Segmentation des clients de Olist

Léo Guillaume
Open Classrooms - projet 5

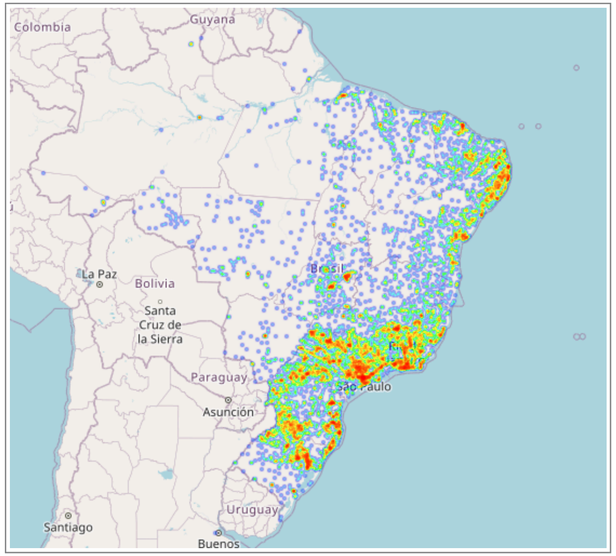
Roadmap

- Introduction
  - Le problème
  - Les données
  - La démarche
- Analyse des données
  - Préparation des données
  - Feature engineering
  - Base client
  - Client type
- Segmentation
  - La démarche
  - Tests de segmentation
  - Sélection et évaluation de la segmentation
  - Contrat de maintenance
- Conclusion

Le problème

Contexte

- Olist est une solution de vente sur les marketplaces en ligne au Brésil



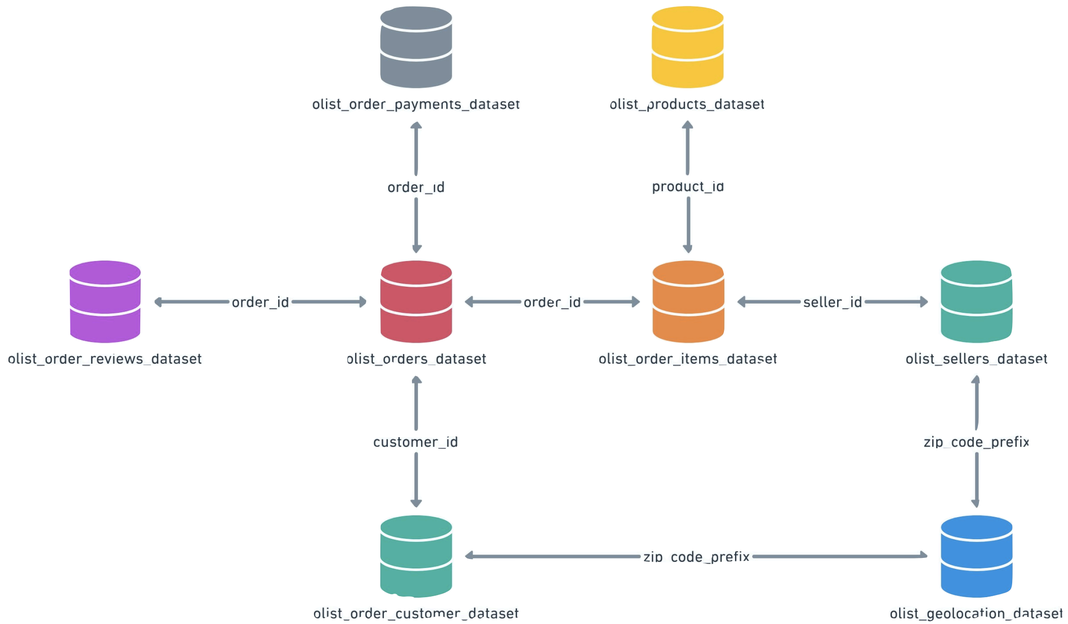
Répartition des clients

Objectifs

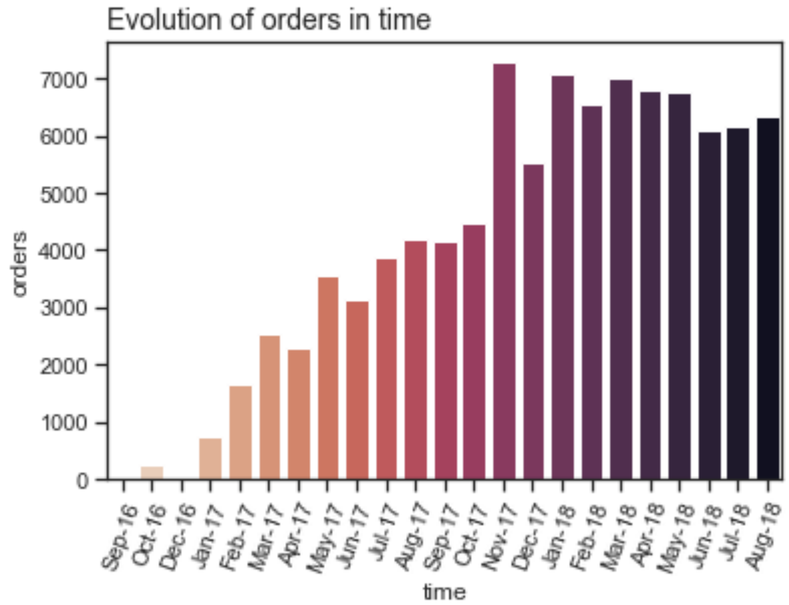
- Comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles
- Créer une segmentation des clients à des fins de campagnes marketing
- Fournir à l'équipe marketing une description actionable des segments
- Etablir la stabilité de la segmentation en vue d'une proposition d'un contrat de maintenance

Les données

- 11 tables de données regroupées par l'ID de la commande
  - 119 151 lignes et 50 variables



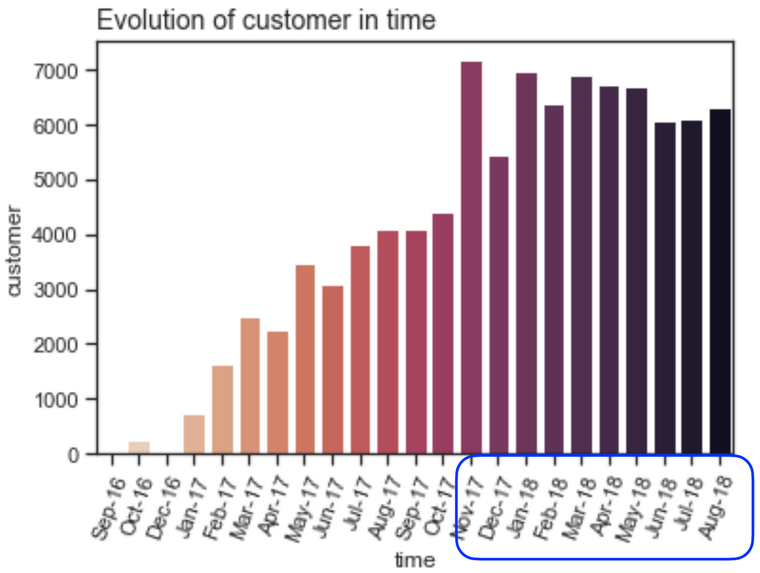
- Période de septembre 2016 à août 2018 (**24 mois**)

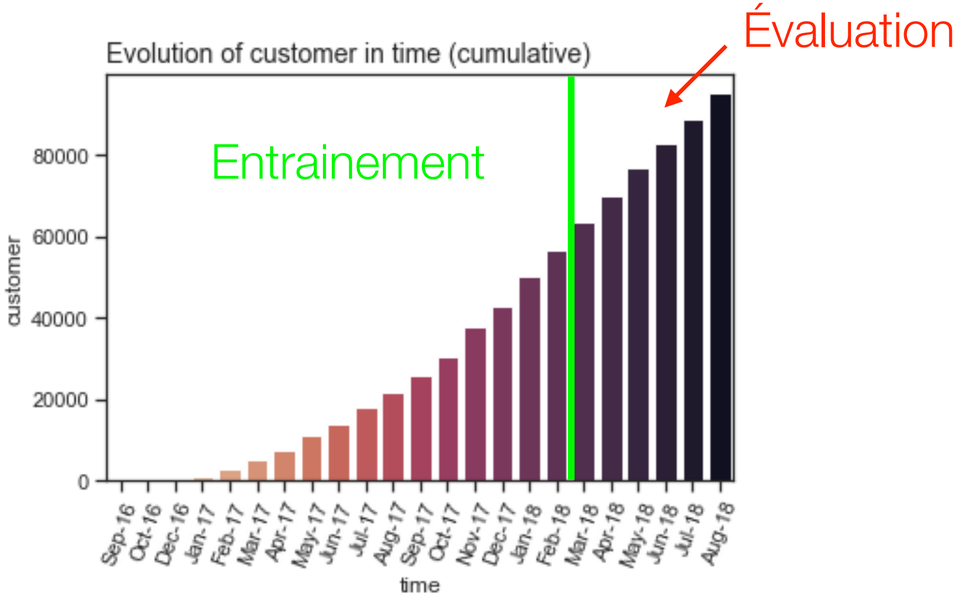


- Base de **96478 clients uniques**

La démarche

1. Groupement des données par ID de commande puis en base client
2. Analyse des données et feature engineering
3. Segmentation des clients sur 18 premiers mois
   Nombre de clients : 55363
4. Interpreter les segments en terme marketing
5. Evaluer la stabilité des segments en ajoutant des périodes de 2 mois (20, 22 et 24 mois)





Préparation des données

Normalisation

- Conversion des textes en minuscule et suppression des accents
- Conversion des variables temporelles en datetime
- Conversion des variables textuelles en variables dichotomique ou en variable numérique
  Exemple : customer city → customer pop

Remplacement

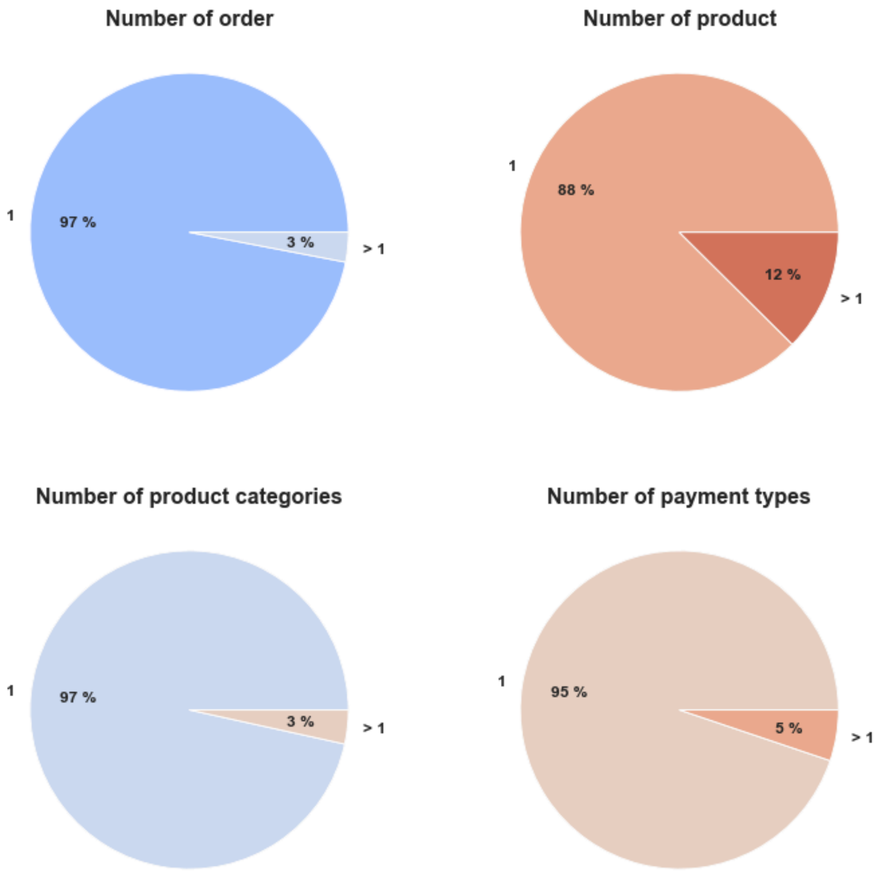
- Remplacement des valeurs manquantes par la moyenne de la variable
- Groupement des 71 catégories produit en 9 catégories selon la nomenclature de Statista
  Leading product categories bought online among internet users in the United States as of November 2017, by gender

Sélection

- Groupement par commandes
- Suppression des commandes non délivrées
- Suppression des lignes de commande dupliquées en raison de review ID différents
- Suppression des lignes de commande dupliquées en raison de type de paiement différents

Analyse préliminaire

- 90 % des clients :
  - Ne passe qu'**une commande**
  - Ne commande qu'**une produit**
  - Ne commande d'**une catégorie de produit**
  - Ne paie qu'avec **un type de paiement**
- Les variables suivantes sont peu discriminantes

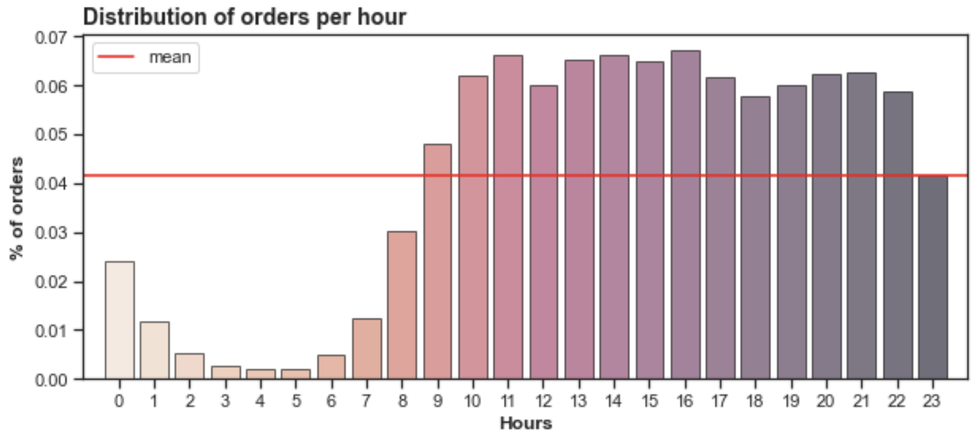


Feature engineering

Heure d’achat

- Deux périodes d’achat : la nuit (avant 9h) et le jour (après 9h)

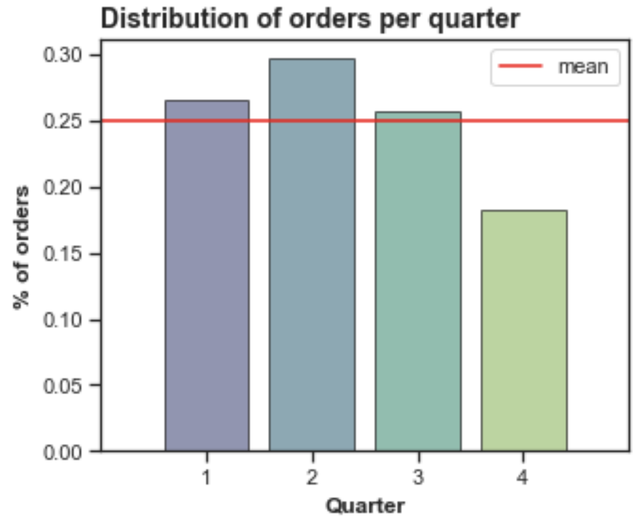
➡ Création d’une variable dichotomique



Trimestre d’achat

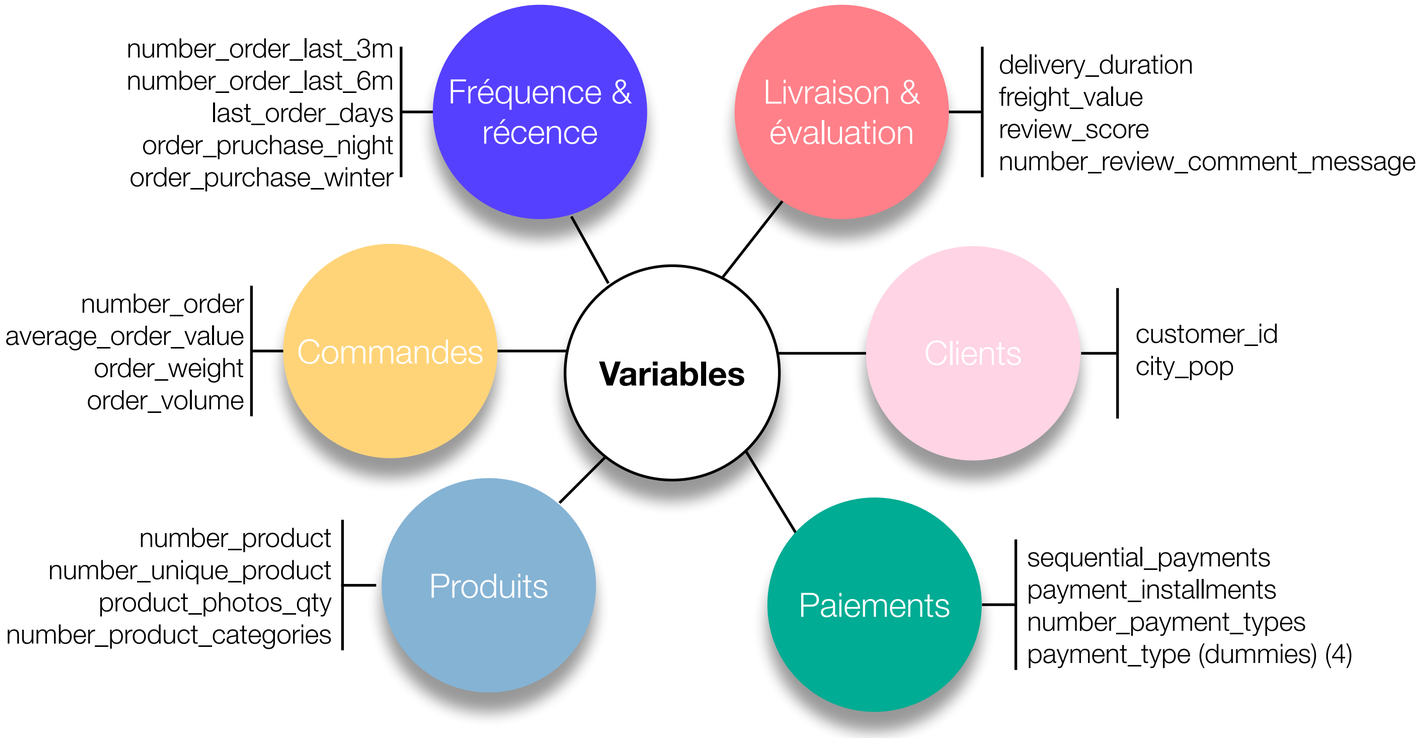
- Baisse des achats sur le dernier trimestre (hiver)

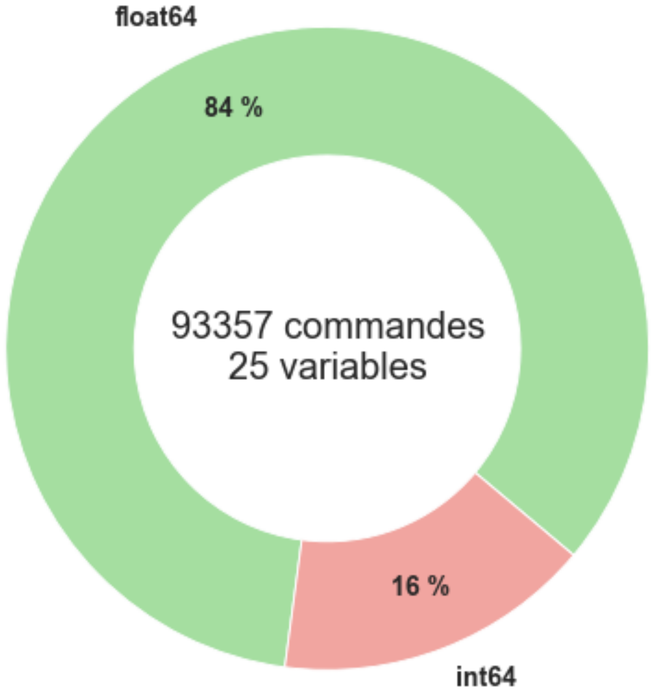
➡ Création d’une variable dichotomique



La base client

- Regroupement des données par ligne de commande

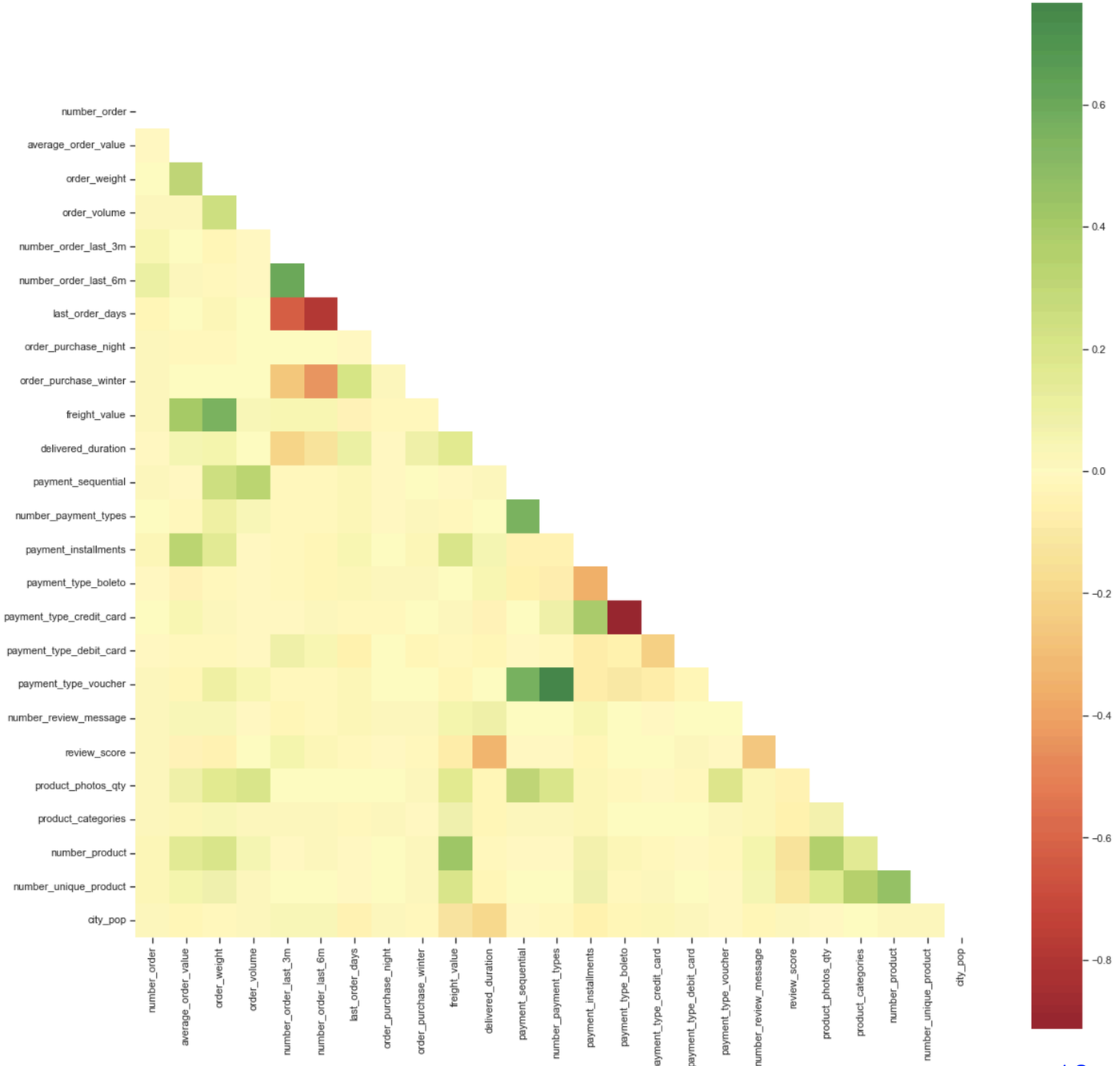




Corrélations

- Le review score est négativement corrélé avec le délai de livraison
- Les frais de livraisons sont positivement corrélés avec le poids de la commande
- Les échelonnements de paiement (payment_installments) sont positivement corrélés avec la valeur de la commande
- La population de la ville est négativement corrélée avec le délai de livraison

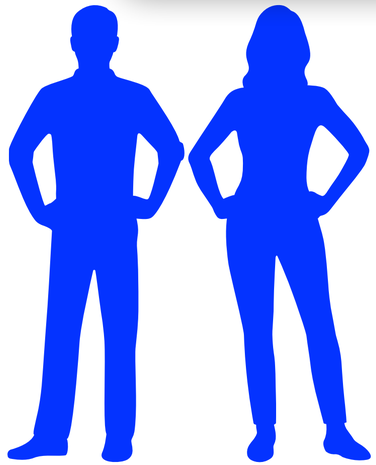
➡ Ne pas associer ces variables pour la segmentation



Client type

- Une commande
- Un produit
- Une catégorie
- Habite à Soa Paulo

- Panier moyen de 2,4 R$
- Frais de transport de 2,5 R$
- Payé en une fois par carte de crédit

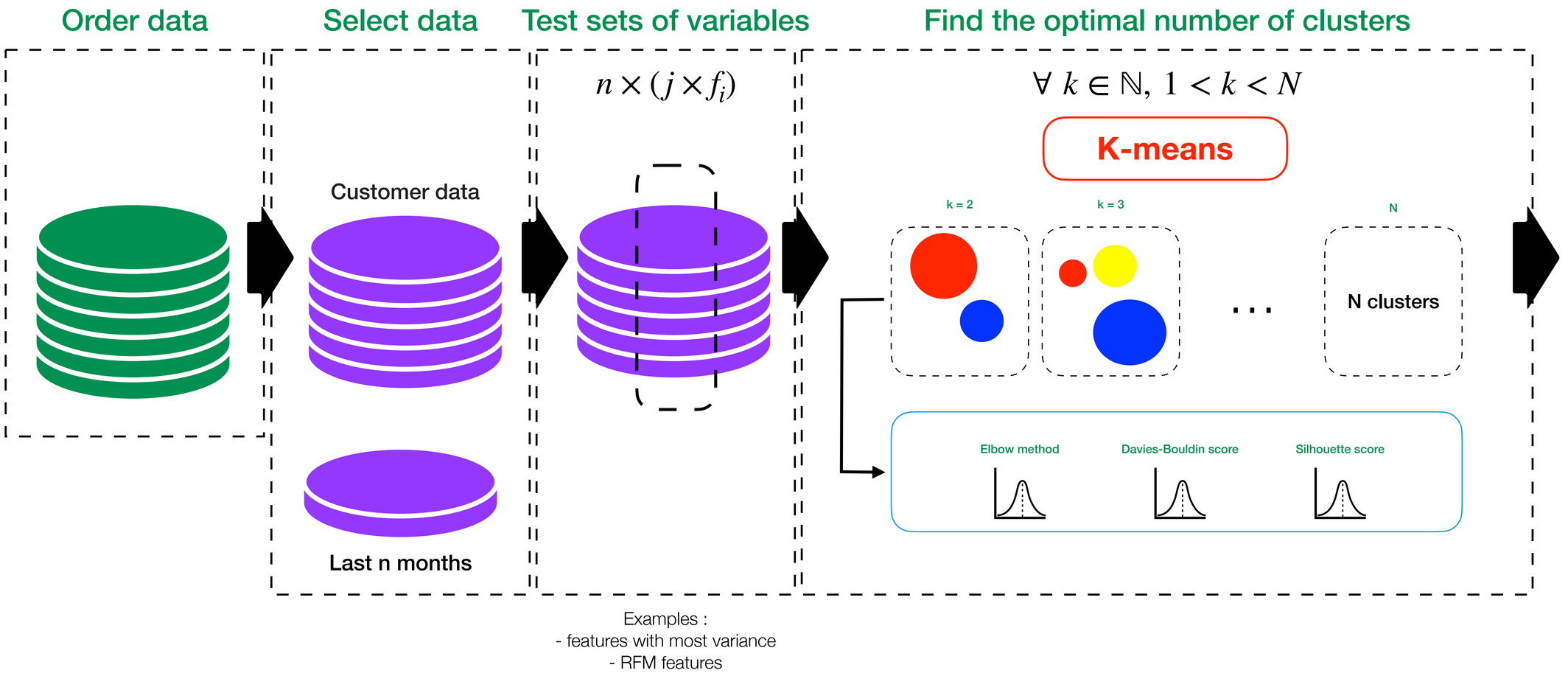


- Livré en 12 jours
- Ne laisse pas de commentaire de review
- Attribut une note de 4 sur 5

- Moins d'une commande au cours des 6 derniers mois
- 236 jours depuis la dernière commande
- Passe ses commandes en journées et avant le dernier trimestre

- Commande des produits avec plus de 2 photos
- Commande de 2,5 kg et de 160 L

Démarche de segmentation

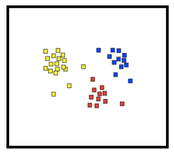


Démarche de segmentation

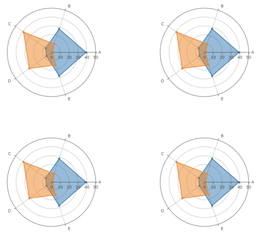
Find the best clustering

Silhouette plot

PCA



Radar plots

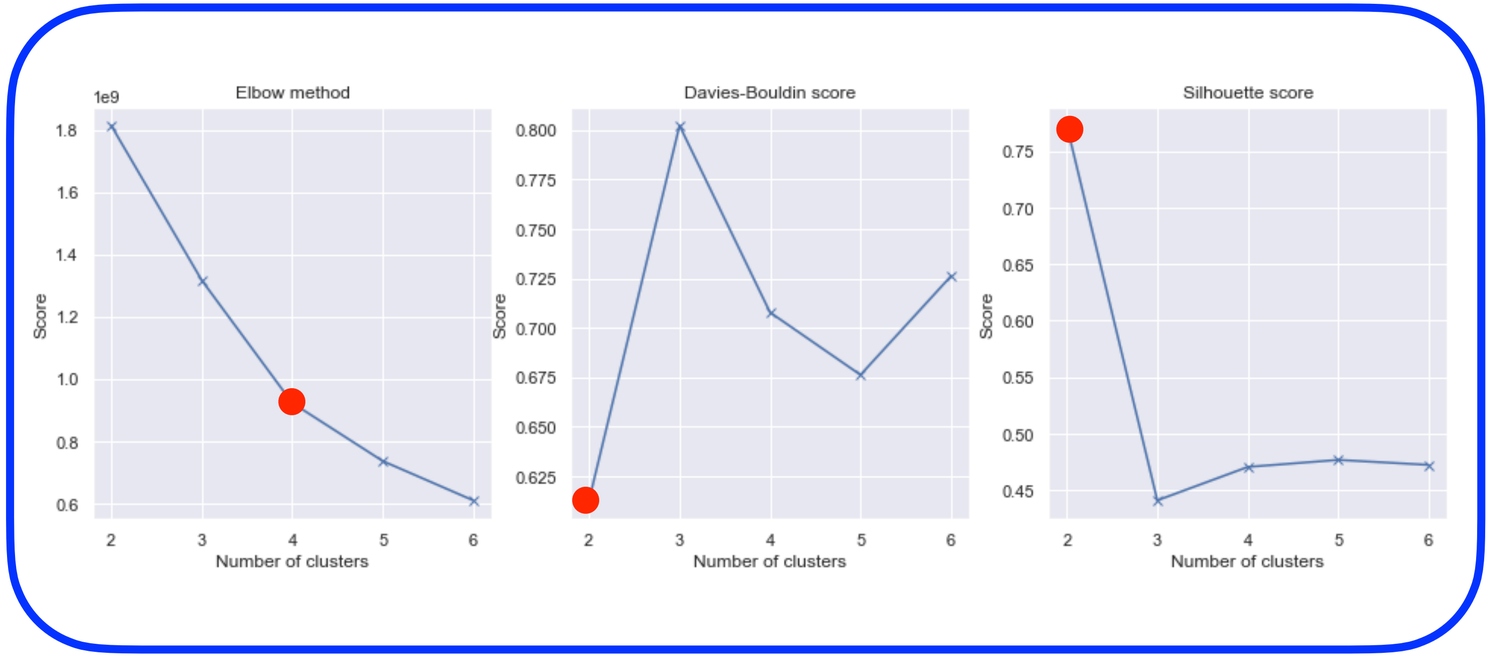


Evaluation and monitoring contract

- Evolution of clusters proportions in time (with new client)
- Evolution of the silhouette scores in time (with new client)
- Evolution of the ARI scores in time (with same clients)

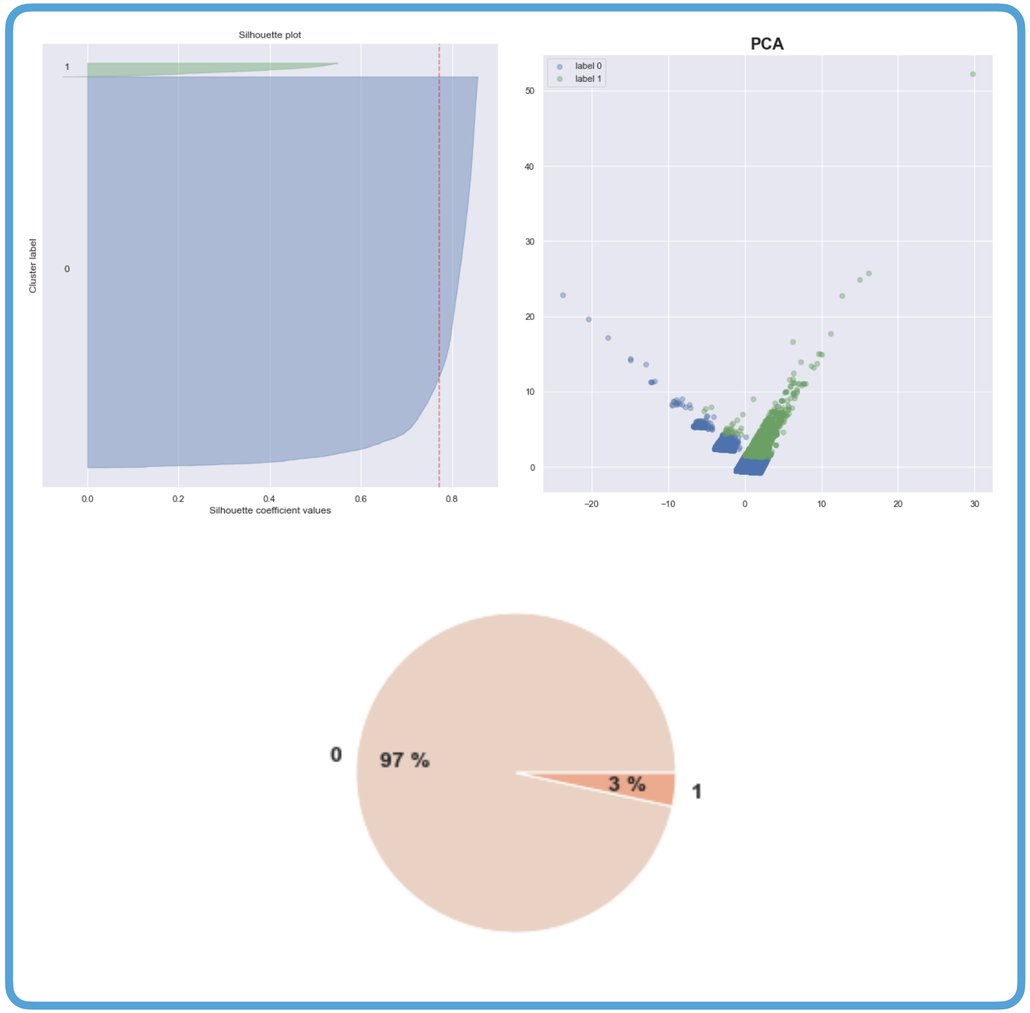
Modélisation : variables RFM

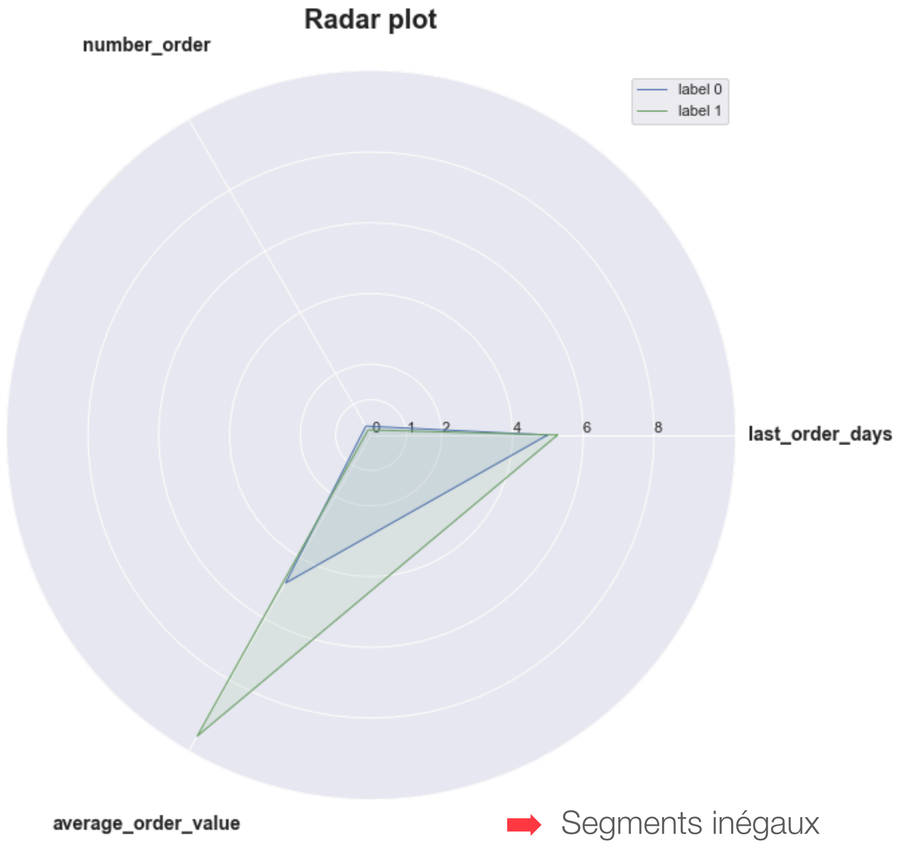
- Récence : date de la dernière commande → last_order_days
- Fréquence : fréquence des commandes → number_order
- Monétaire : panier moyen → average_order_value



➡ Test avec 2 clusters

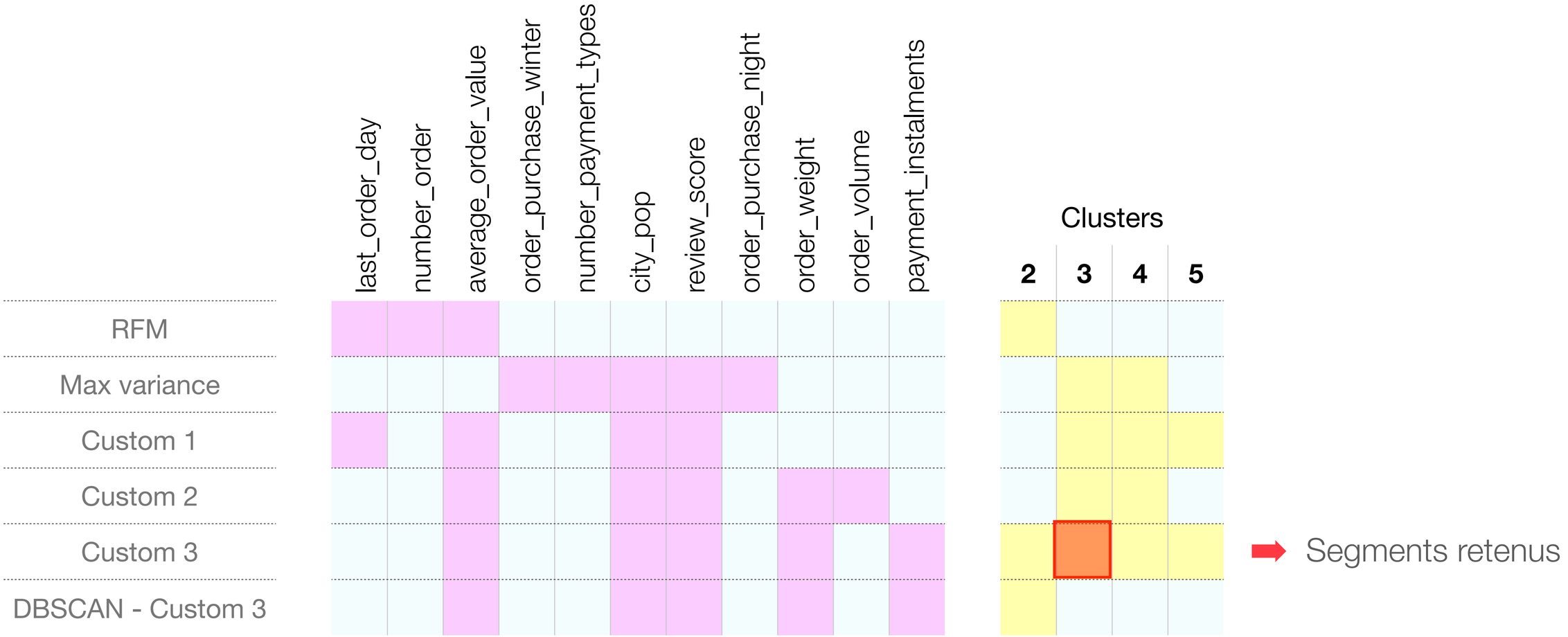
Modélisation : variables RFM





- Segments inégaux
- Variable non discriminate (number_order)
- Segments trop homogènes (pas actionnable)

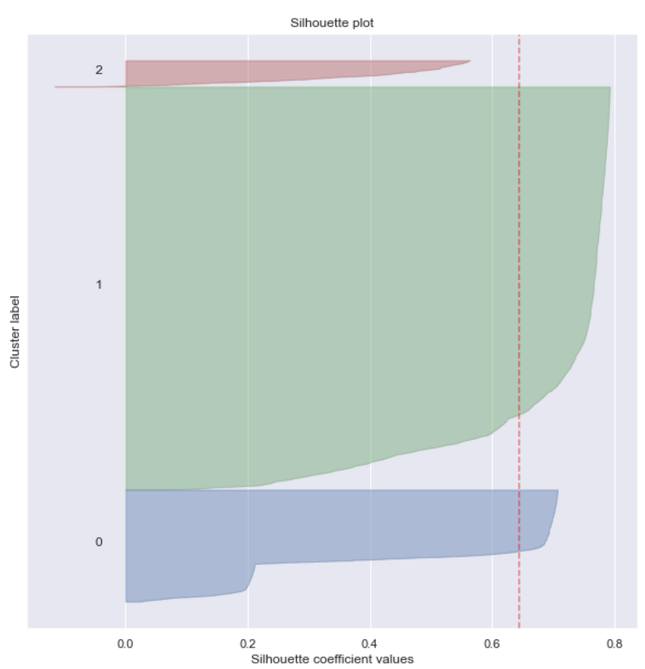
Bilan des modélisations



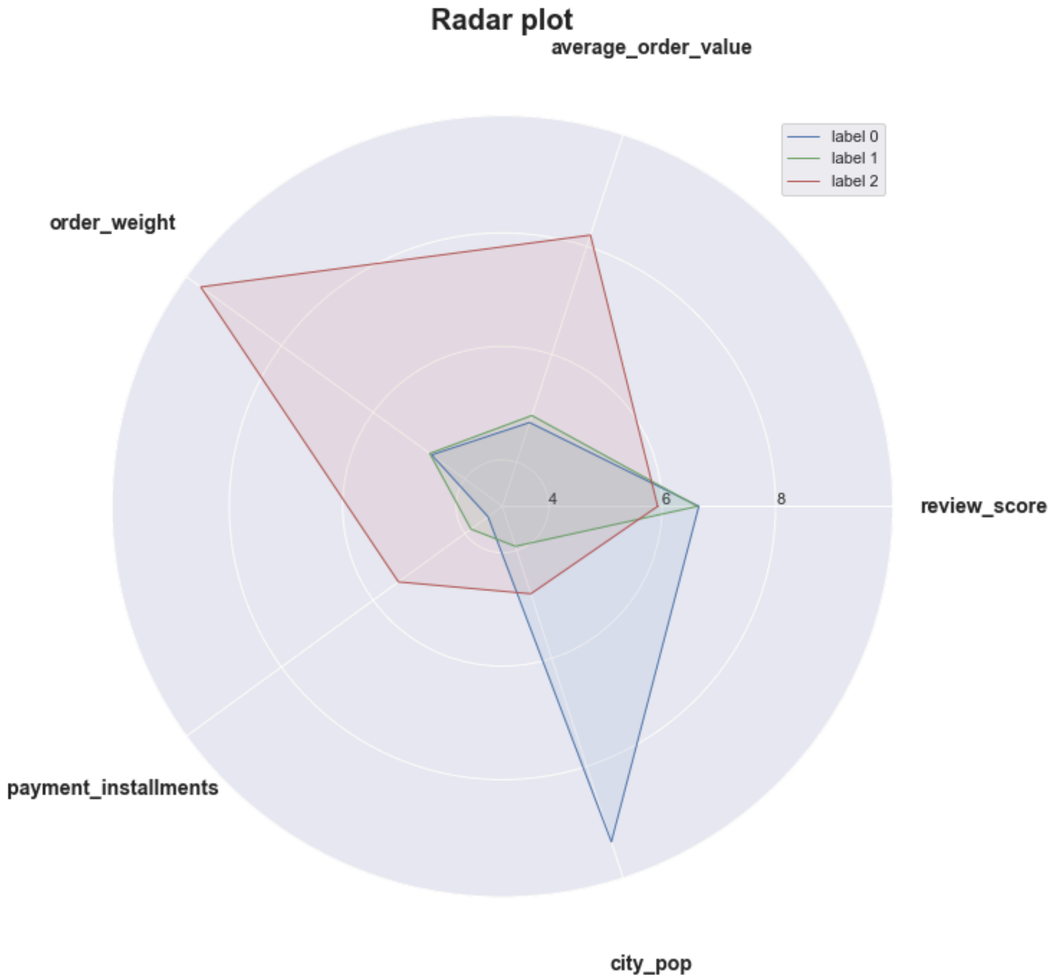
Modélisation sélectionnée

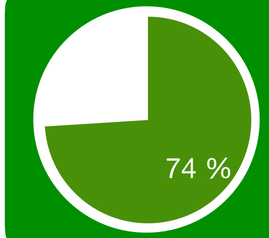
- Review score
- Average order value
- Order weight
- Payment instalments
- City population

➡ **3 segments actionnables en termes marketing**



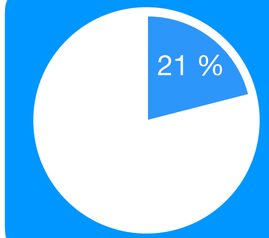
Description des segments





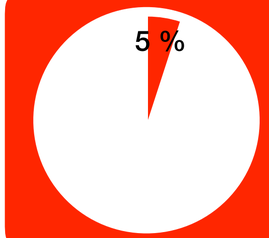
RURAUX

- Dépense peu
- Isolé géographiquement
- Relativement satisfait



CITADINS ÉCONOMES

- Dépense peu
- Habite les grandes villes
- Relativement satisfait



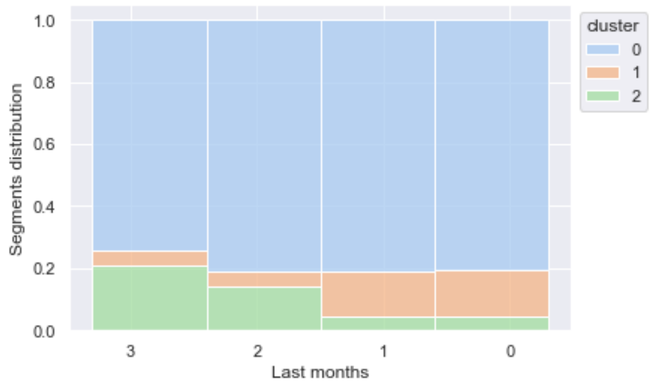
CITADINS DÉPENSIER

- Dépense beaucoup
- Commande importante
- Moyennement satisfait

Contrat de maintenance

Clusters répartitions

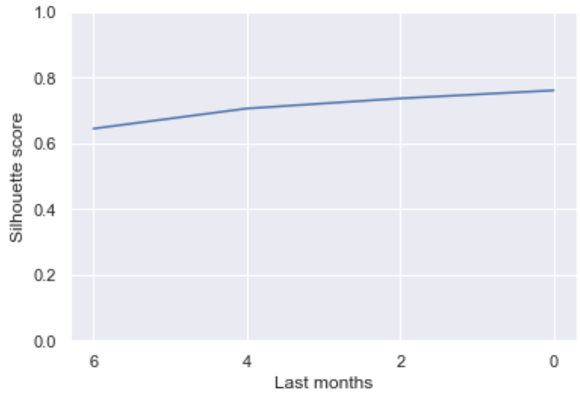
- Stabilité du segment 0 dans le temps
- Les segments 1 et 2 inversent leur répartition



Silhouette score

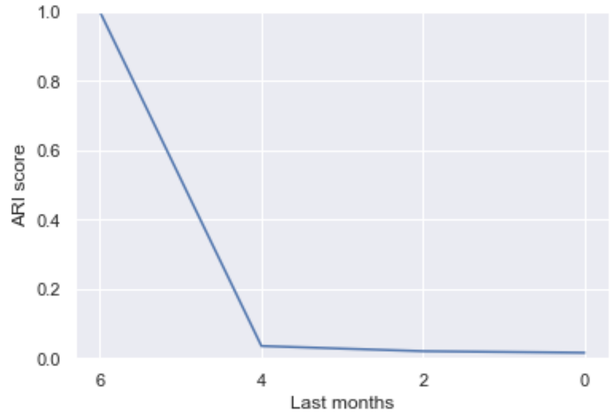
- Augmentation du silhouette score avec la taille de l'échantillon

➡ Segments plus robustes



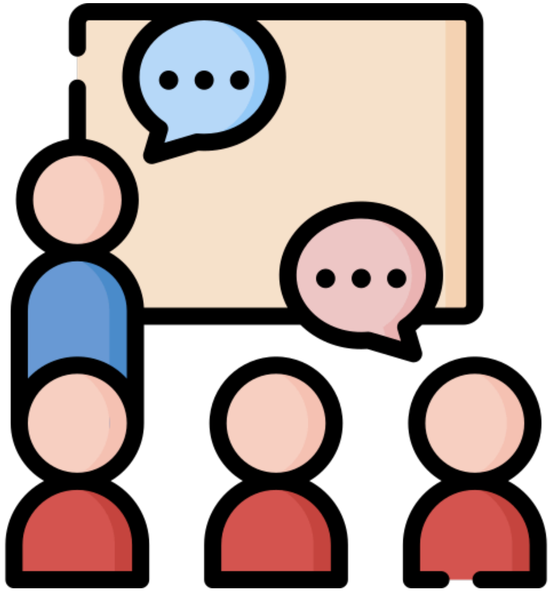
ARI score

- Effondrement du score ARI
- Le modèle est peu stable



Conclusion

- **Réalisations**
  - Nettoyage des données et créations de variables discriminantes
  - Segmentation en 3 groupes de client actionnables
  - Evaluation de la stabilité des segments
- **Amélioration et perspectives**
  - Score silhouette faible, segments inégaux et modèle peu stable
  - Améliorer le jeu de donnée avec plus de données historiques pour augmenter les clients revenants



Merci pour votre attention.