Olist

Segmentation des clients pour notre site de e-commerce

Juin 2024 - Lokman AALIOUI

Feuille de route

01 Contexte

02 Nettoyage, analyse et feature engineering

03 Modélisation

04 Maintenance

Feuille de route

01 Contexte

Qui sommes nous ?

- Entreprise brésilienne spécialisée dans la vente sur les marketplaces en ligne.
- Solution intégrée permettant aux vendeurs de répertorier leurs produits sur plusieurs plateformes de vente.

Notre but ?

- Faciliter l'accès aux marketplaces pour les vendeurs de toutes tailles.
- Optimiser la visibilité et les ventes des produits de leurs clients.

Mission

Réaliser une segmentation client

Comment ?

- Collecter et analyser les données disponibles sur les clients
- Appliquer un algorithme de clustering non supervisés pour regrouper les clients en segments distincts

Pourquoi ?

- Identifier des groupes de clients selon leurs habitudes de consommation
- Personnaliser les Campagnes Marketing
- Cibler les segments les plus rentables
- Intervenir sur les segments à risque de churn
- ...

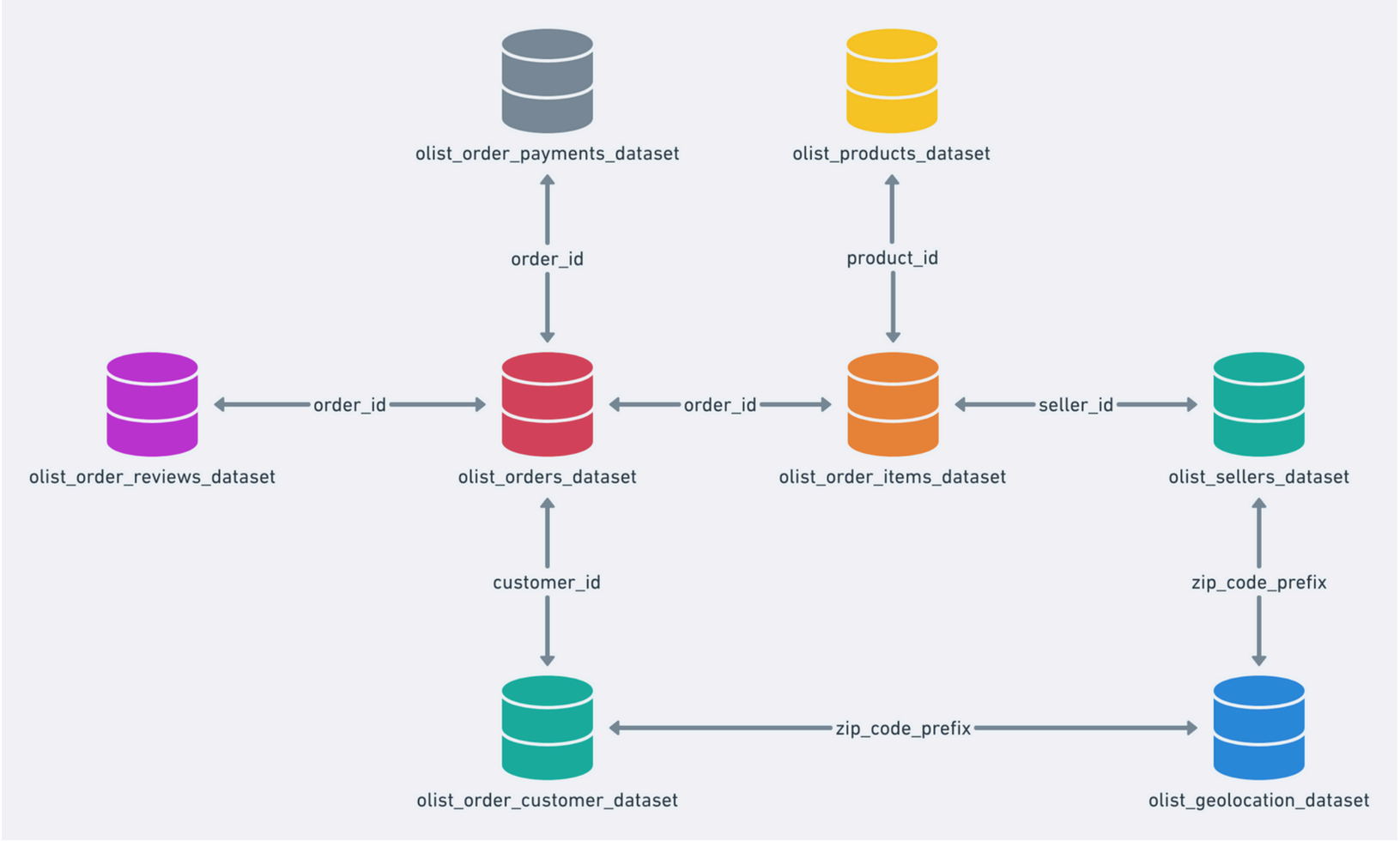
Les données

Structure

- 8 fichiers de données
- Liés par plusieurs clé différentes
- Divisées en 3 types : données clients, données commandes, données produits

Requêtes

- Nous requêtons la BDD via SQL
- Puis nous réalisons la segmentation sur un notebook python



Feuille de route

02

Nettoyage, analyse et feature engineering

Nettoyage des données

Observation des fichiers

- 40 variables
- 100 000 individus

Traitement des valeurs manquantes

- 2.08% : order_delivered_customer_date
- 1.03% : order_delivered_carrier_date
- 0.27% : geolocation_zip_code_prefix / geolocation_lat / geolocation_lng

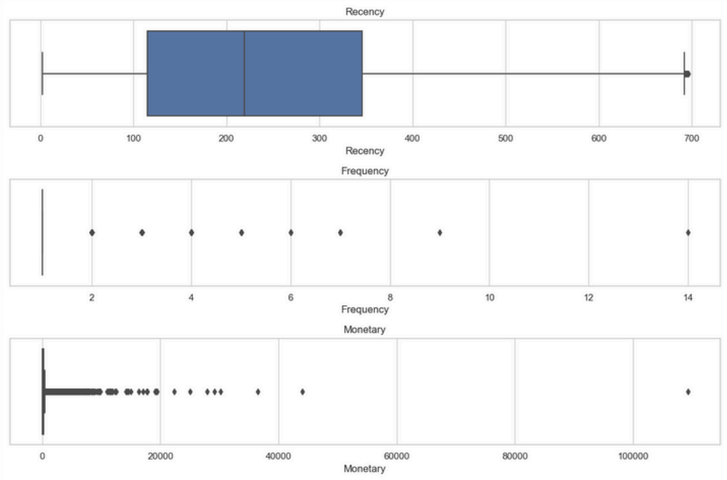
Traitement des valeurs aberrantes

- Les valeurs aberrantes sont justifiées (montant d'achat élevé par exemple)

Feature engineering

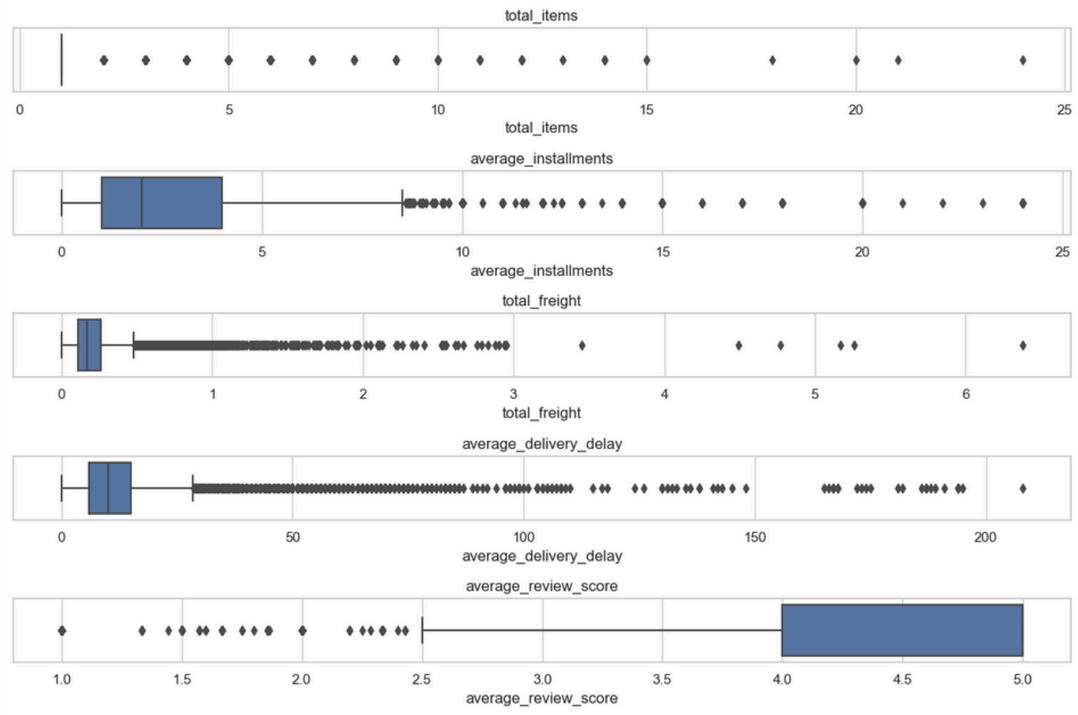
Variables RFM

- Recency
- Frequency
- Monetary



Variables fidélité

- Nombre de produit par client
- Nombre de paiement par client
- Note moyenne de satisfaction



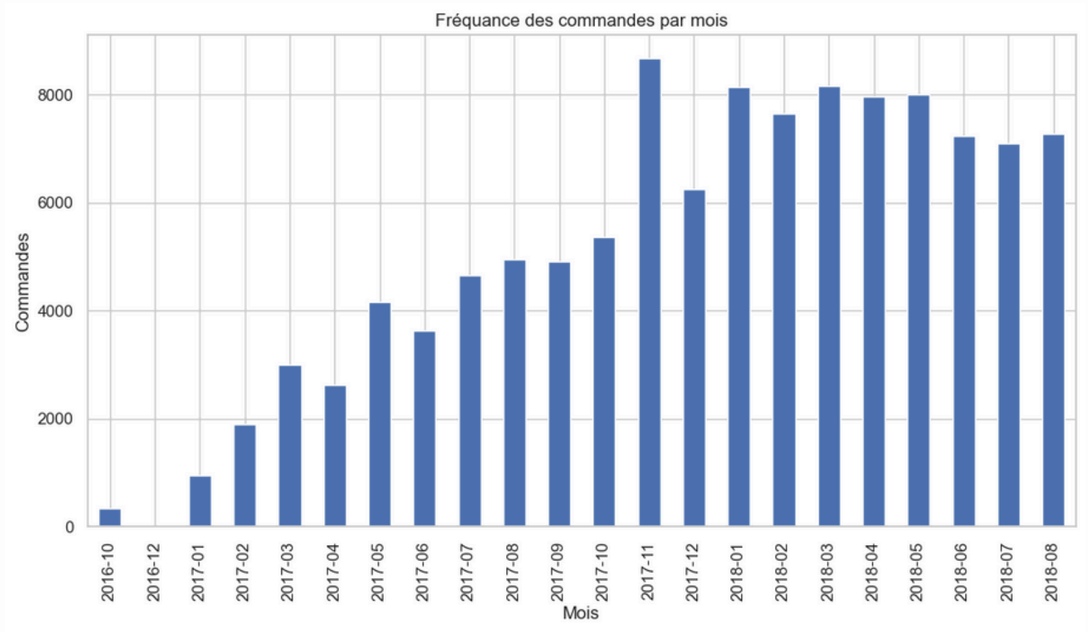
Variables livraison

- Rapport cout d'expedition / cout total
- Délai de livraison moyen

Analyse exploratoire

Analyse des commandes

- Les commandes augmentent avec un pic avant noël

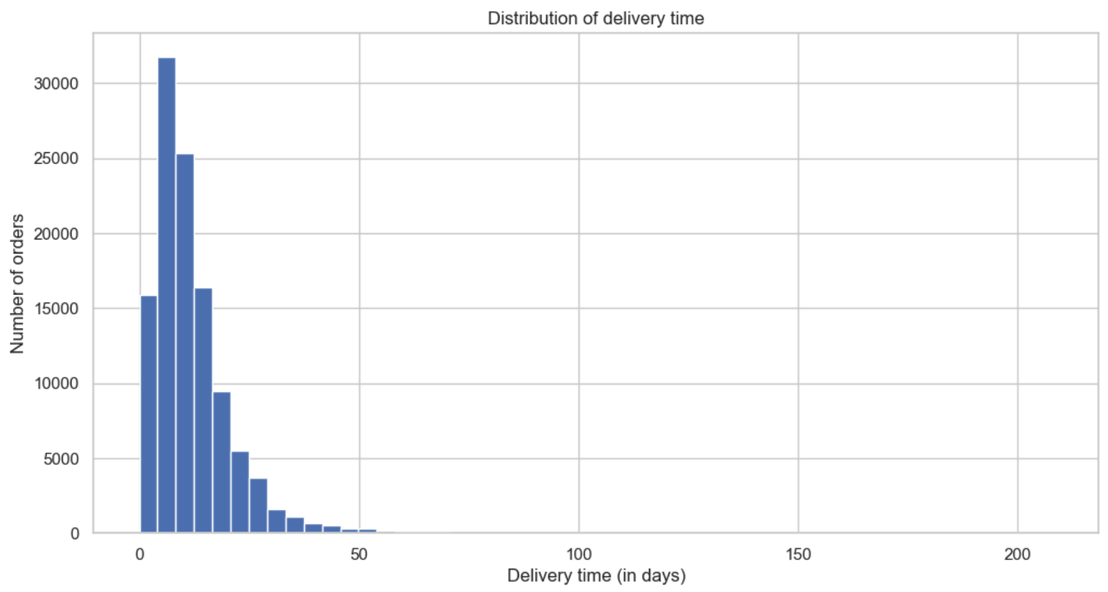


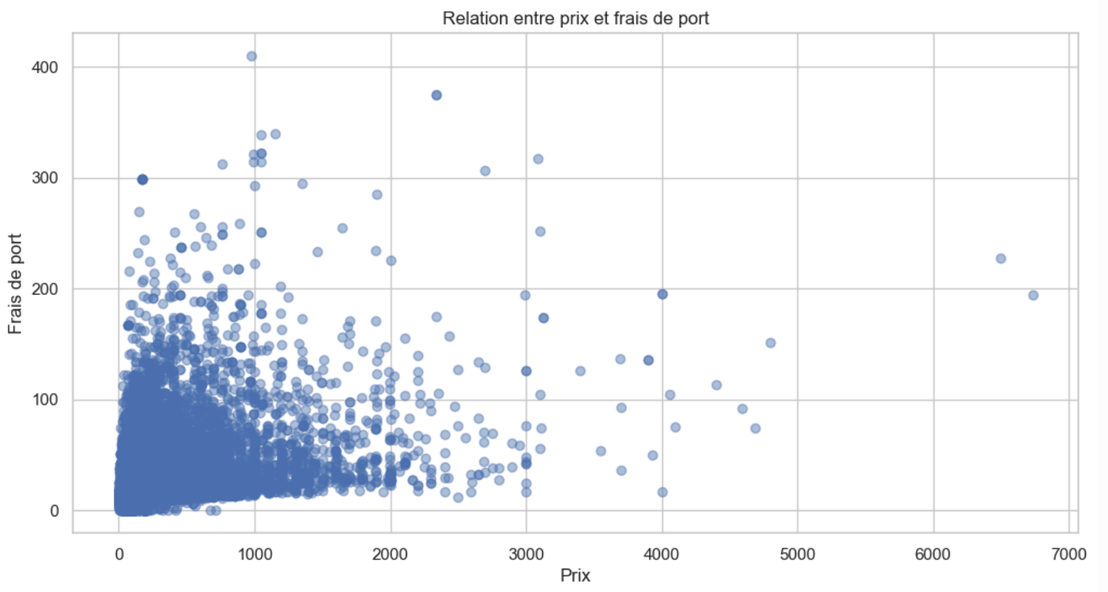
- La majorité des commande ne contiennent qu'un produit

Analyse exploratoire

Analyse des livraisons

- La majorité des commandes sont livrées en moins de 20 jours
- Tendance où les FDP augmentent légèrement avec le prix

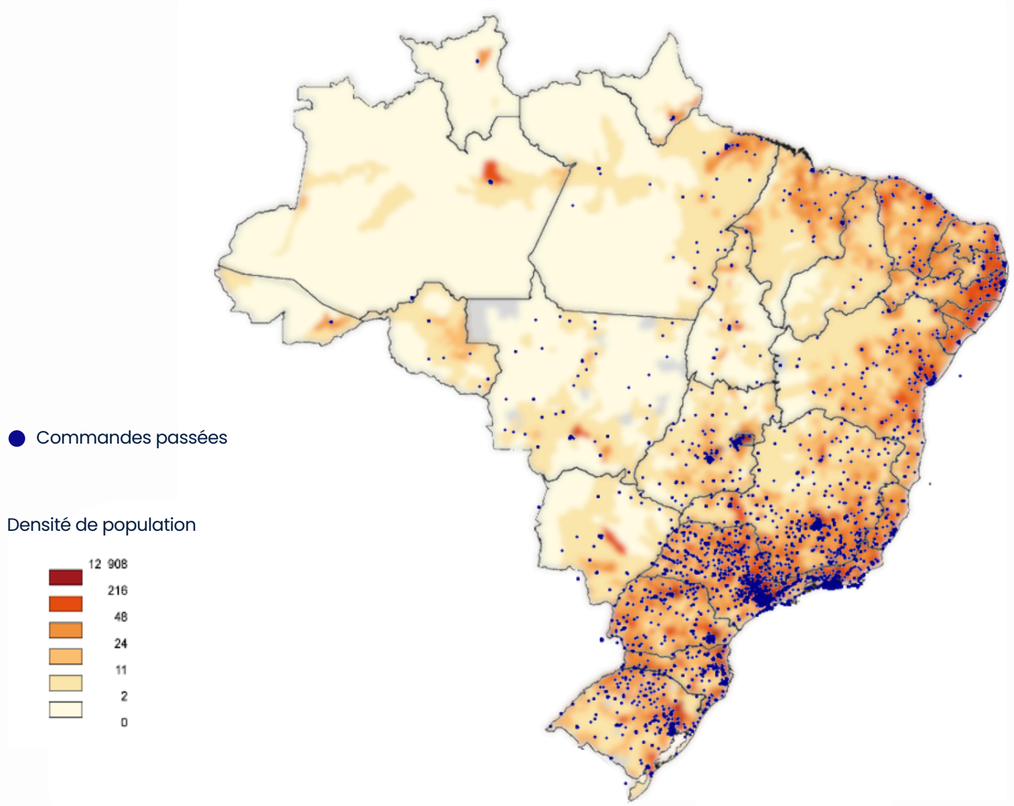




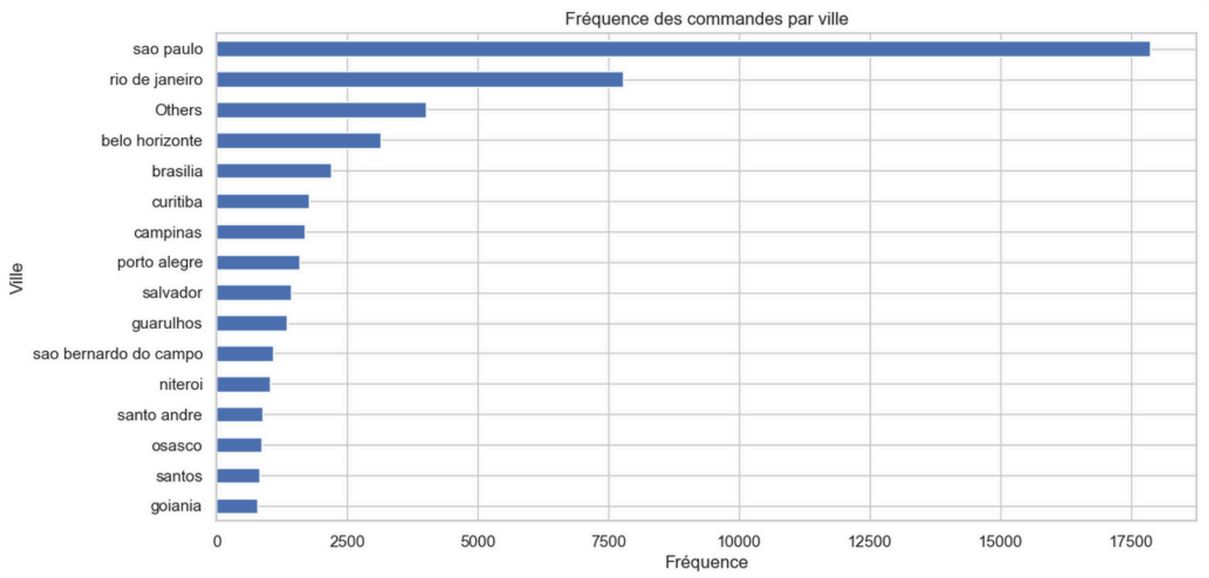
Analyse exploratoire

Analyse des livraisons

- Les commandes se font dans les zones les plus denses



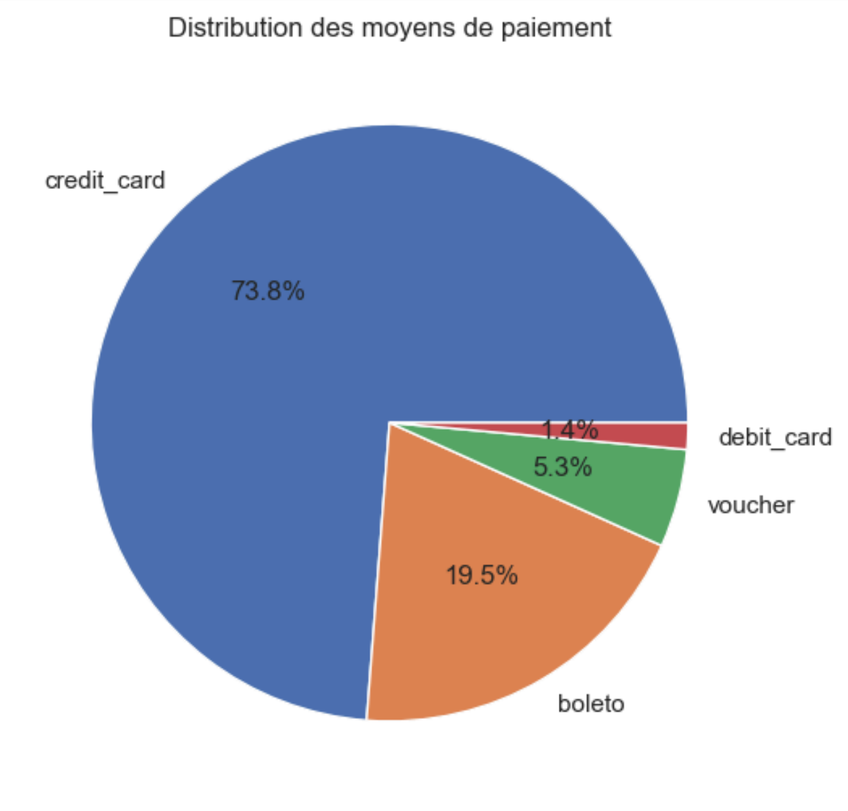
- Sao paulo en tête



Analyse exploratoire

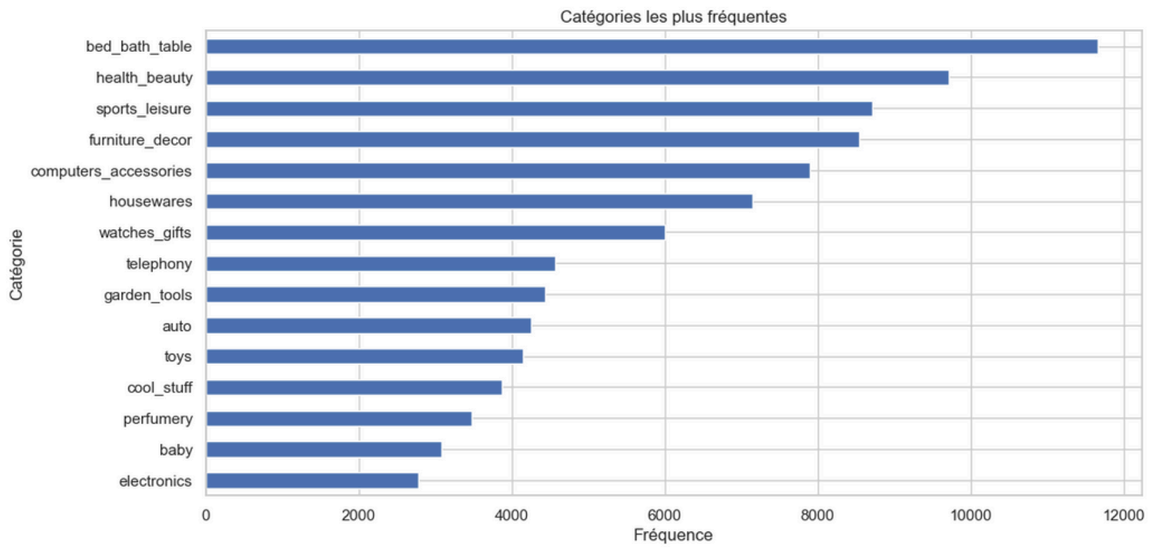
Analyse des paiements

- La CB est de loin le moyen de paiment le plus utilisé



Analyse des catégories

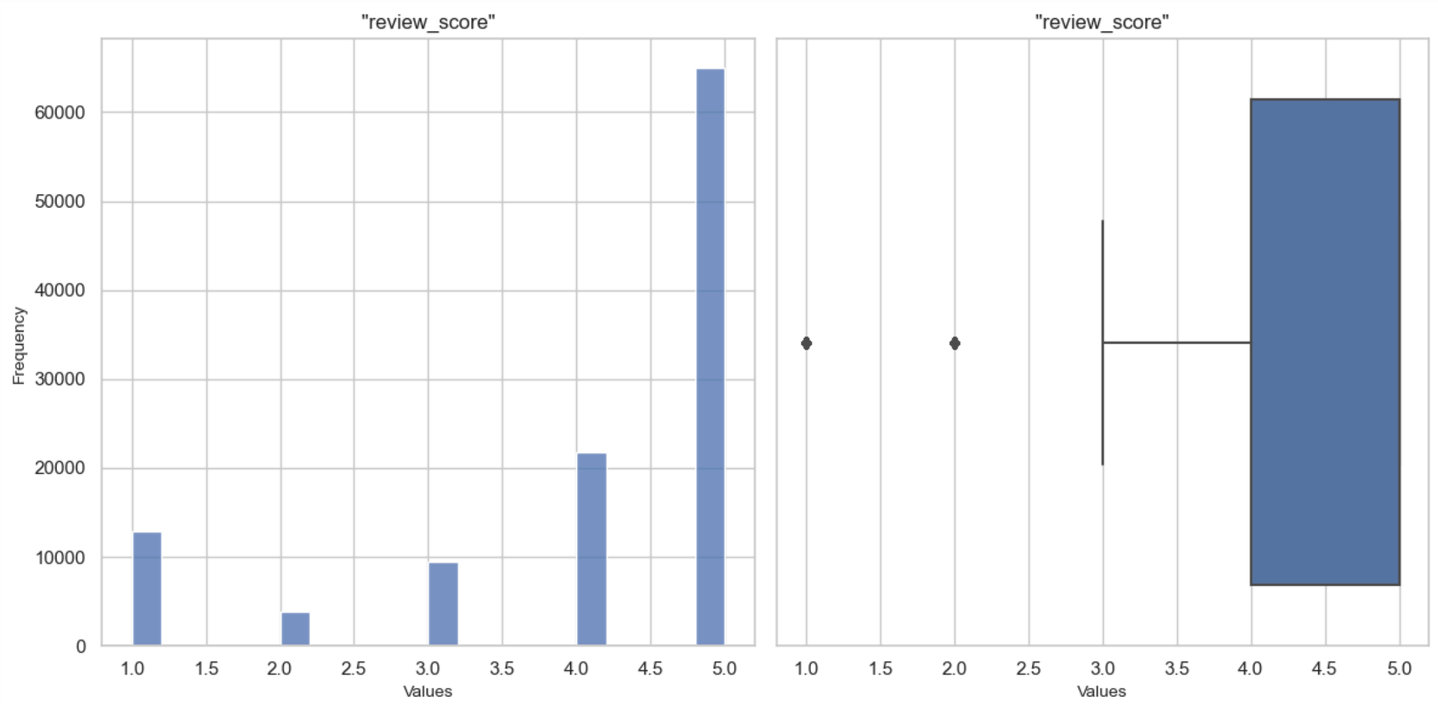
- Les produits les plus fréquemment achetées sont les produits pour la maison, de beauté et de sport



Analyse exploratoire

Analyse de la satisfaction

- La majorité des avis sont très positifs
- Néanmoins les scores de 1 sont plus fréquents que 2 et 3, suggérant que les clients insatisfaits sont plus enclins à laisser des avis



Analyse exploratoire

Analyse des corrélations

- Prix et Valeur de Paiement

Les produits plus chers entraînent des paiements plus élevés

- Prix et Frais de Port

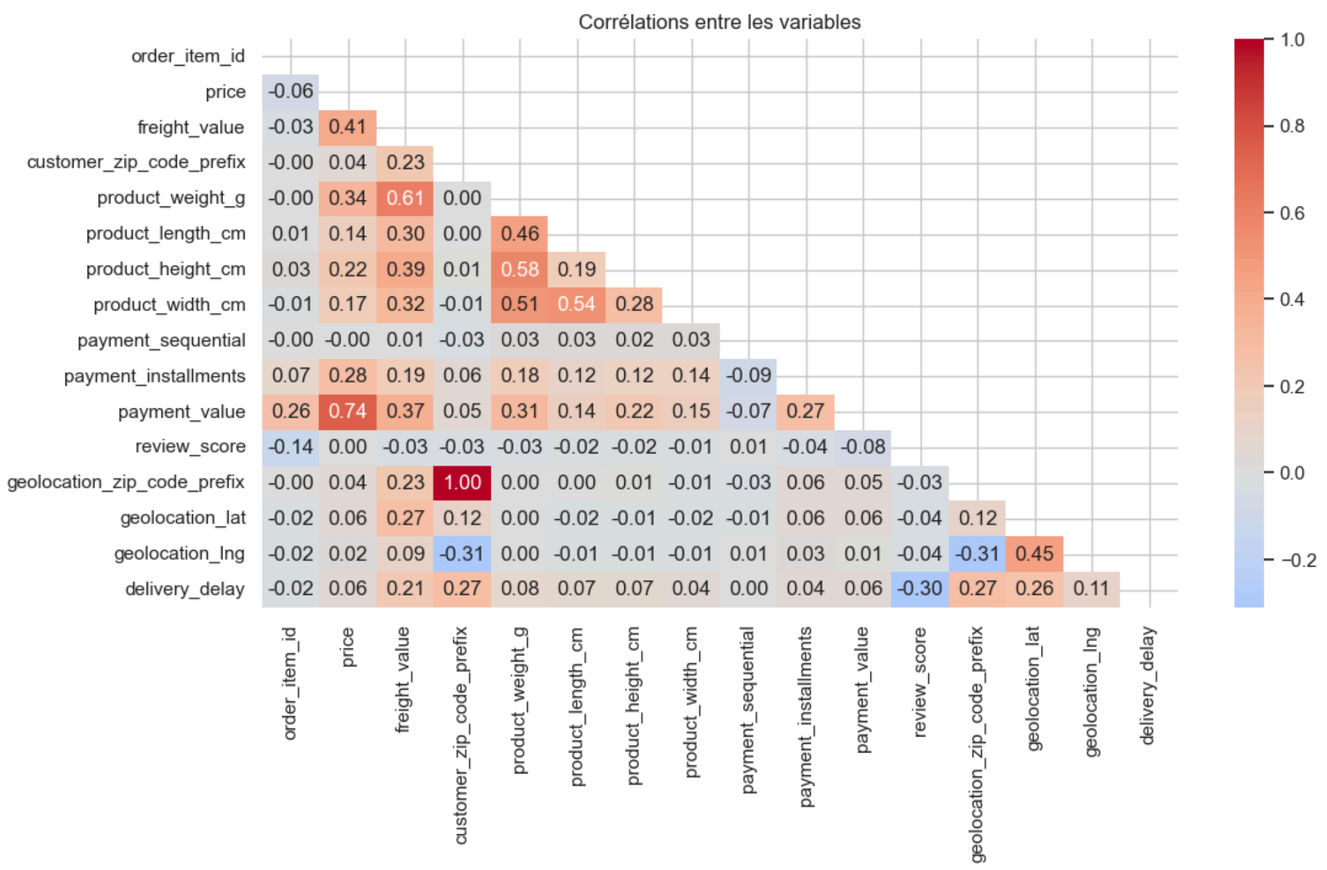
Les produits plus chers ont tendance à avoir des frais de port plus élevés.

- Poids, Dimensions, frais de port

Corrélations élevées

- Délai de Livraison et Coordonnées Géographiques

Corrélation négative, peut indiquer des différences régionales dans les délais de livraison



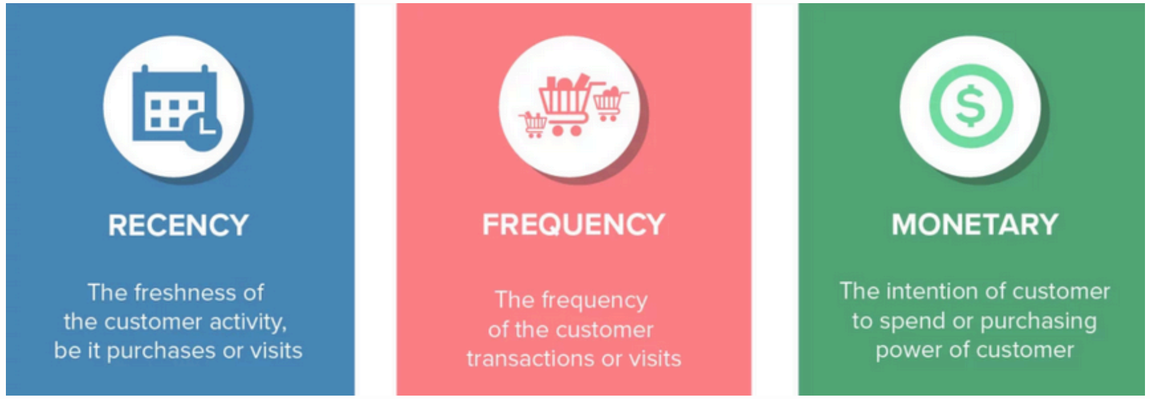
Feuille de route

03 Modélisation

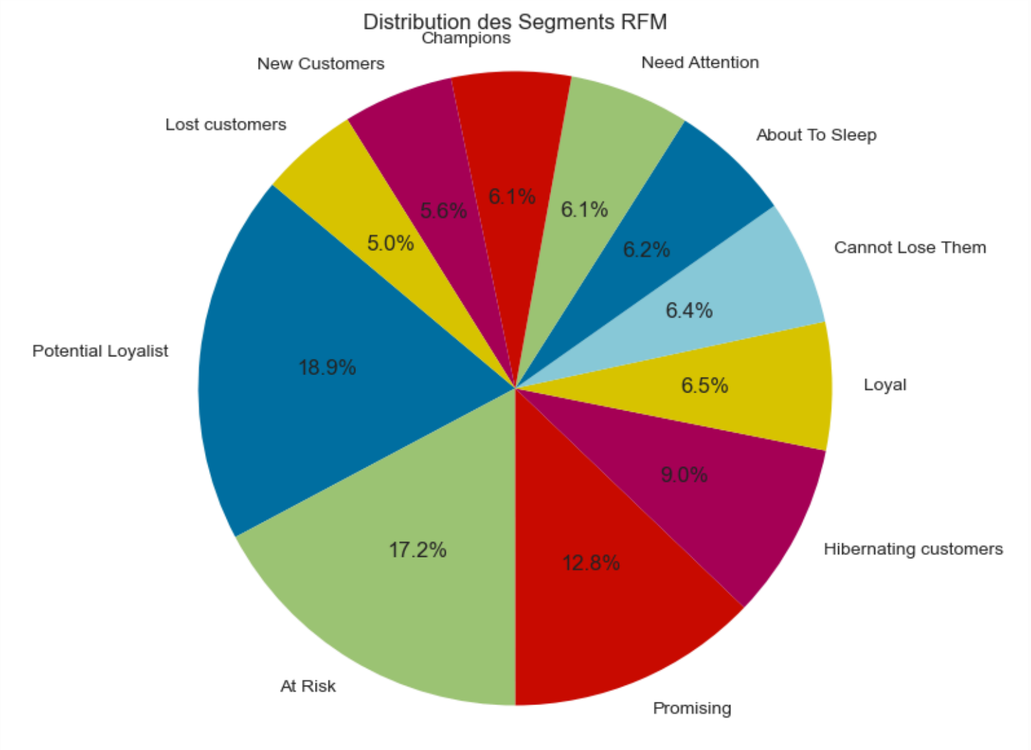
Essai de modélisation

Segmentation RFM

- La segmentation RFM (Récence, Fréquence, Montant) classe les clients en fonction de leur dernière interaction, la fréquence de leurs achats et le montant dépensé



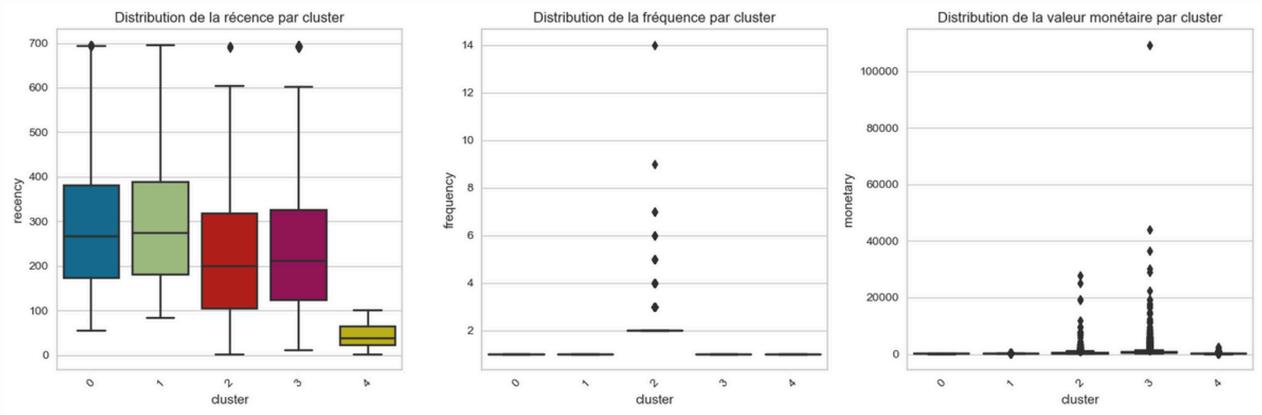
- Nos clients sont divisés en 11 groupes qui demanderont une attention differentes en terme de marketing et de gestion client



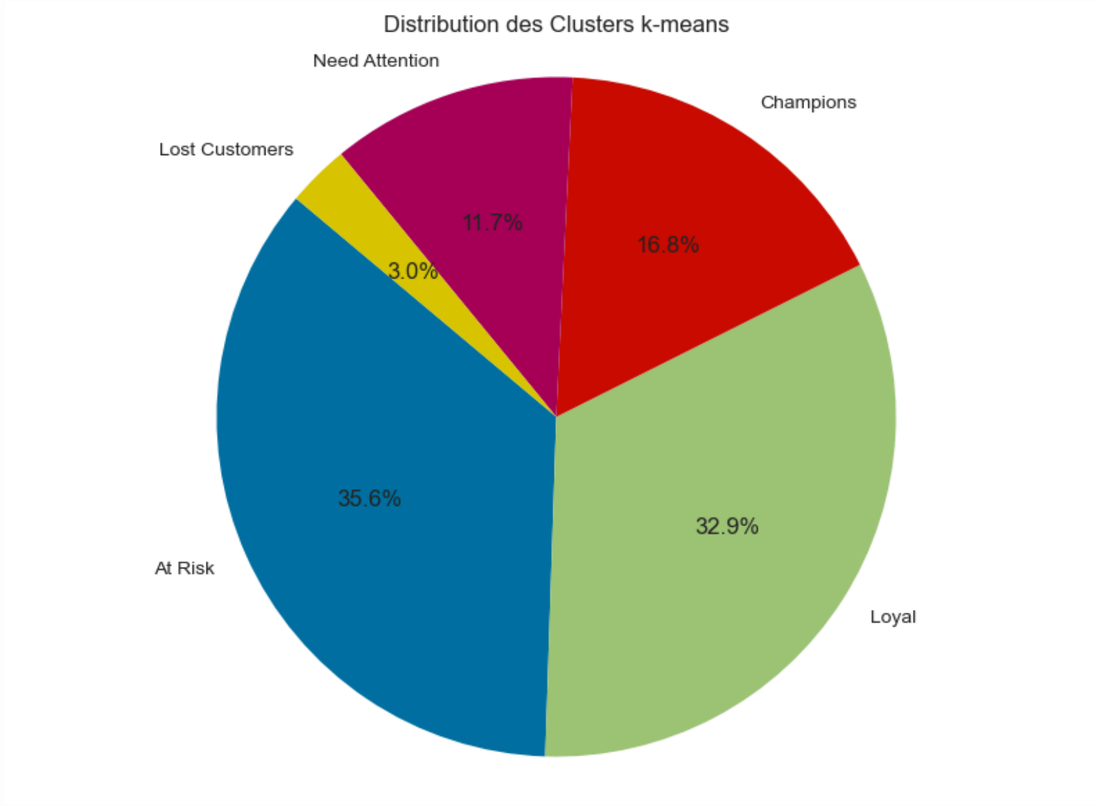
Essai de modélisation

Segmentation K-means RFM

- Les clusters varient en termes de récence, de fréquence et de valeur monétaire, le cluster 4 ayant la récence la plus faible et le cluster 2 les valeurs monétaires les plus élevées



- Nos clients se divisent dont en 5 groupes



Essai de modélisation

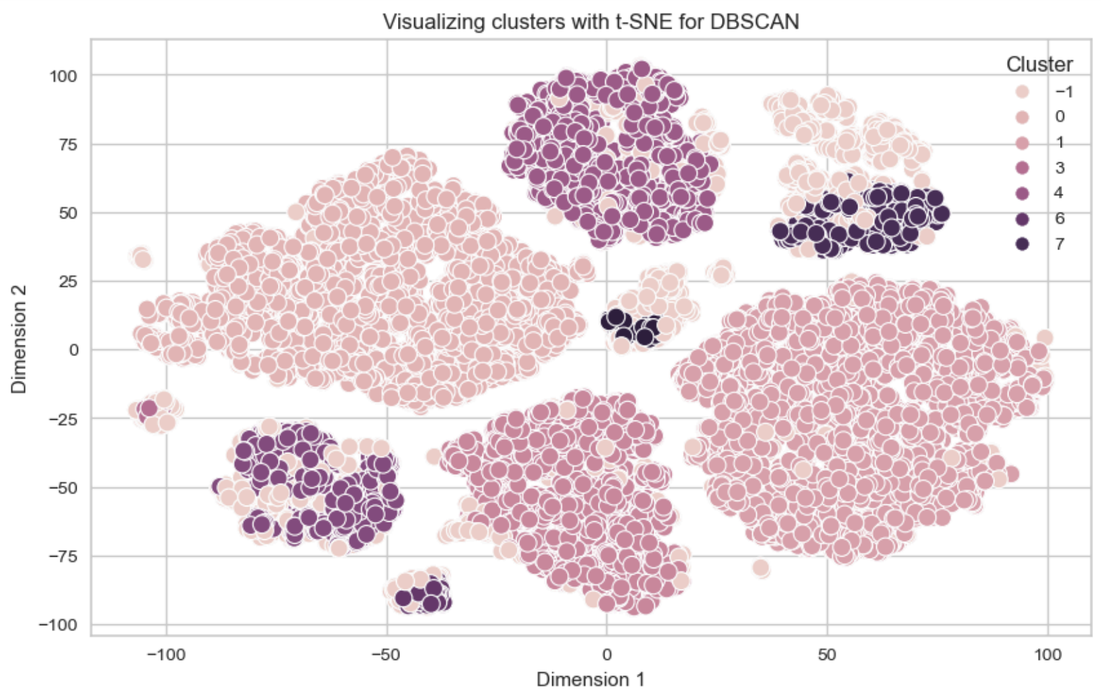
Algorithme DBSCAN

- DBSCAN regroupe des points proches les uns des autres en clusters en fonction de leur densité et considère les points isolés ou faiblement connectés comme du bruit

Epsilon est la distance maximale entre deux points pour qu'ils soient considérés comme voisins : **1**

Min_samples est le nombre minimum de points requis pour former un cluster dense : **25**

- La densité et la distribution des points dans chaque cluster révèlent des zones de forte densité de données, le DBSCAN est plutot efficace. Les points du cluster -1 ne sont pas attribués



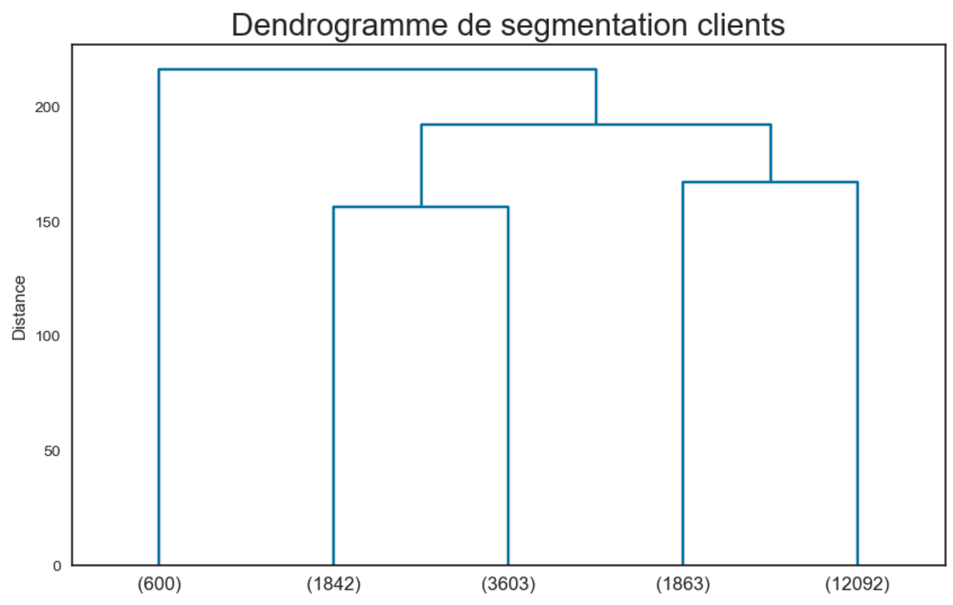

```
Estimated number of noise points: 2388
```

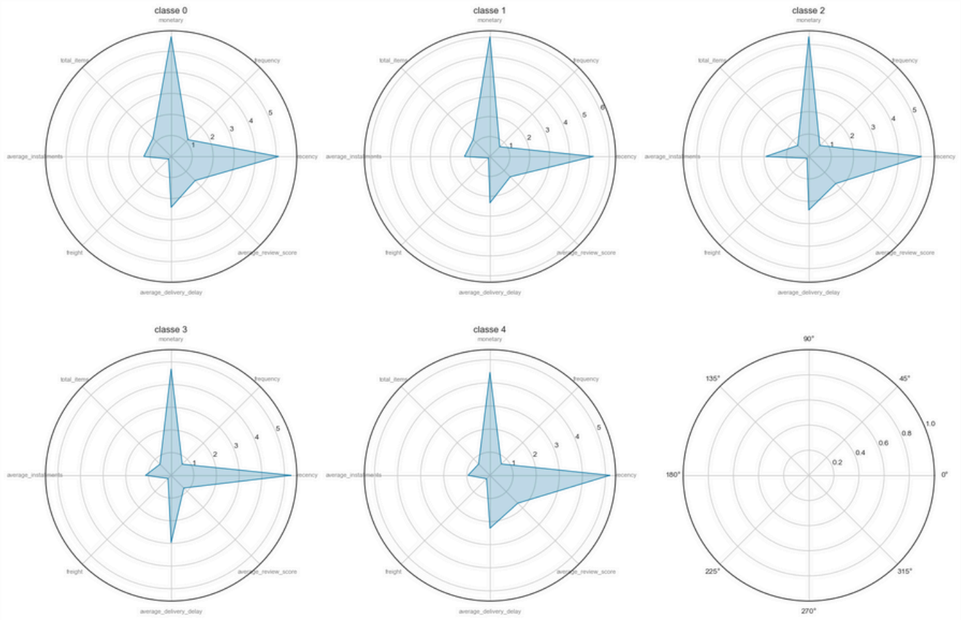
Essai de modélisation

Classification ascendante hiérarchique

- La CAH est réalisée sur un échantillon de 1/5e car très chronophage
- Les clusters sont très désequilibrés avec un cluster 5 contenant 60% des individus

- Il est difficile de discriminer les clusters au travers des variables, ce clustering n'est pas pertinent

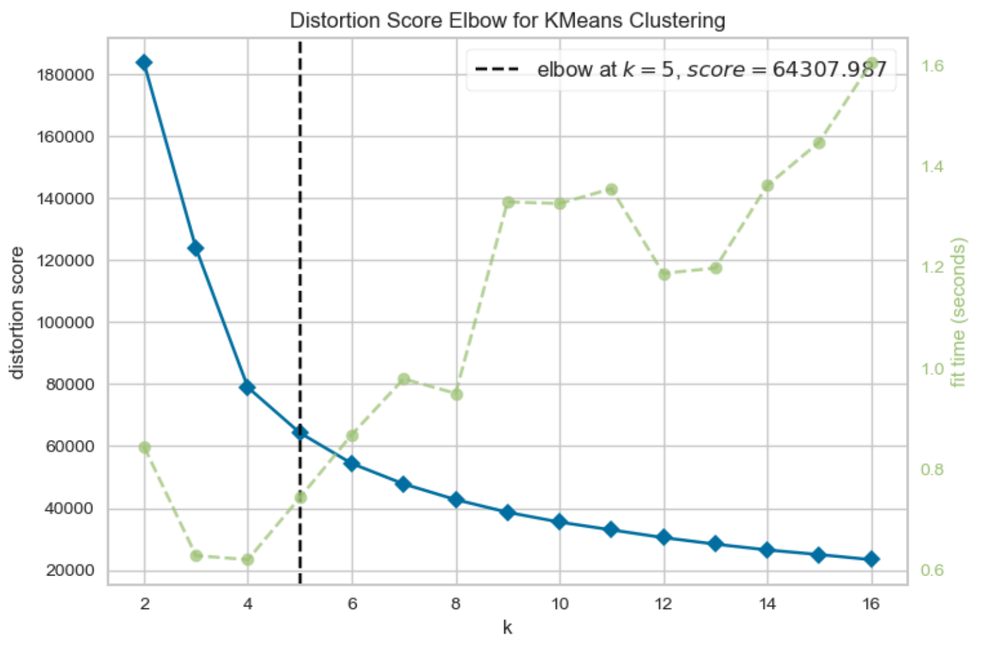




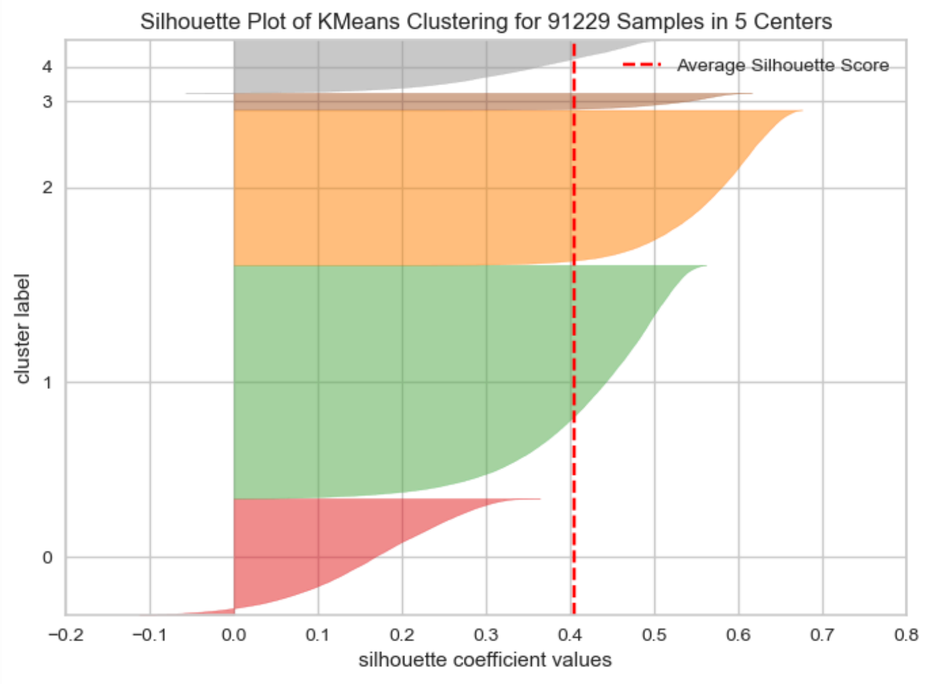
Modélisation

Modélisation K-means 8 features

- Le graphique du coude montre que k=5 est un bon choix pour le nombre de clusters dans l'algorithme K-means, car la distorsion diminue fortement jusqu'à ce point avant de se stabiliser



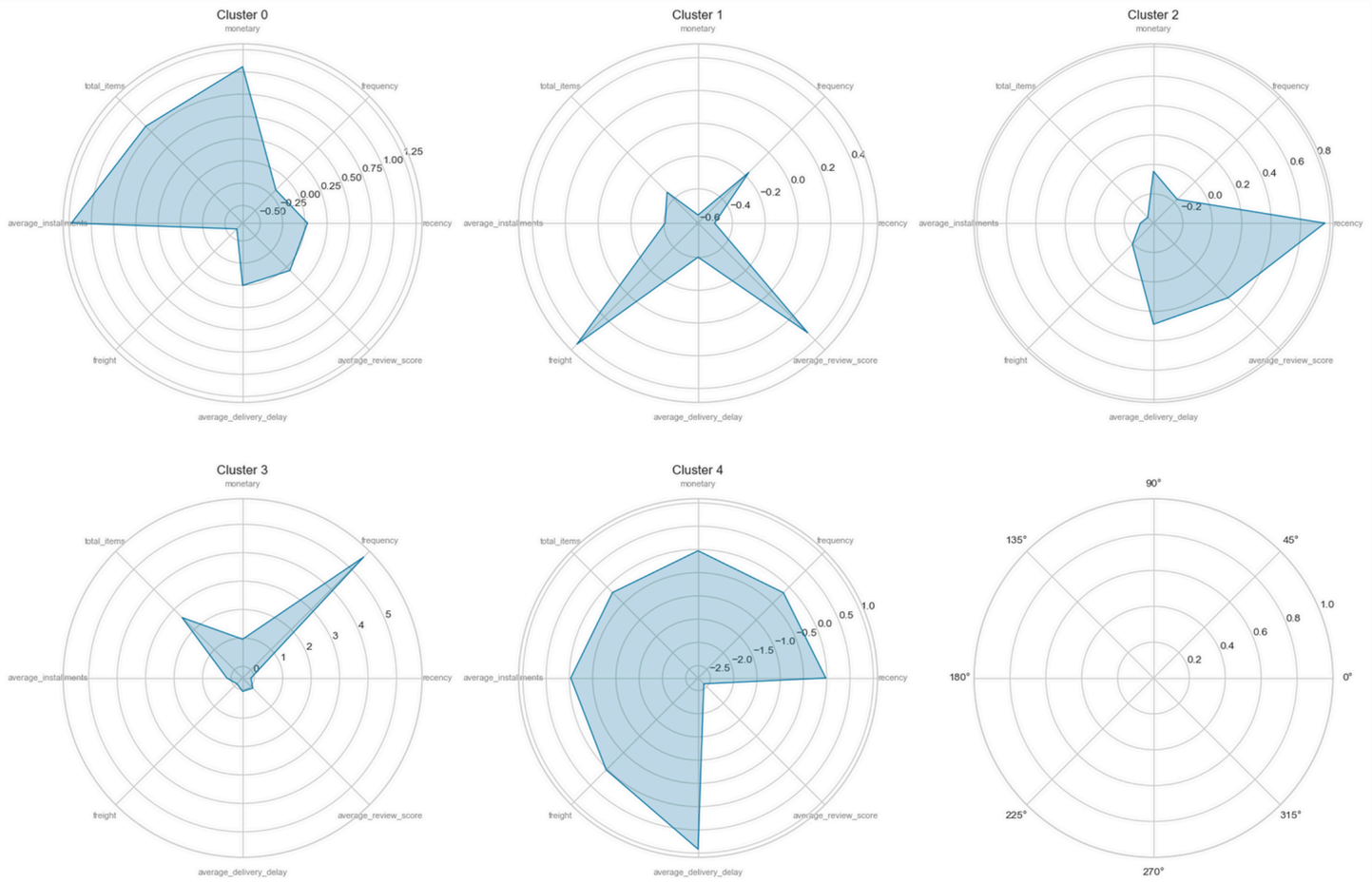
- La segmentation en 5 clusters présente une cohésion modérée avec un score de silhouette moyen d'environ 0,4, la plupart des clusters sont bien définis, bien que certains puissent être améliorés en termes de séparation et de compacité



Modélisation

Modélisation K-means 8 features

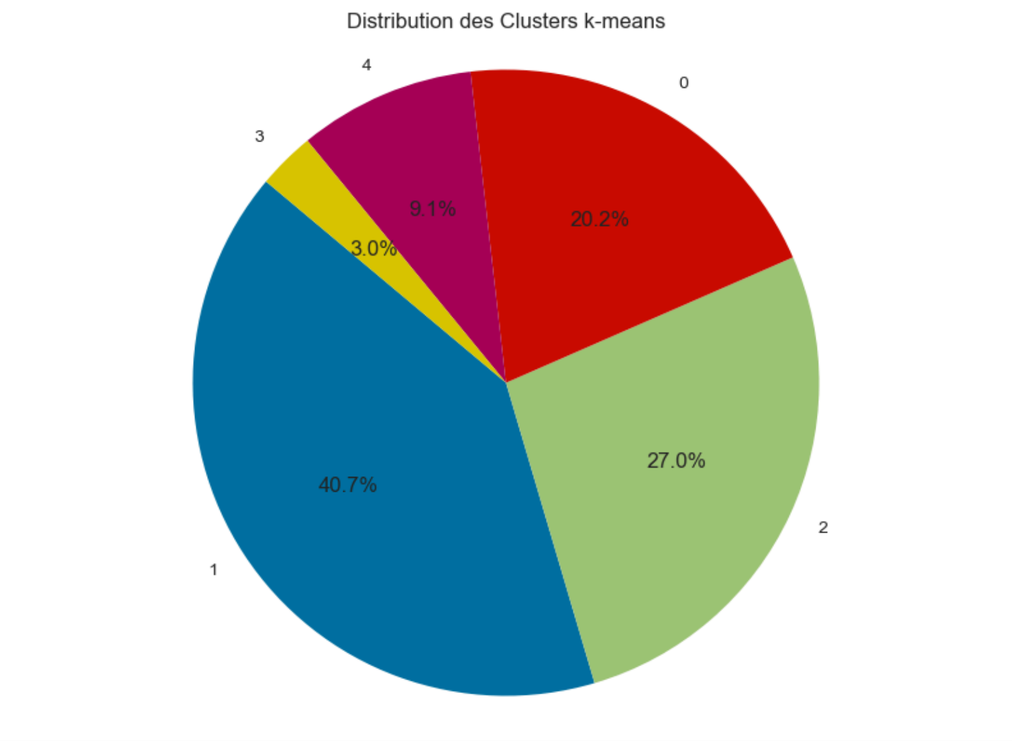
- Cluster 0 : **Clients dépensiers**
- Cluster 1 : **Clients satisfaits**
- Cluster 2 : **Clients loyaux**
- Cluster 3 : **Clients insatisfaits**
- Cluster 4 : **Clients à fort potentiel**



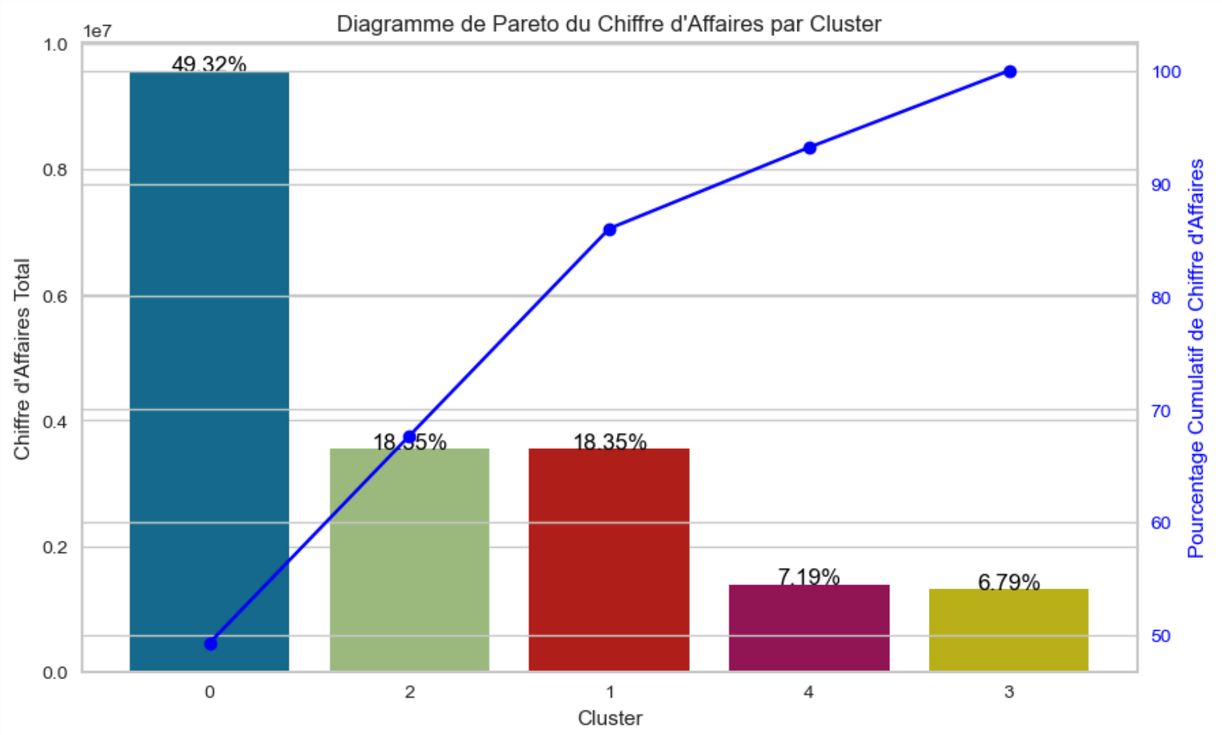
Modélisation

Segmentation K-means 8 features

- Les **clients satisfaits** (1) et **loyaux** (2) sont les plus représentés, suivis par les **clients dépensiers** (0)



- Les **clients dépensiers** réalisent presque 50% du CA, la loi de Pareto est partiellement vérifié, les clusters 0 et 2 réalisent presque 70% du CA cumulés



Feuille de route

04 Maintenance

Maintenance

Utilité

- **Adaptabilité aux nouvelles données :** Les données évoluent constamment, et la maintenance régulière permet d'adapter le modèle aux nouvelles tendances et caractéristiques des données.

- **Performance continue :** L'ajustement du modèle assure que ses performances restent optimales, évitant ainsi une dégradation dans le temps qui pourrait entraîner des décisions erronées

Méthode

- **Adjusted Rand Index (ARI) :** L'ARI est une mesure de similarité entre deux regroupements, utilisée pour évaluer la qualité des clusters produits par un algorithme de clustering. Un ARI élevé indique une forte concordance entre les clusters.

- **Évolution dans le temps :** La mesure de la performance de ce dernier dans le temps permet de définir une baisse de la stabilité de notre modèle.

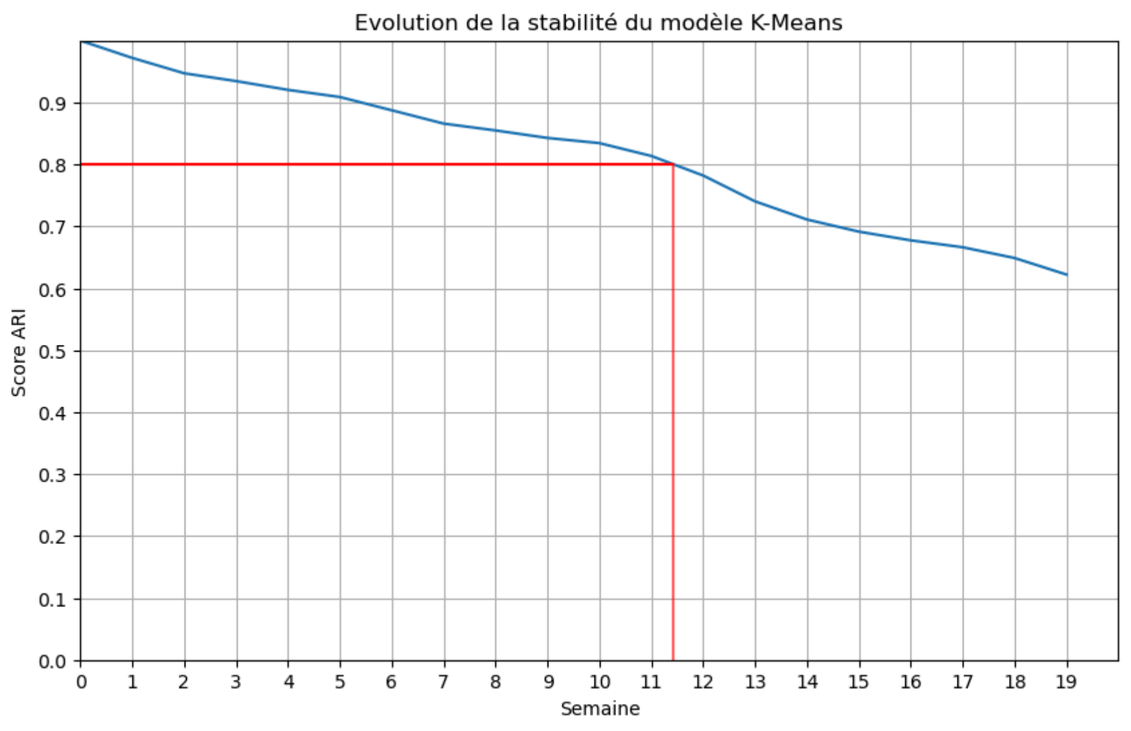
Maintenance

Définition du point critique de Mise à Jour

- Le graphique montre une diminution progressive du score ARI au court des 20 semaines mesurées, indiquant une dégradation de la stabilité du modèle K-means

- À la 12ème semaine, l'ARI-score descend en dessous du seuil critique fixé à 0.8, ce seuil suggère une perte significative de performance

- Il est crucial de réévaluer et de mettre à jour le modèle à ce stade pour maintenir son efficacité



Conclusion

Segmentation des clients

- La segmentation K-means à 5 clusters est plutot pertinente pour une modélisation à partir de 8 variables
- Elle permet de définir des process différents pour chaque types de clients (récompenser les clients dépensiers, ou comprendre les raisons des insatisfactions par exemple)
- La Mise à Jour du modèle devra s'effectuer dans une période de 3 mois, un contrat de maintenance pourra être proposé