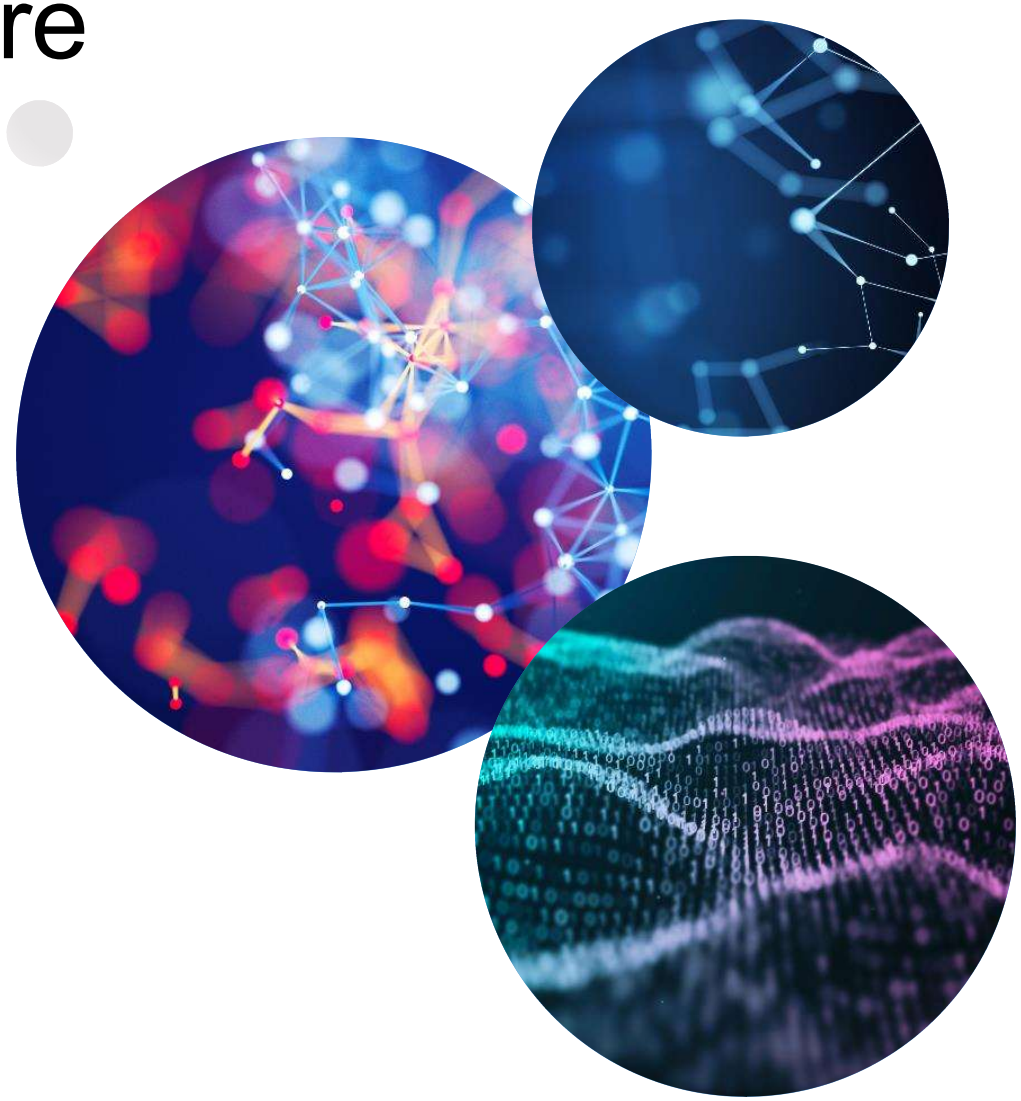
Segmentez des clients d'un site e-commerce

Armand FAUGERE LinkedIn
armand-faugere@live.fr

Sommaire



I) Cadrage du projet et données d’entrée

I) Cadrage du projet et données d'entrée

❑ **Contexte :** Projet de segmentation des Clients du site d'E-commerce Olist.

Olist a besoin d'une segmentation de ses clients à utiliser au quotidien pour ses campagnes de communication.

❑ **But :**

- Créer une segmentation Client pour l'équipe Marketing ainsi qu'une proposition de contrat de maintenance

❑ **Objectifs :**

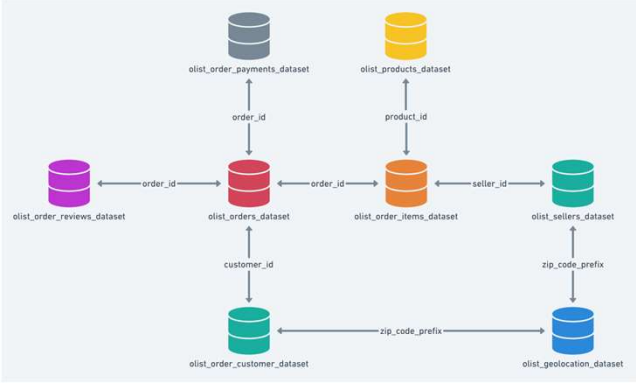
- Traiter et explorer le jeu de donnée
- Réaliser des modélisations de segmentation
- Evaluer les modèles
- Proposer la ou les segmentations pertinentes
- Proposer une maintenance adaptée

❑ **Le jeu de données**

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

❑ Base de données Client anonymisée

8 jeux de données



II) Traitement et analyse des données

2) Traitement et analyse des données

1 Préparation jeu de donnée

Jupiter Notebook, Python, Pandas, Numpy, Matplotlib, Seaborn, sklearn

Analyse rapide des 7 datasets

- ❑ Analyse describe (max, min, moy…)
- ❑ Analyse valeurs nulles
- ❑ *Analyse info sur datsets (nb de lignes, colonnes…)*

Fusion datasets et sélection Clients

- ❑ Fusion des datasets avec colonnes communes
- ❑ Sélection des Clients « delivered »

Traitement NAN

- ❑ Peu de valeurs manquantes,
  ➔ traitement par lignes
- ❑ Les autres valeurs manquantes sont sur des colnnes non utilisées « order_approved_at » , « order_delivered_carrier_date »

Création colonnes

- ❑ Création de multiples indicateurs pour analyse
  - ➔ Volume colis
  - ➔ Délai dernière commande
  - ➔ Délai approbation commande
  - ➔ Délai commande transporteur
  - ➔ Délai livraison
  - ➔ Délai commentaires client
  - ➔ Délai de réponse commentaires client
  - ➔ Retard de livraison
  - ➔ Densité colis

2) Traitement et analyse des données

2 Analyse univariée et bivariée

Jupiter Notebook, Python, *Pandas, Numpy, Matplotlib, Seaborn, sklearn*

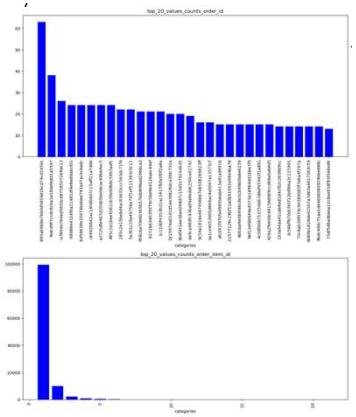
Analyse univariée variables catégorielles et numériques

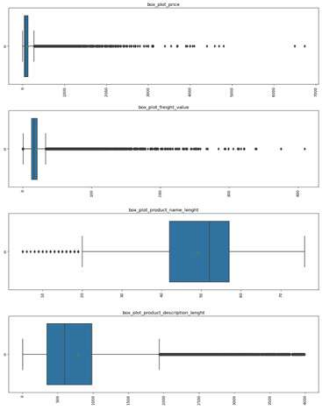
Fonctions pour réalisation des graphiques

Analyse bivariée

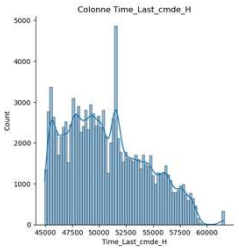
❑ Catégorielles : barplot
➔ *Top 20*
➔ *Autres catégories*

❑ Numériques : boxplot

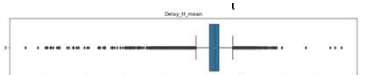


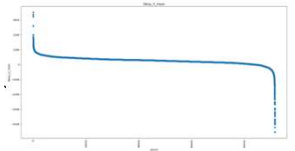


❑ Analyse barplot
➔ Top produits vendus



❑ Analyse boxplot, plot, displot
➔ *price & clients*
➔ *freight_value & clients*
➔ *product_waight_g & clients*
➔ *payment_value & clients*
➔ *volume colis & clients*
➔ *retard & clients*
➔ *time_aswer_review & clients*
➔ *Densité colis et clients*
➔ *délai depuis dernière commande et clients*





2) Traitement et analyse des données

3 Préparation du dataframe pour classification

Jupiter Notebook, Python, *Pandas, Numpy, Matplotlib, Seaborn, sklearn*

Sélection des variables et encodages des variables catégorielles → Agrégation des variables par clients → Normalisation des variables

- ❑ encodage onehotencoder :
  ➔ ['review_score', 'payment_type']
- ❑ Choix des variables

- ❑ groupement par client (groupby) :
  ➔ Sum
  ➔ Mean
  ➔ Min

- ❑ Normalisation des variables (standardscaler)

Business :
- Nb_cmde
- payment_value_sum
- Time_last_cmde_H

Type de colis :
- product_weight_g_mean
- Volume_Colis_cm3_mean

Type de paiement :
- payment_type_boleto_sum
- payment_type_credit_card_sum
- payment_type_debit_card_sum
- payment_type_voucher_sum

Satisfaction Client :
- review_score_1_sum
- review_score_2_sum
- review_score_3_sum
- review_score_4_sum
- review_score_5_sum
- Time_Answer_Review_H
- Delay_H_mean

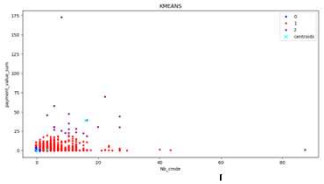
III) Modélisations et Analyses

3) Modélisations

1
Modélisations KMEANS

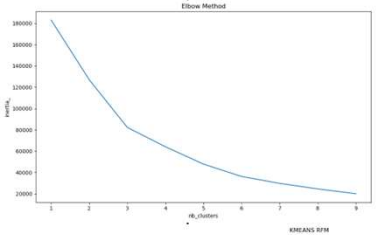
Jupiter Notebook, Python, Pandas, Numpy, Matplotlib, Seaborn, Sklearn

Sans Analyse de composantes principales

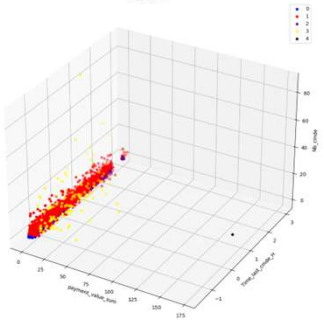


- ❑ *Elbow et coefficient de silhouette pour choix du nombre de clusters*

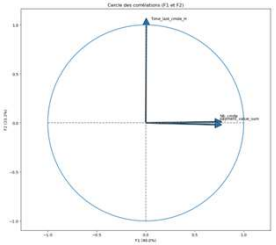
- ❑ *Avec 3 variables*
➔ *[Nb_cmde', 'payment_value_sum', 'Time_last_cmde_H']*



- ❑ *Avec 2 variables*
➔ *['Nb_cmde', 'Time_last_cmde_H']*
➔ *['Nb_cmde', 'payment_value_sum']*
➔ *['payment_value_sum','Time_last_cmde_H']*
➔ *['Nb_cmde', 'review_score_1_sum']*
➔ *['Nb_cmde', 'Time_Answer_Review_H']*
➔ *['Nb_cmde', 'Delay_H_mean']*
➔ *['Nb_cmde', 'Volume_Colis_cm3_mean']*



Avec Analyse de composantes principales



- ❑ *Décomposition PCA*
- ❑ *Cercle des corrélations*
- ❑ *Création de nouveaux indicateurs*
- ❑ *Elbow et coefficient de silhouette pour choix du nombre de clusters*
- ❑ *Nouveaux indicateurs*

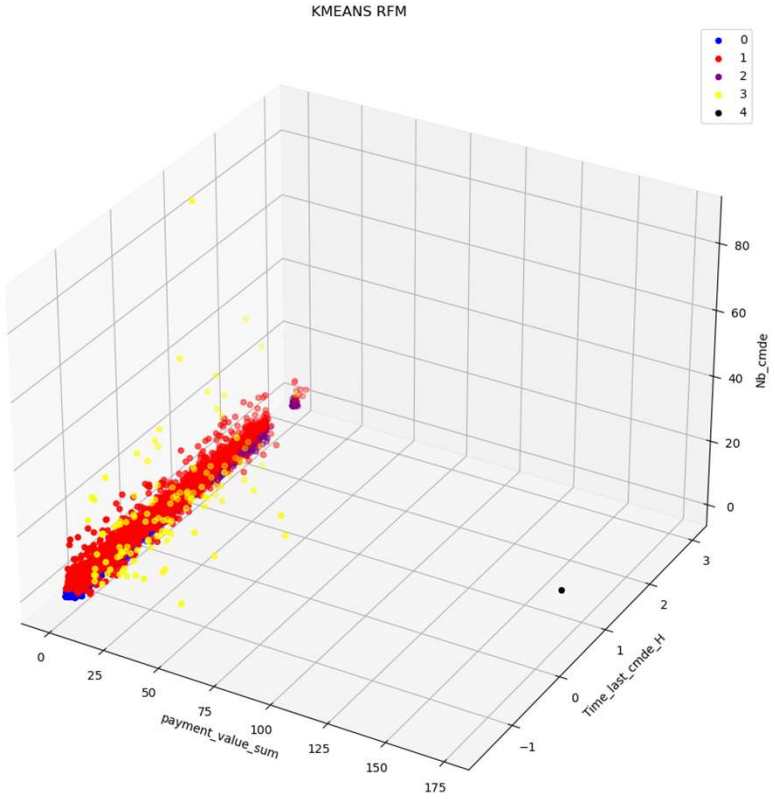
➔ *Economique*
➔ *Satisfaction*
➔ *Performance*

3) Modélisations

[Nb_cmde', 'payment_value_sum', 'Time_last_cmde_H']

coeff_silhouette ➔ 0.482517

Difficile à interpréter



Sans PCA

['Nb_cmde', 'Time_last_cmde_H']

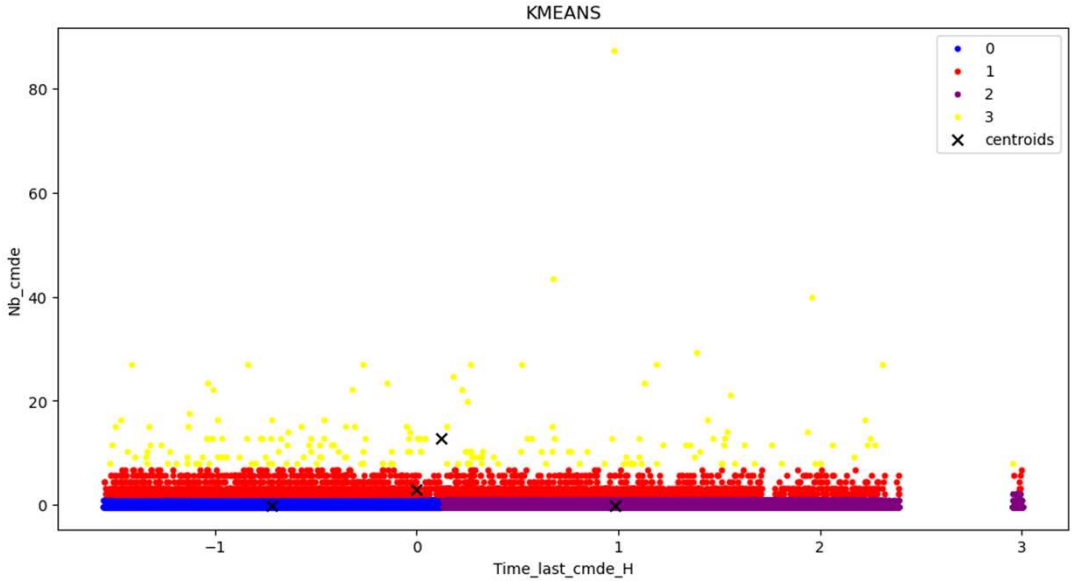
coeff_silhouette ➔ 0.523183

Cluster violet ➔ Clients avec peu de cmde et ayant cmdé il y a très longtemps (Clients perdus)

Cluster rouge ➔ Clients qui ont réalisés plusieurs cmdes dont certaines très récemment (à réactiver pour les plus anciens certains perdus)

Cluster Jaune ➔ Très bon Clients mais certains sont perdus (à réactiver pour les plus anciens)

Cluster Bleu ➔ Clients récents avec peu de cmdes



3) Modélisations

['Nb_cmde', 'payment_value_sum']

coeff_silhouette ➔ 0.853683

Cluster violet ➔ Très bon Clients (Champions)

Cluster rouge ➔ Bon Clients

Cluster Bleu ➔ Clients récents avec peu de cmde

Sans PCA

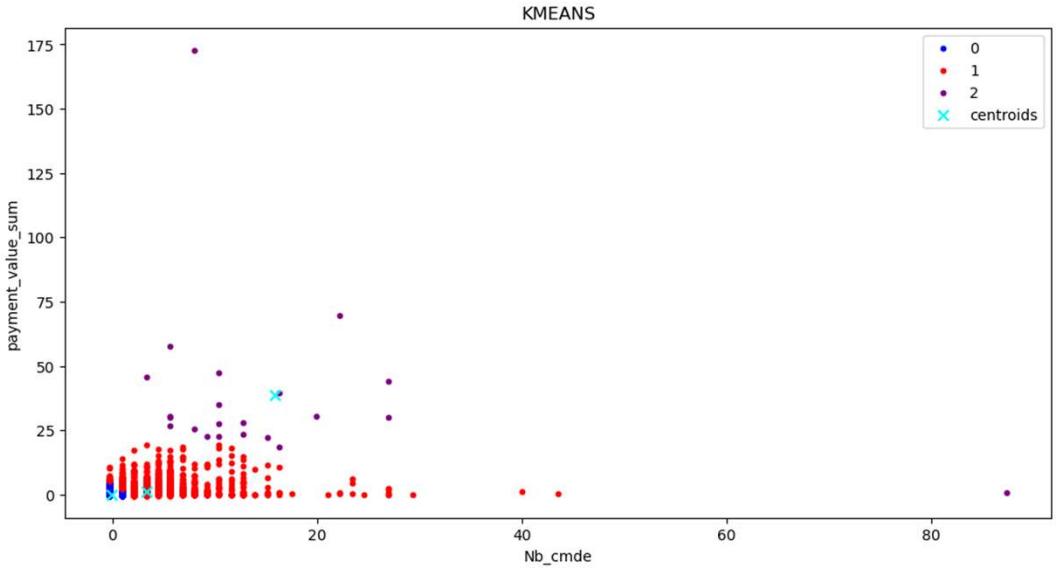
['payment_value_sum', 'Time_last_cmde_H']

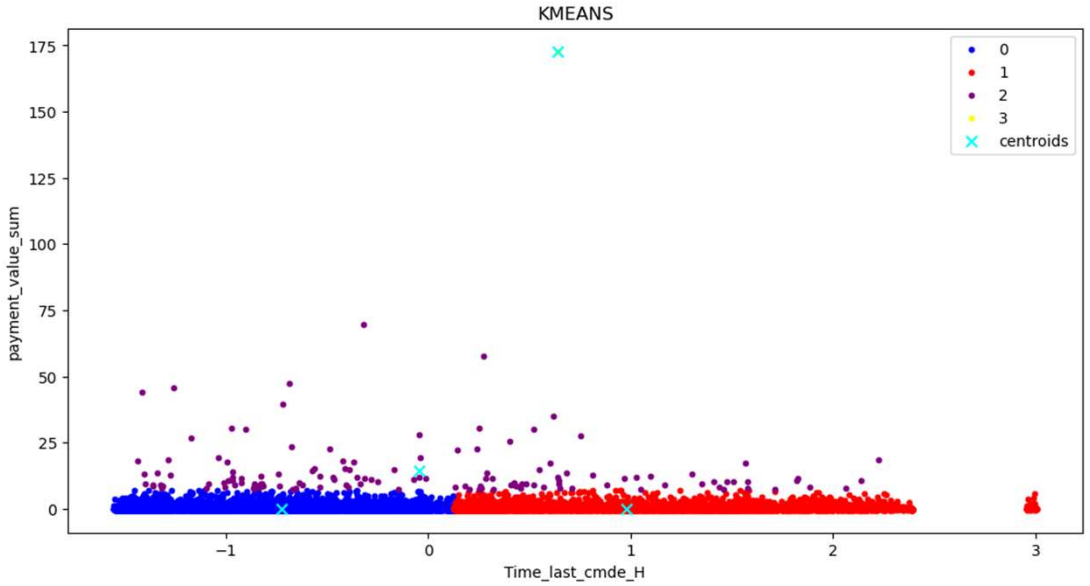
coeff_silhouette ➔ 0.526722

Cluster violet ➔ Très bon Clients (Champions) à réactiver pour les perdus

Cluster rouge ➔ Bon Clients et clients modérés mais perdus

Cluster Bleu ➔ Bon Clients et clients modérés mais récents





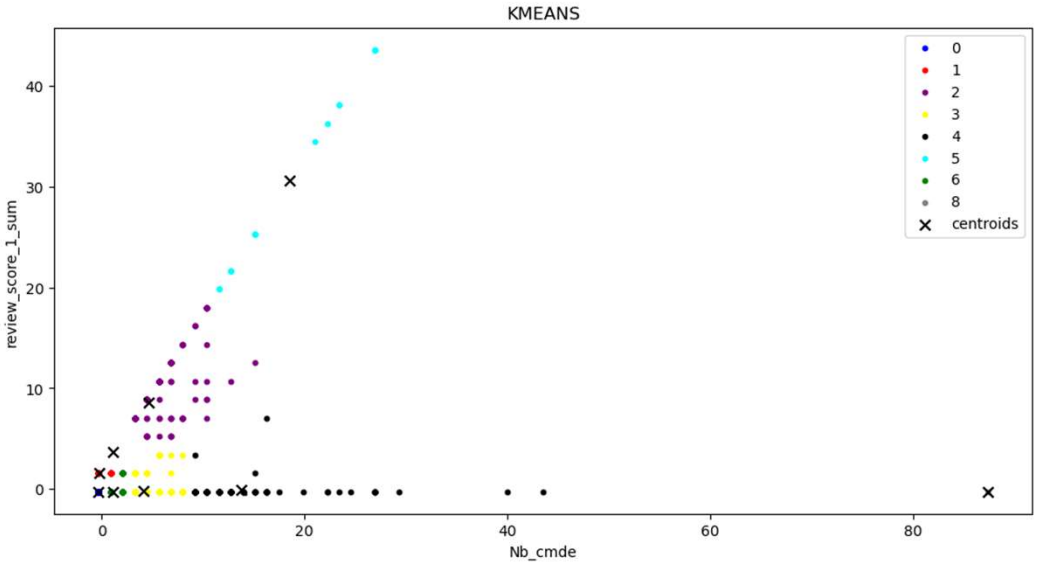
3) Modélisations

['Nb_cmde', 'review_score_1_sum']

coeff_silhouette ➔ 0.9611029465099215

Couleur Cyan ➔ groupe des bons clients qui ont donné de mauvaises notes

Sans PCA



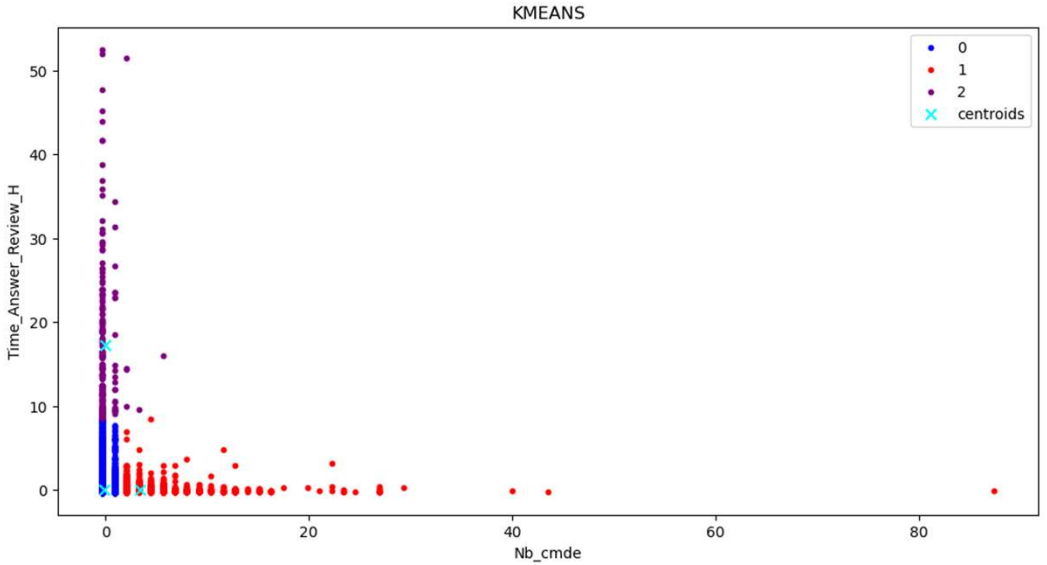
['Nb_cmde', 'Time_Answer_Review_H']

coeff_silhouette ➔ 0.843393

couleur rouge ➔ Clients importants avec réponse rapide

couleur bleu ➔ Clients moins importants avec réponse rapide

couleur violette ➔ Clients moins importants avec réponse lente

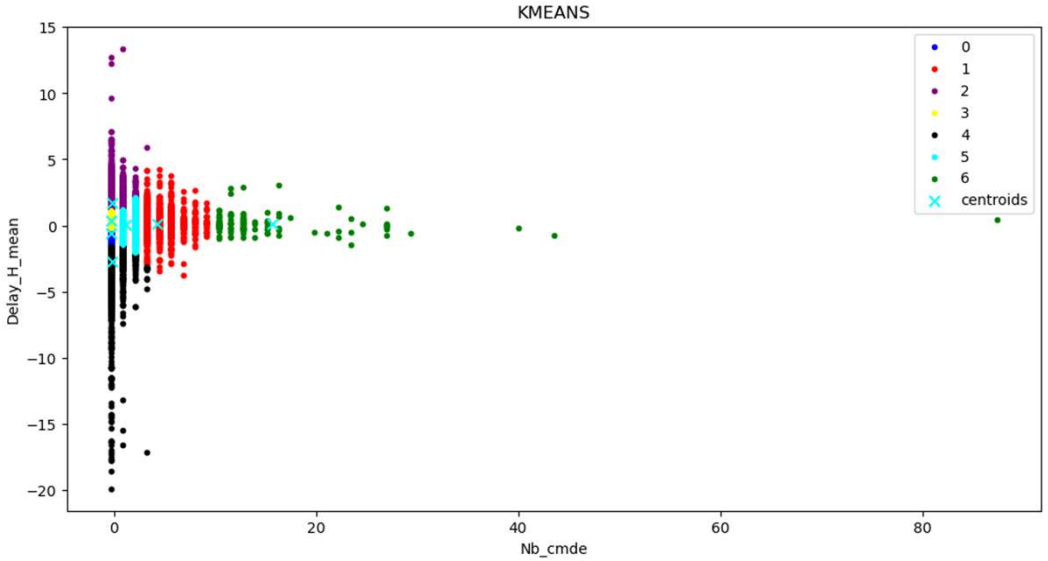


3) Modélisations

Sans PCA

['Nb_cmde', 'Delay_H_mean']

coeff_silhouette ➔ 0.505418



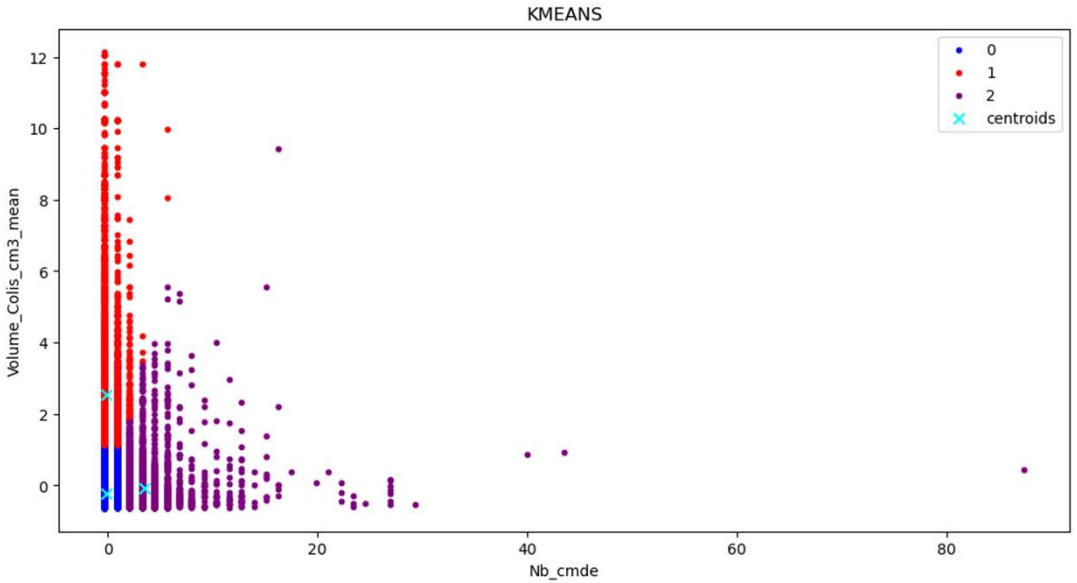
['Nb_cmde', 'Volume_Colis_cm3_mean']

coeff_silhouette ➔ 0.717313

couleur rouge ➔ gros colis petit client

couleur bleu ➔ petit colis petit client

couleur violette ➔ colis moyen client moyen à bon



3) Modélisations

MODELISATION PCA ➔ ECONOMIQUE

coeff_silhouette ➔ 0.509136

Cluster violet ➔ Clients Moyens perdus

Cluster rouge ➔ Très Bon Clients mais certains sont perdus

Cluster Jaune ➔ Bon clients mais certains sont perdus

Cluster Bleu ➔ Clients fidèles mais clients moyens

Avec PCA

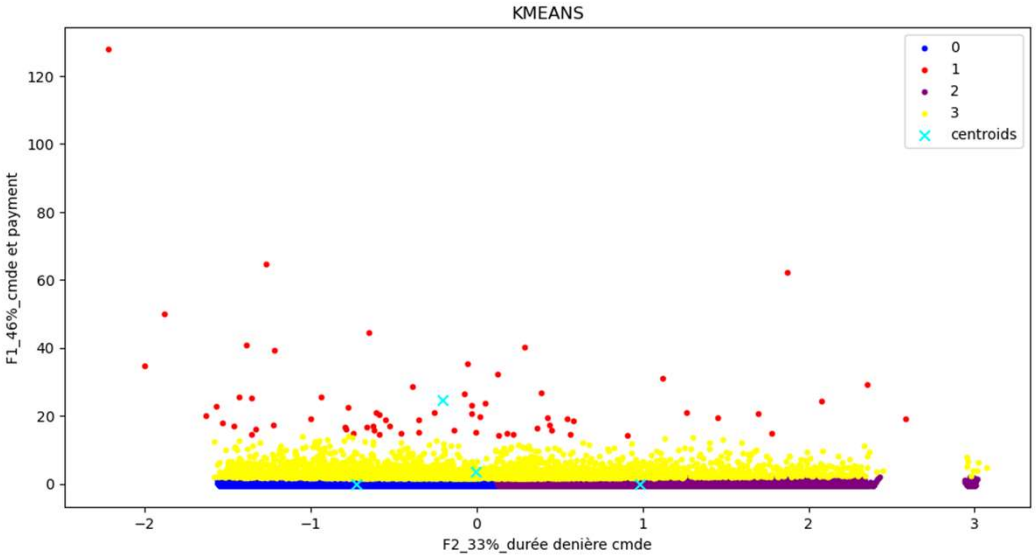
MODELISATION PCA ==> SATISFACTION

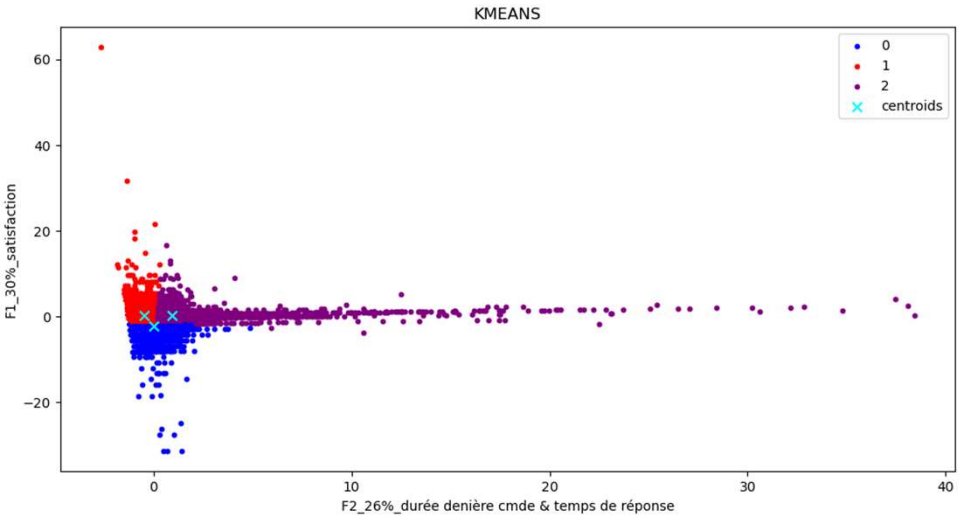
coeff_silhouette ➔ 0.423568

Cluster violet ➔ Clients moyennement satisfaits et perdus ou à risque

Cluster rouge ➔ Clients moyennement satisfaits mais certains sont à risque

Cluster Bleu ➔ Clients pas satisfaits et à risque





3) Modélisations

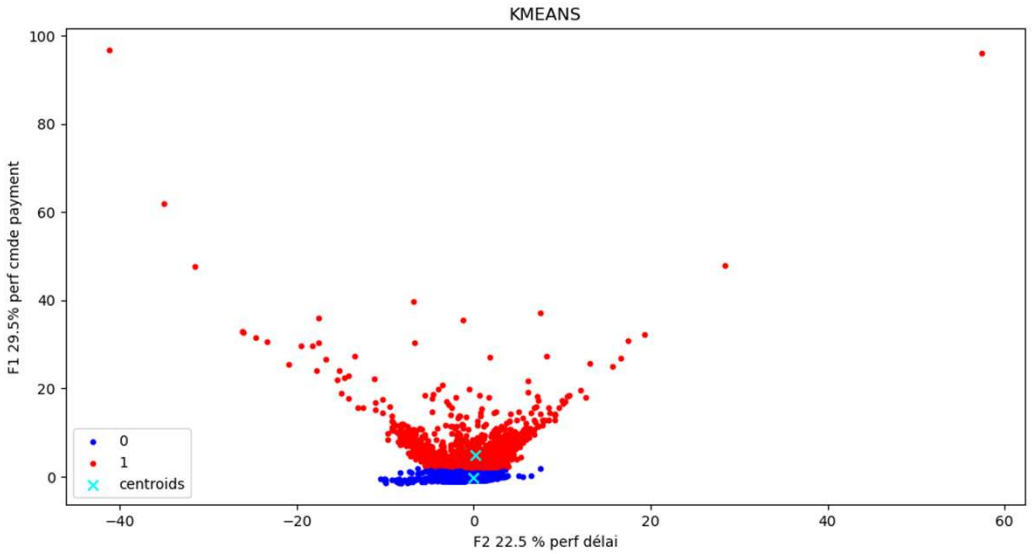
MODELISATION PCA ➔ PERFFORMANCE

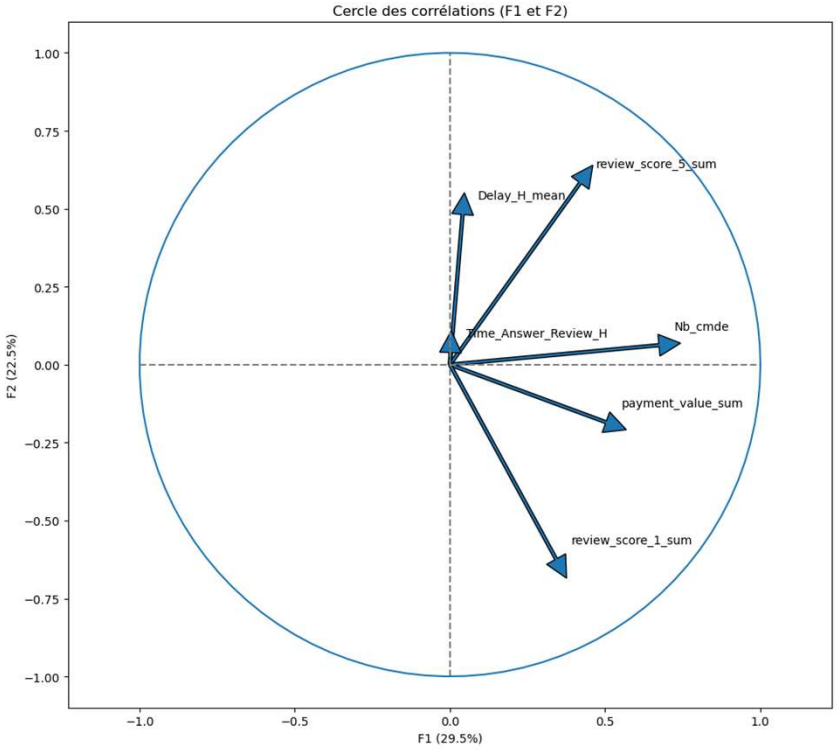
coeff_silhouette ➔ 0.775953

- Le retard n'empêche pas de mettre des bonnes notes
- Les mauvaises notes n'empêchent pas de de réaliser des cmde

➔La qualité des produits et le prix doivent être davantage déterminants pour les clients

Avec PCA





3) Modélisations

2
Autres Modélisations

DBSCAN

CAH

Jupiter Notebook, Python, Pandas, Numpy, Matplotlib, Seaborn, Sklearn

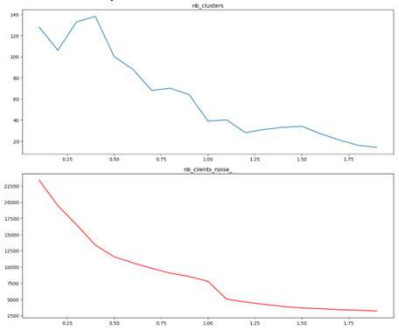
DBSCAN

Fonctions pour DBSCAN avec gestion des paramètres eps et samples_min

- ❑ *Mise en place DBSCAN (pca_économique)*
- ➔ *Pas de convergence qd on prend la totalité des clients*

- ❑ *Mise en place DBSCAN ['Nb_cmde', 'payment_value_sum']*
- ➔ *eps 0,1➔2; samples_min : 10*
- ➔ *eps 0,1➔2; samples_min : 20*
- ➔ *eps 0,1➔2; samples_min : 30*

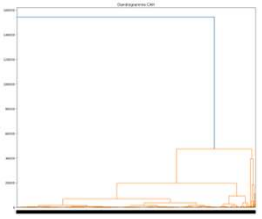
- ➔ *Analyse du nb de cluster*
- ➔ *Analyse du bruit*



CAH

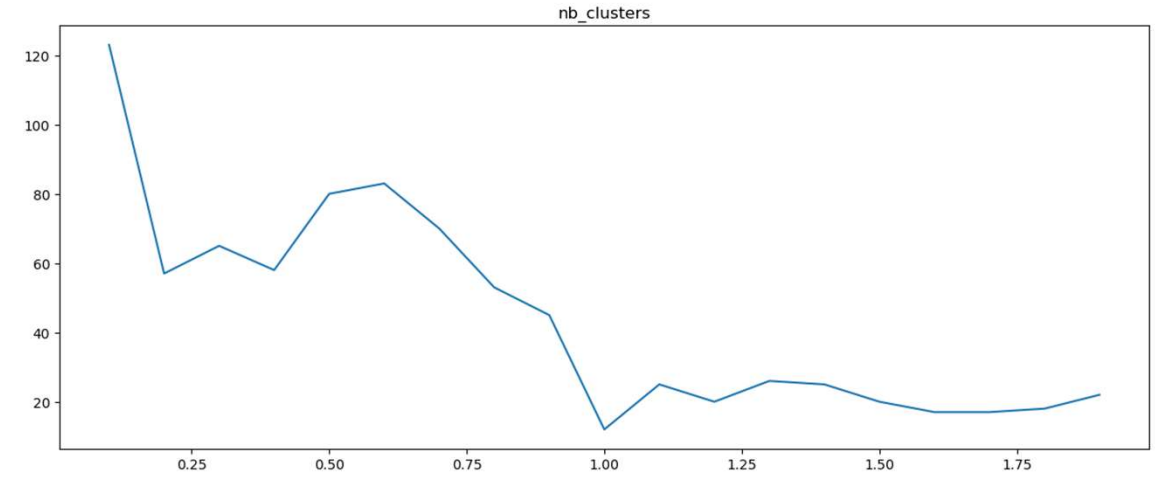
- ❑ *Mise en place de cluster hiérarchique*
- ➔ *Pas de convergence car la base de données est trop grande*

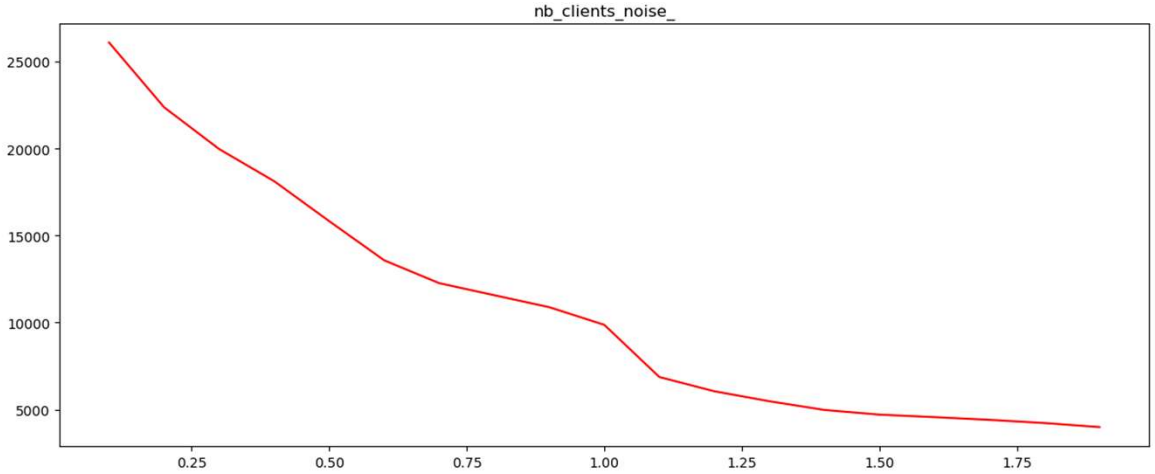
- ❑ *Cluster hiérarchique sur 10000 Clients*
- ➔ *Un résultat mais adapté à la problématique*



Pas adapté

3) Modélisations





- **DBSCAN**
- ➔ eps : 0,1➔ 2
- ➔ samples_min : 30

Algorihtme pas adapté à la problématique

- Beaucoup de clusters ➔ Interprétabilité difficile, représentativité pas pertinente.
- Le bruit exclut des milliers de Clients.

IV) Modèles retenus et analyses

3) Modèles retenus et analyse

Sélection des modèles et analyse

Jupiter Notebook, Python, Pandas, Numpy, Matplotlib, Seaborn, Sklearn

Fonctions pour graphiques

Sélection des modèles pertinents pour bâtir un tableau de bord Client

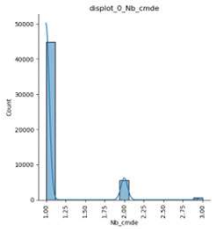
- ❑ *PCA économique*
- ❑ *PCA satisfaction*
- ❑ *Nb_cmde & payment_sum*
- ❑ *Nb_cmde & answer_review_time*
- ❑ *Nb_cmde & volume_colis*

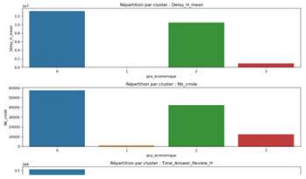
Analyse détaillée

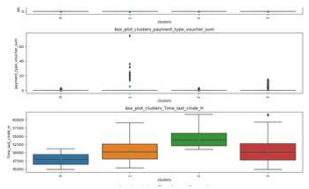
- ❑ *Nombre de clients par clusters*
- ❑ *Agrégation par cluster*
- ❑ *Boxplot des variables par clusters*
- ❑ *Displot des variables*
- ❑ *Barplot répartition variables par cluster*

Validation des modèles

- ❑ ***PCA économique***
- ❑ ***PCA satisfaction***
- ❑ ***Nb_cmde & payment_sum***
- ❑ ***Nb_cmde & answer_review_time***
- ❑ ***Nb_cmde & volume_colis***

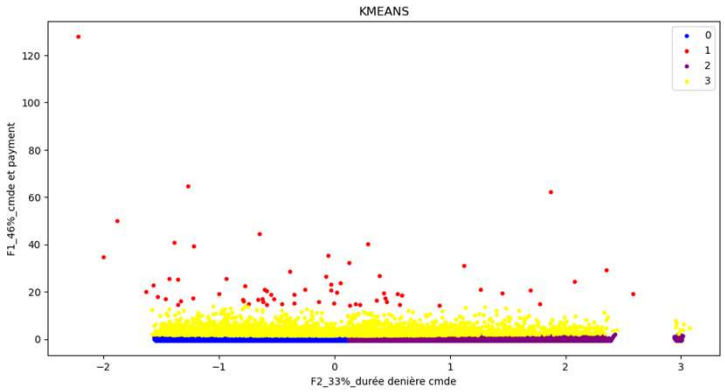




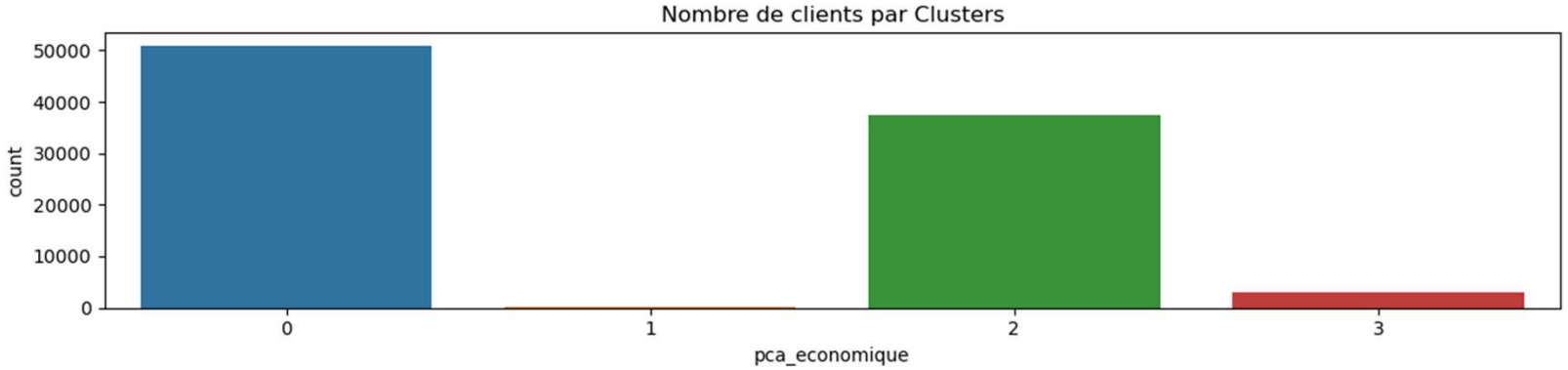


4) Modèles retenus et analyse

❑ Focus PCA économique

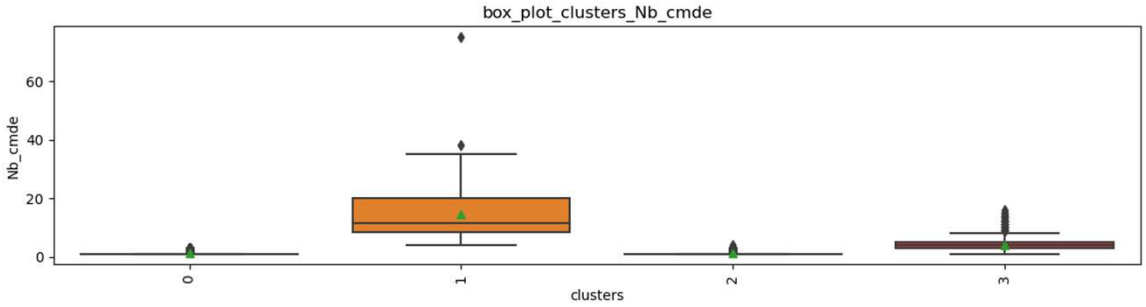


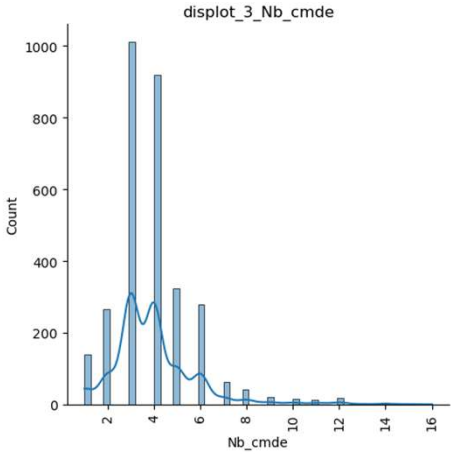
AG MEAN	Nb_cmde	payment_value_sum	review_score_1_sum	review_score_2_sum	review_score_3_sum	review_score_4_sum	review_score_5_sum	payment_type_boleto_sum	payment_type_credit_card_sum	payment_type_debit_card_sum	payment_type_voucher_sum	Time_last_cmde_H	Time_Answer_Review_H	Delay_H_mean	product_weight_g_mean	Volume_Colis_cm3_mean
0	1.129811	164.694358	0.118600	0.035049	0.092205	0.215821	0.668135	0.211062	0.860040	0.022166	0.036544	47960.605569	70.053079	258.386111	1975.806786	13896.908501
1	14.606061	12308.768182	4.696970	0.772727	1.393939	2.469697	5.272727	3.136364	5.651515	0.000000	5.818182	50445.948788	78.130455	278.690606	4376.185303	28086.768939
2	1.134054	164.981740	0.105502	0.036378	0.098584	0.228579	0.665011	0.234028	0.845540	0.010283	0.044204	54196.874032	83.819631	280.570217	2185.110319	16376.039101
3	3.949728	1299.272149	0.823247	0.206532	0.340058	0.679475	1.900416	0.745757	2.488313	0.035223	0.680435	50574.178719	71.895011	295.147928	3190.169635	21897.902696

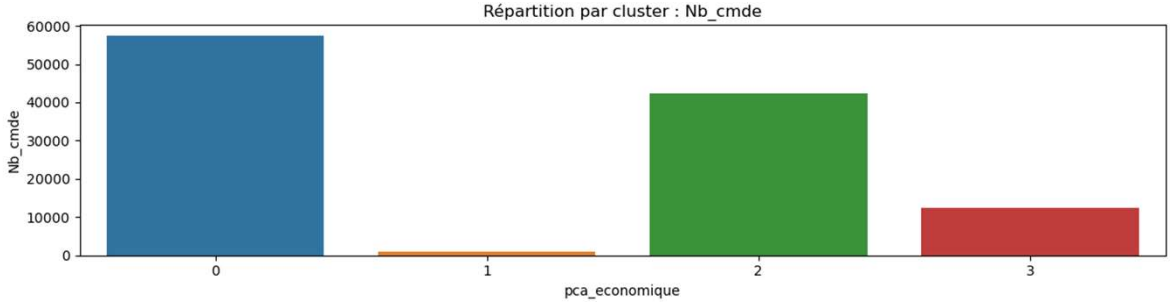


4) Modèles retenus et analyse

- Focus PCA économique sur variable Nb_cmde







V) Maintenance des modèles

5) Maintenance des modèles

Simulations Modèle 0

Jupiter Notebook, Python, Pandas, Numpy, Matplotlib, Seaborn, Sklearn

Fonctions pour simulations

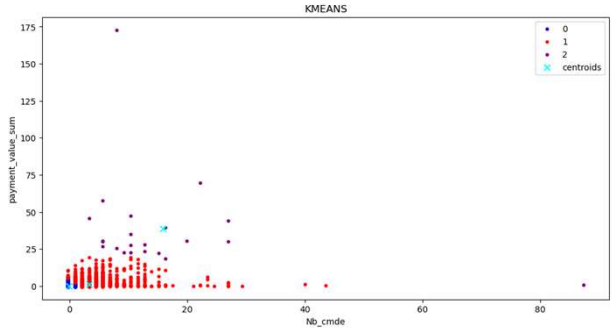
Etablissement modèle M0 à fin 2017 ➔ T0

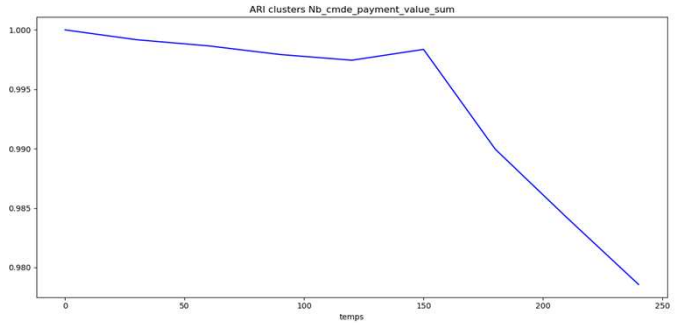
Modélisations tous les 30 jours ➔ Aout 2018

Stratégie de maintenance

- *Comparaison prédictions M0 aux modèles intermédiaires ➔* ***ARI***

5) Maintenance des modèles





['Nb_cmde', 'payment_value_sum']

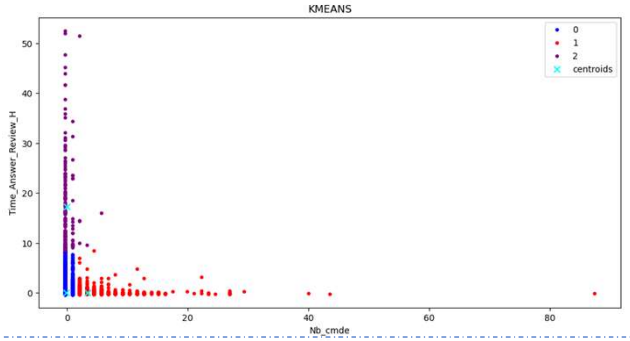
coeff_silhouette ➔ 0.853683

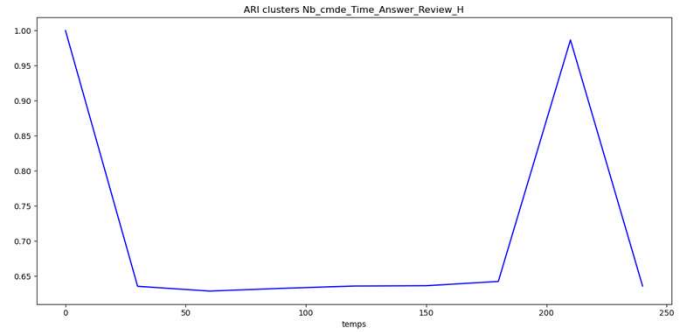
Cluster violet ➔ Très bon Clients (Champions)

Cluster rouge ➔ Bon Clients

Cluster Bleu ➔ Clients récents avec peu de cmde

ARI : Le modèle a une bonne durabilité





['Nb_cmde', 'Time_Answer_Review_H']

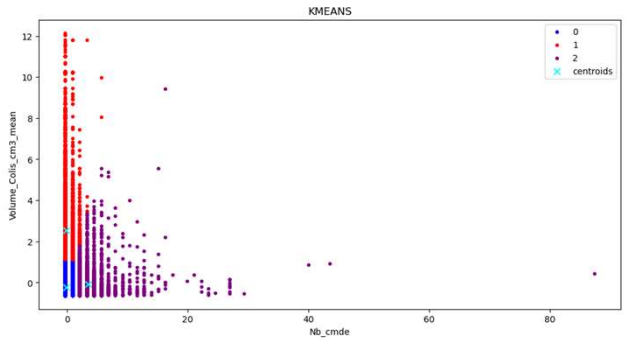
coeff_silhouette ➔ 0.843393

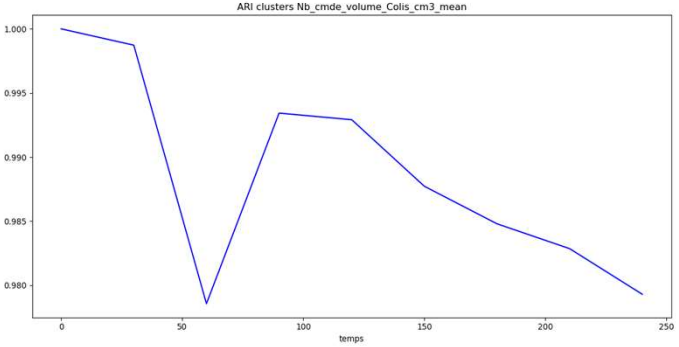
couleur rouge ➔ Clients importants avec réponse rapide

couleur bleu ➔ Clients moins importants avec réponse rapide

couleur violette ➔ Clients moins importants avec réponse lente

ARI : Le modèle devient rapidement obsolète (1 mois)





['Nb_cmde', 'Volume_Colis_cm3_mean']

coeff_silhouette ➔ 0.717313

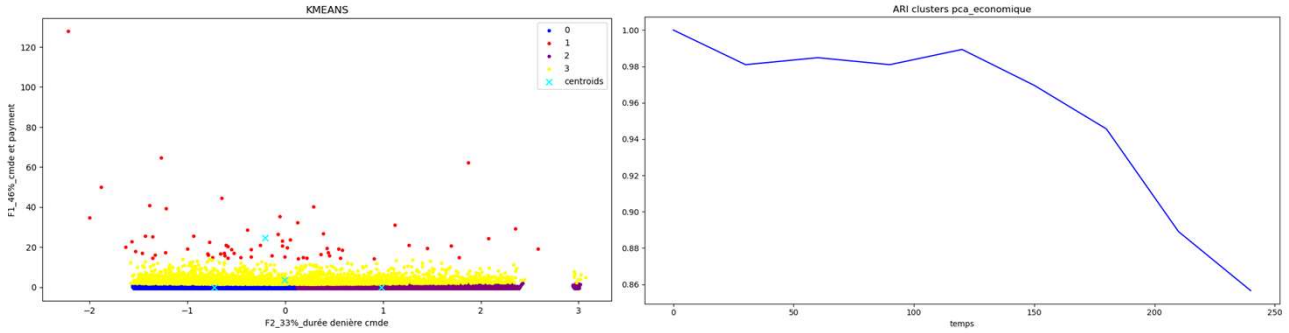
couleur rouge ➔ gros colis petit client

couleur bleu ➔ petit colis petit client

couleur violette ➔ colis moyen client moyen à bon

ARI : Le modèle a une bonne durabilité

5) Maintenance des modèles



MODELISATION PCA ➔ ECONOMIQUE

coeff_silhouette ➔ 0.509136

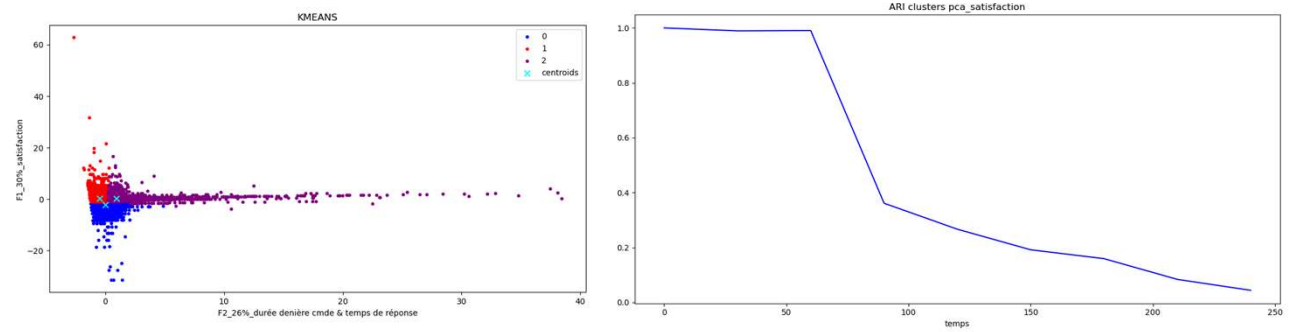
Cluster violet ➔ Clients Moyens perdus

Cluster rouge ➔ Très Bon Clients mais certains sont perdus

Cluster Jaune ➔ Bon clients mais certains sont perdus

Cluster Bleu ➔ Clients fidèles mais clients moyens

ARI : Le modèle tend à être obsolète à partir de 8 mois



MODELISATION PCA ➔ SATISFACTION

coeff_silhouette ➔ 0.423568

Cluster violet ➔ Clients moyennement satisfaits et perdus ou à risque

Cluster rouge ➔ Clients moyennement satisfaits mais certains sont à risque

Cluster Bleu ➔ Clients pas satisfaits et à risque

ARI : Le modèle devient obsolète à partir de 3 mois

5) Maintenance des modèles

Contrat de maintenance

I) Forfait de maintenance premium : mensuel

+ Forte réactivité :
➔ Corriger rapidement des dysfonctionnements en limitant le risque de perdre des clients
➔ Avoir des actions commerciales rapides et pertinentes

II) Forfait de maintenance medium : trimestriel ➔ recommandé

+ Forfait économique :
➔ Corriger des dysfonctionnements
➔ Avoir des opérations commerciales pertinentes, moins réactives mais aussi moins chères

III) Forfait de maintenance minimum : Semestriel

+ Forfait suivi :
➔ Tirer des conclusions et de réajuster la stratégie

VI) Conclusion

6) Conclusion

❑ **But :**

- **Créer une segmentation Client pour l'équipe Marketing ainsi qu'une proposition de contrat de maintenance**

❑ **Objectifs :**

- **Traiter et explorer le jeu de donnée** 
- **Réaliser des modélisations de segmentation** 
- **Evaluer les modèles** 
- **Proposer la ou les segmentations pertinentes** 
- **Proposer une maintenance adaptée** 

➔ **Tableau de bord avec :**

❑ ***PCA Economique***

❑ ***PCA Satisfaction***

❑ ***Nb_cmde & payment_sum***

❑ ***Nb_cmde & answer_review_time***

❑ ***Nb_cmde & volume_colis***

➔ ***Permet d'avoir une approche commerciale ciblée auprès des Clients***

Merci

- Armand FAUGERE
- armand-faugere@live.fr