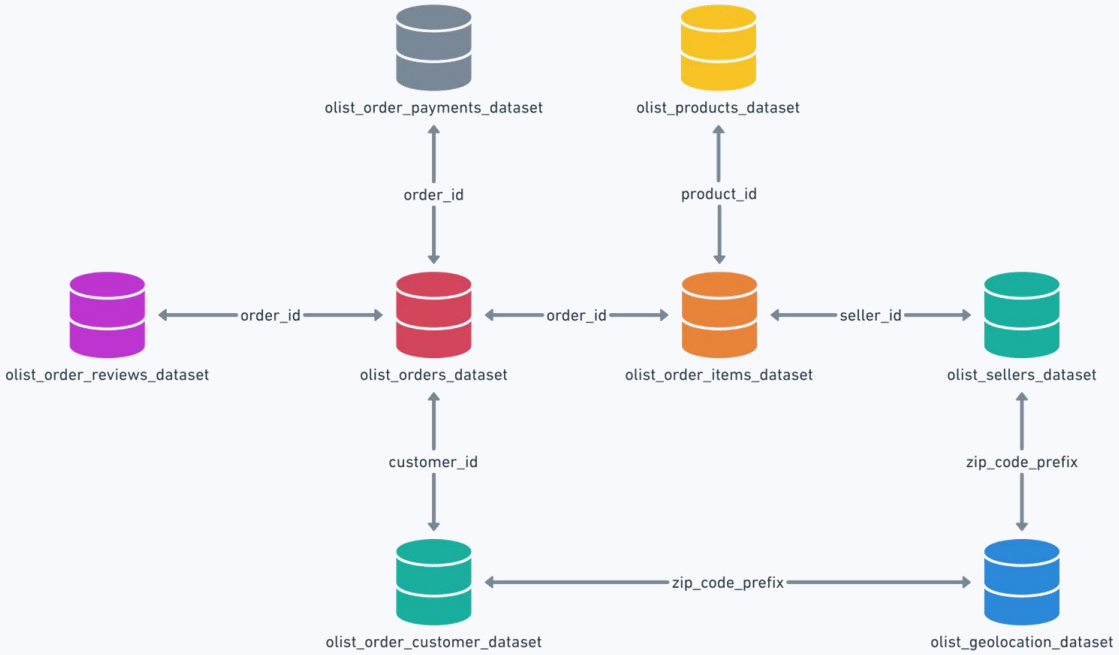
Présentation de l'appel à projets

Fournir une **segmentation des clients**

Environnement de travail

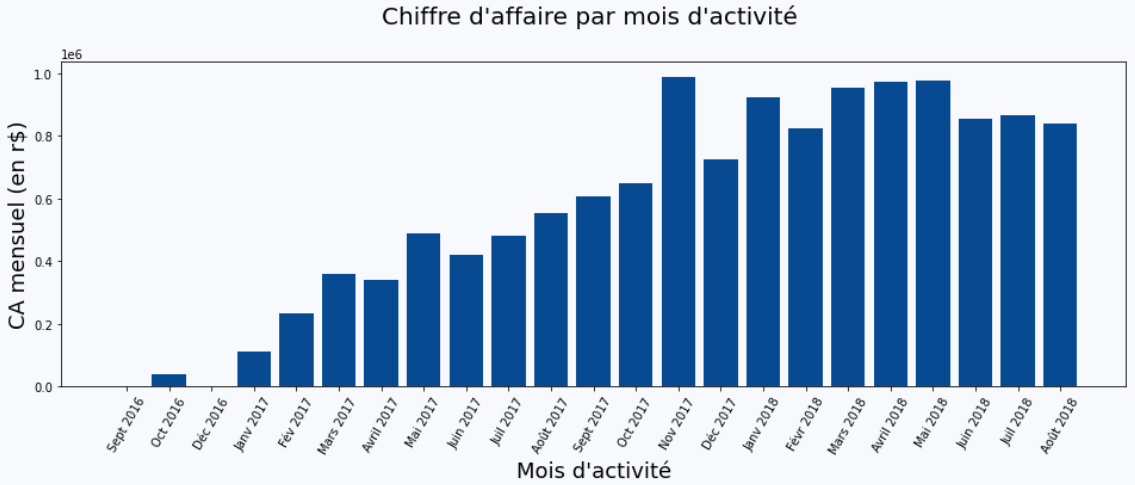
8 bases de données

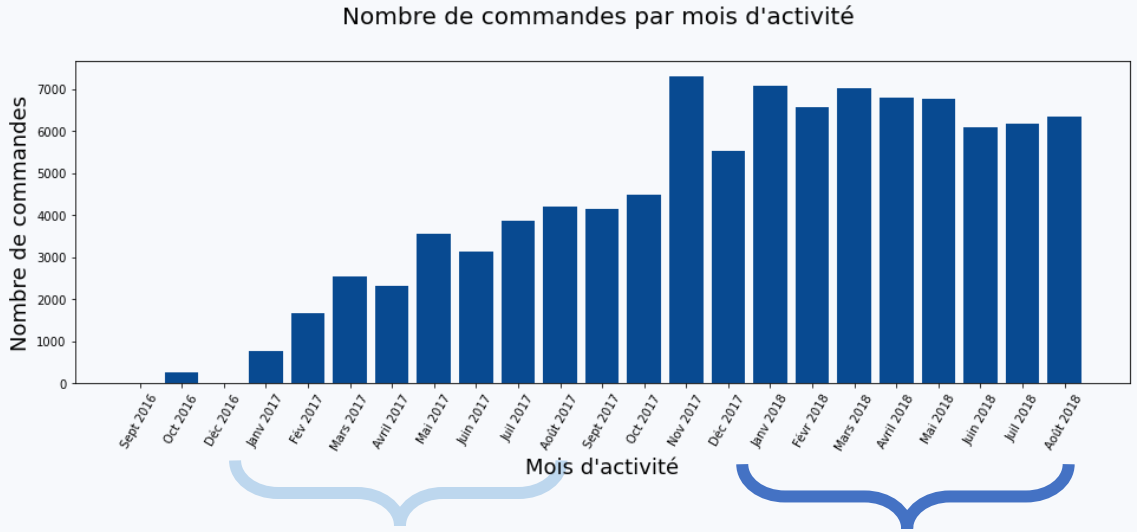


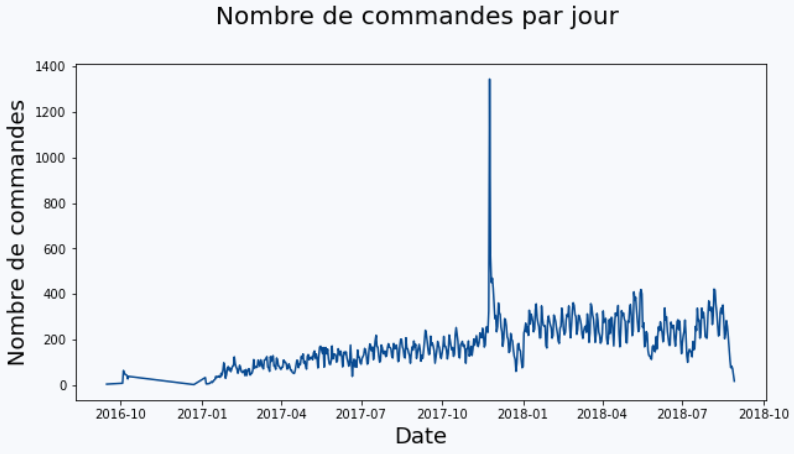
Librairies python spécialisées importées :

- Pandas
- Matplotlib
- Numpy
- Seaborn
- Scikit-learn

Description du jeu de données



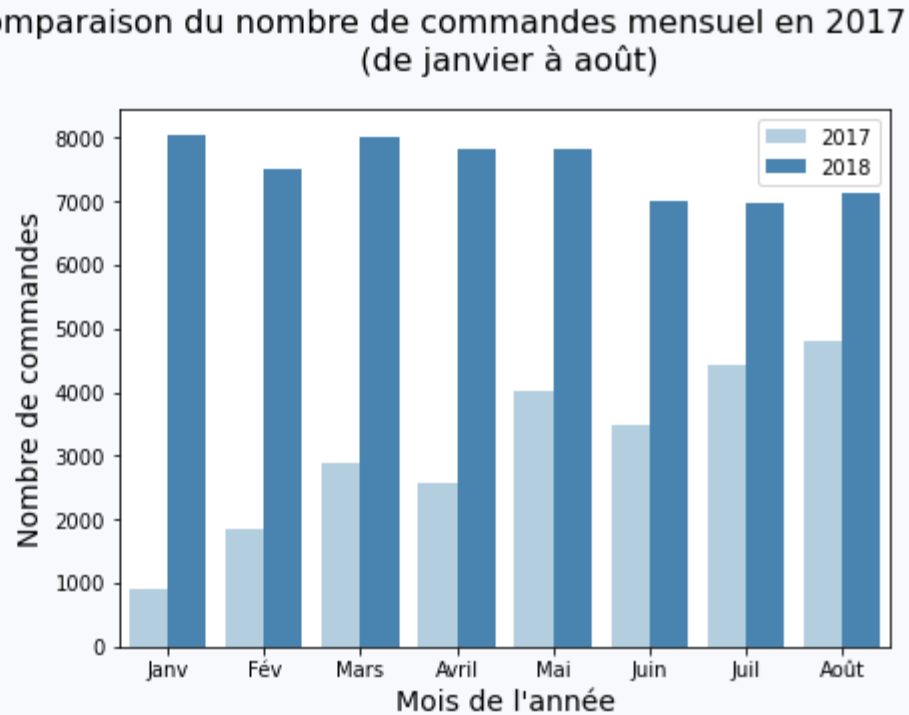




60324
commandes honorées entre Janvier et Août 2018

+141%

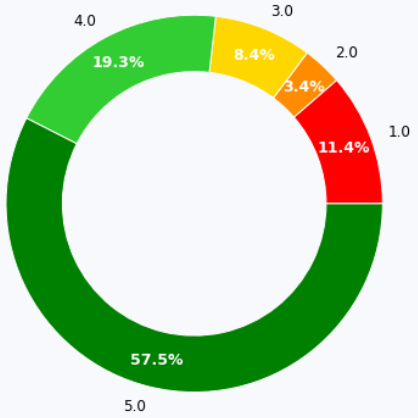
24943
commandes honorées entre Janvier et Août 2017



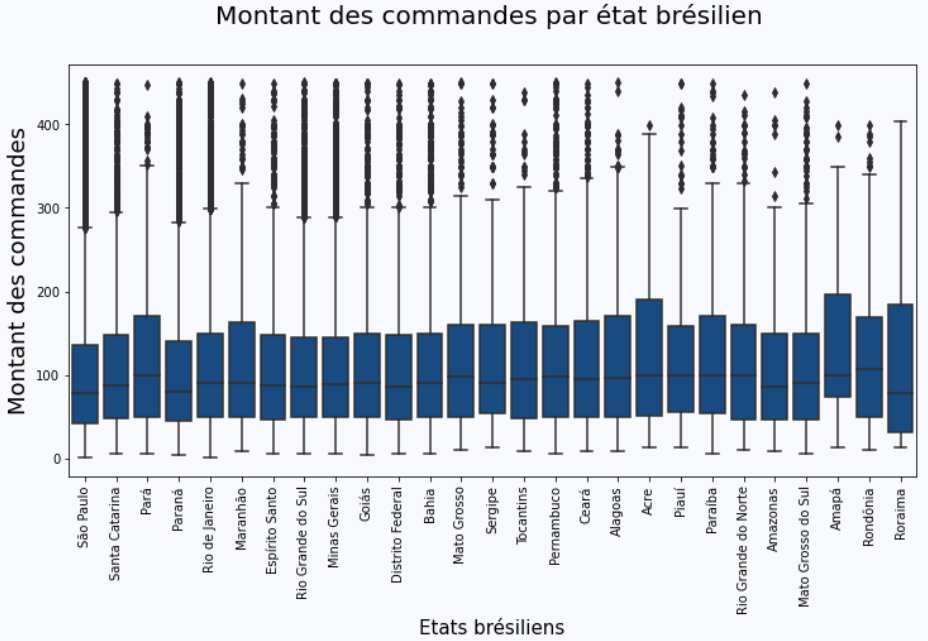
Description du jeu de données

4 fuseaux horaires

Répartition des notes des commandes dans la base de données

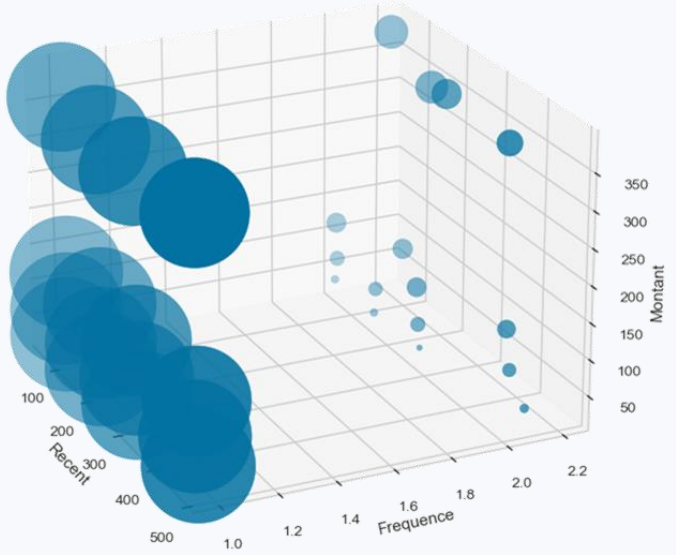


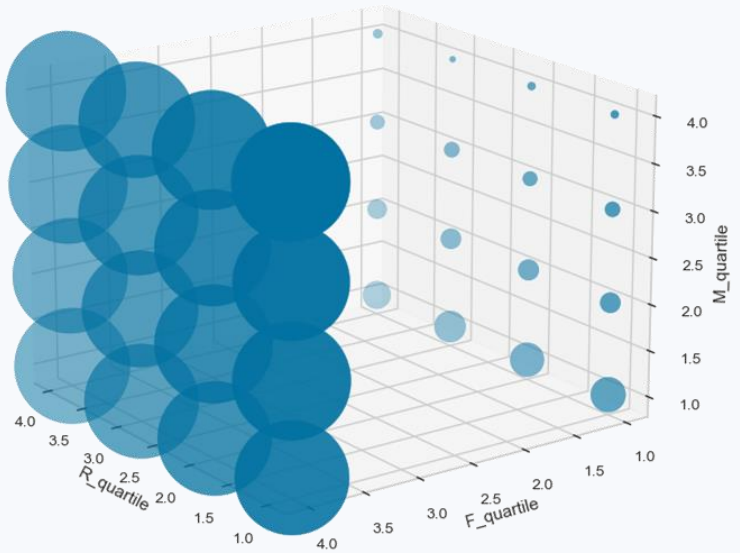
Transformation des variables qualitatives en variables quantitatives :
Distance entre le client et Sao Paolo

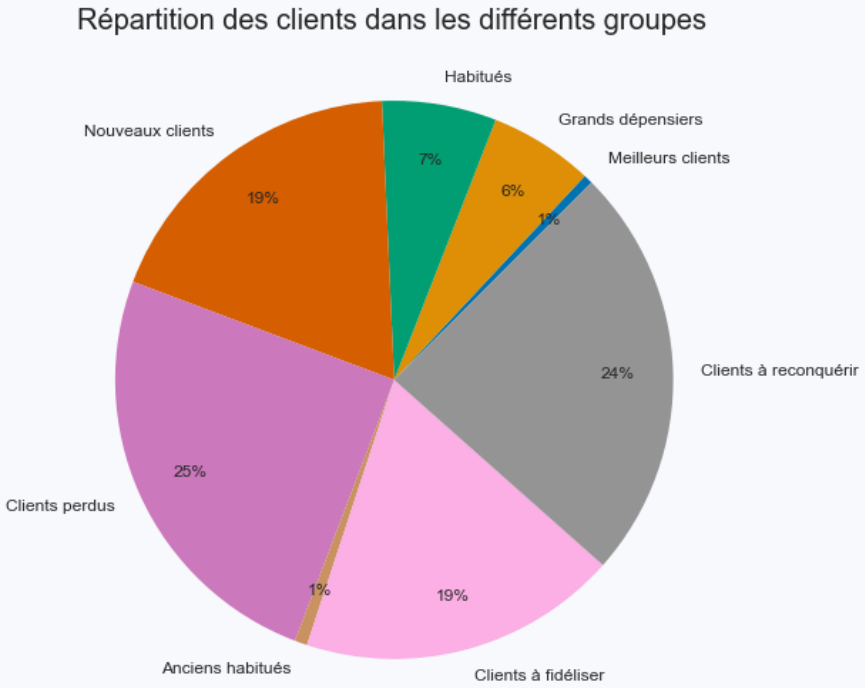


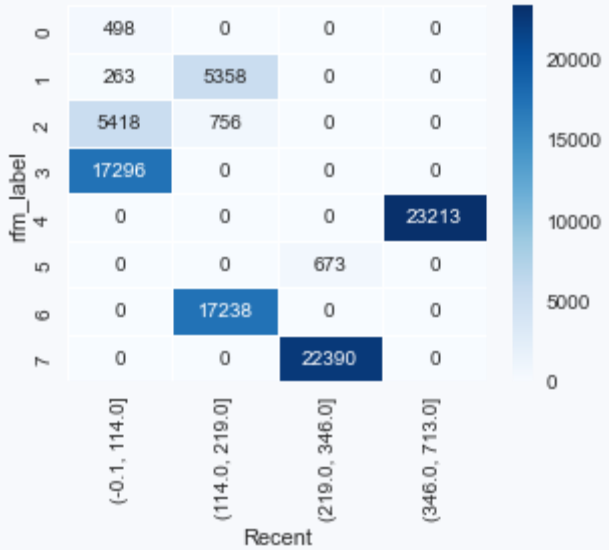
Stratégies de partitionnement

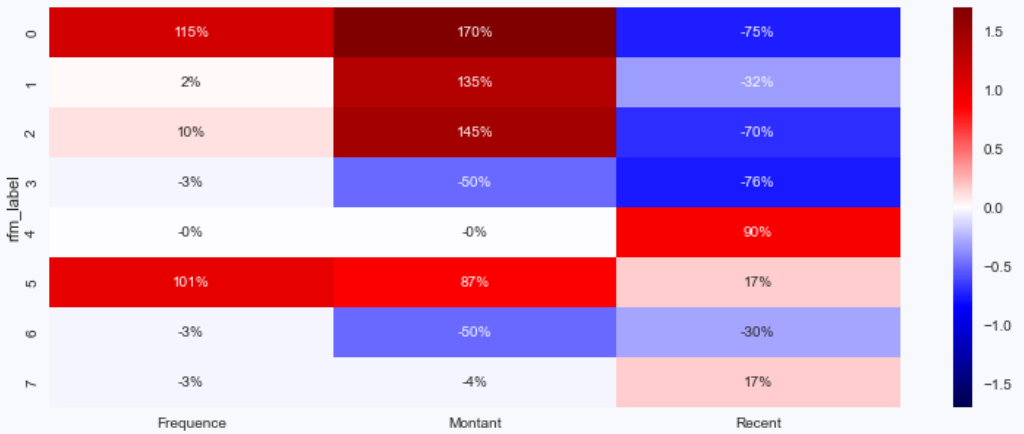
Récent, Fréquence, Montant





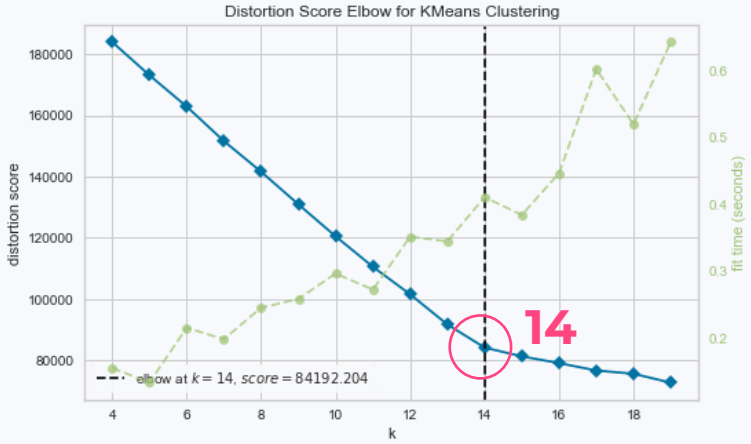




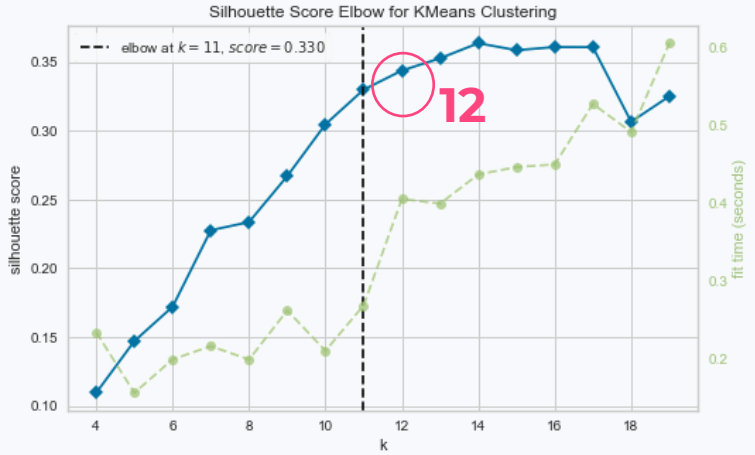


Stratégies de partitionnement

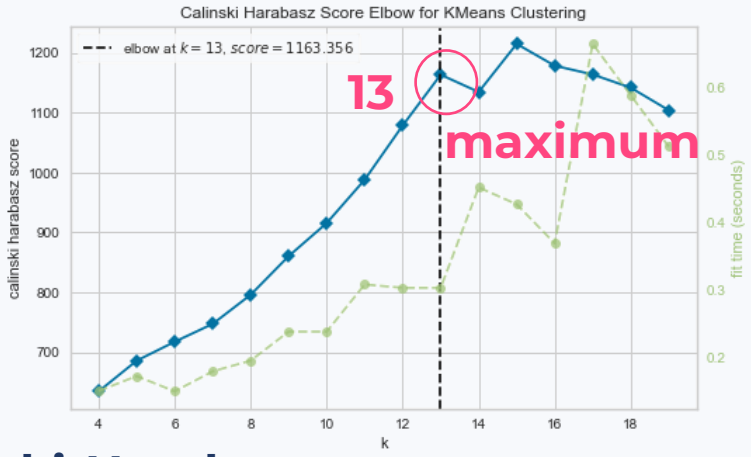
Module Yellow Brick



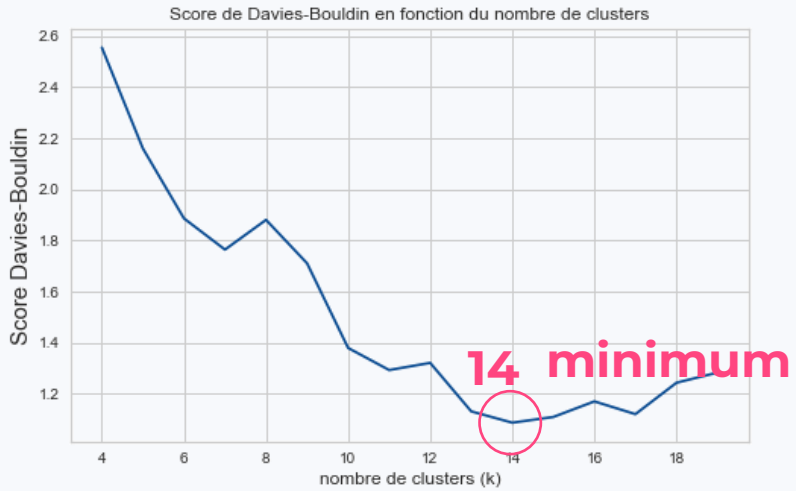
Distortion: moyenne de la somme des carrés des distances au centroïde le plus proche



Silhouette : (distances intra-cluster) – (distances au cluster different le plus proche)



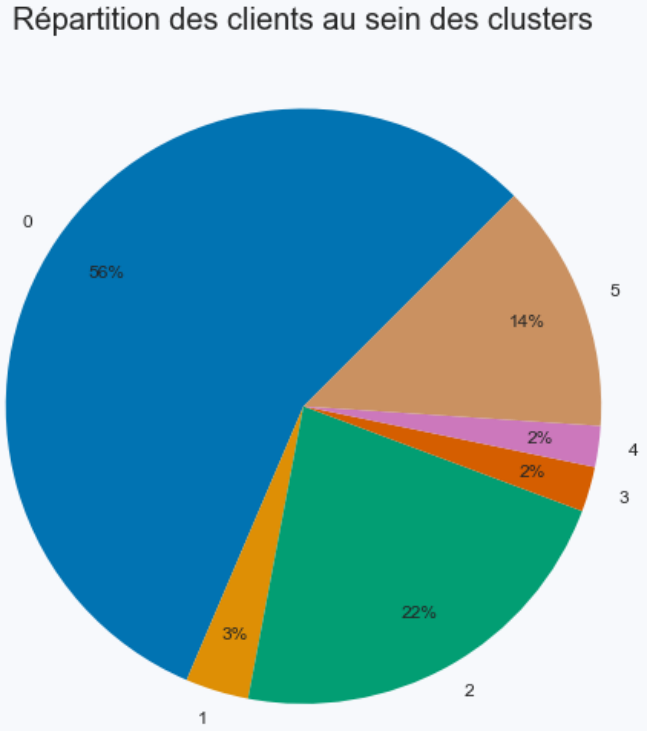
Calinski-Harabasz: variance inter-groupes / variance intra-groupe

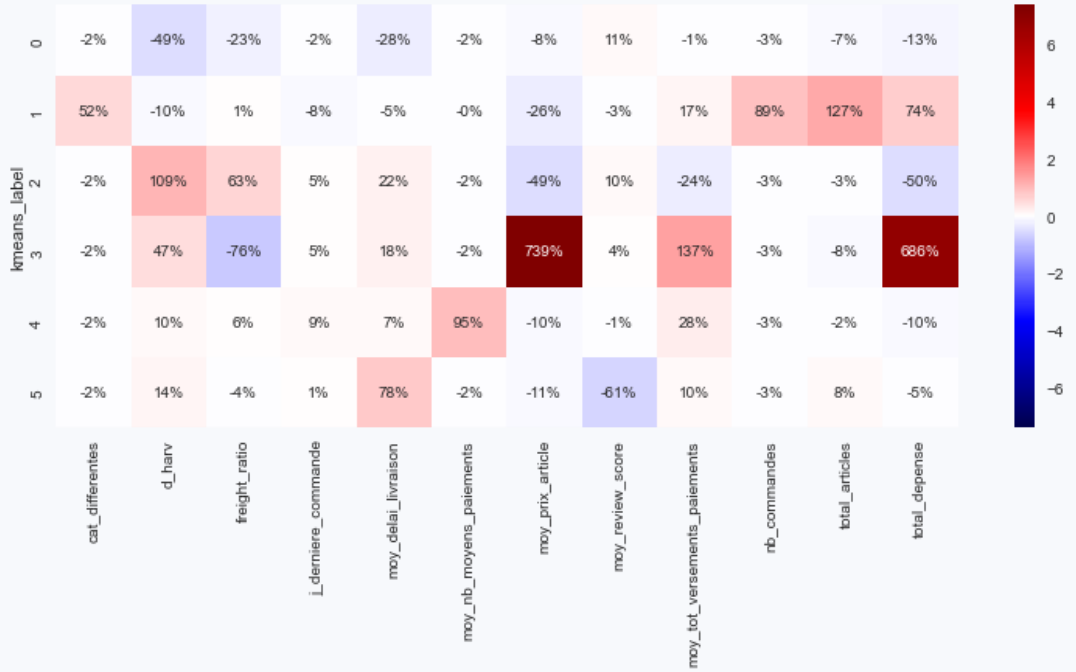


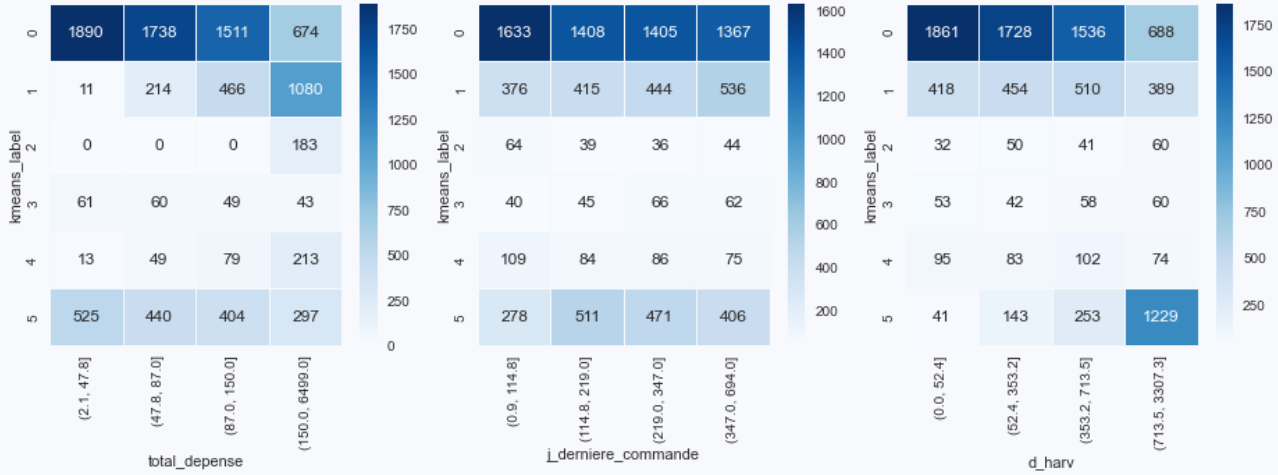
Davies-Bouldin: (distance d'un point au centre de son groupe) / (distance entre deux centres de groupe)

Stratégies de partitionnement

6 clusters

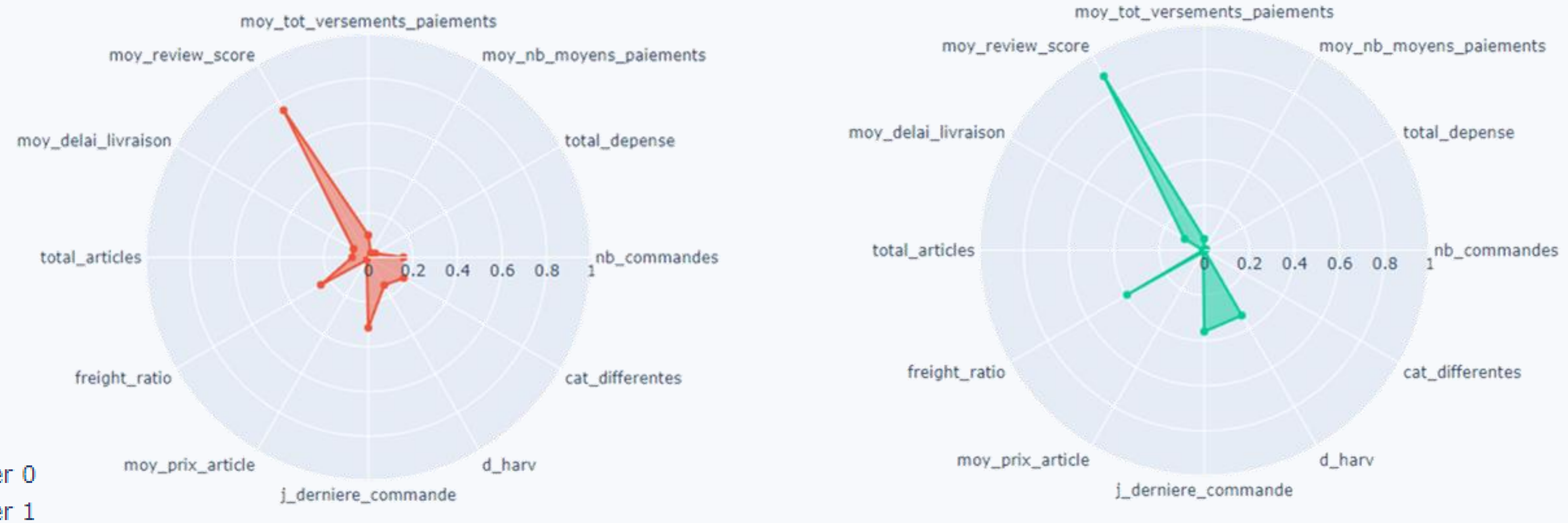




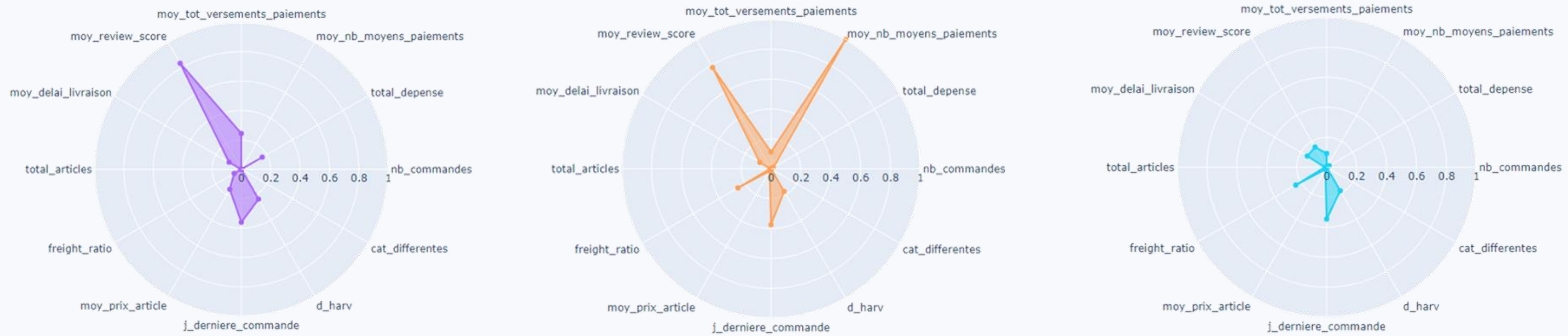


Stratégies de partitionnement

Moyennes normalisées de chaque variable pour les différents clusters

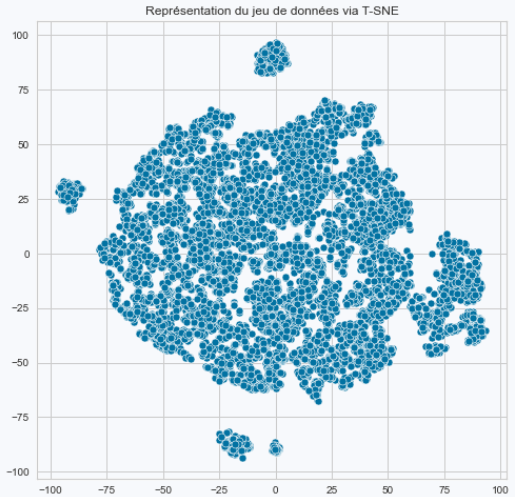


6 ~~clusters~~ → 6 types de clients

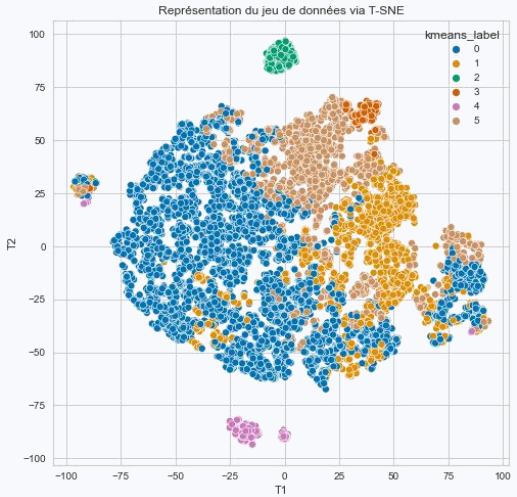


Visualisation des clients grâce à la réduction T-SNE

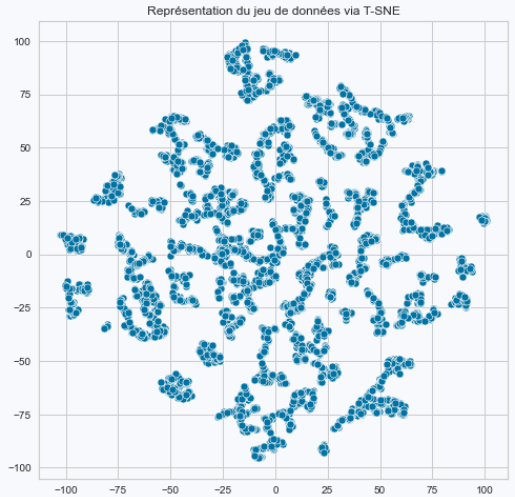
Données standardisées



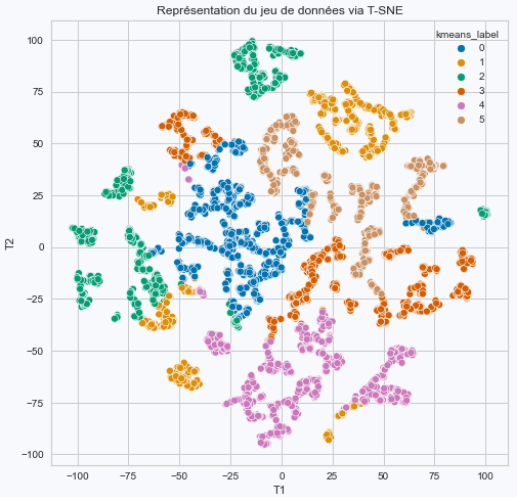
K-means



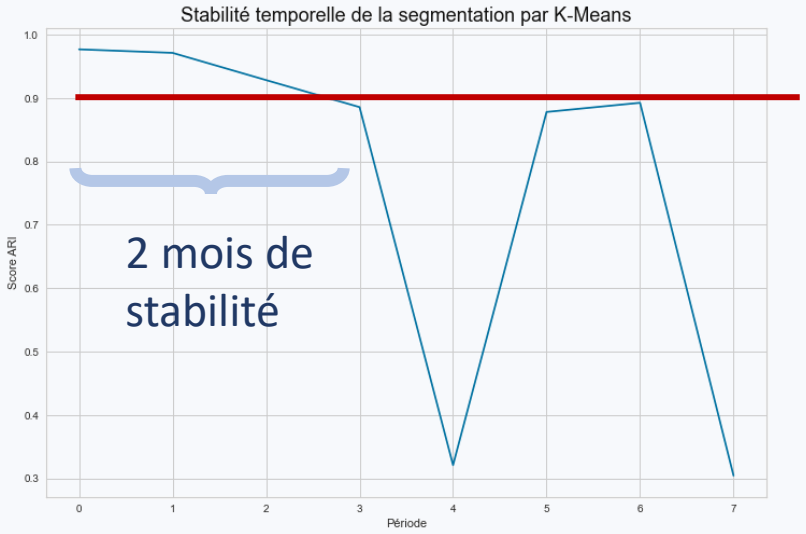
Avec PCA



K-means



Stabilité temporelle de la segmentation



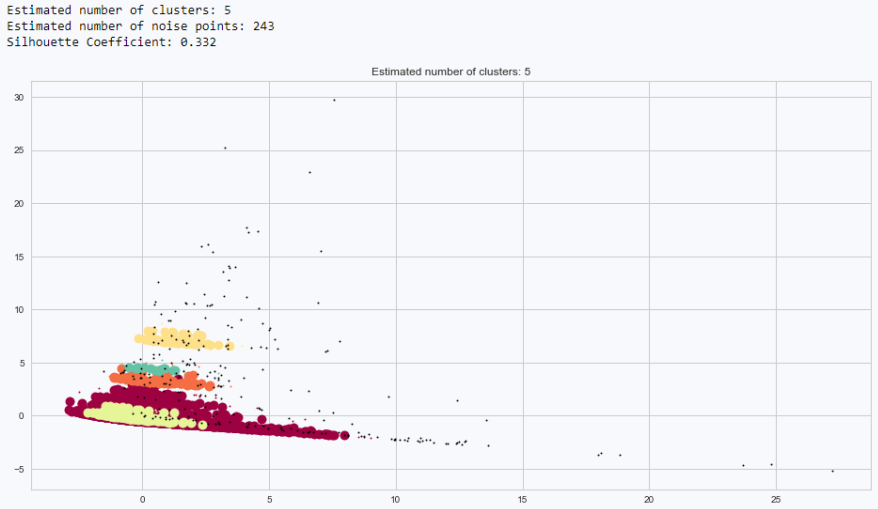
90% de similarité

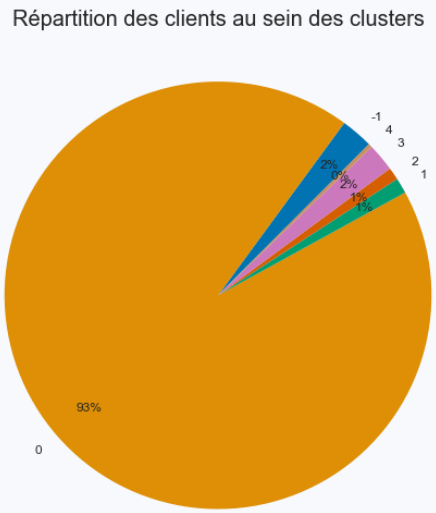
Comparaison des appartenances aux clusters à celle des clients à t0

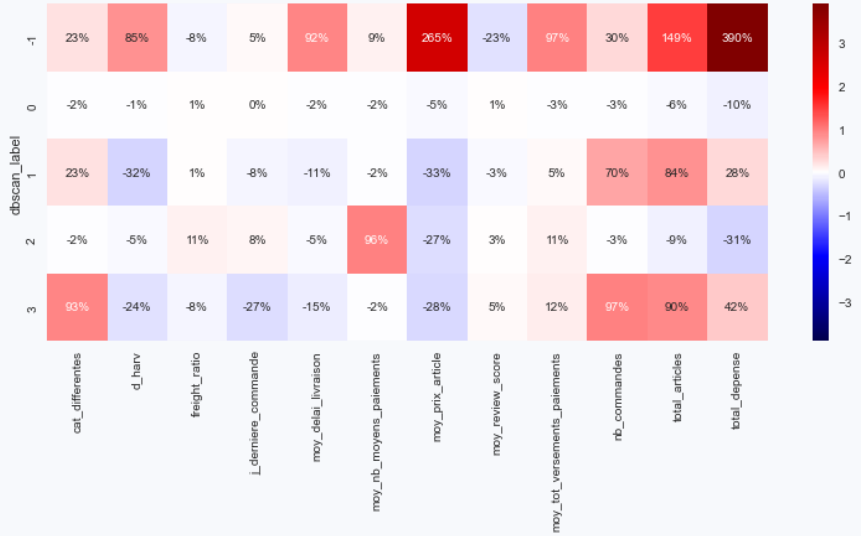
Stratégies de partitionnement

- **K-means Mini Batch** (6 groupes) : Score silhouette et Calinski-Harabasz plus faibles

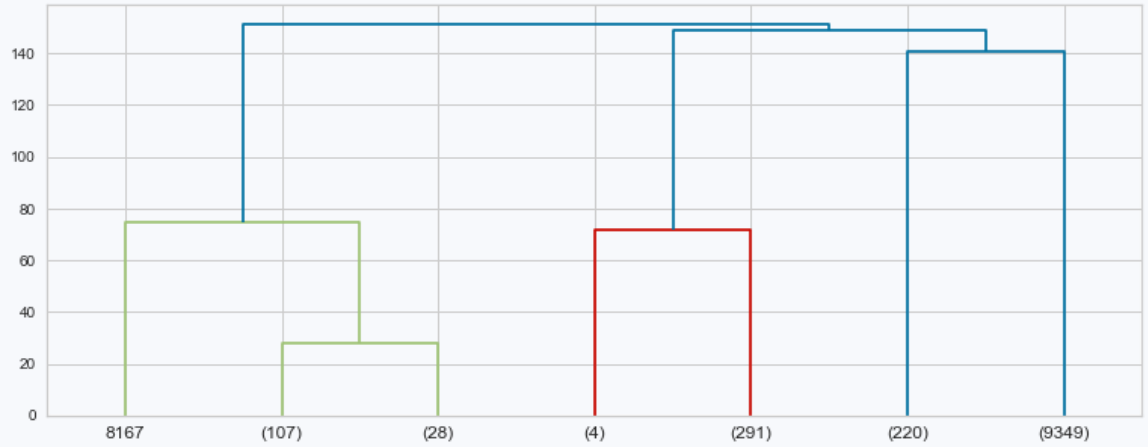
- **DSBSCAN**(5 groupes) : Score silhouette plus important mais …

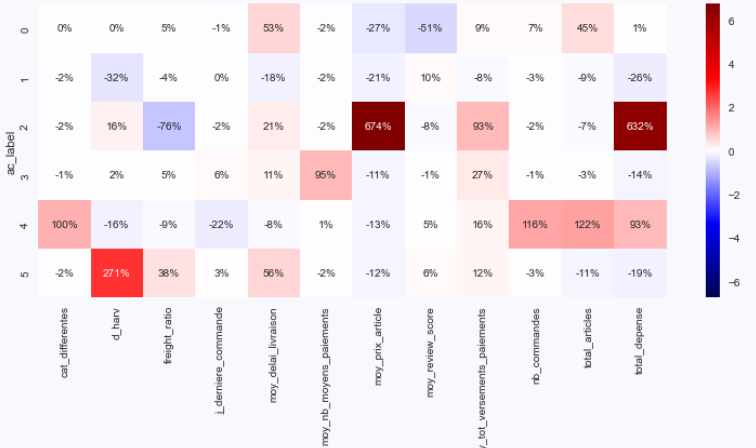






- **Agglomerative Clustering** (6 groupes) : similaire aux résultats avec les K-means





Compétences

- Ajout de nouvelles variables qui améliorent la performance du modèle de segmentation

- Evaluation du modèle :
  - ➢ Score standard : distorsion, silhouette, Calinski-Harabasz, Davies-Bouldin
  - ➢ Comparaison et visualisation des modèles (Répartition des groupes, t-SNE)

- Choix du meilleur modèle de segmentation :
  - Plus simple
  - Plus performant
  - Plus intéressant en temps (et en argent)

- Interprétabilité du modèle :
  Possibilité de faire un portrait robot du client type de chaque groupe

- Evaluation de la stabilité temporelle du modèle :
  Score ARI