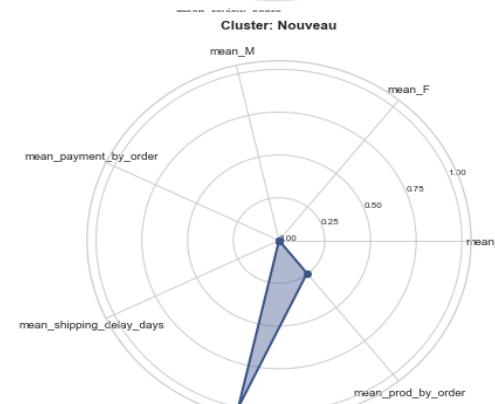
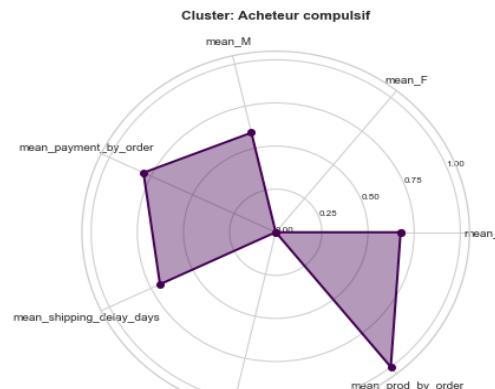
Panier moyen : 73.4 réal

Fréquence : 1

Délai de livraison moyen : 12.4

Nombre de produit par commande : 1

Note moyenne : 4.2



V. Modèle retenu (K_means)

A) Identification des clusters (5 groupes) :

Occasionnel dépensier → 30.3%

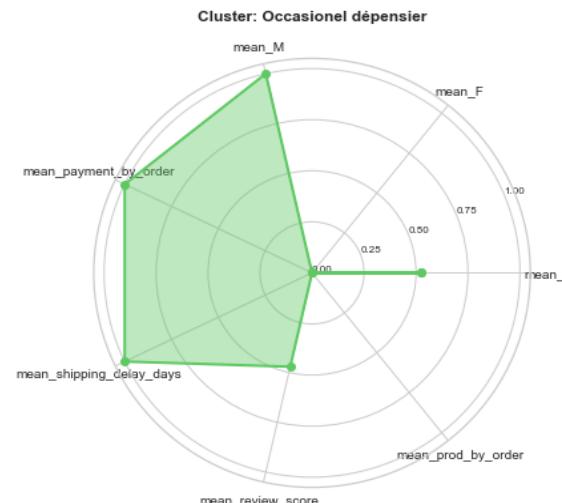
Panier moyen : 258.2 réel

Fréquence : 1

Délai de livraison moyen : 13.9

Nombre de produit par commande : 1

Note moyenne : 4.2



Bon client → 2.3%

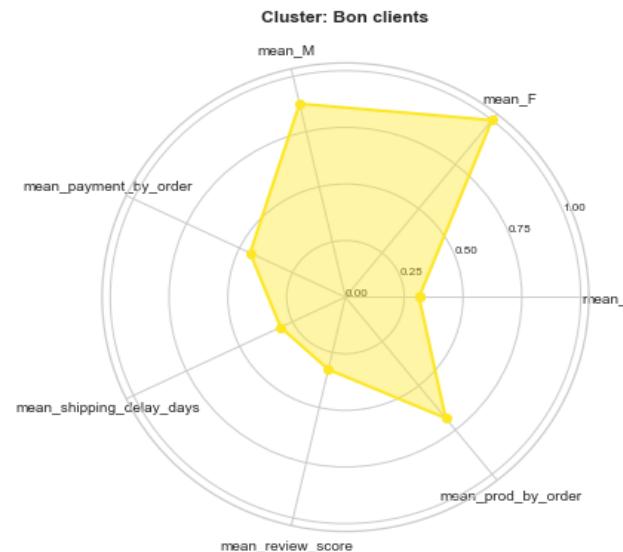
Panier moyen : 121.4 réel

Fréquence : 2.1

Délai de livraison moyen : 11.6

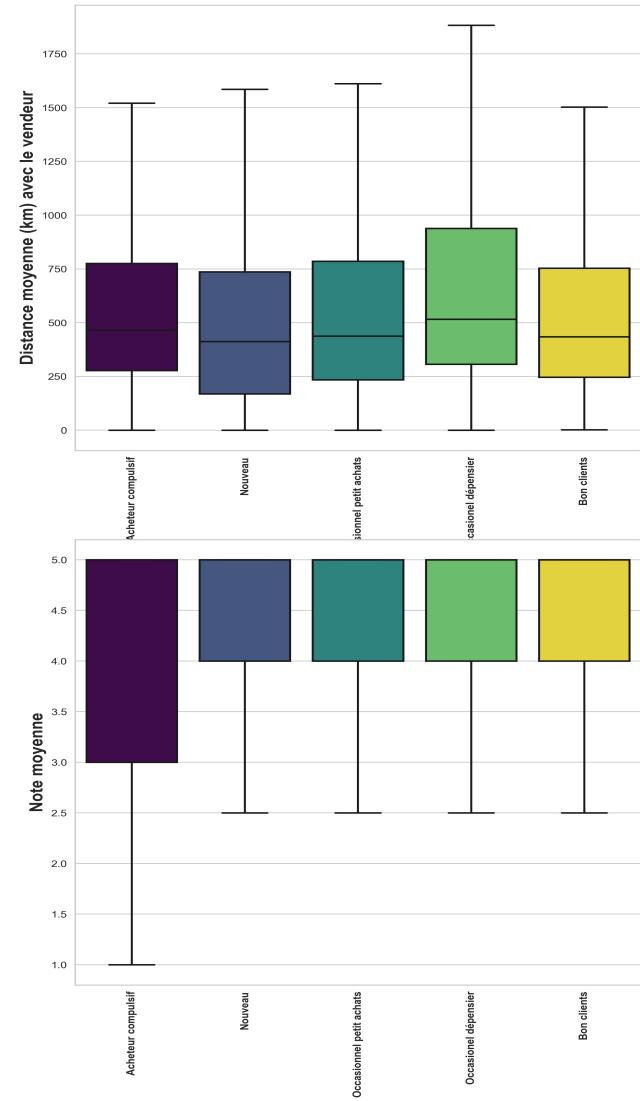
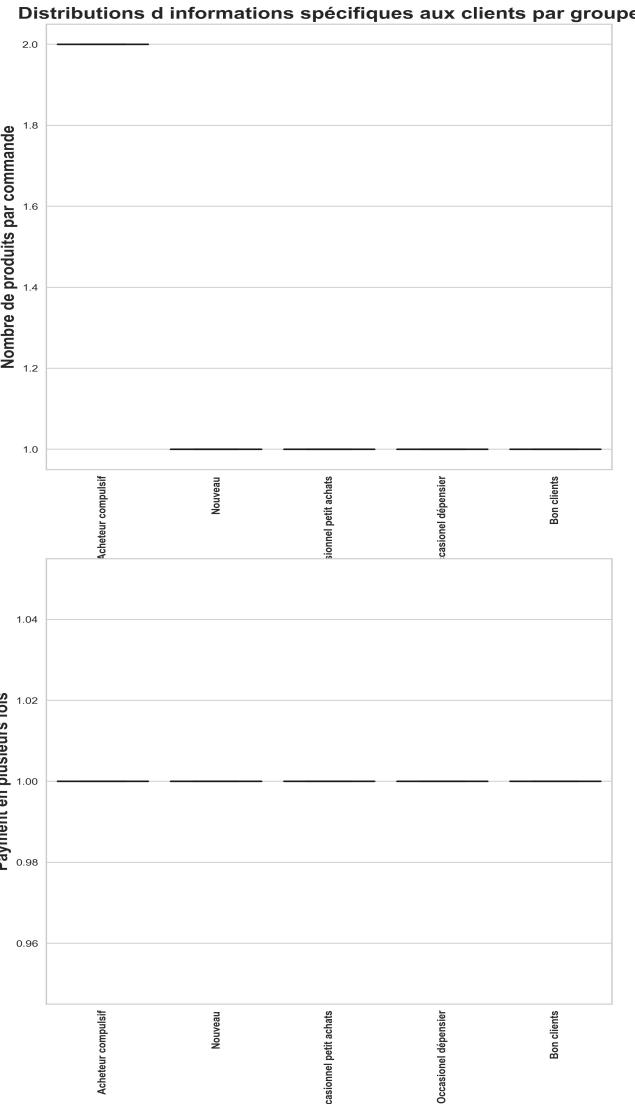
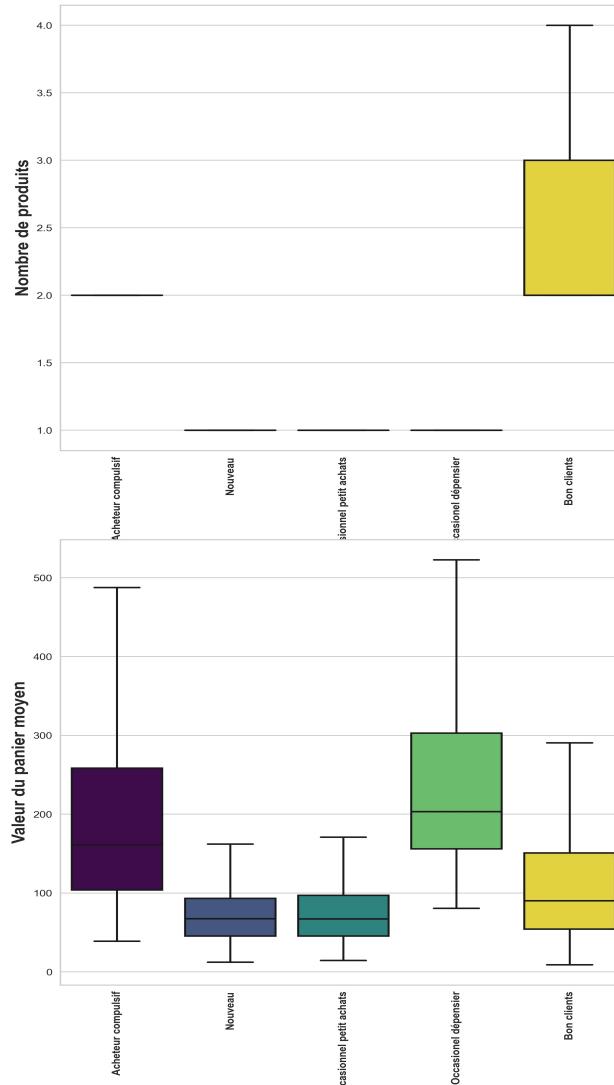
Nombre de produit par commande : 1.2

Note moyenne : 4.2



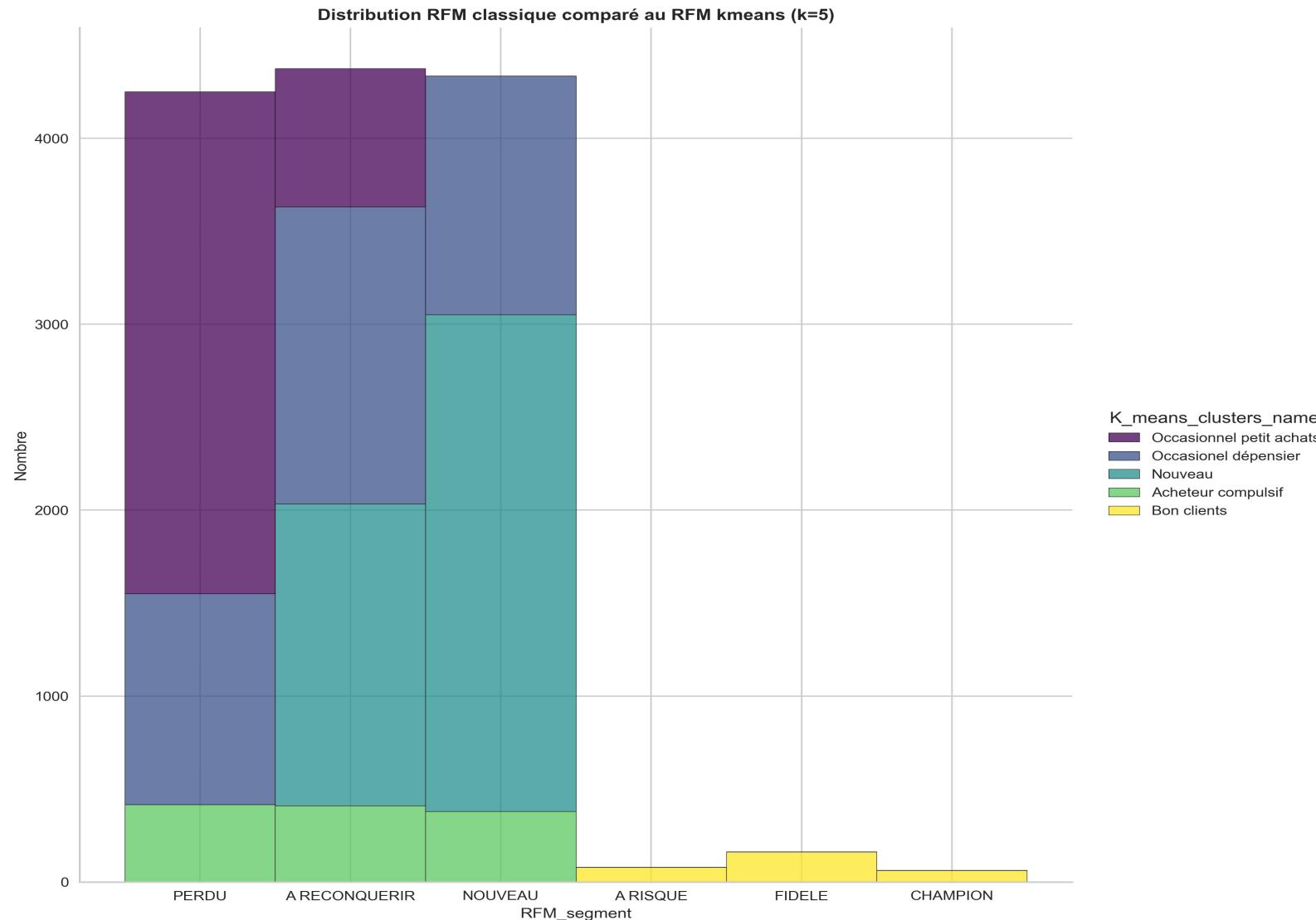
V. Modèle retenu (K_means)

B) Informations complémentaires:



V. Modèle retenu (K_means)

C) Comparaison avec analyse RFM :



VI. Contrat de maintenance

A) Évaluation de la stabilité du cluster :

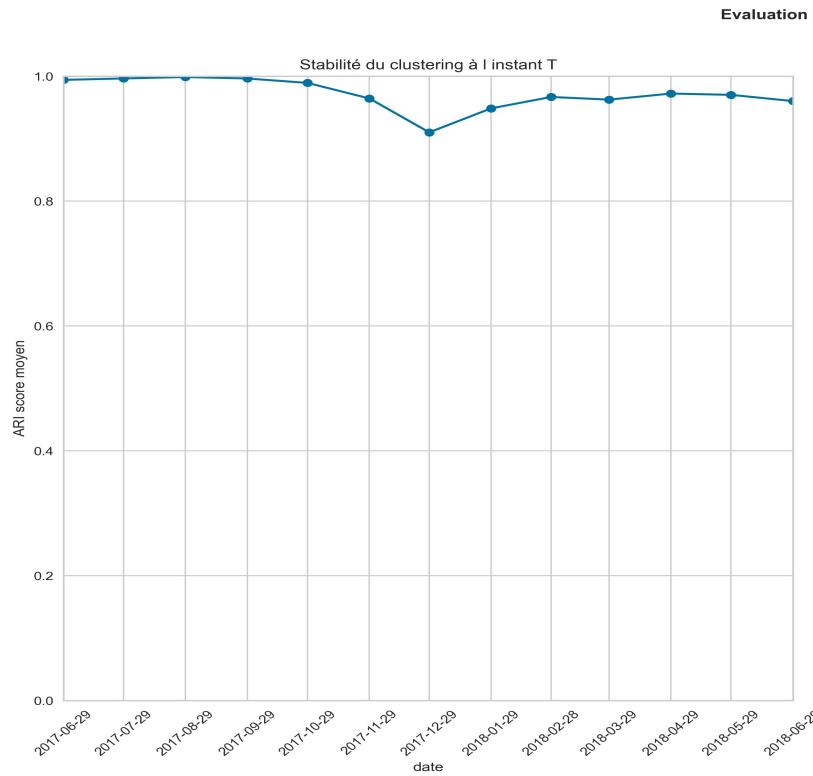
- Fonction pour évaluer (via ARI score) la stabilité du cluster (init method kmeans++) à l'instant T.
- On entraîne 100 modèles de clusters sur notre période puis on évalue l'ARI score pour chaque couple de clusters
- ARI moyen modèle de base = 0.99



VI. Contrat de maintenance

B) Évaluation du modèle dans le temps :

→ On ajoute mois par mois via notre fichier global des données sur les mêmes clients et on compare le modèle de base avec le nouveau modèle (ARI) :



→ Mise à jour du modèle au bout de 7 mois.



VII. Conclusion

- Proposition d'un modèle de clustering pour segmentation client : K-means à 5 clusters (avec leur description)
- Ne pas oublier le cluster des clients qui dépensent beaucoup.
- Contrat de maintenance proposé tous les 7 mois.
- Pour aller plus loin :
 - Plus de features engineering / sélection de variables plus fines (via informations métier)
 - Se pencher sur le K_prototypes?
 - Proposer plus d'informations sur les segments.
 - Proposer l'évolution des clients via Sankey?



Merci pour votre attention !
Questions ?

