Segmenter des clients d'un site e-commerce

Présenté par : *M. Elhadi BELGHACHE*

Plan

1. Contexte & Problématique
2. Nettoyage & Analyse Exploratoire
3. Essais de Segmentation
4. Evaluation & Interprétabilité
5. Synthèse & Conclusion

1. Contexte & Problématique

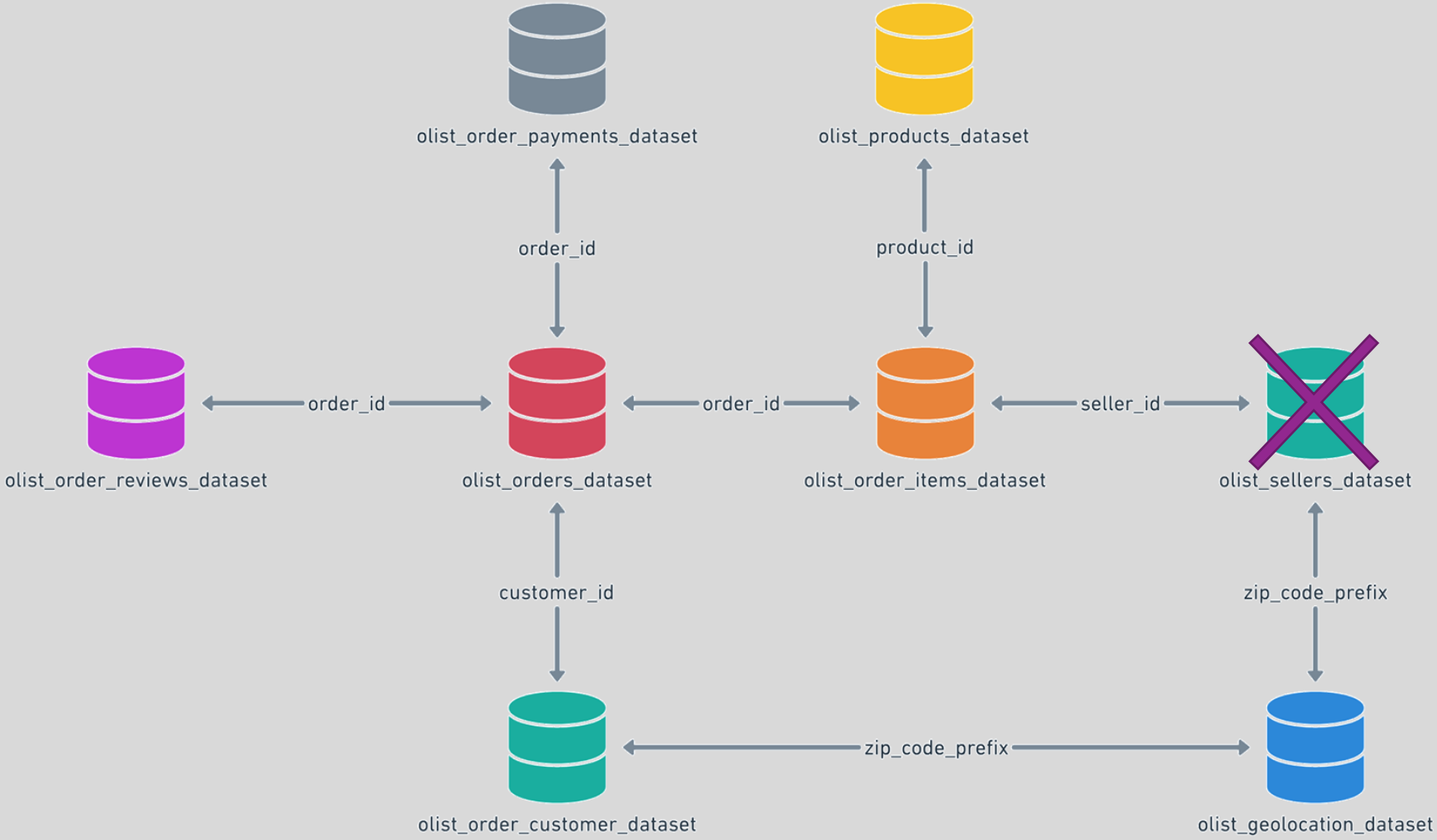
1. Contexte & Problématique

- **Qui ?** Olist
- **Quoi ?**
  - Segmenter clients
    + Fournir description actionnable
  - Contrat maintenance (stabilité)
- **Pourquoi ?**
  - Campagnes de communication ciblées

- **Comment ?**
  "Brazilian E-Commerce Public Dataset by Olist"
  - 2017-2018
  - 100 000 commandes
    + Clients
    + Produits
    + Reviews
    + Localisation

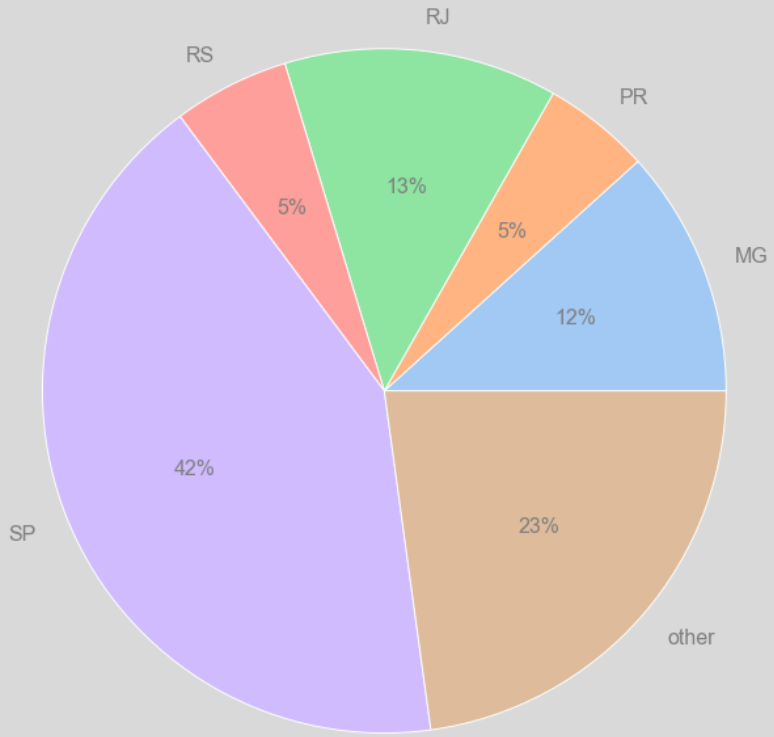
2. Nettoyage & Analyse Exploratoire

2.1. Data Sets : BDD **Olist**

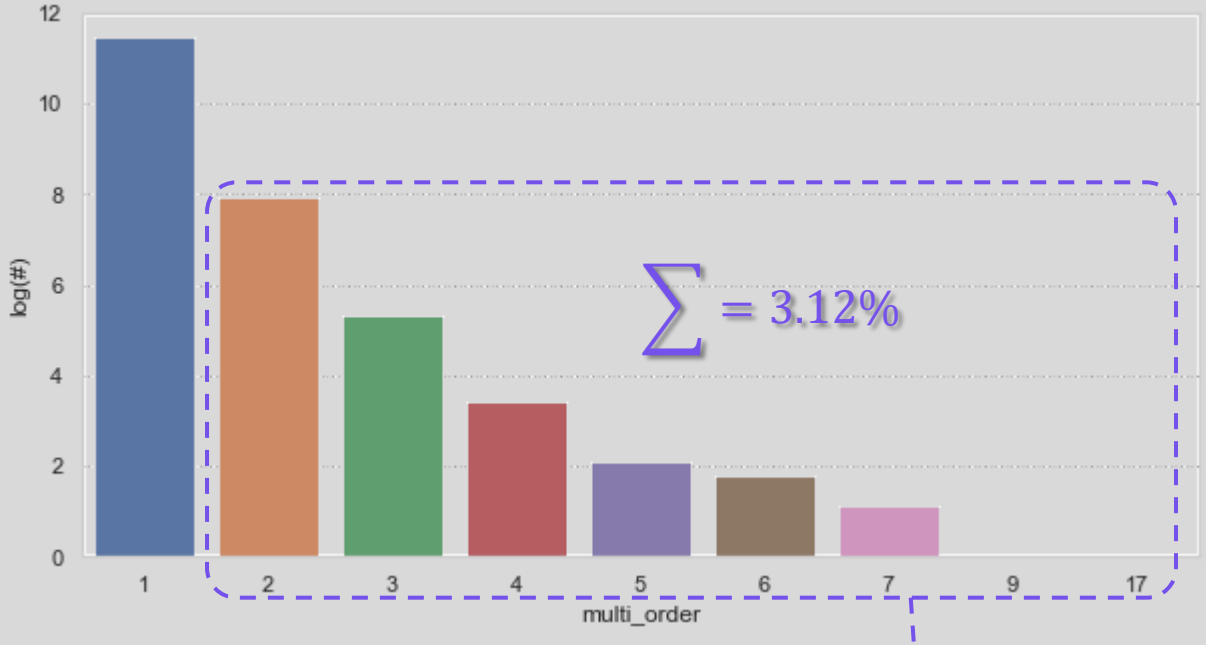


- Schéma *Base De Données* ***Olist***-

2.1.1. Data Sets : Clients



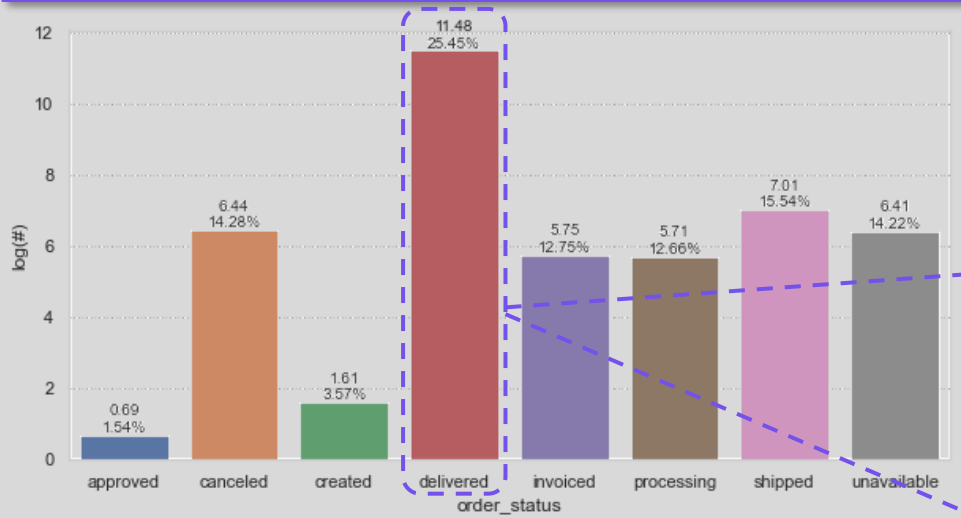
- Clients par Régions -



- Fréquence (au log) des Commandes -

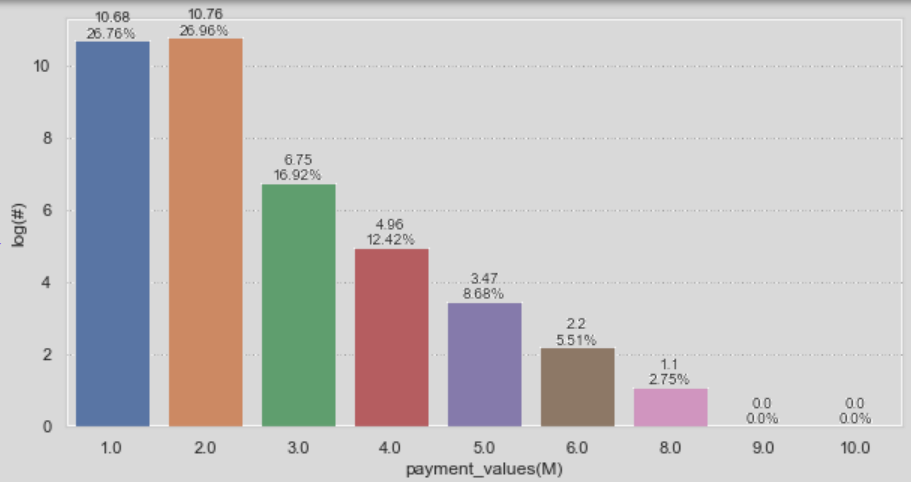
$$Multi\_order_i(Fréquence) = \begin{cases} 1 & si\ plus\ d'une\ commande \\ 0 & sinon \end{cases}$$

2.1.2. Data Sets : Commandes (1/2)



- Nombre des Commandes (au log) par État -

Garder Commandes **livrées**



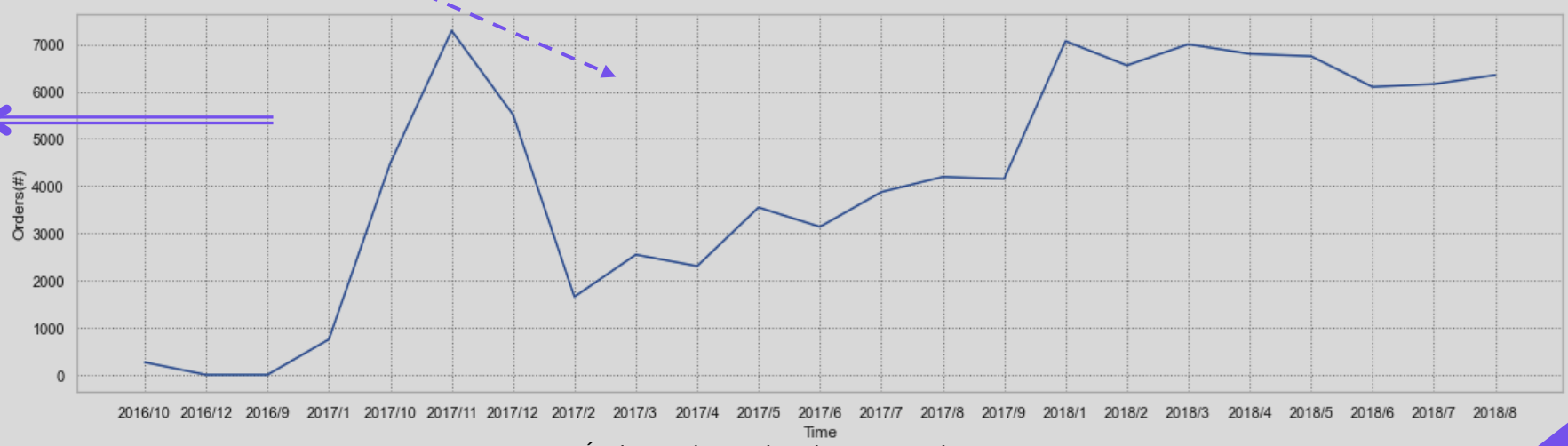
*- Nombre (au log) des **Montants** des Commandes par intervalles (tranches de 1000 REAIS) -*

$$\boldsymbol{Freshness_i(Récence)} = \frac{\max Time}{Time_i}$$

$$\boldsymbol{Freshness} \in [0,1]$$

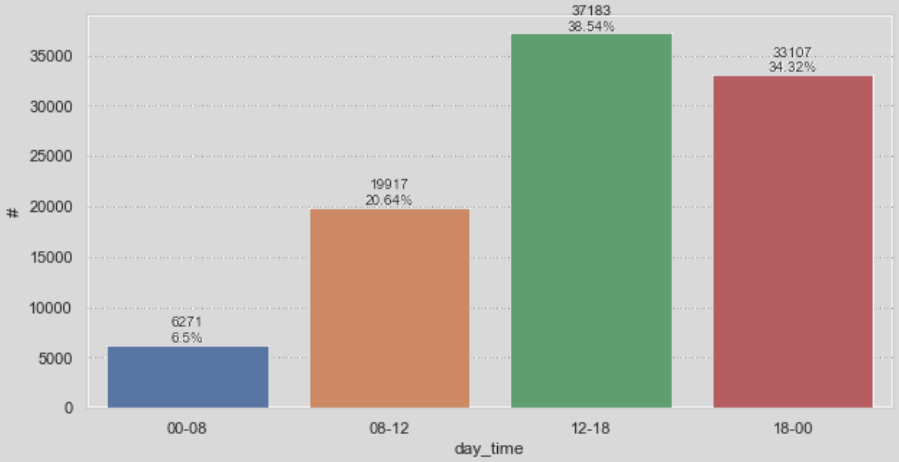
Commandes **moins** Récentes

Commandes **plus** Récentes

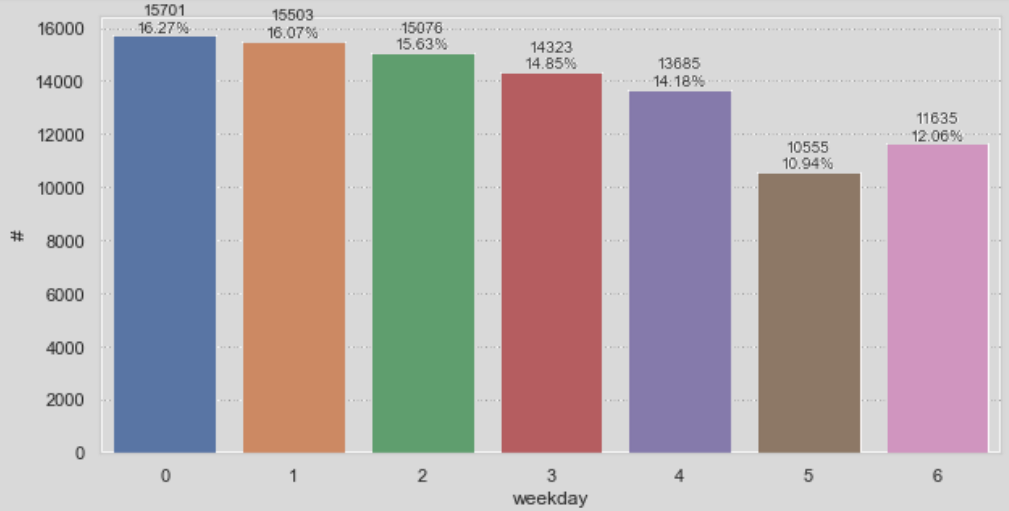


- Évolution du nombre de Commandes par Mois -

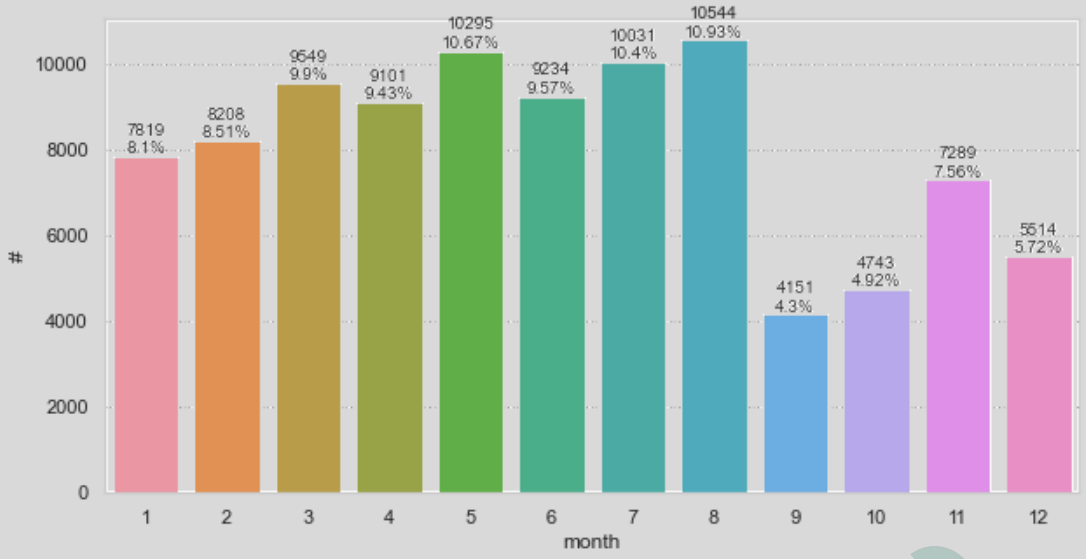
2.1.2. Data Sets : Commandes (2/2)



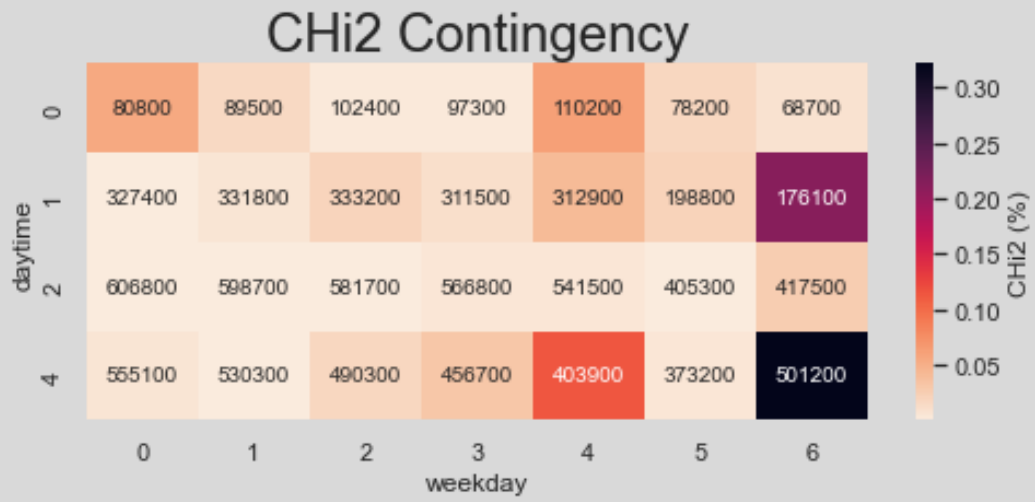
*- Nombre des Commandes par **Moment** du jour -*



*- Nombre des Commandes par **Jour** de semaine -*

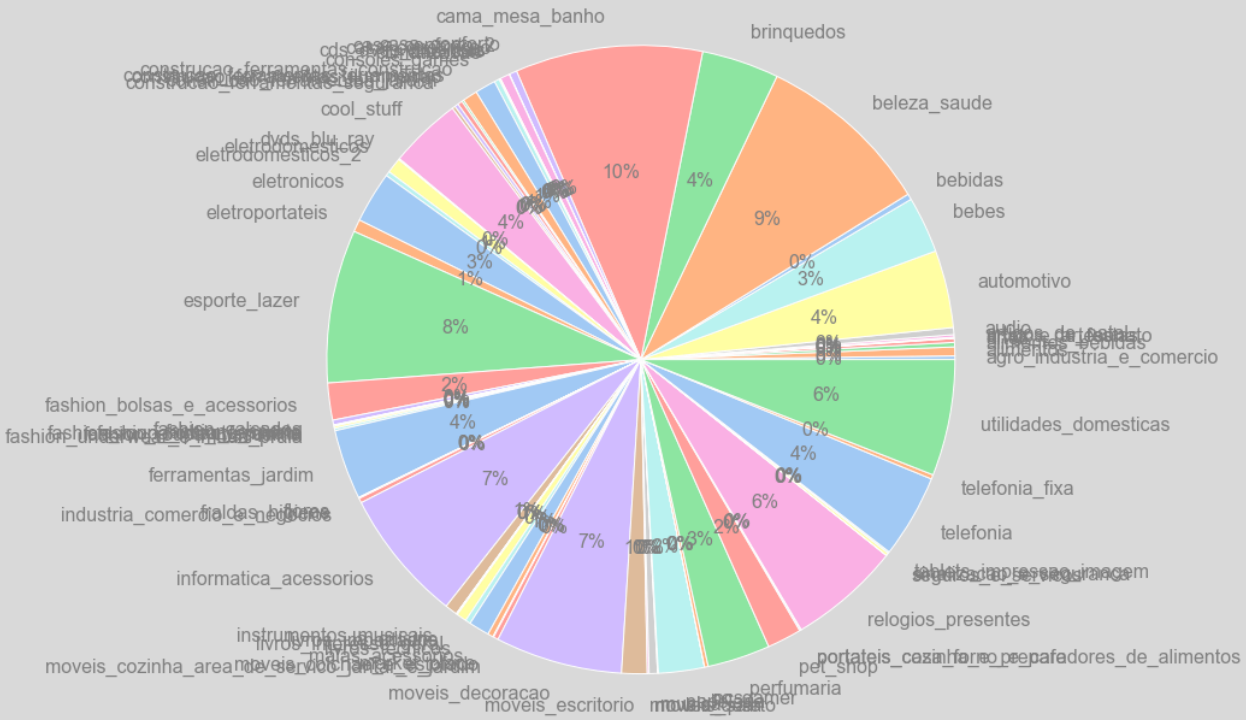


*- Nombre des Commandes par **Mois** de l'année -*

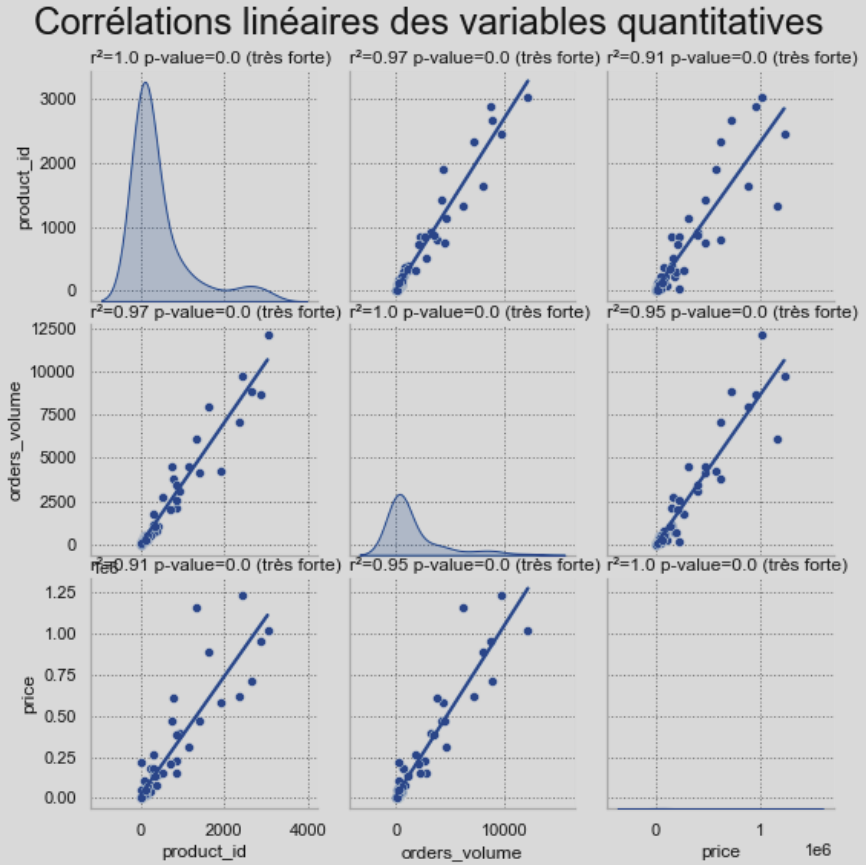


*- Nombre des Commandes par **Moment** du jour et **Jour** de semaine -*

2.1.3. Data Sets : Produits (catégories)

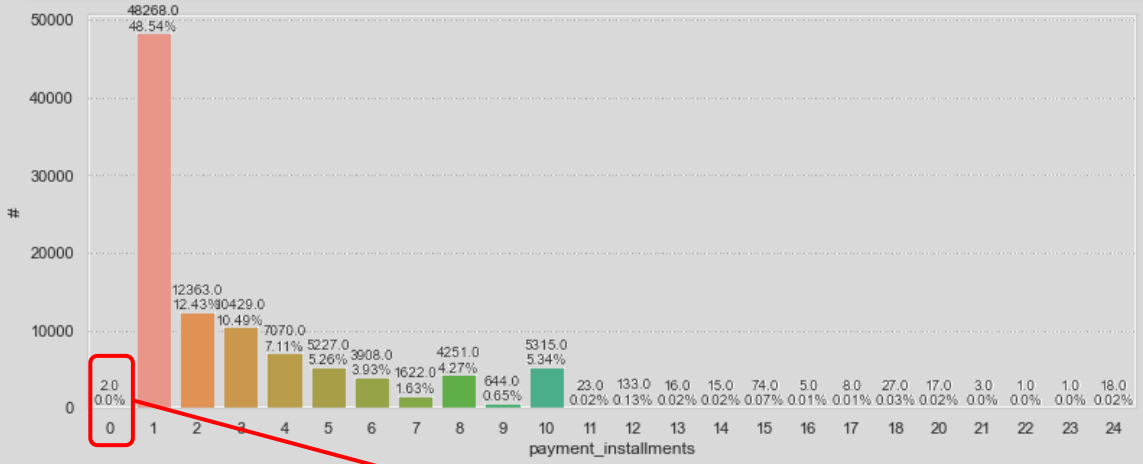


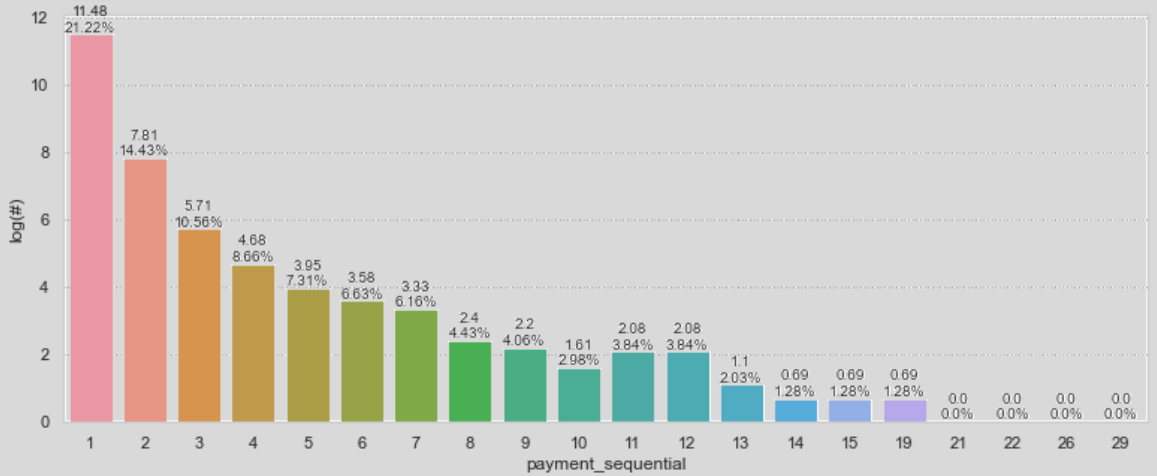
- Produits par Catégorie -



- Corrélations entre le Nombre, le Volume commandé et le Prix total des Produit pat Catégorie -

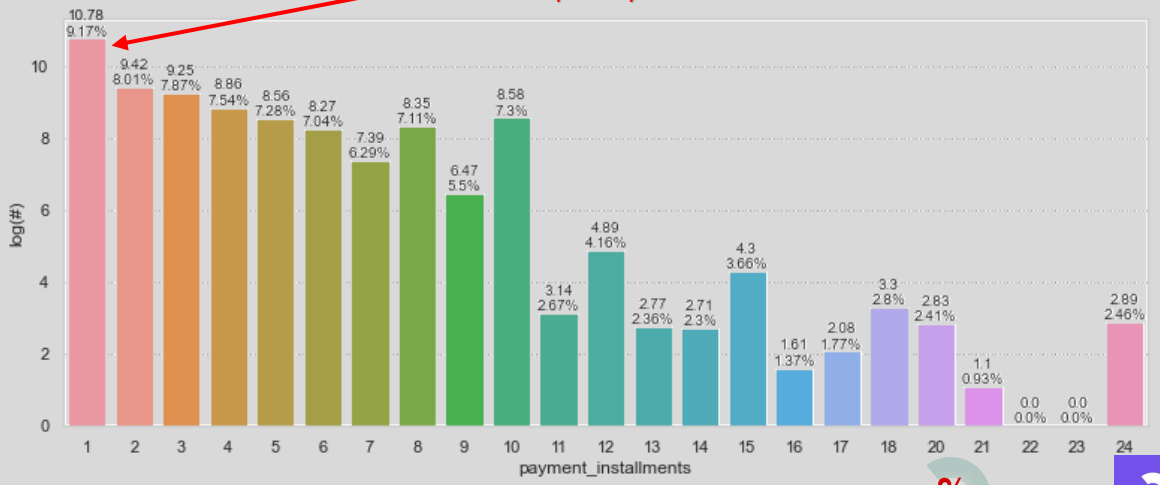
2.1.4. Data Sets : Paiements

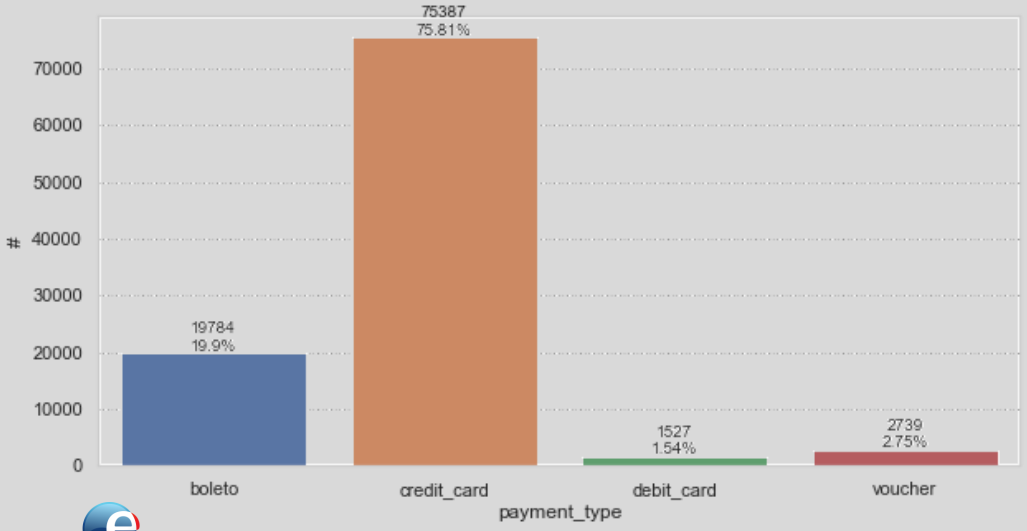




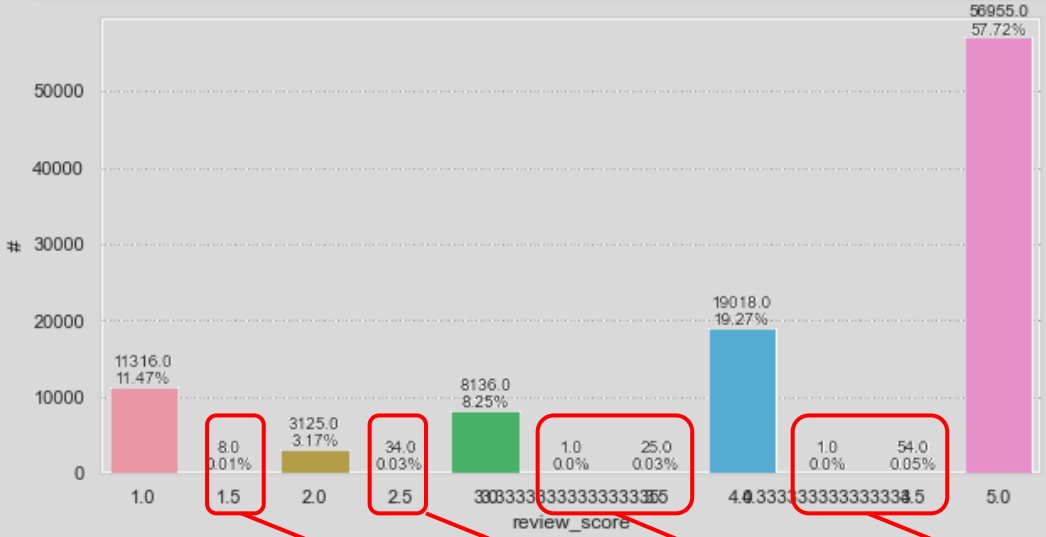
order_id	payment_sequential	payment_installments	payment_value
1a57108394169c0b47d8f876acc9ba2d	2	0	129.94
744bade1fcf9ff3f31d860ace076d422	2	0	58.69

Remplacer par **1**

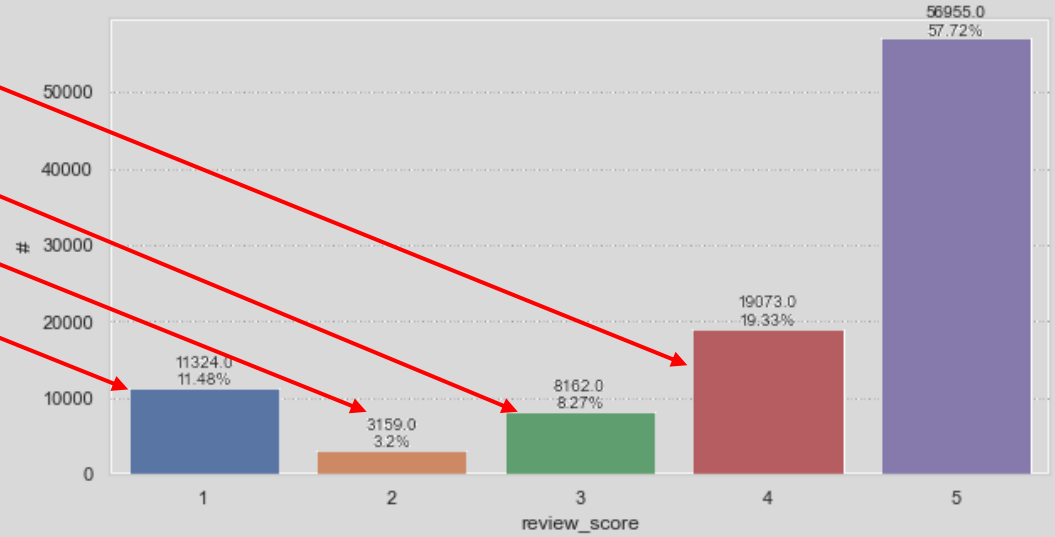




2.1.5. Data Sets : Reviews



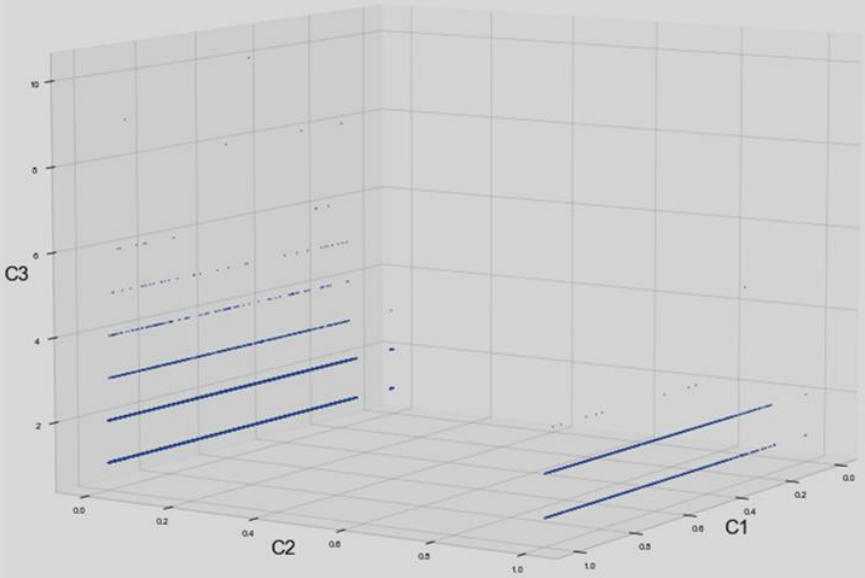
Arrondir les valeurs réelles en entiers



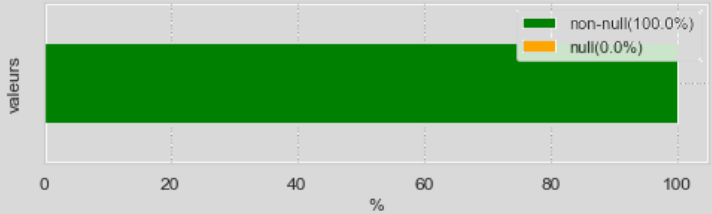
2.2. Data Set Final [91476 × 14]

	customer_unique_id	Freshness(R)	multi_orders(F)	payment_value(M)	review_score(S)	daytime	weekday	month	orders_volume	product_category_name	payment_type	payment_sequential	payment_installments	customer_state
67282	bc0c41b70e1126510082daf85aece147	0.589774	0	1	5	3	2	3	1	telefonia	credit_card	1	3	RJ
39610	6e9359d145dcf22cc8c1d9116f403f17	0.948221	0	1	5	3	5	9	1	telefonia	boleto	1	1	SP
5400	0f0692335c7742aeaa35bc714c812659	0.608716	0	1	5	3	3	3	1	esporte_lazer	boleto	1	1	MG
43480	797c63d3c9502e7f4fbd1e01f26ddd36	0.817307	0	2	5	3	5	6	5	fashion_bolsas_e_acessorios	credit_card	1	2	other
819	0246e728843e079b72ae99cbfe711acd	0.785114	0	1	5	0	6	6	1	eletronicos	boleto	1	1	SP
21662	3c7399e057696d1a324871cb80f46745	0.442015	0	1	5	2	1	11	1	ferramentas_jardim	credit_card	1	3	RS
28690	501a8a414f0ae0164f0da2974541b2b9	0.547212	0	1	1	2	4	1	1	esporte_lazer	boleto	1	1	SP
81504	e3b3e757b651bfba6f0144c1096f9864	0.750666	0	1	5	1	2	8	1	automotivo	credit_card	1	1	RJ
39428	6e09ac06c6d4d178cf0fa8703f7a3732	0.807235	0	2	5	3	5	6	1	cool_stuff	credit_card	1	5	SP
83430	e94315e85e8cd829a025c96474438549	0.741076	0	2	3	3	2	8	1	beleza_saude	credit_card	1	5	SP

*- Échantillon aléatoire du data set final ([**91476 × 14**]) -*



- Projection RFM -

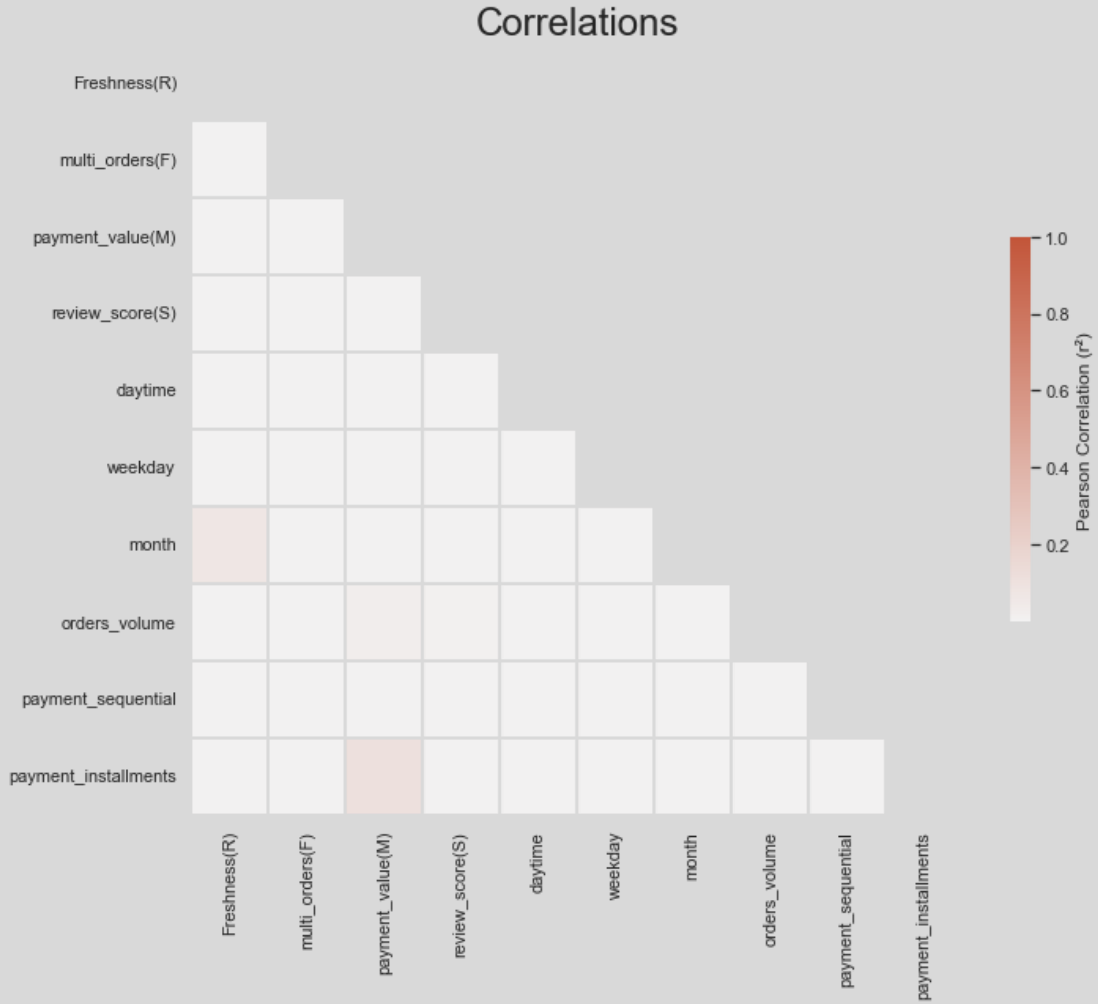


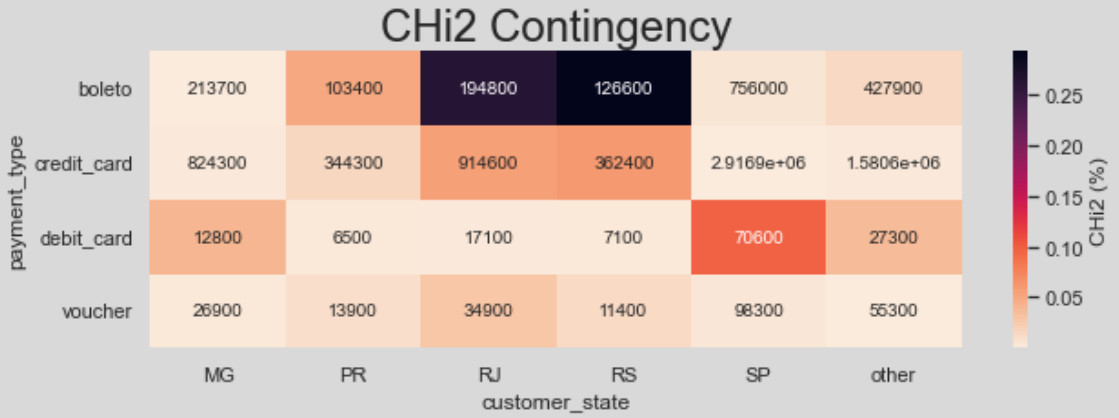
- Taux des valeurs manquantes -

	Freshness(R)	multi_orders(F)	payment_value(M)	review_score(S)	daytime	weekday	month	orders_volume	payment_sequential	payment_installments
count	91476	91476	91476	91476	91476	91476	91476	91476	91476	91476
mean	0.659525	0.029571	1.539683	4.151635	2.046952	3.250077	6.387315	1.188027	1.044602	2.916131
std	0.219601	0.1694	0.534249	1.280064	0.916291	1.942089	3.226861	0.829505	0.368061	2.702679
min	0	0	1	1	0	0	0	1	1	1
25%	0.503616	0	1	4	1	2	4	1	1	1
50%	0.686976	0	2	5	2	3	7	1	1	2
75%	0.837279	0	2	5	3	5	9	1	1	4
max	1	1	10	5	3	6	11	52	26	24

- Distribution des variables -

2.3. Corrélations & CHi2

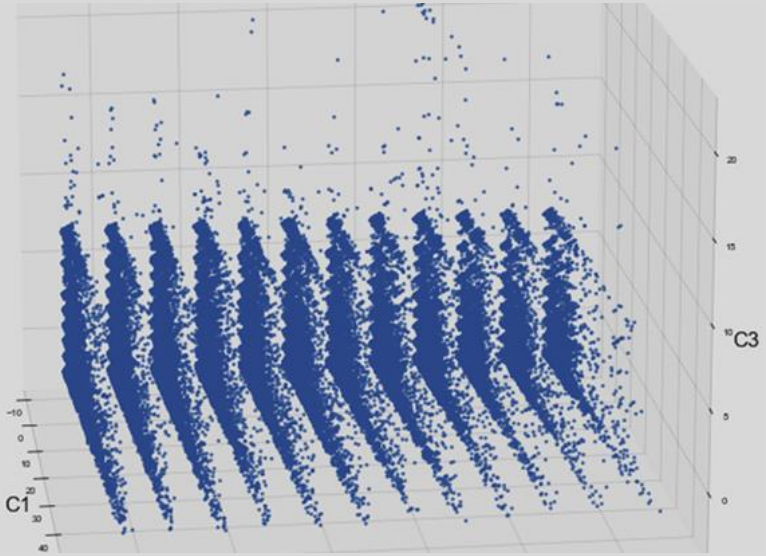




3. Essais de Segmentation

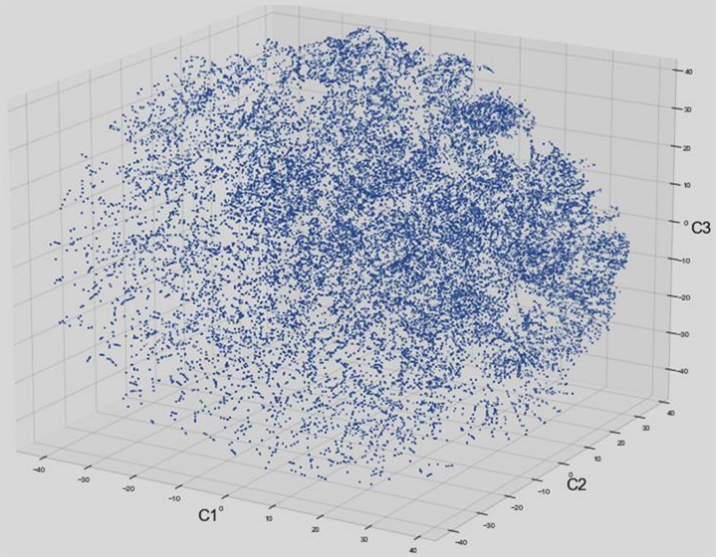
3.1. Réduction Dimensionnelle

PCA



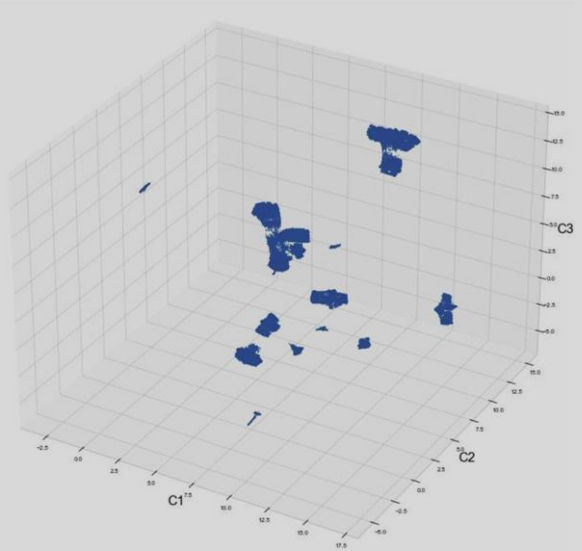
- ✓ Distinction/Séparabilité
- × Non Homogénéité

TSNE



- × Distinction/Séparabilité
- ✓ Non Homogénéité

UMAP



- ✓ Distinction/Séparabilité
- ✓ Non Homogénéité

3.2. Démarche : **Méthodes & Scores**

Méthodes

- *Par Densité* : **DBSCAN**
- *Par Agglomération*
  - Lien Simple (**HCS**)
- *Par Inertie*
  - Sur Données Numérique
    + **KMeans**
    + **KMeans++**
    + **MiniBatchKMeans(++)**
    + **BisectingKMeans(++)**
  - Sur Données Catégorielles : **KModes**
  - Sur Données Mixtes: **KPrototypes**

Scores

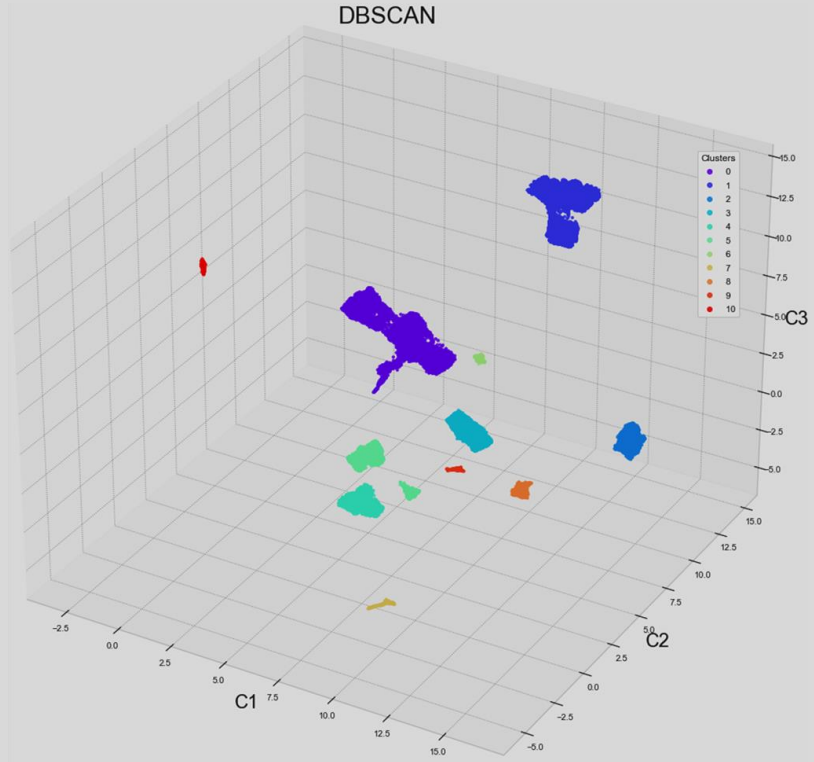
- **Variance (*Calinski-Harabasz Index*)**
  - Élevée pour des clusters denses et bien séparés
- **Similarité (*Davies-Bouldin Index*) [0,1]**
  - Similarité moyenne entre chaque cluster et son cluster le plus similaire
  - Faible pour un meilleur clustering
- **Silhouette [-1,1]**

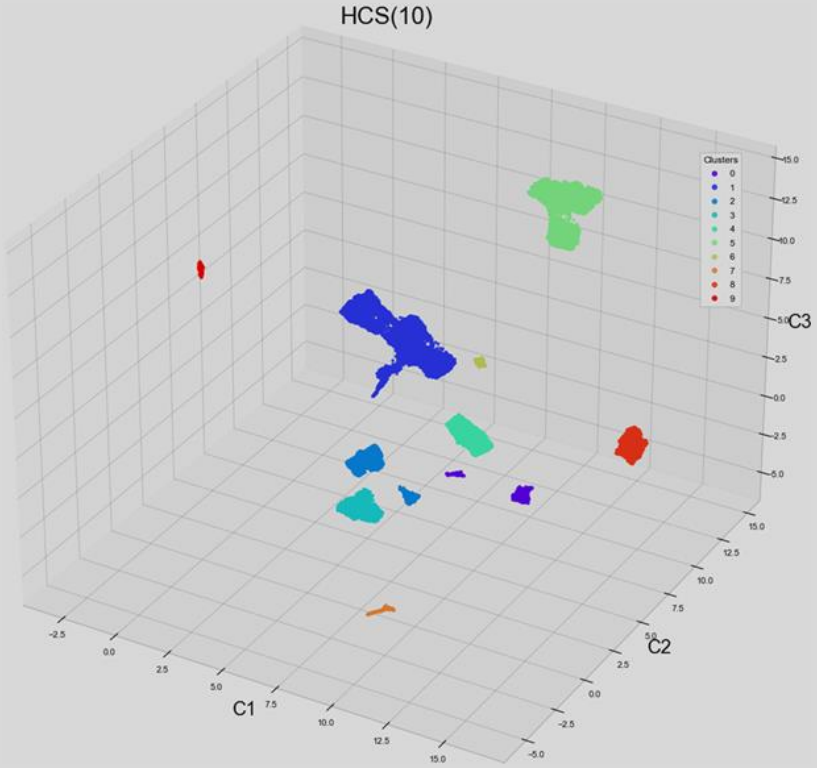
  -1 : mauvais clustering

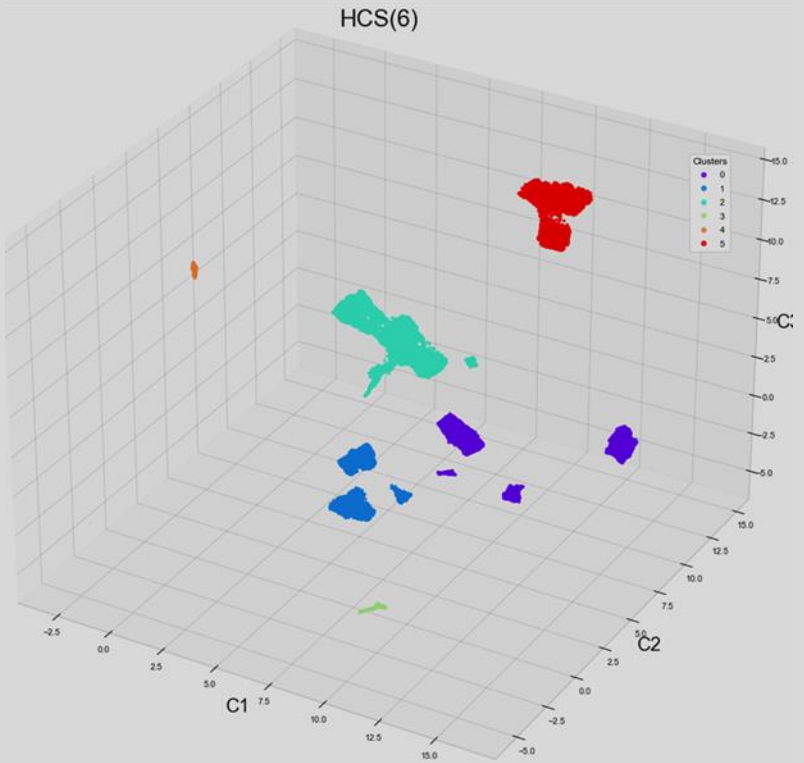
  0 : chevauchement

  +1 : clusters denses
- **Temps (s)**

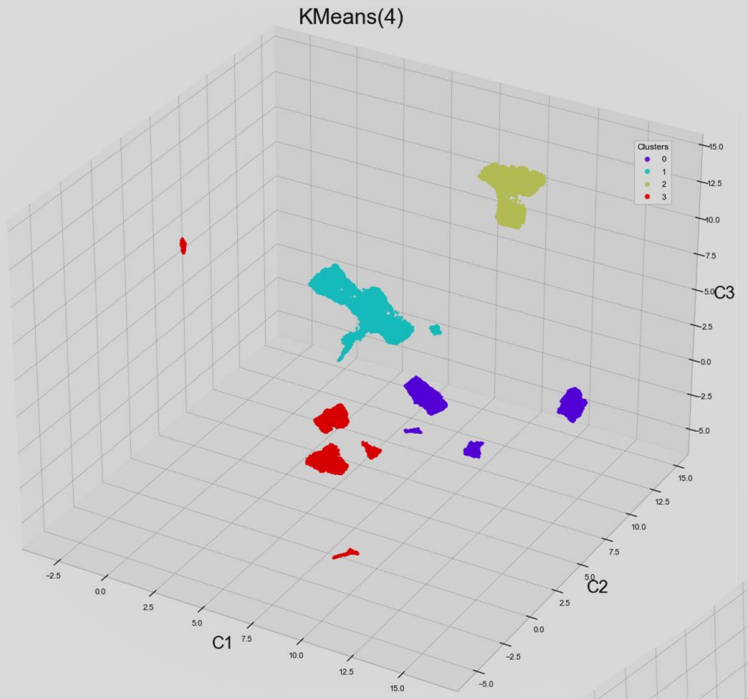
3.3.1. Clustering : **DBSCAN & HCS**

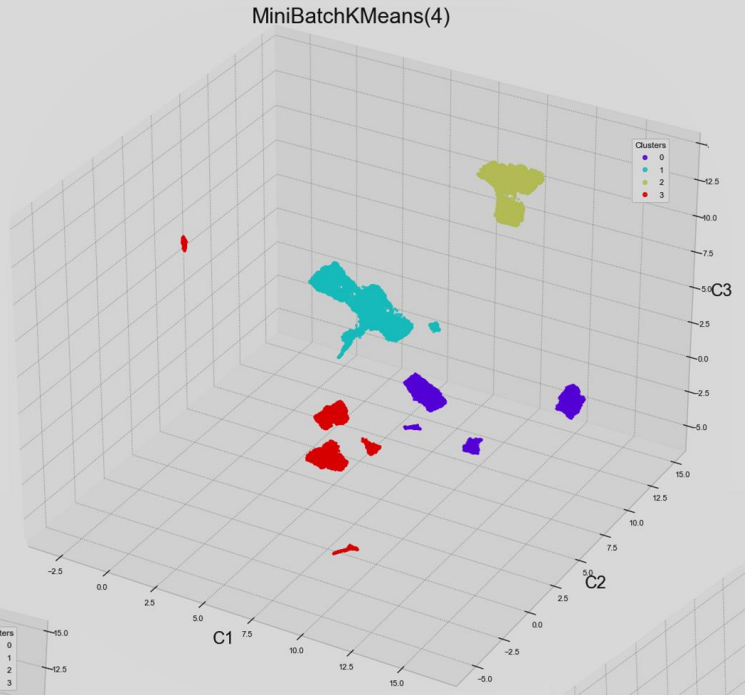


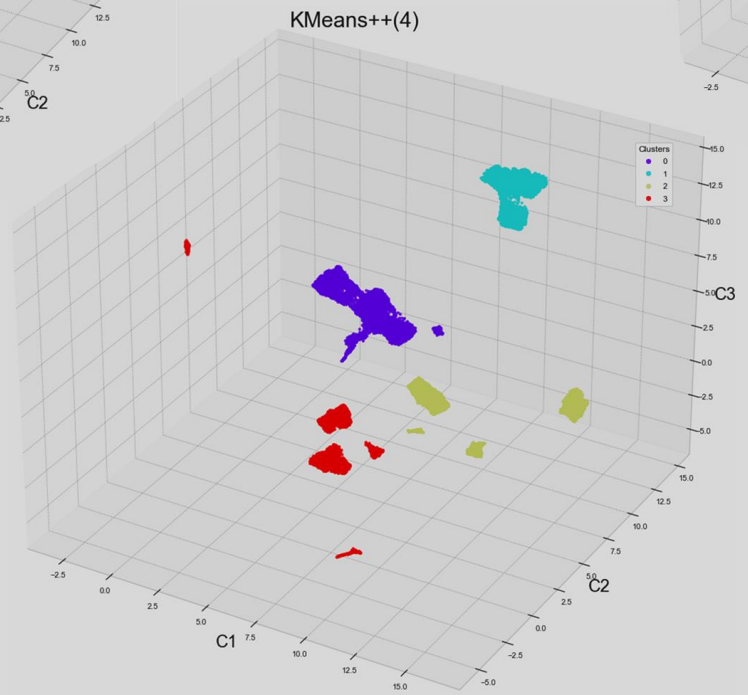


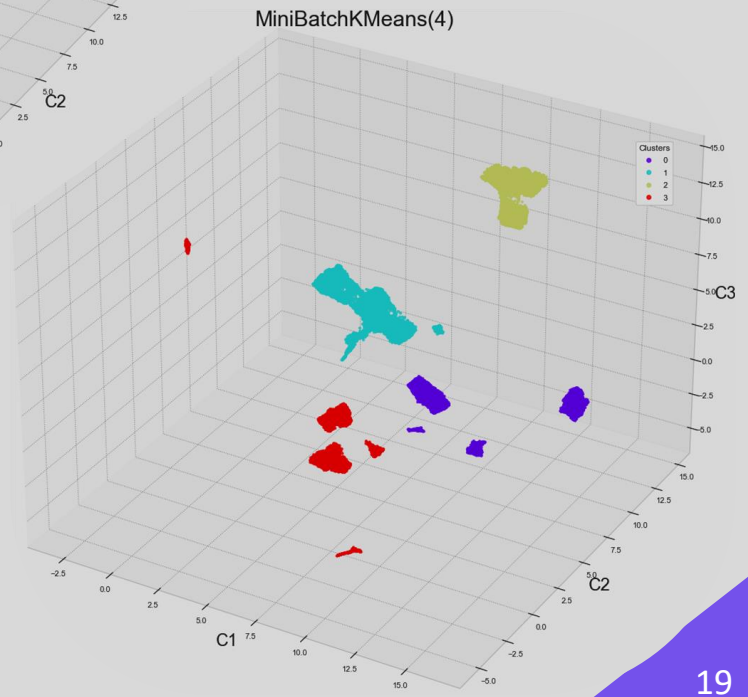


3.3.2. Clustering : **Kmeans, ++, Mini, Bisect**



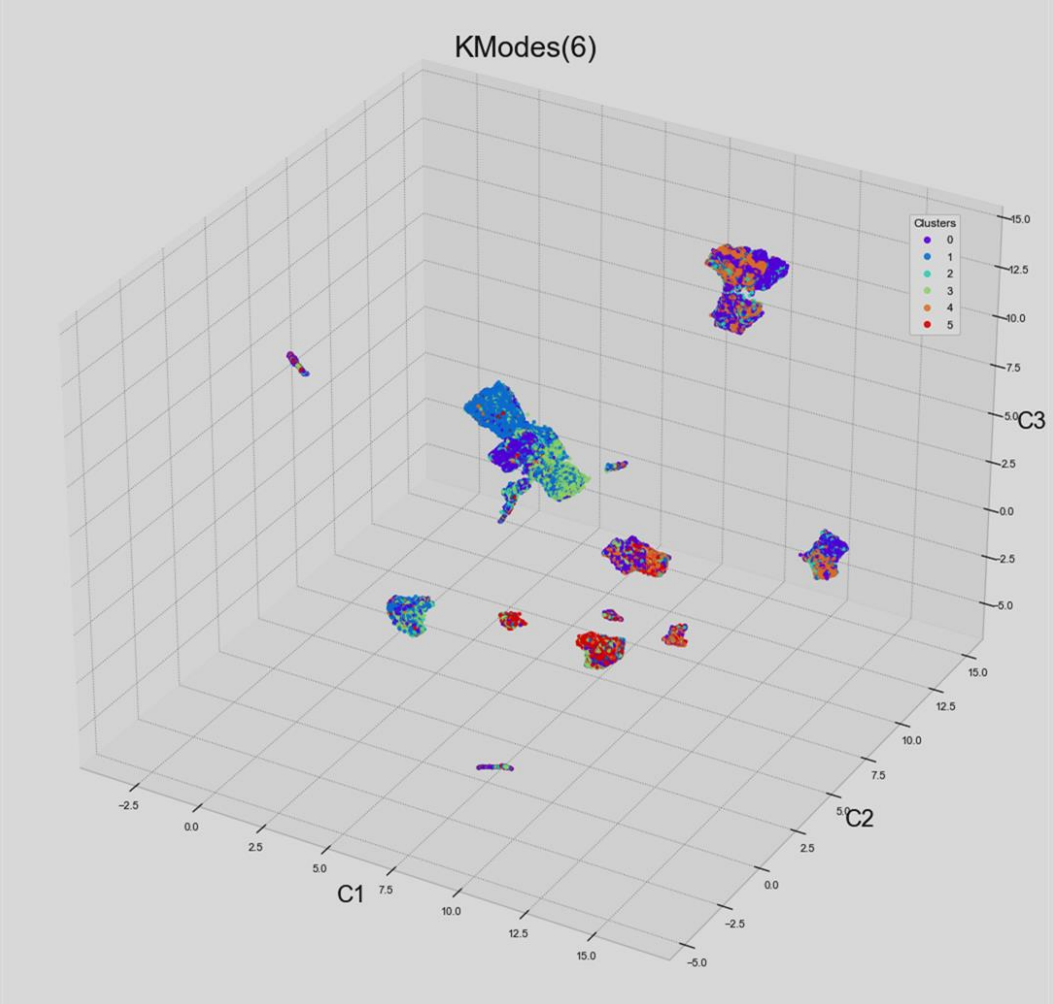


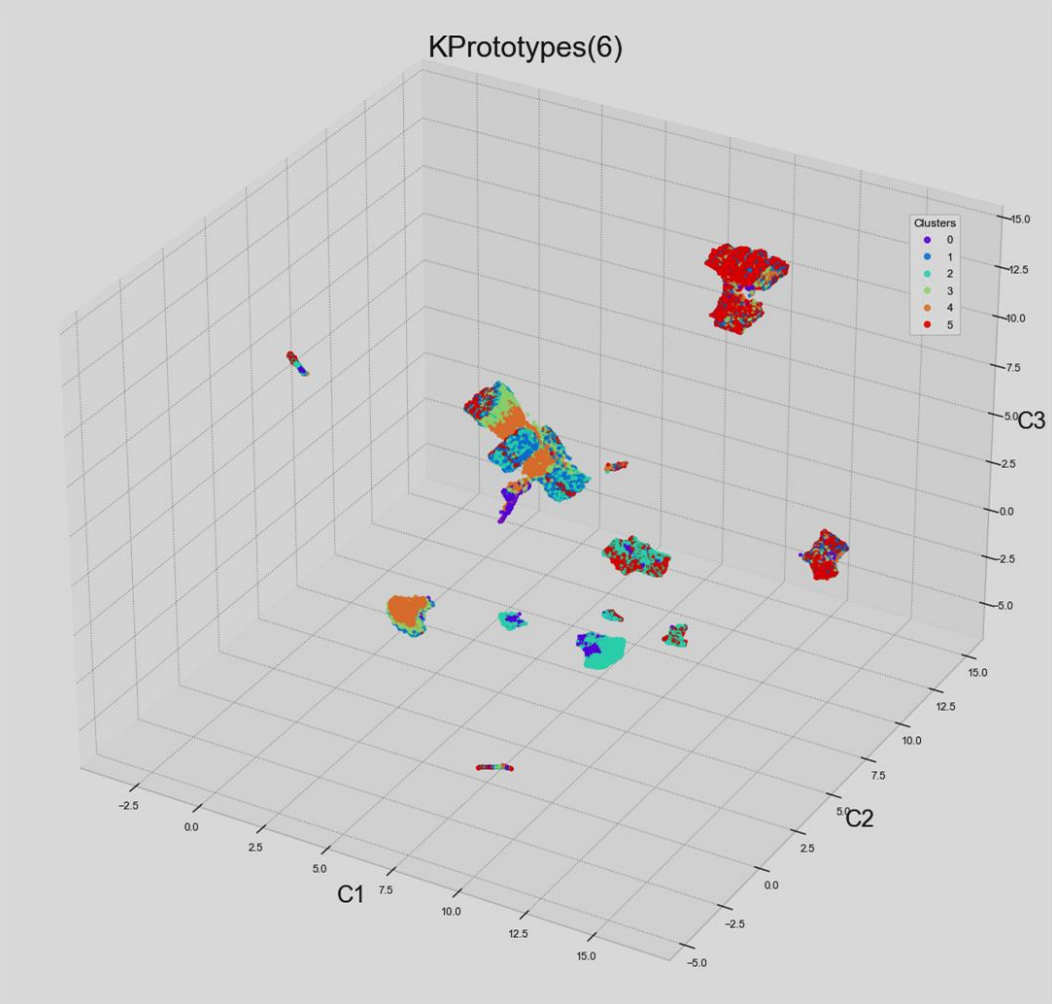




pôle emploi

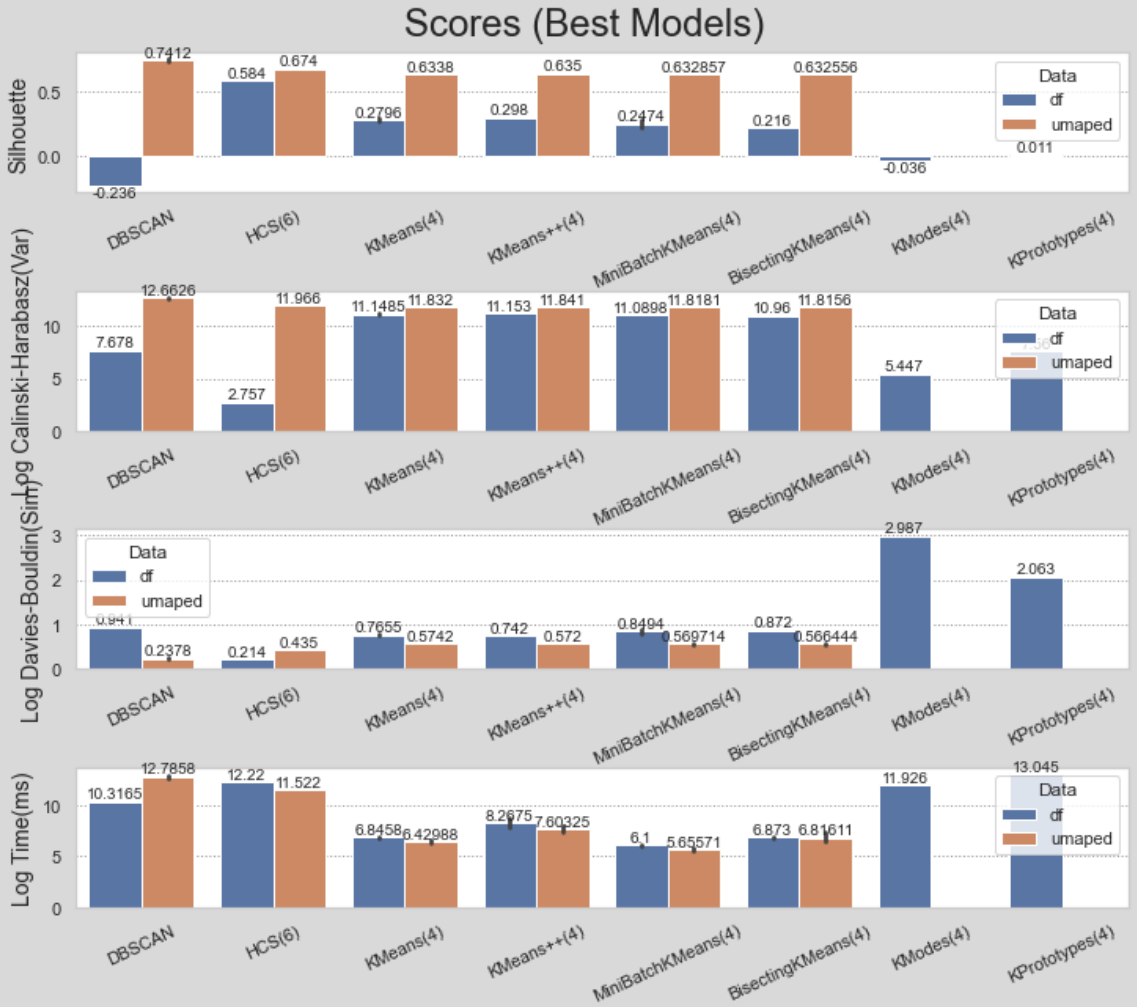
3.3.3. Clustering : **KModes & Kprototypes**





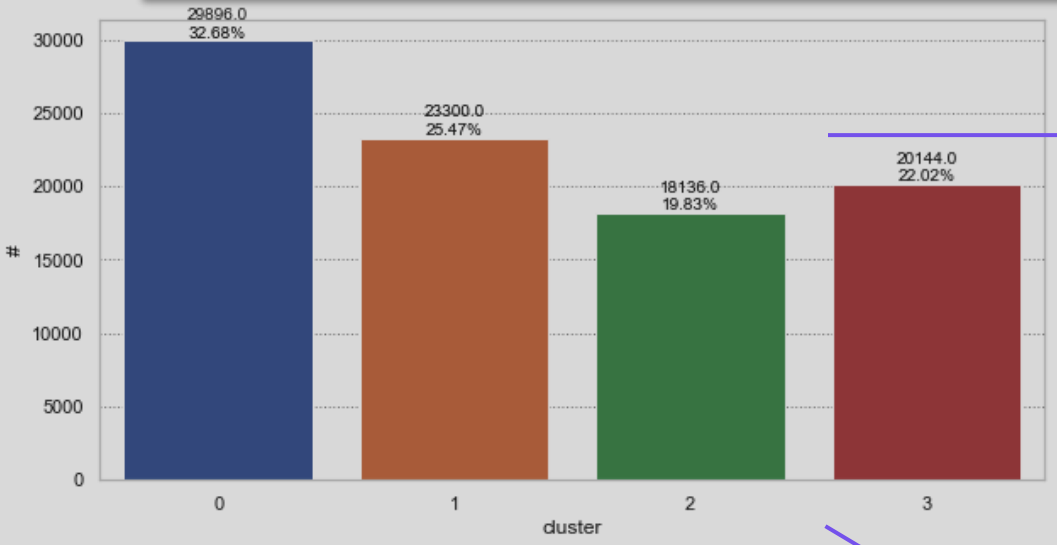
3.4. Comparaison

- **KMeans** :
  KMeans++ << BisectingKMeans++
  BisectingKMeans++ << MiniBatchKMeans++
- **Catégories** :
  Kmodes << Kprototypes
- **Réduction** : df << UMAP
- **Meilleur** : MiniBatchKMeans++(k=4)



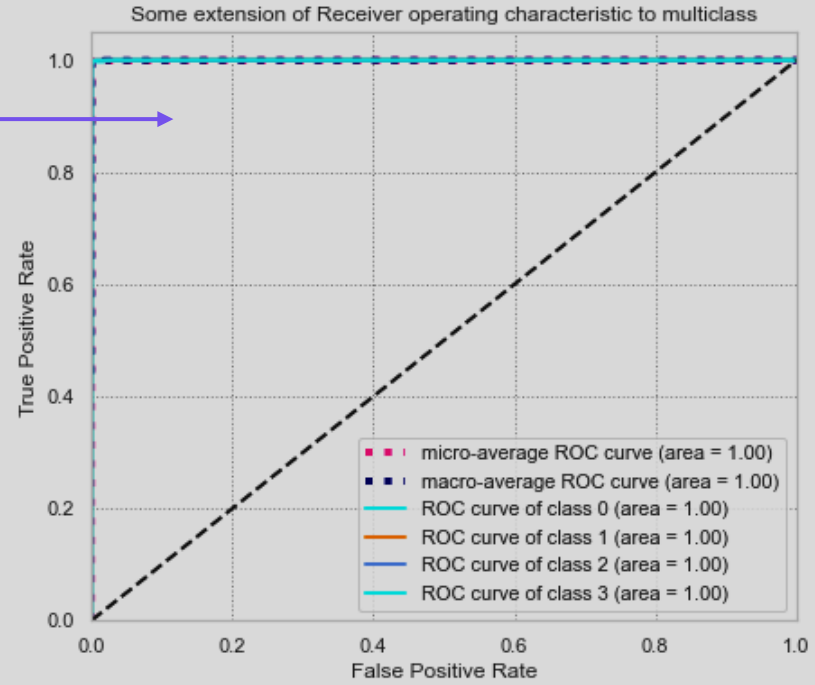
4. Evaluation & Interprétabilité

4.1.1. Evaluation: **OneVsAll & AUC**



- Distribution des Clusters -

Apprentissage du ***Cluster*** par **XGBoost** (1vsA)

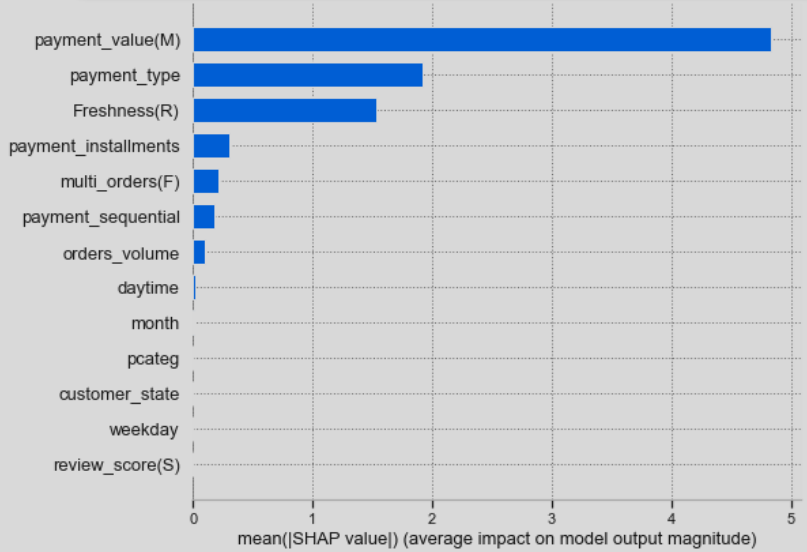


- AUC du modèle 1vsA -

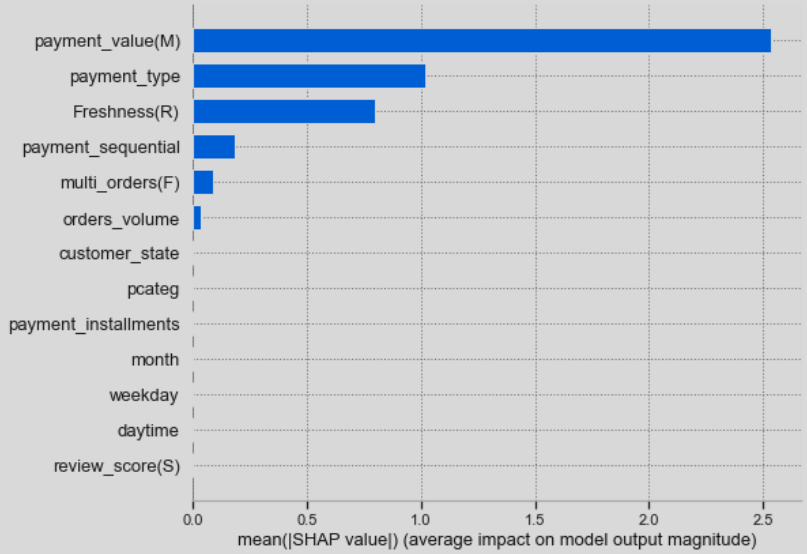
Cluster	Freshness(R)	multi_orders(F)	payment_value(M)	review_score(S)	daytime	weekday	month	orders_volume	payment_sequential	payment_installments	pcateg	customer_state	payment_type
0	0.745553	0	2.013313	4.068872	2.066363	3.212537	5.9039	1.359212	1.045792	4.257693	53.91825	3.645337	3
1	0.75419	0	1	4.197082	2.047682	3.185408	5.897811	1.051674	1	2.269657	54.061202	3.808455	3
2	0.53786	0	1.032201	4.22822	2.040031	3.337671	7.118163	1.058502	1	1.595831	54.143968	3.704124	2.190726
3	0.531888	0.134134	1.917891	4.152899	2.023531	3.301728	7.012957	1.2083	1.134234	2.861547	53.761716	3.589605	2.388701

- Centres des Clusters -

4.1.2. Evaluation: **SHAP**



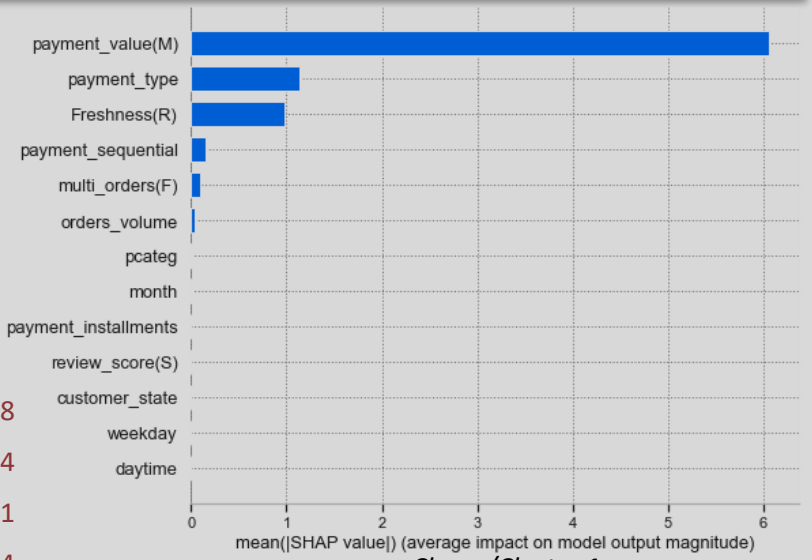
*- Classe/Cluster **0** -*



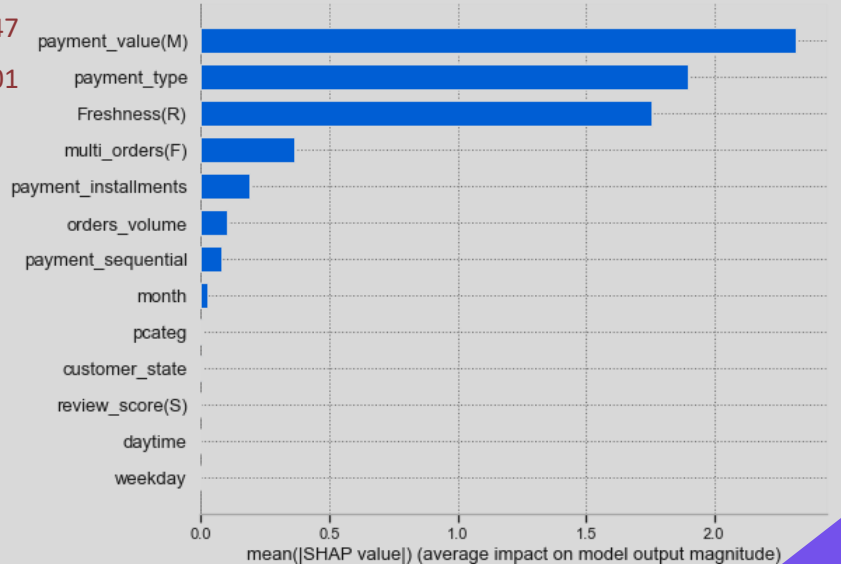
*- Classe/Cluster **2** -*

Cluster	0	1	2	3
Freshness(R)	0.745553	0.75419	0.53786	0.531888
multi_orders(F)	0	0	0	0.134134
payment_value(M)	2.013313	1	1.032201	1.917891
payment_sequential	1.045792	1	1	1.134234
payment_installments	4.257693	2.269657	1.595831	2.861547
payment_type	3	3	2.190726	2.388701

- Centres des Clusters (Attributs Importants) -

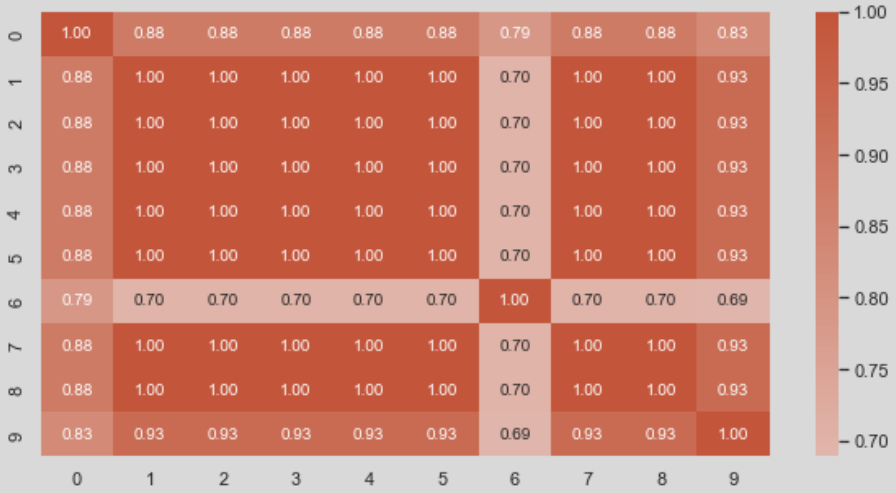


*- Classe/Cluster **1** -*

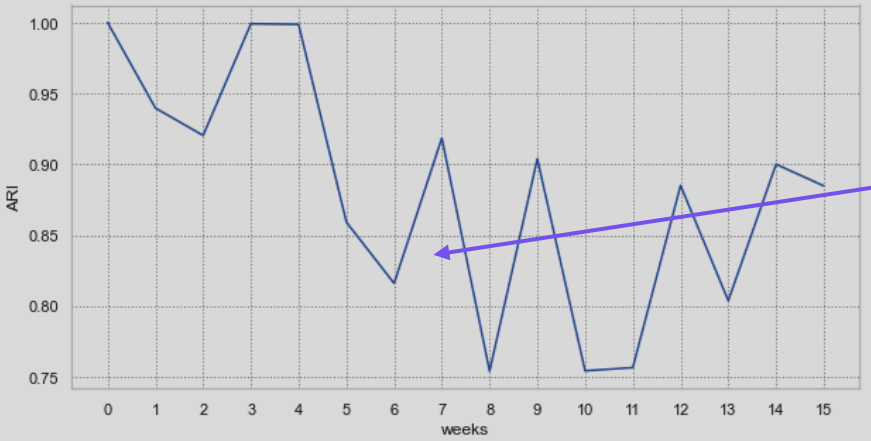


*- Classe/Cluster **3** -*

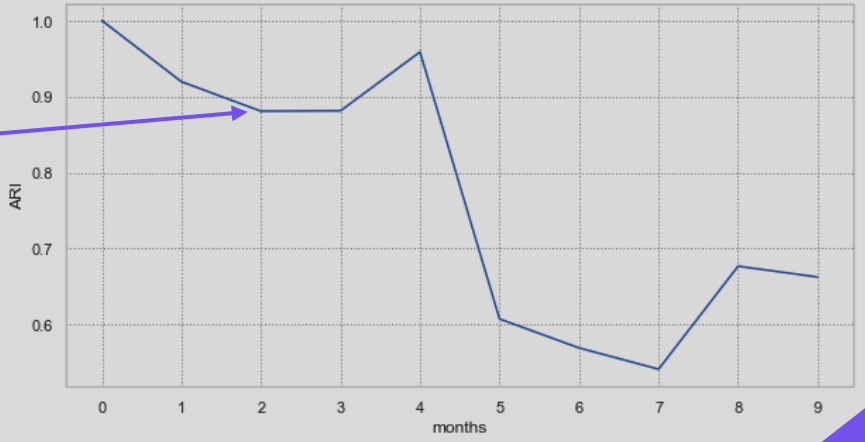
4.2. Stabilité : **Algorithmique & Temporelle**



- Matrice ARI (Seeds X Seeds) -

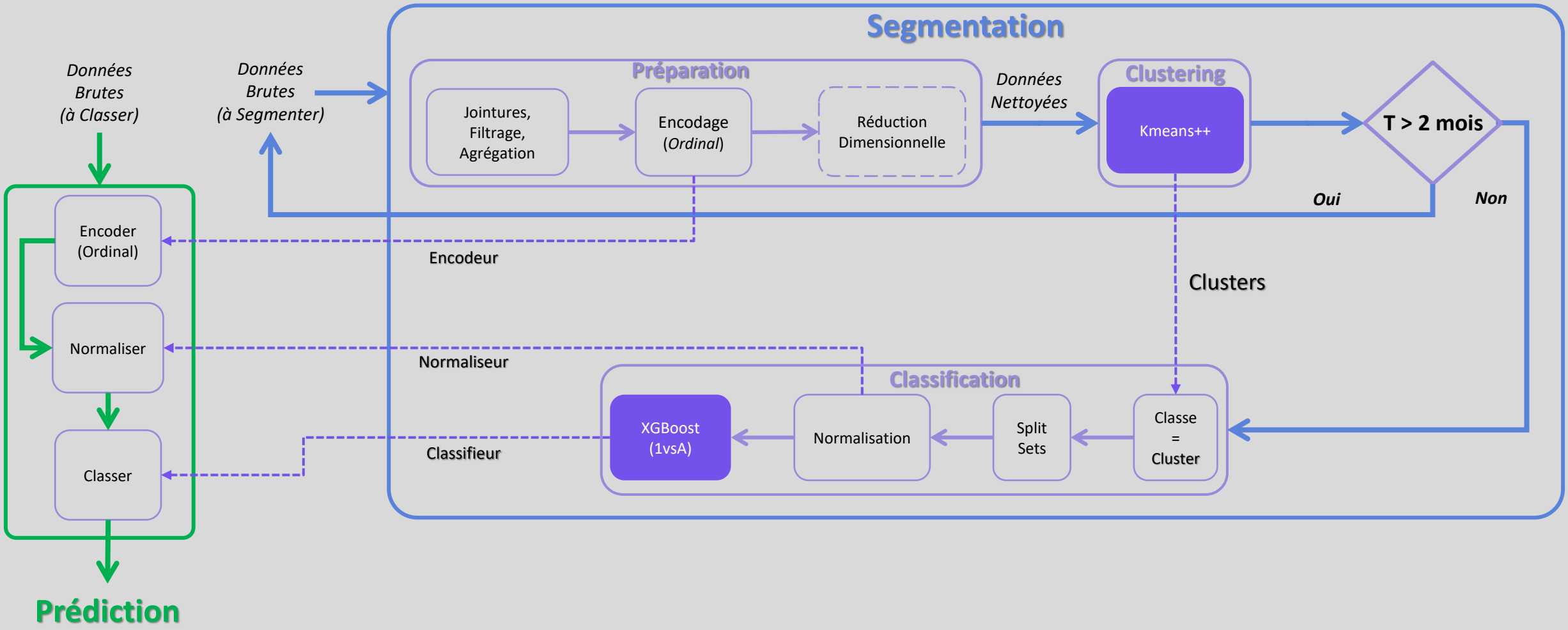


Dégradation Importante (ARI <0.9) après **2 mois**



5. Synthèse & Conclusion

5.1. Synthèse


Segmentation
Préparation
Jointures, Filtrage, Agrégation
Encodage (Ordinal)
Réduction Dimensionnelle
Données Nettoyées
Clustering
Kmeans++
T > 2 mois
Oui
Non
Données Brutes (à Segmenter)
Encodeur
Clusters
Classification
XGBoost (1vsA)
Normalisation
Split Sets
Classe = Cluster
Normaliseur
Classifieur
Données Brutes (à Classer)
Encoder (Ordinal)
Normaliser
Classer
Prédiction

5.2. Conclusion

- *Réduction dimensionnelle* & *Clustering* limités
  - *Complexité temporelle*
  - *Complexité spatiale*
  - Cause: **volume** des données.
- Informations Importantes
  - **RFM**
  - **Paiements**
- Stable *Algorithmiquement*
- Instable *Temporellement* après **2 mois** (mise à jour)