Segmentez des clients d’un site d’e-commerce

Parcours Data Scientist - *OPENCLASSROOMS*

Plan de la présentation

1. Exploration des données
    a. Présentation de la problématique
    b. Nettoyage des données
2. Modélisation des données
    a. Réduction dimensionnelle avec l'ACP
    b. Utilisation du k-means
    c. Autres approches
    d. Etude de stabilité dans le temps
3. Conclusion

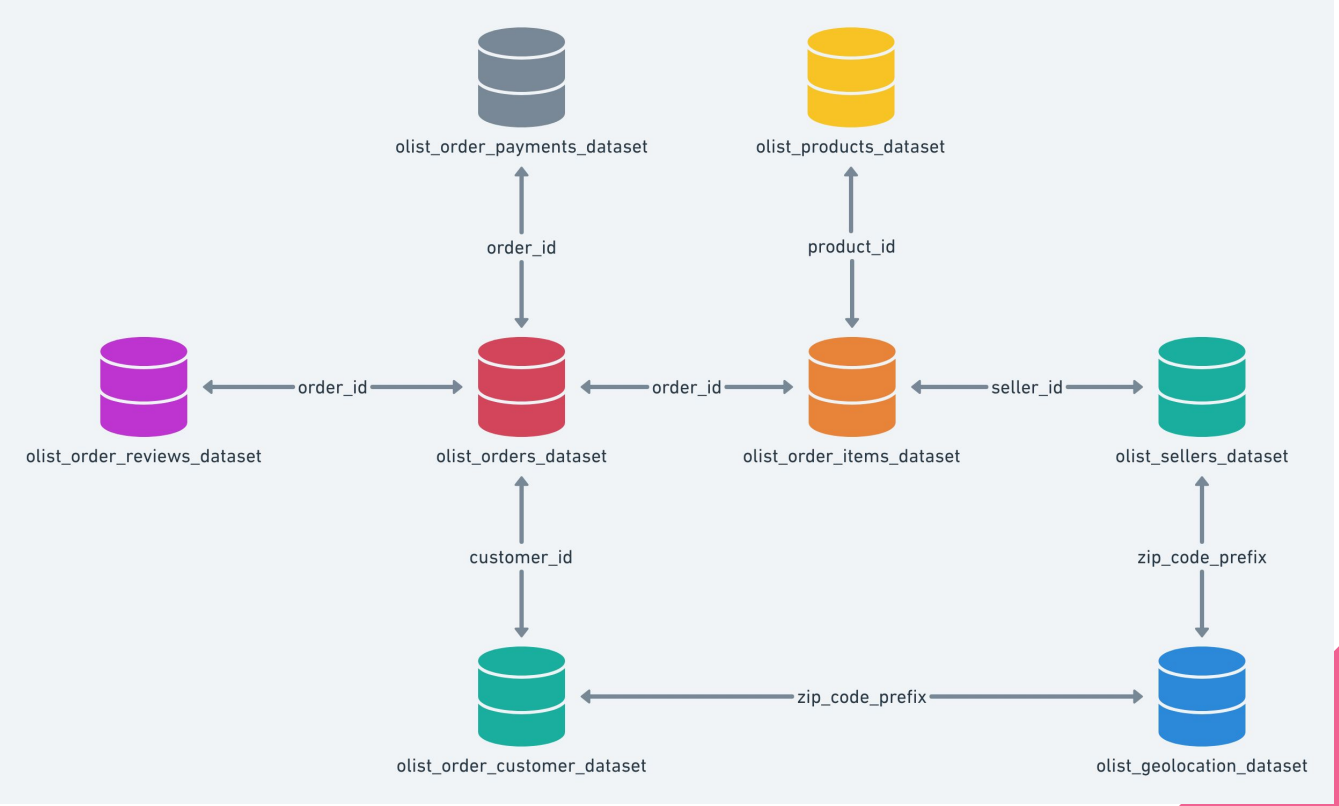
Exploration des données

Présentation de la problématique

Présentation de la problématique

- Olist => solution de vente sur les marketplaces en ligne
- Comprendre les profils des différents utilisateurs pour aider le marketing
- Les données fournies:
  - Un fichier d'utilisateurs
  - Un fichier de localisation
  - Un fichier des produits commandés par ordre
  - Un fichier des ordres
  - Un fichier de géolocalisation
  - Un fichier de paiements
  - Un fichier de produits
  - Un fichier de vendeurs
  - Et un fichier de traduction des noms de produits

Présentation de la problématique

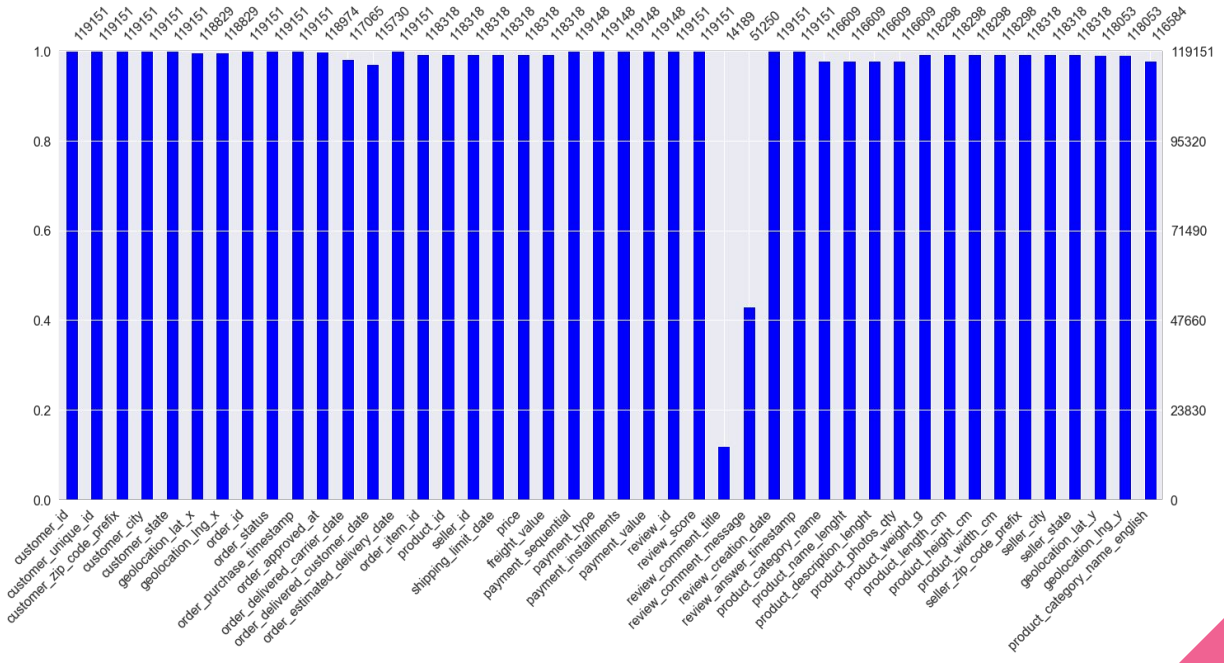


Exploration des données

Nettoyage des données

Nettoyage des données

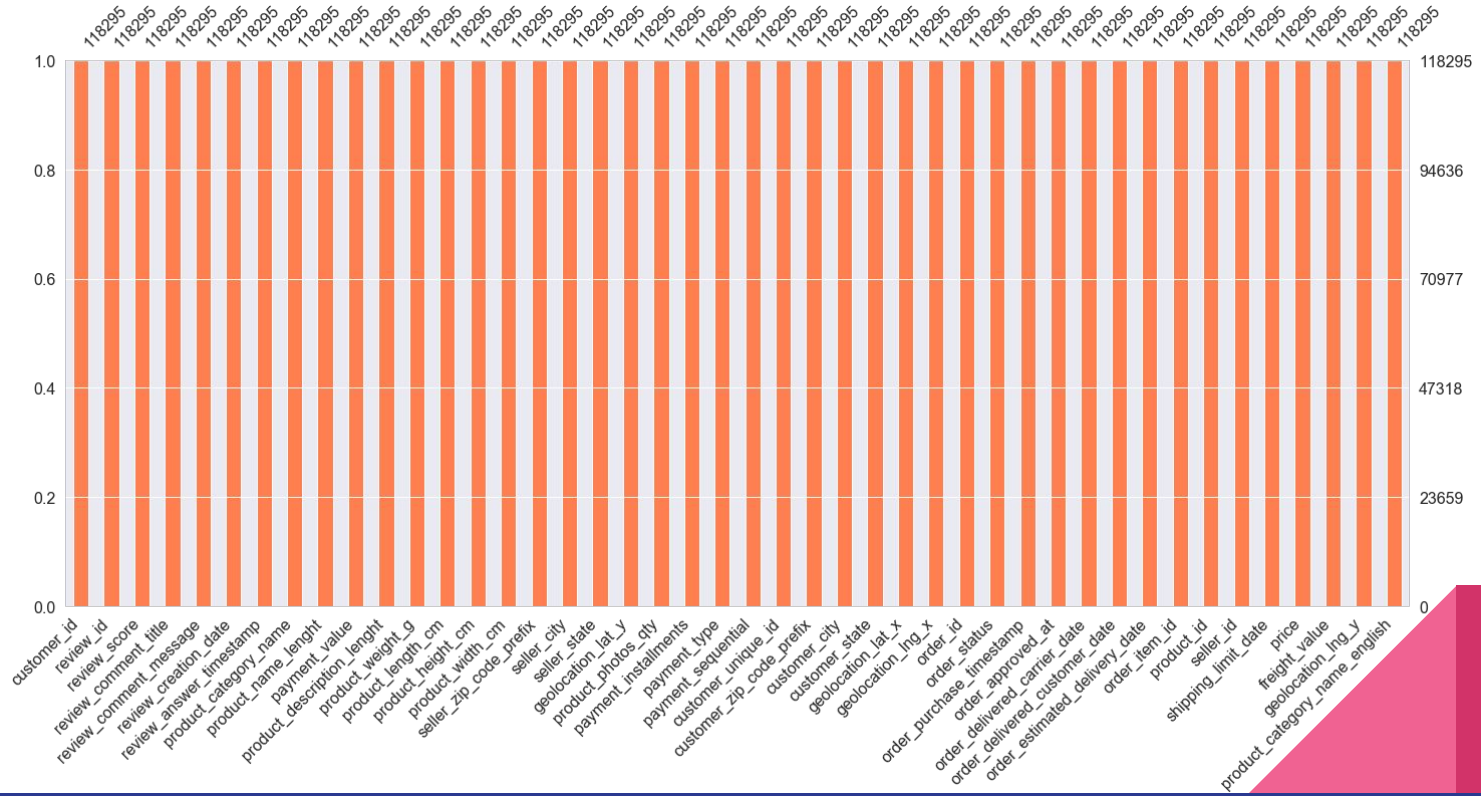
- Agrégation de tous les différents CSVs ⇒ 119151 lignes et 44 colonnes



Nettoyage des données

- On retire les lignes pour lesquelles y a pas de seller_id
- Imputation du product name, review message et review title ...
- Imputation des dates pour le ‘approved at’, ‘delivered date’, et ‘delivered carrier date’
- Imputation de valeurs à 0 pour les photos, la longueur de description du produit

Nettoyage des données



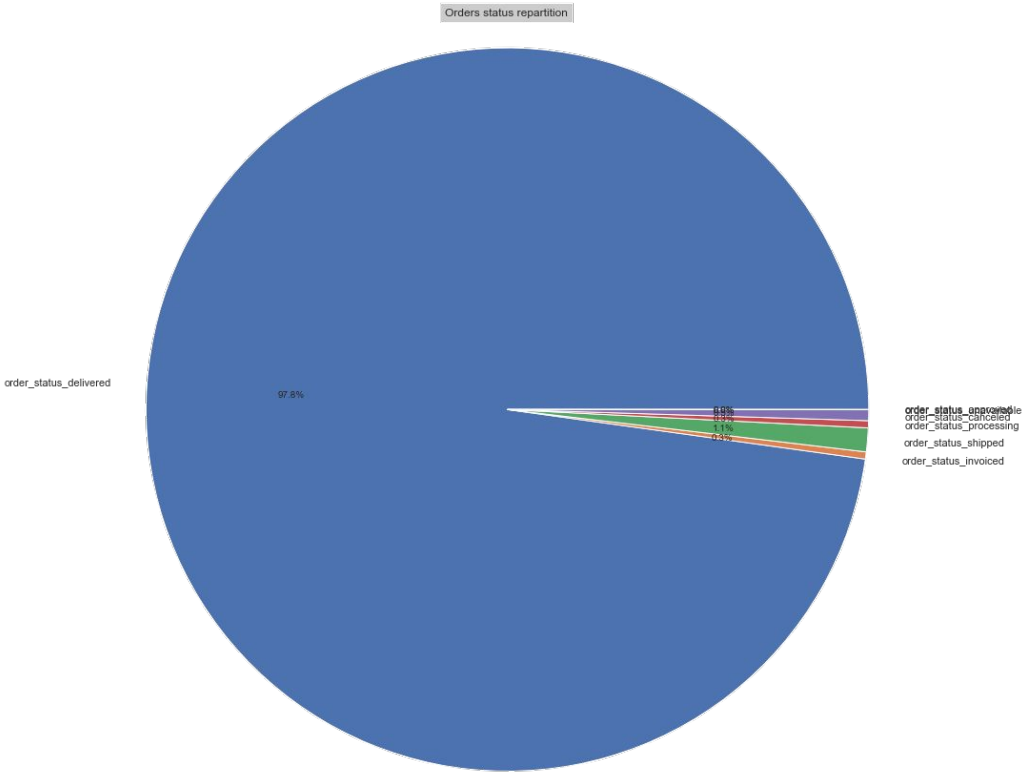
Nettoyage des données

- Transformation des colonnes ‘date’ en ‘délai’
- Créer 9 catégories de produits:
  - Misc
  - Electronic
  - Office
  - Construction
  - Food
  - House
  - Mode
  - Leisure
  - Auto

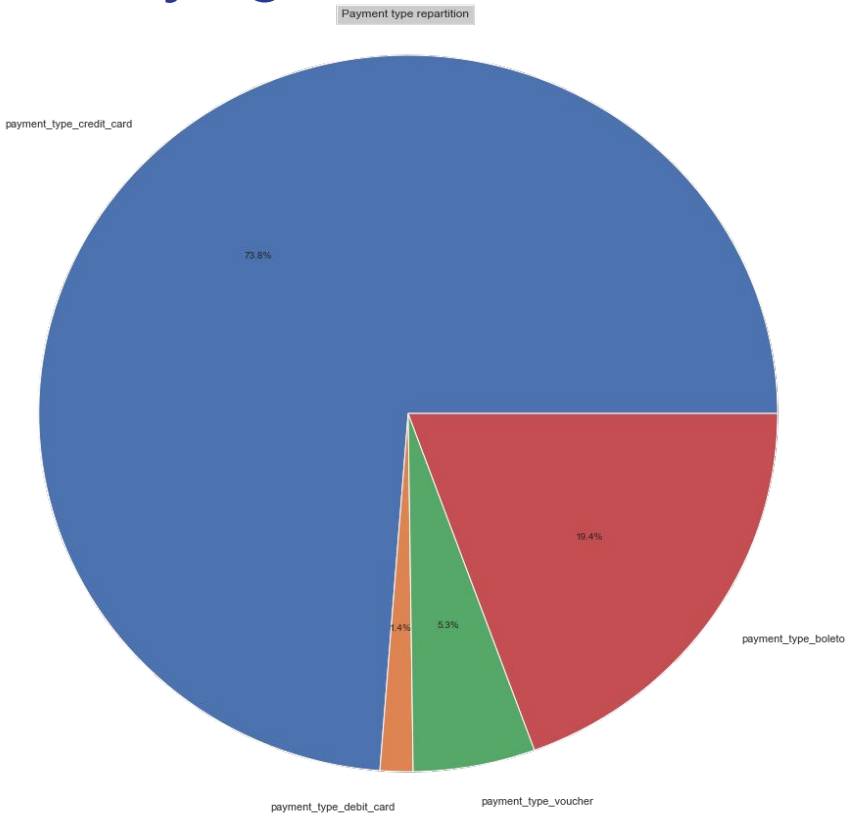
Nettoyage des données

- Utilisation du volume de produit (au lieu de longueur, largeur et hauteur)
- Création d’une variable distance (pour remplacer les coordonnées de départ et d’arrivée)
- Création d’une colonne par type de paiement, et par type d’ordre

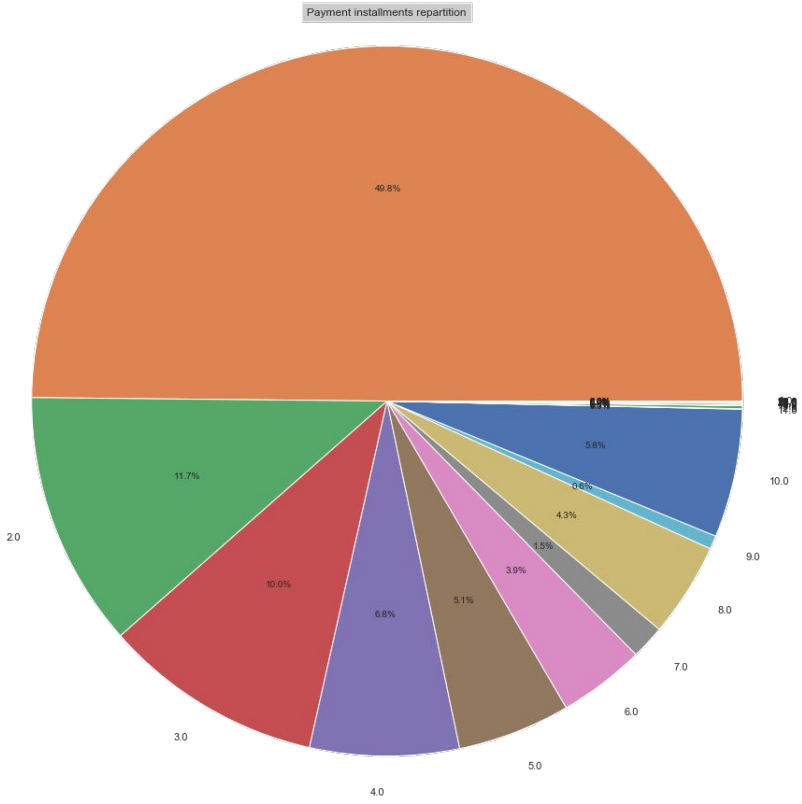
Nettoyage des données



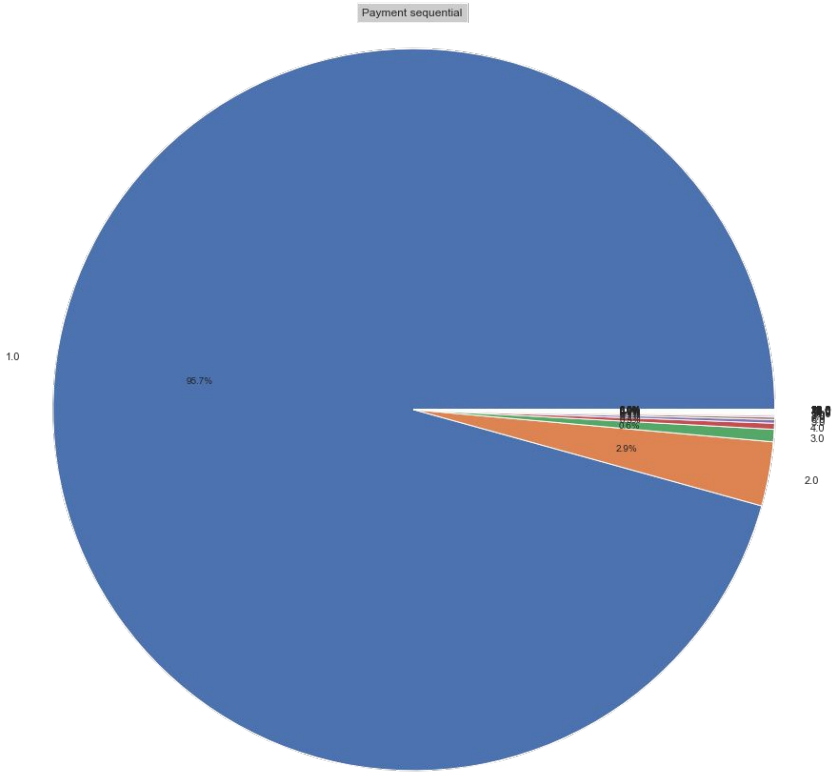
Nettoyage des données



Nettoyage des données

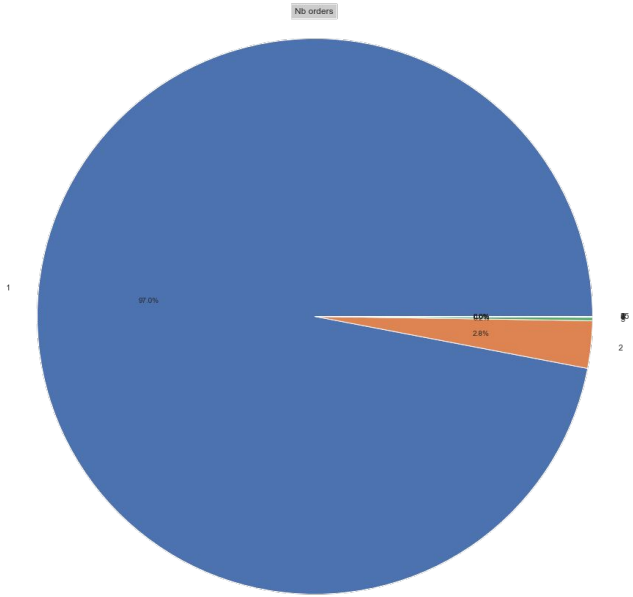


Nettoyage des données

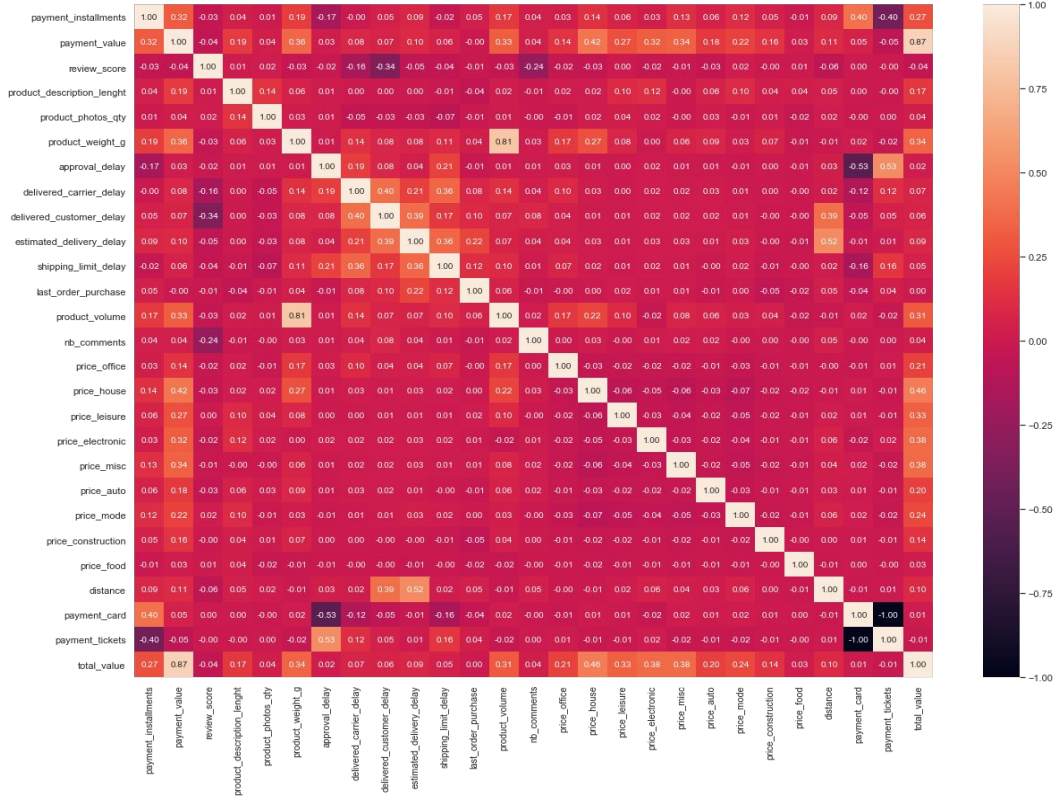


Nettoyage des données

- Nos données après avoir appliqué un grouping : 96461 lignes et 35 colonnes
- Grouper le price et les frais de livraison en une seule valeur total value

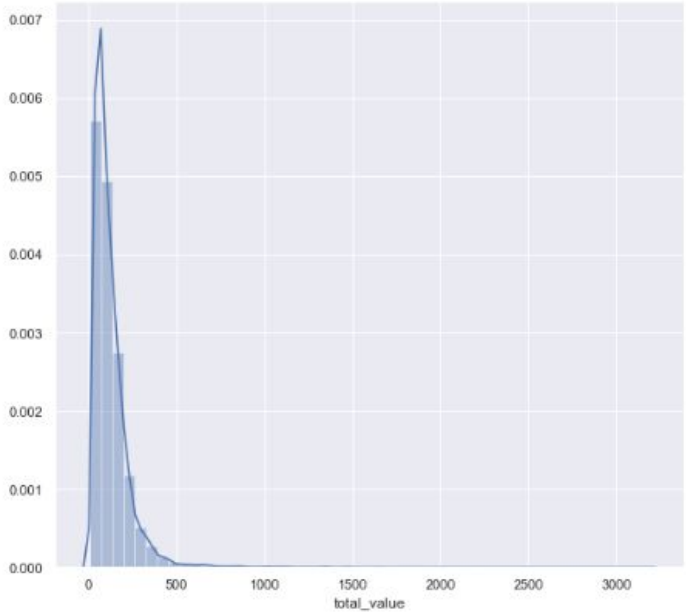


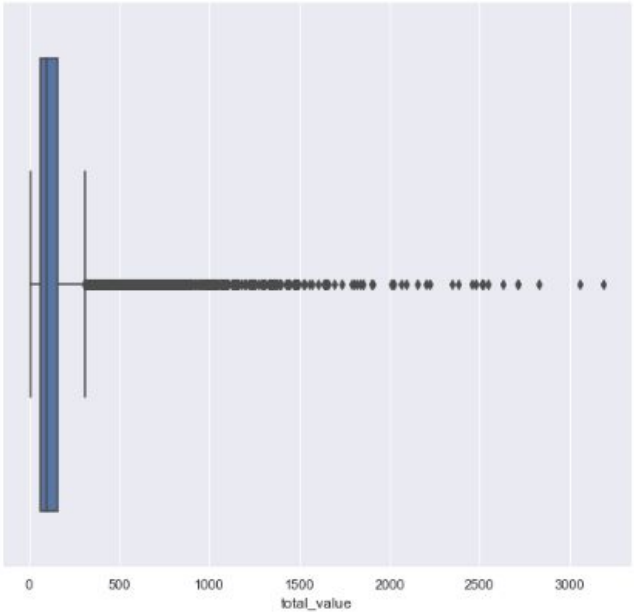
Nettoyage des données



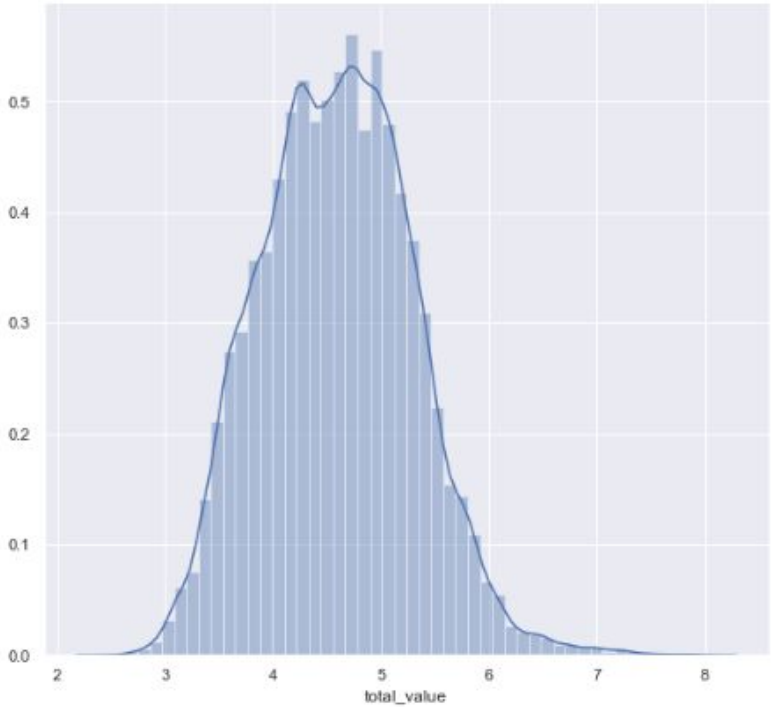
Nettoyage des données

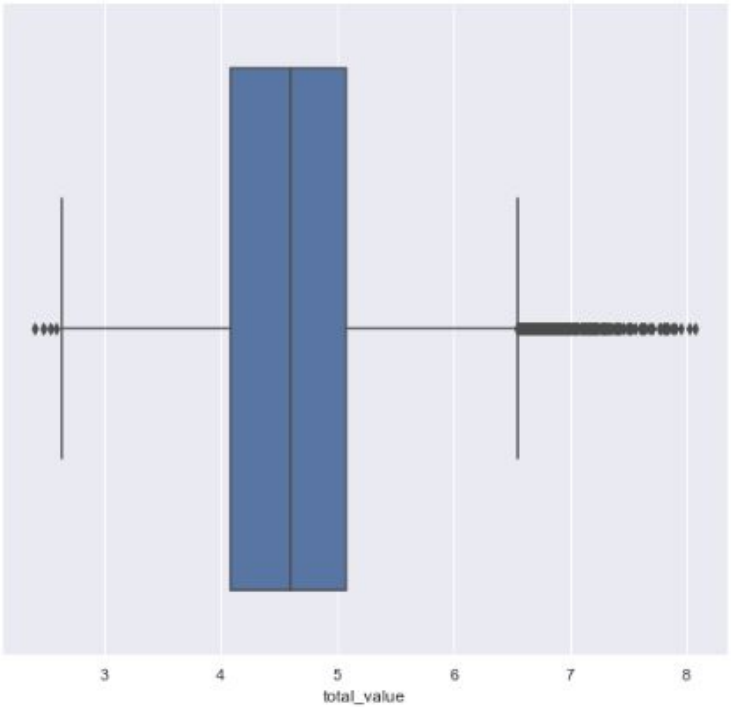
- Distribution des données





Nettoyage des données

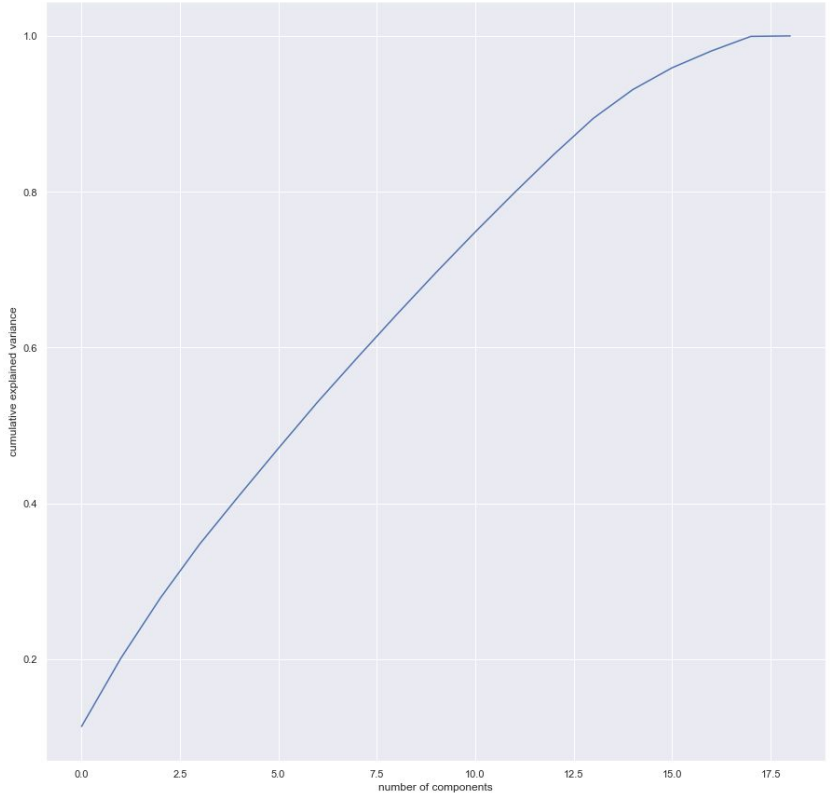


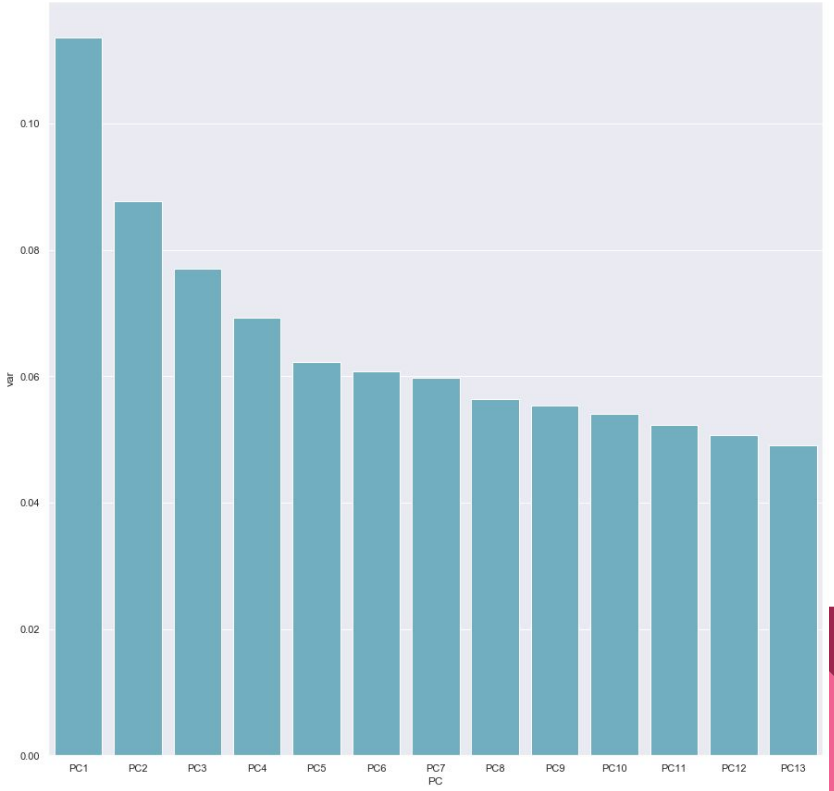


Modélisation des données

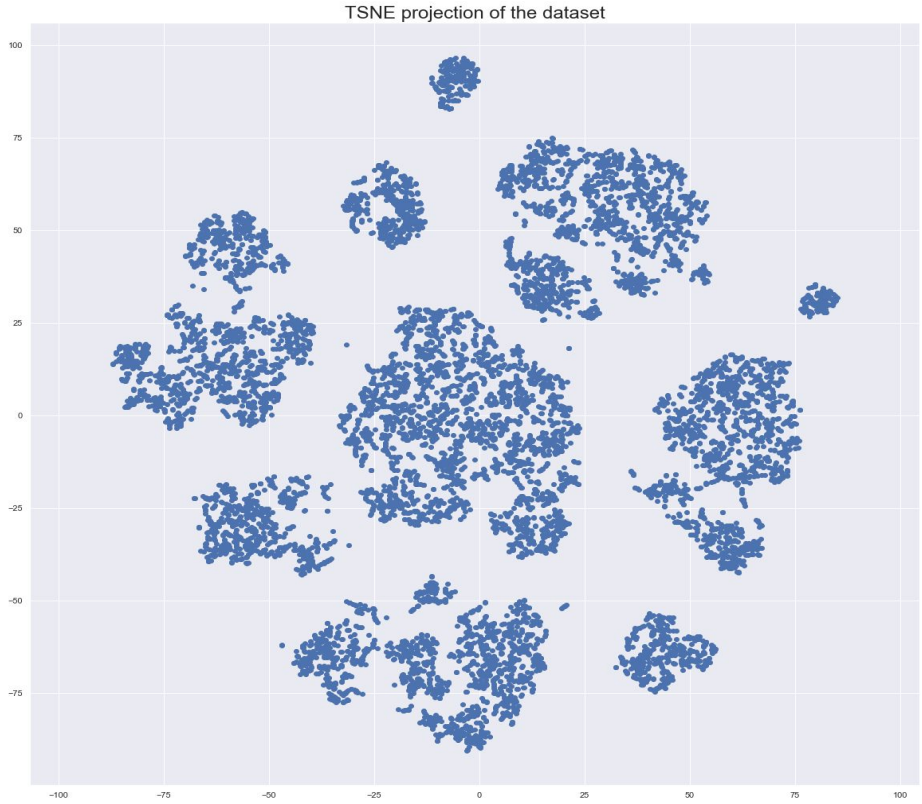
Réduction dimensionnelle avec l’ACP

Réduction dimensionnelle avec l’ACP

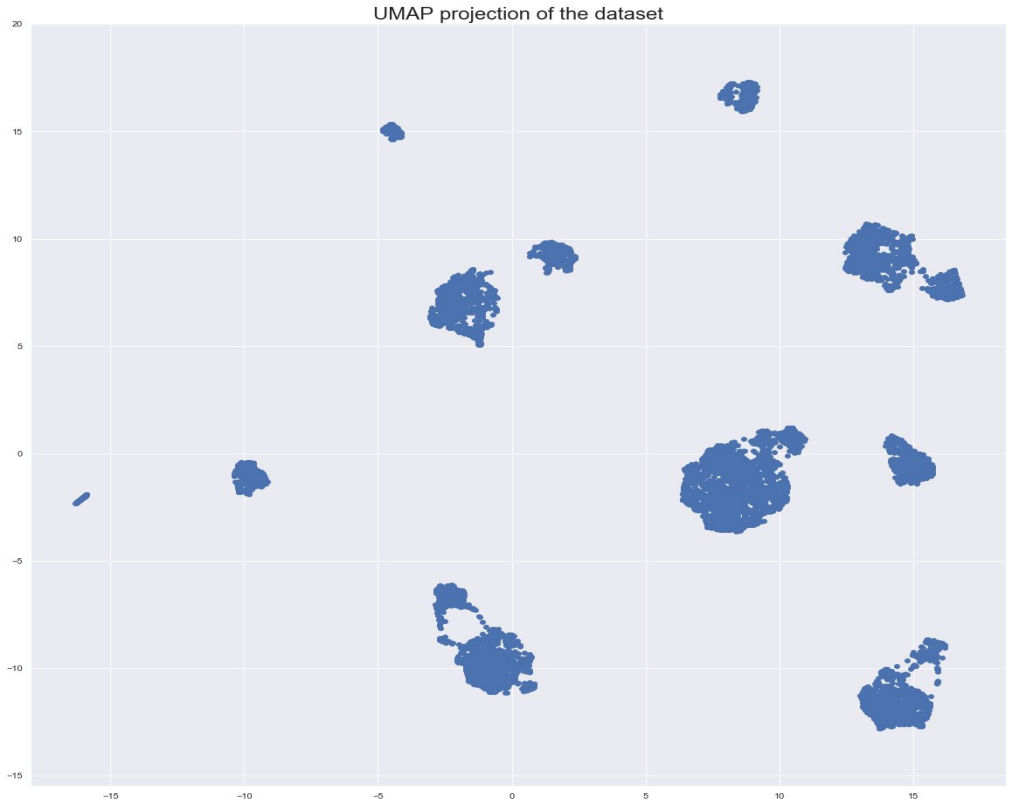




Réduction dimensionnelle avec l’ACP



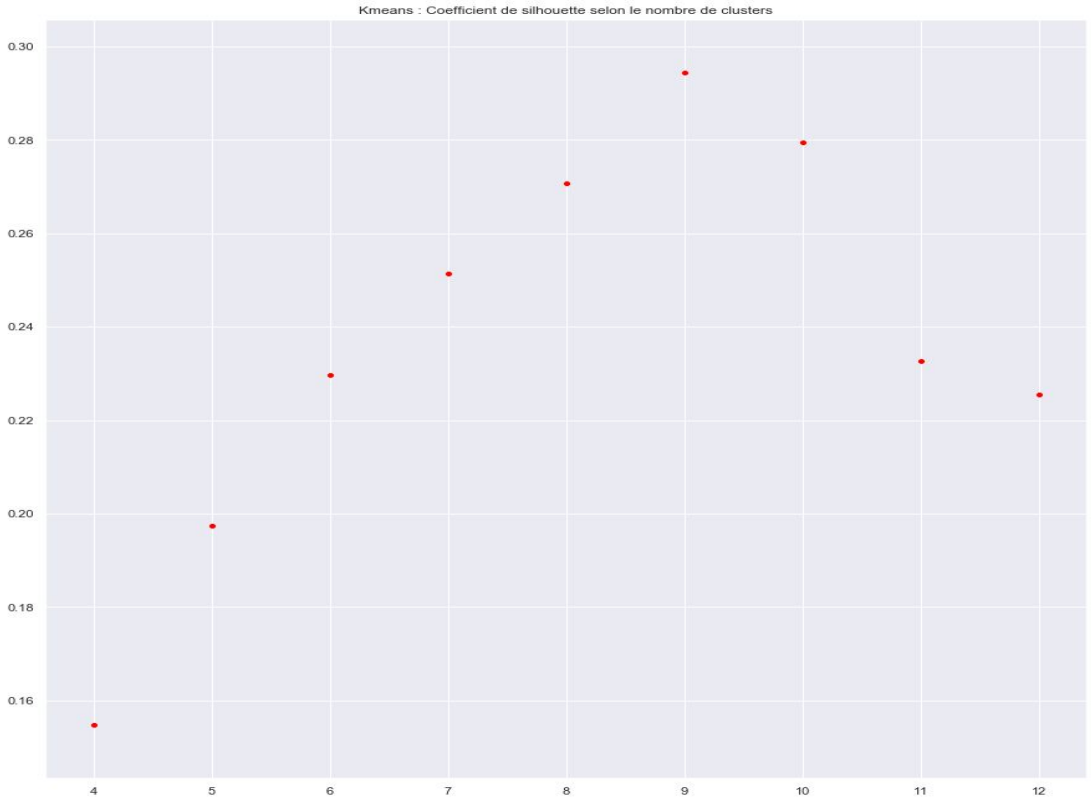
Réduction dimensionnelle avec l’ACP



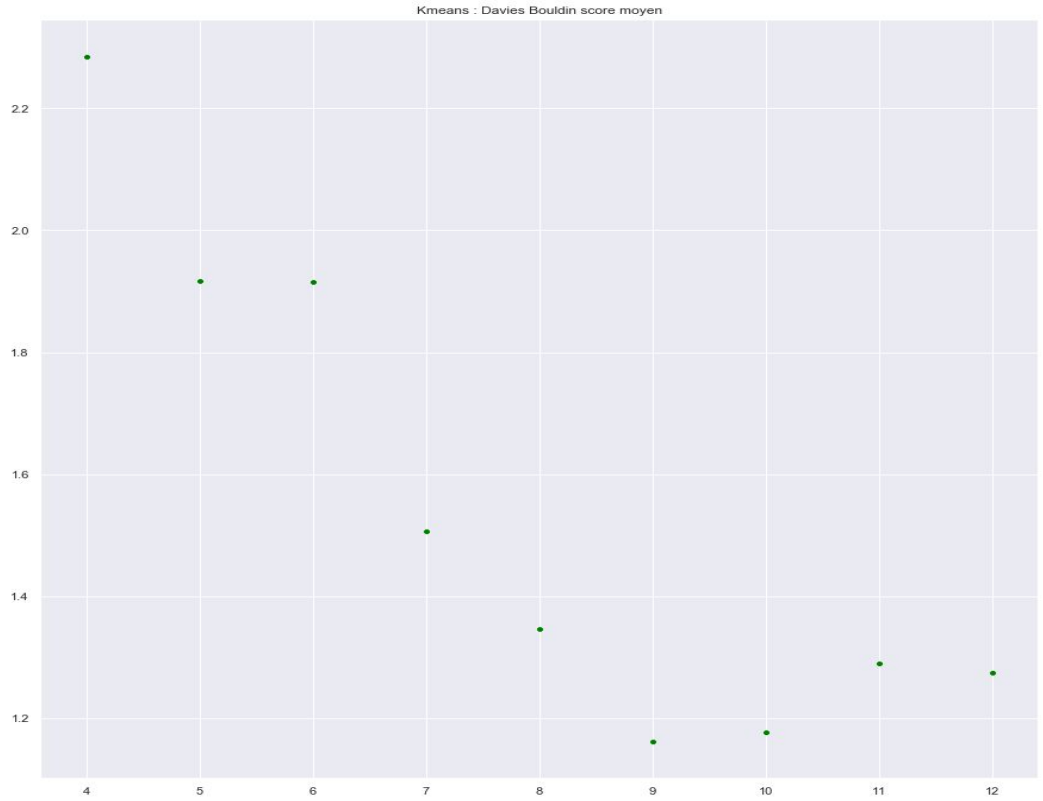
Modélisation des données

Utilisation du k-means

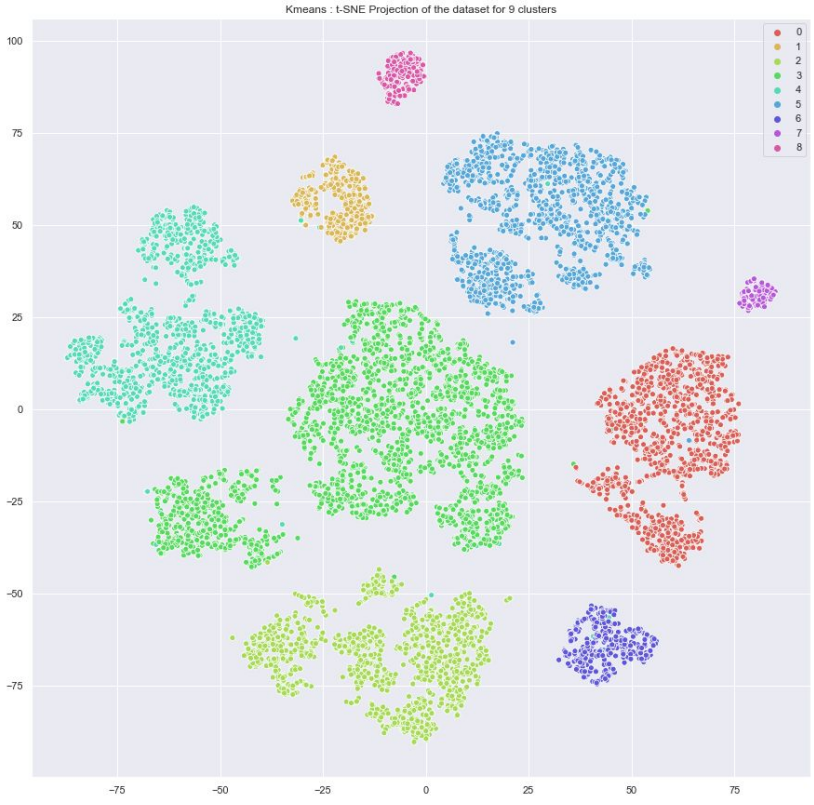
Utilisation du k-means



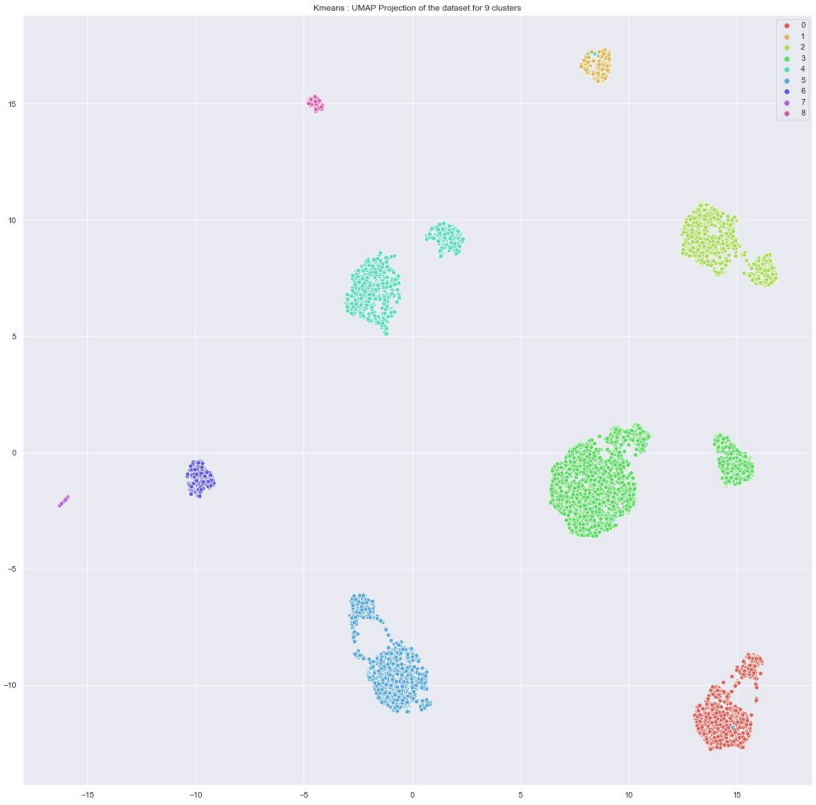
Utilisation du k-means



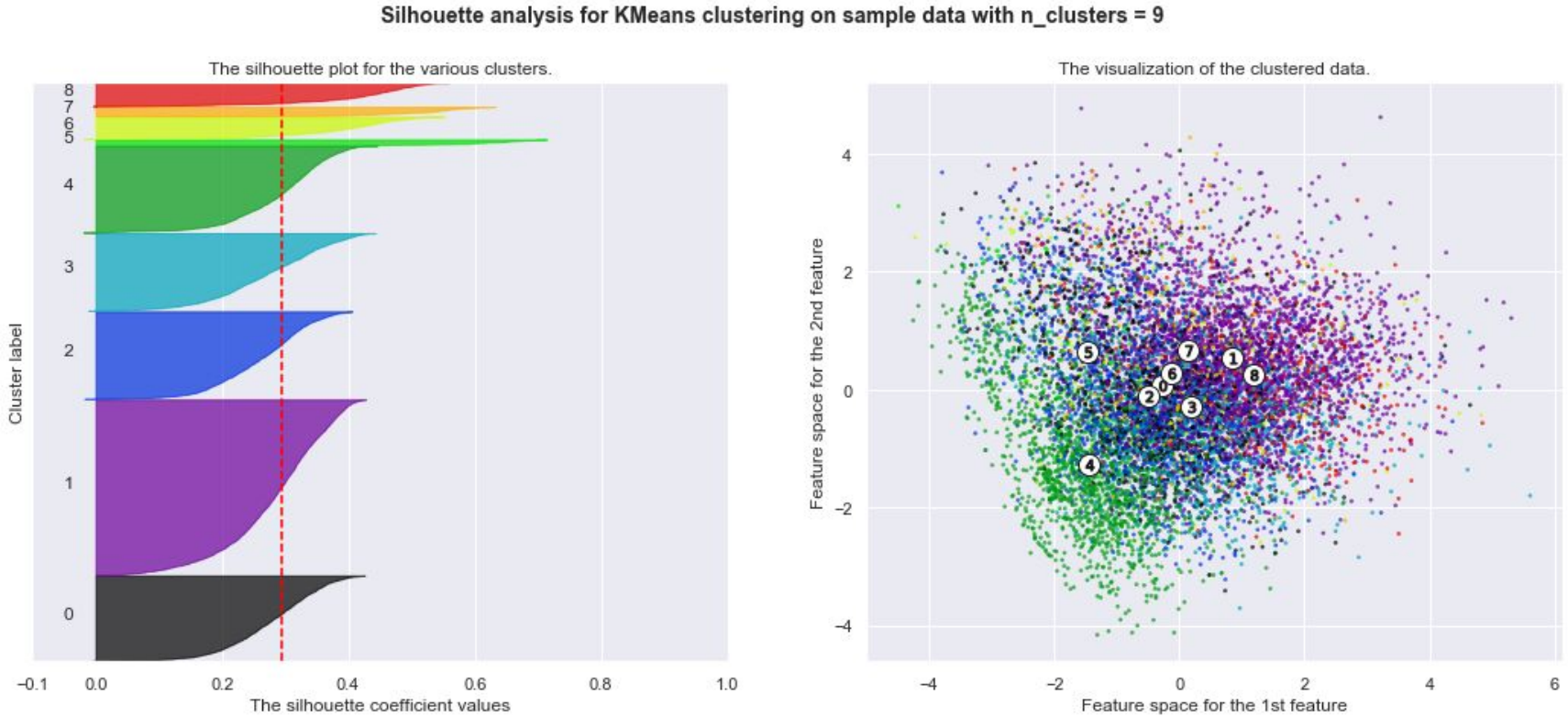
Utilisation du k-means



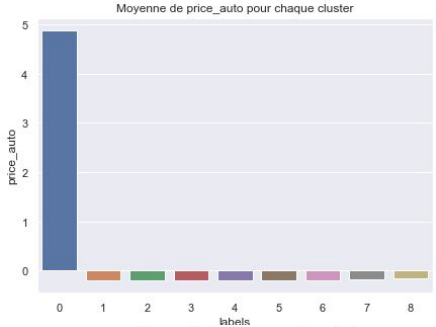
Utilisation du k-means

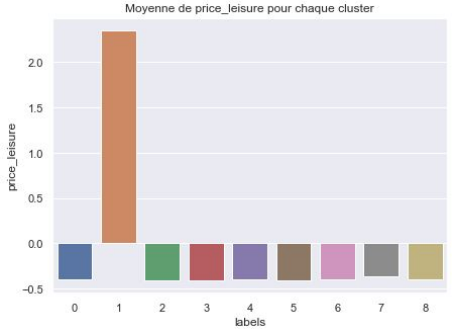


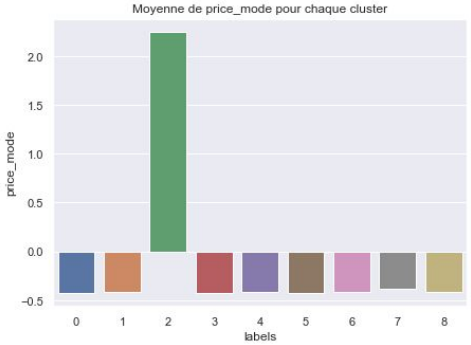
Utilisation du k-means

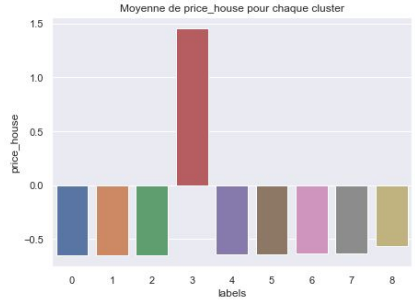


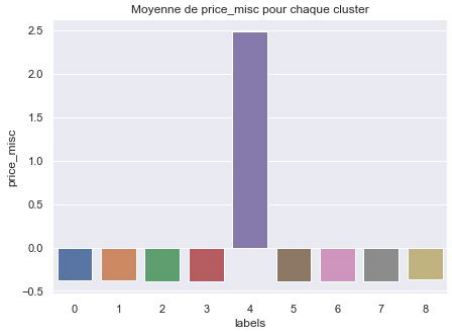
Utilisation du k-means

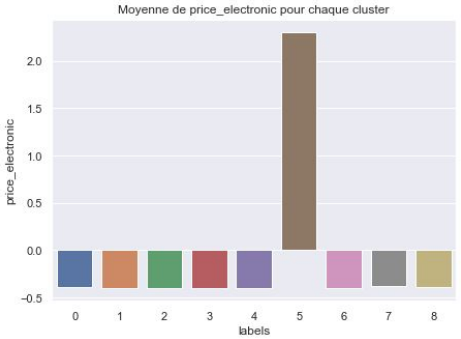






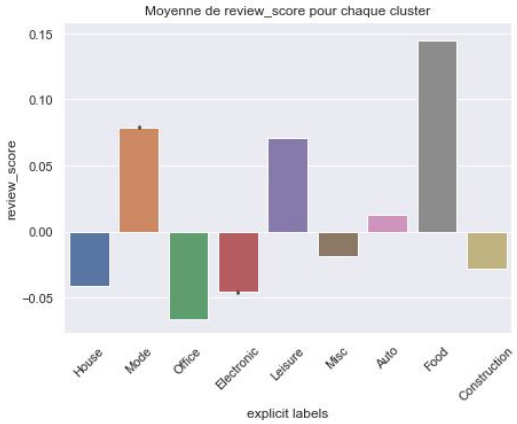


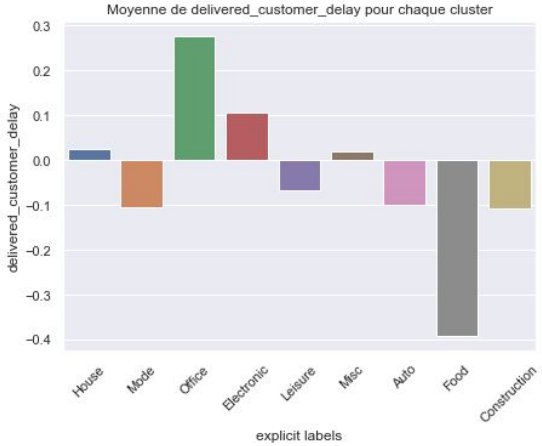


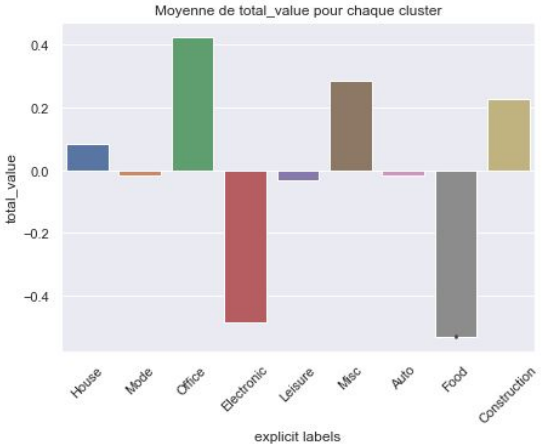


Utilisation du k-means

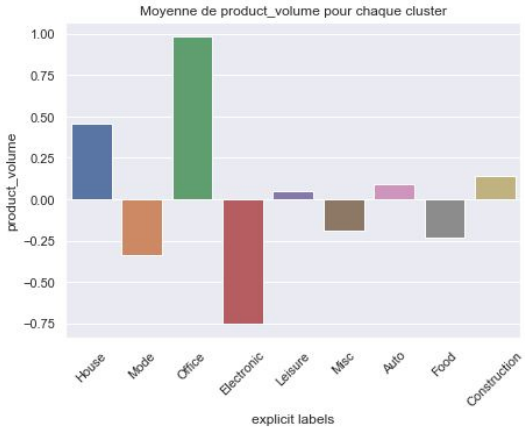
On va rajouter des labels plus explicites à chaque cluster

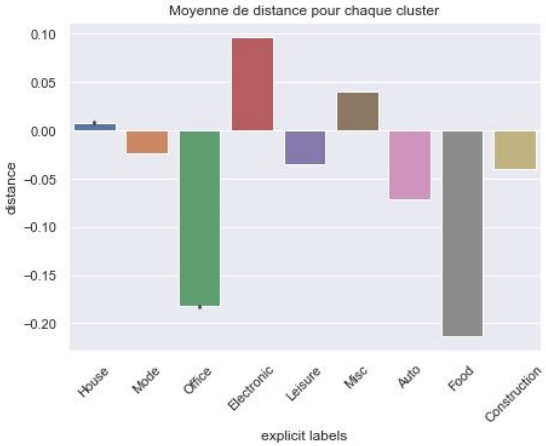


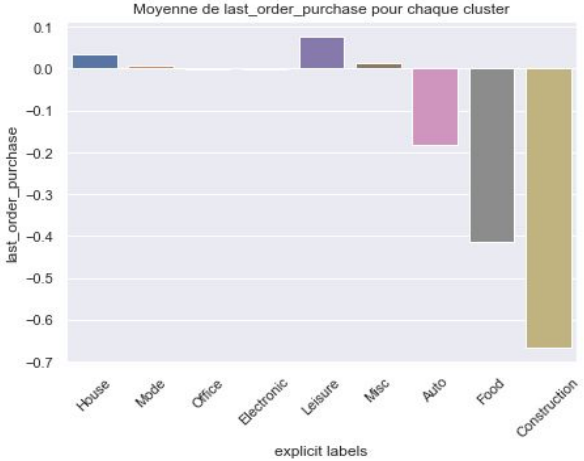




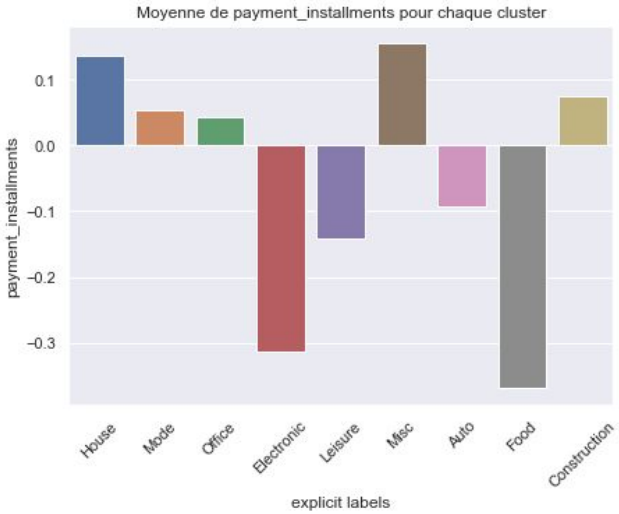
Utilisation du k-means

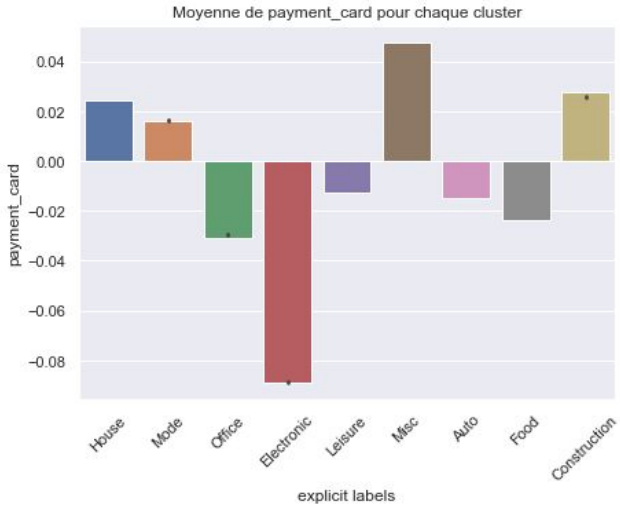






Utilisation du k-means



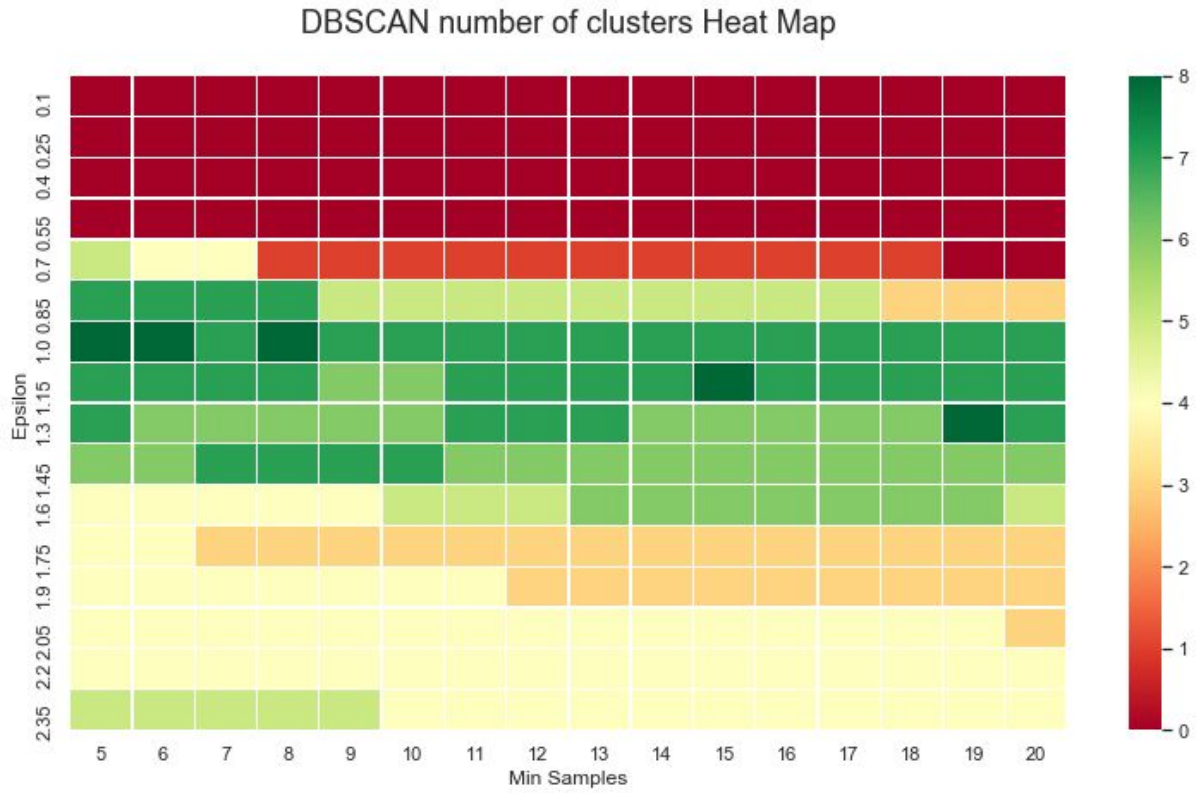


Modélisation des données

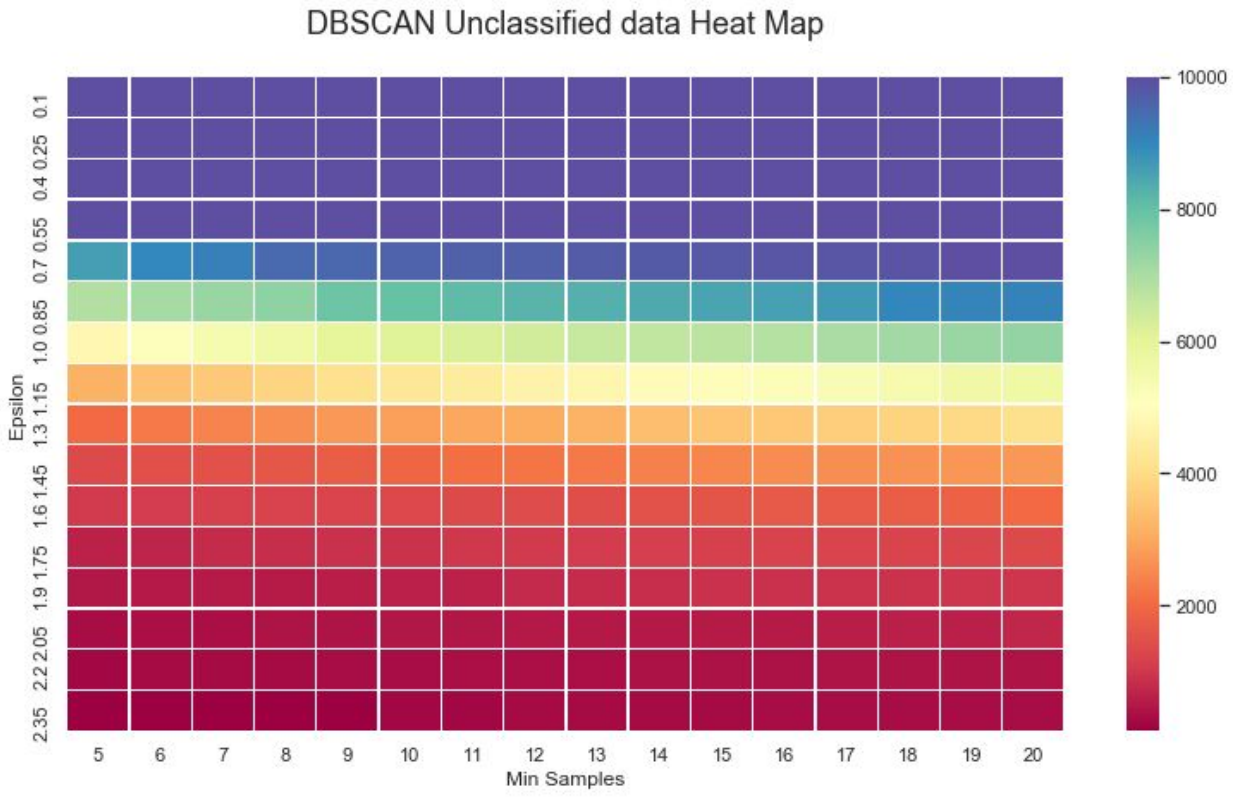
Autres approches

Autres approches : Clustering hiérarchique

Autres approches: DBScan



Autres approches: DBScan

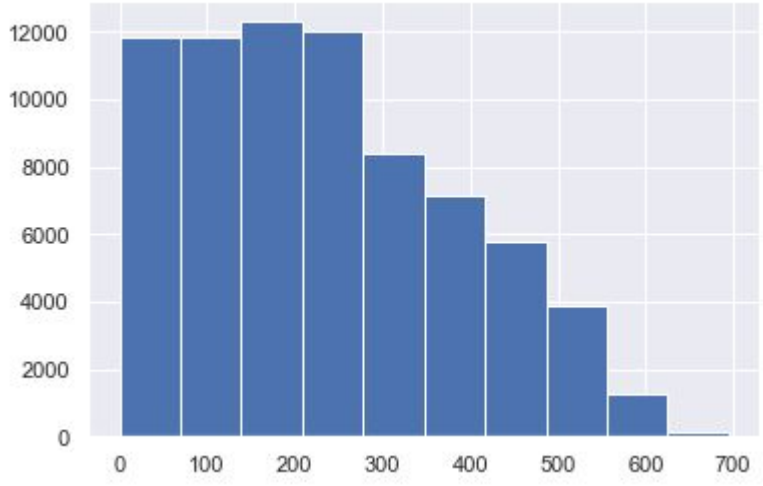


Modélisation des données

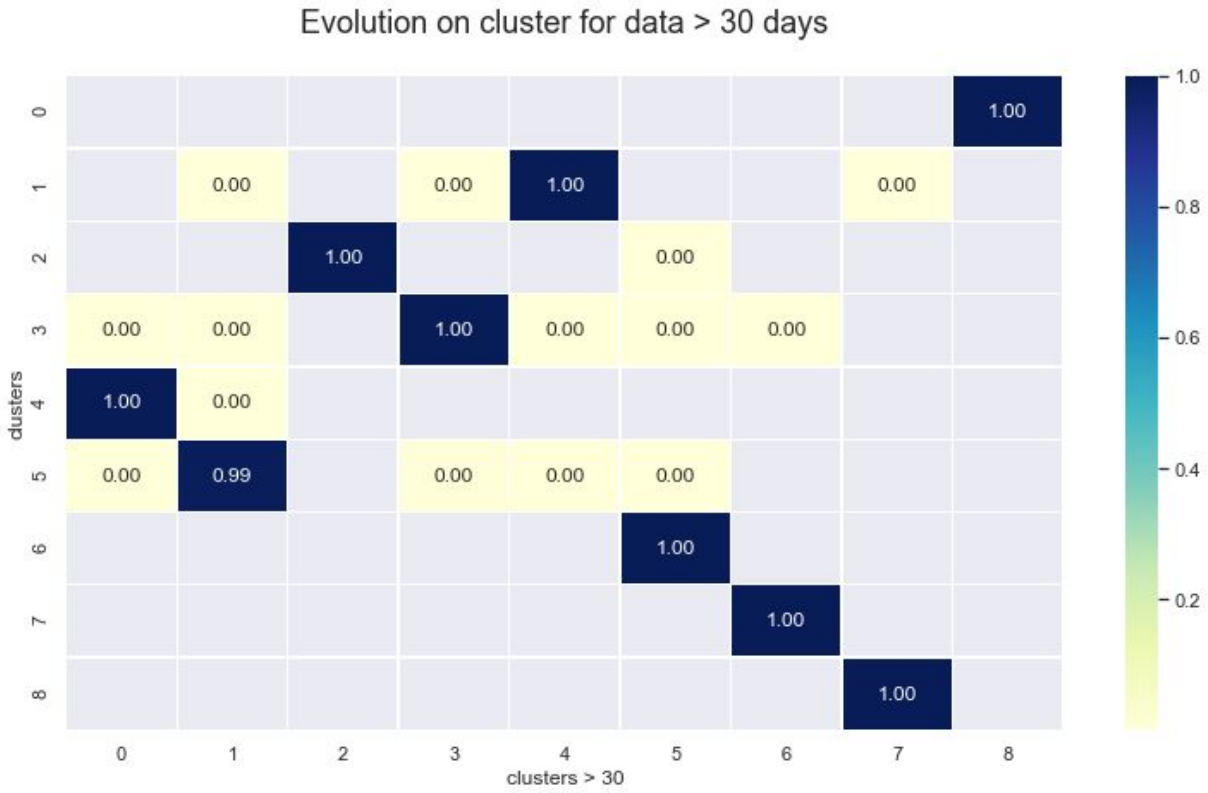
Stabilité dans le temps

Stabilité dans le temps

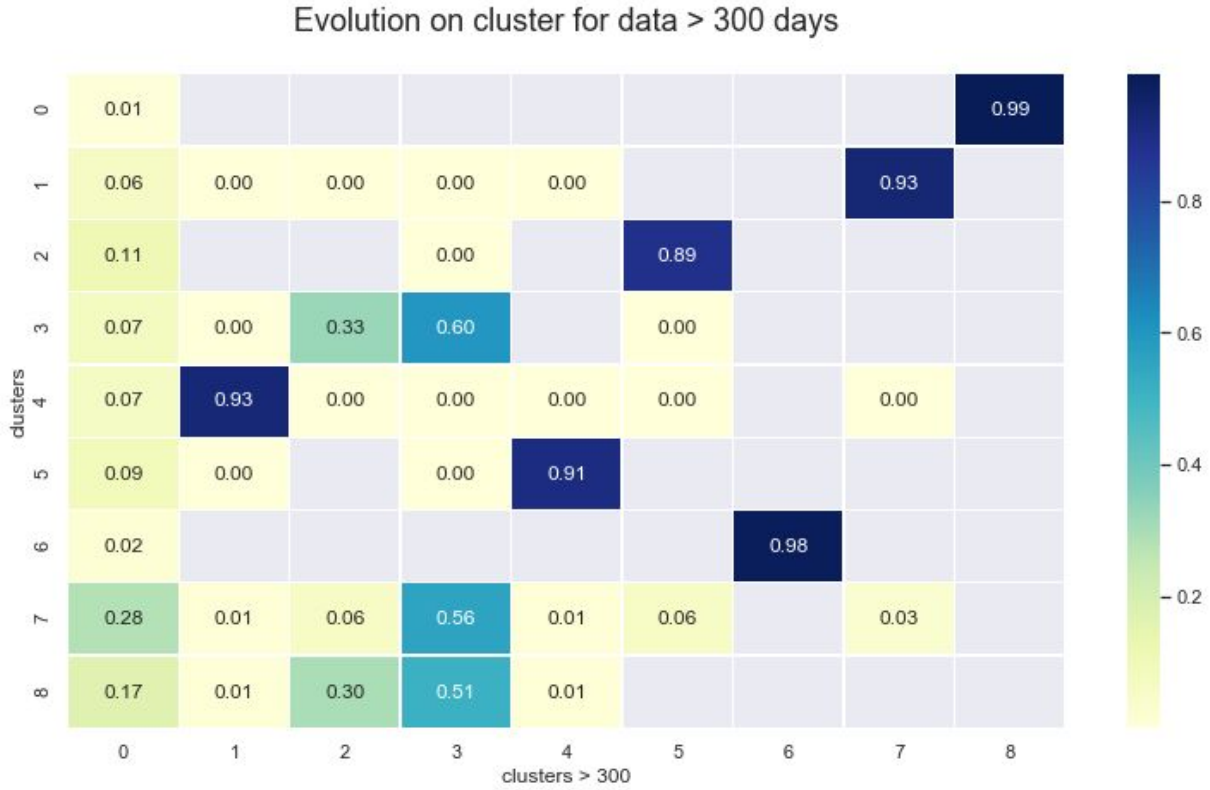
- Utilisation de la colonne purchase_delay, qu'on va "décaler":



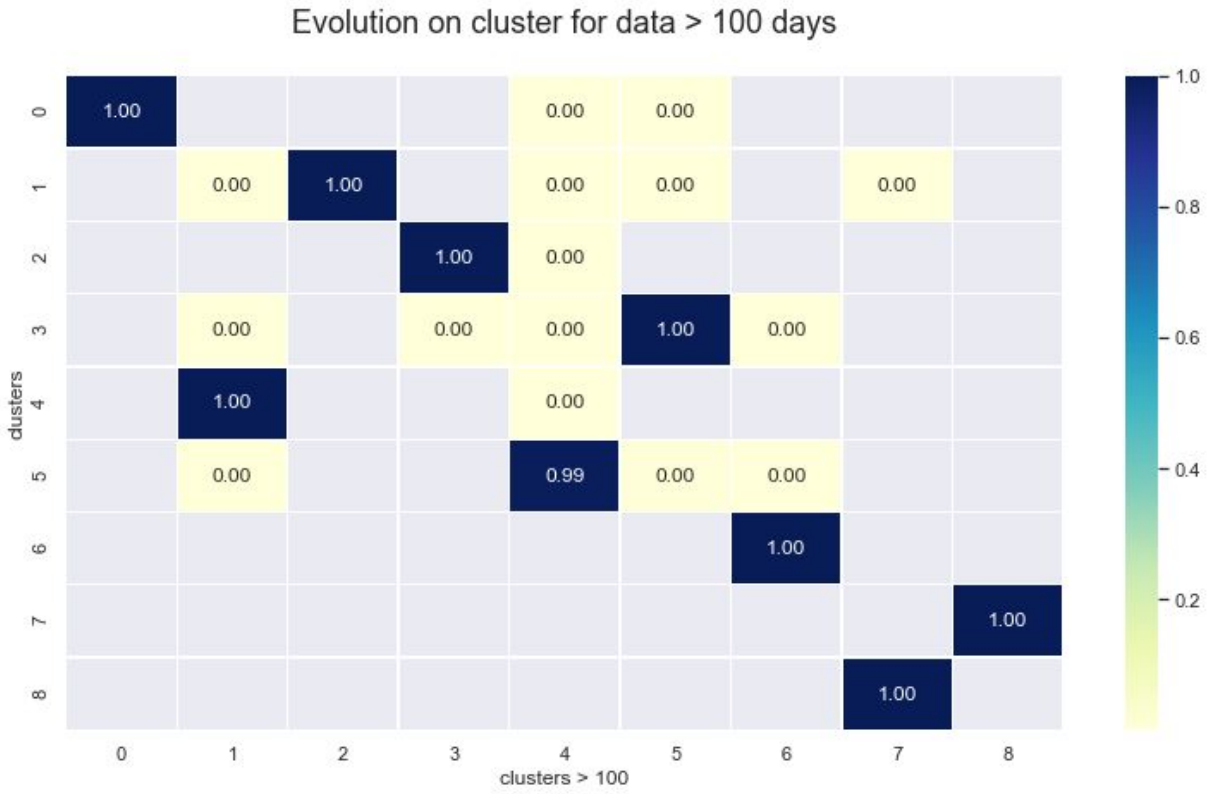
Stabilité dans le temps



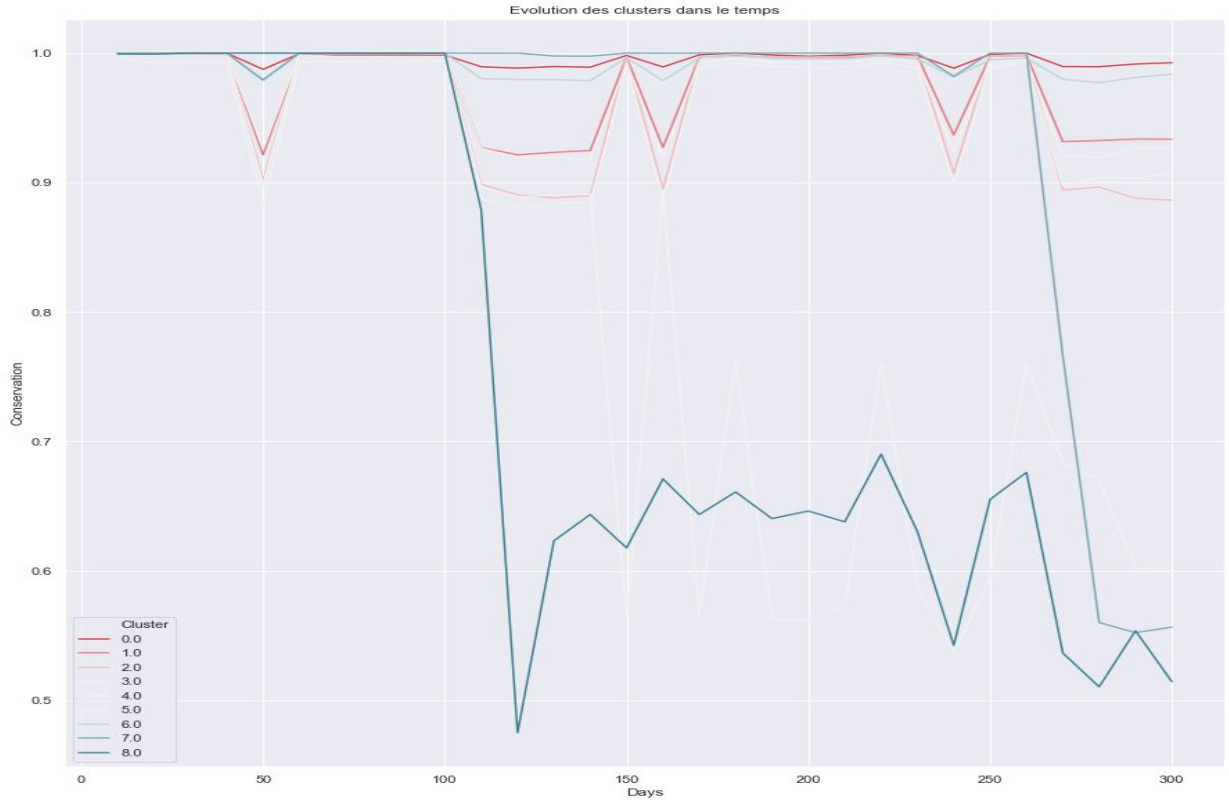
Stabilité dans le temps



Stabilité dans le temps



Stabilité dans le temps



Conclusion

Conclusion

- Utilisation du k-means permet de détecter quelques cluster intéressants 9 au total.
- Les Clusters sont stables sur 100 jours à priori
- Pistes d'amélioration:
  - Tenter d'autres catégorisations des produits, ou pouvoir avoir des tags depuis le site
  - Avoir un échantillon plus représentatif (actuellement la majorité des clients ont commandé une seule fois)