Soutenance
Projet 5

Segmenter des clients d'un e-commerce

Formation Data Scientist
avec
OPENCLASSROOMS

olist
empowering commerce

Ordre du jour:

1) Introduction
   - ⇒ *Problématique*
   - ⇒ *Découverte des jeux de données*
2) Transformation et exploration des données
   - ⇒ *Feature engineering et jointure des bases de données*
   - ⇒ *Analyse exploratoire*
3) Différents modèles étudiés
   - ⇒ *Rappel de la problématique et choix de différents modèles*
   - ⇒ *Etude des modèles*
   - ⇒ *Modèle retenue et études des clusters*
4) Etude du délai de maintenance pour le modèle sélectionné
5) Conclusion

Introduction

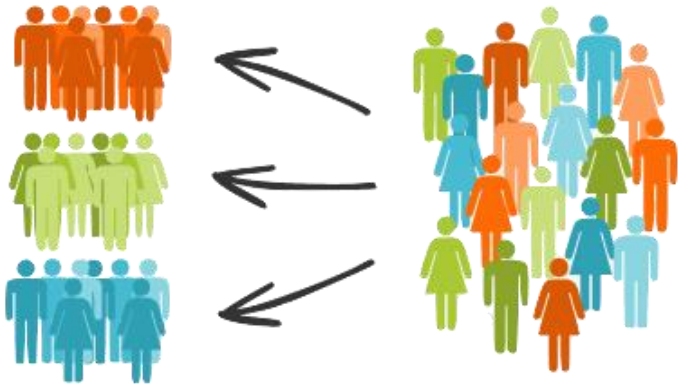
$\Rightarrow$ *Problématique*

$\Rightarrow$ *Découverte des jeux de données*

Introduction: Problématique

Olist: Entreprise brésilienne proposant une solution de vente sur les marketplaces en ligne

Mission: Fournir une segmentation des clients sur des données passées, avec une description des clusters et une proposition des contrat de maintenance



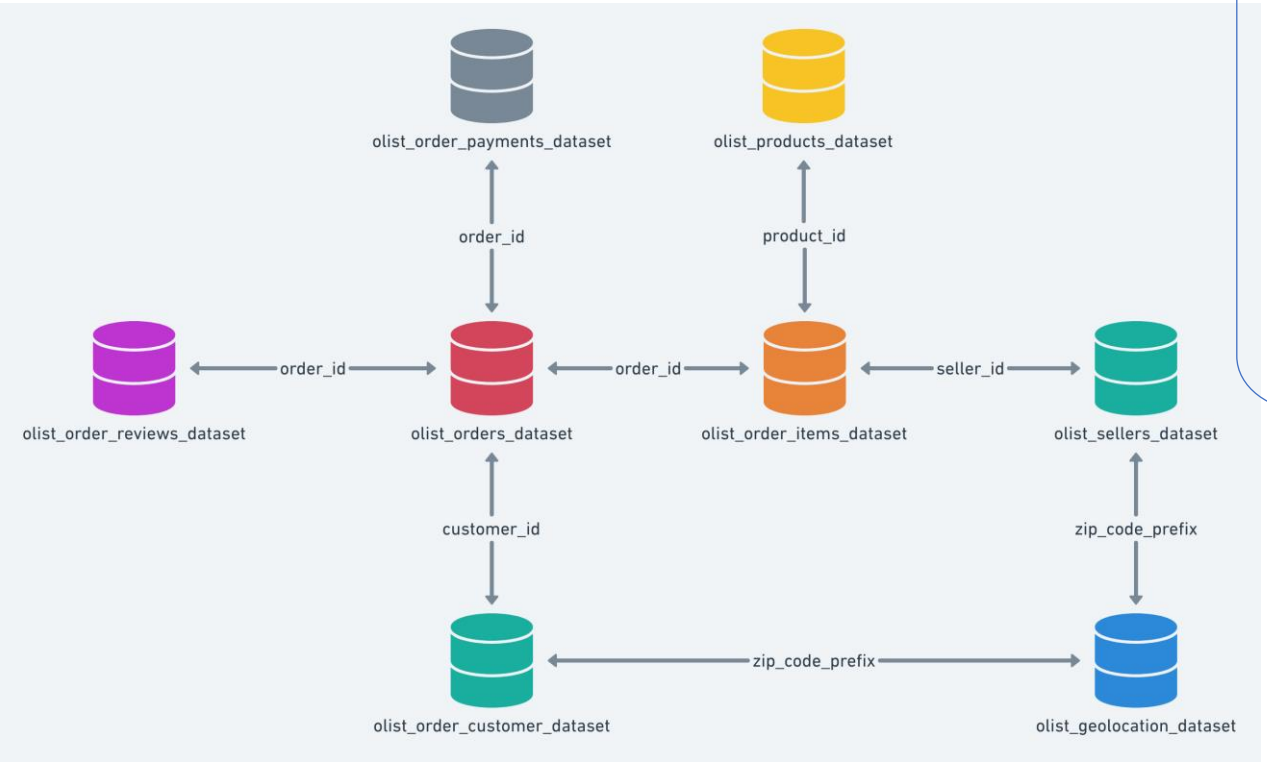
Solution:

- Etude des données et joindre par client
- Etude de modèle de clustering
- Choix d'un modèle et description des clusters
- Etude délai de maintenance sur la stabilité du modèle

Introduction:

Découverte des jeux de données

9 bases de données



- olist_customer_dataset: 99 441 clients 5 variables (info client…)
- olist_orders_dataset: 99 441 commandes 8 variables (info commandes..)
- olist_order_items_dataset: 112 650 produits des commandes 7 variables (info composition commande…)
- olist_order_payment: 103 886 lignes 5 variables (info payement commandes)
- olist_order_reviews: 99 224 reviews 7 variables (détails des reviews)
- olist_products: 32 951 produits 9 variables (descriptions des produits)
- Product_category_name_translation: 71 lignes 2 variables (traduction portugais-anglais)
- olist_sellers: 3 095 vendeurs 4 variables (info vendeurs)
- olist_geolocation: 738 332 localisations 5 variables

Transformation et exploration des données

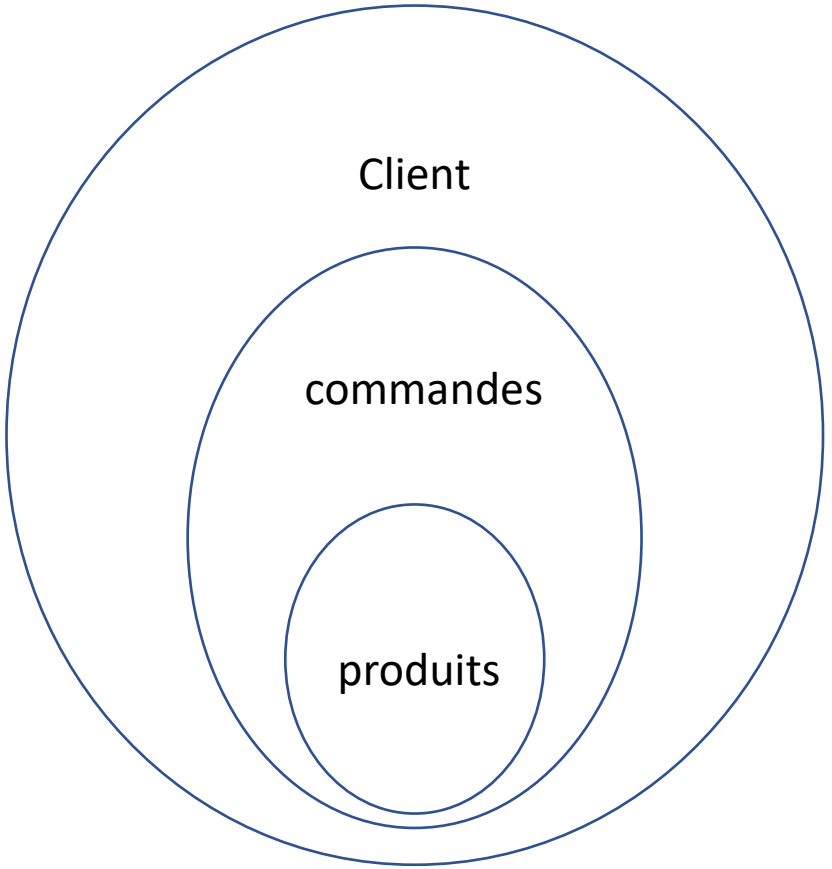
⇒ *Feature engineering et jointure des bases de données*

⇒ *Analyse exploratoire*

Transformation et exploration des données:

Feature engineering et jointure des bases de données

Construction base de données



Feature engineering :

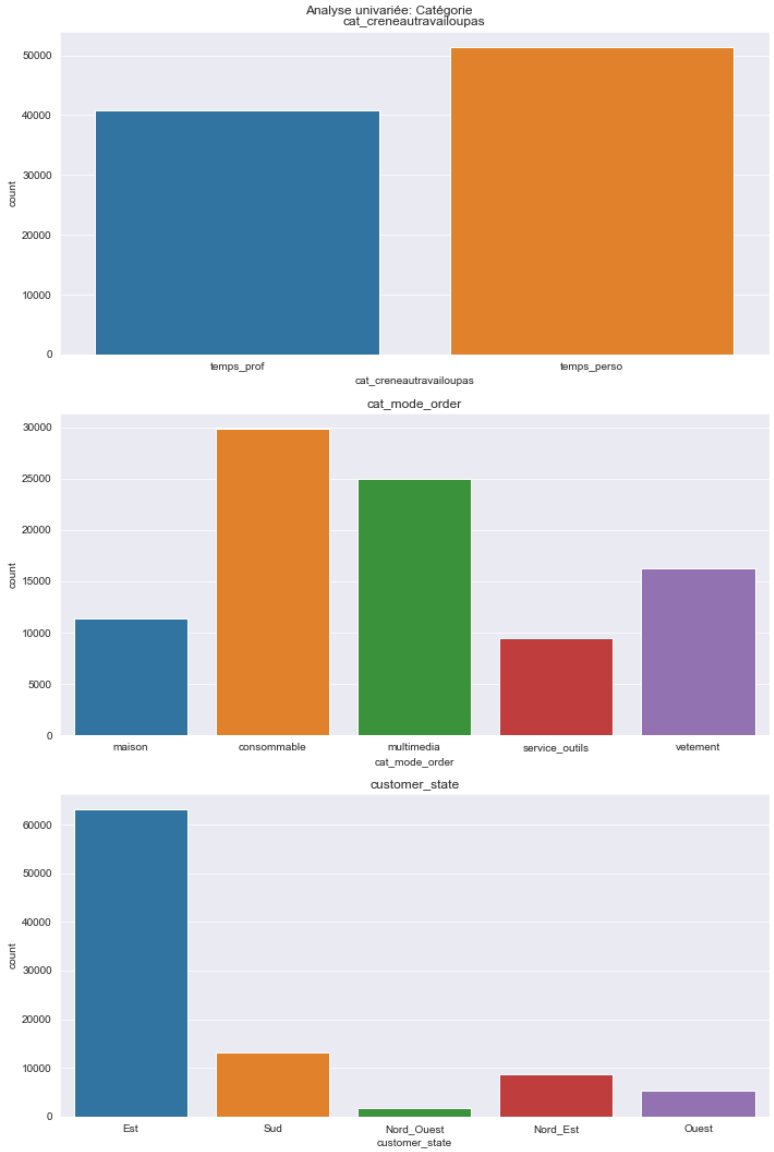
- 73 Catégorie de produits regrouper en 5 (consommable, service outil, multimédia, maison & vêtement)
- Calcul volume par commande
- Calcul poids du colis par commande
- Calcul pourcentage prix livraison
- Création catégories: semaine/weekend et temps travail/temps perso
- Calcul de la distance en Km entre vendeur et acheteur
- Rassemblement par région à partir des états (27->5)

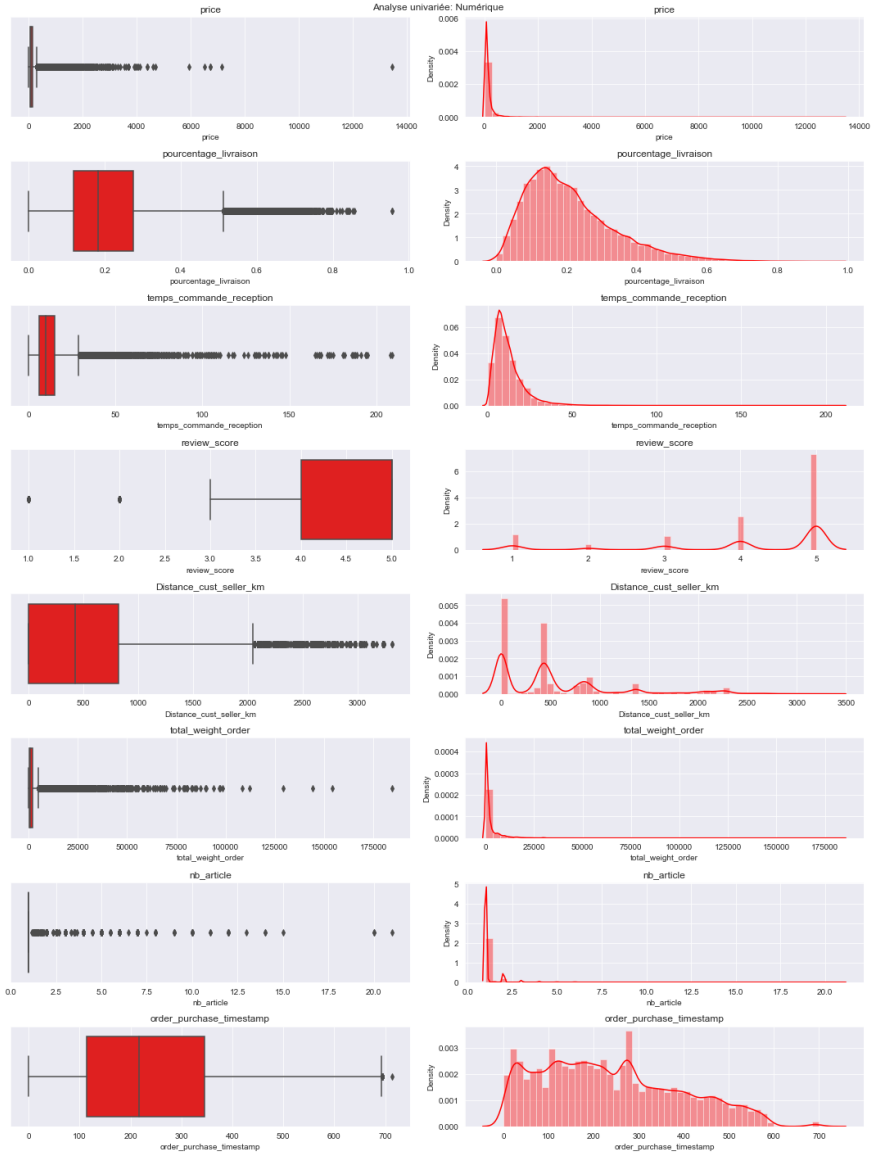
Gestion des valeurs manquantes

- Suppression individu pour certaine variables
- Remplissage par moyenne pour d'autre (review)

Transformation et exploration des données:

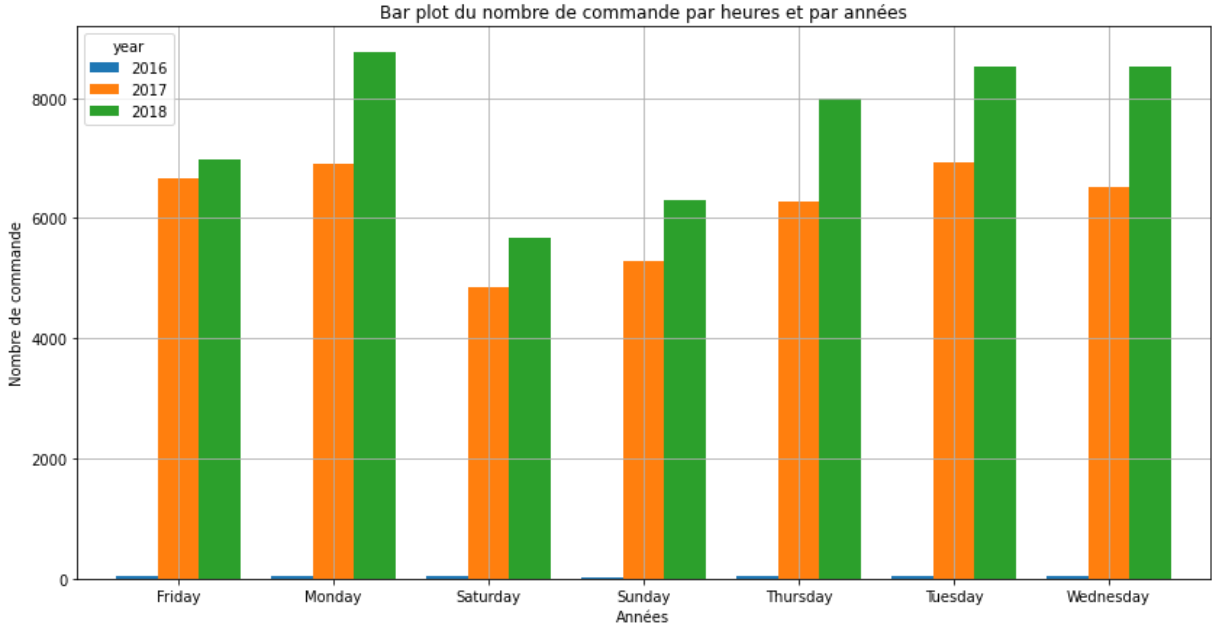
Analyse exploratoire (1-univariée)

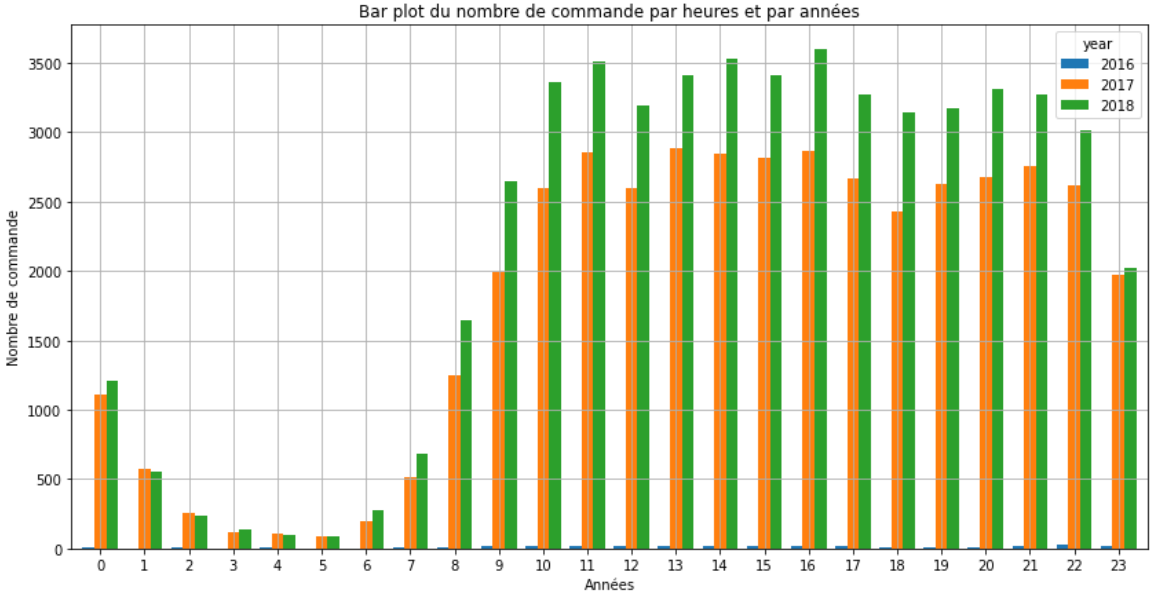


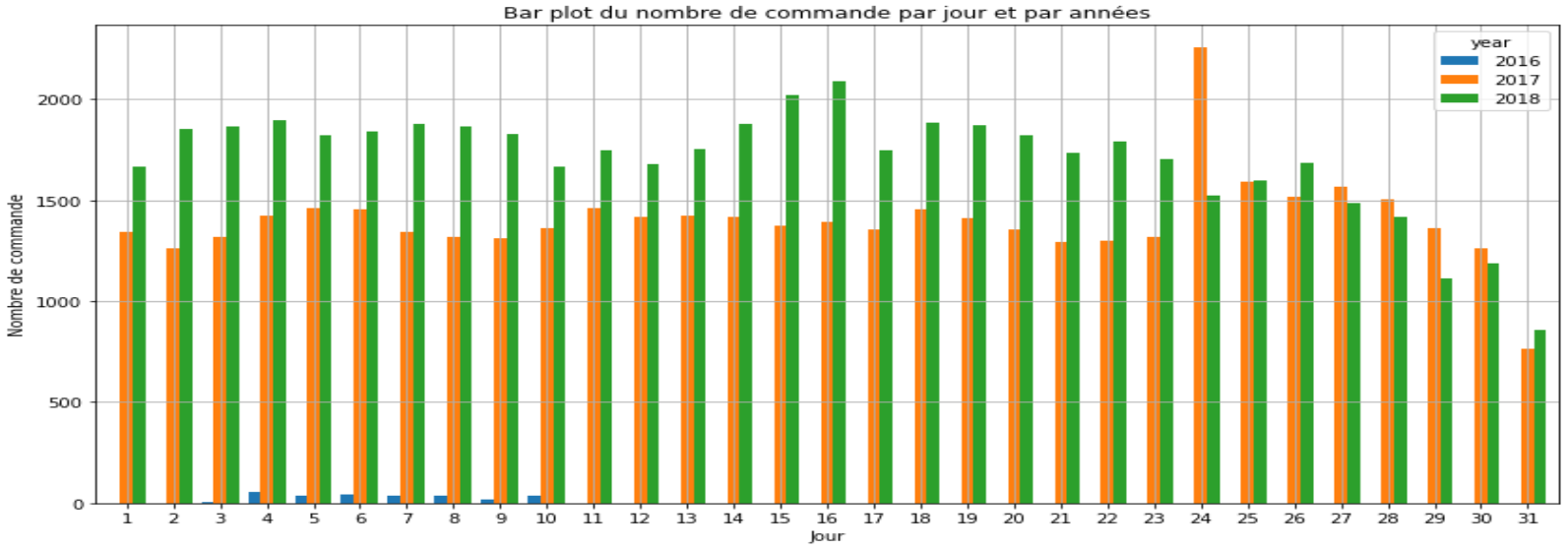


Transformation et exploration des données:

Analyse exploratoire (2-temporelle)

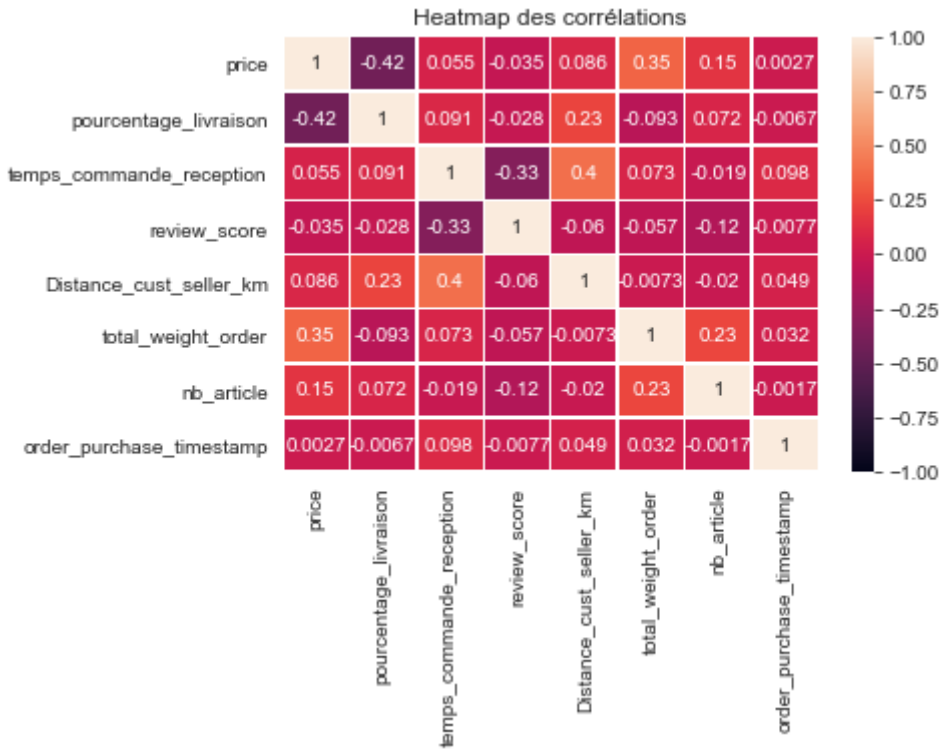


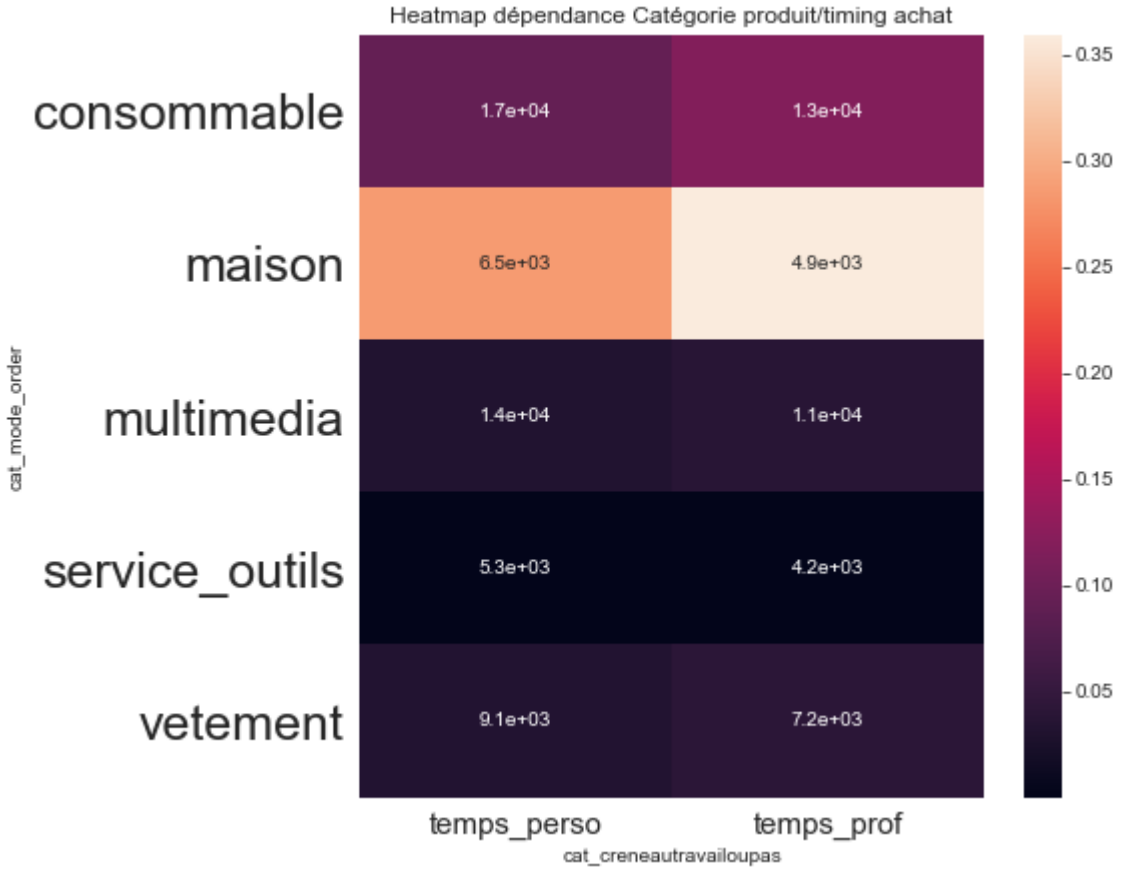




Transformation et exploration des données:

Analyse exploratoire (3-Multivariée)





Possible dépendance entre l'heure et la catégorie de produit acheté par test du chi-2

Différents modèles étudiées

⇒ *Rappel de la problématique et choix de différents modèles*

⇒ *Etude des modèles*

⇒ *Modèle retenue et études des clusters*

Différents modèles étudiés

Rappel de la problématique et choix de différents modèles

Objectif:

- Choisir un modèle de segmentation
- Choisir des variables pertinentes
- Analyser les différents clusters pour le besoin marketing
- Choisir un modèle final

Données:

95 104 orders
92 057 clients uniques

19 variables

Infos:

- Clients (région, review..)
- Commande (période d'achat, nb produit…)
- Produits (catégorie, prix..)

Modèles étudiés:

- RFM (Recency, Frequency & Monetary) avec attribution de note
- DBSCAN
- Agglomerative Hierarchical Clustering
- KMEANS avec différents K sélectionnés

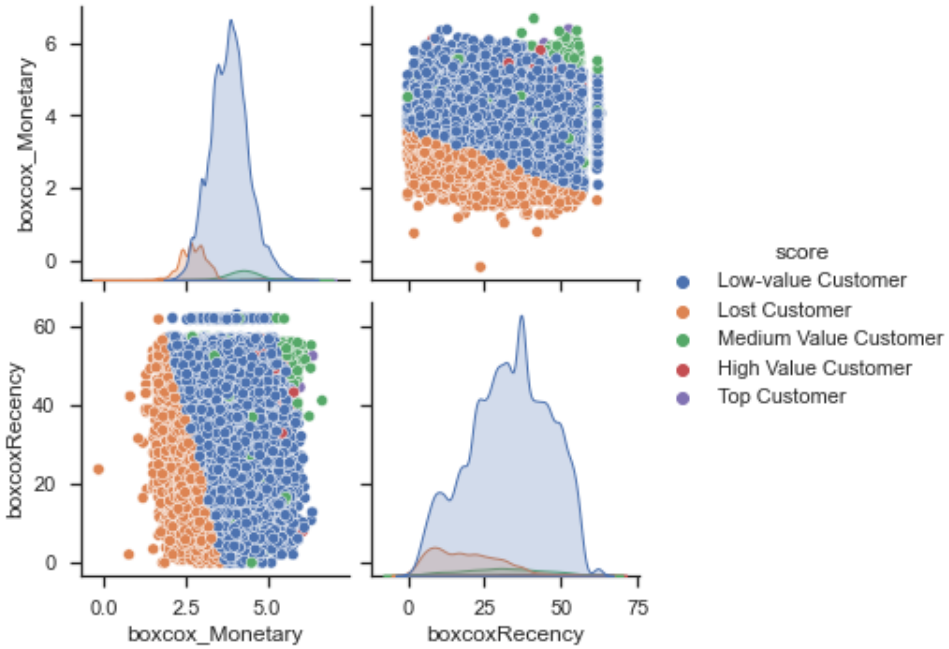
Différents modèles étudiés

Etude des modèles: RFM

Recency: jours écoulés depuis dernière commande
Frequency: fréquence d'achat, nombre de commande
Monetary: dépense total du client

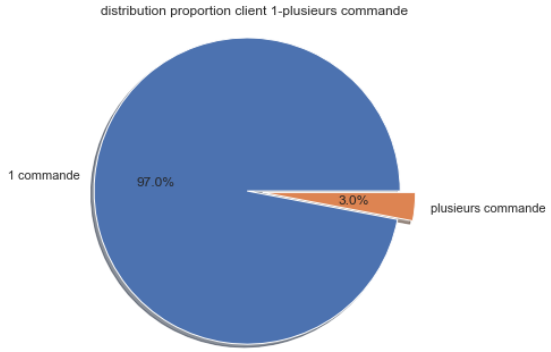
Regroupement par client → Transformation gaussienne (BoxCox)
→ Scaler 0-100 → score RFM

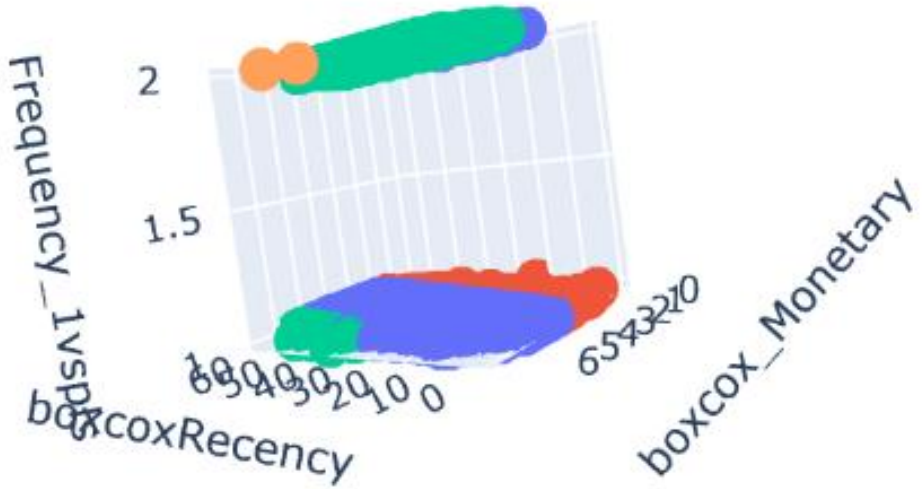
0,15.Recency score + 0,28.Frequency score + 0,57.Monetary score



Rating Customer based upon the RFM score

- rfm score >4.5 : Top Customer
- 4.5 > rfm score > 4 : High Value Customer
- 4>rfm score >3 : Medium value customer
- 3>rfm score>1.6 : Low-value customer
- rfm score<1.6 :Lost Customer





Conclusion RFM:

- Problème frequency: très peu de client reviennent
- distribution très peu similaire

⇒ Prendre une autre approche

Différents modèles étudiés

Etude des modèles: Clustering sur de nouvelle variables

Etude sur 5 features:

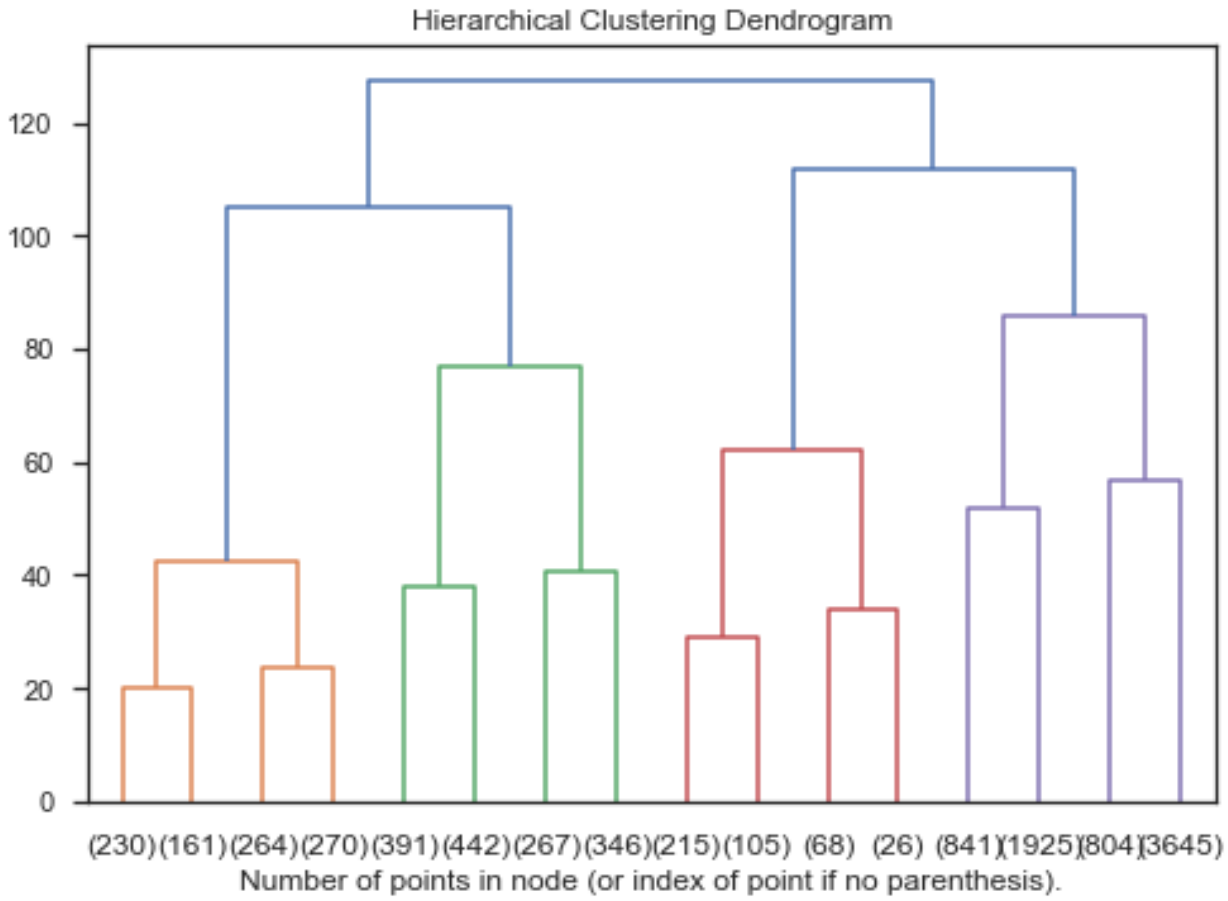
- Distance km client/vendeur moyenne
- Review_score moyen
- Temps entre commande et réception moyen
- Prix moyen
- Pourcentage prix livraison moyen

✗ DBSCAN: problème densité des données, trop ou trop peu de cluster, beaucoup de bruit…

✓ KMeans: Etude de plusieurs k et comparaison sur plusieurs métriques. Analyse des clusters et correspondance agglomerative hierarchical algorithm

Différents modèles étudiés

Etude des modèles: Clustering



KMEANS:

Opération sur les données: StandardScaler()

Choix du paramètre k:

- Observation distorstion score avec la technique du « coude »
- Evaluation par rapport au silhouette score

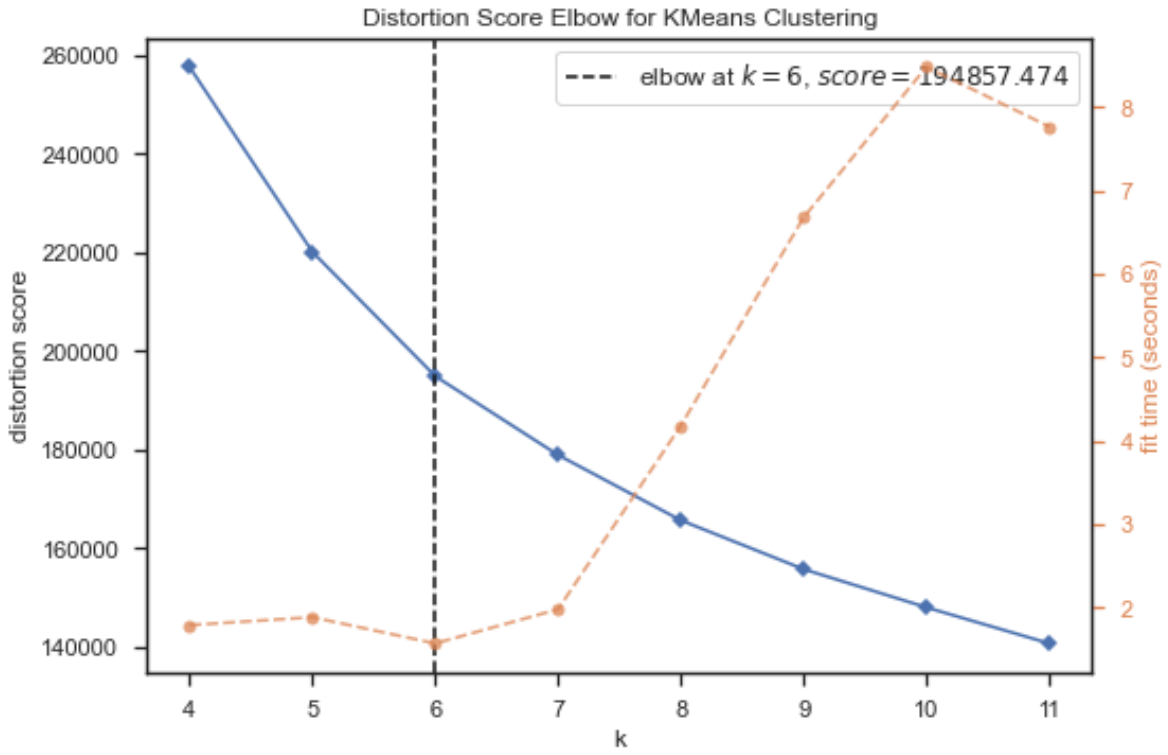
Observation sur les clusters:

- Interprétation des clusters pour répondre objectifs Marketing

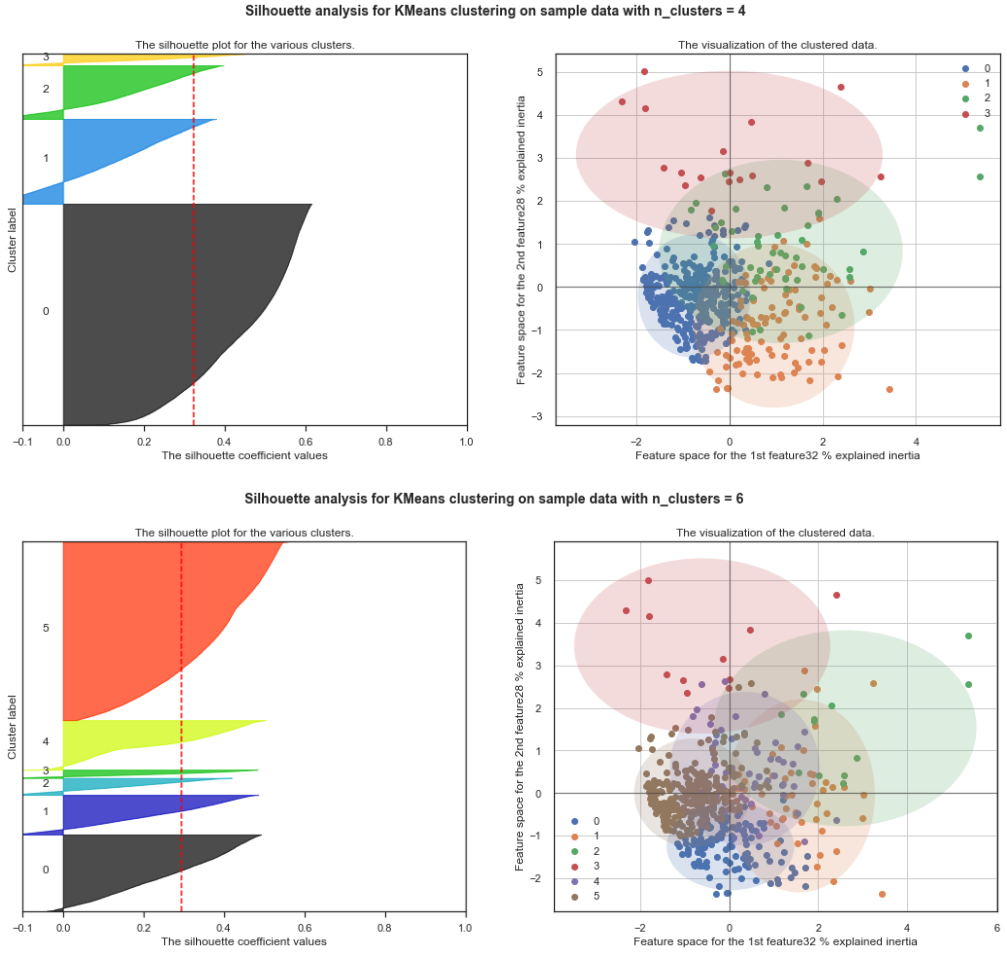
Différents modèles étudiés

Etude des modèles: KMEANS k: méthodes du « coude » vs Silhouette Score

Méthode du « coude »



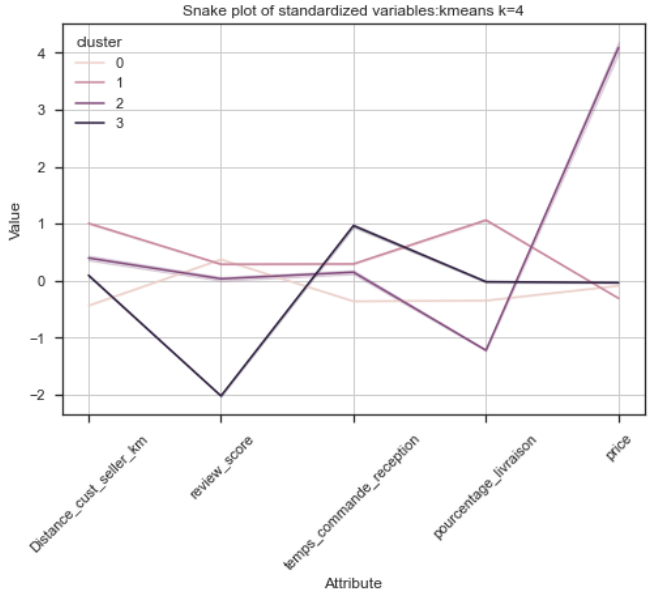
Analyse silhouette score k∈ [3,7]

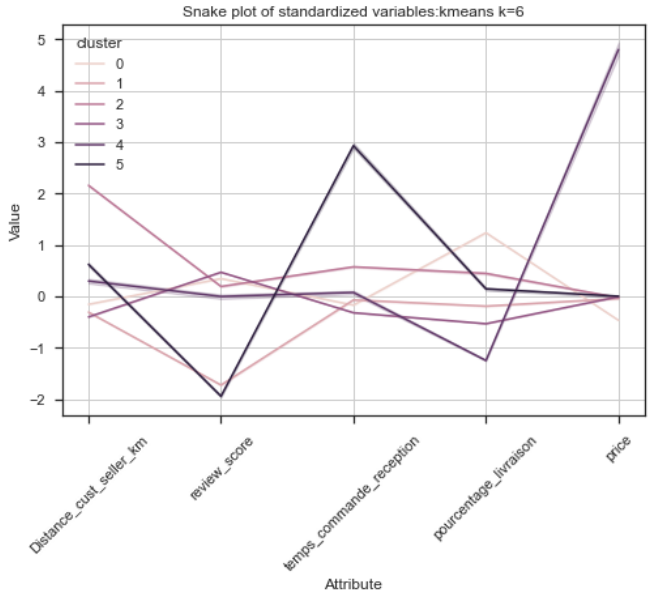


⇒ K intéressant pour 4 et 6 donc analyse des clusters

Différents modèles étudiés

Etude des modèles: KMEANS k: méthodes du « coude » vs Silhouette Score





4means / 6means	0	1	2	3
0	7593	11267	0	167
1	3477	108	68	8668
2	262	8931	224	445
3	43387	667	451	65
4	0	0	2059	0
5	9	221	42	3946

Test d'indépendance du χ_2

Pas d'indépendance et corrélation entre les clusters trouvés

- Analyse des différents clusters pour k=4 et 6
- Les clusters de 6 sont une sous divisions des clusters de 4

→ Choix de garder le kmeans 4

Différents modèles étudiés

Etude des modèles: KMEANS hyperparamètrage avec une recherche par grille

Hyperparamètre de base:

-n_init = 10

-max_iter = 300

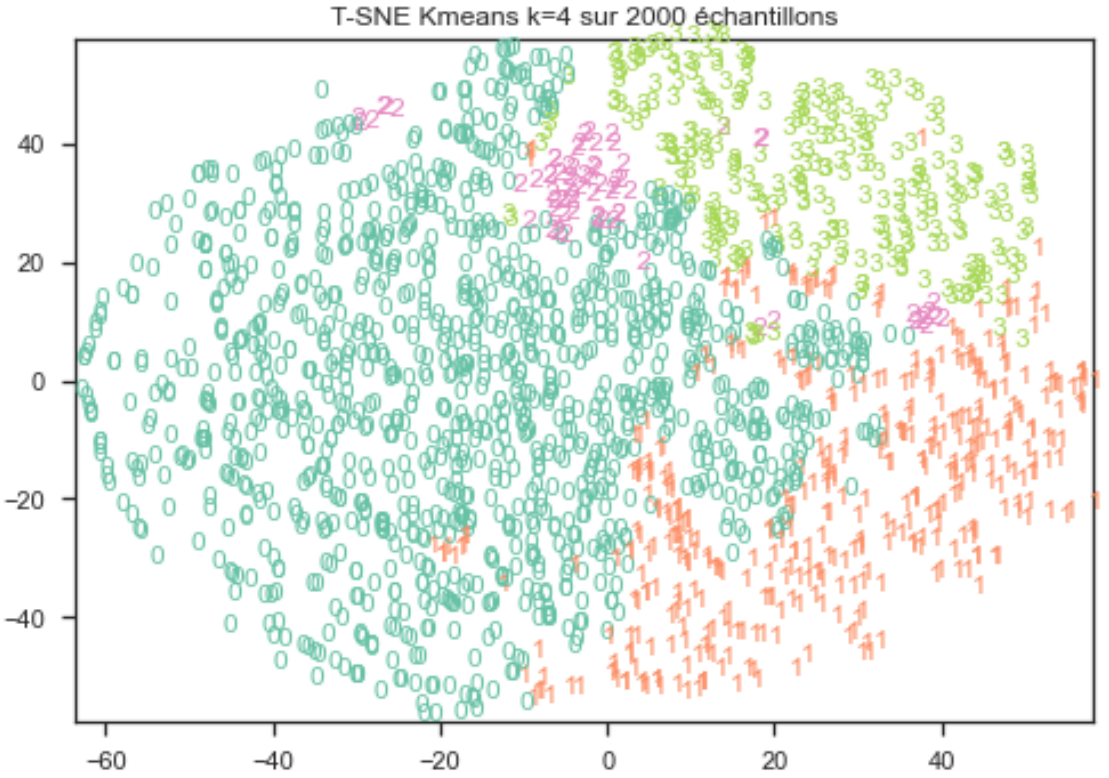
Test avec différentes combinaisons de n_init et max_iter puis on observe le résultat que les changements de ces hyperparmètres ne changent pas la prédictions et les clusters.

	silhouette_score	inertia_	ARI
model_k4_n_init20_max_iter600	0.323826	257598.396842	0.992043
model_k4_n_init20_max_iter100	0.323826	257599.435037	0.992043
model_k4_n_init20_max_iter100	0.323826	257598.396842	0.992043
model_k4_n_init20_max_iter300	0.323826	257598.396842	0.992043
model_k4_n_init10_max_iter600	0.323067	257598.788366	0.997597
model_k4_n_init5_max_iter300	0.323067	257599.435037	0.997597
model_k4_n_init5_max_iter600	0.323067	257599.435037	0.997597
model_k4_n_init10_max_iter100	0.323067	257599.435037	0.997597
model_k4_n_init10_max_iter300	0.323067	257599.435037	0.997597
model_k4_n_init10_max_iter600	0.323067	257599.435037	0.997597
model_k4_n_init5_max_iter100	0.323067	257598.788366	0.997597
model_k4_n_init5_max_iter300	0.323067	257598.788366	0.997597
model_k4_n_init5_max_iter600	0.323067	257598.788366	0.997597
model_k4_n_init10_max_iter100	0.323067	257598.788366	0.997597
model_k4_n_init10_max_iter300	0.323067	257598.788366	0.997597
model_k4_n_init5_max_iter100	0.323067	257599.435037	0.997597
model_init	0.322702	257599.435037	1.0

Différents modèles étudiés

Etude des modèles: Interprétation des 4 clusters

Interprétation: Centroïdes & comparaison des points (scatter plot, distributions par clusters boxplot…)



0: des bons clients qui dépensent correctement et sont satisfaits par la rapidité de la livraison
1: clients satisfaits par leur commande et qui sont prêt à mettre beaucoup d'argent dans la livraison
2: les clients les plus dépensiers, plutôt satisfaits avec peu de % de livraison
3: clients peu contents, qui ont attendu le plus, avec une dépense moyenne

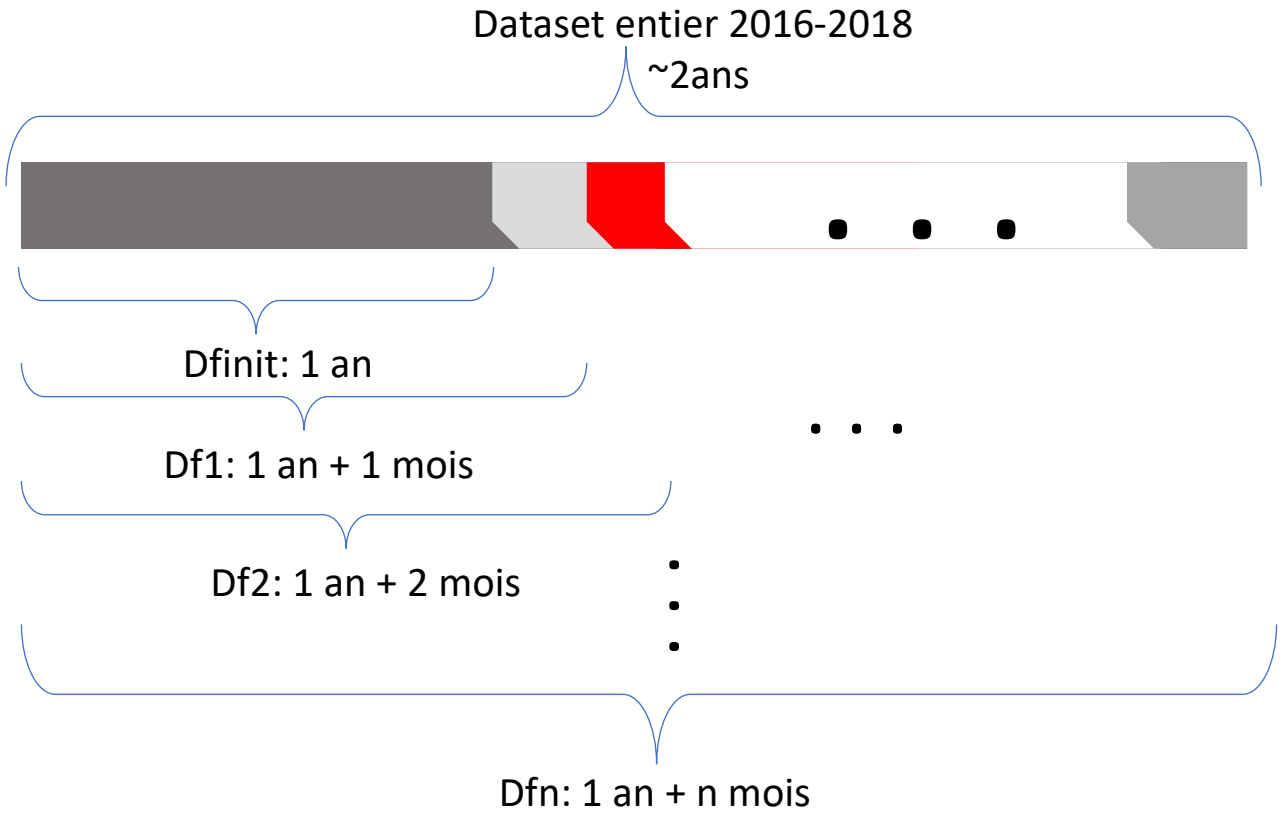
	Distance_cust_seller_km	review_score	temps_commande_reception	pourcentage_livraison	price	nb_cust_clust
0	266.086624	4.633977	8.679637	0.165563	119.548500	54728
1	1157.792924	4.526629	14.922509	0.341787	73.518144	21194
2	783.496198	4.199398	13.550176	0.056699	996.237728	2844
3	595.944214	1.580764	21.352489	0.206881	130.871750	13291

Etude du délai de maintenance pour le modèle sélectionné

Etude du délai de maintenance pour le modèle sélectionné:
Idée d'analyse du concept drift

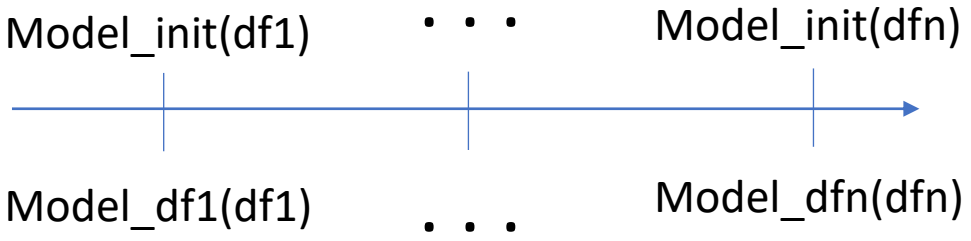
Concept drift: Changement de propriété des données (tendance…)

A. L. F. D. F. G. J. G. a. G. Z. Jie Lu, «Learning under Concept Drift : a review,» IEEE, 2020.



Score Adjusted rand index:

Score similarité entre deux prédictions (1: semblable, 0: hasard)



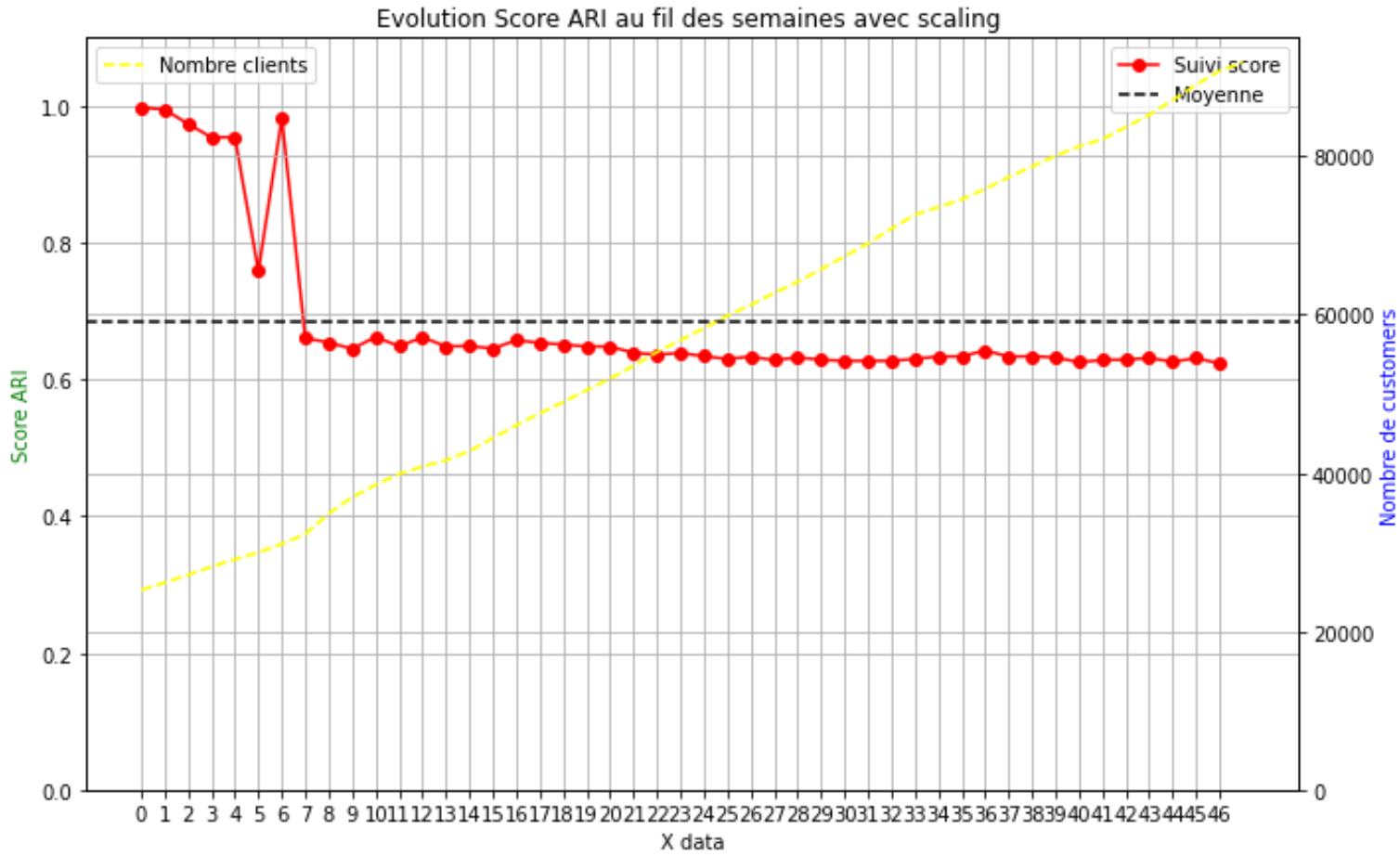
Etude du délai de la maintenance pour le modèle sélectionné

Evolution ARI

Df_init: du 2016-09-15 12:16:38 au 2017-10-01 12:16:38
25 000 clients

Df_k: on ajoute les k semaines suivantes de la fin de df_init de données

Finalement on voit qu'une mise à jour de notre modèle peut se faire toutes les 4 semaines.



Conclusion

- ✓ Découverte d'une base de données définissant un Marketplace:
  - Travail de jointure pour définir des profils clients selon leurs commandes
  - Feature engineering sur les données en créant des variables
  - Analyse exploratoire pour mieux comprendre les données
- ✓ Recherche de modèle pouvant définir et segmenter les clients selon un besoin marketing:
  - Recherche d'un modèle cohérent et interprétable
  - Distinguer les différents clusters obtenus
- ✓ Recherche du délai de maintenance sur un modèle choisi:
  - Trouver une méthode de détection de dérive du modèle
  - Séparer et analyser sur graph l'évolution d'une métrique de comparaison
  - Proposer une fréquence de maintenance
- ✓ Utiliser la norme PEP8 pour le code:
  - Connaitre les contraintes (nombre de caractère par ligne, syntaxe déclaration, indentation…)
  - Utilisation d'une librairie Black (pour jupyternb) pour le dernier notebook

Pour aller plus loin

- Avoir un retour de l’équipe marketing pour connaître plus précisément le besoin
- Essayer d’autres modèles comme K-Prototypes qui prend en compte les variables catégorielles
- Analyser et comprendre mieux le changement de comportement observer pour le concept drift et la nécessité de maintenance

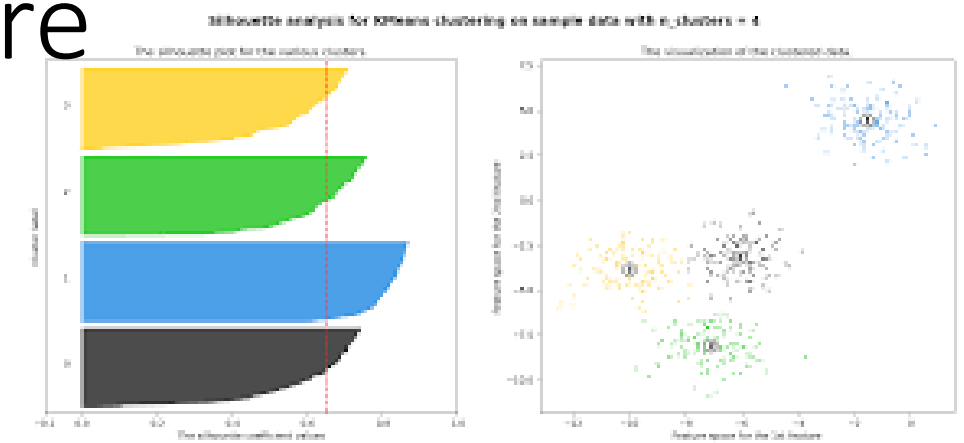
Merci!

Questions ?

Silhouette score

Mesure qualité de partition de données pour la classification automatique.

On attribue a chaque point un coefficient, différence entre distance moyenne avec les points du même groupe et la distance moyenne avec les points des autres groupes. (cohésion-séparation)



Distortion score

Calcul de la variance des individus d'un même cluster.
C'est-à-dire la distance moyenne des points avec leur cluster (variance intra-classe)