Projet 5 Openclassrooms

Segmentation Client d’un site de e-commerce

Contexte Projet

Consultant pour olist entreprise brésilienne de mise en relation de vendeur avec des marketplaces

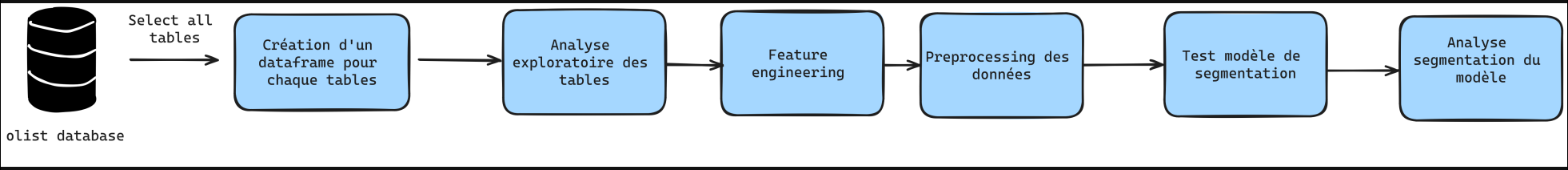
Objectifs

1. Requêtes SQL pour un tableau de bord sur l'expérience client
2. Analyse exploratoire des clients d'olist
3. Segmentation des clients avec des algorithmes de clustering non supervisés
4. Simulation d'un contrat de maintenant de notre modèle pour assurer son efficacité

Requête SQL pour Dashboard expérience client

```sql
-- Question 1 ${1}
SELECT *
FROM orders
WHERE order_status != 'canceled'
    AND order_purchase_timestamp >= (
        SELECT DATETIME(MAX(order_purchase_timestamp), '-3 months')
        FROM orders
    )
    AND order_estimated_delivery_date > DATETIME(order_delivered_customer_date, '+3 days');
```

Etapes modelisation clustering



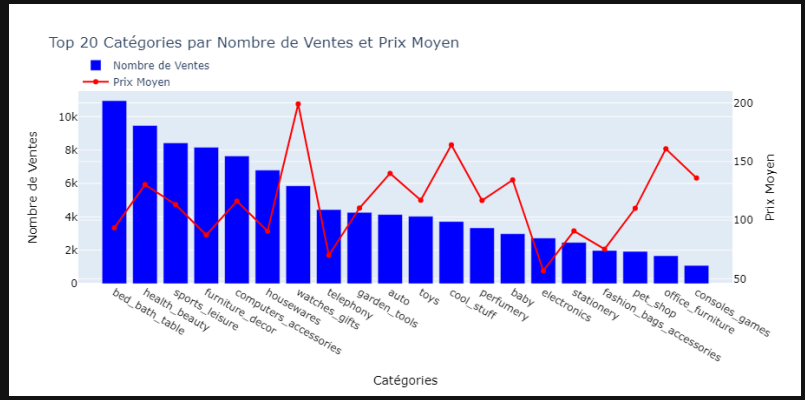
Analyse Exploratoire des données

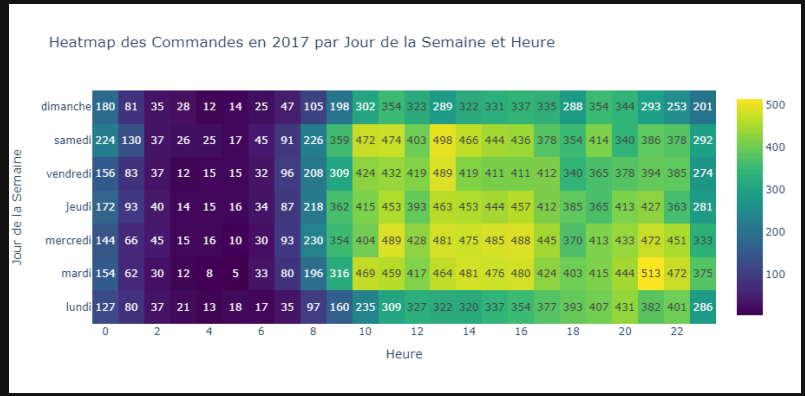
Informations sur les données

- 9 *tables* Contenant :
  - Informations sur les clients,vendeur,produits,avis
  - Historique de données de **2 ans**
  - Données sur **100 000 clients**

Analyser les *features* pertinantes pour notre analyse

Exemple: analyse graphique effectuer sur nos données





Création d’un jeu de données pour modelisation

Hypothèse formulée

- Regroupé nos données en fonction des commandes des clients `unique_id`
  - Requetâge directement sur notre Base de données (limitation temps traitement)

Nettoyage des données

- Suppression des valeurs dupliquées
- Analyse/suppression des outliers (2750 lignes supprimés)
- Remplacement des valeurs manquantes :
  - Méthode 'Most fréquent values' en fonction de la `catégorie préférée`

Jeu de données obtenu
96682 lignes et 18 colonnes pour qualifier nos données

Feature enginerring et selection des Features

Transformation de variable dans une optique métier

- `Récence` `Fréquence` `Montant` (analyse de la qualité client)
- Dépenses par commandes clients par catégories :
  - Réduction du nombre de catégories à 10
- `type_de_paiement_préféré` `catégorie_préférée` `jour_avec_plus_de_commandes` par client

Extrait jeu de donnée

identifiant_client	récence	dernière_date_achat	fréquence	montant	type_de_paiement_préféré	ville_client	état_client	jour_avec_plus_de_commandes	catégorie_préférée
0000366f3b9a7992bf8c76cfdf3221e2	111.0	2018-05-10 10:56:27	1	129.90	credit_card	cajamar	SP	Jeudi	bed_bath_table
0000b849f77a49e4a4ce2b2a4ca5be3f	114.0	2018-05-07 11:11:27	1	18.90	credit_card	osasco	SP	Lundi	health_beauty
0000f46a3911fa3c0805444483337064	537.0	2017-03-10 21:05:03	1	69.00	credit_card	sao jose	SC	Vendredi	stationery
0000f6ccb0745a6a4b88665a16c9f078	321.0	2017-10-12 20:29:41	1	25.99	credit_card	belem	PA	Jeudi	telephony
0004aac84e0df4da2b147fca70cf8255	288.0	2017-11-14 19:45:42	1	180.00	credit_card	sorocaba	SP	Mardi	telephony
...	...	...	...	...	...	...	...	...	...
fffbf87b7a1a6fa8b03f081c5f51a201	245.0	2017-12-27 22:36:41	1	149.00	credit_card	fortaleza	CE	Mercredi	bed_bath_table
fffea47cd6d3cc0a88bd621562a9d061	262.0	2017-12-10 20:07:56	1	64.89	credit_card	feira de santana	BA	Dimanche	baby
ffff371b4d645b6ecea244b27531430a	568.0	2017-02-07 15:49:16	1	89.90	credit_card	sinop	MT	Mardi	auto
ffff5962728ec6157033ef9805bacc48	119.0	2018-05-02 15:17:41	1	115.00	credit_card	bom jesus do norte	ES	Mercredi	watches_gifts
ffffd2657e2aad2907e67c3e9daecbeb	484.0	2017-05-02 20:18:45	1	56.99	credit_card	campo largo	PR	Mardi	perfumery

Qu’esceque le score RFM ?

Technique qui permet de determiner les habitudes du consommateurs pour mieux cibler les actions marketing à venir

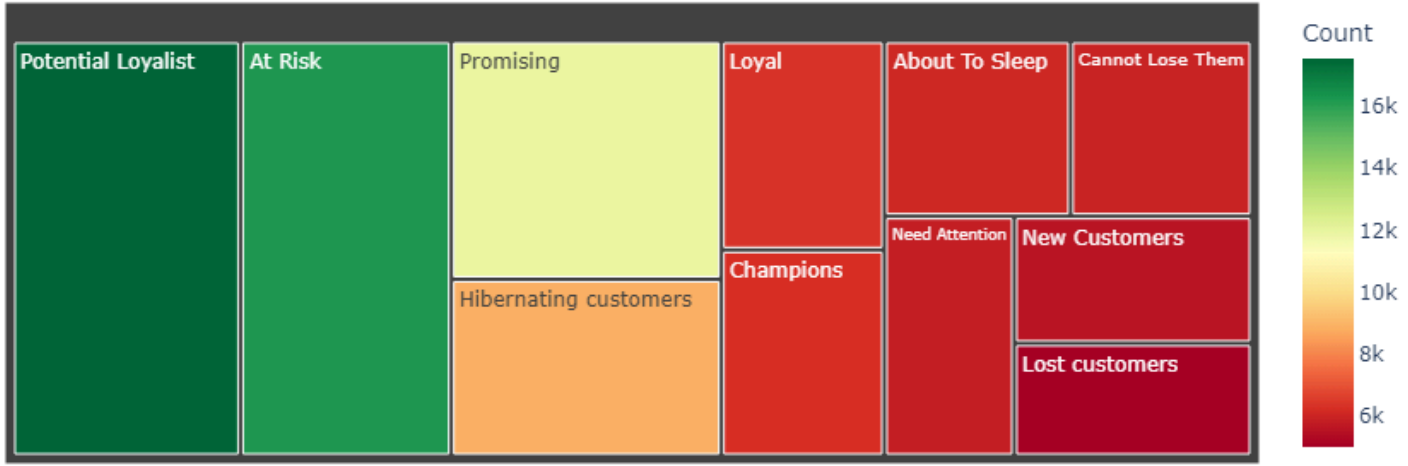
RFM signification :

- Récence : quelle est la date du dernier achat ?
- Fréquence : quel est le nombre d’achats effectués ?
- Montant : quelle est la somme d’achats cumulées réalisés ?

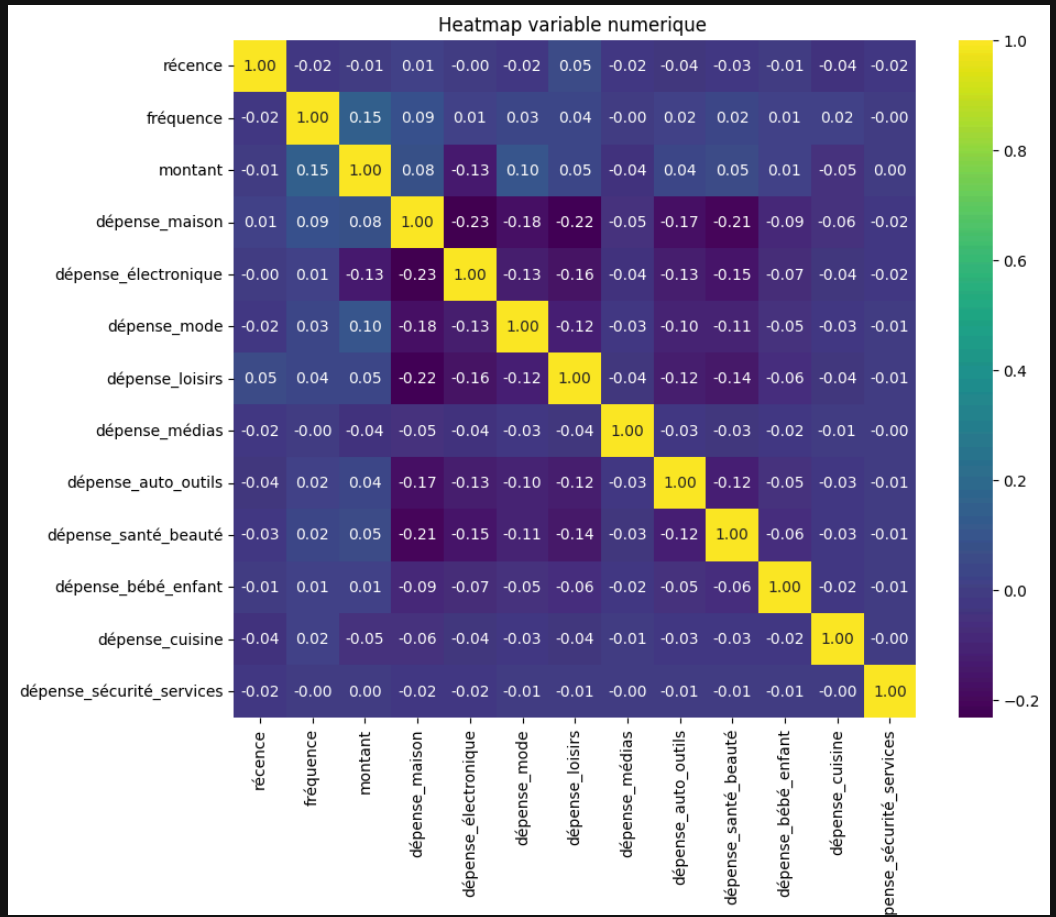
Le scoring RFM permet grouper les clients en segment

Analyse Scoring RFM sur nos données

Treemap des segments de clients



Analyse corrélation entre nos variables



Préparation des données pour modelisation

Encodage des variables numérique

- StandardScaler

Encodage des variables catégorielles

- LabelEncoder

identifiant_client	récence	fréquence	montant	type_de_paiement_préféré	ville_client	état_client
0000366f3b9a7992bf8c76cfdf3221e2	-0.825585	-0.200448	-0.074179	1	648	25
0000b849f77a49e4a4ce2b2a4ca5be3f	-0.805902	-0.200448	-0.574410	1	2571	25
0000f46a3911fa3c0805444483337064	1.969343	-0.200448	-0.348630	1	3486	23
0000f6ccb0745a6a4b88665a16c9f078	0.552197	-0.200448	-0.542459	1	444	13
0004aac84e0df4da2b147fca70cf8255	0.335688	-0.200448	0.151602	1	3724	25

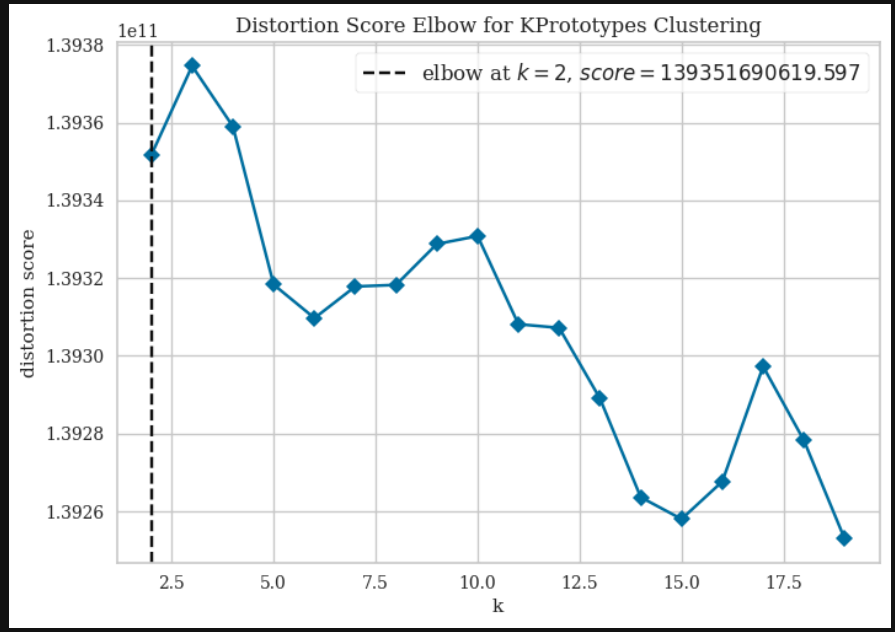
Modelisation clustering : Test de différents modèles

Modèle Kprototypes

- Modèle permettant mixer des données numériques et catégorielles
- Test Elbow pour trouver le nombre optimal de cluster

Modèle non retenu car il propose un nombre trop restrint de cluster pour analyse marketing

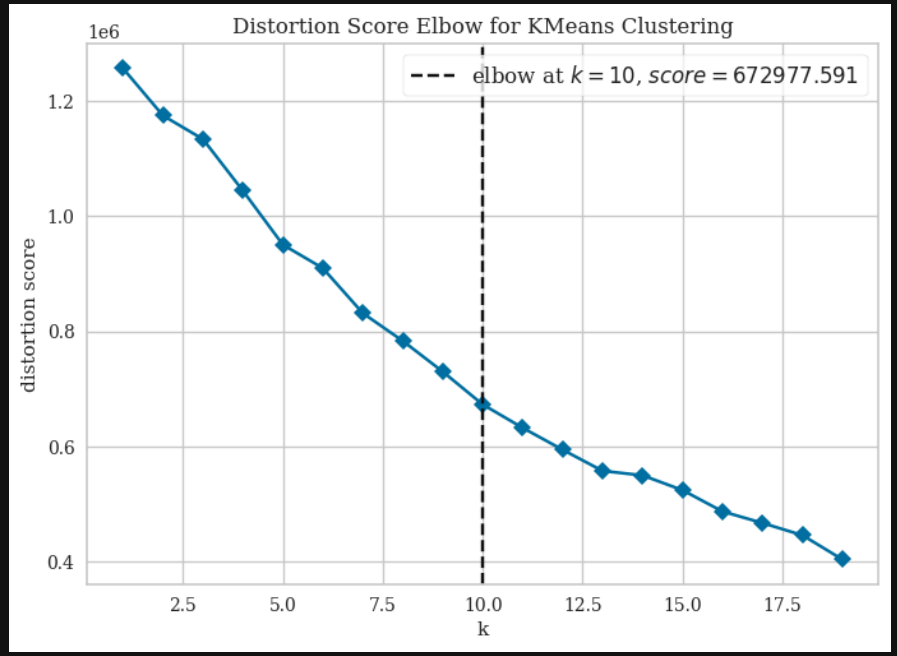
Methode Elbow Kprototypes



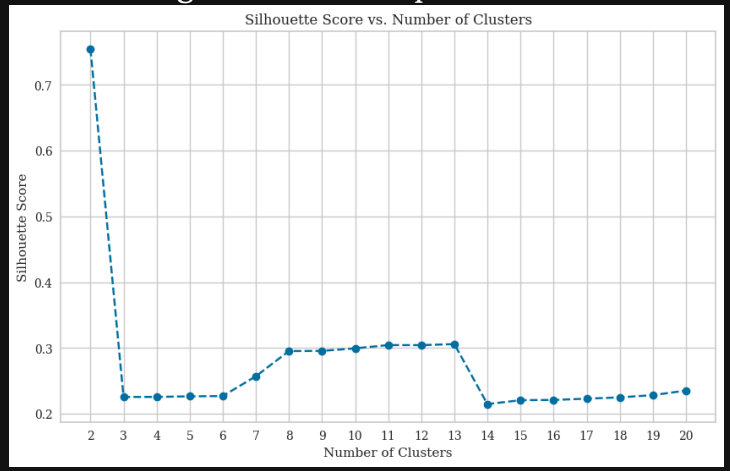
Modelisation clustering : Test de différents modèles

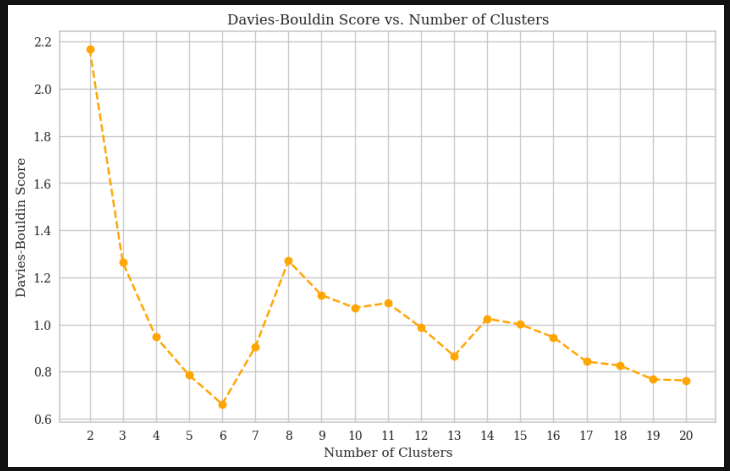
Kmeans

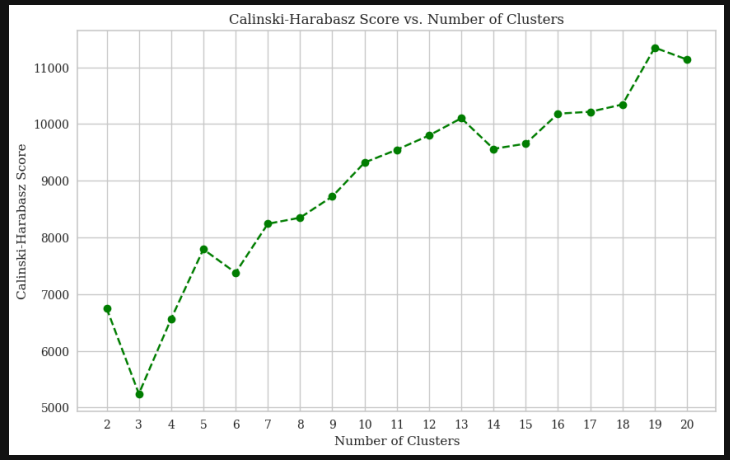
- Selection des features numérique
- Test Elbow pour trouver le nombre optimal de cluster
- Nombre de cluster optimum 10

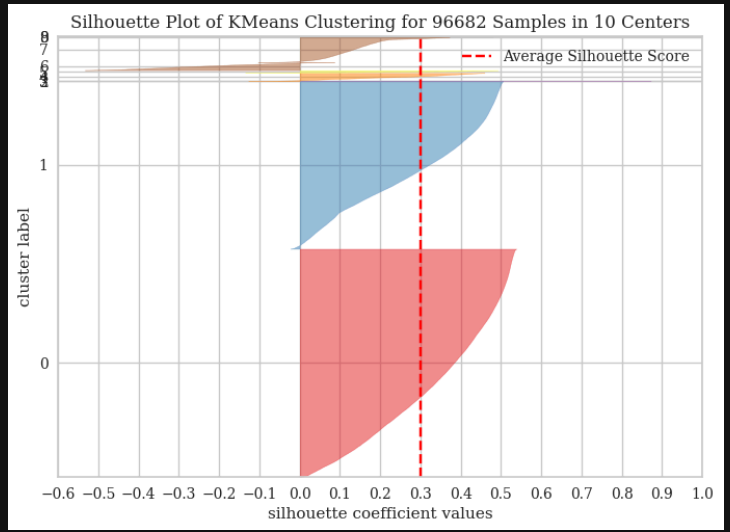


Kmeans clustering : autres métriques choix du nombre de clusters

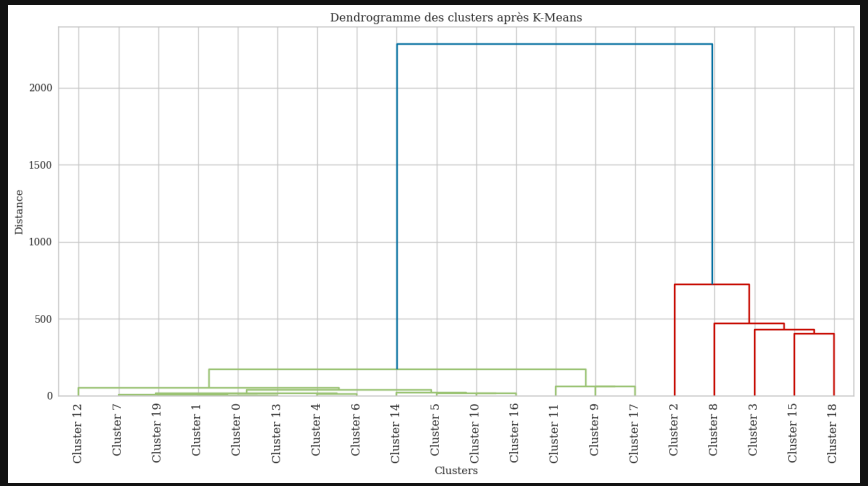


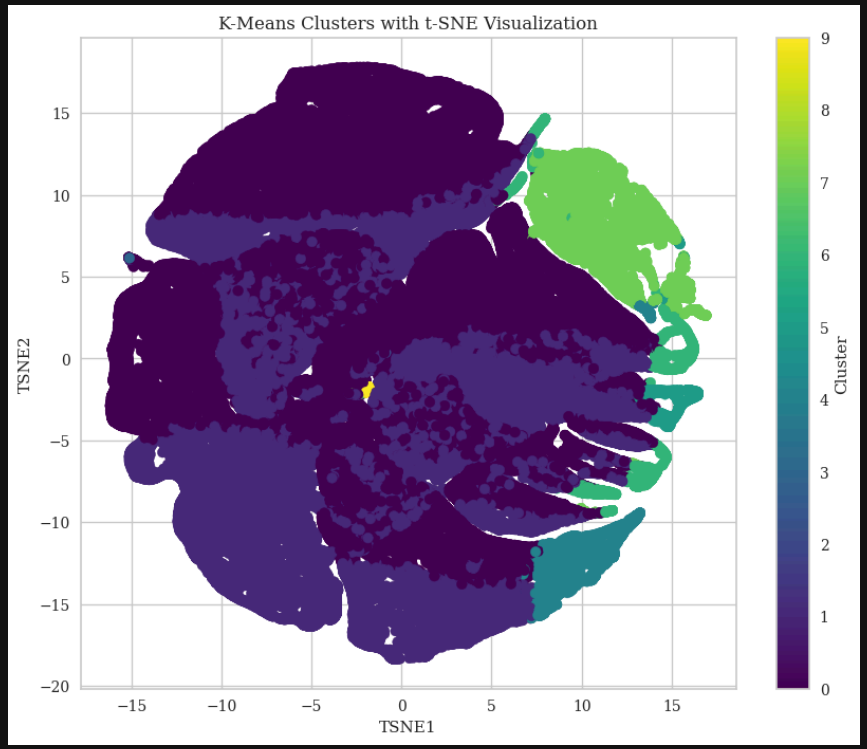






Visualisation des cluster kmeans k=10

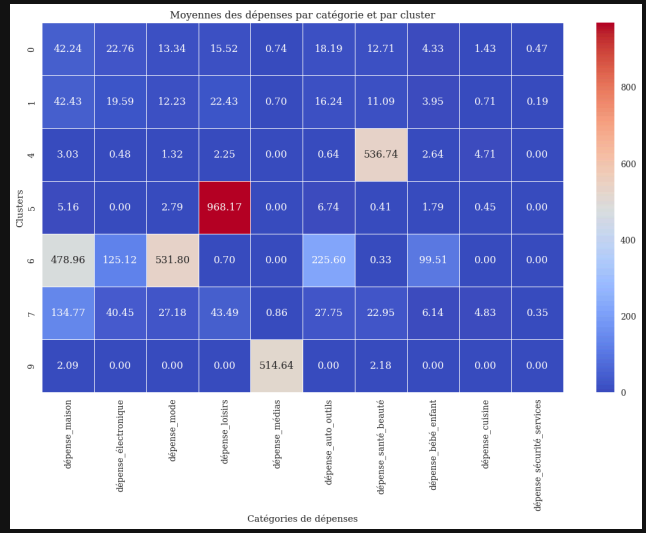


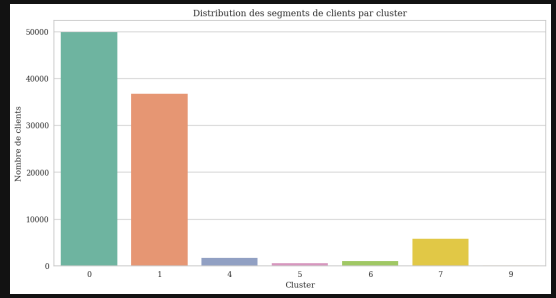


Analyse des clusters kmeans

- Un nombre trop important de clusters pour une analyse marketing
- **Variable** `dépense catégories` qui détermine trop la formation des clusters

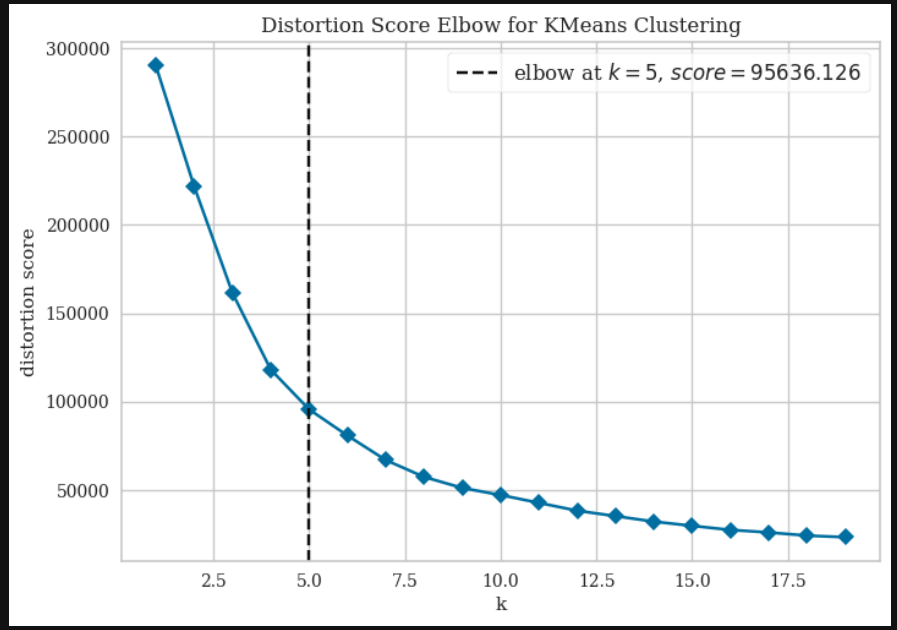
Tester kmeans sans les variables `dépense catégories`



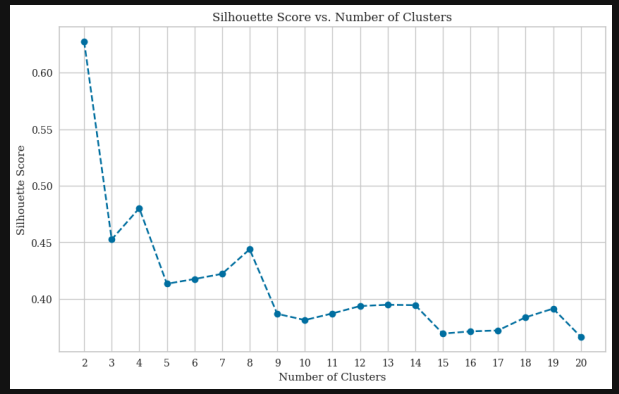


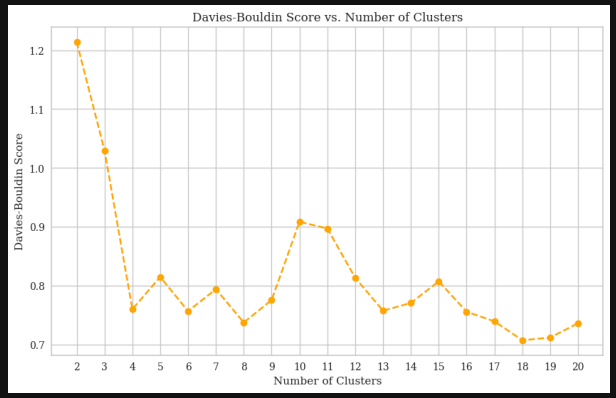
Kmeans suppression des features

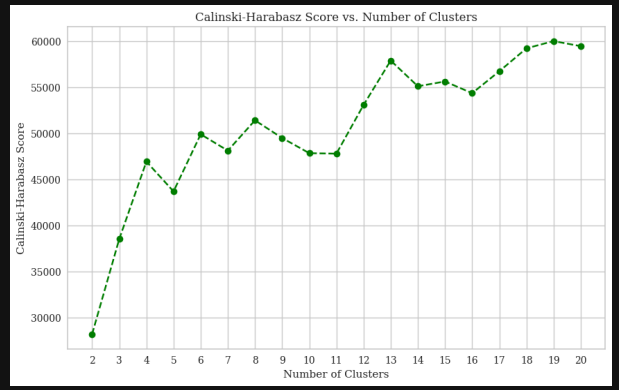
- Recherche du cluster optimum k=5
- Silhouette score meilleur que sur le précédent modèle

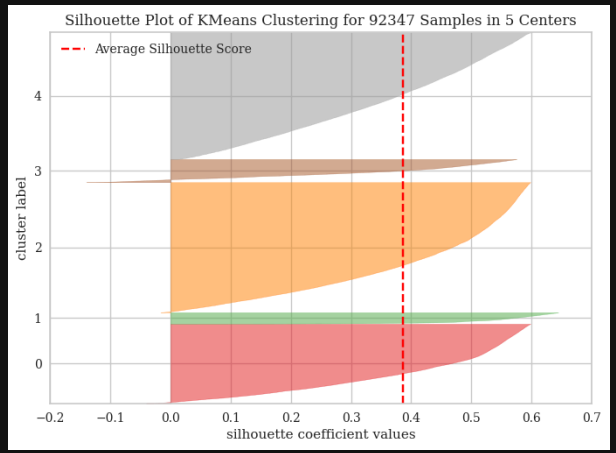


Kmeans clustering : autres métriques choix du nombre de clusters

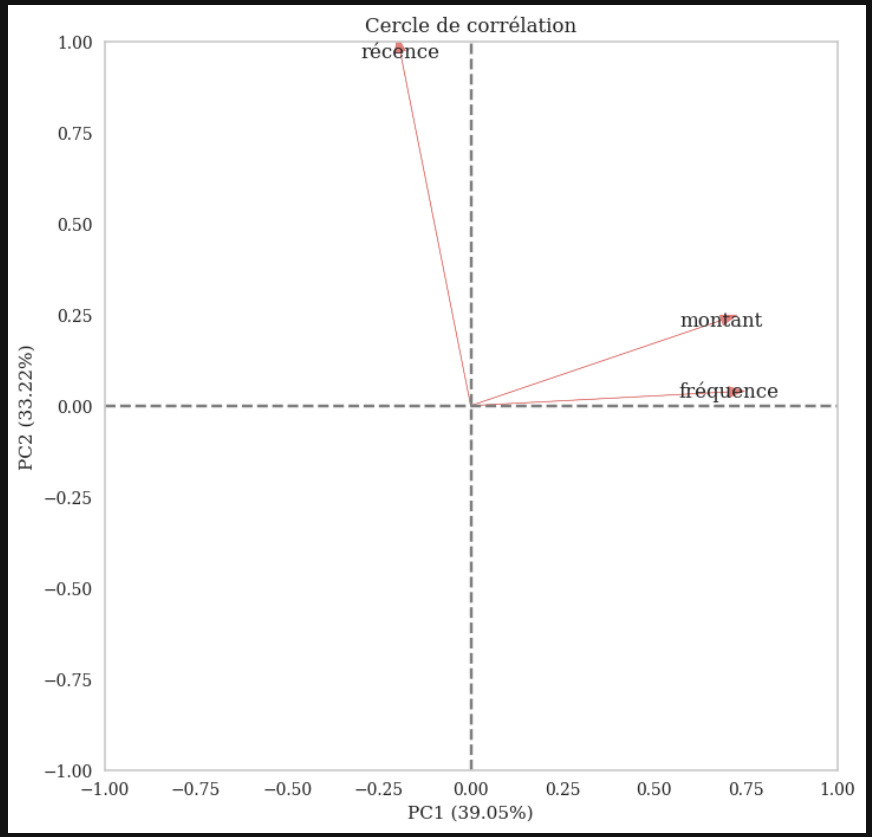


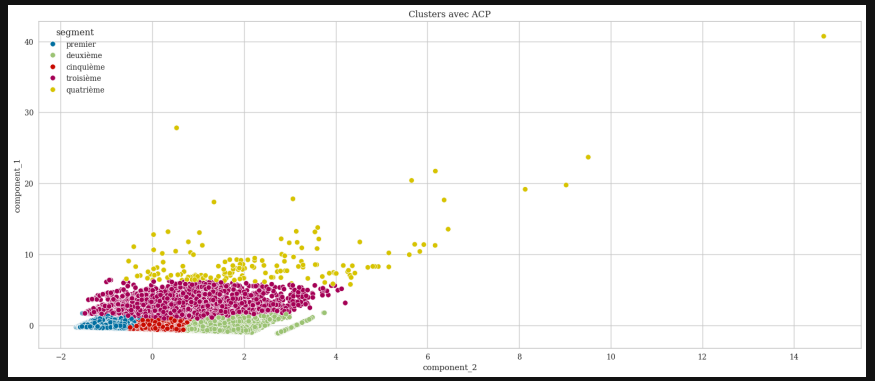


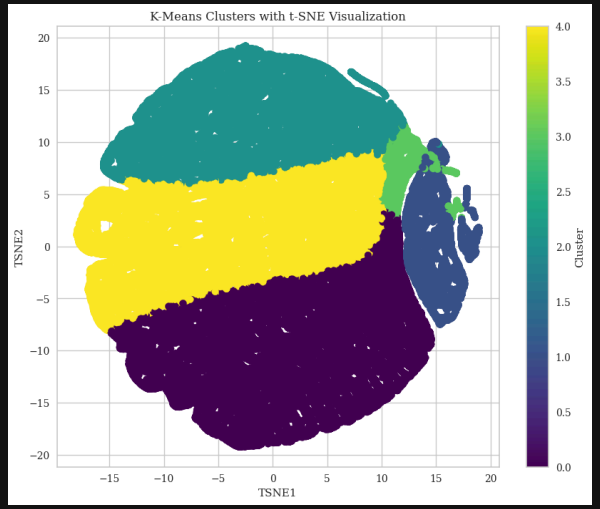




Visualisation des cluster kmeans k=5

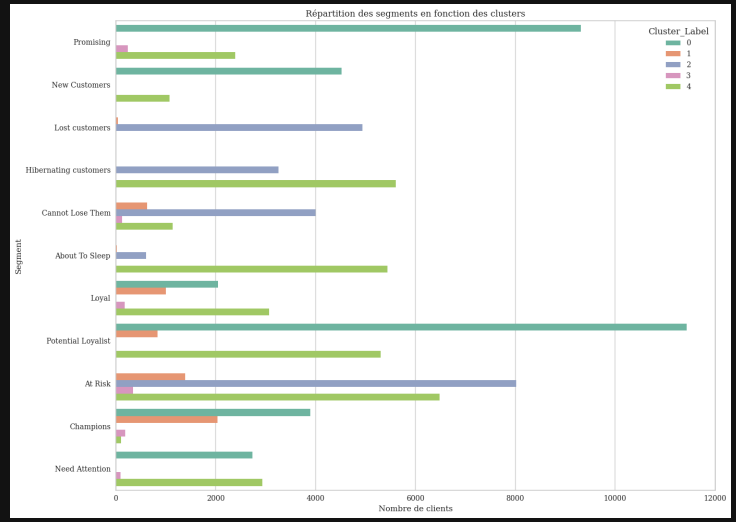


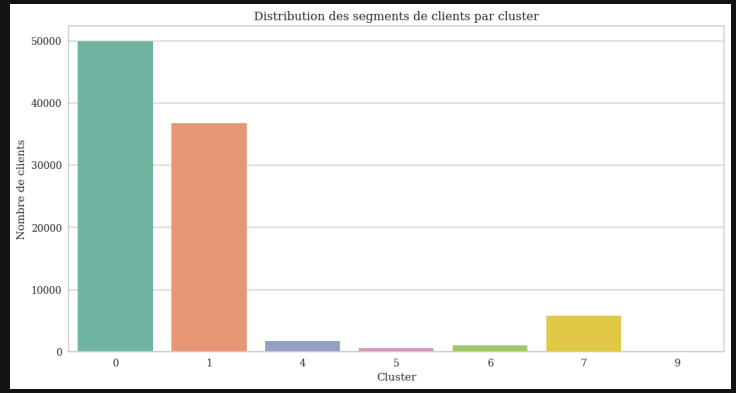




Analyse des clusters kmeans

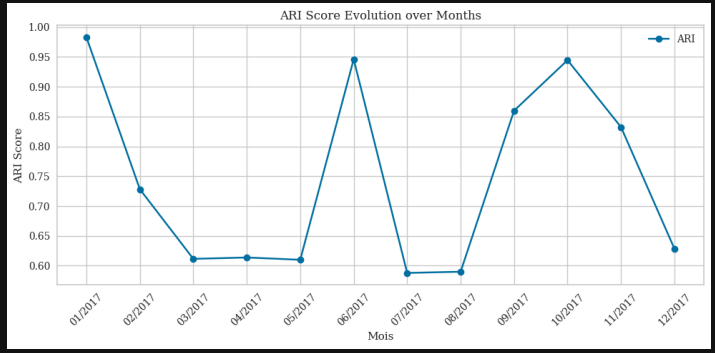
Cluster	Caractéristiques	Actions Marketing
Cluster 0	- Taille Importante : Le plus grand cluster en nombre de clients. - Dépenses Modérées : Dépenses faibles concentrées sur "dépense_maison" et "dépense_électronique". - Segments Dominants : "Potential Loyalist" et "Promising".	- Renforcement de la Fidélisation : Offres exclusives, programmes de fidélité. - Campagnes Ciblées : Stimuler les dépenses dans "mode" et "loisirs".
Cluster 1	- Taille Modérée : Nombre significatif de clients. - Dépenses Élevées : Catégories "dépense_maison", "dépense_mode", "dépense_loisirs". - Segments Représentés : "Cannot Lose Them" et "Champions".	- Offres Premium : Services ou produits premium. - Engagement Personnalisé : Avantages exclusifs pour les "Champions".
Cluster 2	- Taille Moyenne : Dépenses faibles à modérées. - Dépenses Uniformes : "loisirs" et "santé_beauté". - Segments Inclus : "At Risk" et "Hibernating Customers".	- Campagnes de Réactivation : Promotions ciblées. - Programmes de Récupération : Remises pour les clients "At Risk".
Cluster 3	- Petit Taille, Dépenses Élevées : "dépense_électronique" et "dépense_maison". - Segments Principaux : "Champions" et "Loyal".	- Programme VIP : Avantages premium. - Cross-Selling et Upselling : Produits complémentaires de haute valeur.
Cluster 4	- Taille Importante : Dépenses modérées. - Dépenses Uniformes : "dépense_maison" et "dépense_loisirs". - Segments Dominants : "Promising" et "At Risk".	- Incitations à la Fidélité : Programmes de récompenses. - Optimisation des Campagnes : Cibler les segments "At Risk".





Simulation d’un contrat de maintenance

- Réentrainement du modèle : 3 mois
- Test de Kolmogorov-Smirnov, pas de différence entre les distributions testées et celle initiale



	Mois	KS_récence	KS_fréquence	KS_montant
0	01/2017	(0.0, 1.0)	(0.0, 1.0)	(0.0, 1.0)
1	02/2017	(0.0, 1.0)	(0.0, 1.0)	(0.0, 1.0)
2	03/2017	(0.0, 1.0)	(0.0, 1.0)	(0.0, 1.0)
3	04/2017	(0.0, 1.0)	(0.0, 1.0)	(0.0, 1.0)
4	05/2017	(0.0, 1.0)	(0.0, 1.0)	(0.0, 1.0)
5	06/2017	(0.0, 1.0)	(0.0, 1.0)	(0.0, 1.0)
6	07/2017	(0.0, 1.0)	(0.0, 1.0)	(0.0, 1.0)
7	08/2017	(0.0, 1.0)	(0.0, 1.0)	(0.0, 1.0)
8	09/2017	(0.0, 1.0)	(0.0, 1.0)	(0.0, 1.0)
9	10/2017	(0.0, 1.0)	(0.0, 1.0)	(0.0, 1.0)
10	11/2017	(0.0, 1.0)	(0.0, 1.0)	(0.0, 1.0)
11	12/2017	(0.0, 1.0)	(0.0, 1.0)	(0.0, 1.0)

Conclusion

- Mise en place d’un modèle de clustering qui soit stable et interprétable

Amélioration du modèle

- Obtenir plus de renseignement sur nos client (âge,sexe…)