



olist
empowering commerce

Segmentez des clients d'un site e-commerce

Data & Analytics
Eric Blanvillain - 21-12-2021

Problématique

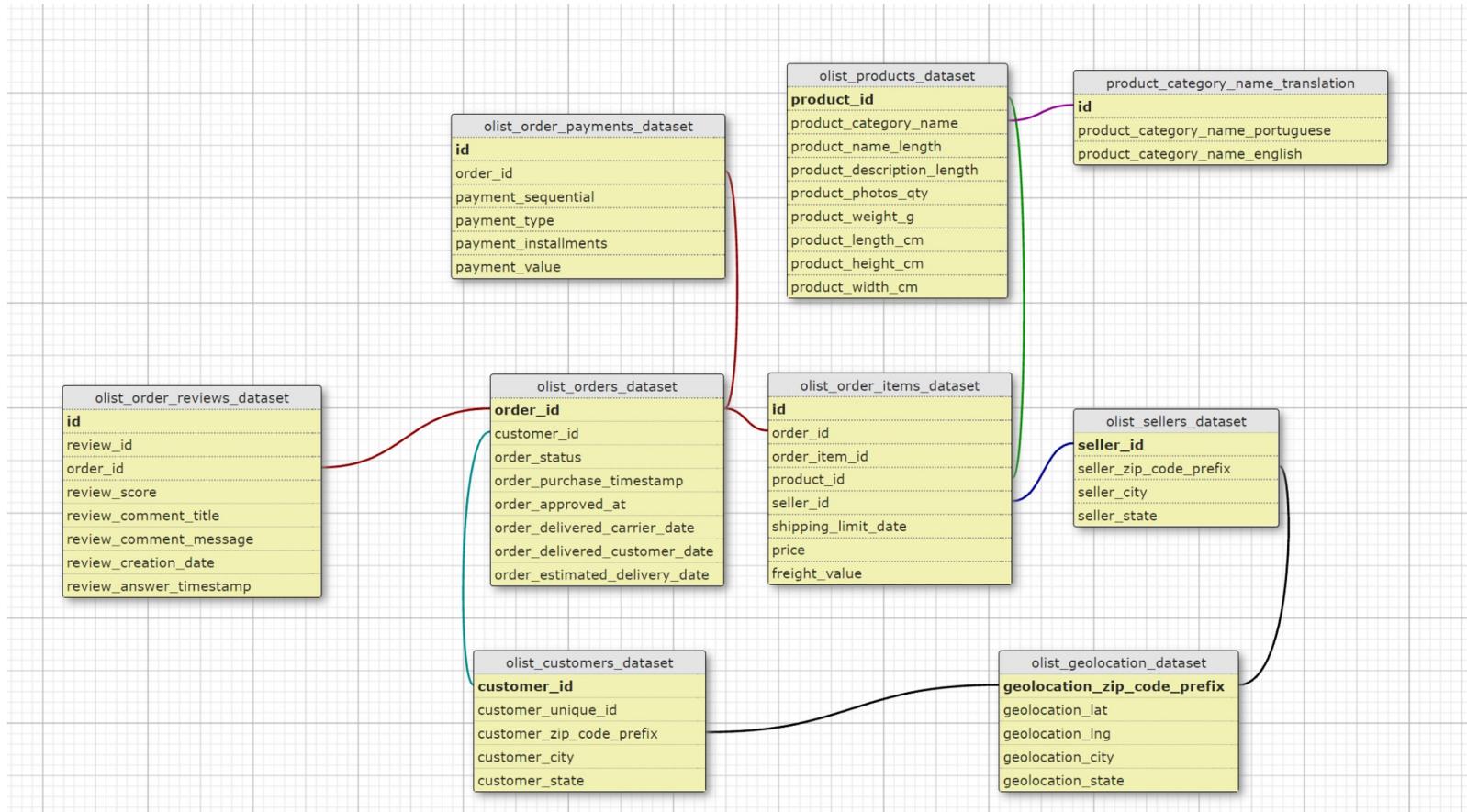
Mon rôle :

- Olist souhaite que je fournisse à leurs équipes d'e-commerce une **segmentation des clients** qu'elles pourront utiliser au quotidien pour leurs **campagnes de communication**.
- L'objectif est de **comprendre les différents types d'utilisateurs** grâce à leur comportement et à leurs données personnelles.
- Je devrais **fournir à l'équipe marketing une description actionable** de votre segmentation et de sa logique sous-jacente pour une utilisation optimale, ainsi qu'une **proposition de contrat de maintenance** basée sur une analyse de la stabilité des segments au cours du temps.

Les points à aborder :

- Présentation de la base de donnée Olist et de la donnée étudiée
- Résumé de l'analyse de la donnée (chiffres d'affaires, saisonnalité, catégories de produits, régions clefs)
- Analyse de segmentation RFM puis CLV
- Etudes des différents modèles de clustering non-supervisé
- Clustering Kmeans avec YellowBrick
- Etude de stabilité des clusters avec le ARI Score

Structure de la base Base de Donnée



Information de la base Base de Donnée

Information de la donnée étudiée

Période étudié	23 mois (Oct 2016 - Oct 2018)
nombre de commandes	112 650
nombre de clients	99 441
nombre de produits	32 951
nombre de vendeurs	3 095

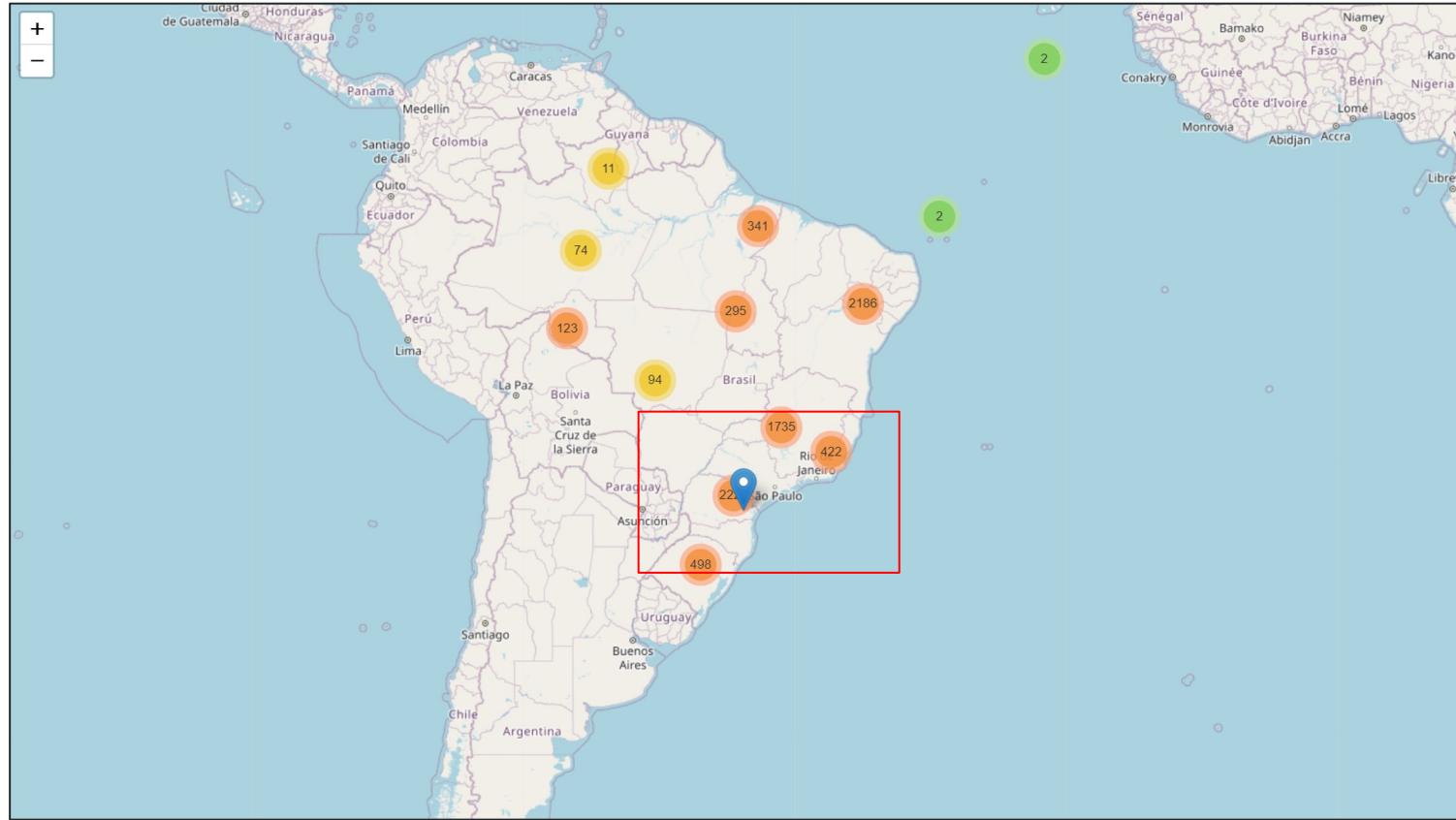


Base de donnée restreinte !

Résumé de l'exploration des données

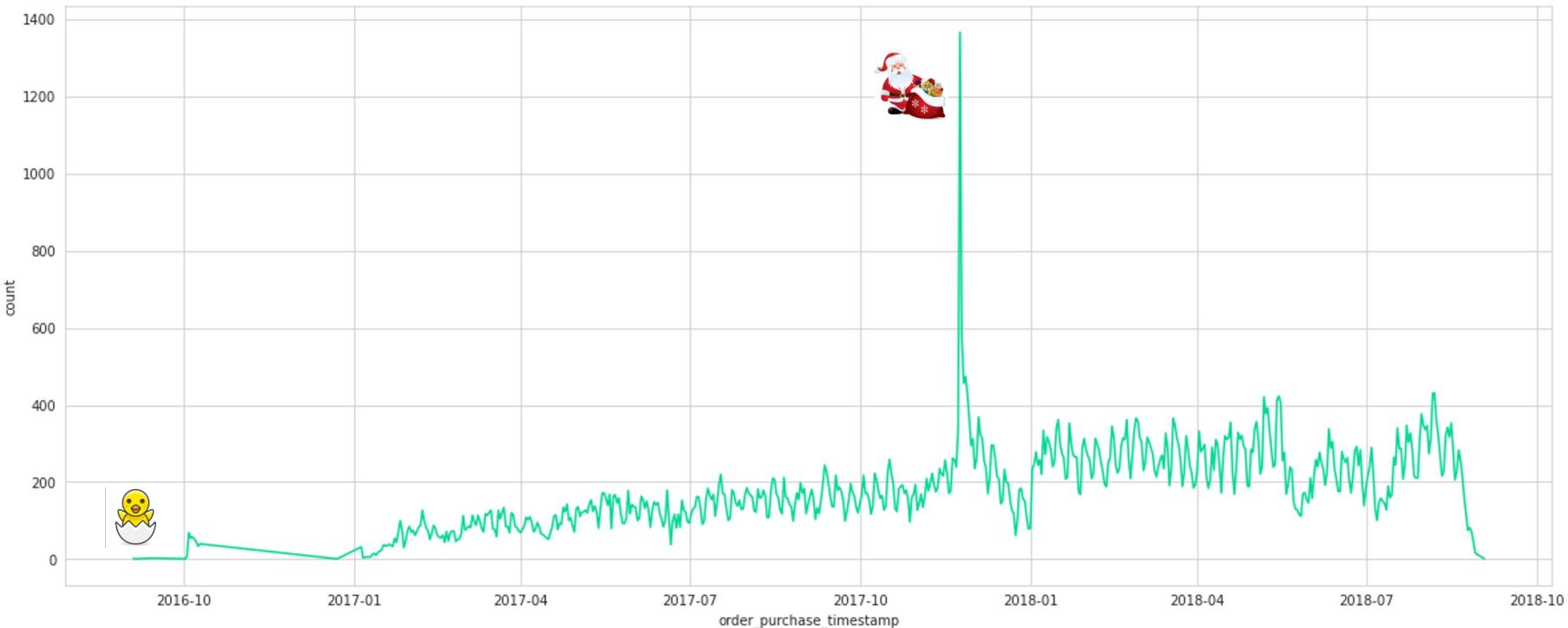
La donnée	Nouvelles features	Exemples de catégories possibles
Les commandes par mois / jours / heures	favorite_sale_day favorite_sale_month	Préférence pour le matin ou le soir / le weekend ou la semaine / la fin ou le début d'année
Le chiffre d'affaire par mois / états / clients Prix moyens des commandes / Nombre d'article par commandes	total_spend, mean_price_order total_items, mean_nb_items	Client profitable / non profitable / dépensier / non dépensier
Les moyens de paiement	mean_payment_sequential mean_payment_installments	Carte bancaire / autres moyens de paiement Paie en une fois / plusieurs fois
Les délais de livraison / Les coûts de livraison	mean_delivery_days order_mean_delay / freight_ratio	Délais de livraison courts / longs Coûts de livraison élevés / faibles
Les notes attribuées	mean_review_score	Client satisfait / non satisfait
Nombre de commandes par clients	nb_orders	Client récurrent / non récurrent

Localisation de Olist et des villes du dataset



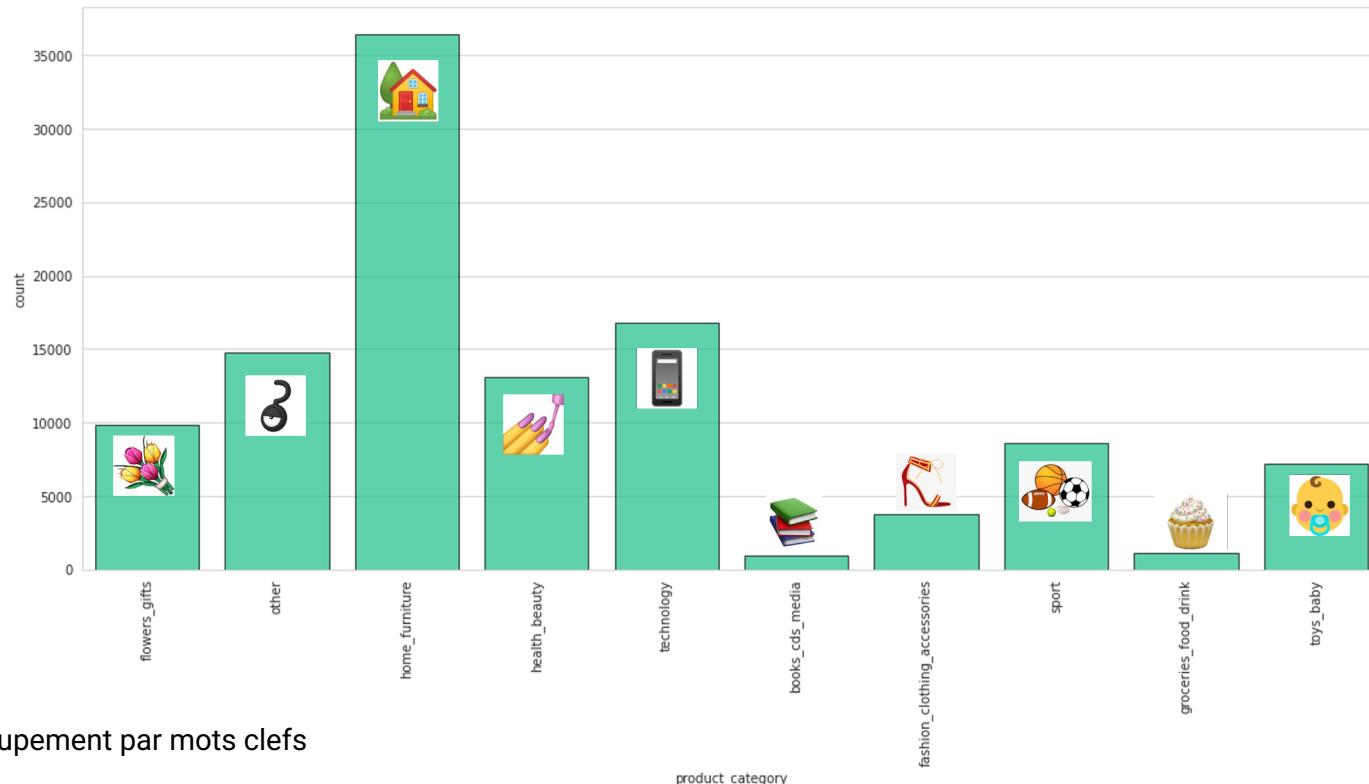
Analyse des données - Les commandes journalières

Evolution du nombre de commandes journalières



Analyse des données - Les catégories de produits

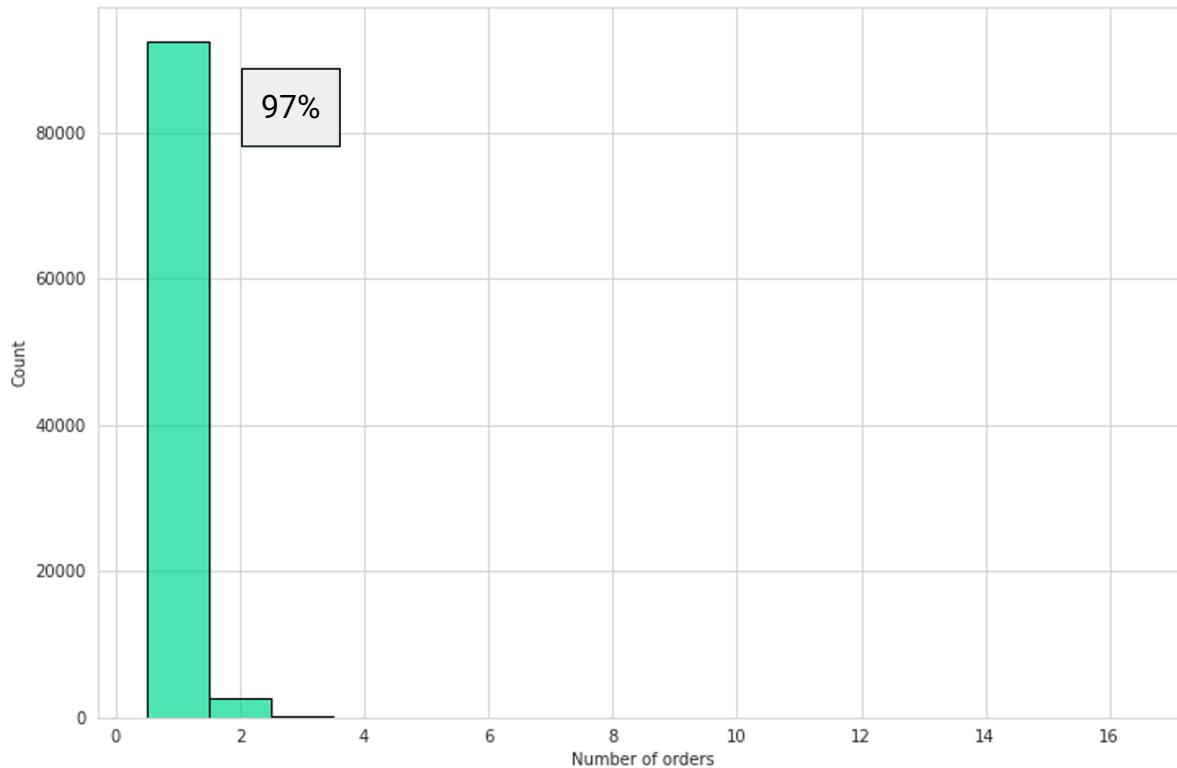
Les catégories de produits les plus représentées



regroupement par mots clefs

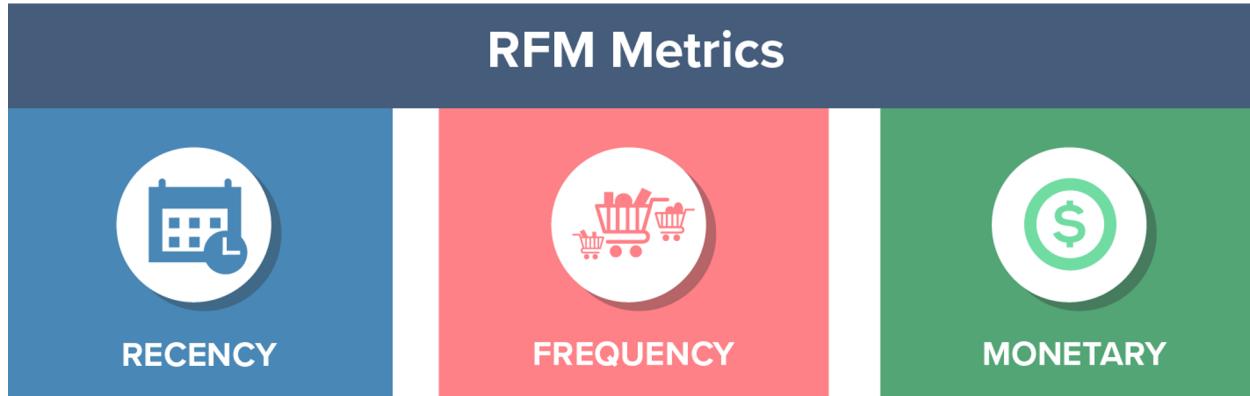
Analyse des données - Les commandes par client

Nombre de commandes par client



Et maintenant ?
- Modélisation
- Recommendations

Framework d'analyse de base (RFM)

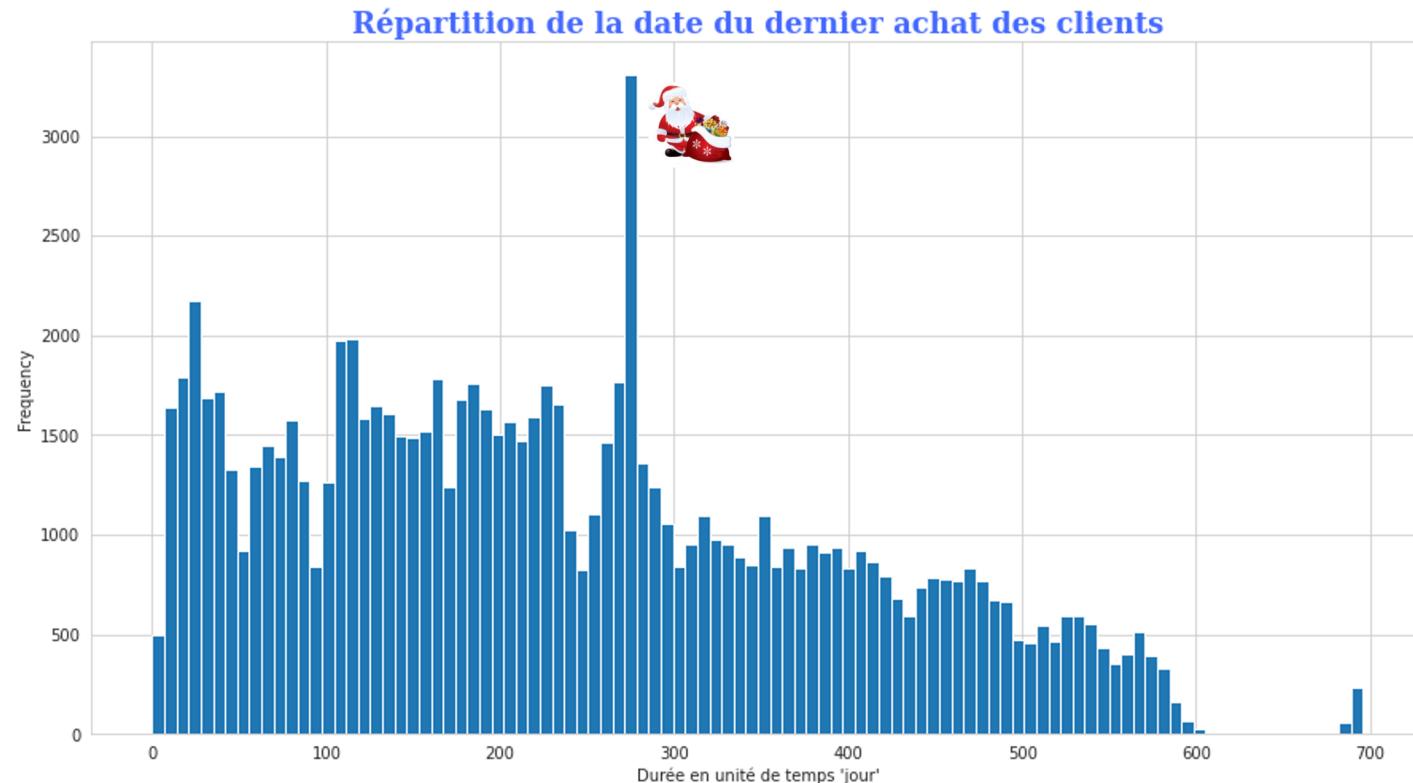


- Récence : Représente la date ou le client a effectué son dernier achat
- Fréquence : Représente le nombre d'achats effectués par le client
- Valeur monétaire : Représente la valeur moyenne des achats d'un client donné

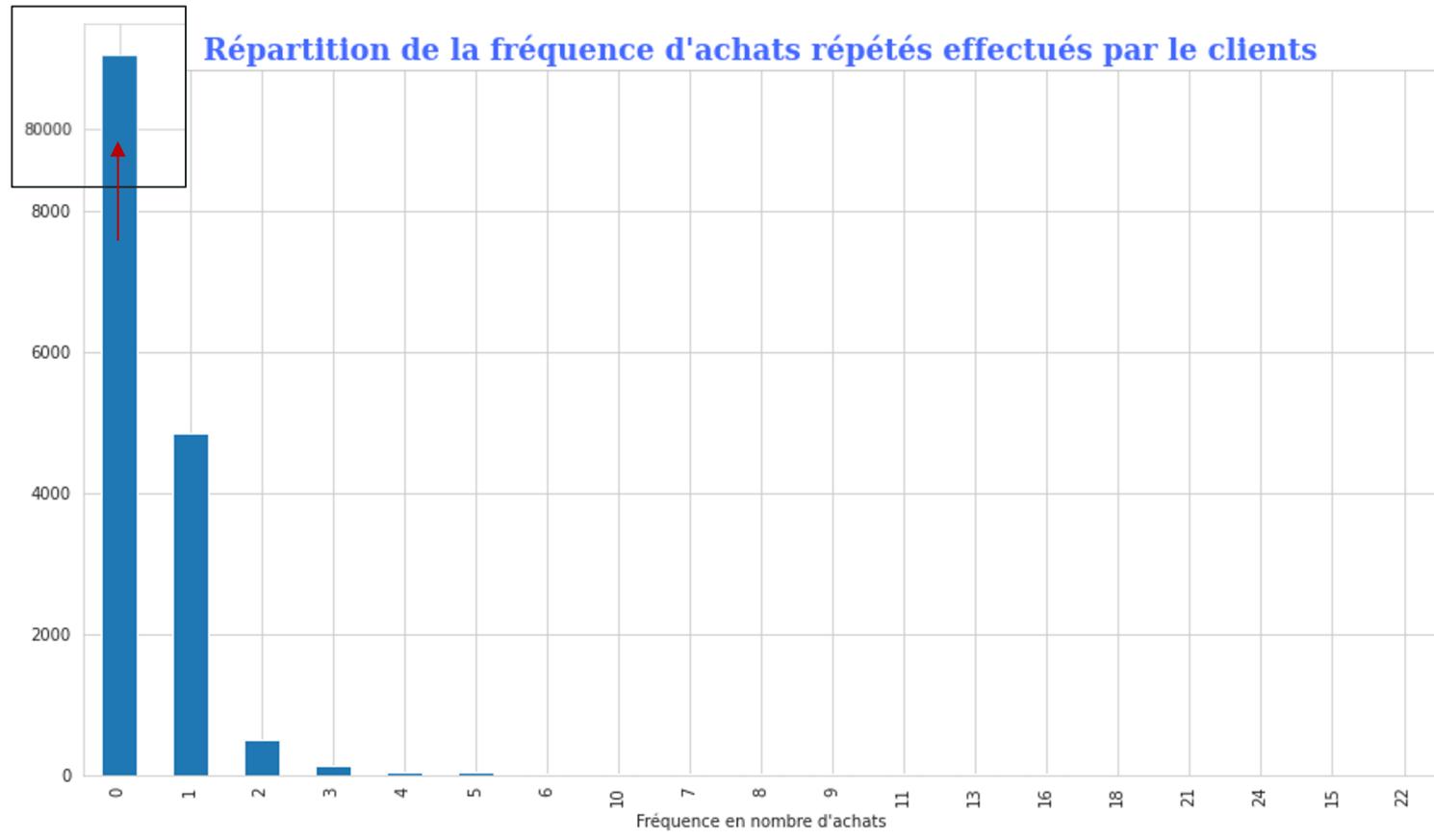
Autres axes d'analyse :

- Lifetime monetary value (CLV)
- Préférence de produits (categories)
- Satisfaction du client (review score)
- ...

RFM - Récence des clients

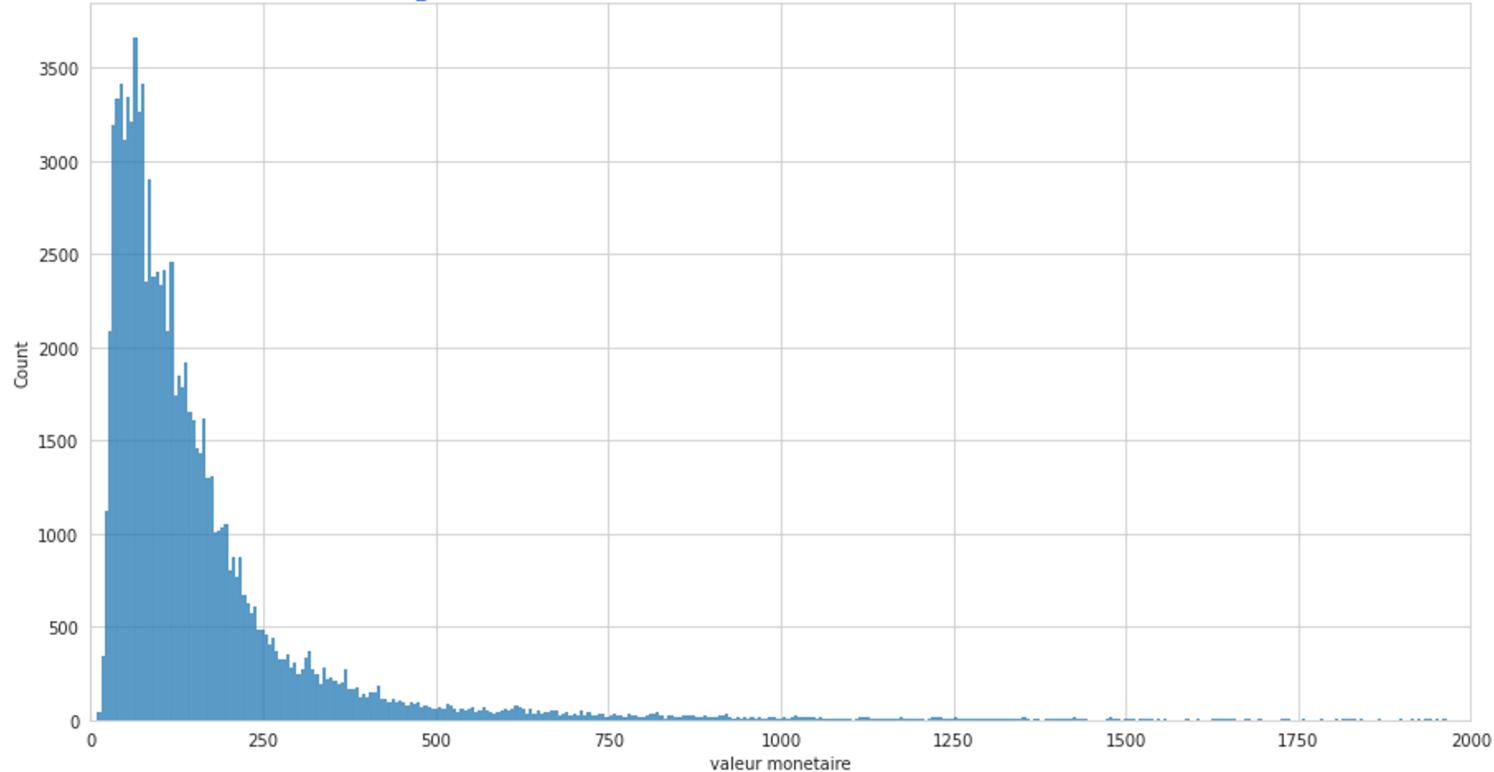


RFM - Fréquence des achats



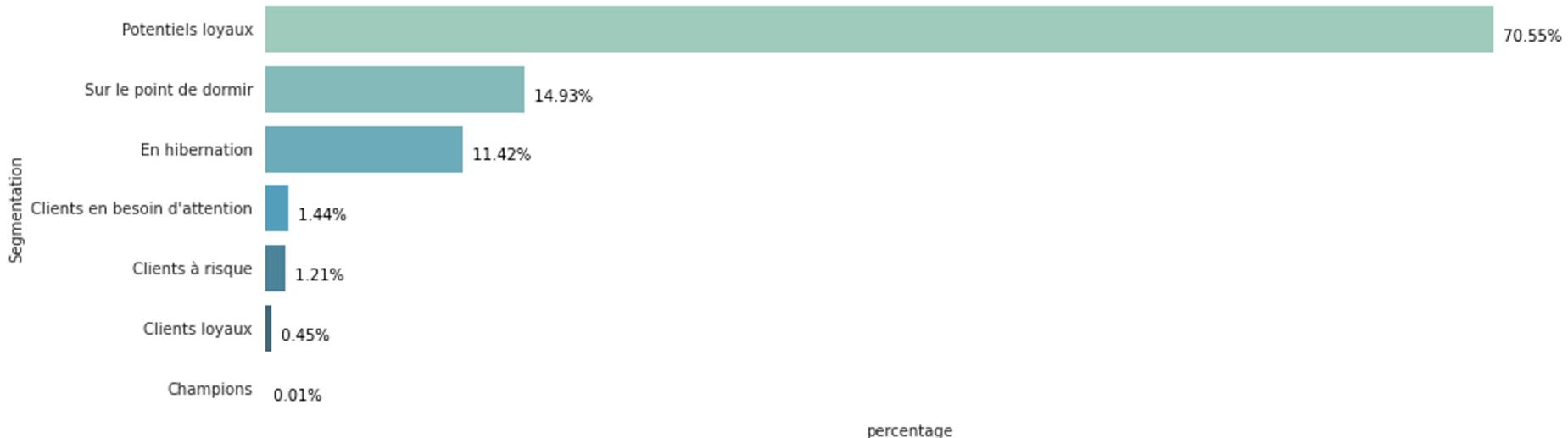
RFM - Valeur monétaire

Répartition de la valeur monétaire de nos clients



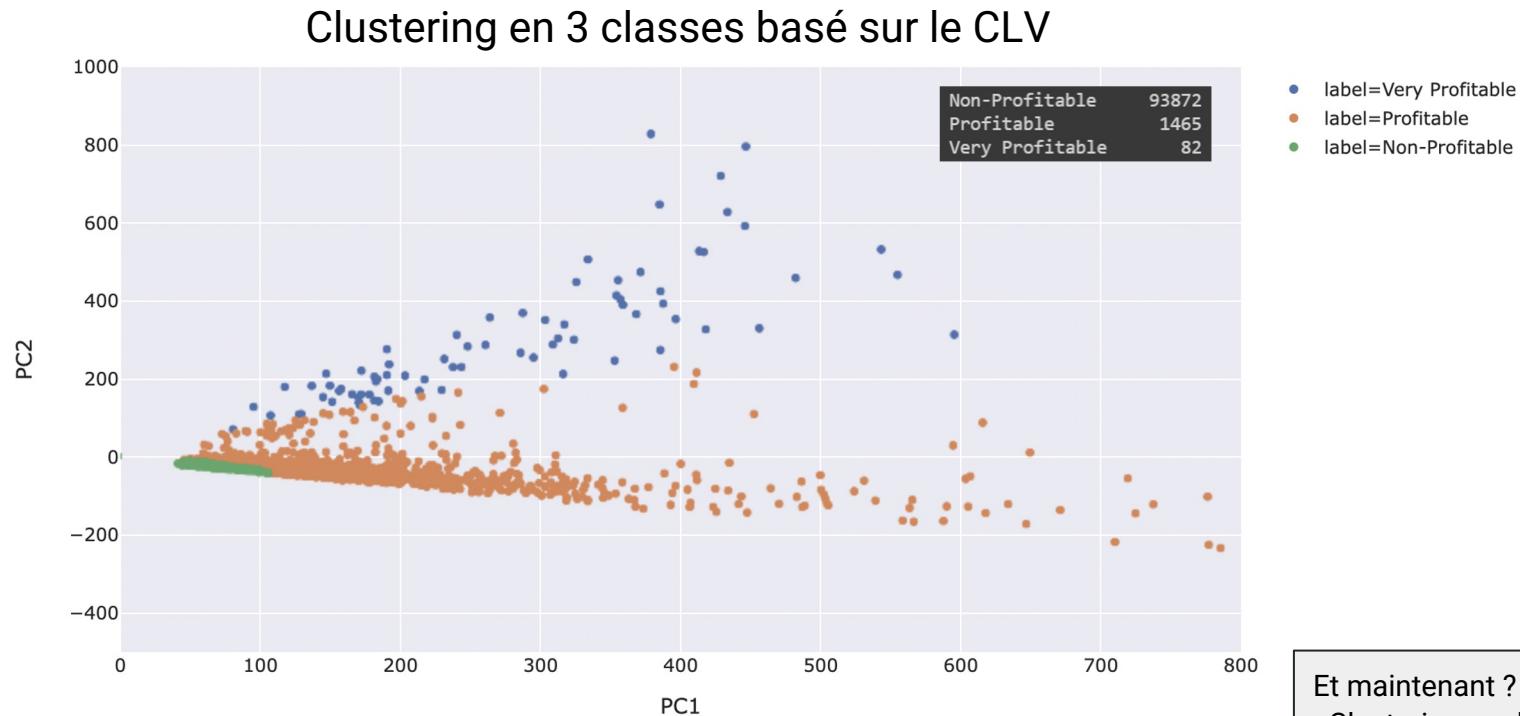
RFM - les groupes d'utilisateurs

Ségmentation des utilisateurs



→ Recommandations sur le site de CleverTap ! 😎😎

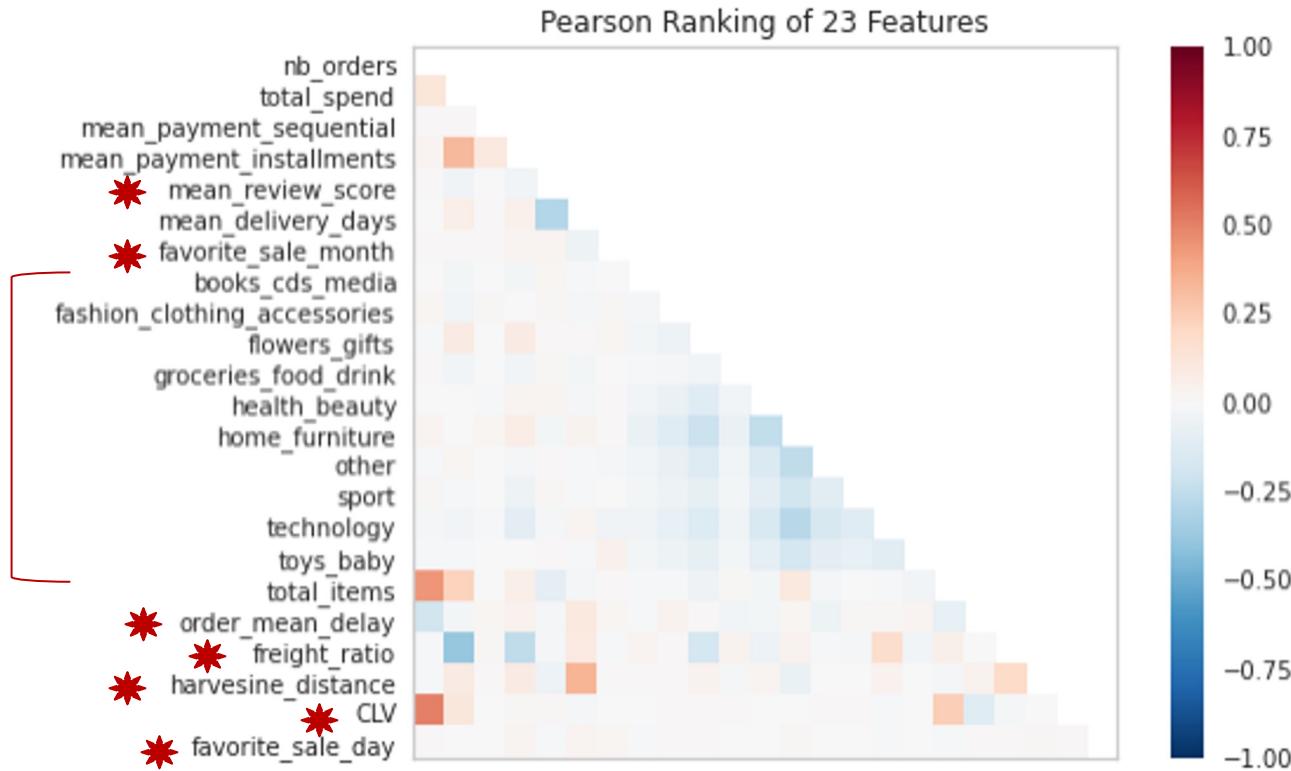
Analyse de profitabilité des clients à l'aide du CLV



→ PCA réalisée en 2D à des fins de visualisations 🌟

Et maintenant ?
- Clustering sur l'ensemble
des variables

Données - Heatmap des corrélations linéaires



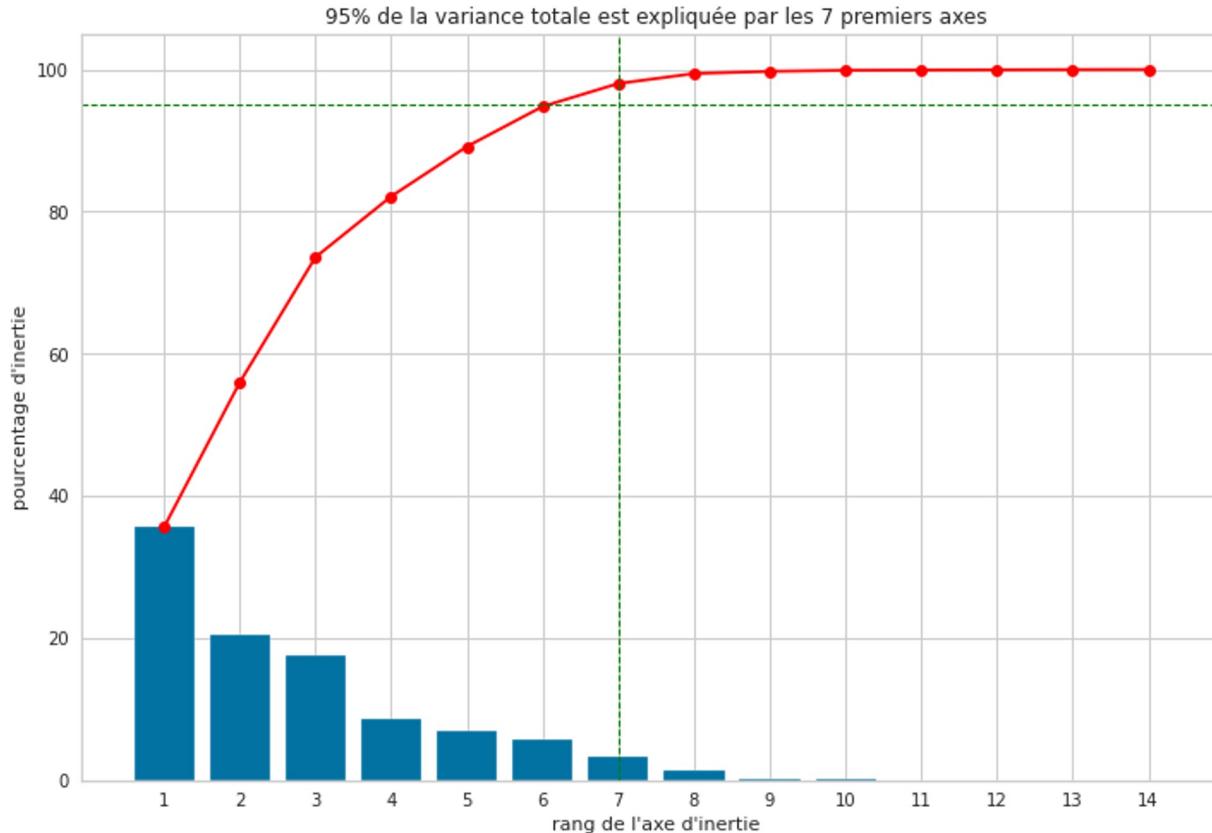
Les algorithmes de clusterings

	Non Hierarchical	Hierarchical
Centroid Based	K Means K Prototypes GMM	Agglomerative Clustering Complete Linkage Ward
Density Based	DBSCAN Mean Shift	HDBSCAN

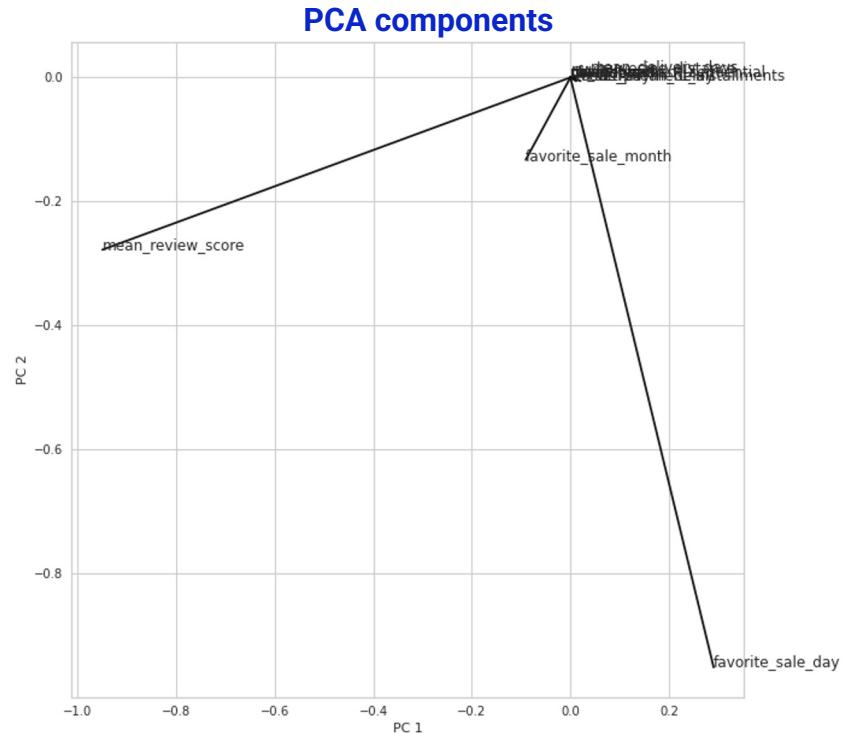
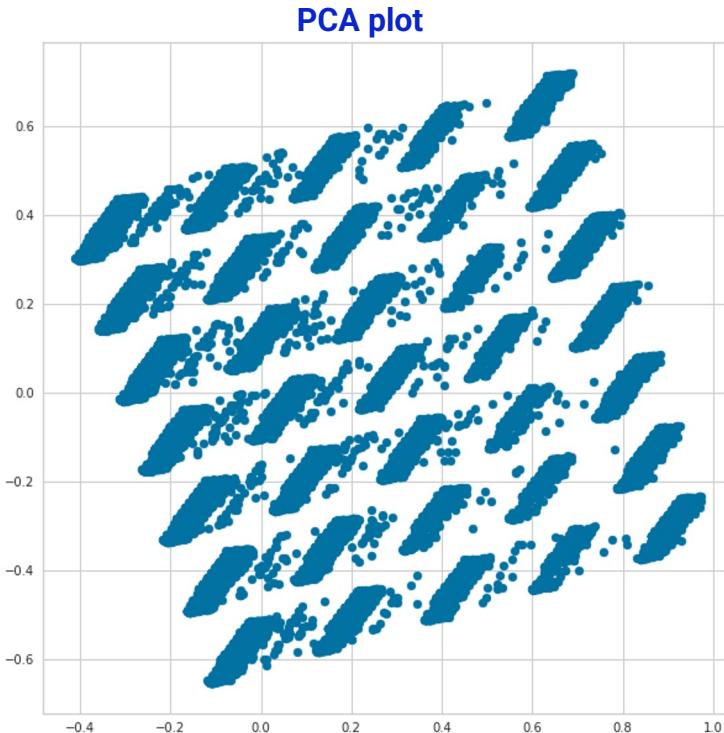
Et maintenant ?
- Etude du PCA
- Etude des modèles de clusterings

PCA - Variance des axes de réductions

olist



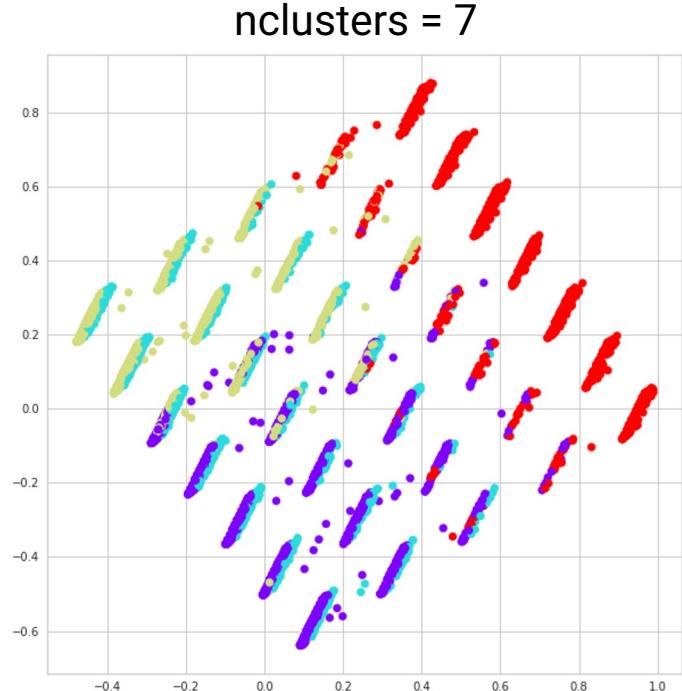
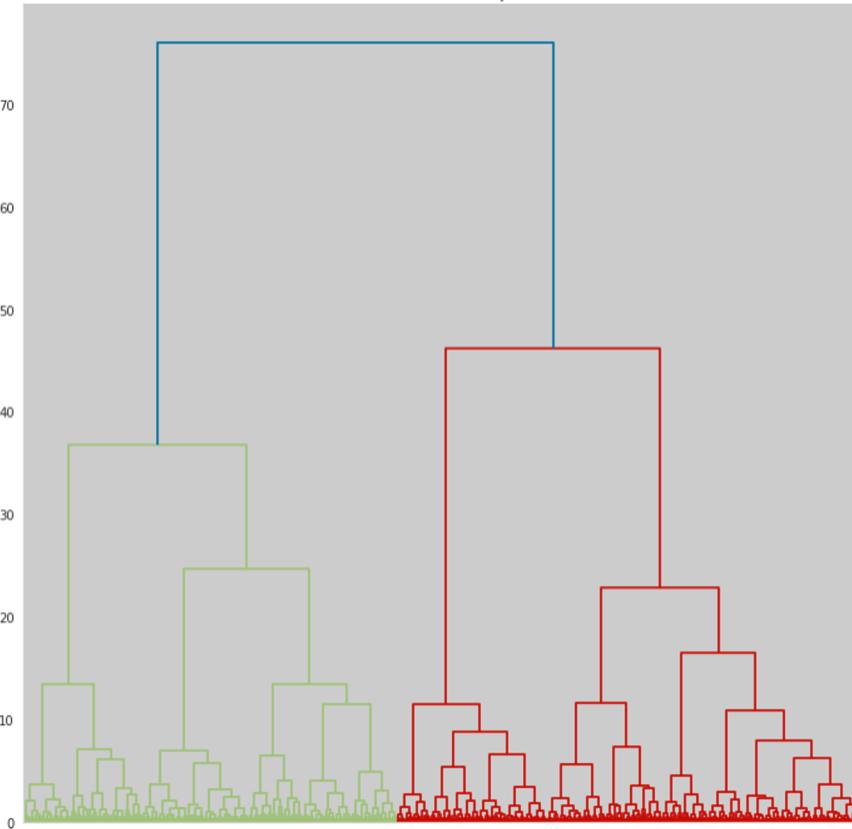
Visualisation de la donnée PCA



Algorithmes de Clustering Hiérarchique (Centroid Based)

olist

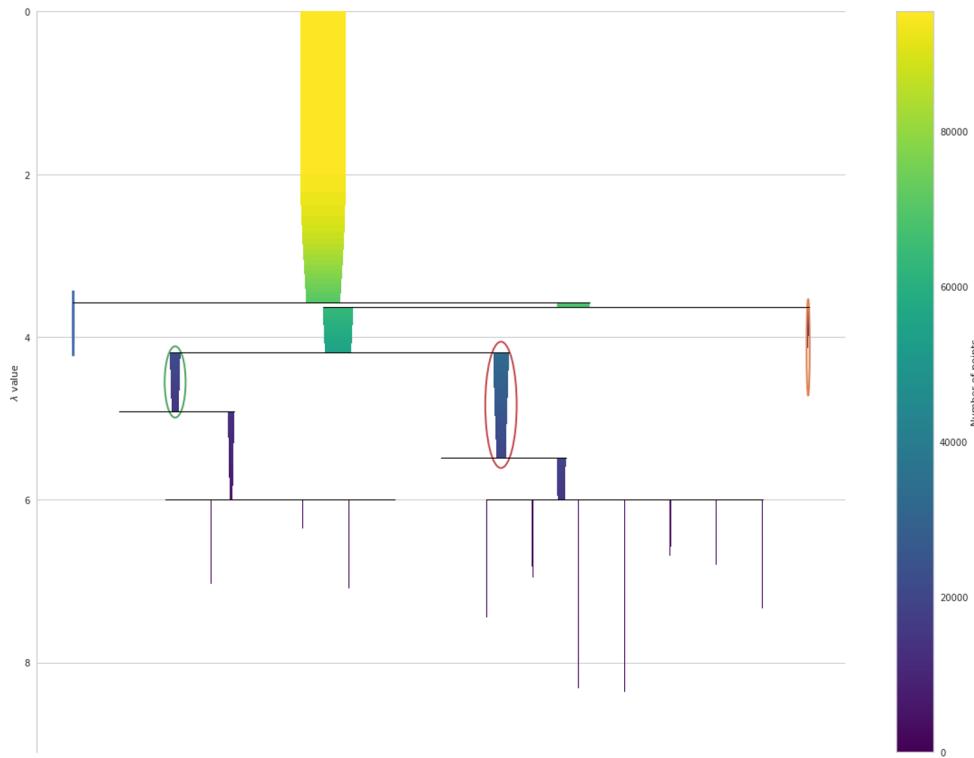
Hierarchy des clusters (Agglomerative Clustering)



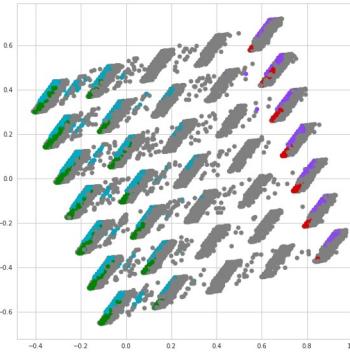
Algorithmes de Clustering Hiérarchique (Density Based)

olist

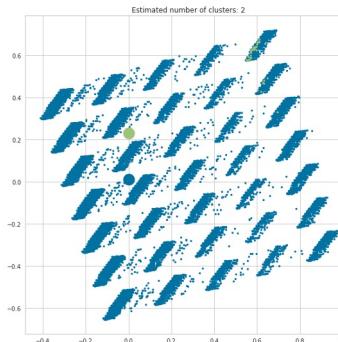
Hierarchy des clusters avec HDBSCAN



HDBSCAN Clusters

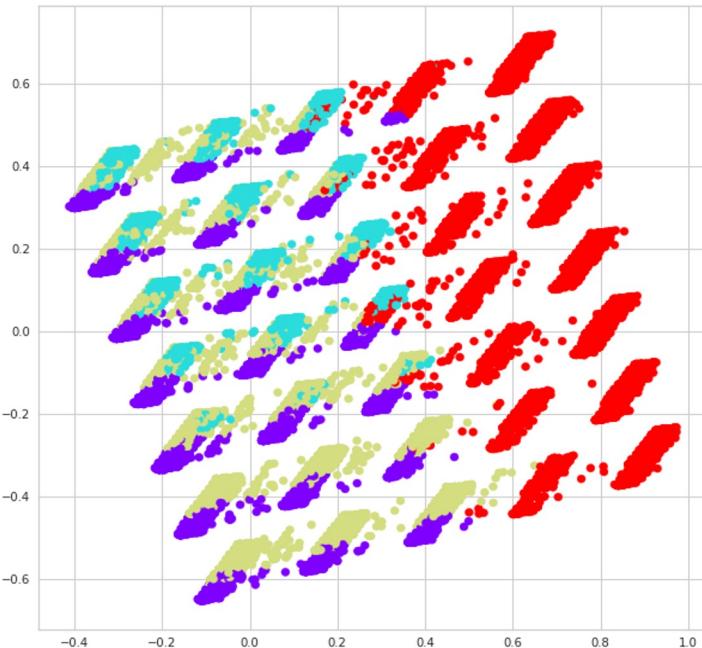


Mean Shift Clusters

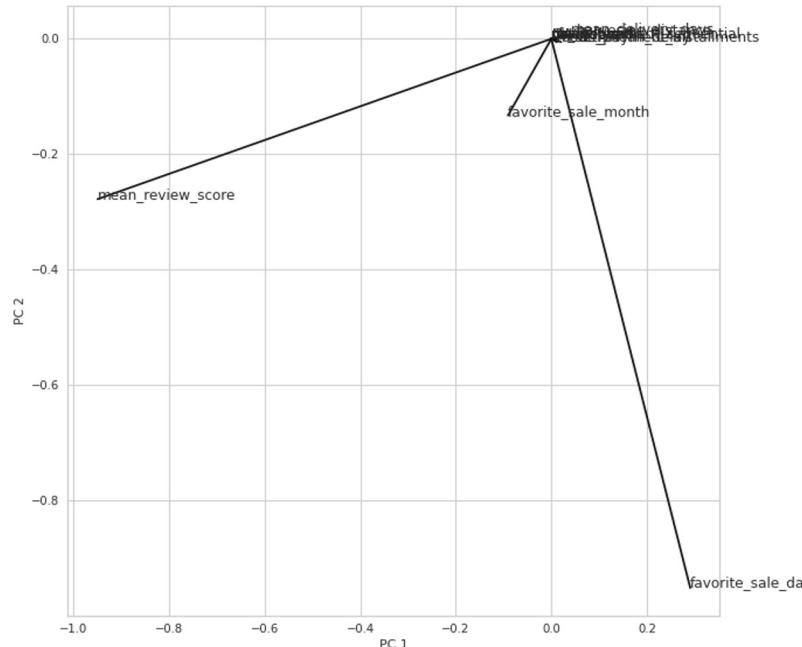


Visualisation de la donnée PCA, Kmeans

n clusters = 4

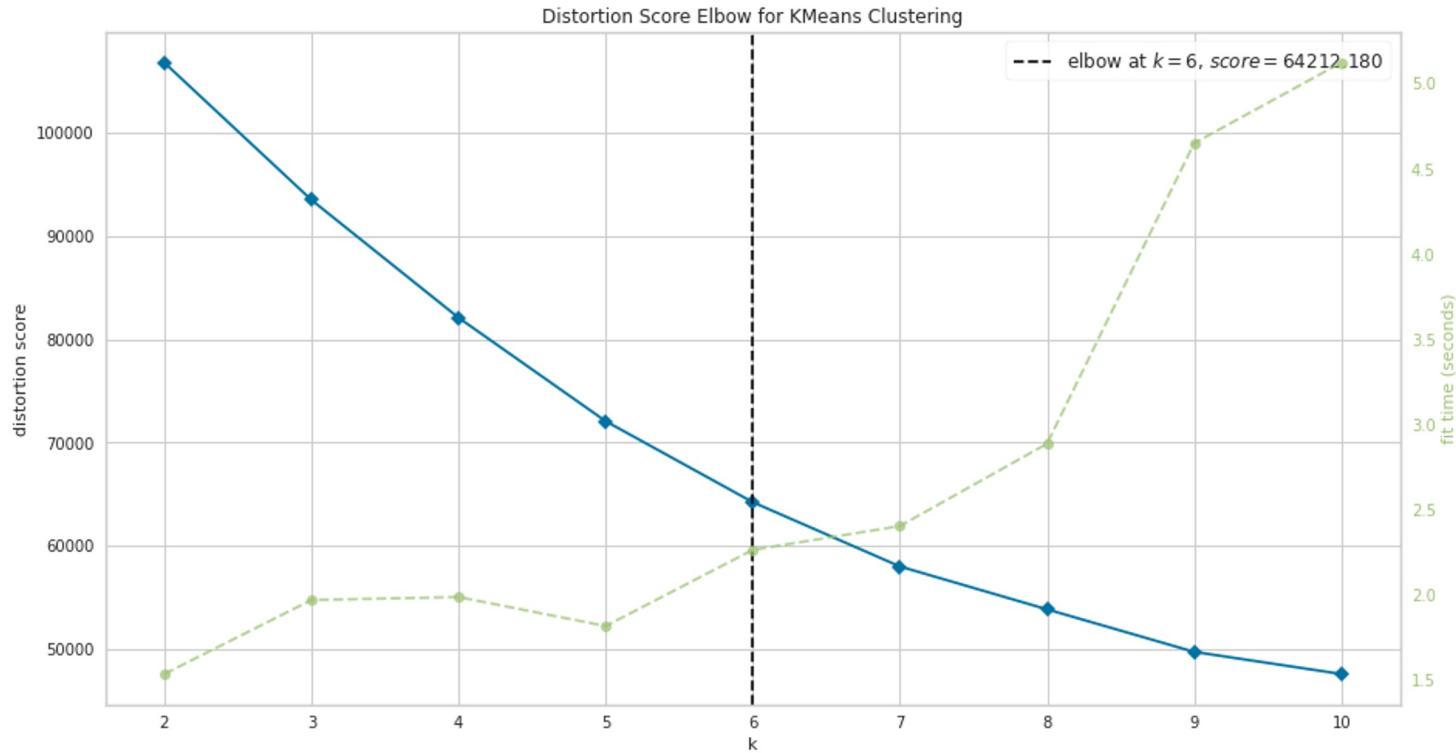


PCA components



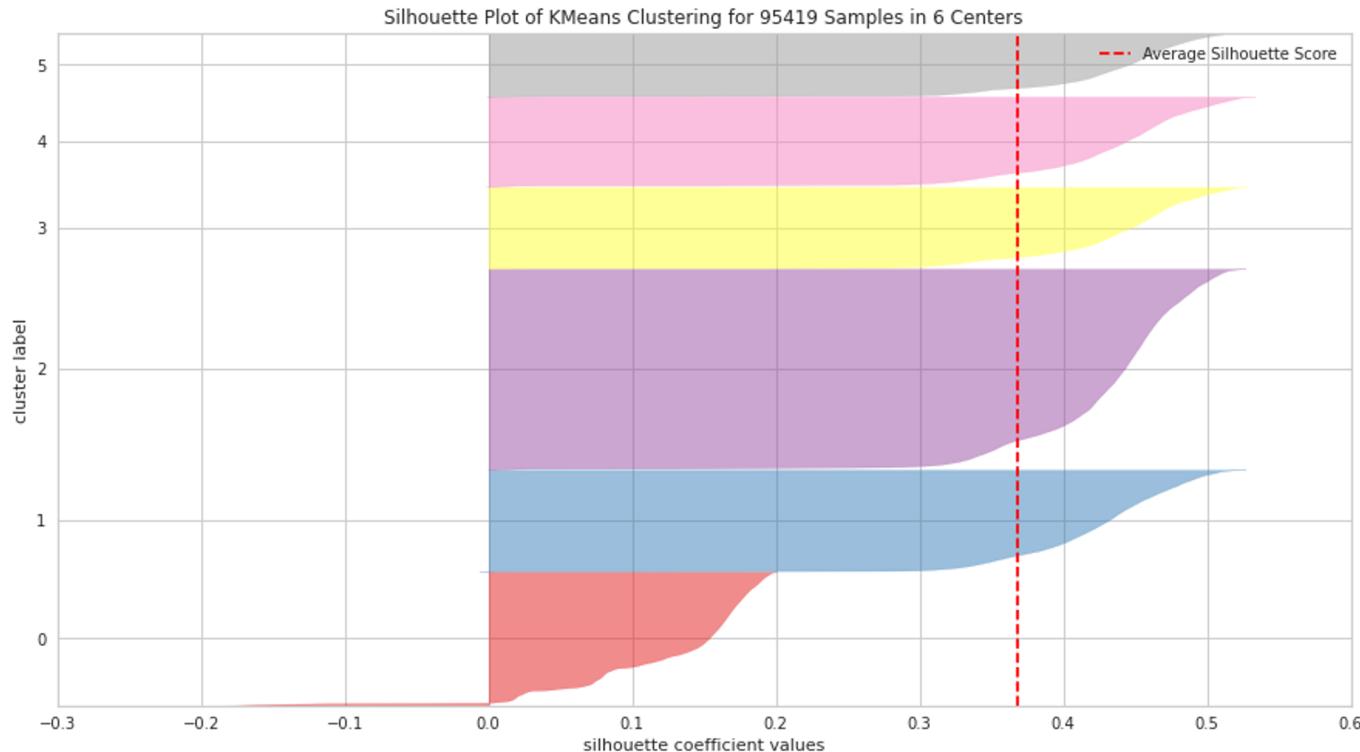
Et maintenant ?
- Etude Kmeans avec Yellowbrick
- Etude des clusters

KMeans Clustering - Elbow Method



Kmeans avec toute les données

KMeans Clustering - Silhouette Plot K = 6



KMeans Clustering - Distance inter-cluster



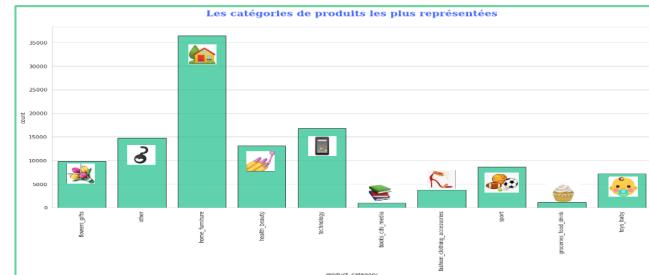
KMeans Clustering - Clusters

Comparaison des moyennes par variable des clusters

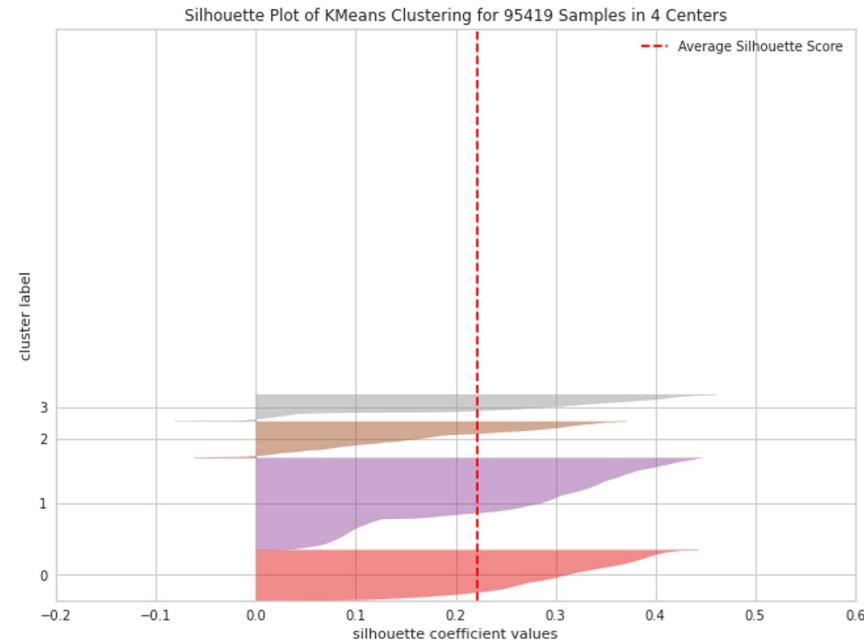
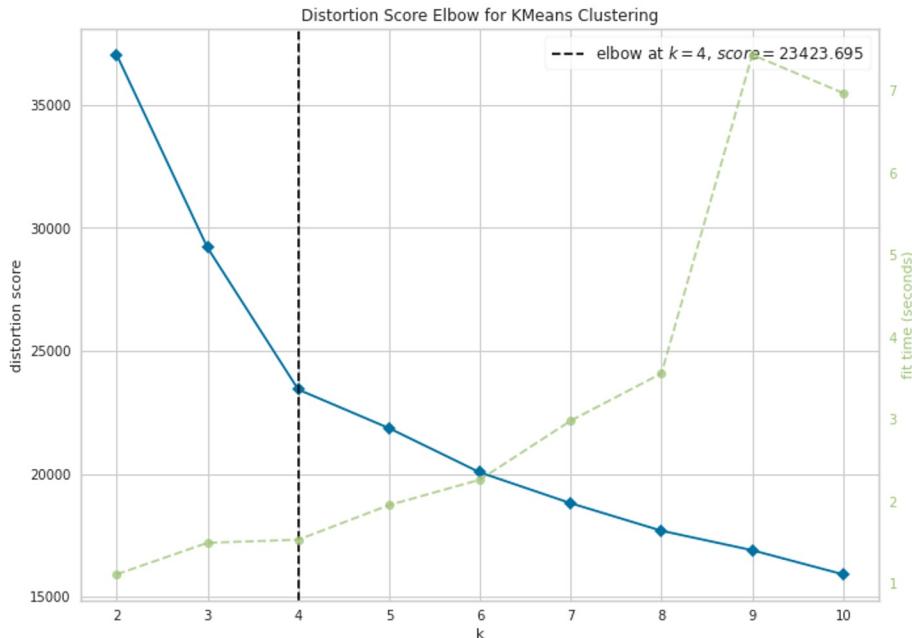
- Cluster 0
- Cluster 1
- Cluster 2
- Cluster 3
- Cluster 4
- Cluster 5



On retire les catégories de produits

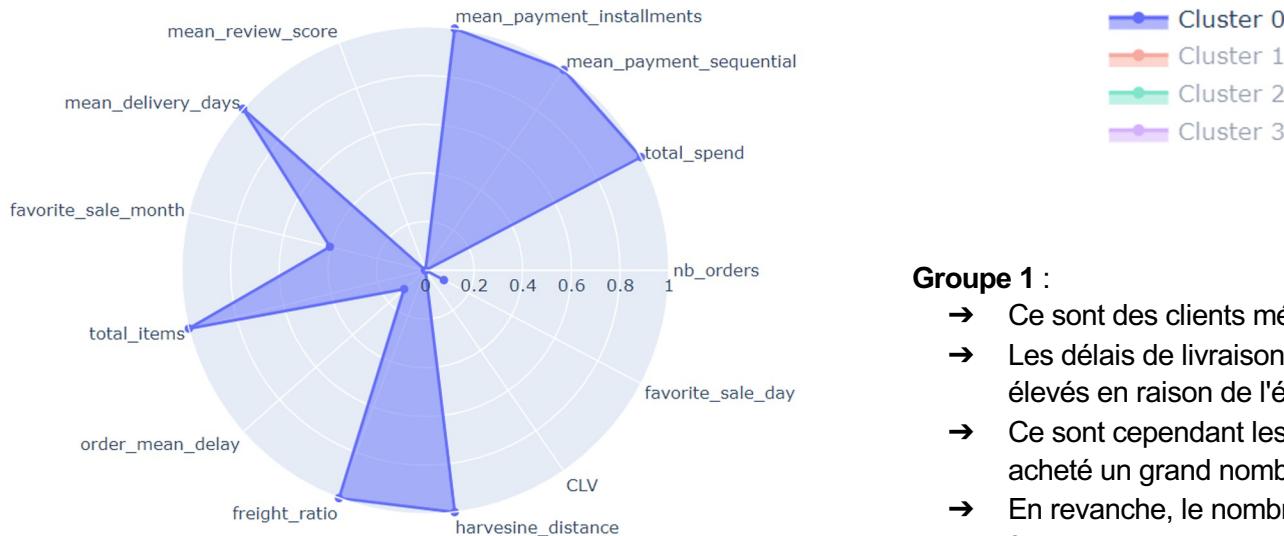


KMeans Clustering - Elbow Method 2



KMeans Clustering - Explications des clusters (3/3)

Comparaison des moyennes par variable des clusters



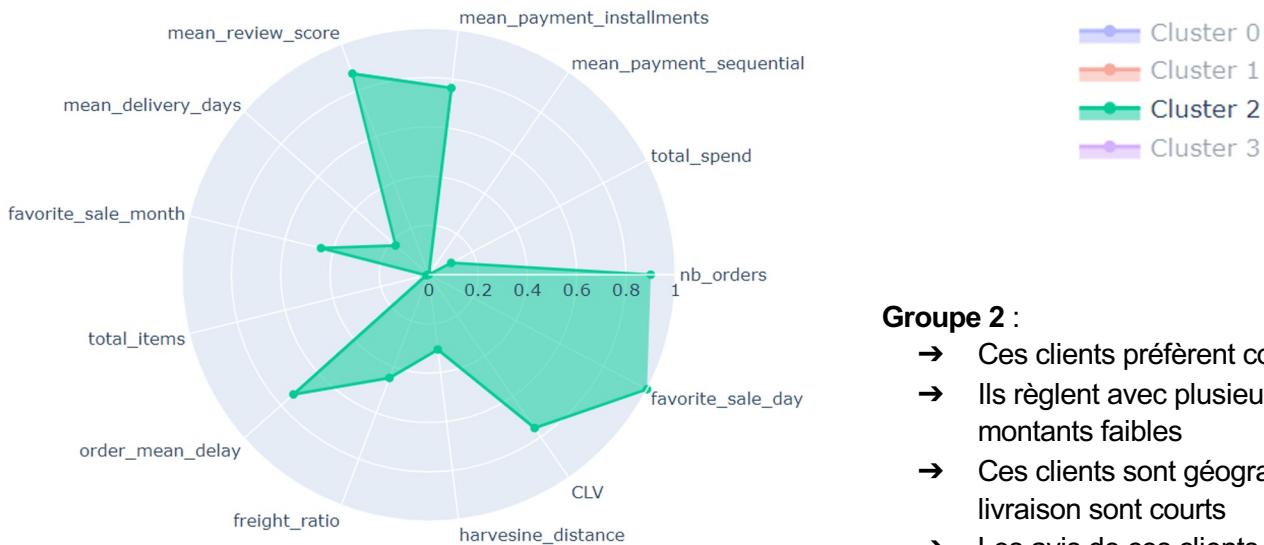
Groupe 1 :

- Ce sont des clients mécontents (les avis sont mauvais)
- Les délais de livraison sont très importants et les frais de port élevés en raison de l'éloignement géographique
- Ce sont cependant les clients qui ont le plus dépensé et ont acheté un grand nombre d'articles
- En revanche, le nombre de commandes passées sur le site est faible
- Ils ne sont pas rentable

Ma recommandation : Créez une option "livraison rapide" payante ! ⚡⚡⚡

KMeans Clustering - Explications des clusters (3/3)

Comparaison des moyennes par variable des clusters



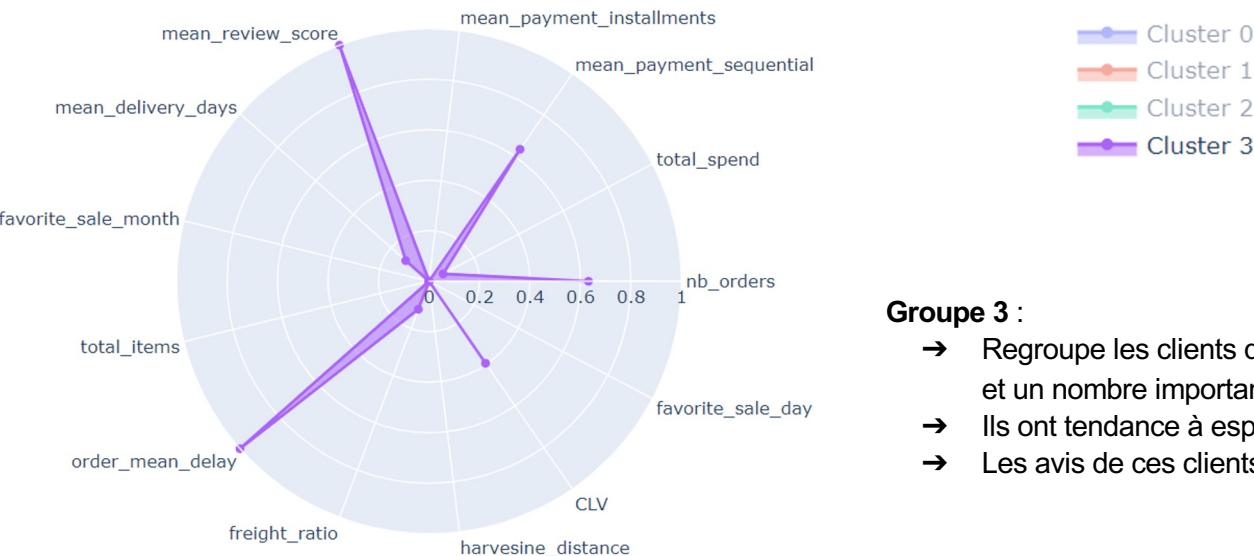
Groupe 2 :

- Ces clients préfèrent commander le weekend
- Ils règlent avec plusieurs moyens de paiement pour des montants faibles
- Ces clients sont géographiquement proches et les délais de livraison sont courts
- Les avis de ces clients sont globalement bons

Ma recommandation : Proposer des offres promotionnelles spécial weekend ! 😎😎

KMeans Clustering - Explications des clusters (3/3)

Comparaison des moyennes par variable des clusters



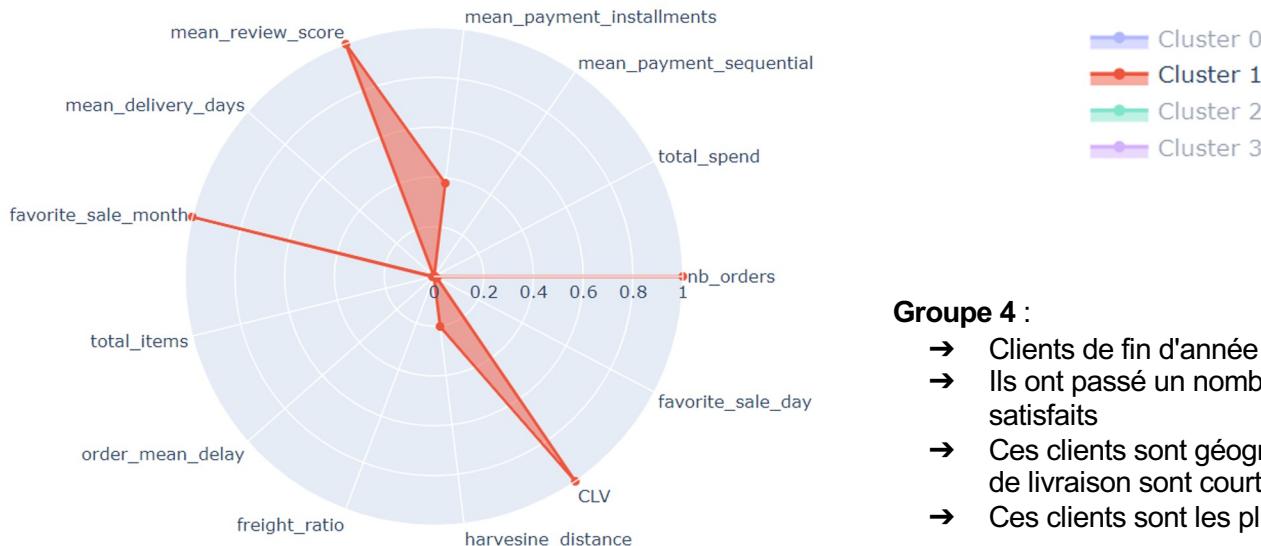
Groupe 3 :

- Regroupe les clients qui utilisent plusieurs moyens de paiement et un nombre important d'échéances
- Ils ont tendance à espacer les délais entre deux commandes
- Les avis de ces clients sont également très bons

Ma recommandation : Envoyer des newsletters fréquemment afin de mieux engager ces clients ! 🎉

KMeans Clustering - Explications des clusters (3/3)

Comparaison des moyennes par variable des clusters



Groupe 4 :

- Clients de fin d'année
- Ils ont passé un nombre important de commandes et sont satisfaits
- Ces clients sont géographiquement peu éloignés et les délais de livraison sont courts
- Ces clients sont les plus rentable

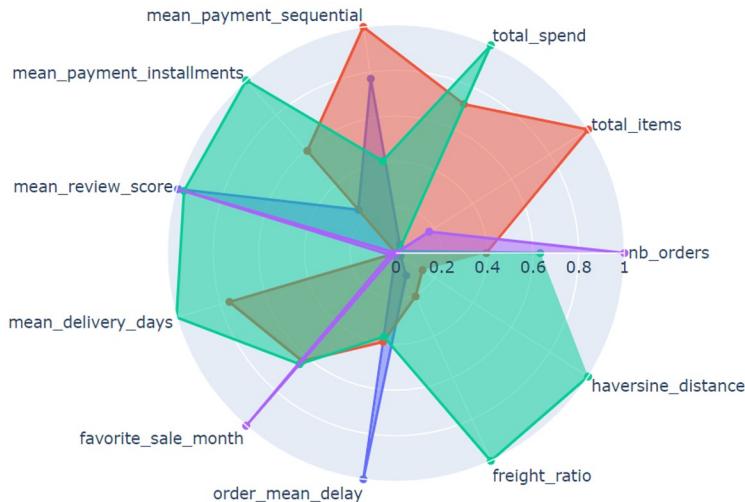
Ma recommandation : Créer une offre "Premium" afin d'avoir des prix tout au long de l'année !

Et maintenant ?
- Etude de stabilité du clustering

Stabilité temporelle de la segmentation

1er période

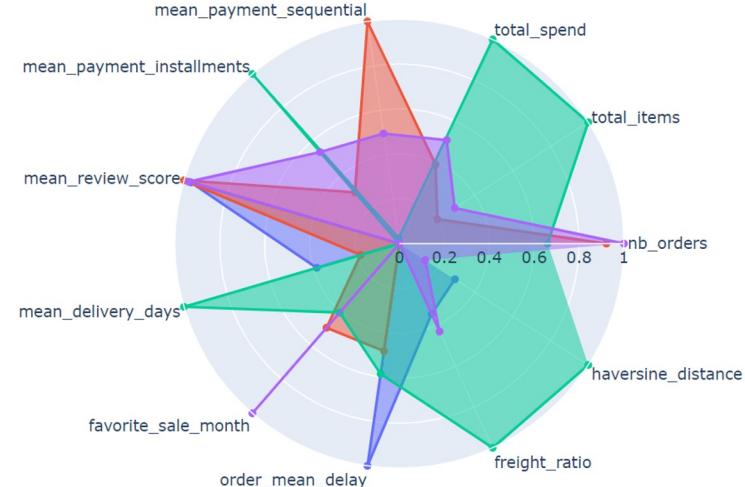
Comparaison des moyennes par variable des clusters



adjusted_rand_score
= 0.915

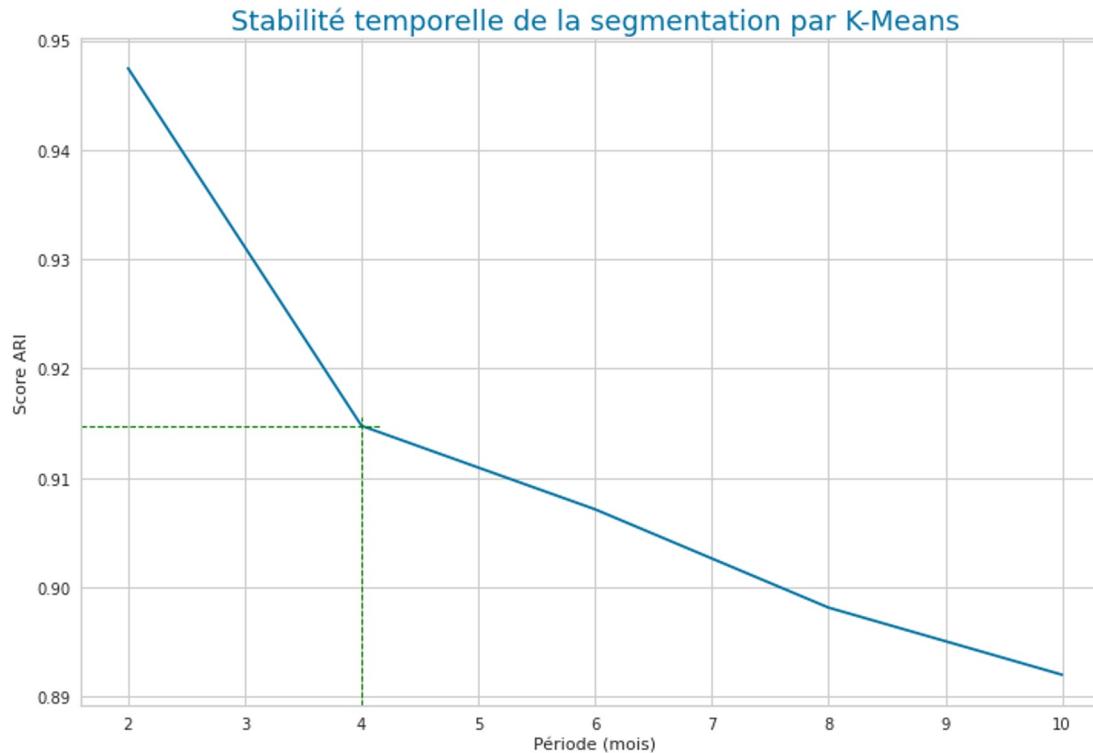
2nd période

Comparaison des moyennes par variable des clusters



adjusted_rand_score
= 0.892

Stabilité temporelle de la segmentation (ARI)



Itérations par période de 2 mois

Et maintenant ?
- Conclusions

Conclusions

1 - Quels sont les différents type d'utilisateurs du site Olist :

-> La segmentation Kmeans nous permet de catégoriser les utilisateurs en 4 groupe, suivant des attributs bien définis

2 - Comment cette segmentation peut elle nous permettre de définir des actions marketing actionnable :

-> Chaque groupe peut faire l'objet d'une campagne commerciale ciblée

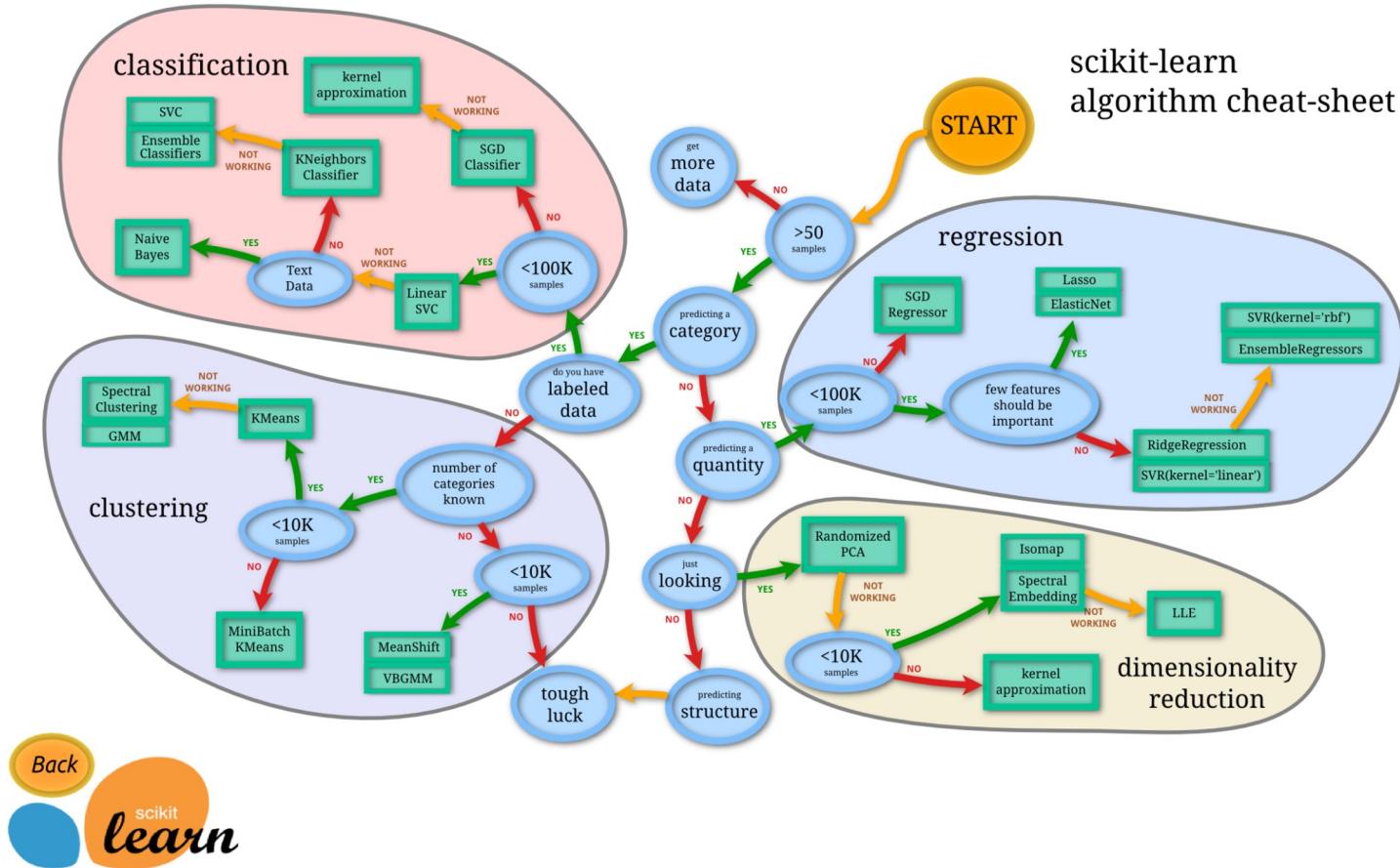
3 - A quelle fréquence dois-je mettre à jour ma segmentation :

-> La stabilité de la segmentation est de 4 mois, il nous faut donc mettre à jour la segmentation 3 fois par an

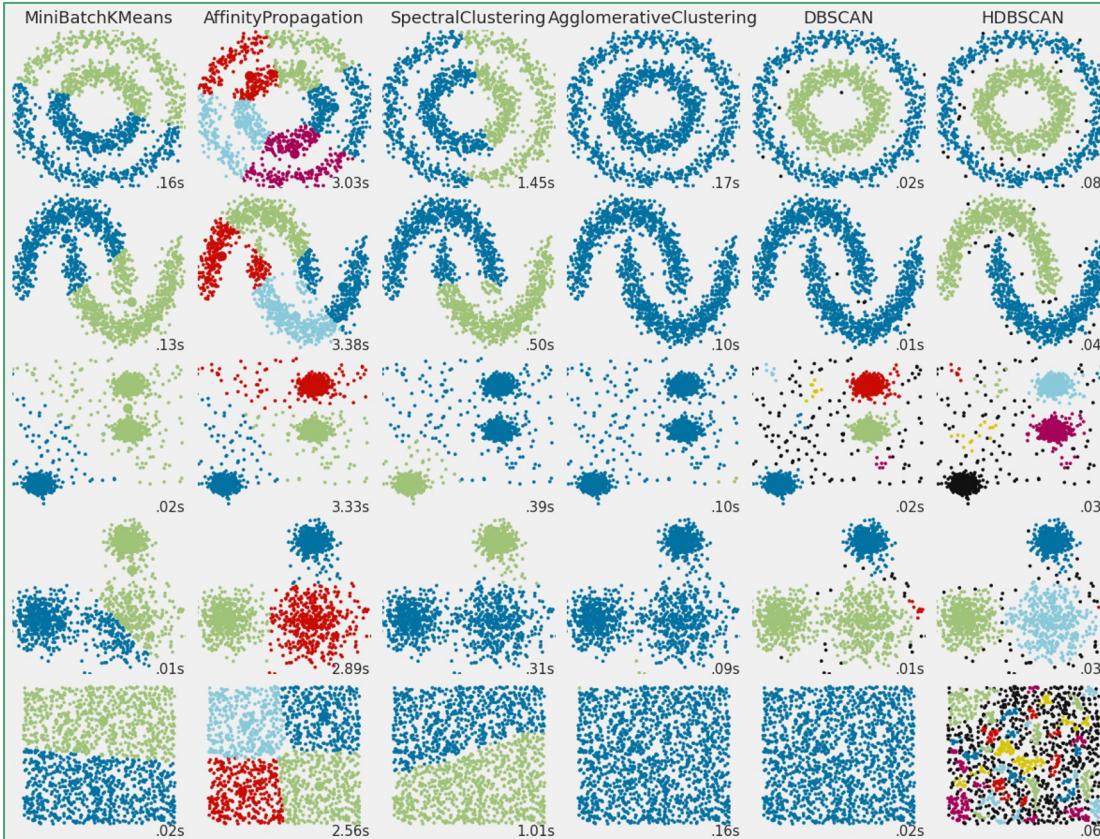
Bibliographie

- Boitmobile. "Segmentation RFM." Définitions Marketing " L'encyclopédie illustrée du marketing. Accessed December 8, 2021. <https://www.definitions-marketing.com/definition/segmentation-rfm/>.
- CleverTap. "RFM Analysis for Customer Segmentation." Accessed December 8, 2021. <https://clevertap.com/blog/rfm-analysis/>.
- kautumn06. "Yellowbrick - Clustering Evaluation Examples." Kaggle. Kaggle, August 17, 2018. <https://www.kaggle.com/kautumn06/yellowbrick-clustering-evaluation-examples>.
- Maklin, Cory. "DBSCAN Python Example: The Optimal Value For Epsilon (EPS)." Towards Data Science, July 14, 2019. <https://towardsdatascience.com/machine-learning-clustering-dbscan-determine-the-optimal-value-for-epsilon-eps-python-example-3100091cfbc>.
- Frenzel, Charles. "Tuning with HDBSCAN." Medium. Towards Data Science, September 11, 2021. <https://towardsdatascience.com/tuning-with-hdbscan-149865ac2970>.

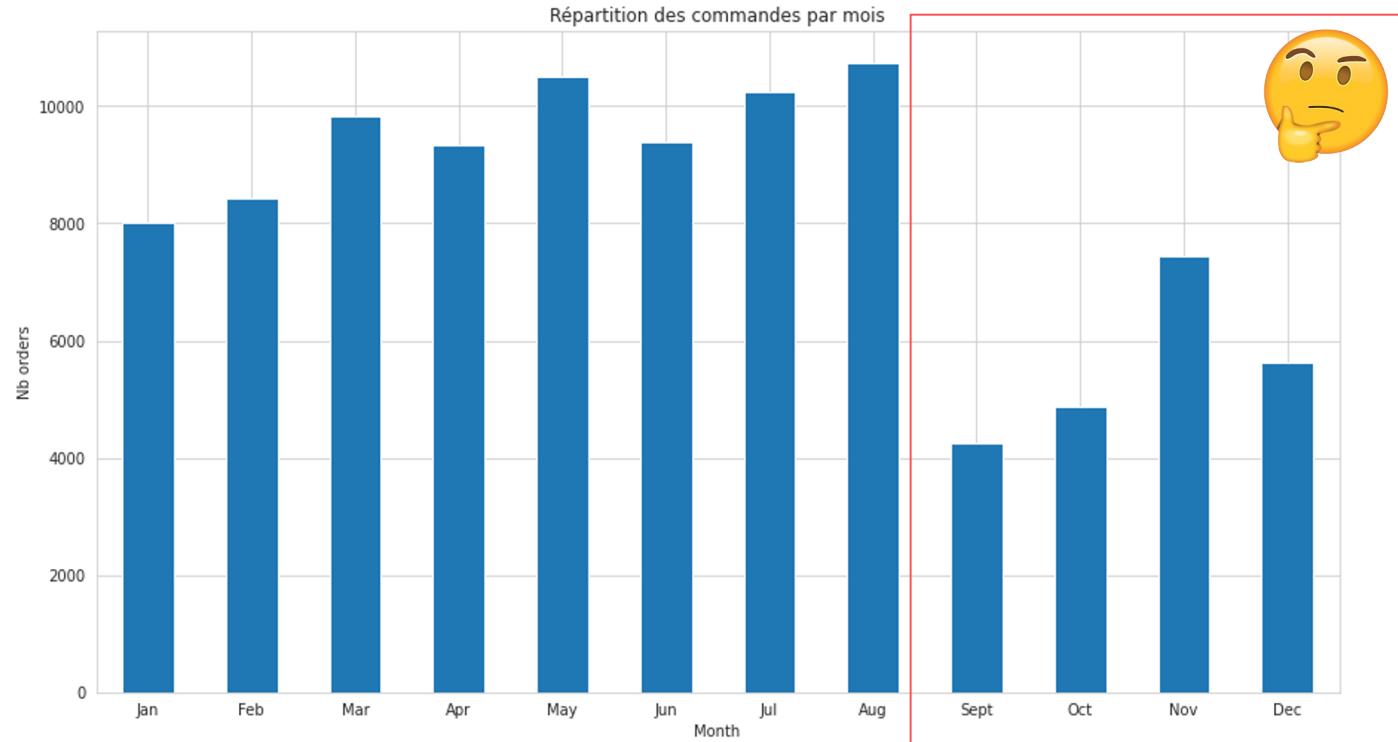
Annexes - sklearn cheat sheet



Test Clusterings

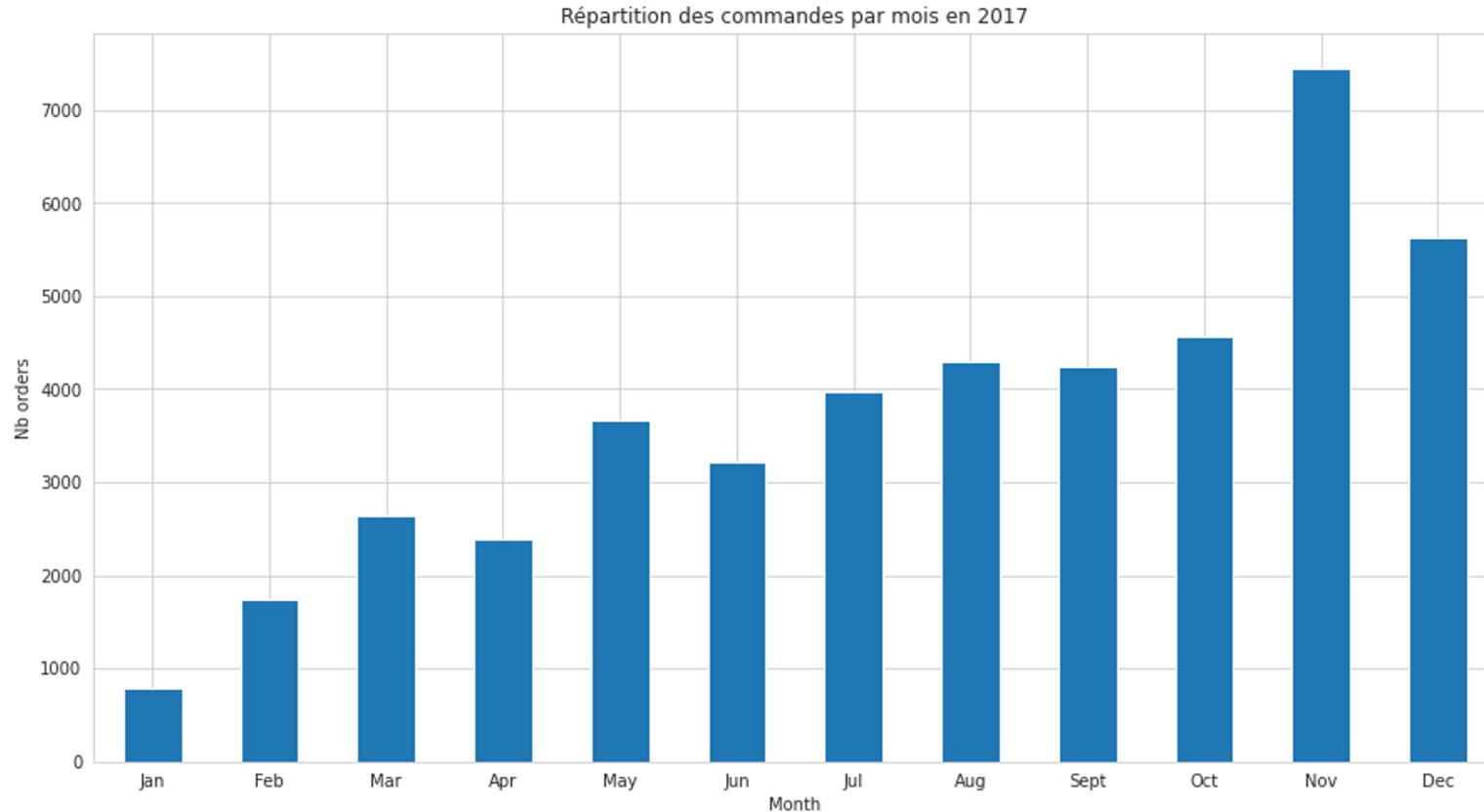


Analyse des données - Les commandes par mois (all)

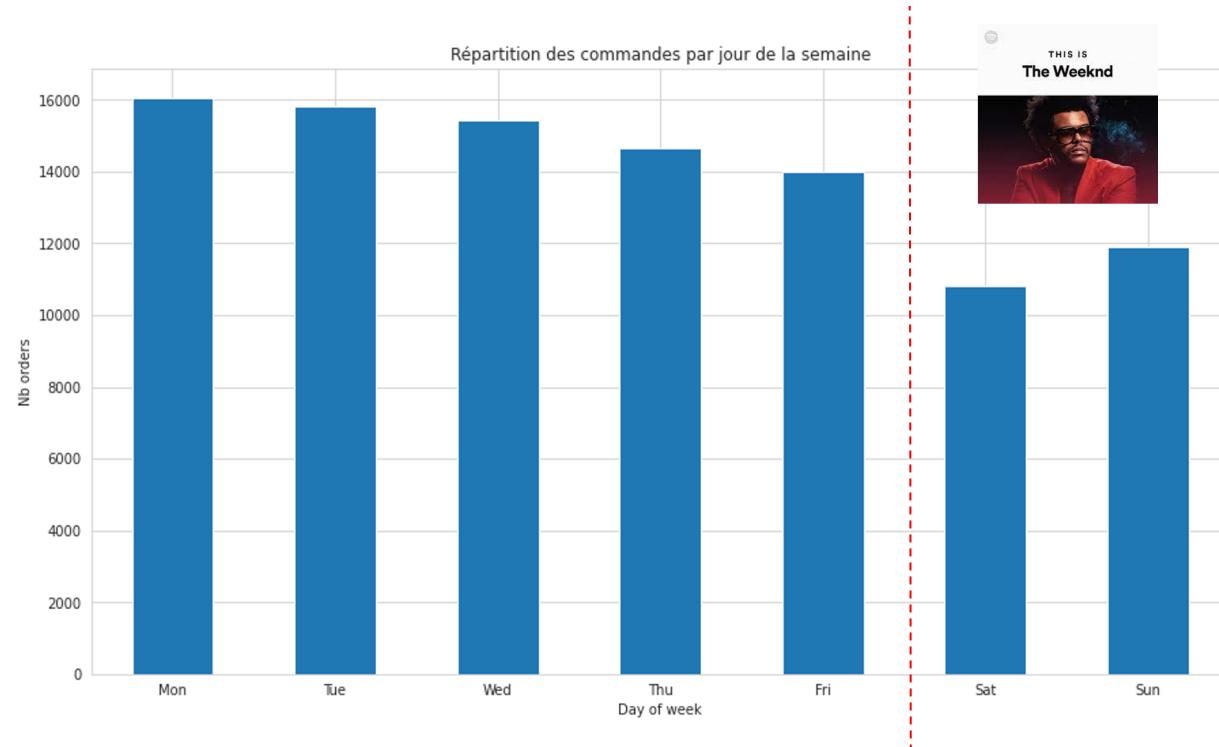


Analyse des données - Les commandes par mois (2017)

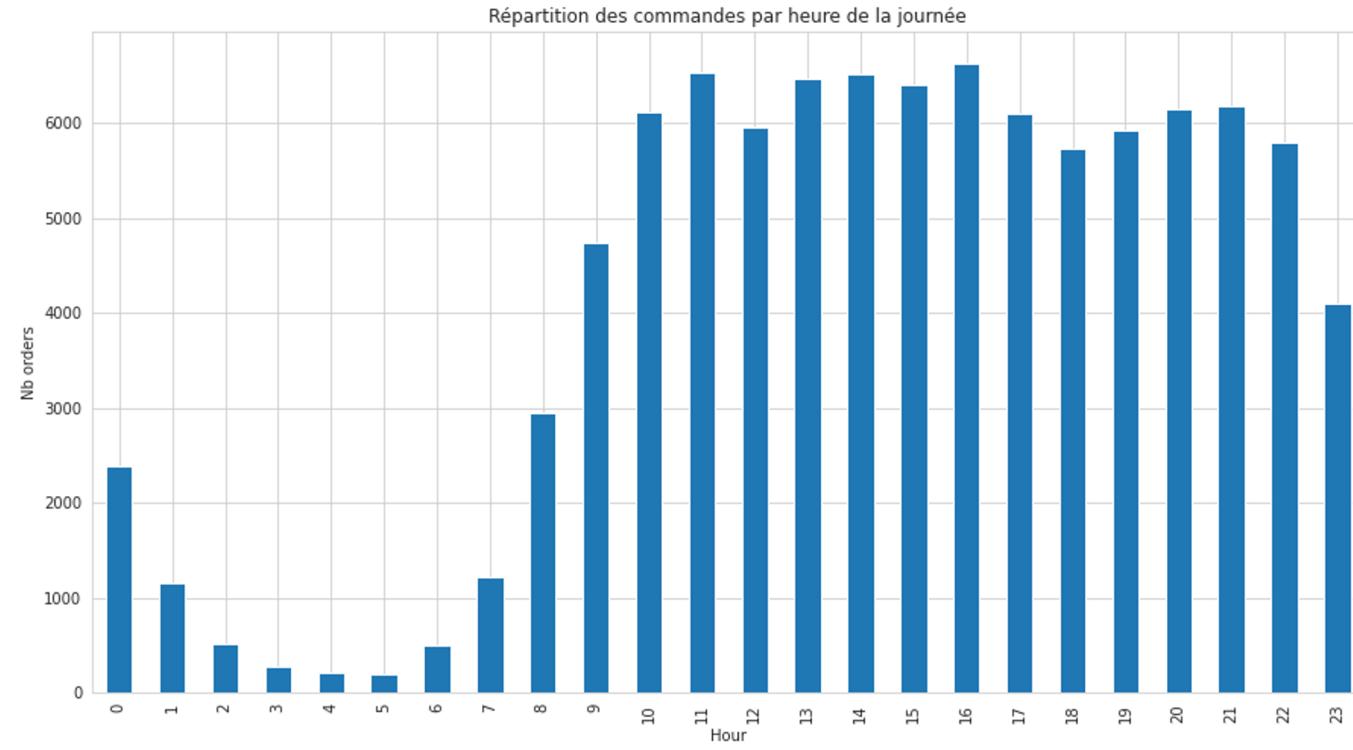
olist



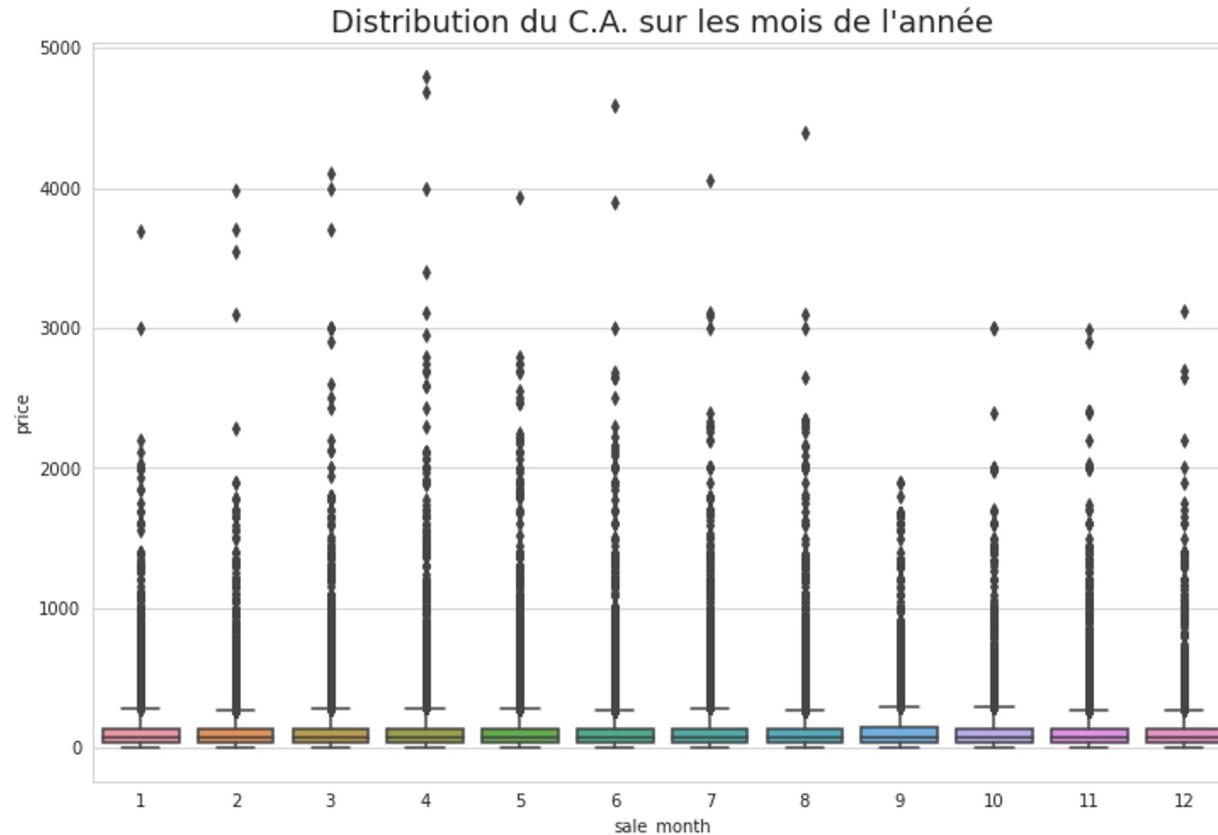
Analyse des données - Les commandes par jour de la semaine olist



Analyse des données - Les commandes par heure de la journée

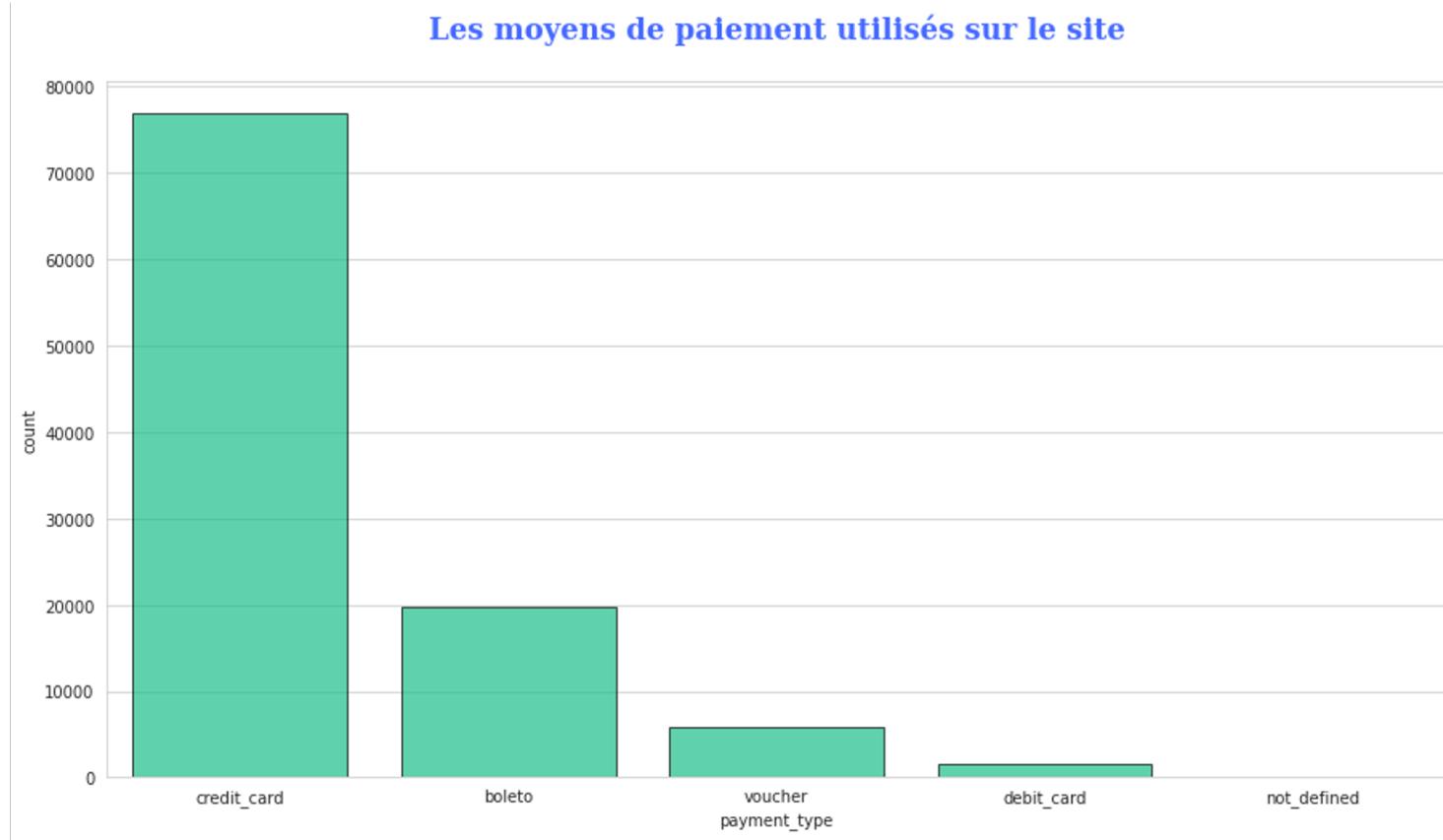


Analyse des données - Le chiffre d'affaire



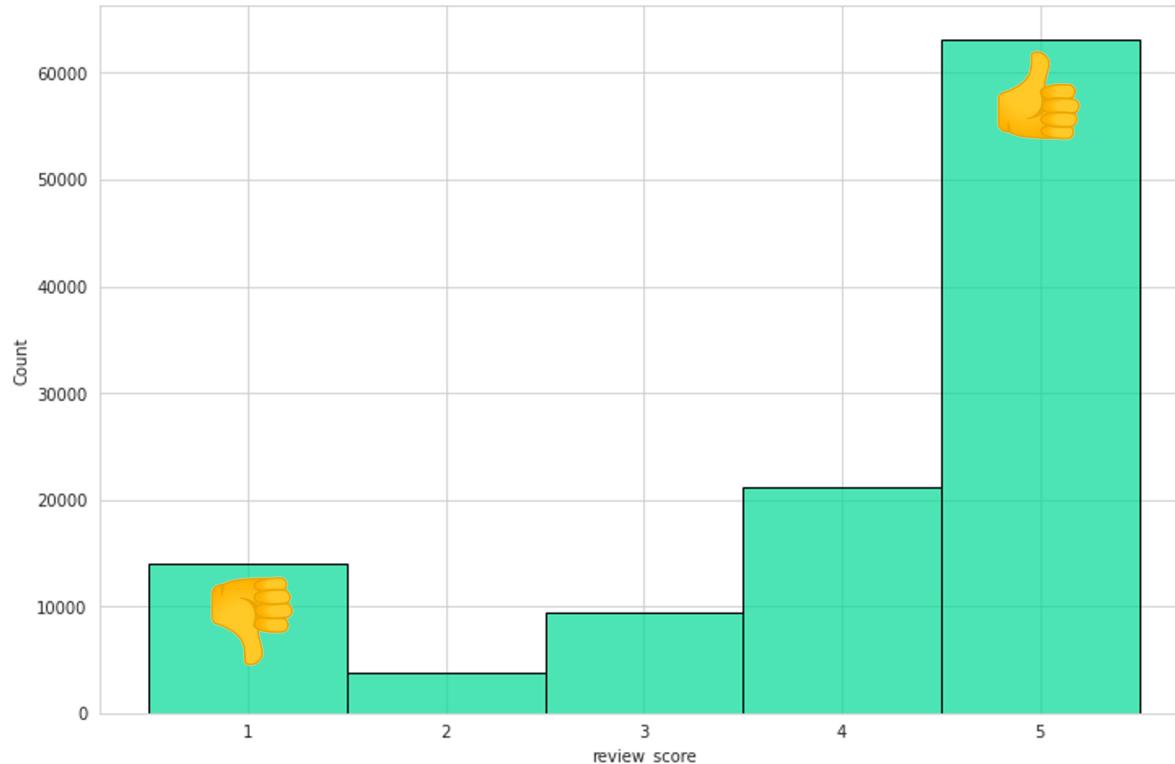
Analyse des données - Les moyens de paiements

olist



Analyse des données - Les notes de reviews

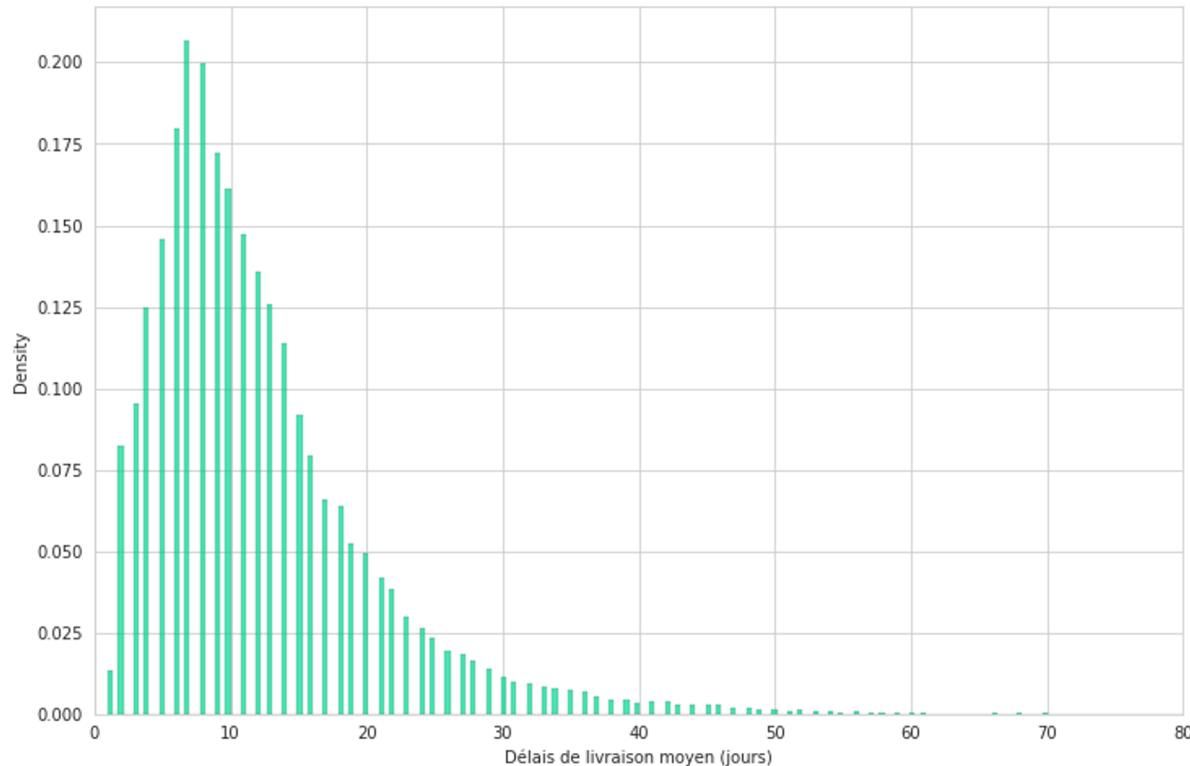
Répartition des notes attribuées aux commandes



Analyse des données - Les délais de livraisons

olist

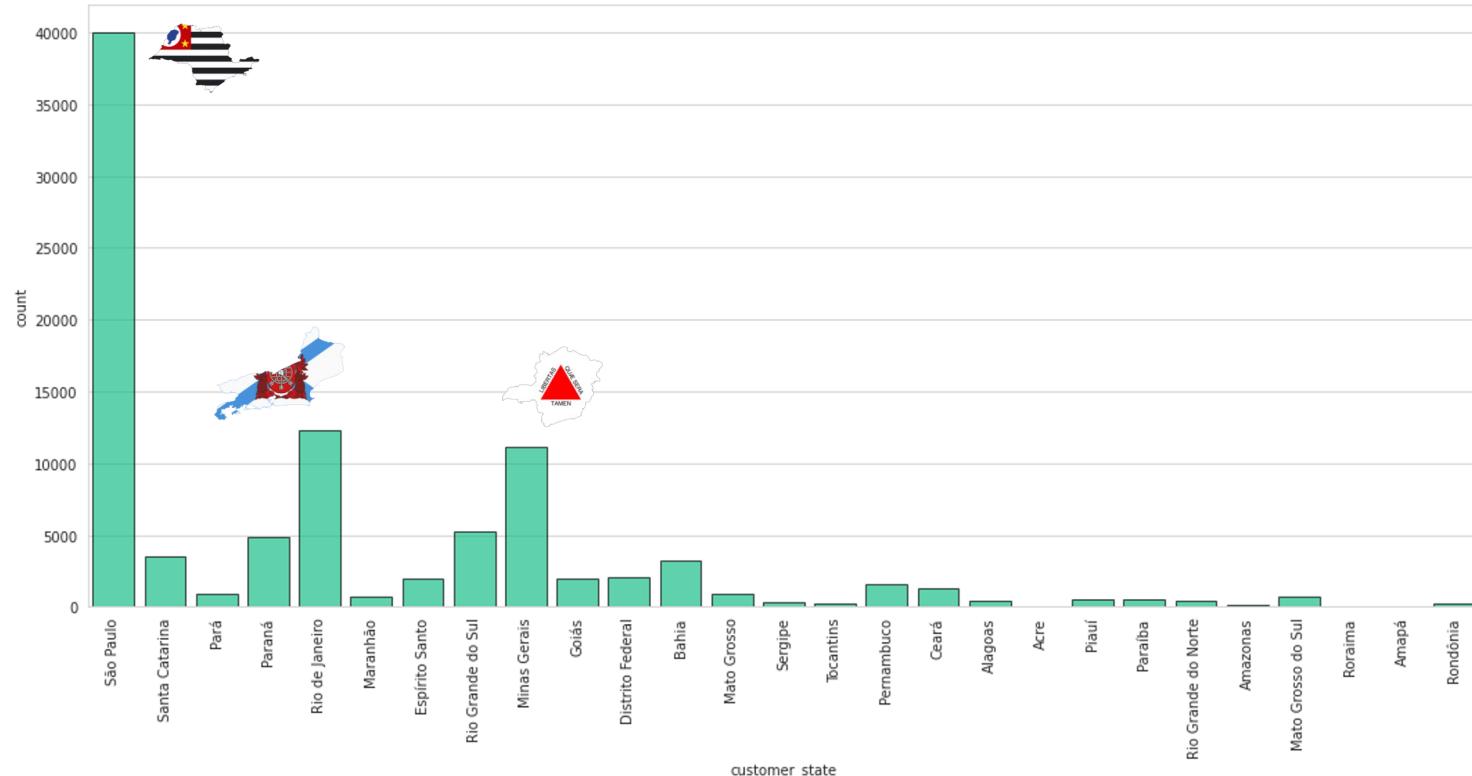
Répartition des délais de livraison moyens



Analyse des données - Les états brésiliens (1/2)

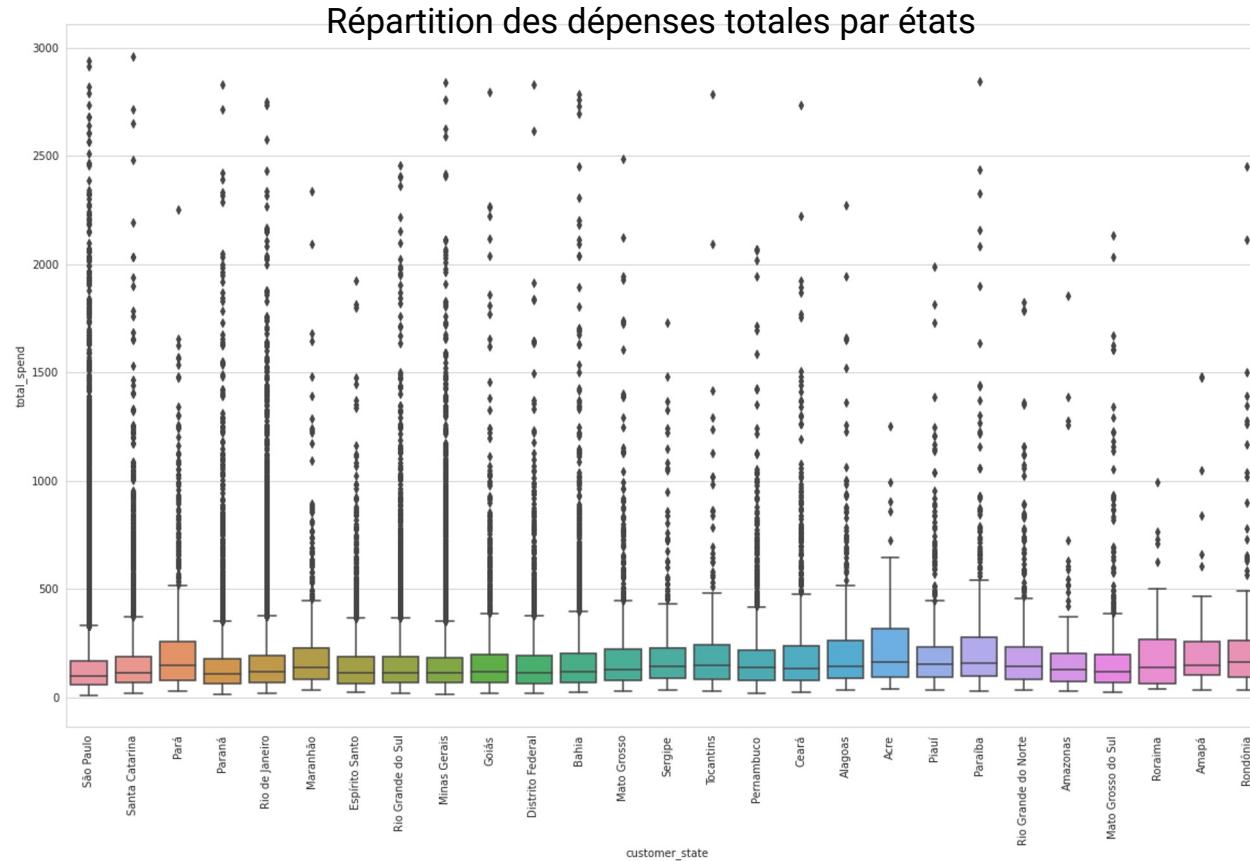
olist

Les états Brésiliens les plus représentées



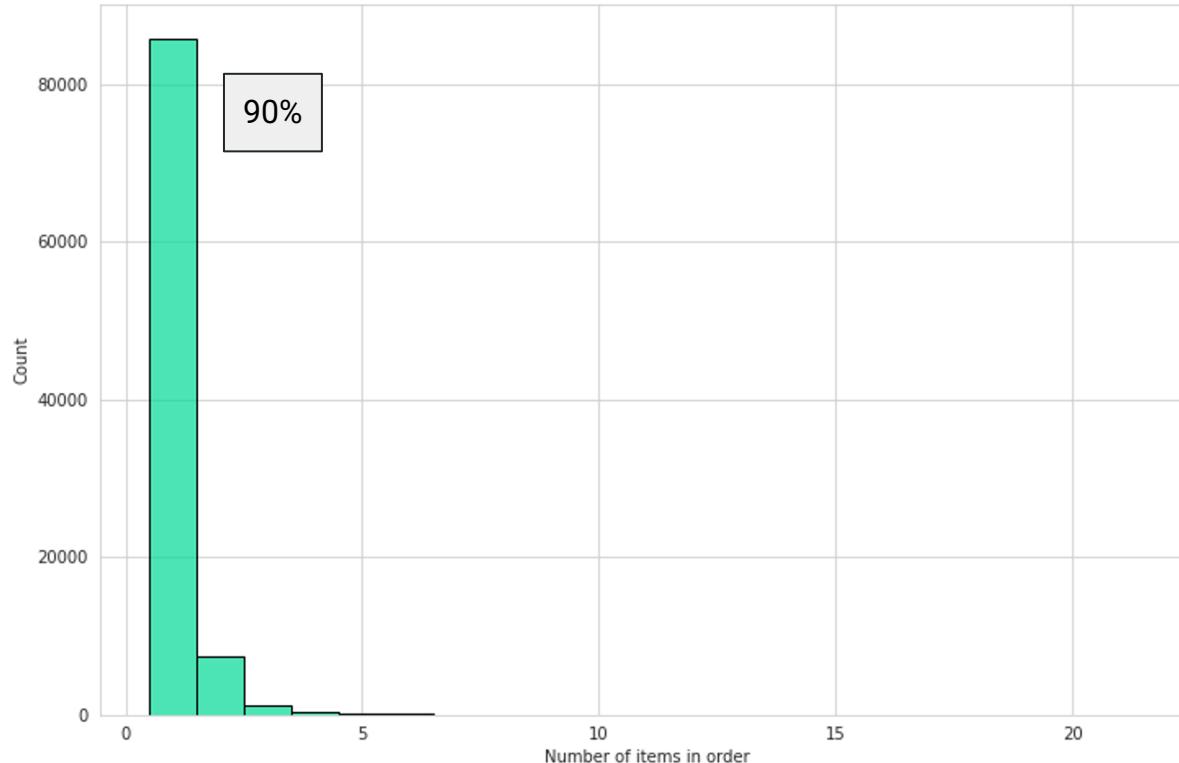
Analyse des données - Les états brésiliens (2/2)

olist



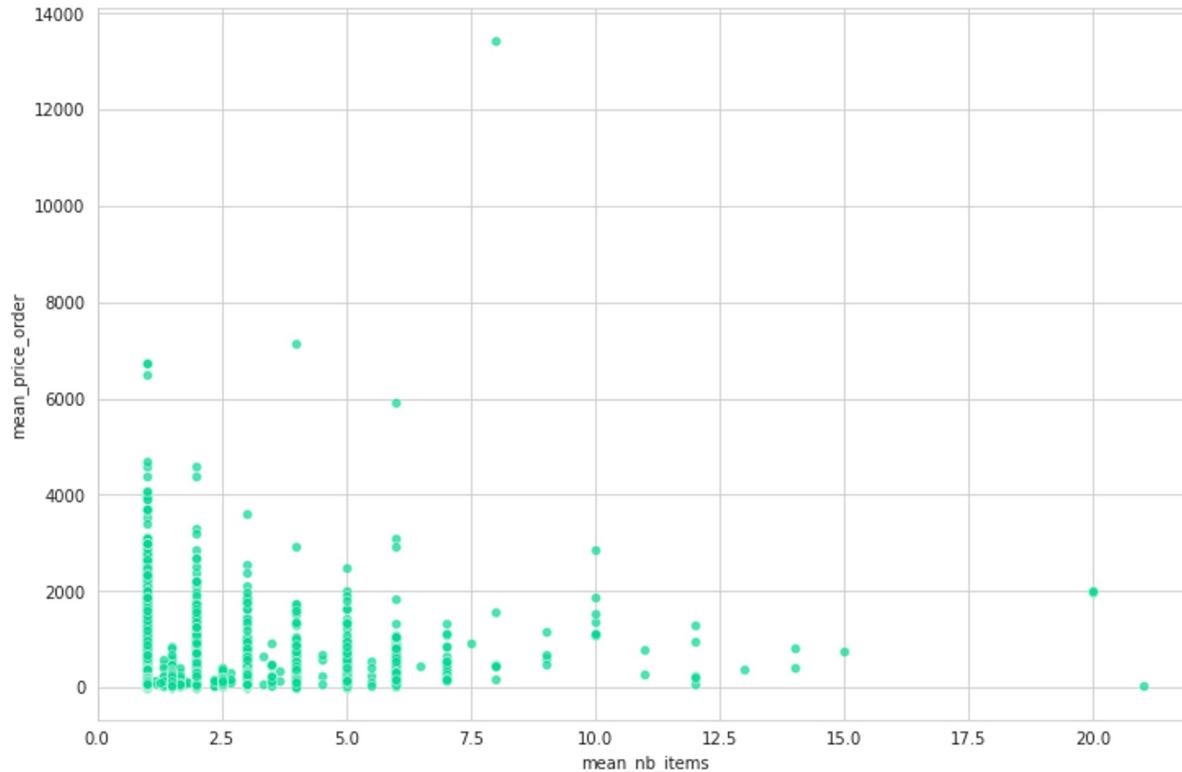
Analyse des données - Les articles (1/2)

Nombre moyen d'articles par commande



Analyse des données - Les articles (2/2)

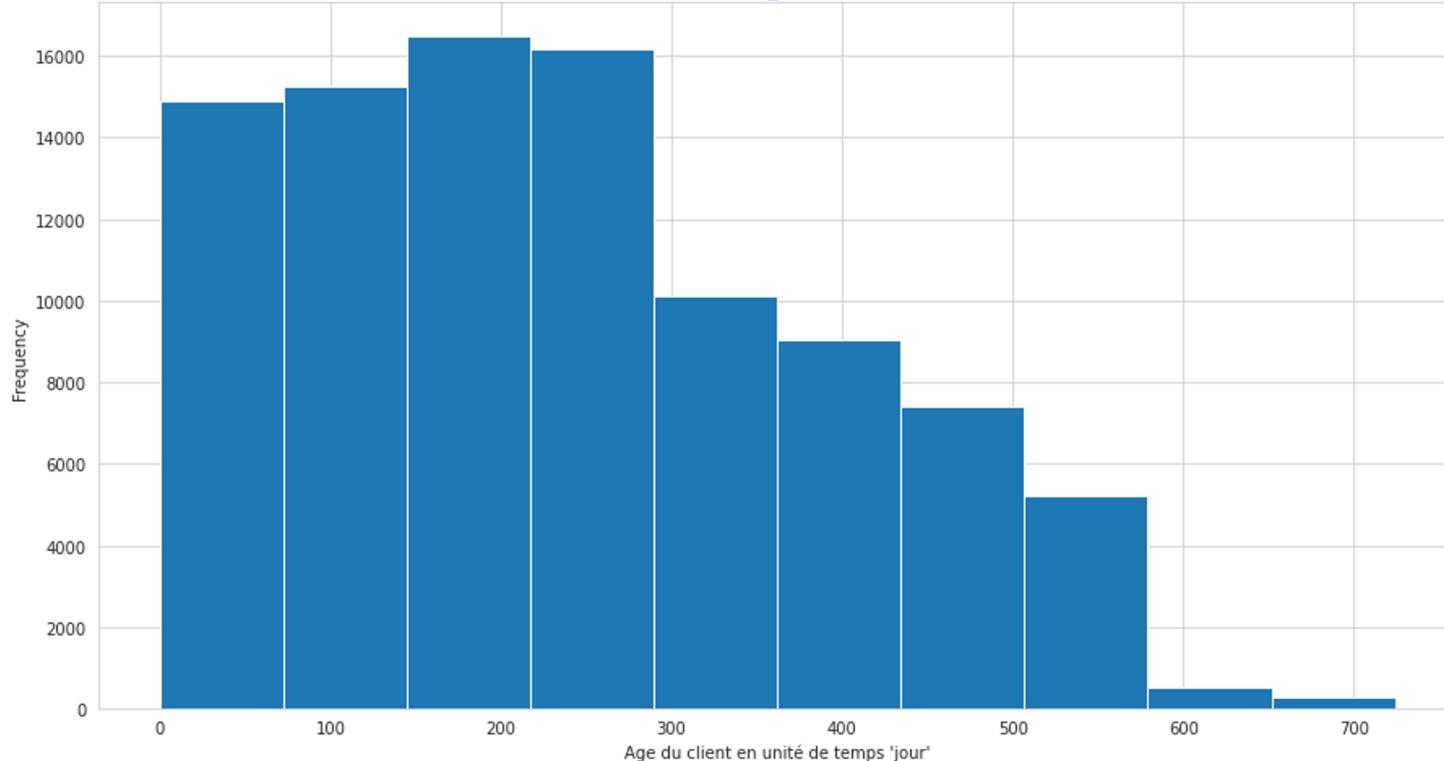
Répartition des prix moyen de commandes en fonction du nombre d'articles



RFM - Age des clients sur le site Olist

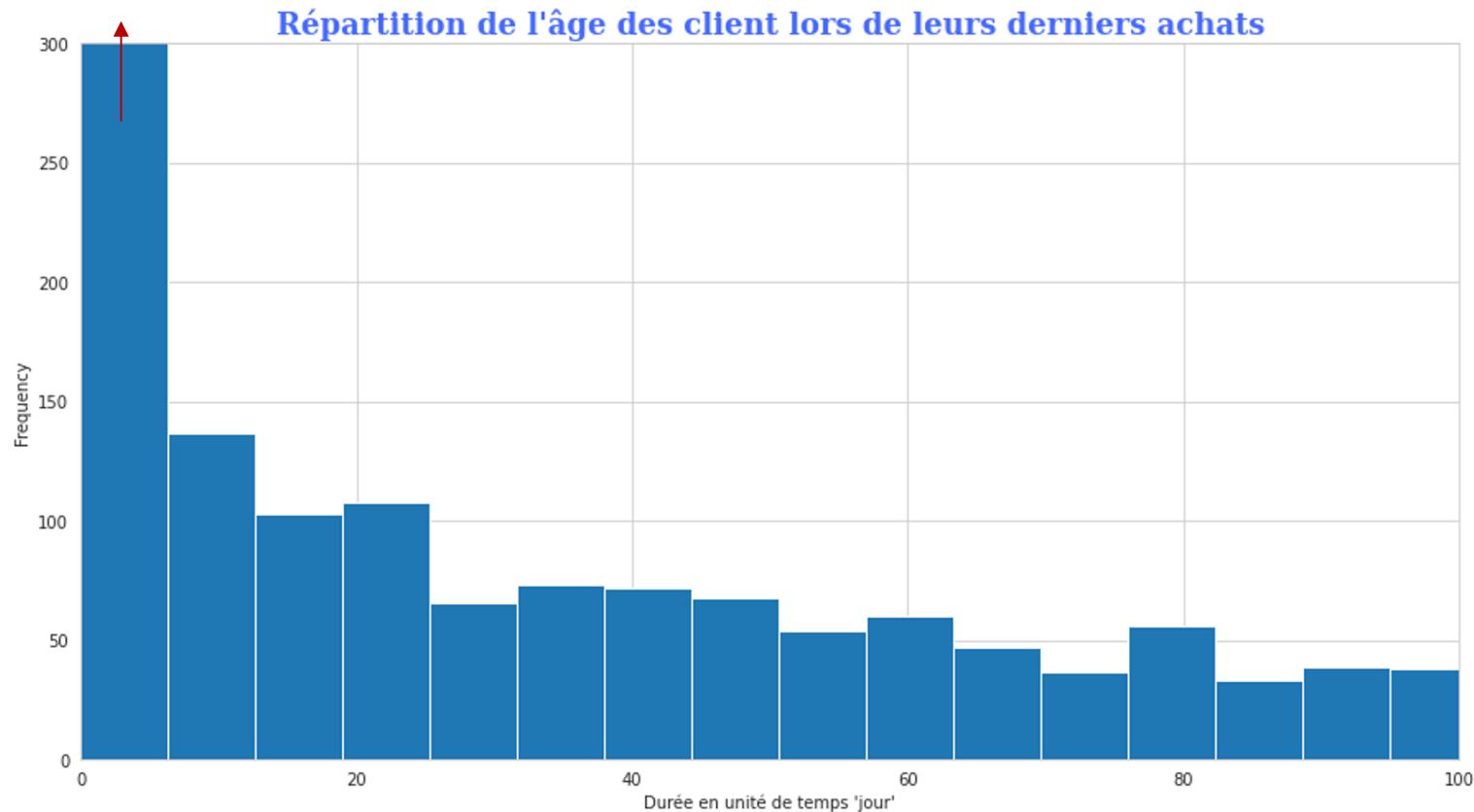
olist

Répartition de la durée entre le premier achat d'un client
et la fin de la période étudiée

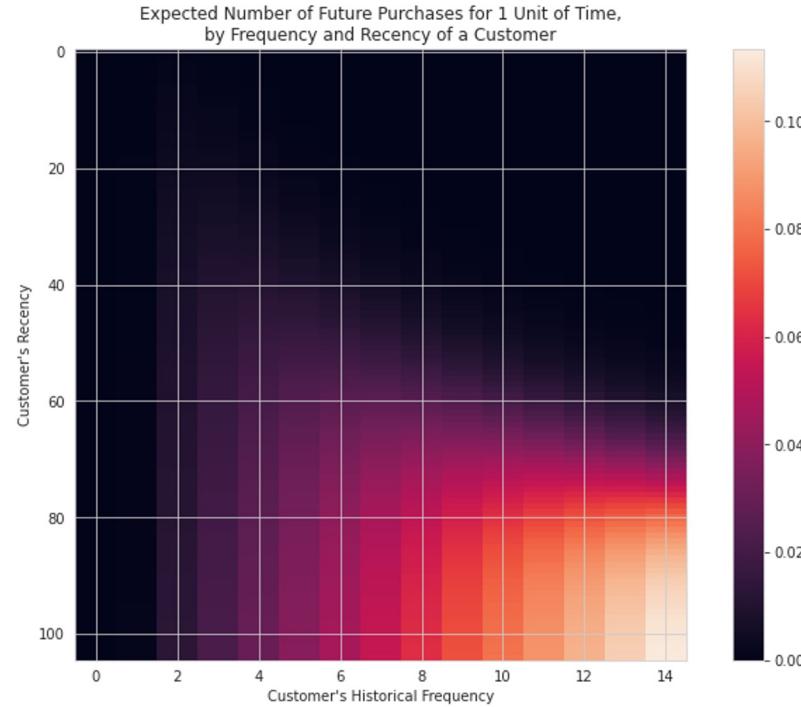
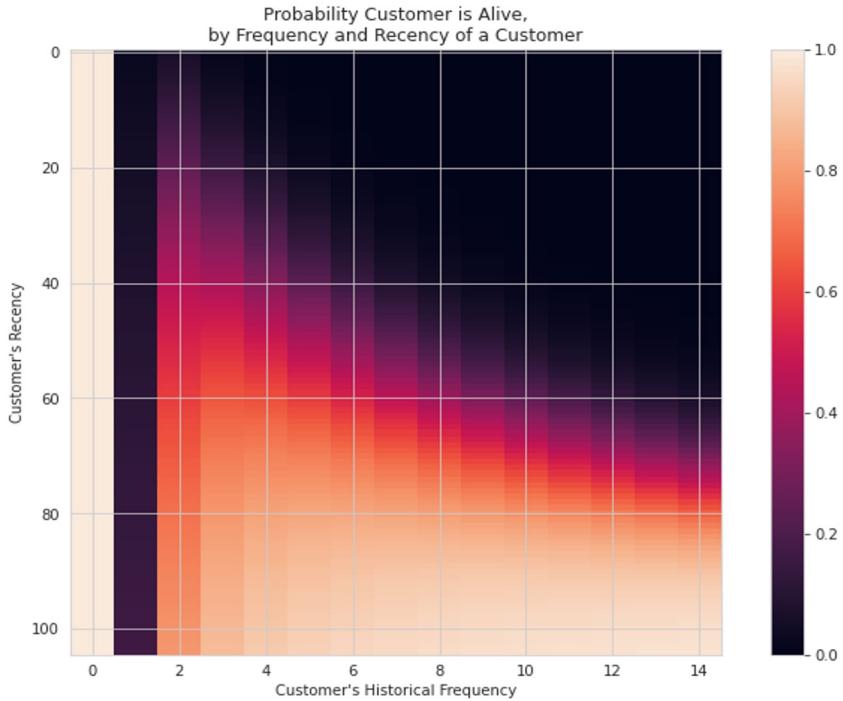


RFM - Age des clients lors du dernier achat

olist

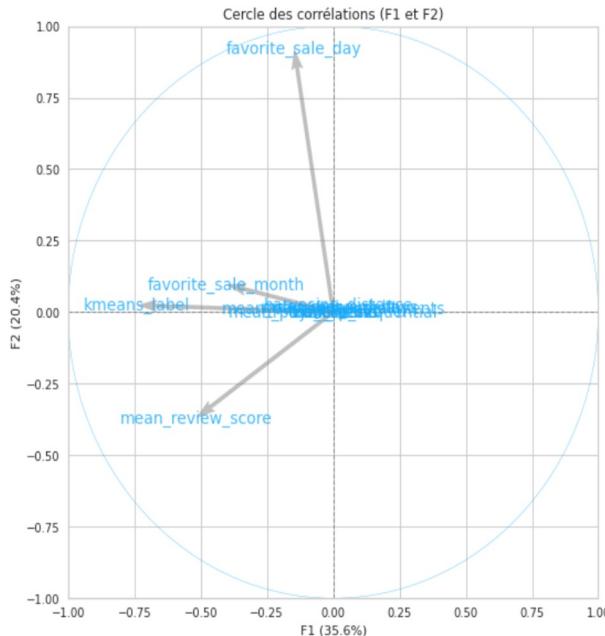


Analyses à l'aide du RFM

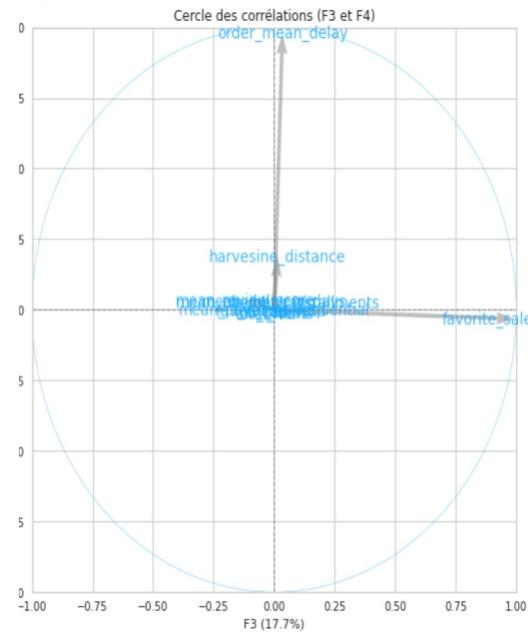


PCA - Cercles de corrélations

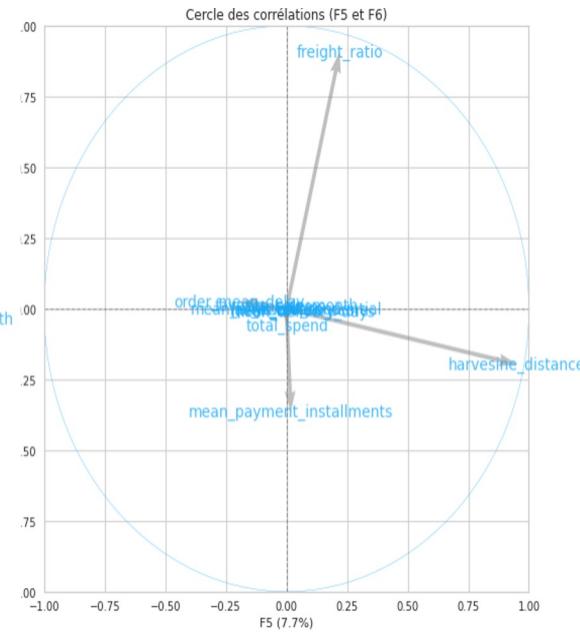
Cercles de corrélations des 6 premiers axes



F1 = mean_review_score
F2 = favorite_sale_day

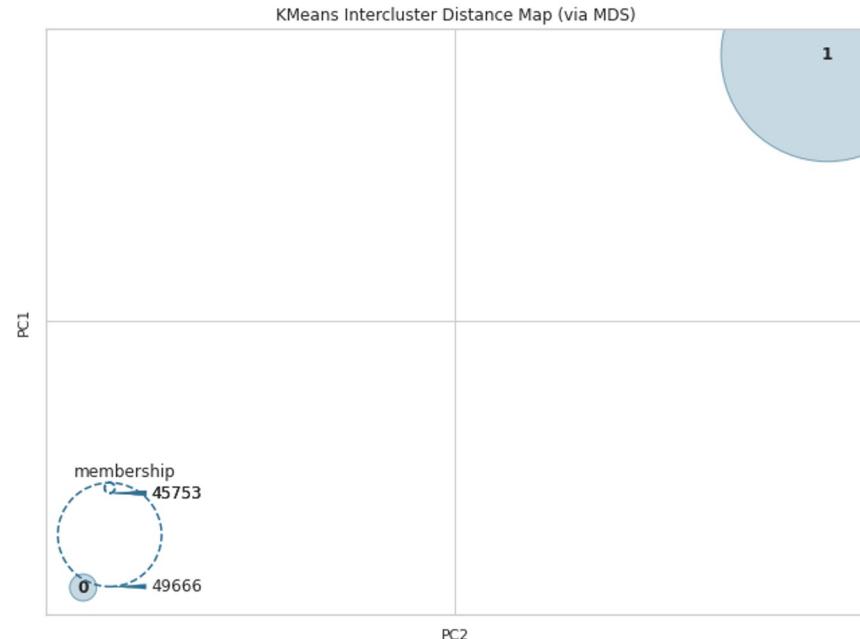
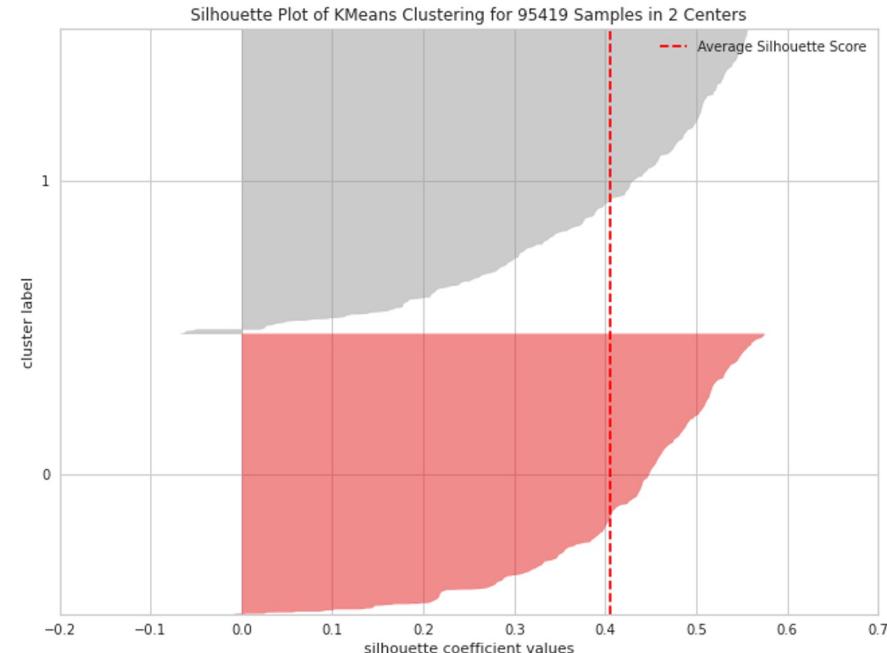


F3 = order_mean_delay
F4 = favorite_sales_month



F5 = freight_ratio
F6 = harvesine_distance

PCA - Réduction des dimensions K = 2

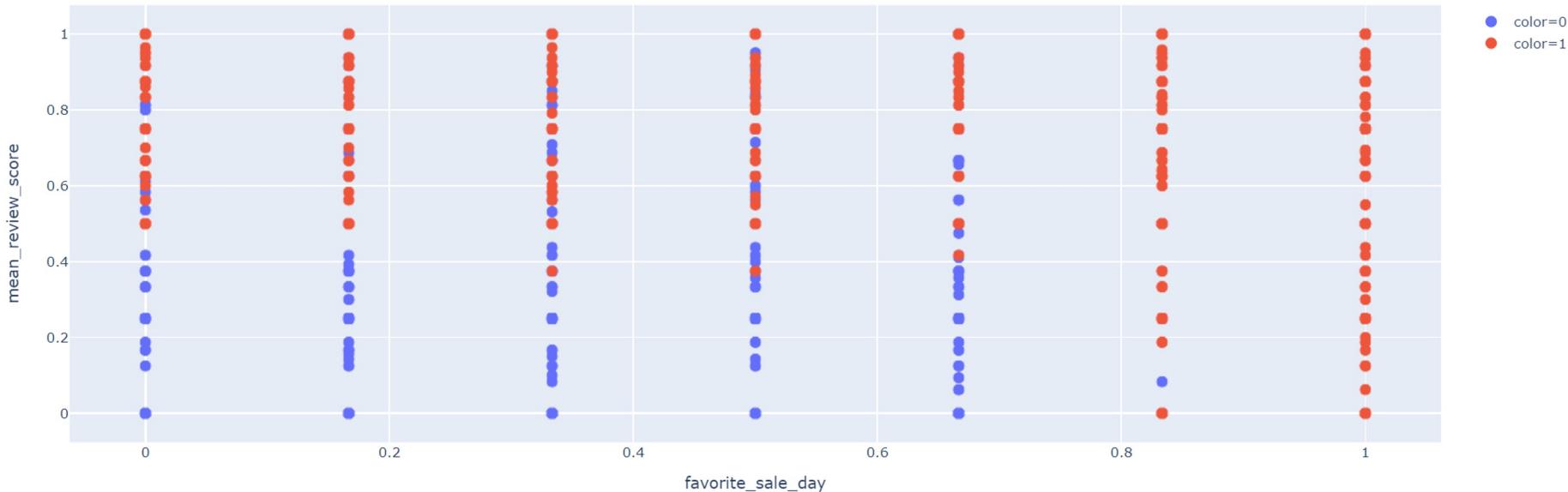


La réduction de dimension offre les mêmes axes de segmentation
les scores silhouette sont ici meilleurs comparés au données brutes



PCA - Exemple de segmentation avec PCA (Kmeans)

olist



HDBSCAN - Visualisation des clusters

olist

Clusters HDBSCAN - 3 clusters identifiés

