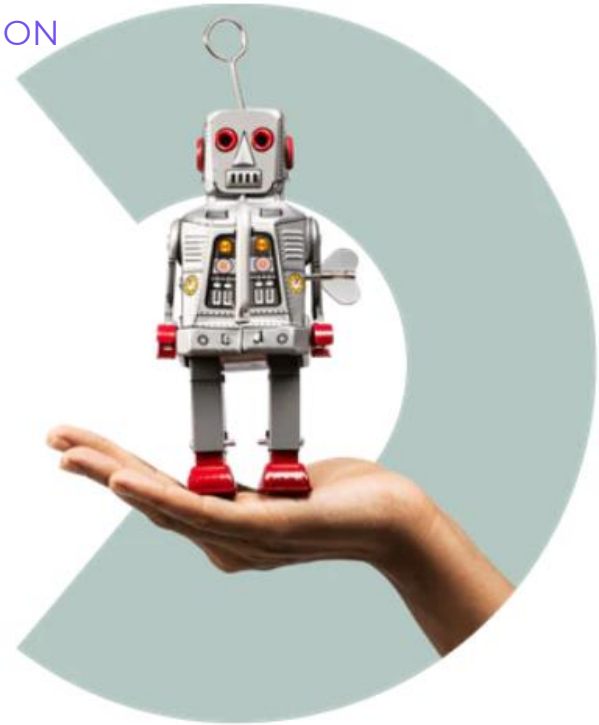
PROJET 5

FOURNIR UNE **SEGMENTATION DES CLIENTS** D'UN SITE E-COMMERCE , FACILE D'UTILISATION POUR CIBLER DES CAMPAGNES DE COMMUNICATION:

MODELISATION : CLASSIFICATION NON SUPERVISEE

#PEP8 #SEGMENTATION RFM #ARI #CLUSTERING #NORMALIZATION #FEATURES ENGINEERING #YELLOWBRICK #SKLEARN



Ingénieur IA

Développez et intégrez des algorithmes de Deep Learning au sein d'un produit IA

OPENCLASSROOMS

OUDDANE NABIL

SOMMAIRE

Projet 5

Segment de clients

A INTRODUCTION

1. Contexte

- ENJEU global:
  - Fournir une **segmentation des clients , facile d'utilisation** pour cibler des campagnes de communication
    - **Fournir une description** de votre segmentation et de sa logique sous-jacente pour une utilisation optimale
    - **comprendre les différents types d'utilisateurs** grâce à leur comportement et à leurs données personnelles
  - Fournir une **proposition de contrat de maintenance** basée sur une analyse de la stabilité des segments au cours du temps
- ENJEU DU P5: Modélisation **: Classification non supervisée**
- DONNEES SOURCES
  - le jeu de données :
    - https://www.kaggle.com/olistbr/brazilian-ecommerce
      - anonymisée comportant des informations sur l'historique de commandes, les produits achetés, les commentaires de satisfaction, et la localisation des clients depuis janvier 2017
      - créer les meilleures features pour les exploiter

2. Objectifs

- SCRIPT:
  - Le code fourni doit respecter la convention PEP8
    - https://pep8.org/
    - Pycodestyle / flake8
    - Autopep8 extension pour jupyter
  - Notebook d'analyse exploratoire non cleané
  - Notebook d essais des différentes approches de modélisation

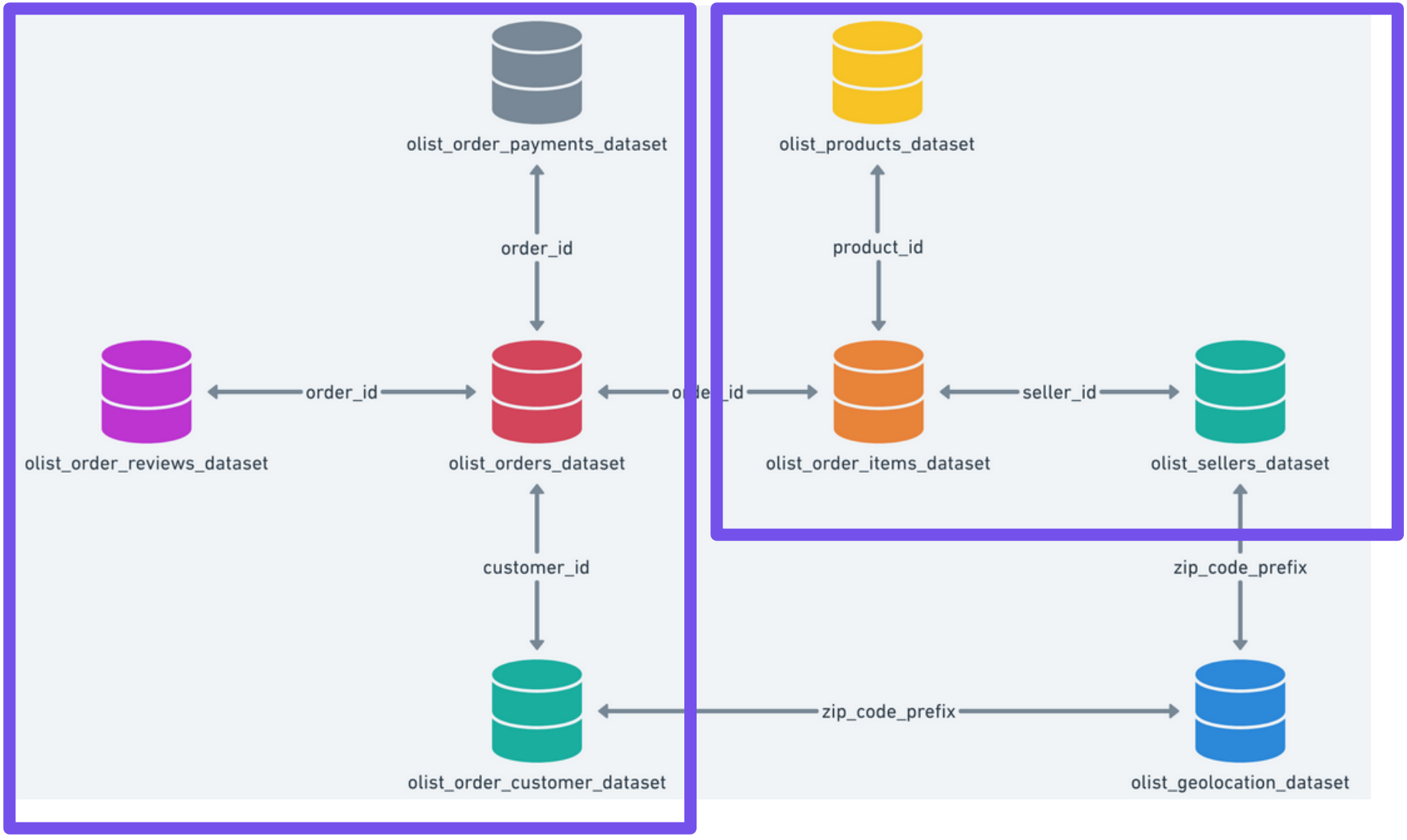
3. Ressources complémentaires

- Segmentation RFM
  - https://www.definitions-marketing.com/definition/segmentation-rfm/
- ARI index
  - https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html
- Kaggle: exemple d'evaluation de clustering
  - https://www.kaggle.com/kautumn06/yellowbrick-clustering-evaluation-examples

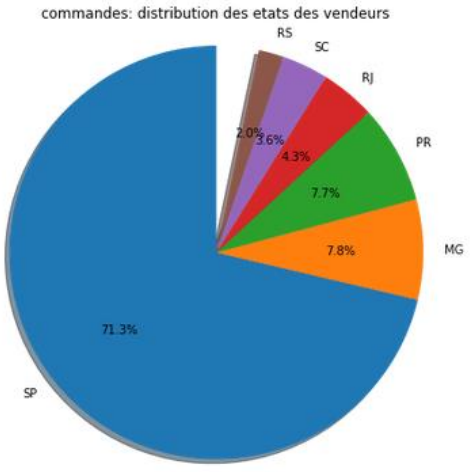
B

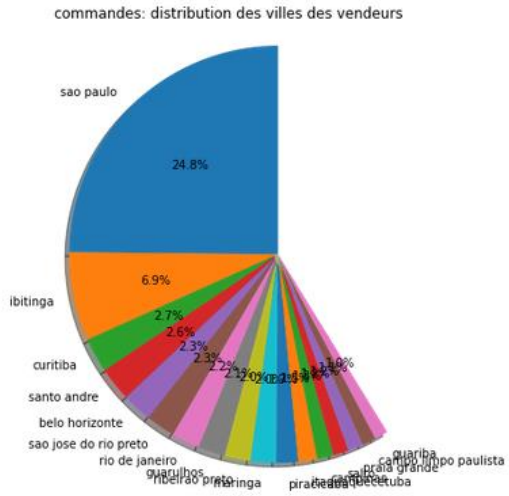
PROJET: ANALYSE EXPLORATOIRE

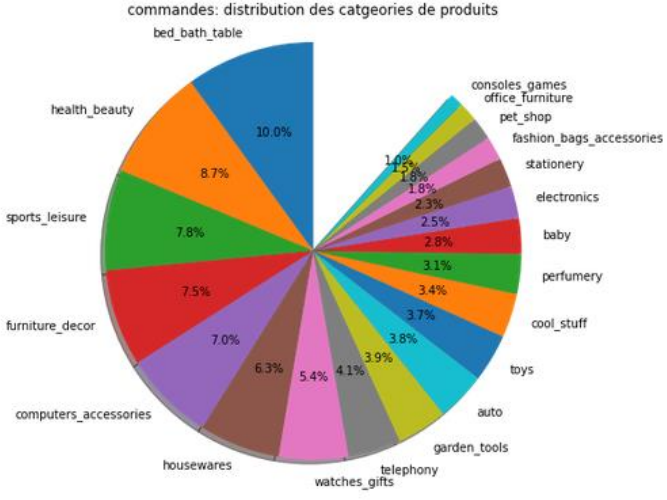
B: Analyse exploratoire
1-Schéma de la base de donnée d'e-commerce

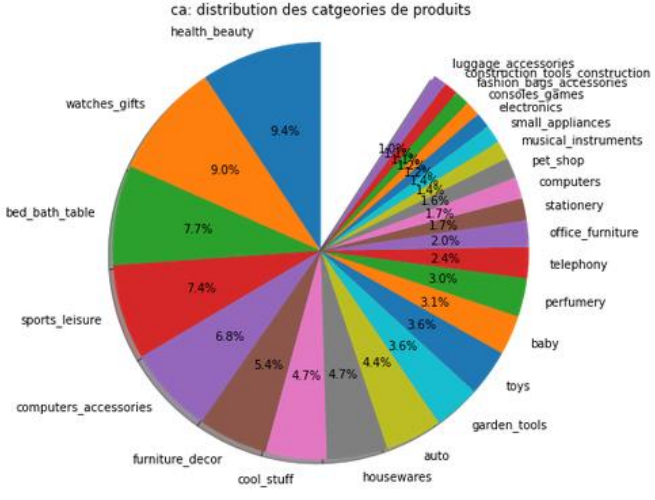


B: Analyse exploratoire
2-Produits et vendeurs





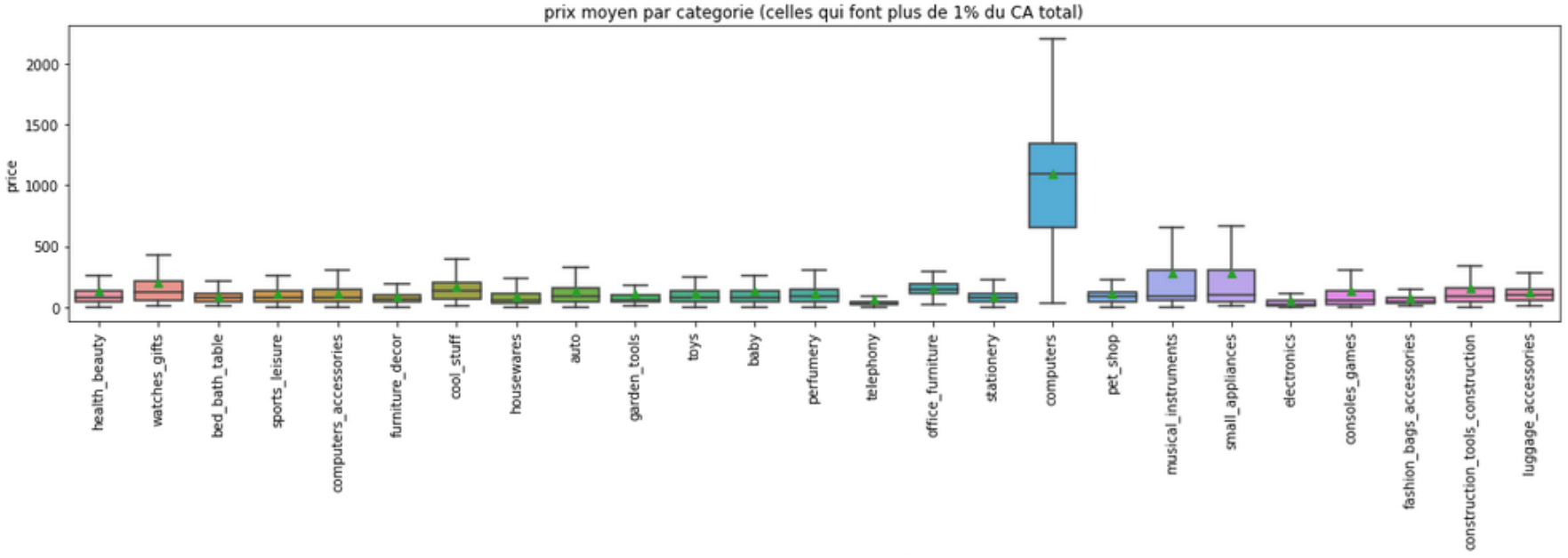




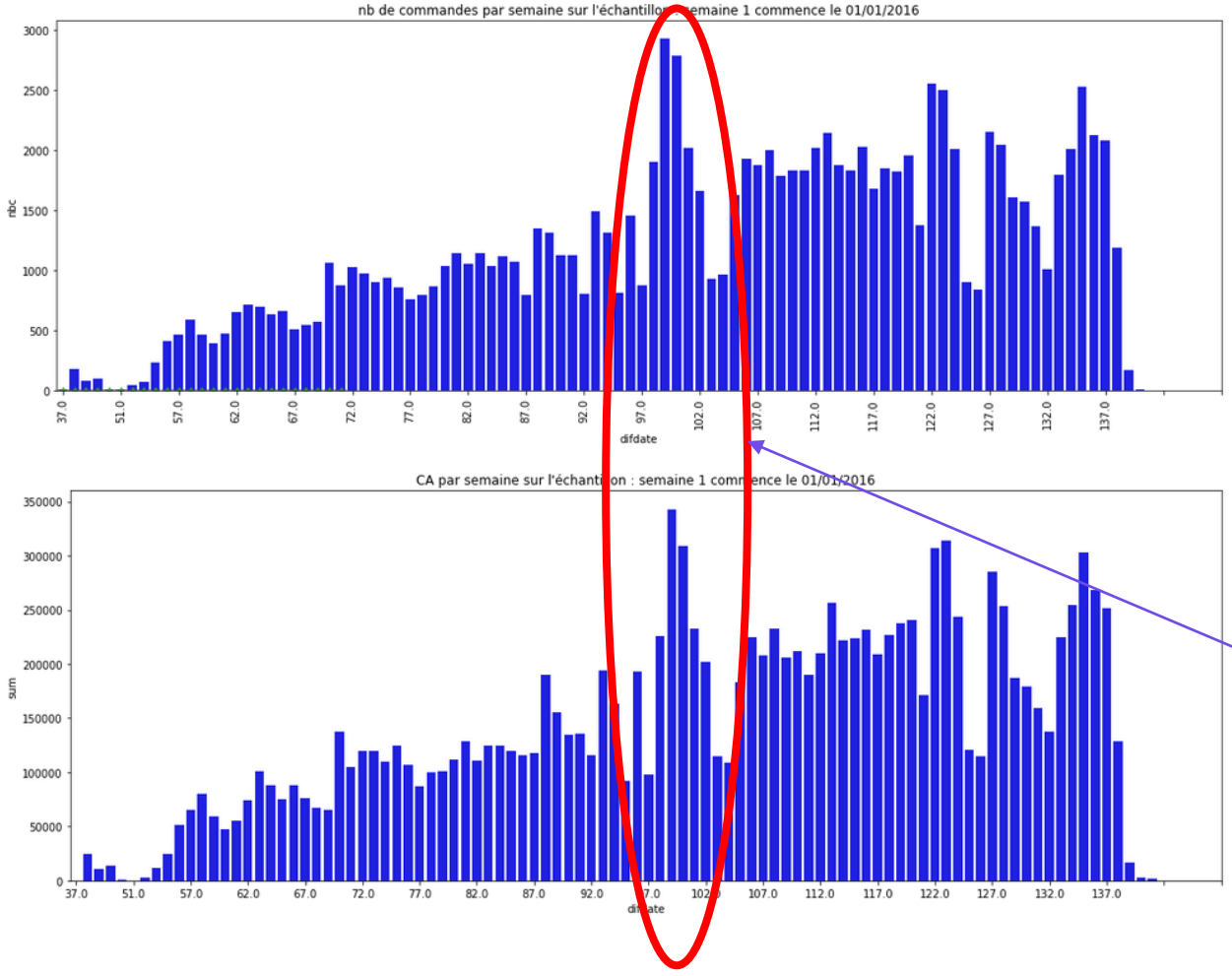
- **¾ des vendeurs se trouvent dans l'état de Sao Paulo et même ¼ dans la ville de Sao Paulo**
- **Les categories de produits les plus commandées en nombre et en chiffre d'affaires sont: les produits de beauté, les accessoires de decoration, les articles de sport, les produits manager ainsi que les montres**

B: Analyse exploratoire
2-Produits et vendeurs

- Les prix des produits dont les categories représentent plus de 1% du chiffre d'affaires total se situent **entre 0 et 250/300 real ou dollar**
- Seul le prix moyens/medians des **ordinateurs arrivent jusqu'à 1000**



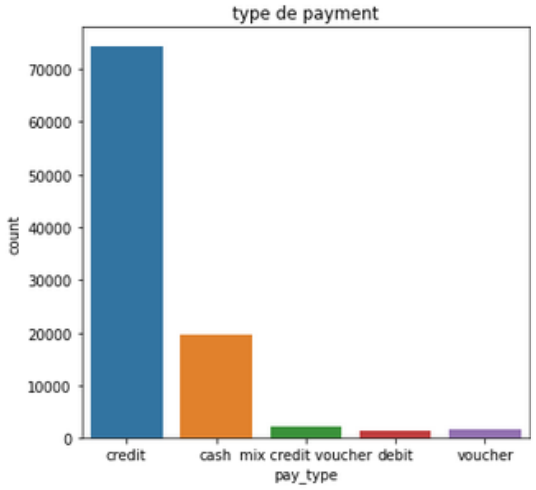
B: Analyse exploratoire: 2-Produits et vendeurs

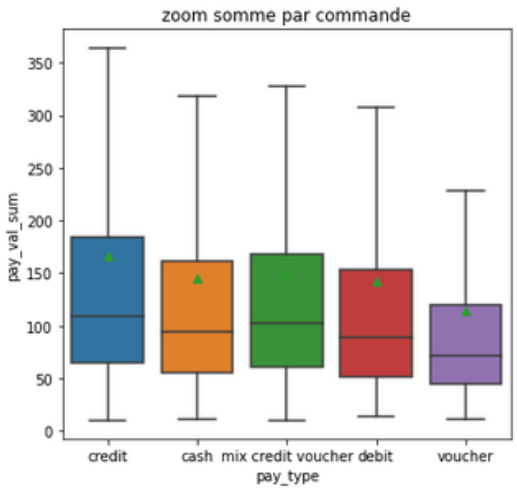


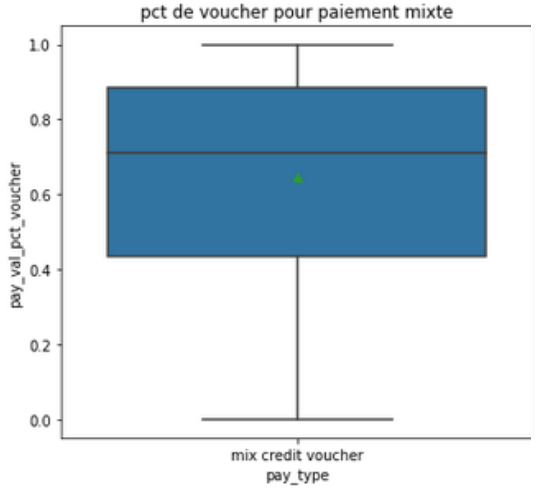
- Sur la période 2017-2018, on observe clairement un trend haussier du nombre de commandes et du chiffre d'affaires

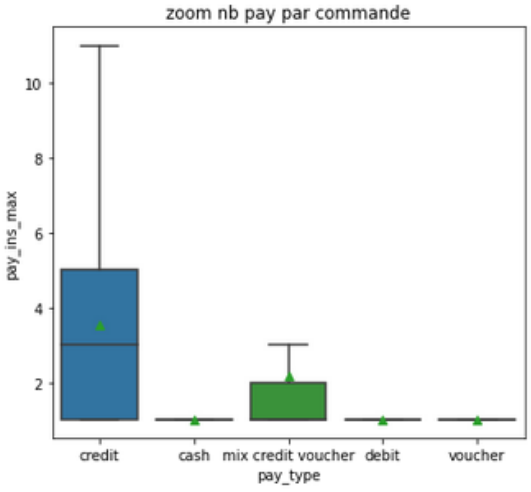
- On observe un pic sur la période precedent noel

B: Analyse exploratoire: 3-Types de paiement



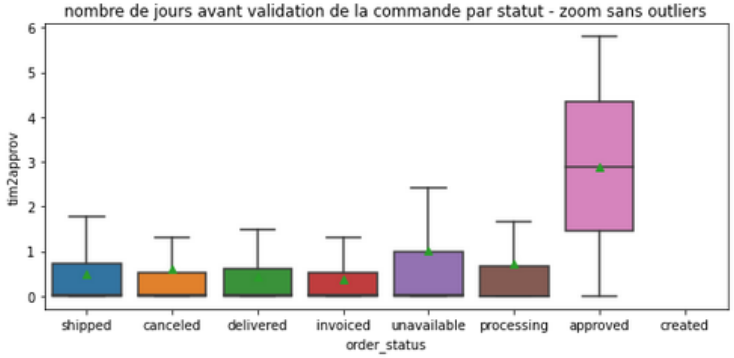


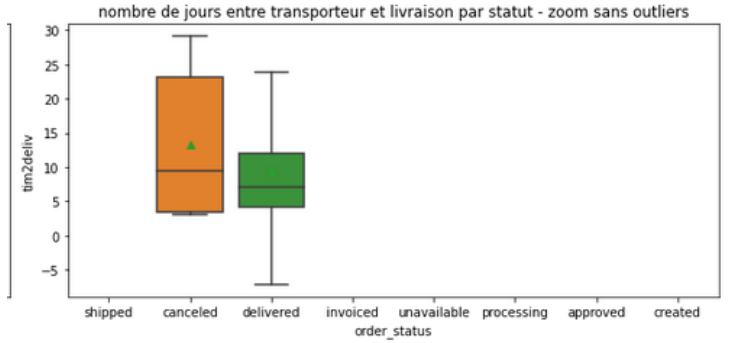


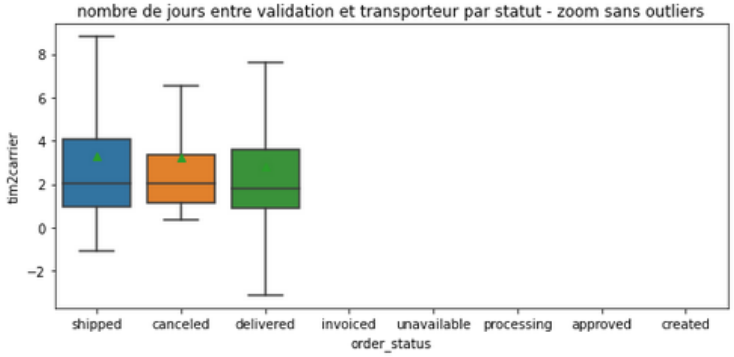


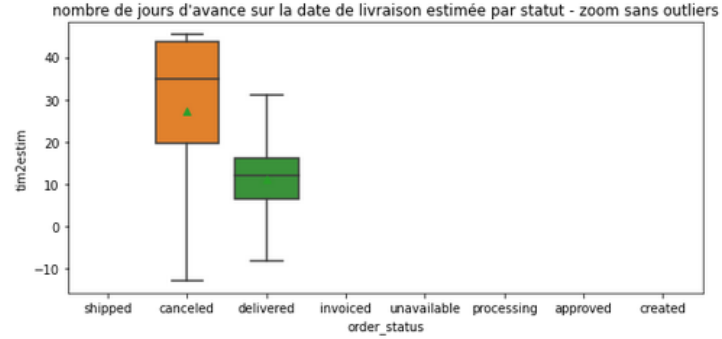
- **75% des paiement se font par credit**
- **20% cash**
- **Les personnes payant à credit depensent un peu plus en Moyenne**
- **Les personnes utilisant les vouchers payent parfois la difference à credit**
- **Au final on peut separer le mode de paiement en 2 categories distinctes : credit-voucher / cash-debit**

B: Analyse exploratoire:
4-Temps de traitement des commandes



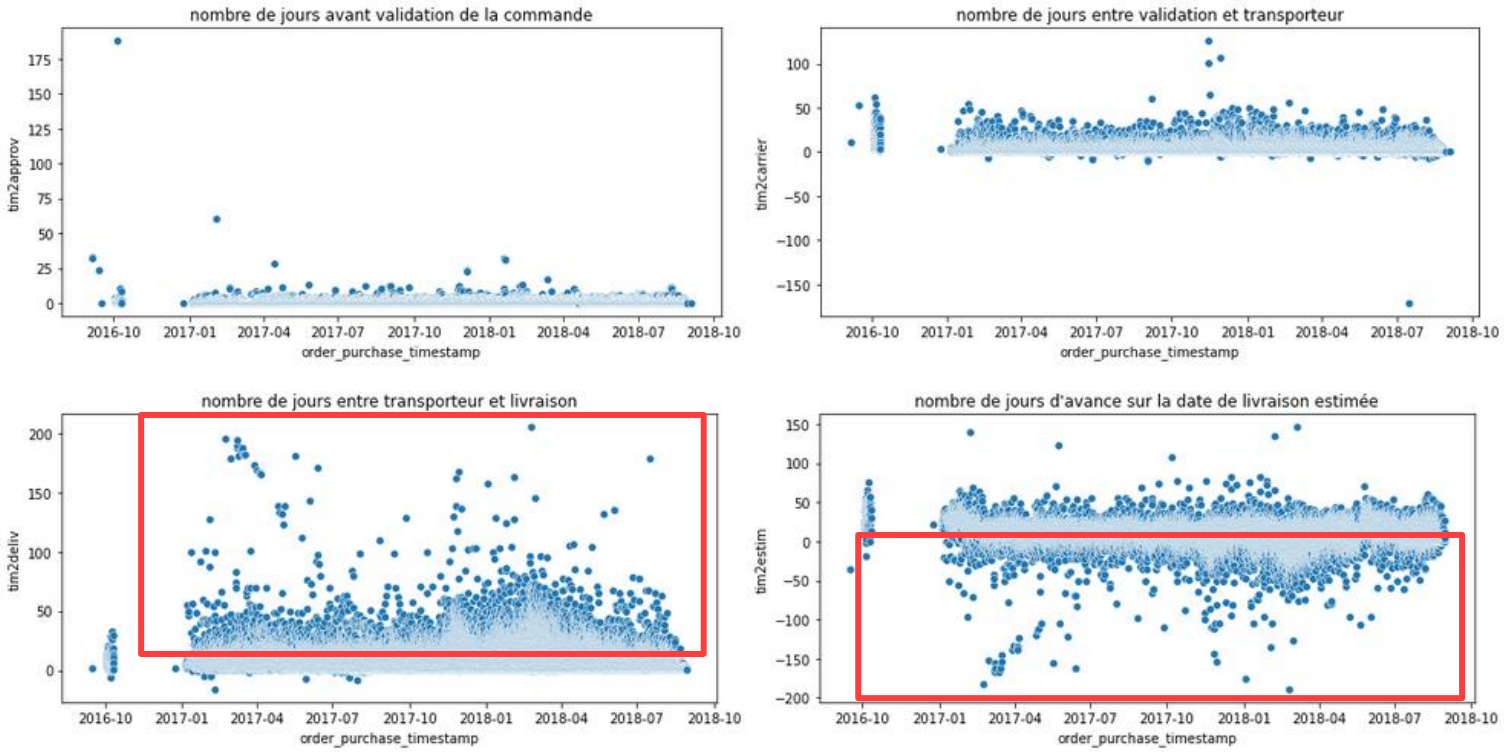






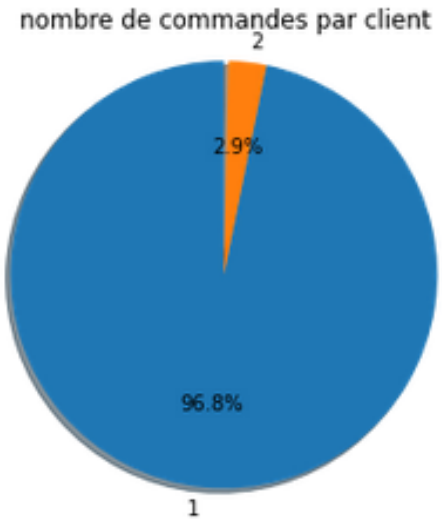
- **Validation de la commande en moins d'un jour en Moyenne**
- **Remise au transporteur au bout de 1 à 4 jours après validation**
- **Livraison de la commande entre 5 et 10 jours en moyenne**
- **Les livraisons sont souvent en avance sur la date estimée au départ**

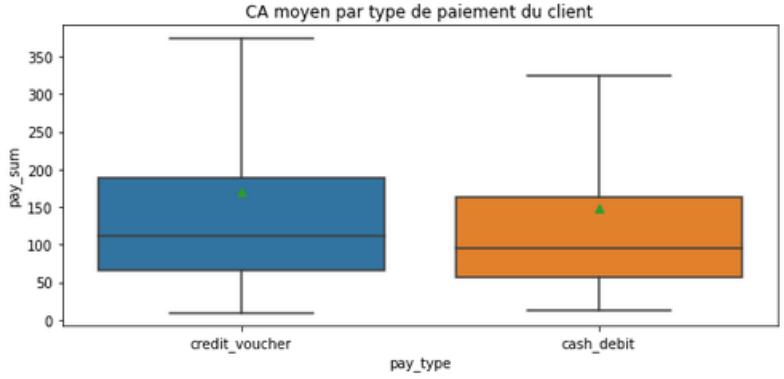
B: Analyse exploratoire:
4-Temps de traitement des commandes



- **Il n'est pas rare d'avoir des livraisons tres lentes**
- **Ou des livraisons qui prennent beaucoup plus de temps que prévu**

B: Analyse exploratoire: 5-Clients et commandes

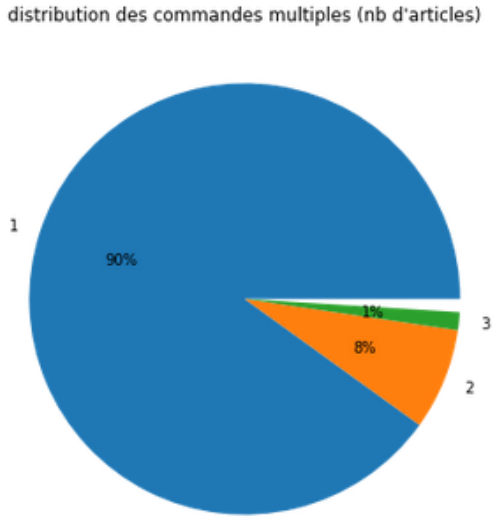


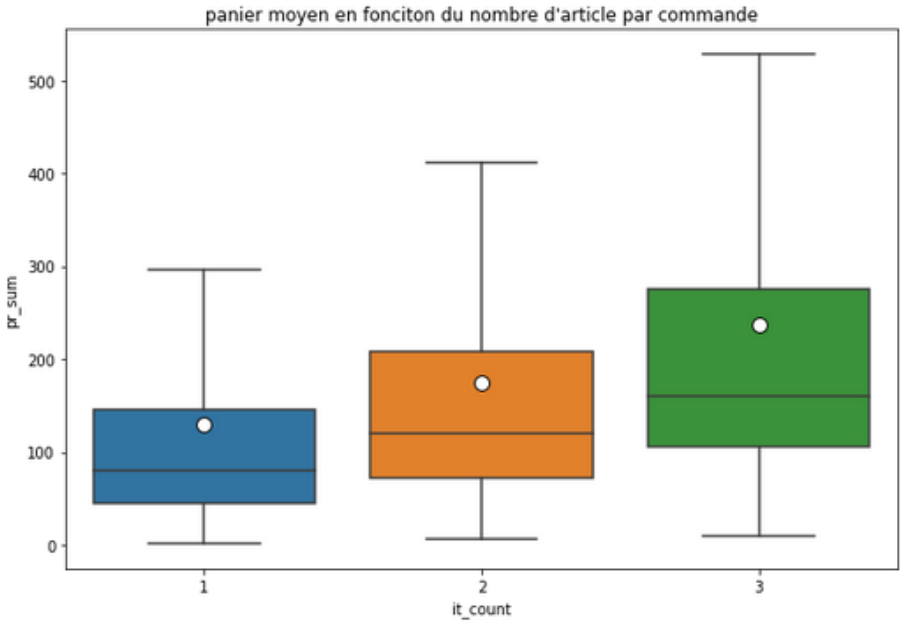


- **97% des clients n'ont passé qu'une seule commande sur la période**
- **Le panier moyen des clients à credit est legèrement plus élevé que pour les clients payant en cash ou debit immédiat,**
- **Le panier moyen des clients à crédit est autour de 170 alors qu il tourne autour de 150 pour les clients à cash/debit**

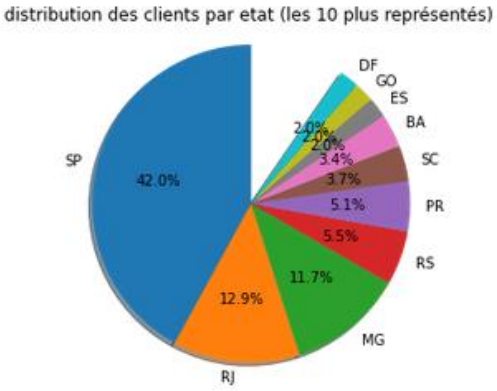
B: Analyse exploratoire:
5-Clients et commandes

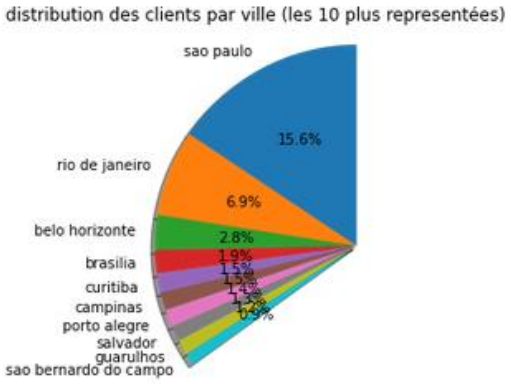
- **90% des commandes ne portent que sur un seul article, 8% sur 2 et 1% sur 3**
- **Le panier moyen est logiquement croissant avec le nombre d'articles commandés**

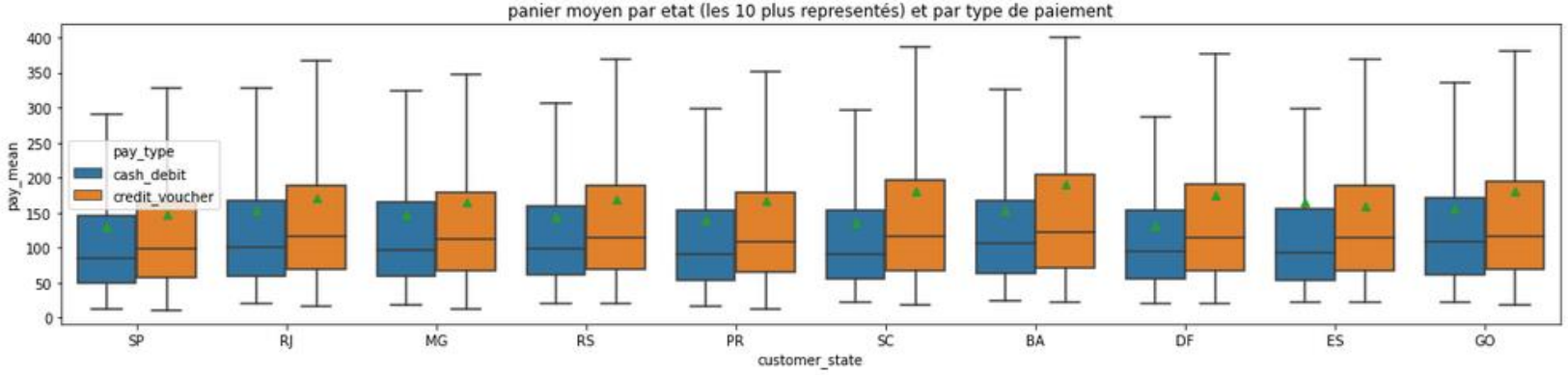




B: Analyse exploratoire: 6-Clients et geographie

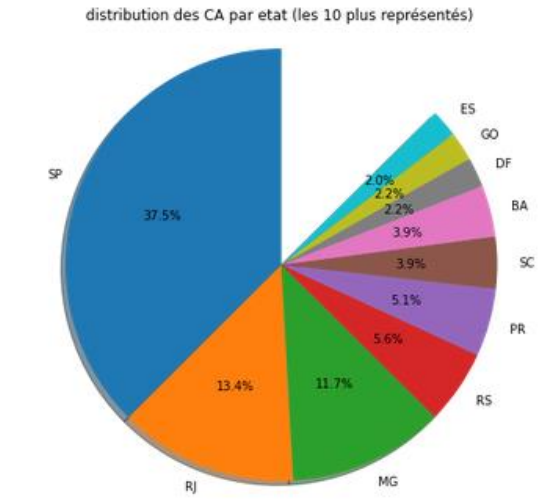


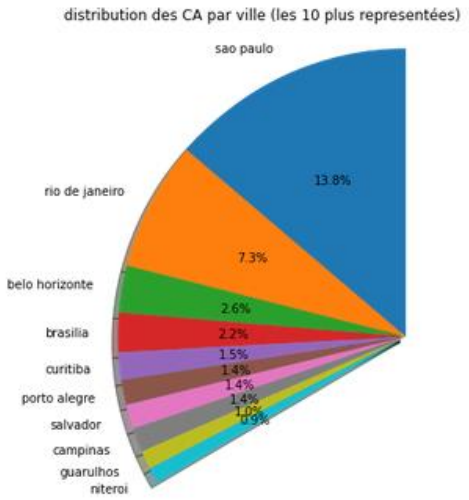


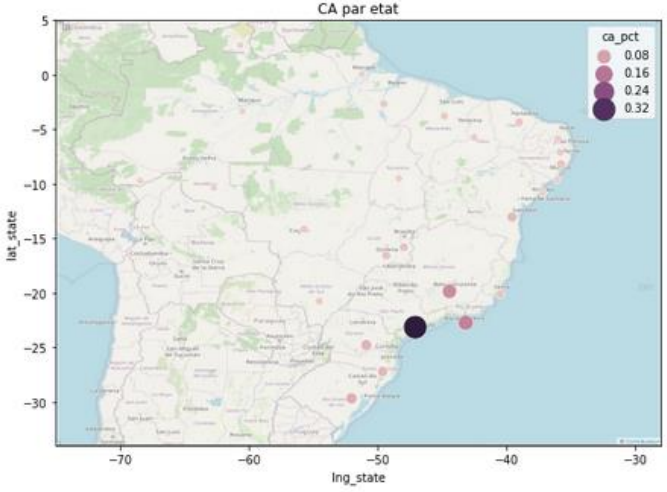


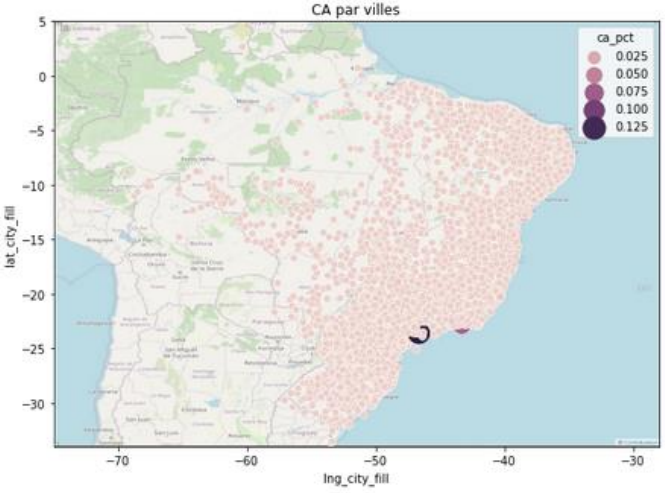
- **42% des clients se trouvent dans l'état de Sao Paulo (15,6% dans la ville de Sao Paulo)**
- **13% des clients proviennent de l'état de Rio de Janeiro et semblent dépenser légèrement plus que les clients de Sao Paulo**

B: Analyse exploratoire: 7-Geographie et CA



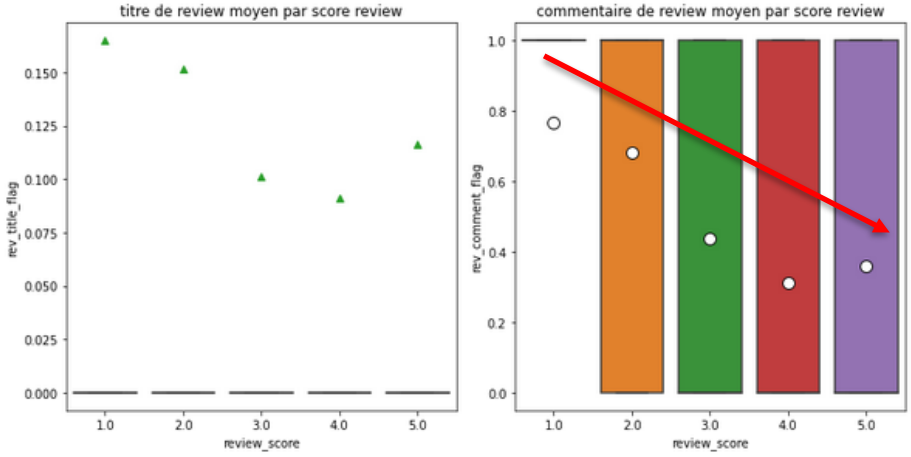






- **37,5% du chiffre d'affaires provient de l'état de Sao Paulo**
- **13,4% du CA provient de l'état de Rio de Janeiro,**
- **La plupart du CA provient des états cotiers**

B: Analyse exploratoire:
8-Clients et score review



- **57% à 76% des clients sont ravis de leur experience**
- **11 à 14% des clients sont assez mécontents**
- **70 à 80% des clients mécontents ont posté un commentaire alors que moins de 40% des clients satisfaits l'ont fait**

C

PROJET: SEGMENTATION MODELISATION

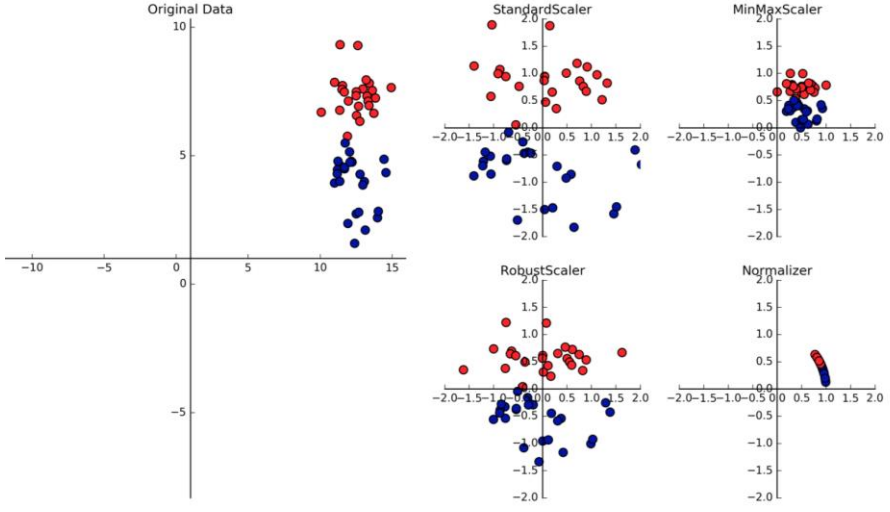
C: Modelisation
1-Préambule - choix de l'algo: Kmeans un bon choix de départ

- **Favoriser une approche inductive permettant de prédire un label pour de nouvelles données (nouveaux clients) sans refit**
- **Pour une segmentation de clients , nous ne voulons pas trop de clusters mais de taille proche si possible afin de faciliter le travail des équipes marketing**
- **Kmeans semble dans notre cas le 1er algorithme à tester (de plus adapté à un usage general)**

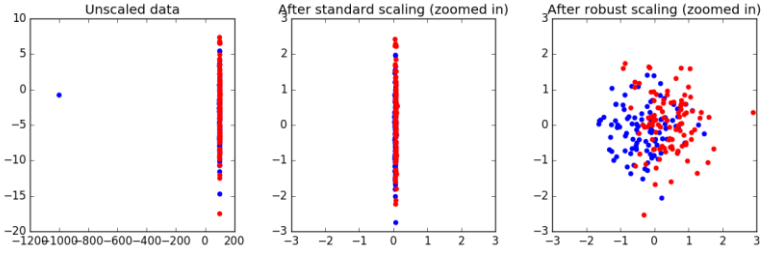
Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters, inductive	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry, inductive	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry, inductive	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry, transductive	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters or distance threshold	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, transductive	Distances between points
Agglomerative clustering	number of clusters or distance threshold, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances, transductive	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes, outlier removal, transductive	Distances between nearest points
OPTICS	minimum cluster membership	Very large `n_samples`, large `n_clusters`	Non-flat geometry, uneven cluster sizes, variable cluster density, outlier removal, transductive	Distances between points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation, inductive	Mahalanobis distances to centers
BIRCH	branching factor, threshold, optional global clusterer.	Large `n_clusters` and `n_samples`	Large dataset, outlier removal, data reduction, inductive	Euclidean distance between points

C: Modelisation
1-Préambule - choix de la normalisation

Preprocessing Type	Scikit-learn Function	Range	Mean	Distribution Characteristics	When Use	Definition	Notes
Scale	MinMaxScaler	0 to 1 default, can override	varies	Bounded	When want a light touch.	Add or substract a constant. Then multiply or divide by another constant. MinMaxScaler subtracts the mimimum value in the column and then divides by the difference between the original maximum and original minimum.	Preserves the shape of the original distribution. Doesn't reduce the importance of outliers. Least disruptive to the information in the original data. Default range for MinMaxScaler is 0 to 1.
Standardize	RobustScaler	varies	varies	Unbounded	Use if have outliers and don't want them to have much influence.	RobustScaler standardizes a feature by removing the median and dividing each feature by the interquartile range.	Outliers have less influence than with MinMaxScaler. Range is larger than MinMaxScaler or StandardScaler.
Standardize	StandardScaler	varies	0	Unbounded, Unit variance	When need to transform a feature has zero mean and unit standard deviation. It's my go-to.	StandardScaler standardizes a feature by removing the mean and dividing each value by the standard deviation.	Results in a distribution with a standard deviation equal to 1 (and variance equal to 1). If you have outliers in your feature (column), normalizing your data will scale most of the data to a small interval.
Normalize	Normalizer	varies	0	Unit norm	Rarely.	An observation (row) is normalized by applying l2 (Euclidian) normalization. If each element were squared and summed, the total would equal 1. Could also specify l1 (Manhatten) normalization.	Normalizes each sample observation (row), not the feature (column)!



- **Necessité de rendre comparables les variables en les normalisant**
- **MinMax scaler preserve la forme de la distribution initiale mais ne réduit pas l'influence des outliers:**
  - **En presence de données positives et d'outliers , ce scaling va concenter les données autour du point (0.0.0....0)**
- **En terme de reduction d'influence des outliers, Robust scaler est le Meilleur . Vient ensuite Standard scaler souvent bon choix de départ**

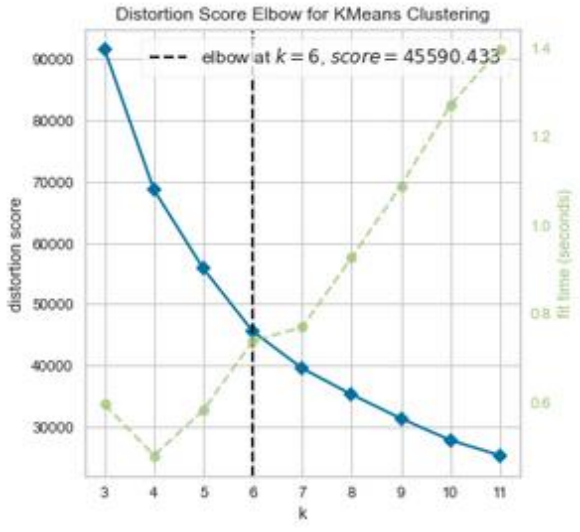


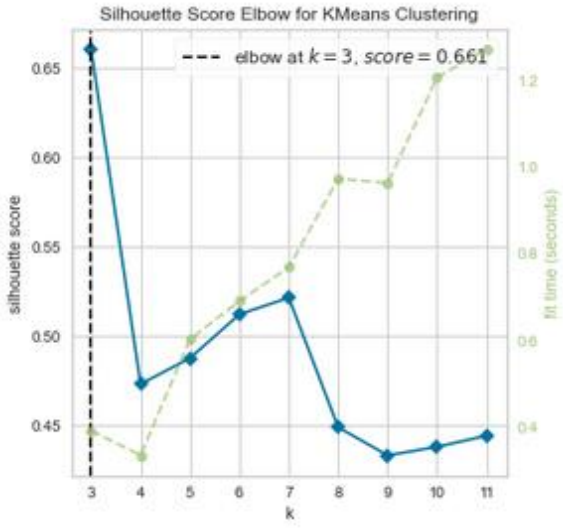
C: Modelisation
2-Features engineering:

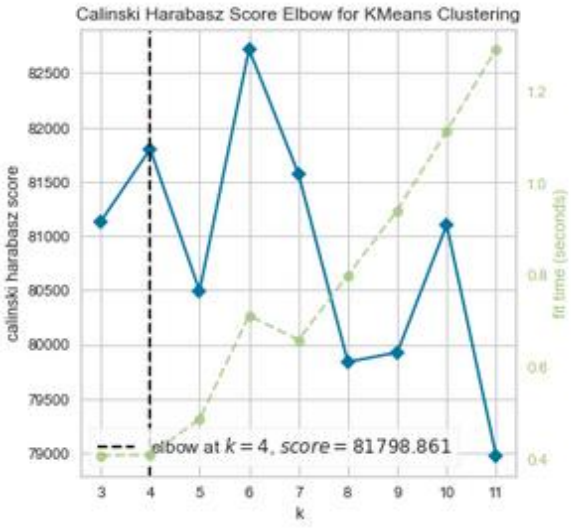
- **Jointure des tables et creation de features par aggregations successives :**
  - **'nb_com' :** *nombre de commandes par client unique*
  - **'item_nb_last_com':** *nombre d'articles de la dernière commande du client*
  - **'item_nb_total':** *nombre total d'articles commandés*
  - **'pr_sum','pr_last','pr_mean':** *CA client, prix de la derniere commande, panier moyen du client*
  - **'fdp_sum','fdp_last','fdp_mean':** *frais de port totaux du client, frais de port de la derniere commande, frais de port moyen par commande du client*
  - **'w_sum','w_last','w_mean':** *poids total des commandes du client, poids de la dernière commande, poids moyen par commande du client*
  - **'vol_sum','vol_last','vol_mean':** *volume total des commandes du client, volume de la dernière commande, volume moyen par commande du client*
  - **'pay_ins_sum','pay_ins_last','pay_ins_mean':** *nombre total de paiements du client, nombre de paiement de la dernière commande du client, nombre moyen de paiements par commande du client*
  - **'pct_voucher':** *pourcentage des commandes du client payees en voucher*
  - **'note_last','note_mean':** *dernière note laisse par le client et note Moyenne laissée par le client*
  - **'com_last','com_mean***': flag de commentaire sur la dernière commande du client, flag moyen de commentaires*
  - **'dist_zip', 'dist_state':** *distance du client par rapport a Sao Paulo : (zip et état)*
  - **'cat_w_last','cat_ca_last':** *poids et CA moyen de la catégorie principale de la dernière commande*
  - **'cat_pr_last','cat_fot_last':** *prix moyen et nombre moyen de photos de la catégorie principale de la dernière commande*
  - **'rec':** *recence*

C: Modelisation
3-Segmentation RFM - Kmeans:

- **3 features:**
  - **Recence de la dernière commande du client**
  - **nombre de commandes sur les 12 derniers mois**
  - **panier moyen du client**
- **Détermination du nombre de clusters: Technique du coude : entre 4 et 8 clusters**
- **On souhaite minimiser la distance intra cluster et augmenter la distance inter clusters en ayant un nombre limité et compréhensible de clusters**
  - **Distorsion score (intra clusters) :** somme des distances au carré entre chaque point et son centroide
  - **Silhouette score (yellowbrick) :** ratio entre la distance Moyenne intra cluster et la distance Moyenne avec le plus proche cluster
  - **Calinski score (yellowbrick)** : ratio entre la dispersion intra cluster et la dispersion inter cluster

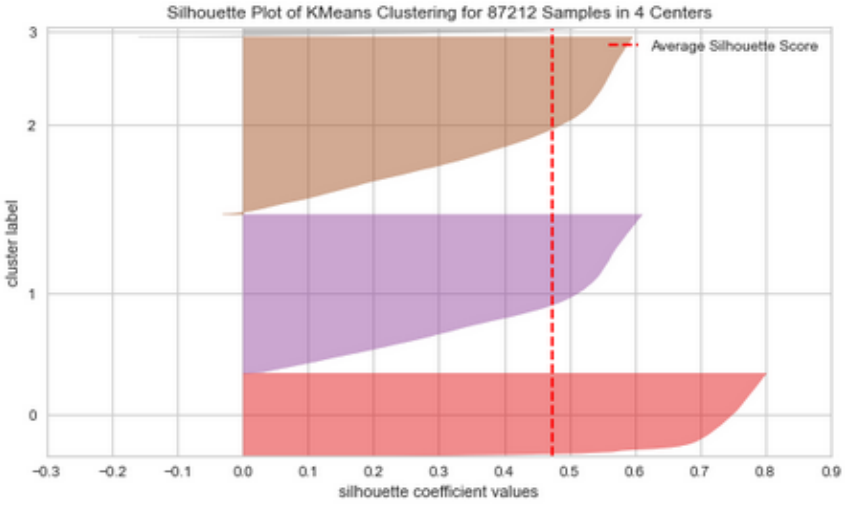


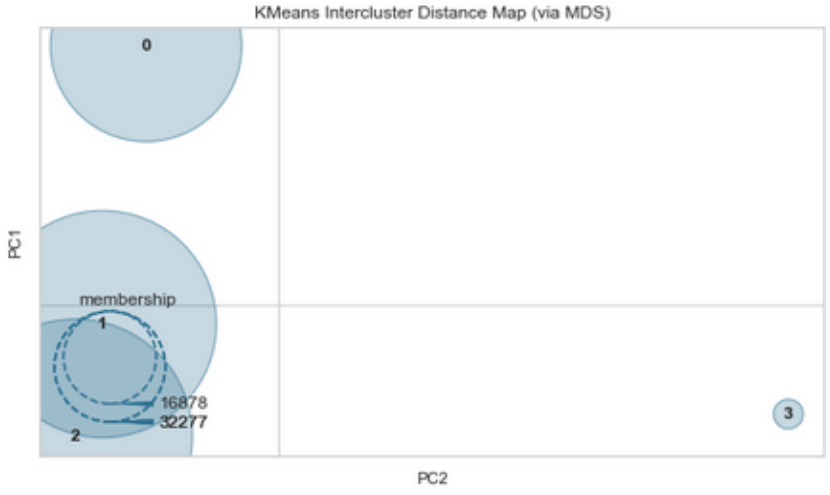




C: Modelisation
3-Segmentation RFM – Kmeans(4):

- **Cas à 4 clusters**
  - **Un score de silhouette autour de 0,5**
  - **Les clusters ont tous un score qui dépasse la moyenne**
  - **Un des clusters est assez petit (2,2% des clients) mais cela reste acceptable**
  - **Prendre un K plus important rajoute des clusters de trés petites tailles**

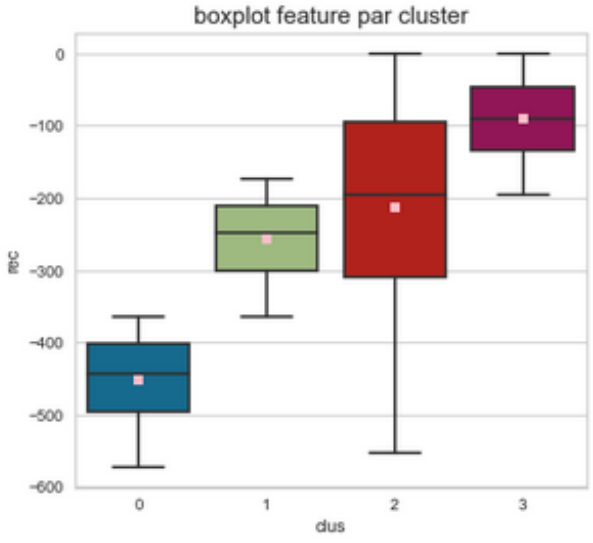


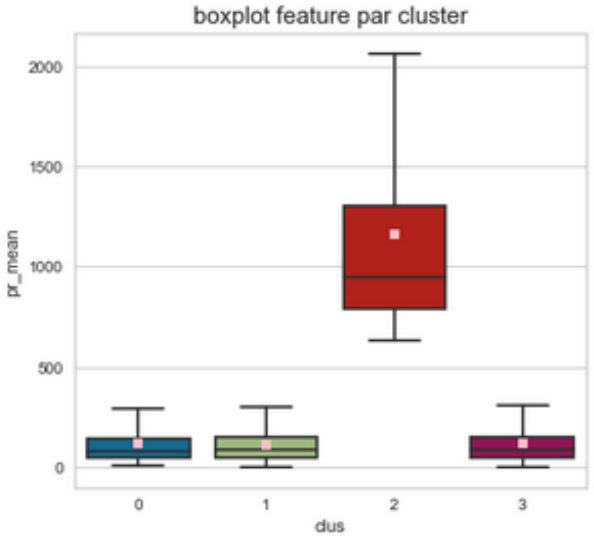


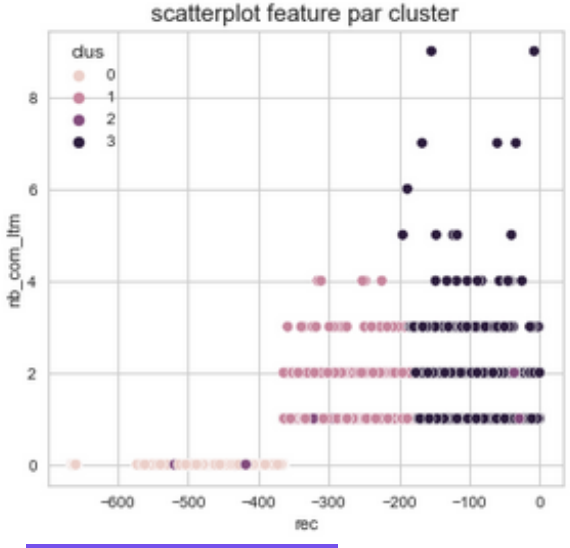
**les labels se rapportent à la page suivante et non aux graphes ci dessus*

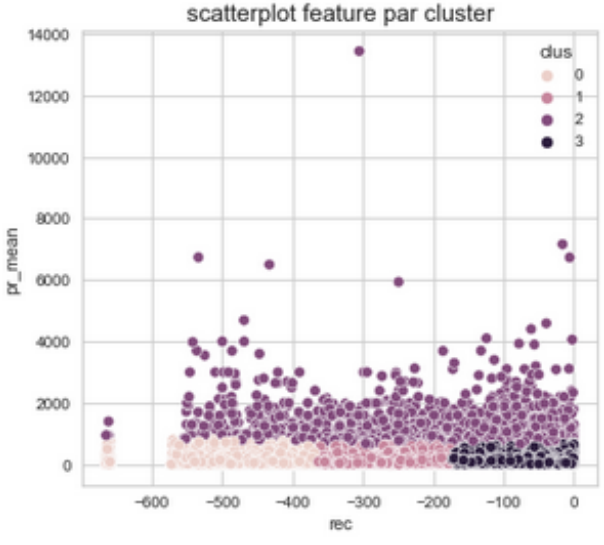
- **Recence**
  - **Le cluster 0 représente des clients dont les commandes sont assez vieilles**
  - **Le cluster 3 représente des clients dont les commandes sont relativement récentes**
- **Frequence**
  - **Le cluster 0 représente des clients sans commande dans les 12 derniers mois**
- **Panier moyen**
  - **Le cluster 2 représente des clients avec un panier moyen assez elevé**

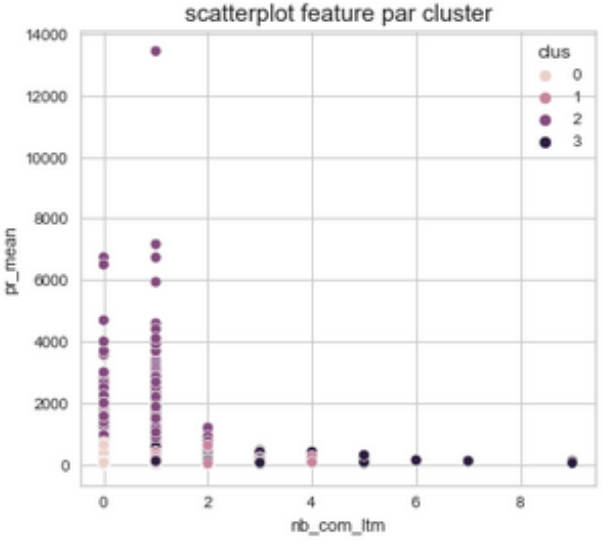
C: Modelisation
3-Segmentation RFM - Kmeans(4):



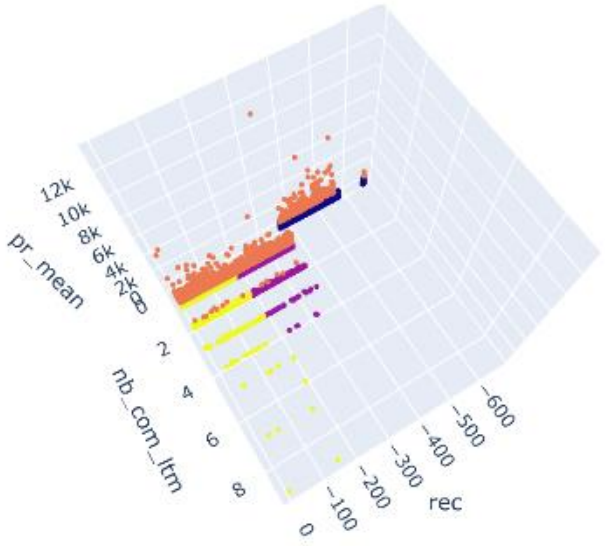


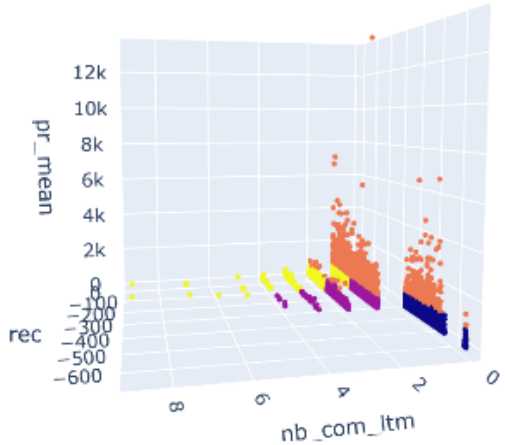






C: Modelisation
3-Segmentation RFM - Kmeans(4):

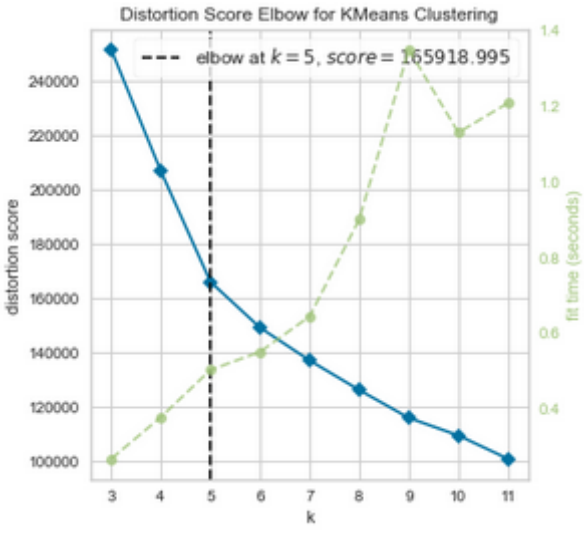


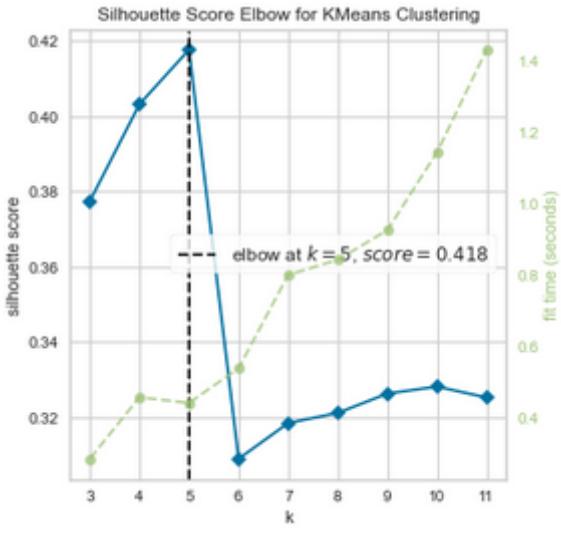


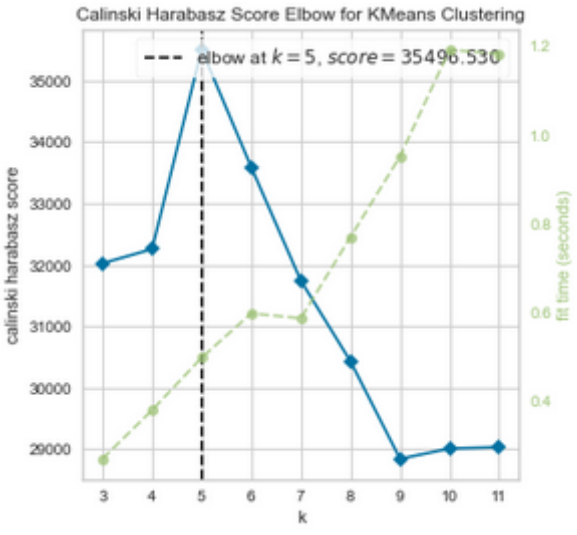
- **Cas à 4 clusters**
  - **La récence segmente en 3**
  - **Le panier moyen segmente en 2**

C: Modelisation
4-Segmentation RFM+distance+note - Kmeans:

- **5 features:**
  - **Recence de la dernière commande du client**
  - **nombre de commandes sur les 12 derniers mois**
  - **panier moyen du client**
  - **Distance entre le client et sao Paulo**
  - **Note du client**
- **Détermination du nombre de clusters: Technique du coude : entre 5 et 6 clusters**

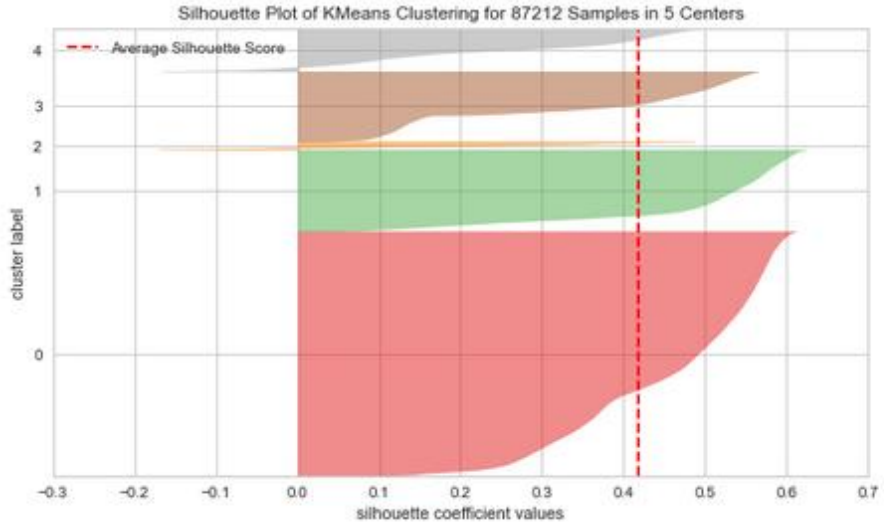


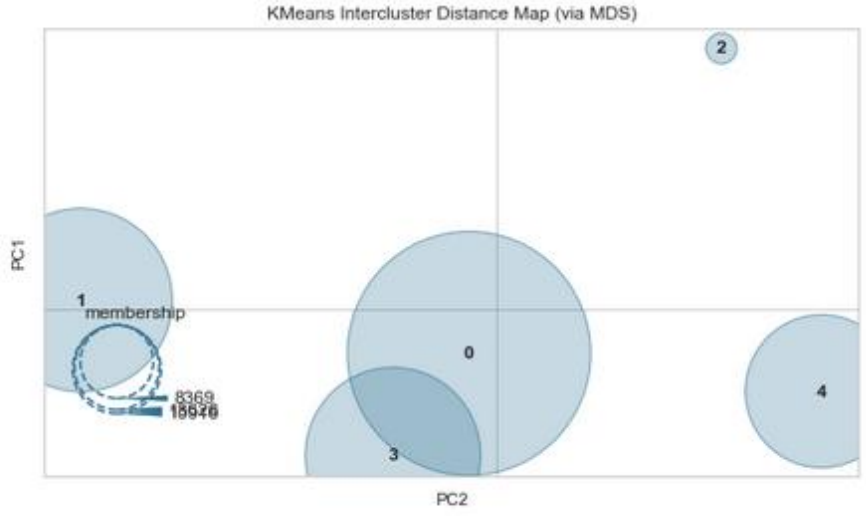




C: Modelisation
4-Segmentation RFM+distance+note – Kmeans(5):

- **Cas à 5 clusters**
  - **Un score de silhouette au dessus de 0,4**
  - **Les clusters ont tous un score qui dépasse la Moyenne du score de silhouette**
  - **Un des clusters est assez petit (1,9% des clients) mais cela reste acceptable**
  - **Prendre un K plus important rajoute des clusters de trés petites tailles ou diminue le score**

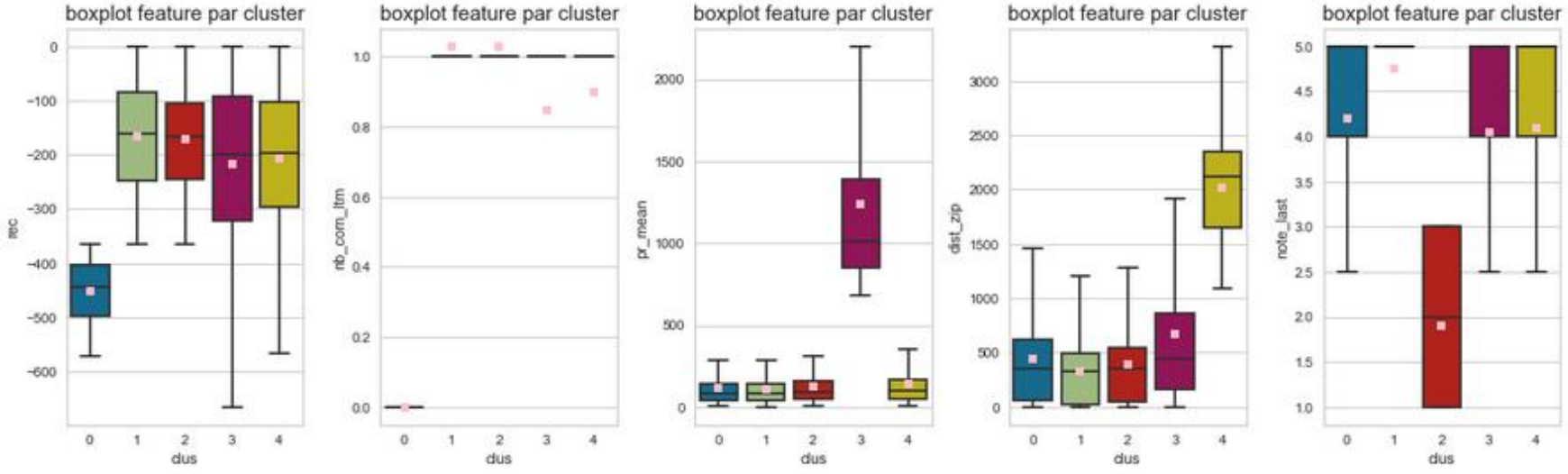


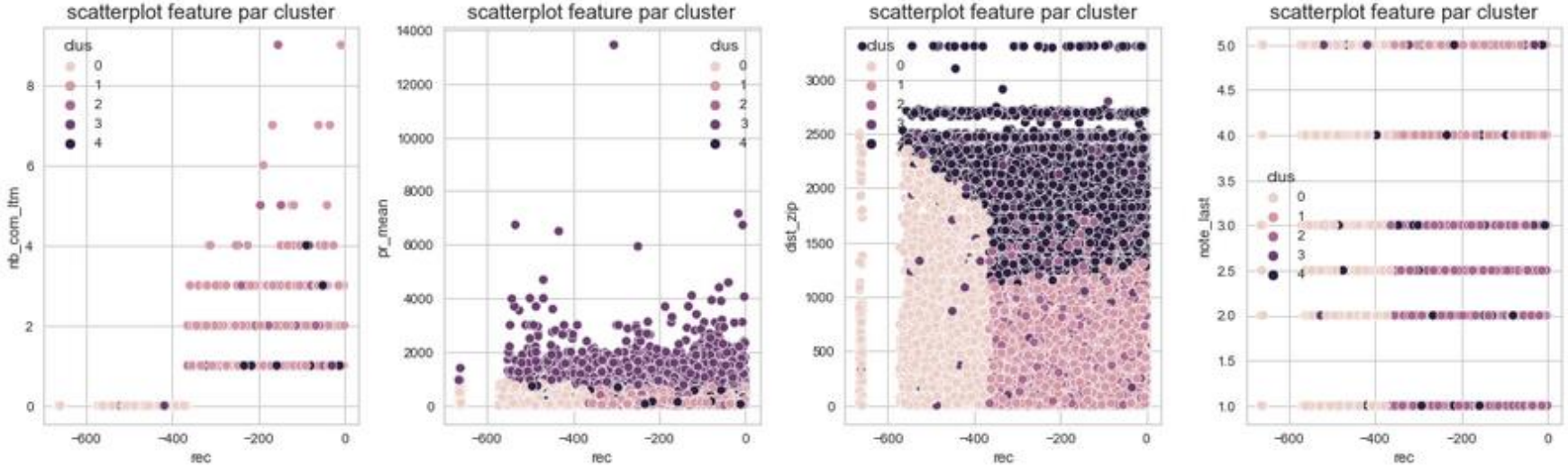


les labels se rapportent à la page suivante et non aux graphes ci dessus

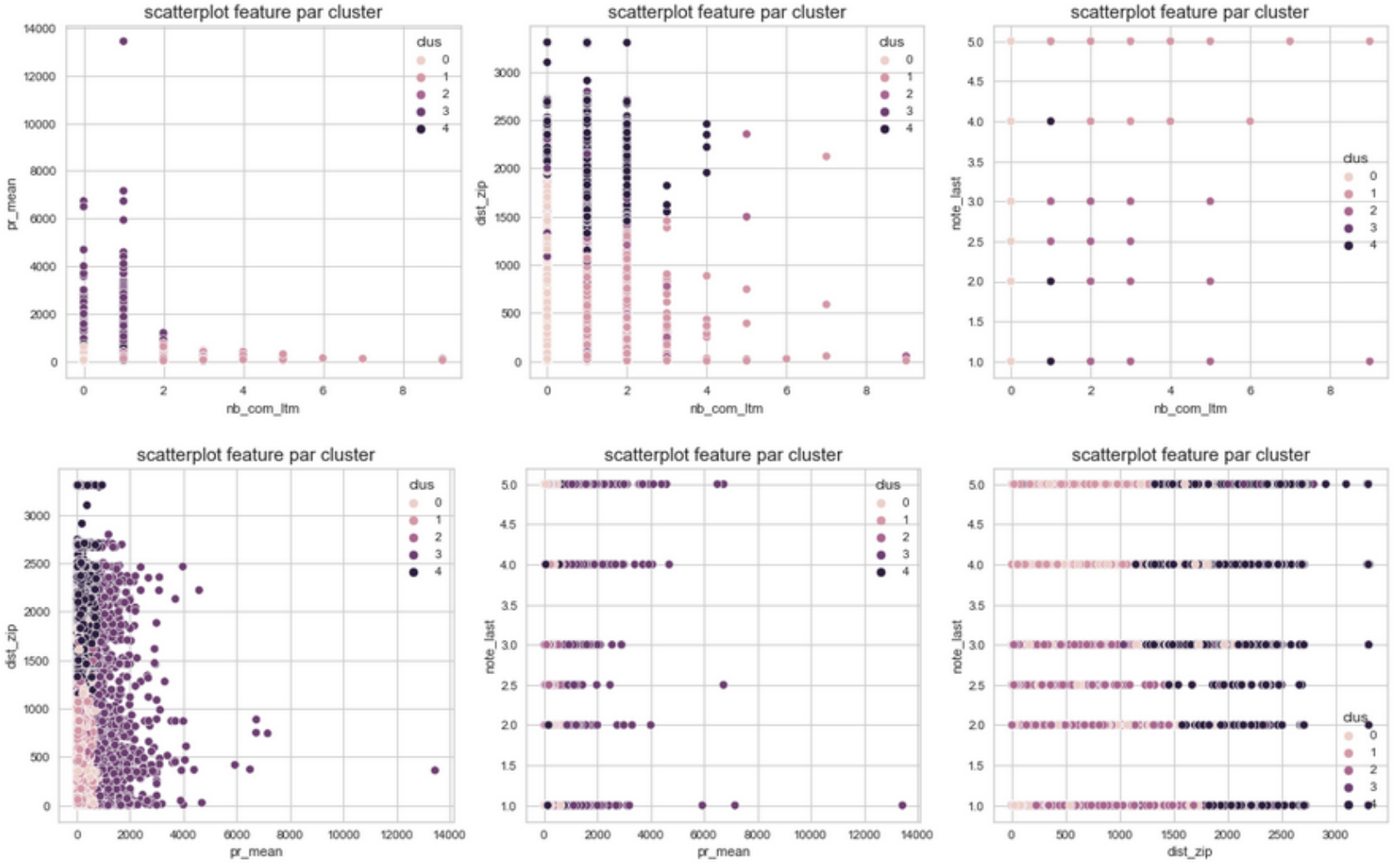
- **Recence**
  - **Le cluster 0 représente des clients dont les commandes sont assez vieilles**
- **Frequence**
  - **Le cluster 0 représente des clients sans commande dans les 12 derniers mois**
- **Panier moyen**
  - **Le cluster 3 représente des clients avec un panier moyer assez elevé**
- **distance**
  - **Le cluster 4 représente des clients eloignés de sao Paulo**
- **note**
  - **Le cluster 2 représente des clients mécontents**

C: Modelisation
4-Segmentation RFM+distance+note – Kmeans(5):

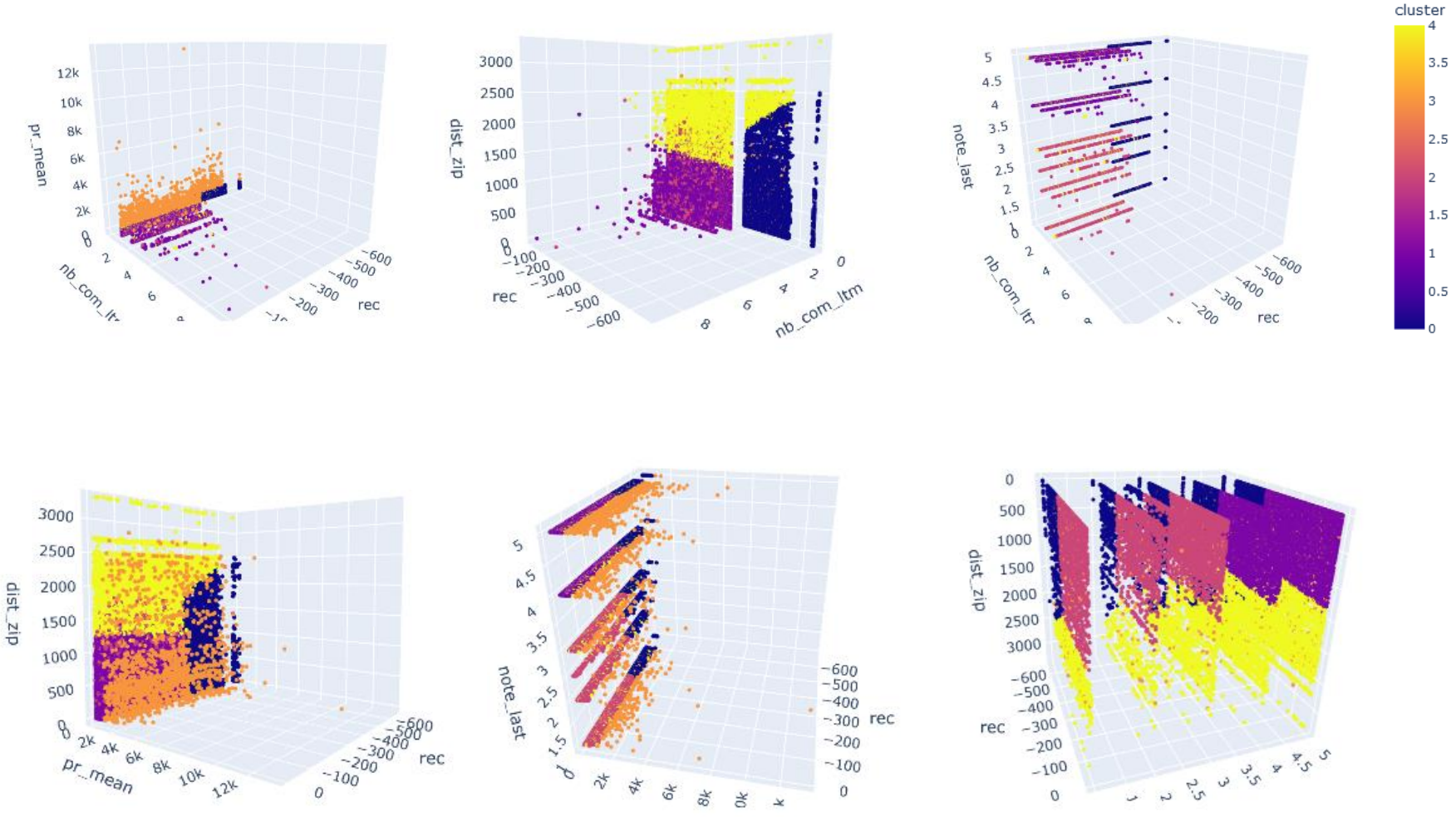




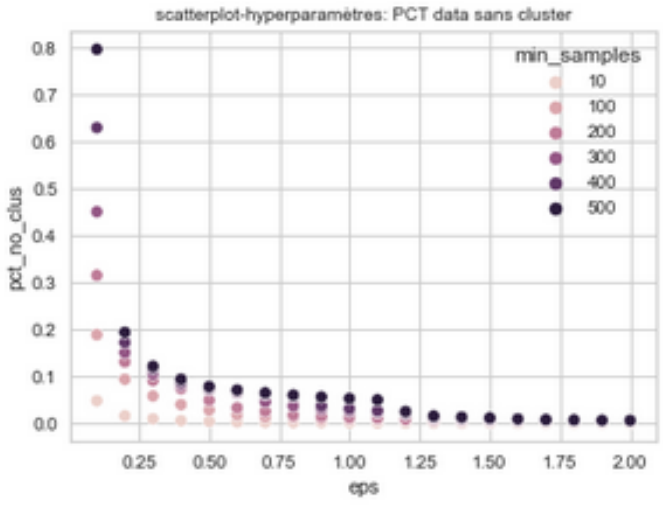
C: Modelisation
4-Segmentation RFM+distance+note – Kmeans(5):

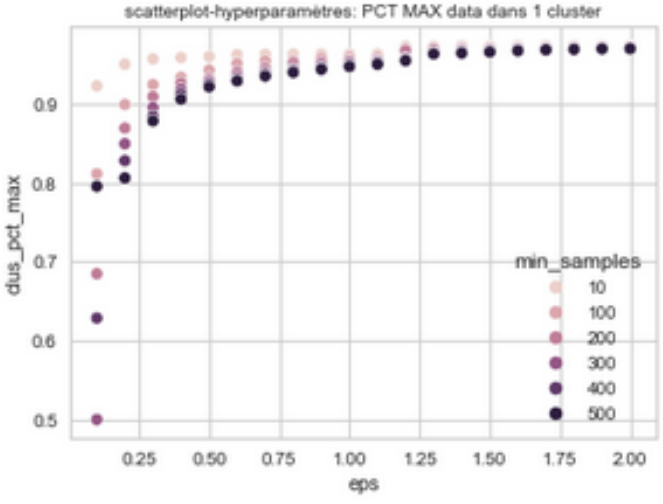


C: Modelisation
4-Segmentation RFM+distance+note – Kmeans(5):



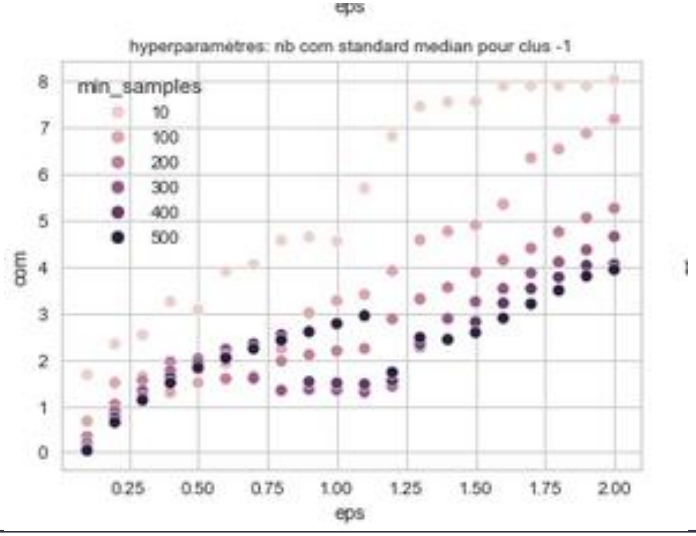
C: Modelisation
5-Segmentation RFM - DBSCAN





- **Contraintes à notre cas**
  - Choix du nombre de clusters dépendant des hyperparamètres
  - **Augmenter L'eps –** controle de la taille du voisinage (plus tolérent à une densité plus faible) **converge vers un cluster de forte taille** (entre 90 et 99% des clients), Ce phénomène est amplifié par un faible MIN SAMPLES – tolerance au bruit
  - **Diminuer L'eps** (une densité plus forte est necessaire pour former un cluster) **augmente le nombre de clients non affecté à un cluster**

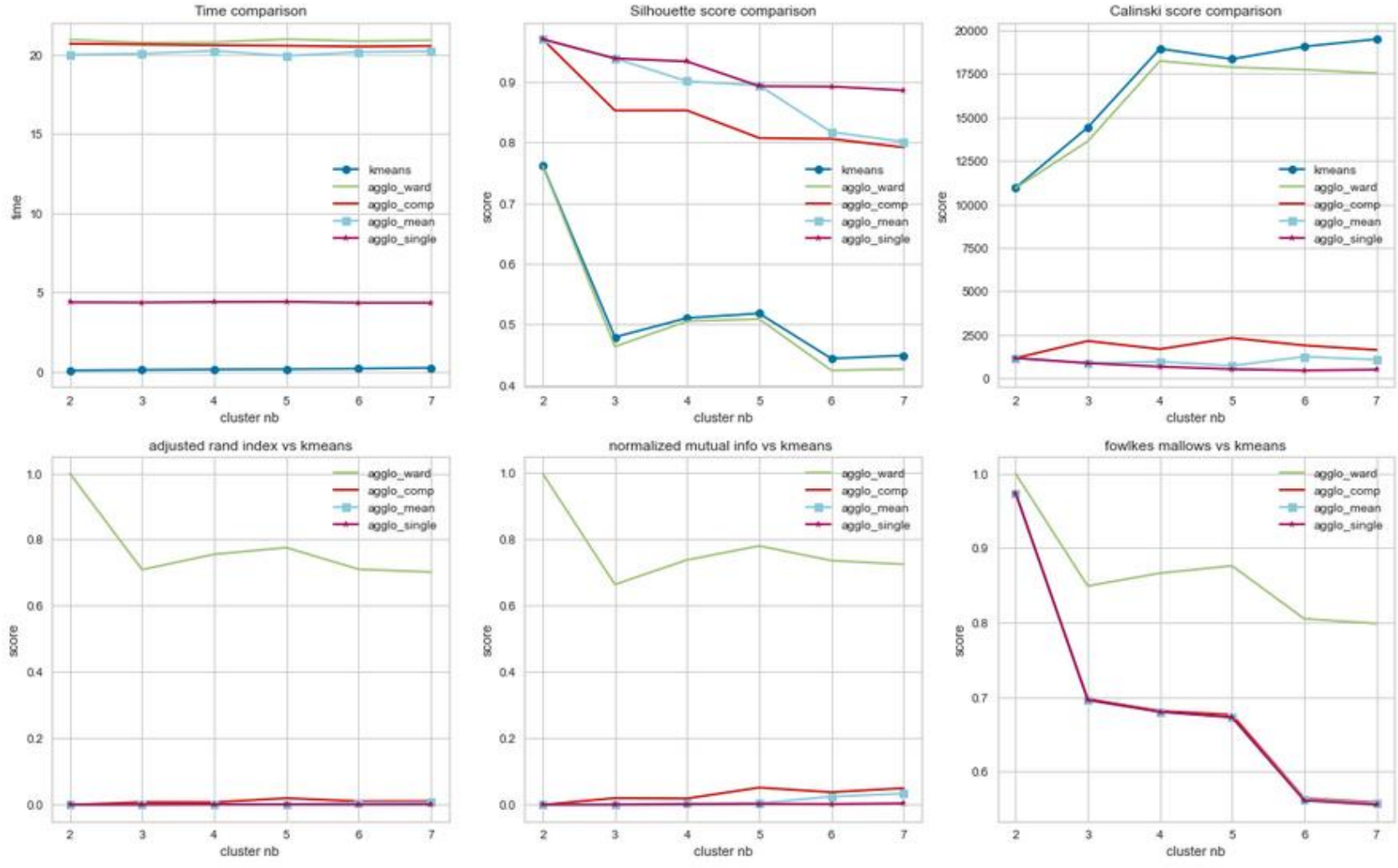
- **Le DBSCAN n'est donc pas adapté à nos données. Il nécéssite des clusters bien denses séparés par des zones vides ou peu denses**
  - Les clients **non affectés** à des clusters sont en partie **des clients à plusieurs commandes**



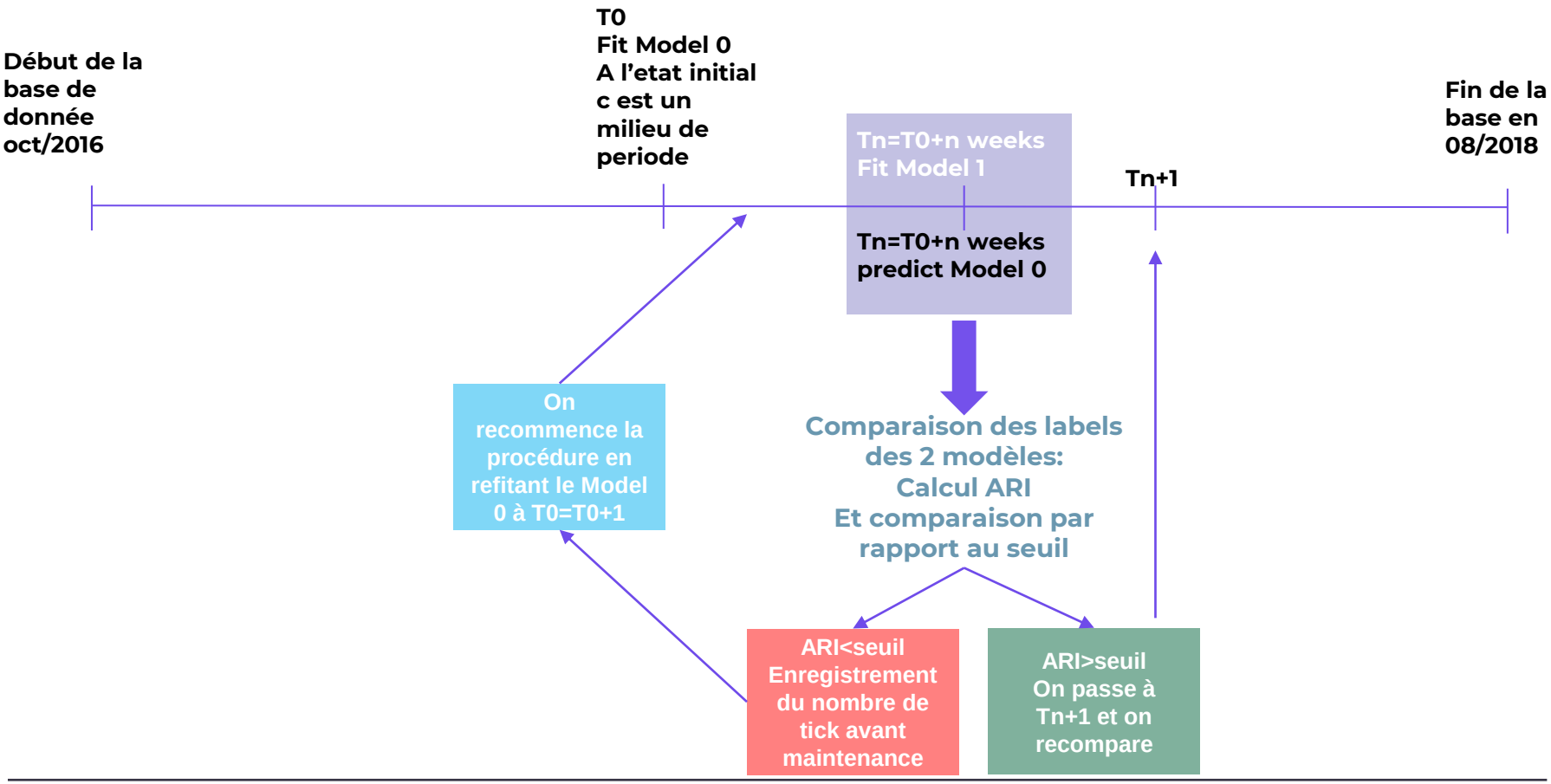
C: Modelisation
6-Segmentation RFM – KMEANS vs HIERARCHIQUE

- **Comparaison:**
  - **Temps de calcul**: avantage KMEANS
  - **Score de silhouette:** L'algorithme de ward est trés proche des KMEANS
    - Les metriques 'complete', 'mean' et 'single' donnent des scores plus faibles
  - Même constat avec le score de Calinski
  - **Comparaison par calcul de l'adjusted rand index:**
    - L'algorithme de ward est trés proche des KMEANS au contraire des autres métriques qui donnent des résultats trés different
- **Conclusion:**
  - **Avec des distances euclidiennes , mieux vaut utiliser les kmeans et ensuite Ward si besoin**
  - **Si d'autres métriques sont nécéssaires , la richesse du clustering hierarchique reste une bonne alternative**

C: Modelisation
6-Segmentation RFM – KMEANS vs HIERARCHIQUE



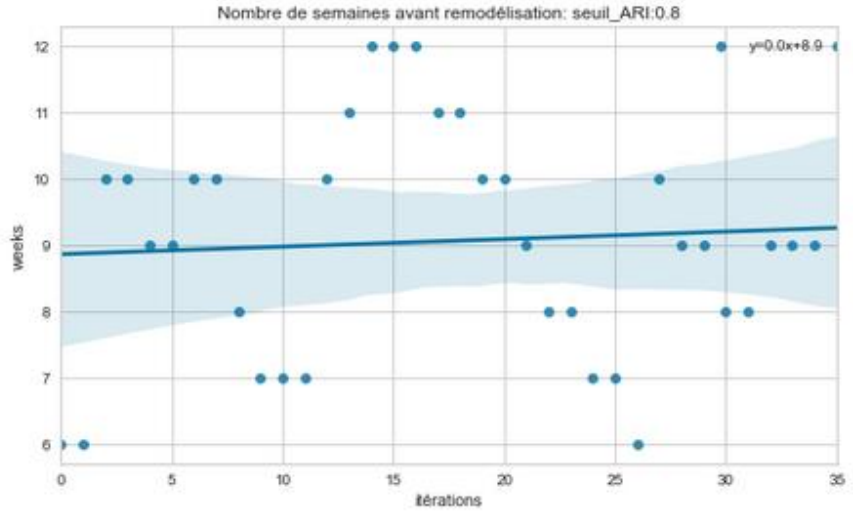
C: Modelisation
7-Segmentation RFM – Stabilité:
Présentation de l'algo de recherche de divergence du kmeans de segmentation

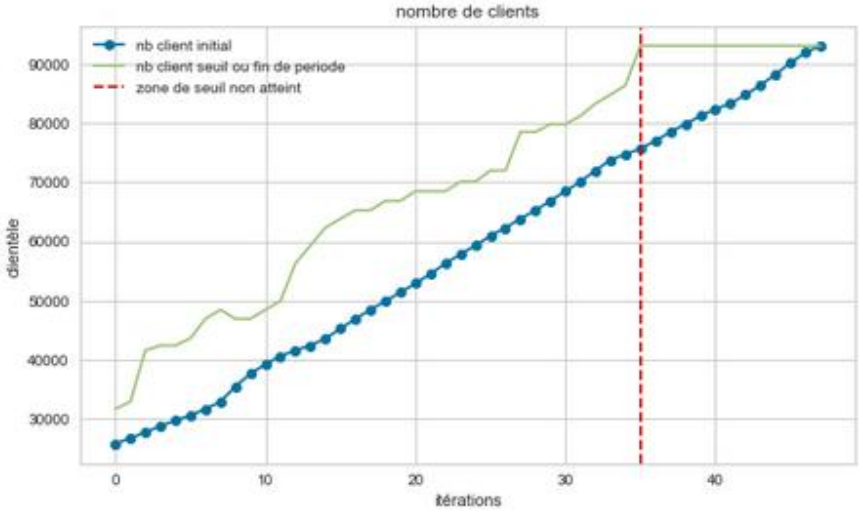


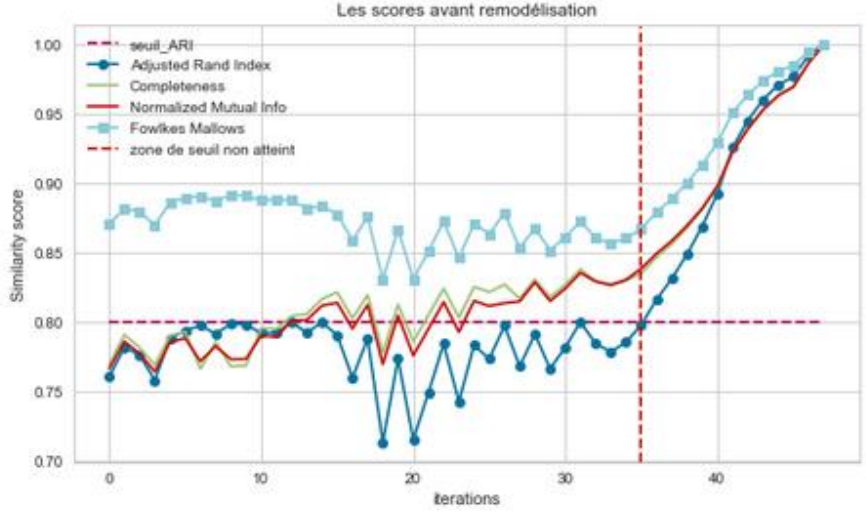
C: Modelisation

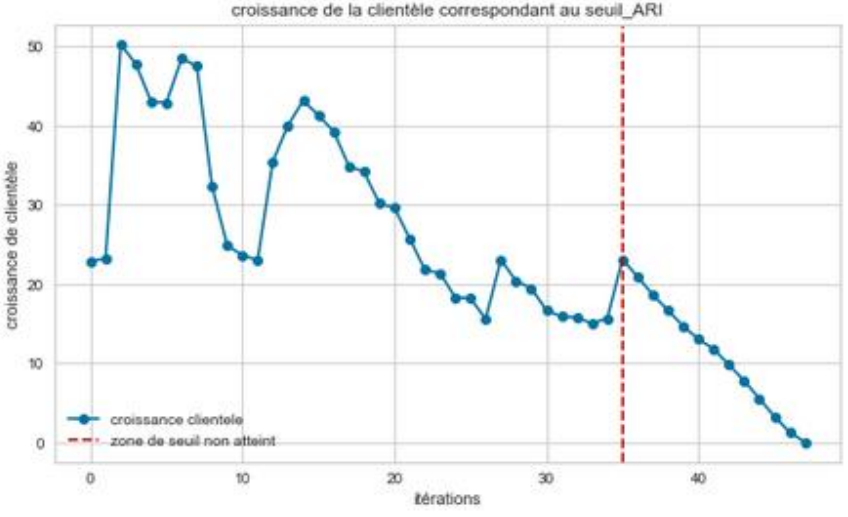
7-Segmentation RFM – Stabilité:

Seuil ARI fixé à 0,8: Maintenance au bout de 9 semaines ou d'une hausse minimale de clientèle de 20%









OPENCLASSROOMS