



Projet 5 :

Segmentez des clients d'un site e-commerce

Raphaël GIRAUDOT

Sommaire



1. Contexte et objectifs de la mission
2. Exploration des données à disposition
3. Préparation des données pour la modélisation
4. Pistes de modélisation & Méthodologie
5. Présentation des algorithmes de clustering
 - a. K-Means
 - b. DBScan
 - c. Performances des modèles
6. Modélisation : Exemple 1er fichier
 - a. K-Means
 - b. DBScan
 - c. Comparaison des performances et des clusters
7. Fréquence de mise à jour
8. Mode d'emploi du Clustering
 - a. Interprétation du clustering
 - b. Ajout de nouveaux clients
9. Conclusion & Pistes d'Amélioration

1. Contexte et objectifs de la mission

olist



1. Réaliser une segmentation des clients de Olist à partir de la base de données des clients.
2. Présenter la segmentation à l'équipe marketing pour une utilisation optimale.
3. Proposer un contrat de maintenance pour les mise à jour de la segmentation.

2. Exploration des données à disposition

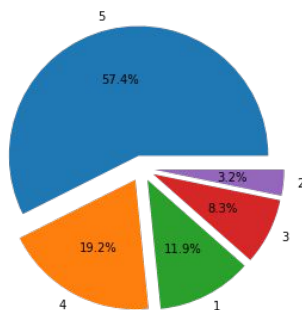
Statistiques Générales

- 96 096 clients
- 32 951 produits / 71 catégories de produits différentes
- 3 095 vendeurs

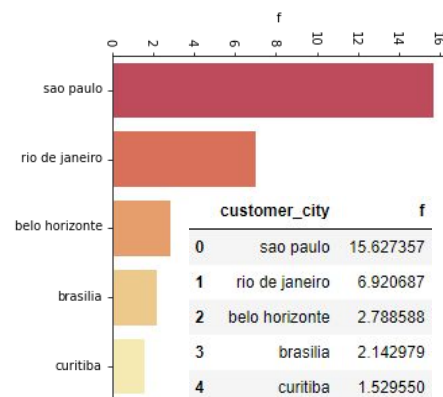
Période couverte :

- 04/09/2016
- 17/10/2018
- +/- 2ans

Note de satisfaction



Top 5 Villes



3. Préparation des données pour la modélisation

Fichier des travail




customer_unique_id	feature_1	feature_2	feature_n
<i>client_1</i>			
<i>client_2</i>			
<i>client_x</i>			

Features :

Générales

“RFM” :

- Récence : Date dernier achat 
- Fréquence : Fréquence d'achat 
- Montant : Montant des achats 



Périodes considérées :

Base : 01/01/2017 -> 31/12/2017

Décalage : +1 & +3 Mois

3. Préparation des données pour la modélisation

Feature Engineering

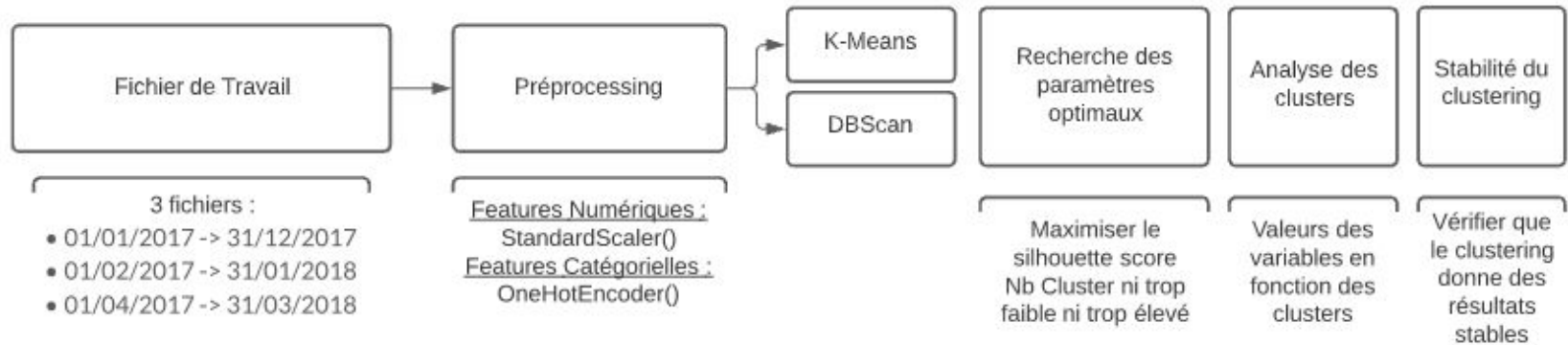
Variables Générales :

- **nb_orders** : nombre de commandes
- **mean_review_score** : moyenne de notes des avis de satisfaction
- **most_freq_payment_type** : moyen de paiement le plus fréquent
- **nb_voucher** : nombre de bons d'achats utilisés
- **voucher_value** : montant des bons d'achats utilisés
- **mean_payment_installments** : nombre moyen de paiement en plusieurs fois

Variables "RFM" :

- Récence
 - **nb_days_last_order** : nombre de jour depuis la dernière commande
- Fréquence
 - **most_freq_cat_name** : catégorie de produit les plus commandés
- Montant
 - **payment_value** : valeurs de tous les paiements
 - **average_basket** : montant moyen du panier
 - **min_basket** : panier minimum
 - **max_basket** : panier maximum
 - **mean_fdp** : frais de ports moyen par commandes

4. Pistes de modélisation & Méthodologie



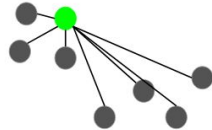
5. Présentation des algorithmes de clustering

a. K-Means : Fonctionnement

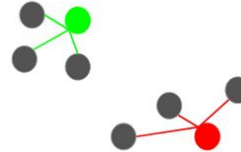
Paramètres : nombre de clusters

Initialisation K-Means : k-means++

1. Sélection aléatoire du 1^{er} centroïde & calcul des distances aux individus

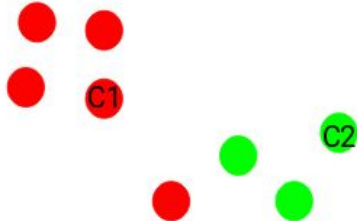


2. 2^{ème} centroïde = individus le + éloigné du 1^{er} centroïde etc...

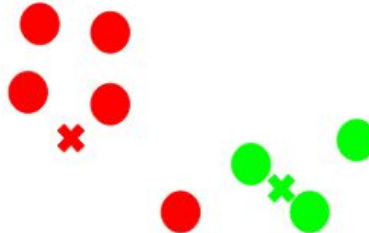


Étapes K-Means :

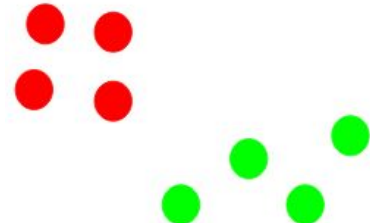
1. Association des individus au centroïde du cluster le plus proche



2. Re-Calcul des centroïdes : barycentre des individus



3. Itérer sur 1 & 3 jusqu'à convergence



5. Présentation des algorithmes de clustering

b. DBScan : Fonctionnement

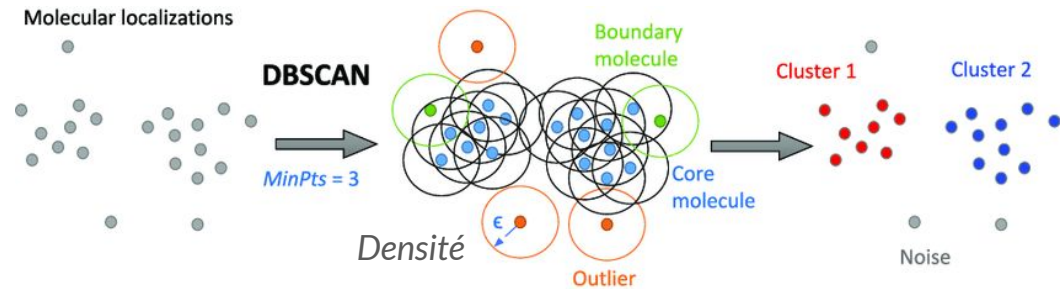
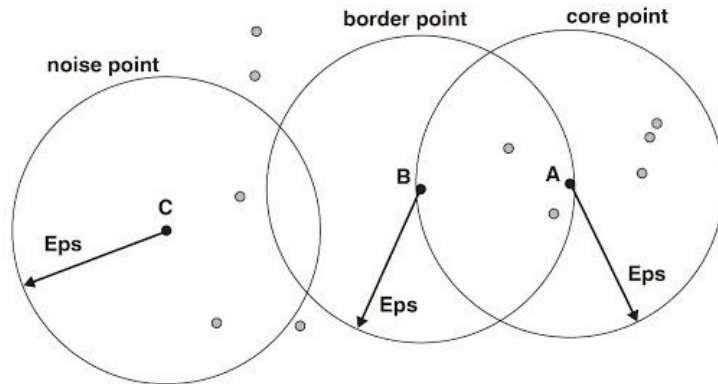
Paramètres :

- epsilon : rayon du voisinage du point
- min_samples : nb minimum de points dans le voisinage pour l'intégrer dans un cluster

Fonctionnement :

- Classifier les points, individus, en fonctions des paramètres fixés
- Associer clusters

exemple : min_samples = 3



5. Présentation des algorithmes de clustering

c. Performances des modèles

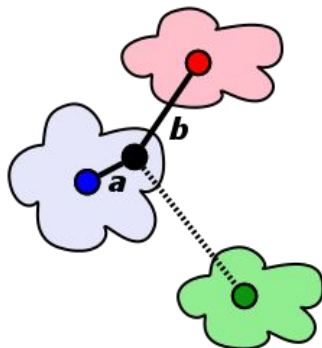
Bon clustering = Individus dans des groupes distincts & éloignés



Minimiser a
Maximiser b

Score Silhouette, [-1:1]:

- a : distance d'un point au centroïde de son groupe
- b : distance d'un point au groupe le plus proche

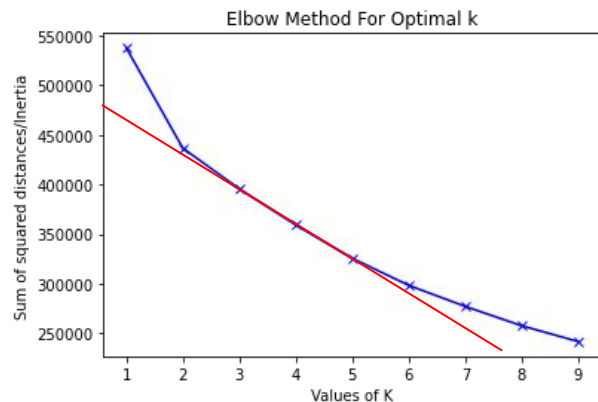


$$SSI_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

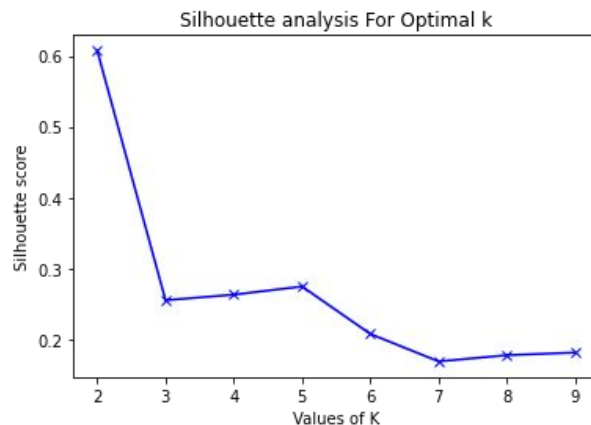
6. Modélisation : Exemple 1er fichier

a. K-Means

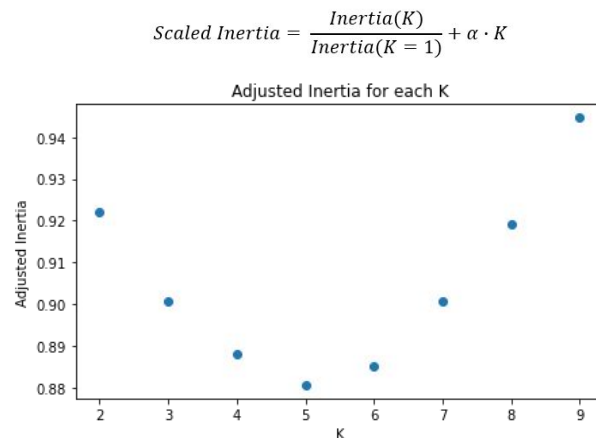
Choix du nombre de cluster :



Changement de pente significatif



Maximiser le score de silhouette



Minimiser l'inertie pondérée
Source

➡ 5 clusters

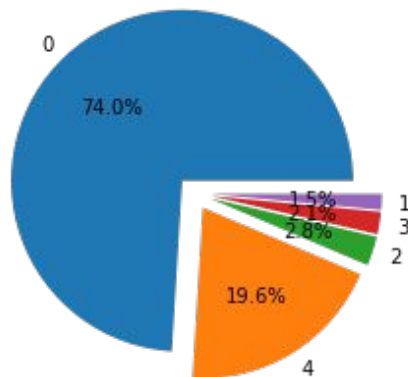
6. Modélisation : Exemple 1er fichier

a. K-Means

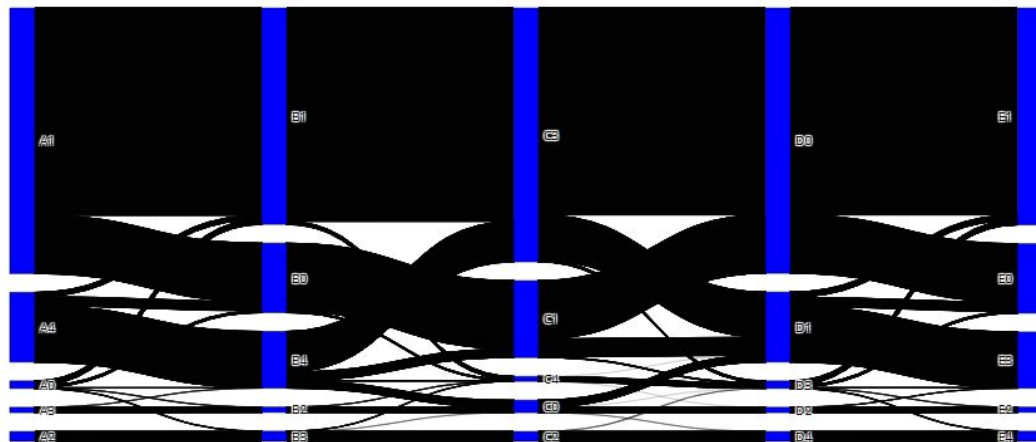
Performance :

- Silhouette Score : 0.27
- Adjusted Rand Score (5 itérations) : 0.85

Répartition des clusters :



Stabilité du clustering



6. Modélisation : Exemple 1er fichier

b. DBScan

Recherche des paramètres optimaux :

- Epsilon
- Min_samples

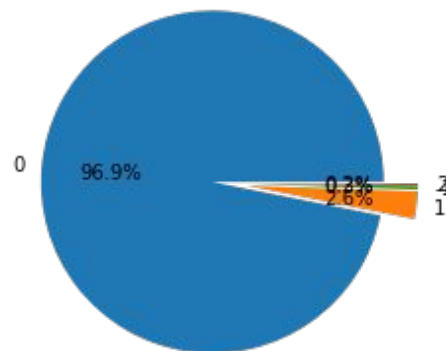
1. Itération sur plusieurs valeurs pour chaque paramètres
2. Recherche du couple de paramètres maximisant le silhouette_score tout en ayant un nombre de clusters et un % de bruit raisonnable

	epsilon	n_samples	n_clusters	n_noise	noise_percent	silhouette_score
15	5.0	50.0	3.0	151.0	0.346259	0.431713
14	5.0	40.0	3.0	144.0	0.330207	0.431701
11	4.5	50.0	3.0	180.0	0.412759	0.430589
5	4.0	20.0	3.0	183.0	0.419638	0.430478

Performance :

- Silhouette Score : 0.43

Répartition des clusters :

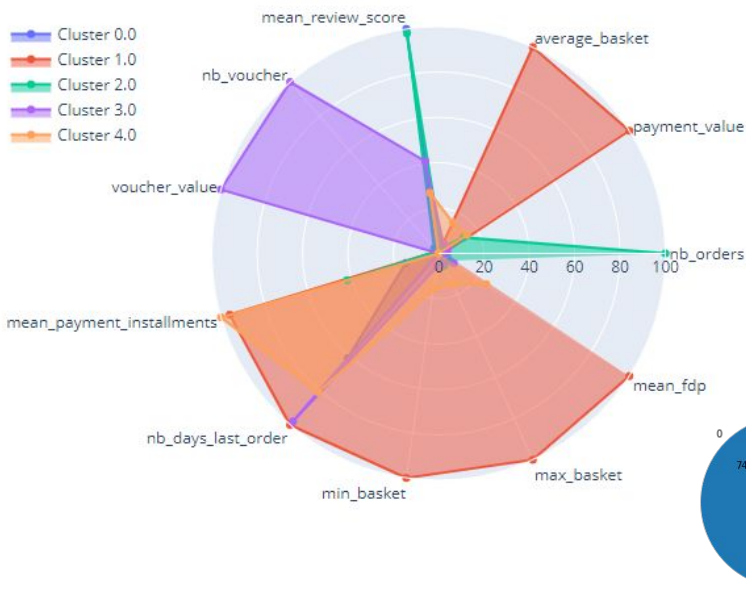


6. Modélisation : Exemple 1er fichier

c. Comparaison des performances et des clusters

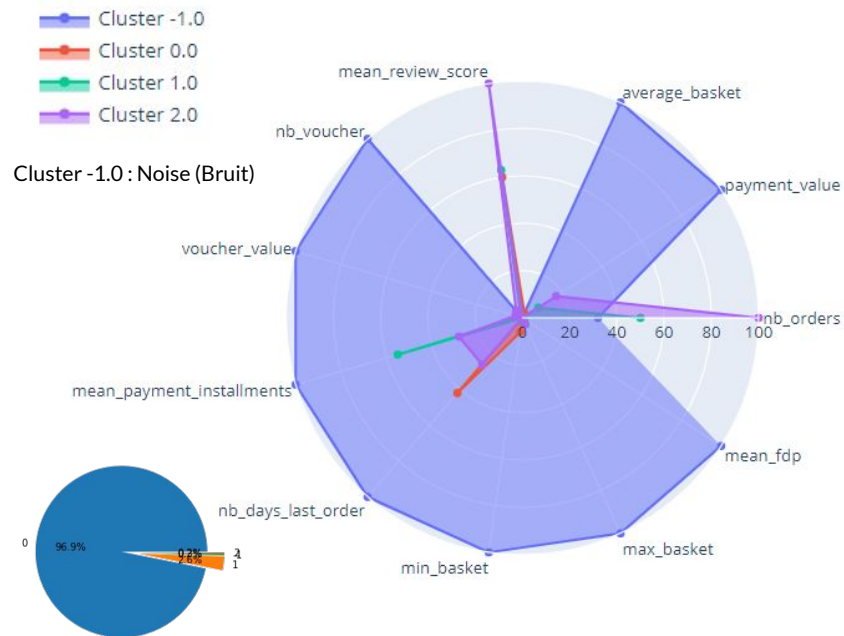
K-Means

Silhouette Score : 0.27



DBScan

Silhouette Score : 0.43

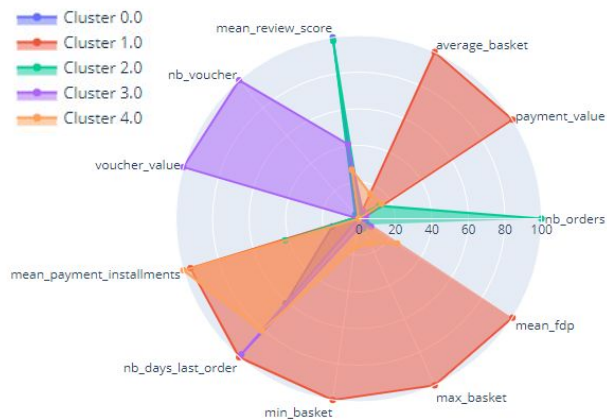


7. Fréquence de mise à jour

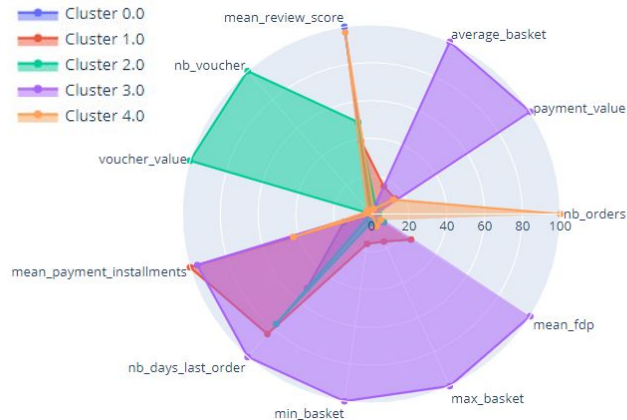
Comparaison des clusterings : Features numériques



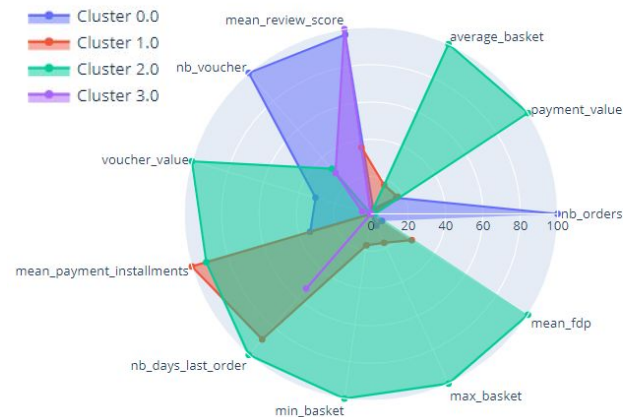
1er fichier :
nb_clust = 5



2ème fichier :
nb_clust = 5



3ème fichier :
nb_clust = 4



7. Fréquence de mise à jour

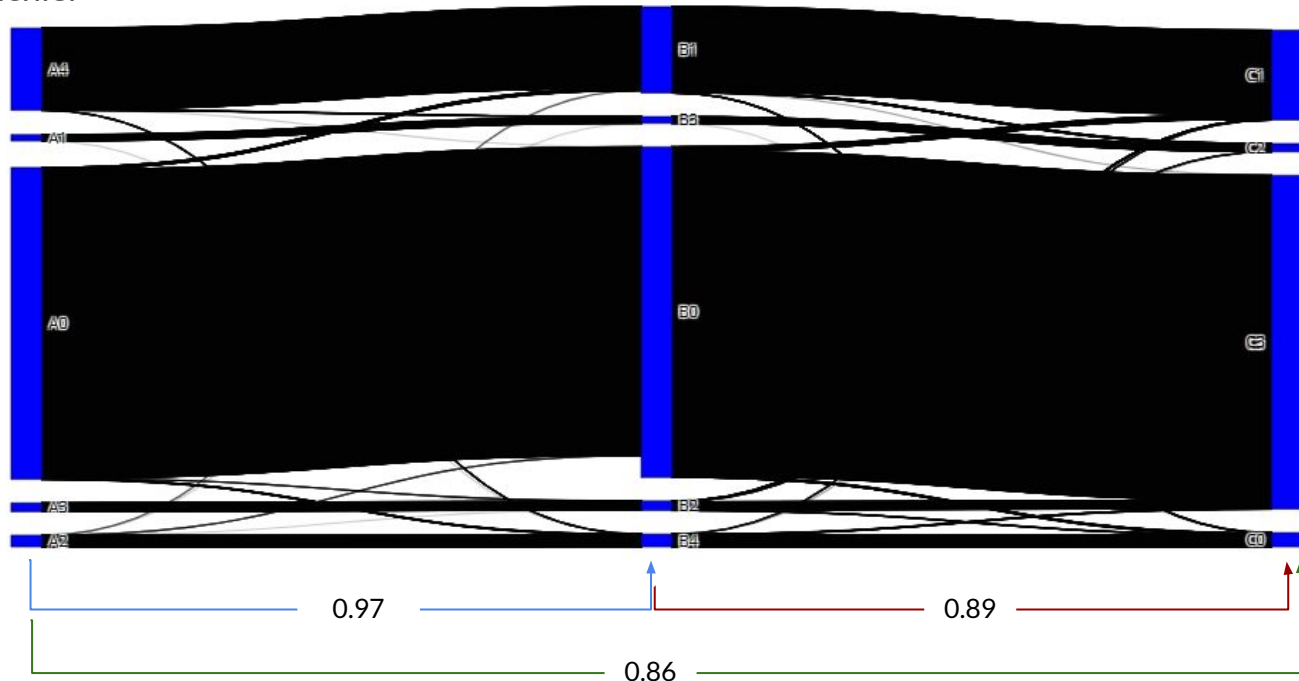
Comparaison des clusterings : Stabilité au fil du temps



1er fichier

2ème fichier

3ème fichier



Adjusted Rand
Score :

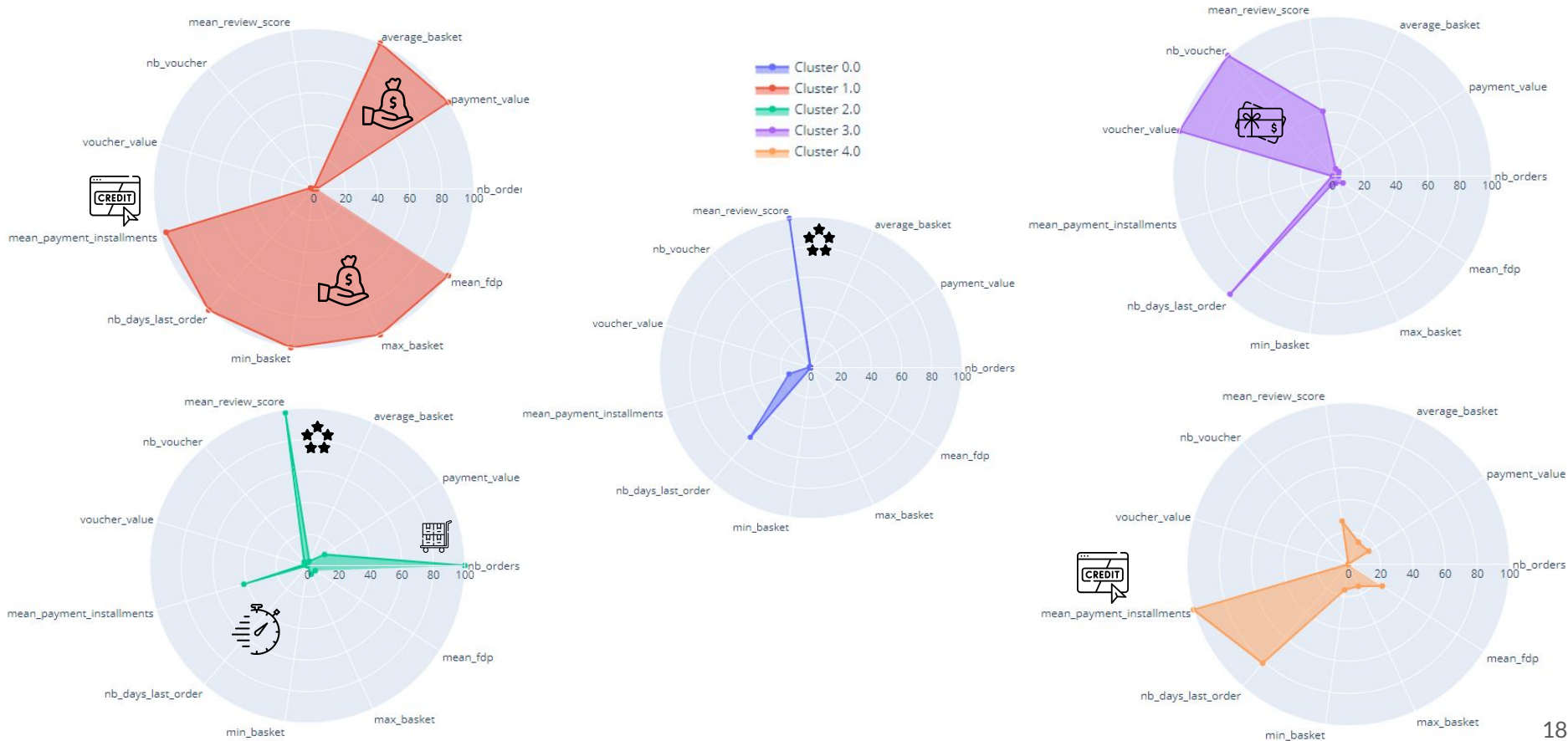
0.97

0.86

0.89

8. Mode d'emploi du clustering

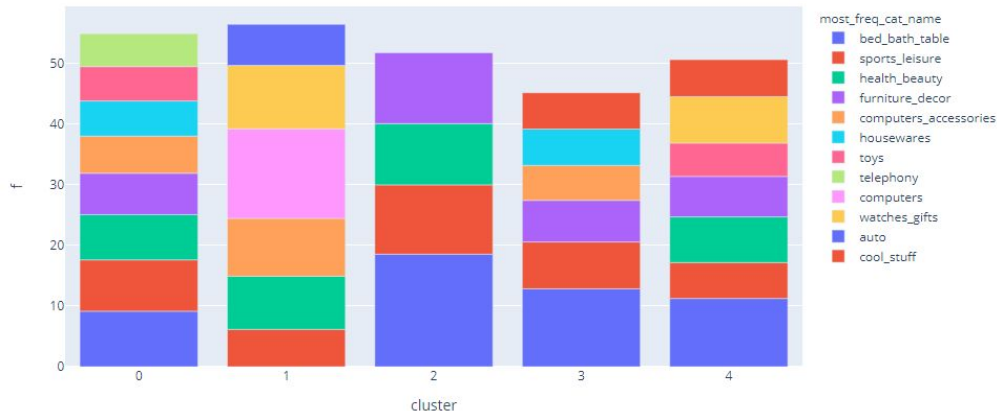
a. Interprétation du clustering



8. Mode d'emploi du clustering

a. Interprétation du clustering

Catégories de produits les plus commandés $f > 5\%$



- 8/12 catégories
- Cluster le + hétérogène
- Unique cluster avec Telephony



- 6/12 catégories
- Unique cluster avec Auto & Computers



- 4/12 catégories
- Maison & Corps



- 6/12 catégories
- Gadgets & Maison

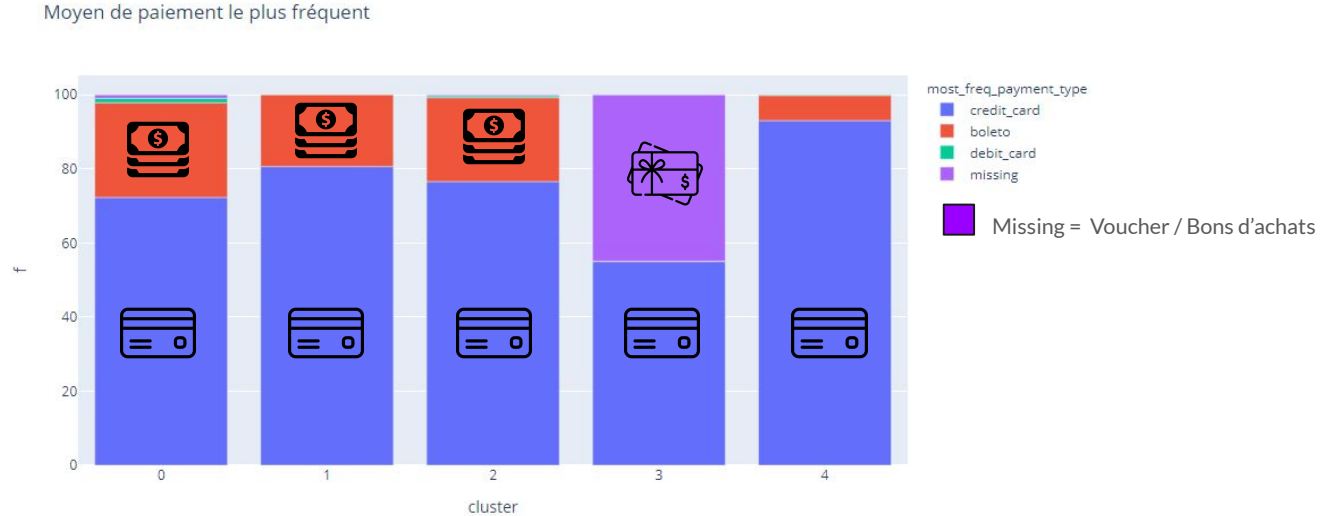


- 7/12 catégories
- Maison & Corps



8. Mode d'emploi du clustering

a. Interprétation du clustering



0. Tous moyens de paiements
1. 20% Billets, 80% CB
2. 20% Billets, 75% CB, <5% Debit Card
3. Utilisation de bons d'achats (+/-40%)
4. Utilisation de CB (+/- 90%)

8. Mode d'emploi du clustering

b. Ajout des nouveaux clients

```
k_means.predict(df_nouveaux_clients)
```

Vérification de la stabilité sur de nouvelles données :

```
# silhouette score sur le train set  
silhouette_score(train, k_means['kmeans'].labels_)
```

```
0.2049283130016531
```

```
# silhouette score sur le test set  
silhouette_score(test, k_means.predict(test))
```

```
0.19841880312355437
```

9. Conclusion & Pistes d'Amélioration



Bons résultats mais comment augmenter les performances ?

Features Engineering :

- Changer la période d'observation : trimestre, semestre, mensuel, etc...
- Regrouper les catégories de produits pour réduire le nombre de modalités
- Avis du “métier” pour le choix des variables