Segmentez des clients d'un site e-commerce

Formation Data Scientist – Projet 5

Octave POUILLOT Juin 2023

Sommaire

- Mission
- Présentation du jeu de données
- Nettoyage & Exploration
- Modélisation
- Maintenance
- Conclusions

Présentation de la mission

Présentation de la mission

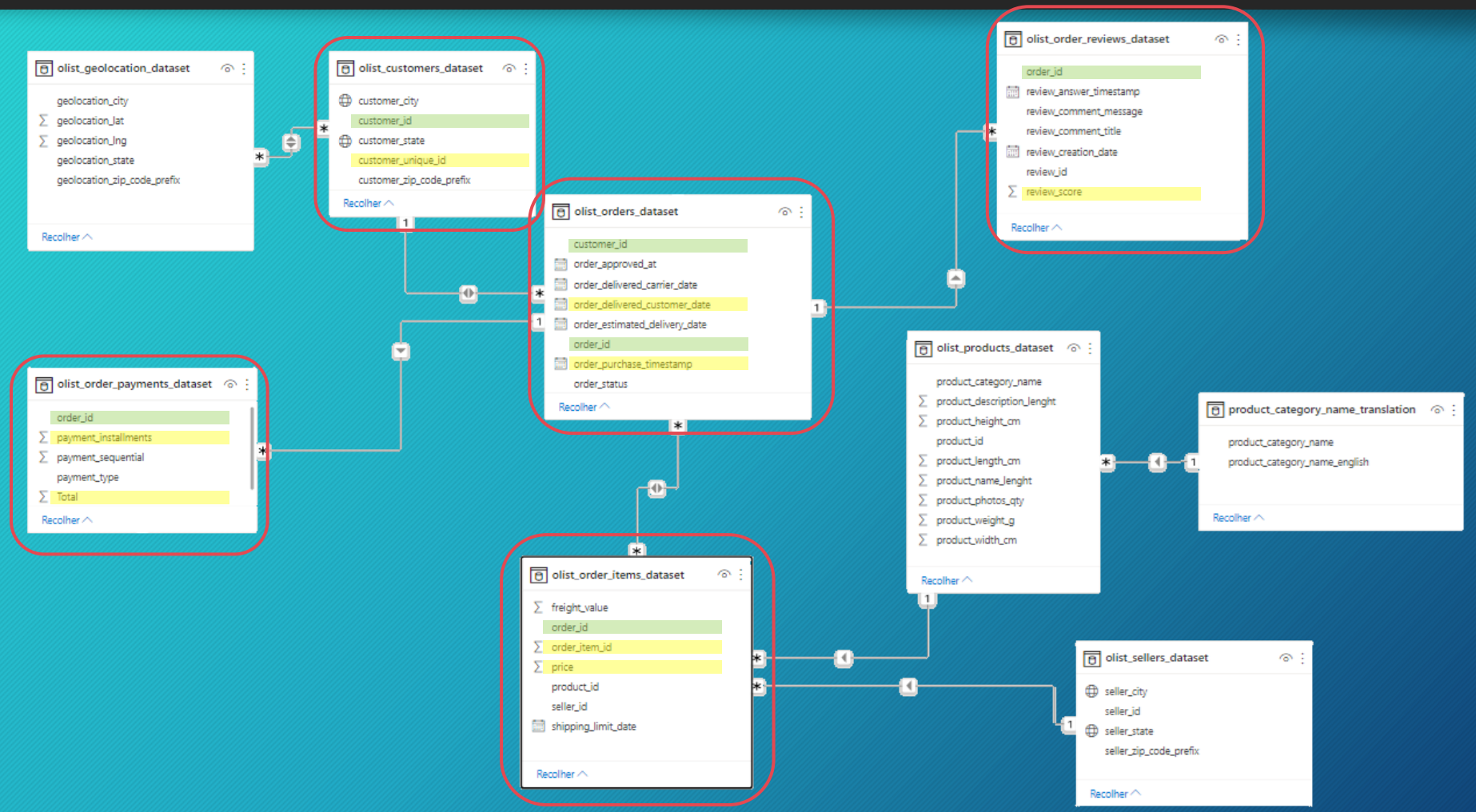
Olist

- Campagne de communication
  ➔ Segmentation des clients
- Comprendre les différents types d'utilisateurs
- Fournir à l'équipe marketing une description actionnable
  ➔ Segmentation compréhensible
- Analyse de la stabilité des segments au cours du temps
  ➔ Proposer un contrat de maintenance

Présentation du jeu de données

olist

Présentation du jeu de données



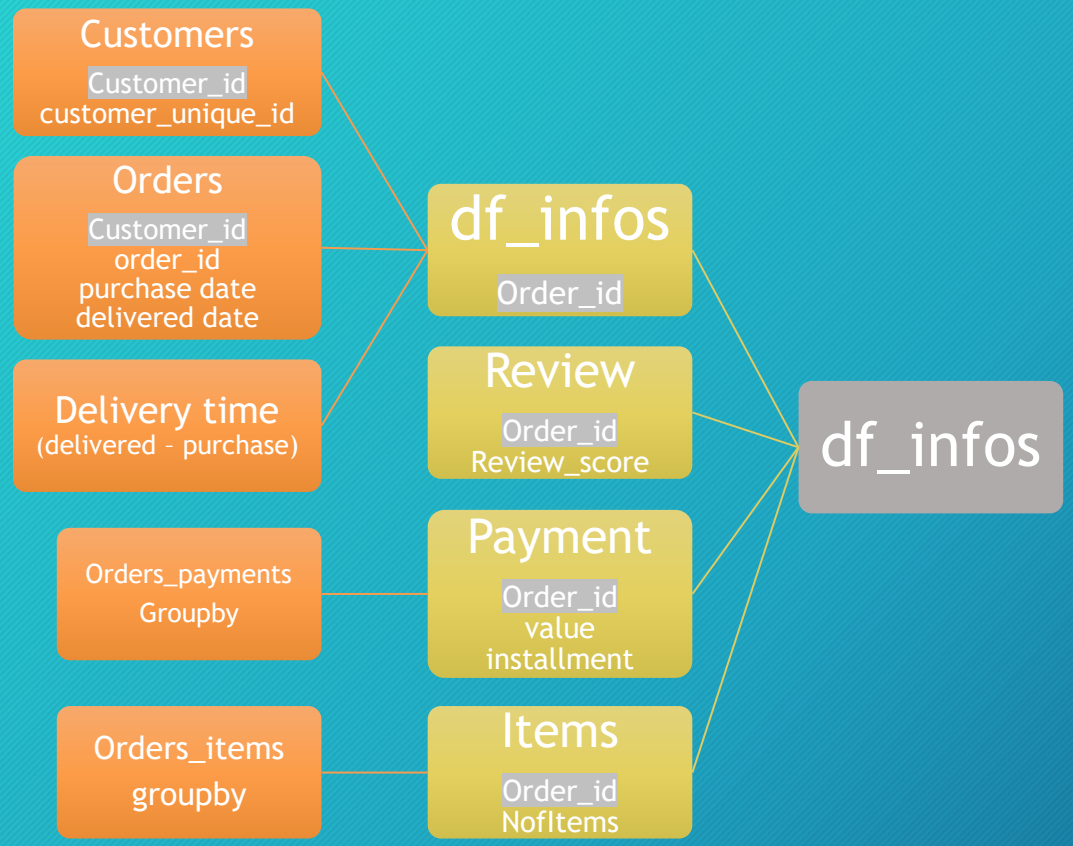
Nettoyage & Exploration

RFM — RFM Review — RFM « full »

Nettoyage & Exploration

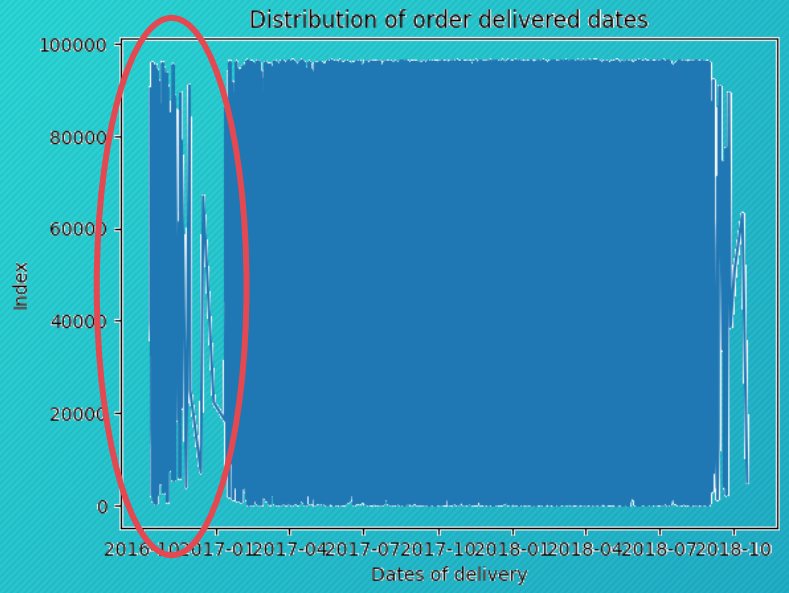
- Imports
- Chargement des fichiers de dataset :
  - orders
  - orders_payements
  - orders_items
  - orders_reviews
  - customers
- Premières vérifications :
  - head
  - duplicated
  - isna

Nettoyage & Exploration

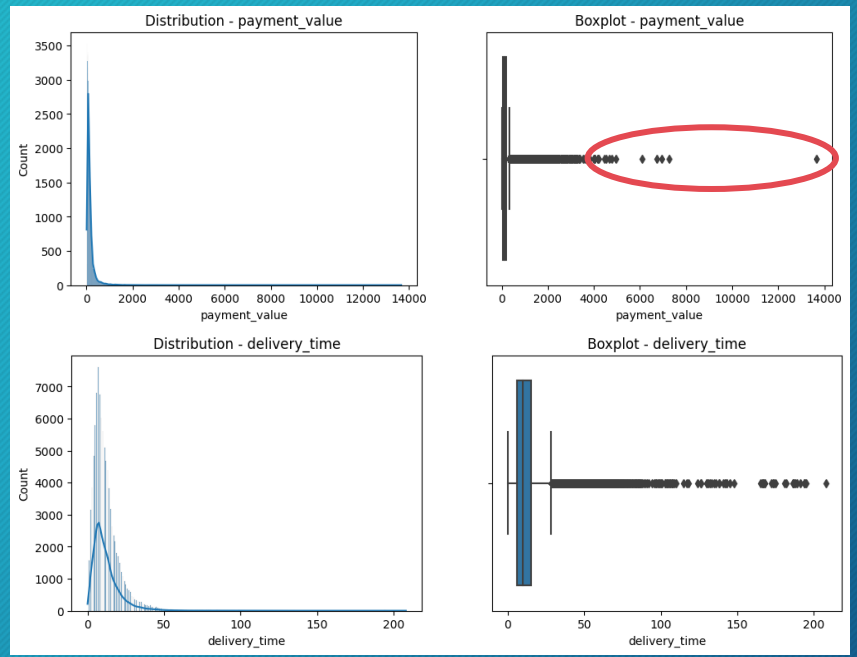


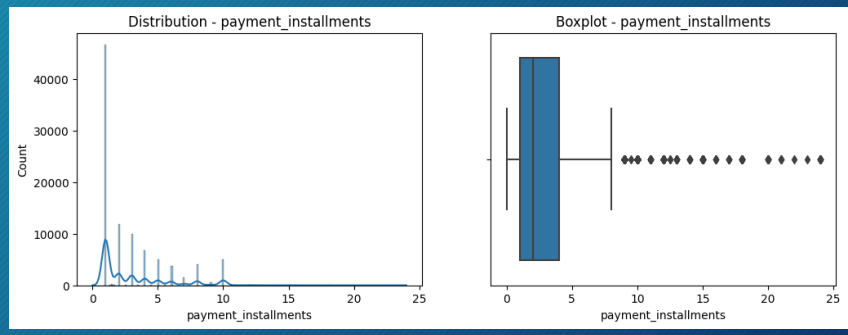
- Nettoyage :
  - Sélection des commandes livrées
  - Date de livraison NA = date estimée
  - Payment : group by order_id
    value: sum & *installment*: mean
  - Items : group by order_id
    NofItems: count
  - Merge des informations

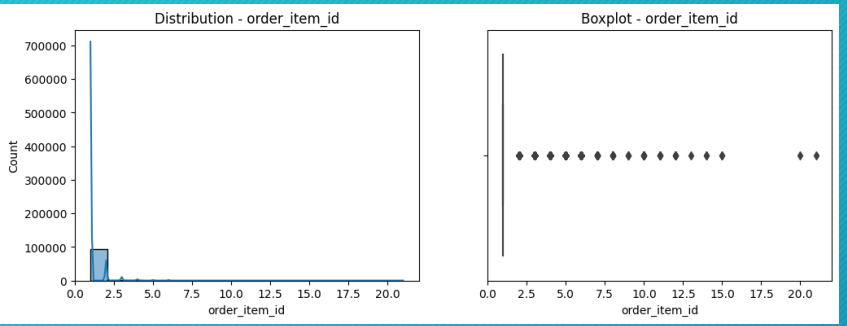
Nettoyage & Exploration

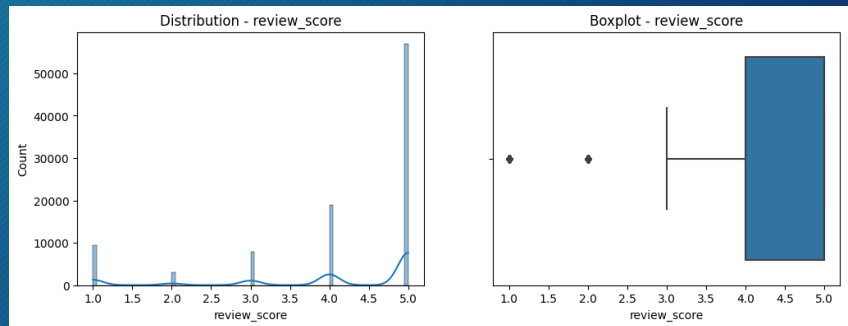


- Concentration sur 01-2017 -> 10-2018
- Suppression des achats spécifiques à gros montants

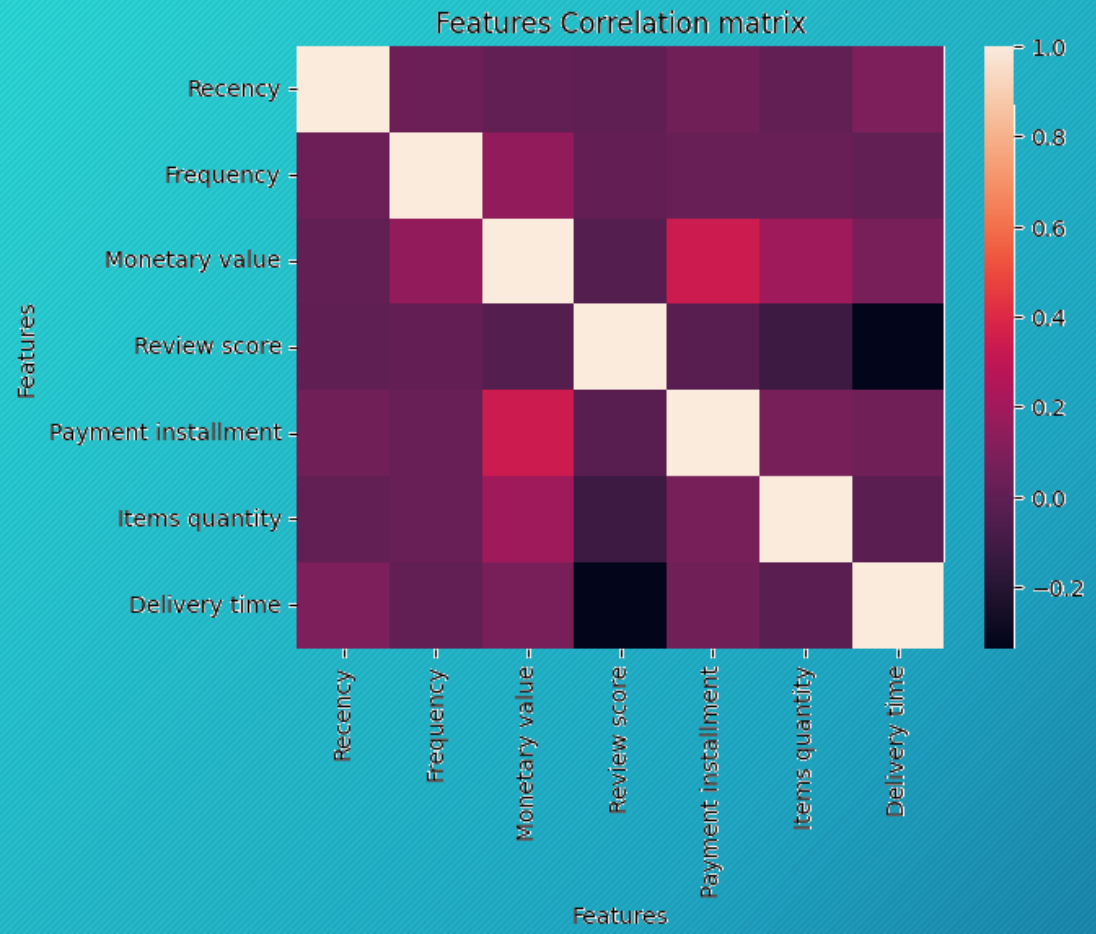








Nettoyage & Exploration



- Création « RFM » : pour chaque « customer_unique_id »
  - Recency
  - Frequency
  - Monetary value
  - Review score
  - Payment installment
  - Items quantity
  - Delivery time

- Sauvegarde :
  - RFM
  - RFM + Review
  - RFM + Review + Installement + Items + Delivery

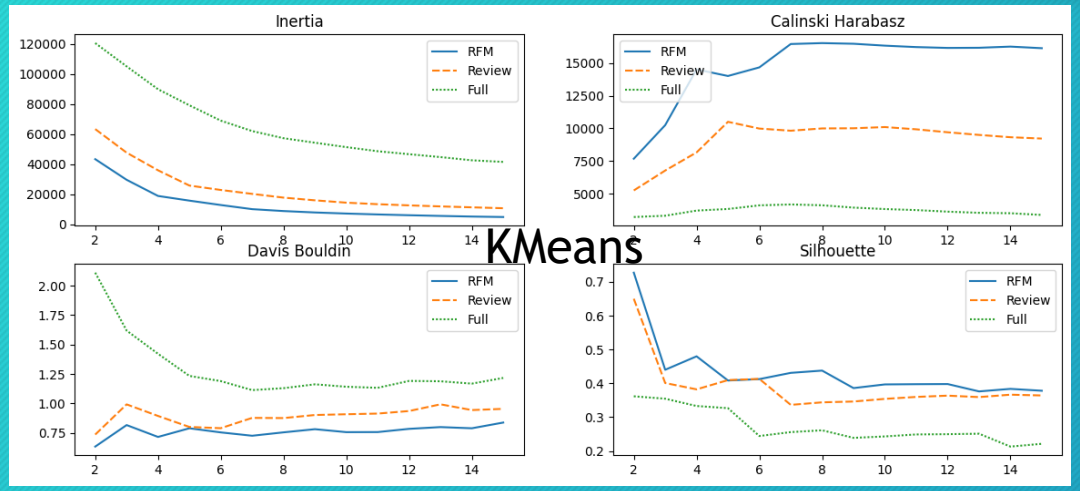
Modélisation

Algorithmes :	KMeans	-	CAH	- DBSCAN
Datasets :	RFM	-	RFM Review	- RFM Full
Scores :	Silhouette	-	Davis Bouldin	- Kalinski Harabasz

Modélisation

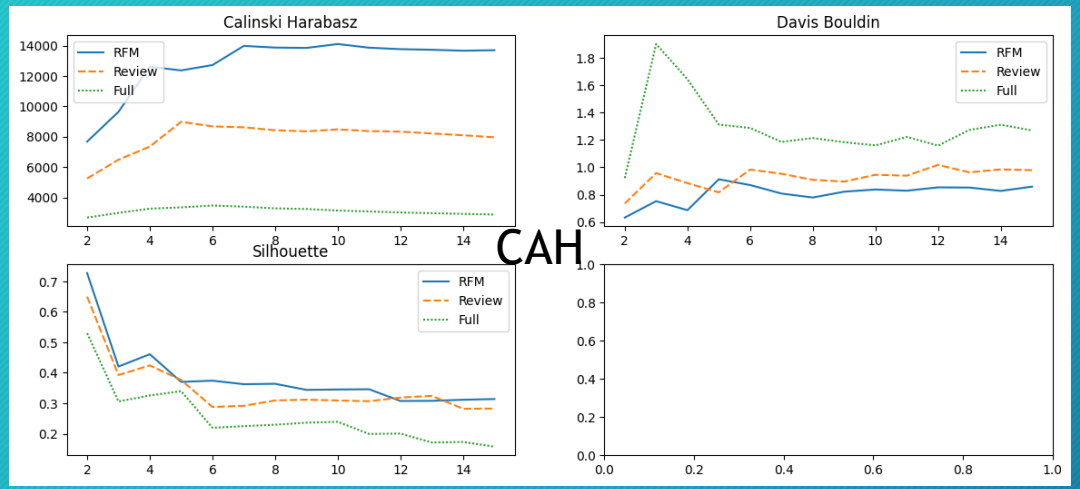
- Echantillon pour chaque dataset :
  - 20 000 samples
  - Graine fixée (random state)
- Sélection de l'algorithme :
  - Hyperparamètres
  - Scores
  - t-SNE
- Sélection du dataset :
  - Barplot
  - t-SNE
  - radar plot

Modélisation – Sélection de l'algorithme

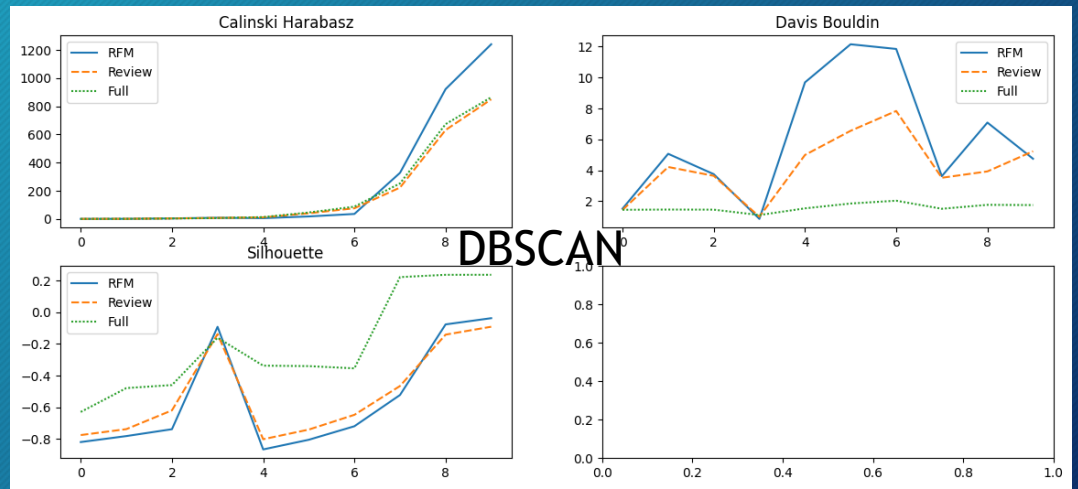


KMeans

- Kmeans & CAH
  - RFM : n_clusters=4
  - RFM review : n_clusters=5
  - RFM full : n_clusters=5
- DBSCAN
  - eps=0.056, min_samples=5



CAH

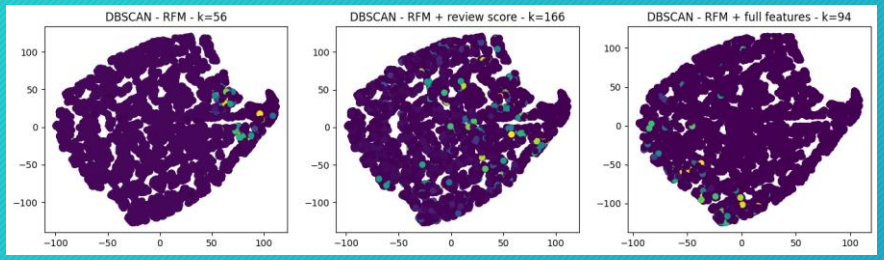


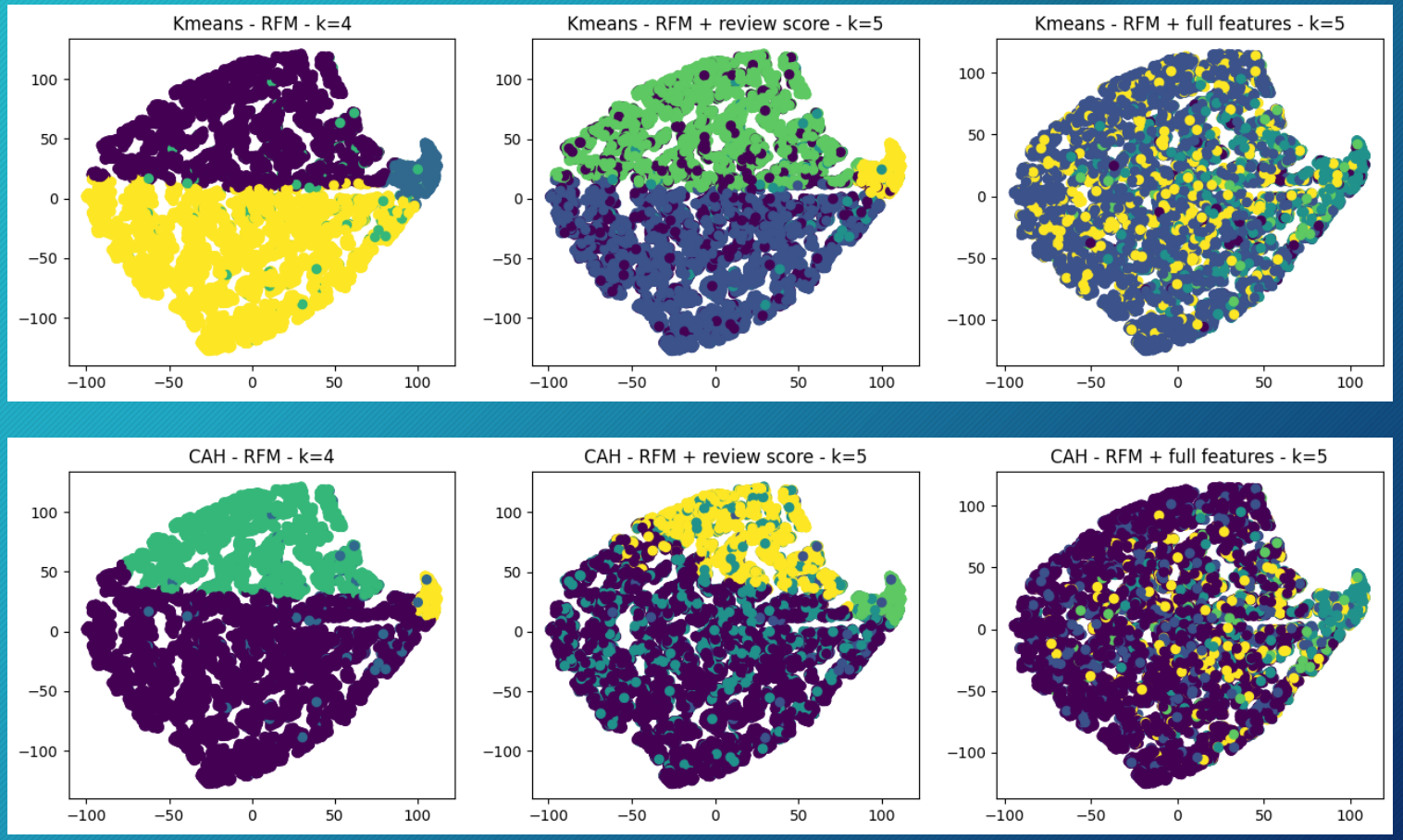
DBSCAN

Modélisation – Sélection de l'algorithme

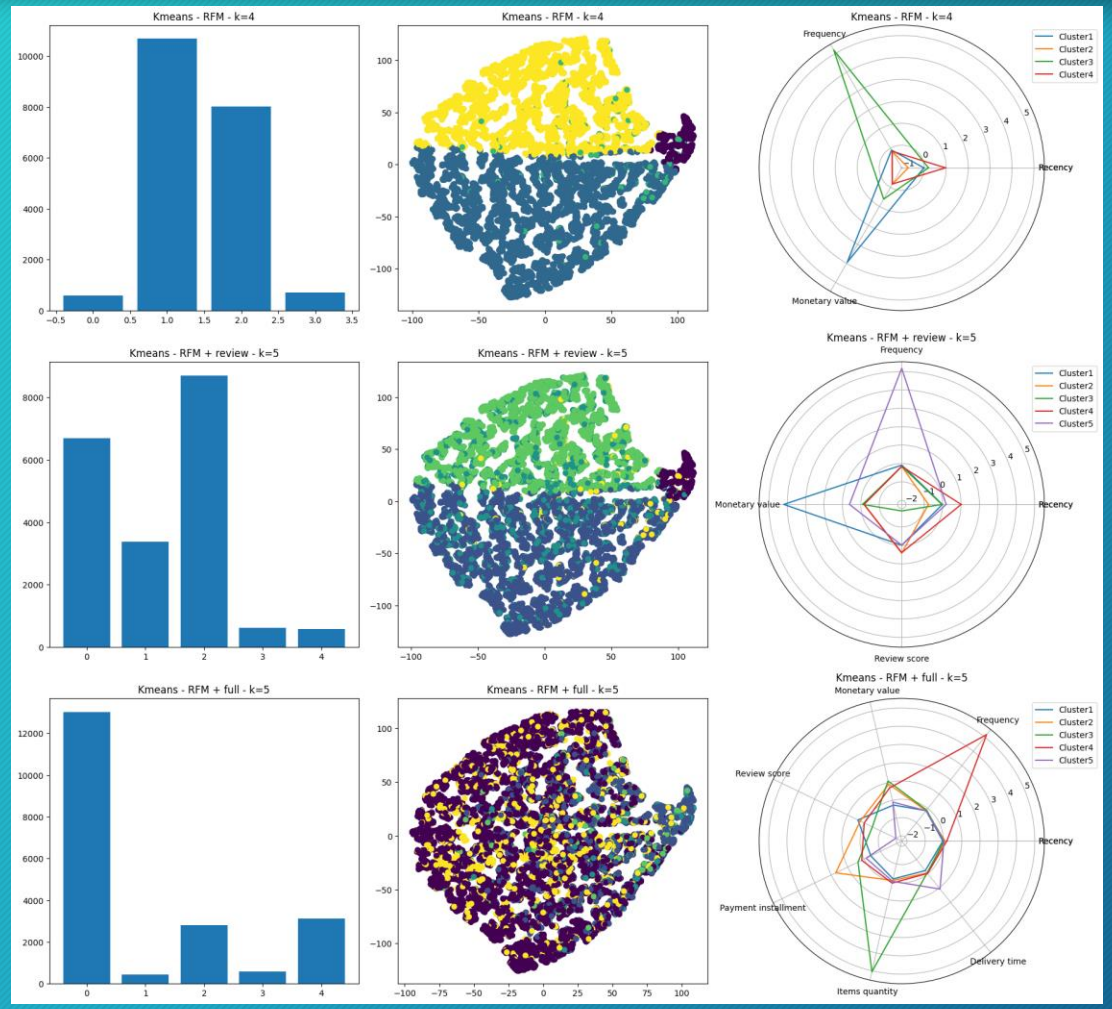
Meilleur algorithme :

- KMEANS





Modélisation – Sélection du dataset

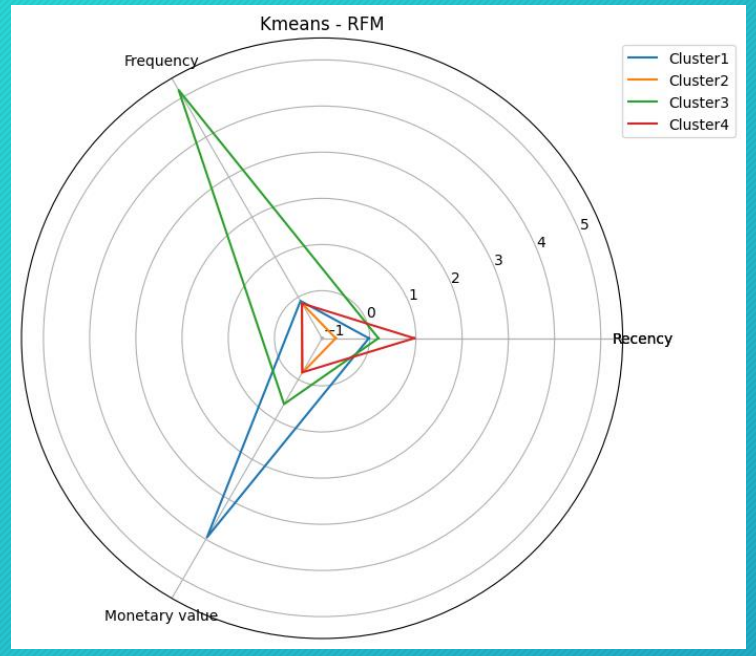


RFM

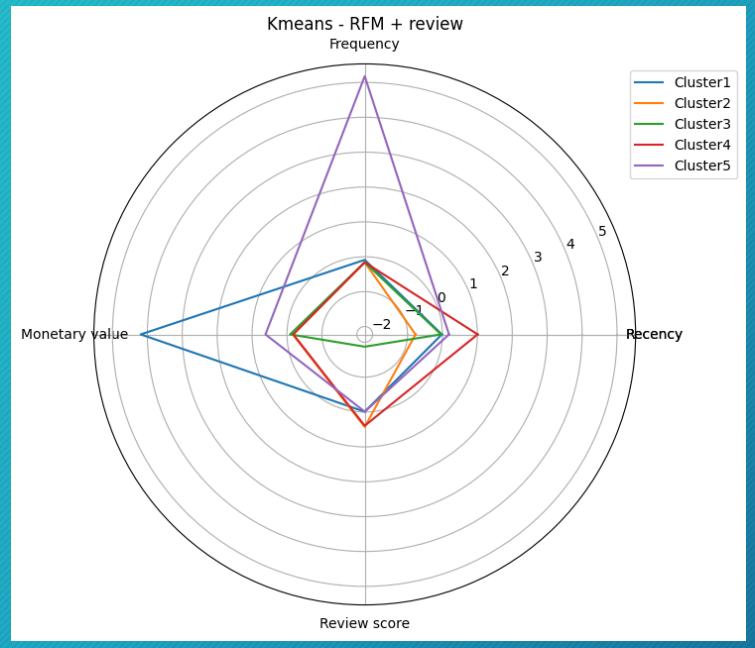
RFM review

RFM full

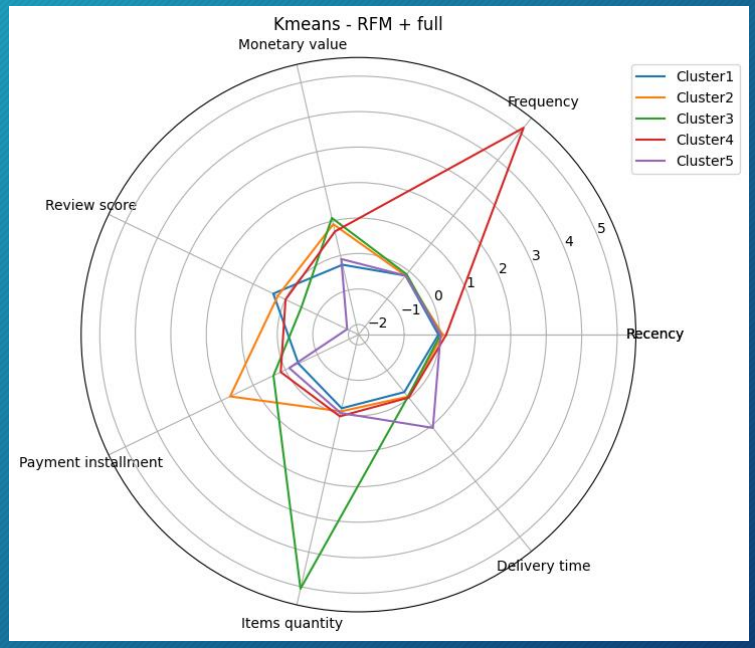
Modélisation – Sélection du dataset



Cluster1 : Gros acheteur (occasionel)
Cluster2 : Acheteur "classique"
Cluster3 : Acheteur récent
Cluster4 : Acheteur régulier



Cluster1 : Gros acheteur (occasionel)
Cluster2 : Acheteur "classique"
Cluster3 : Acheteur mécontent
Cluster4 : Acheteur récent
Cluster5 : Acheteur régulier



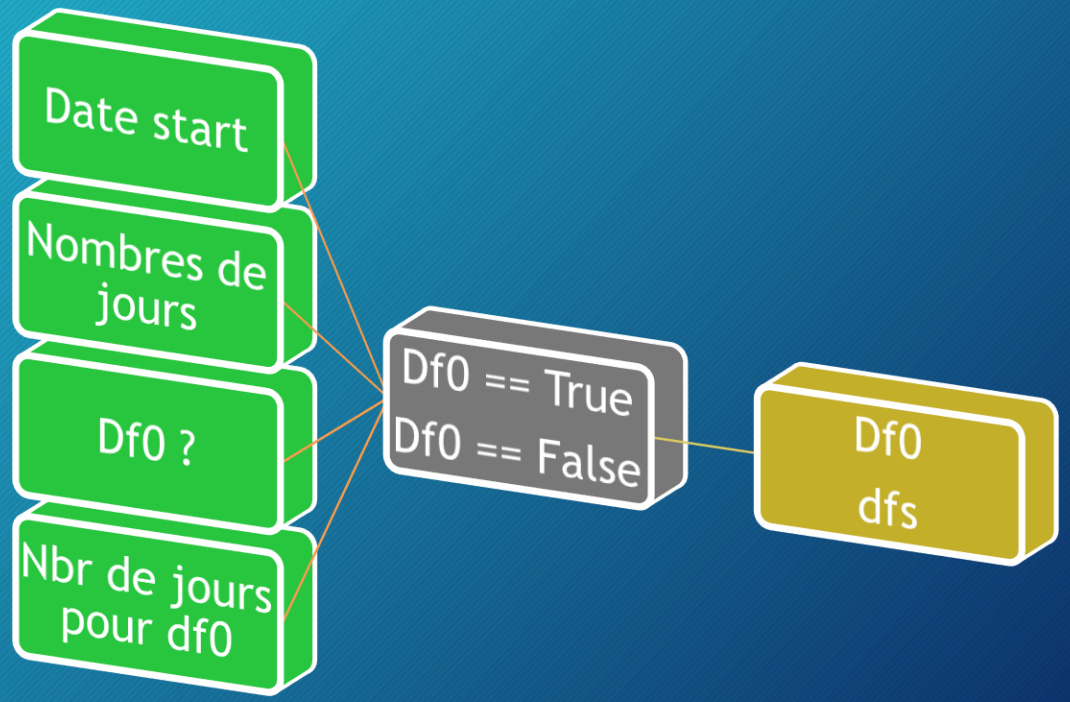
Cluster1 : Acheteur "classique" content
Cluster2 : Gros acheteur (occasionel)
Cluster3 : Acheteur en quantité
Cluster4 : Acheteur régulier
Cluster5 : Acheteur mécontent

Maintenance

Simulation de décalage temporel

Maintenance

- Création de fonction :
  - Automatisation Nettoyage & Création dataframe
  - Date start = 01-01-2017
  - N_days = 15
  - N_days_d0 = 365
- Création de df0
  01-01-2017 + 365 jours
- Création de dictionnaire dfs
  01-01-2017 ➔ 01-01-2018
  +15 jours / df



Maintenance

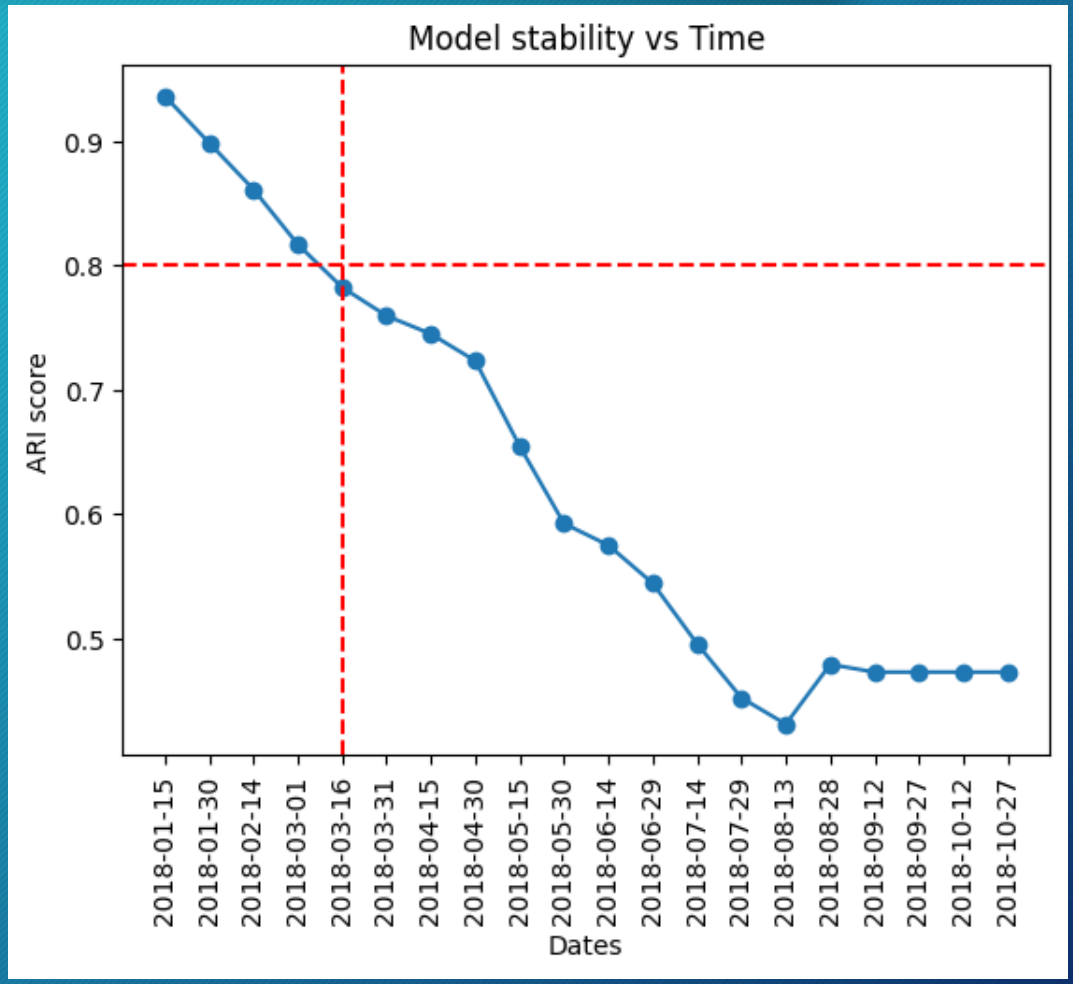
olist

- StandardScaler
- Kmeans(n_clusters=5)
- Adjusted Rand Index score

dataframe	Df0	df1	...	dfn
scaler	Sc0 (fitted on df0)	Sc1 (fitted on df1)		
Scaled dataframe	Df0_sc0	Df1_sc1 Df1_sc0		
Models	Km0 kmeans fitted on df0_sc0	Km1 kmeans fitted on df1_sc1		
ARI score		Label0 vs label1 ←→ km0.predict(df1_sc0) vs km1.predict(df1_sc1)		

- Boucle for : score ari (predict Km0(dfx), predict Km(dfx))

- Limite de performance fixée à 0,80
- Modèle obsolète au bout de ~ 3 mois



Conclusion

Conclusion

- Modèle : Kmeans avec 5 clusters
- Dataset : RFMR

- Maintenance tout les 3 mois
  - Changement du jeu d'entrainement
  - Vérification de la cohérence du modèle
  - Vérification de la segmentation
  - Vérification du rythme de maintenance