OPENCLASSROOMS

Kevin

Parcours Data Scientist

Problématique

L'entreprise

- Entreprise brésilienne
- Propose une solution de vente e-commerce

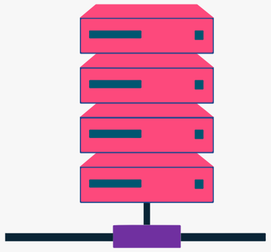
olist

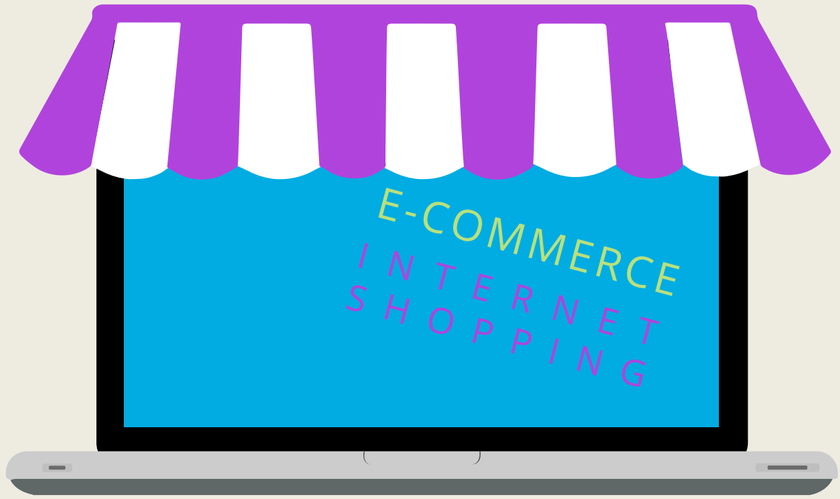
L'objectif

- Segmenter des clients dans un but marketing
- Fournir une description de la segmentation
- Eviter l'effet « boite noire »
- Evaluation de la stabilité temporelle de la segmentation

Les données

- Base de données anonymisée
- Peu de données
- 3% de clients avec plus d'une commande





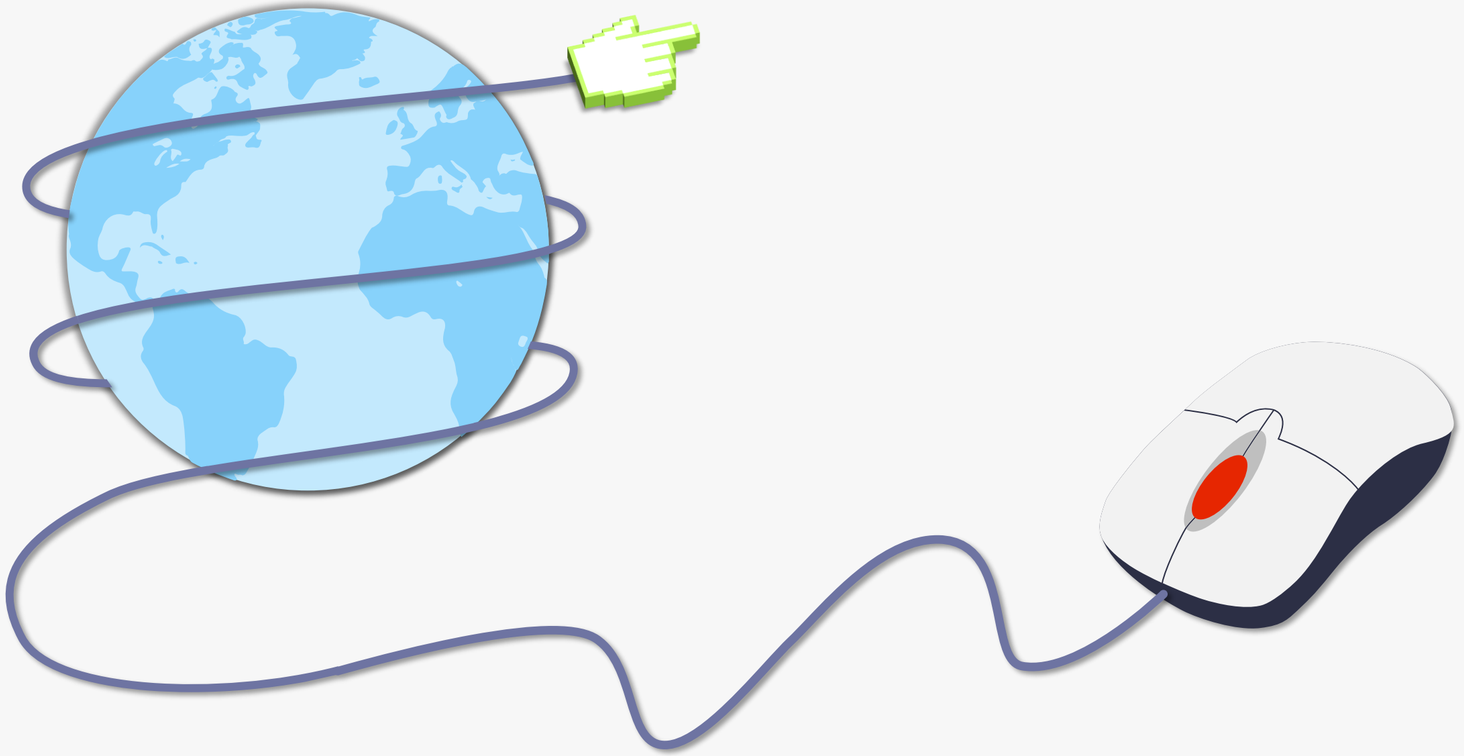
Data

Modélisation

Contrat de maintenance

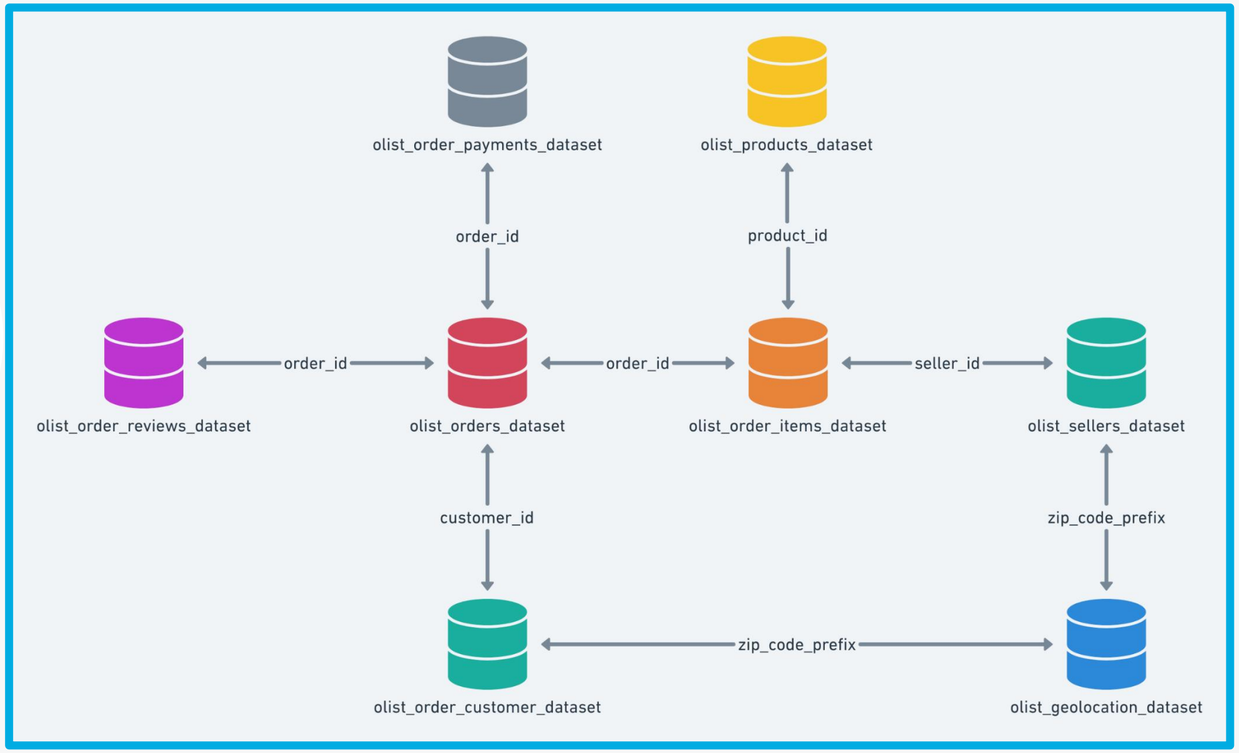
1

Data



Data


Customers
Sellers
Geolocation
Products translated
OLIST
Order_items
Products
Order_payments
Orders
Order_reviews

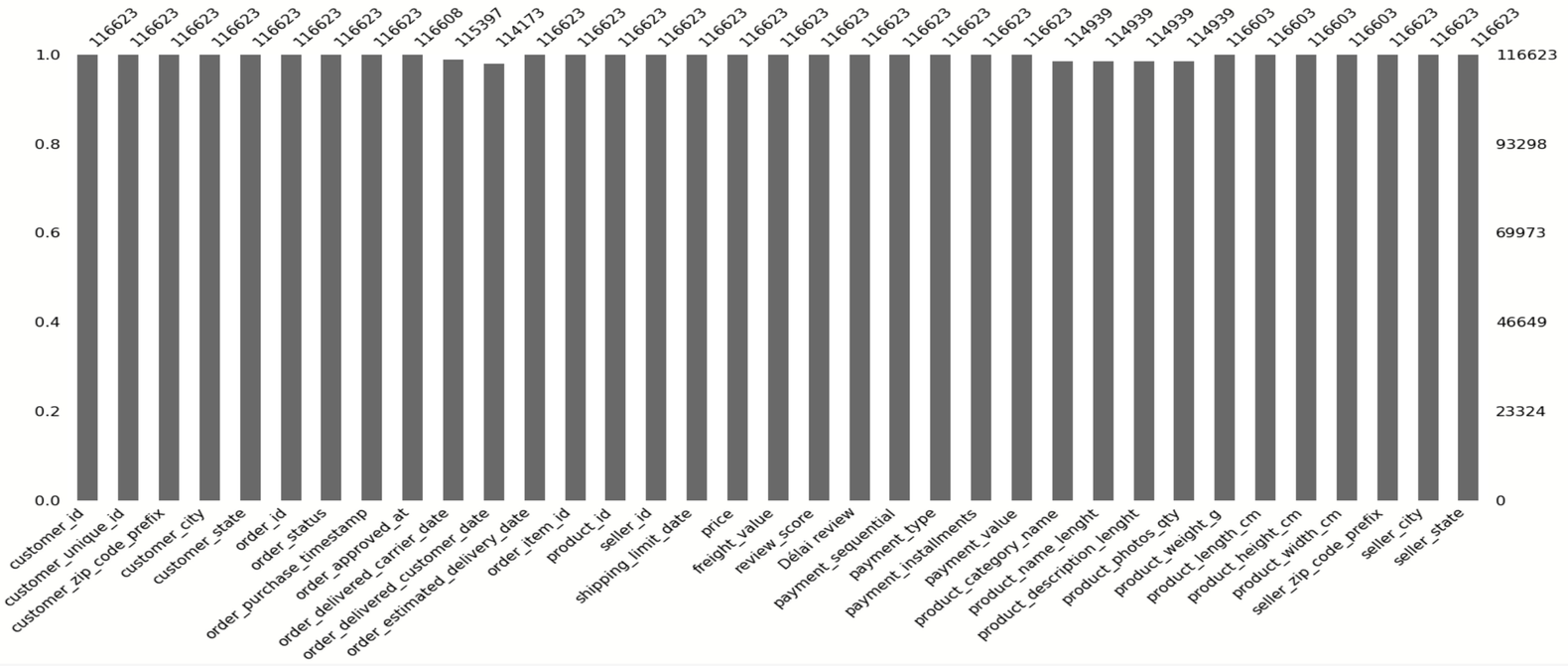

olist_order_payments_dataset
olist_products_dataset
order_id
product_id
order_id
order_id
seller_id
olist_order_reviews_dataset
olist_orders_dataset
olist_order_items_dataset
olist_sellers_dataset
customer_id
zip_code_prefix
zip_code_prefix
olist_order_customer_dataset
olist_geolocation_dataset

Caractéristiques de la data

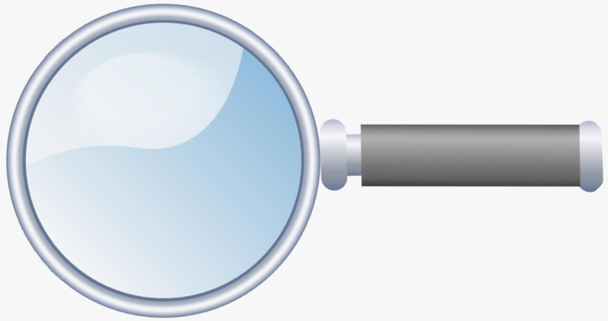
Describe

Valeurs manquantes

Doublons



Commande en doublon



Commande

Produit

Review

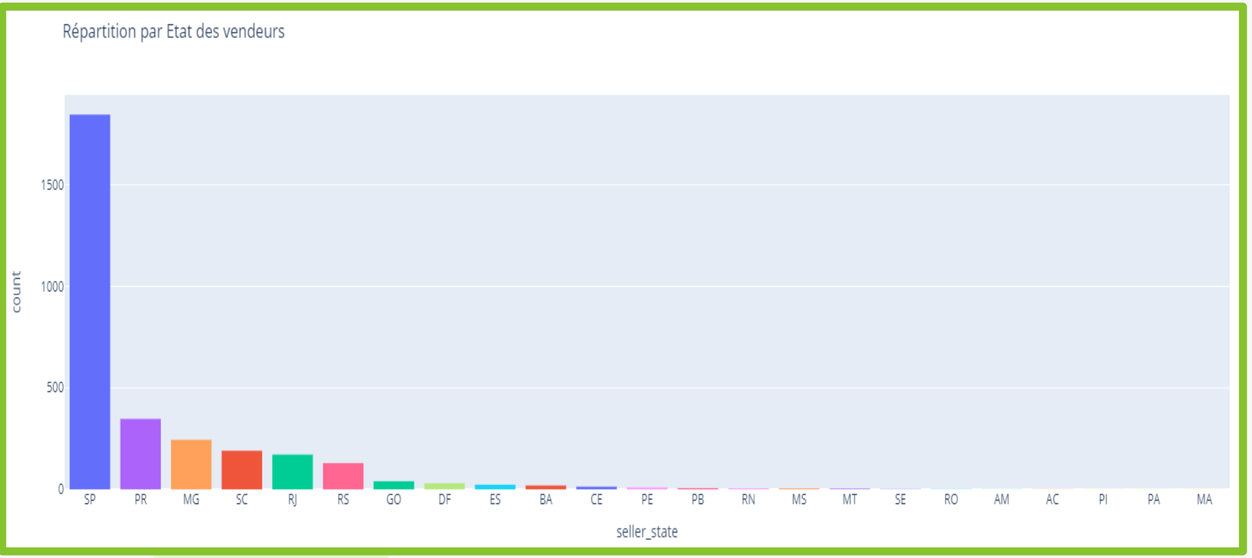
Paiement plusieurs fois

Base de données

1 commande = 1 ligne ?

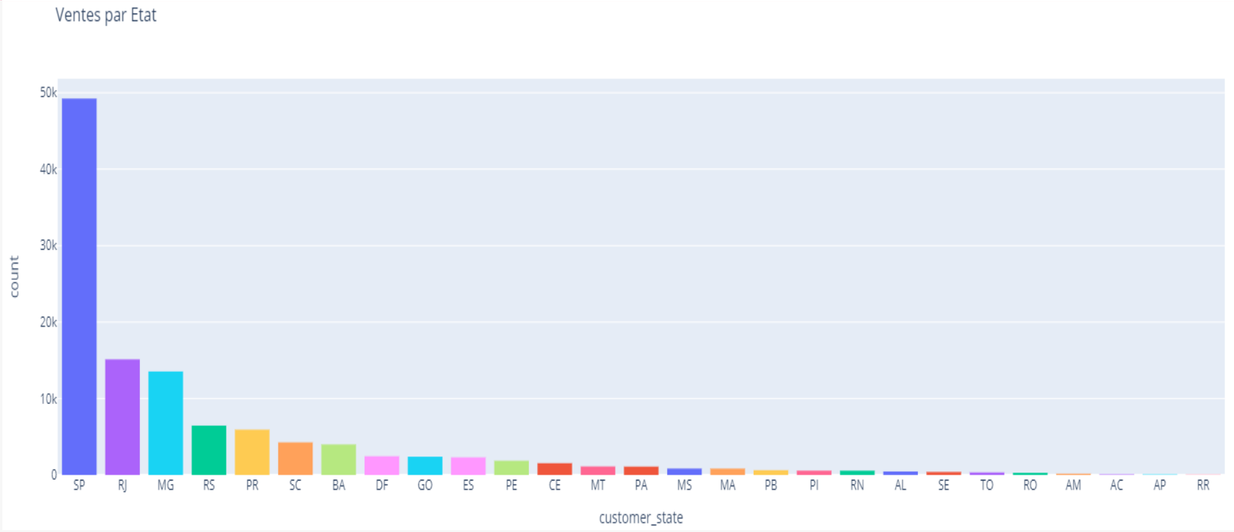
16% identifiants commande doublon

Visualisation des transactions



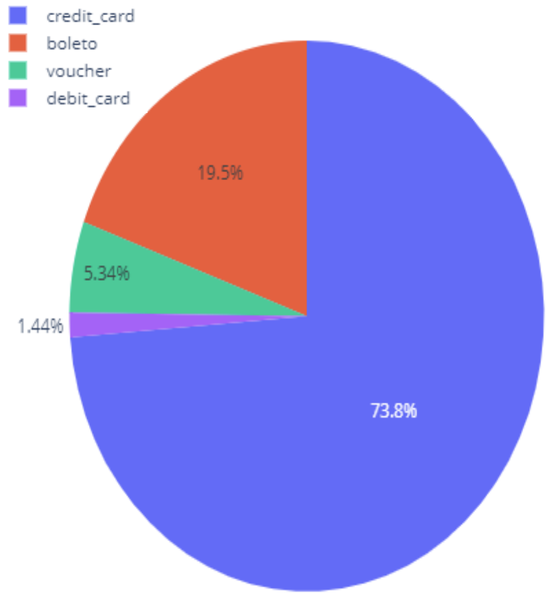
Localisation des vendeurs

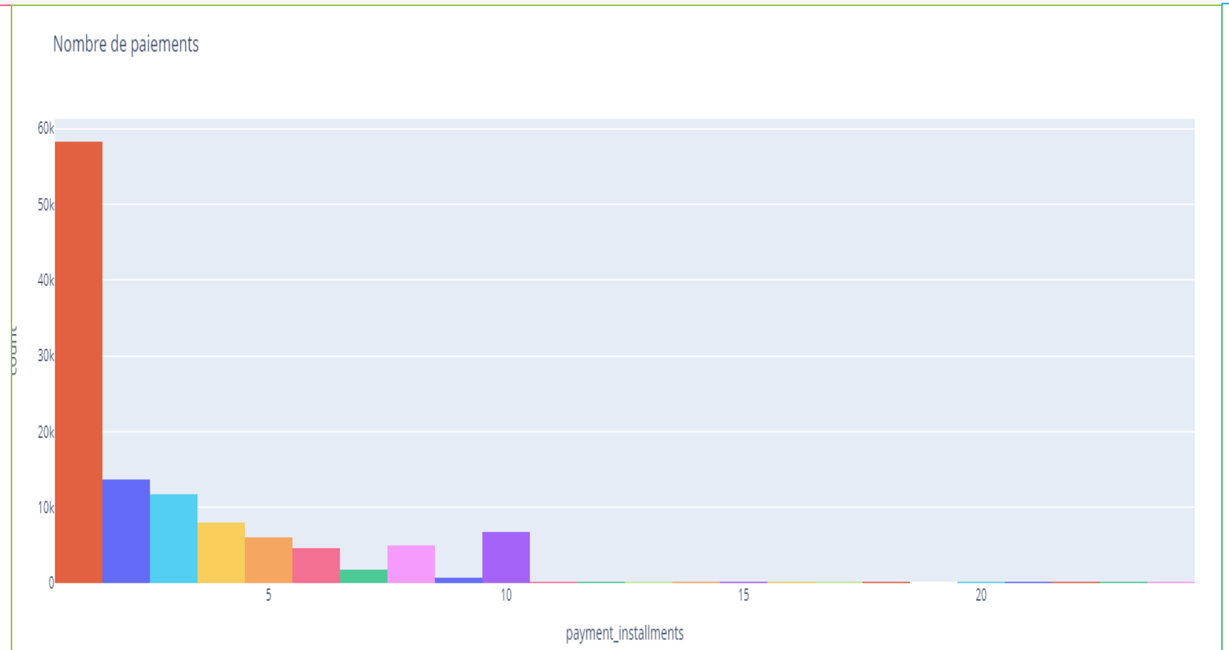
Localisation des acheteurs

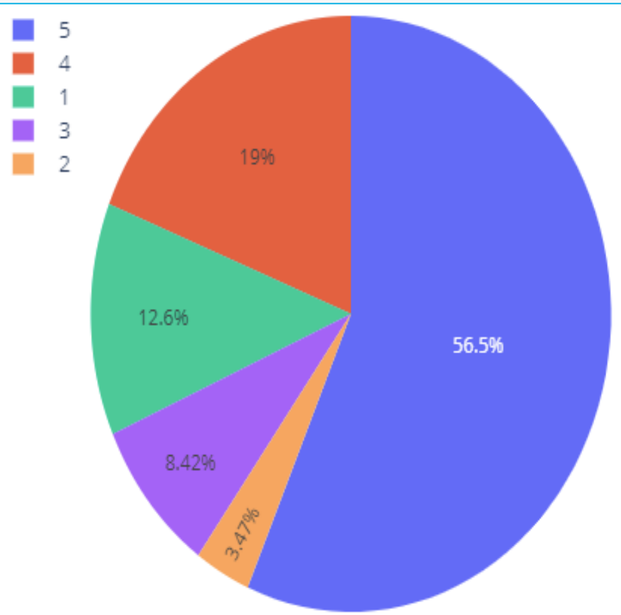


27 Etats où il y a des ventes, réparties dans 4108 villes

Variables discriminantes







Moyen de paiements

Nombre de paiements

Review

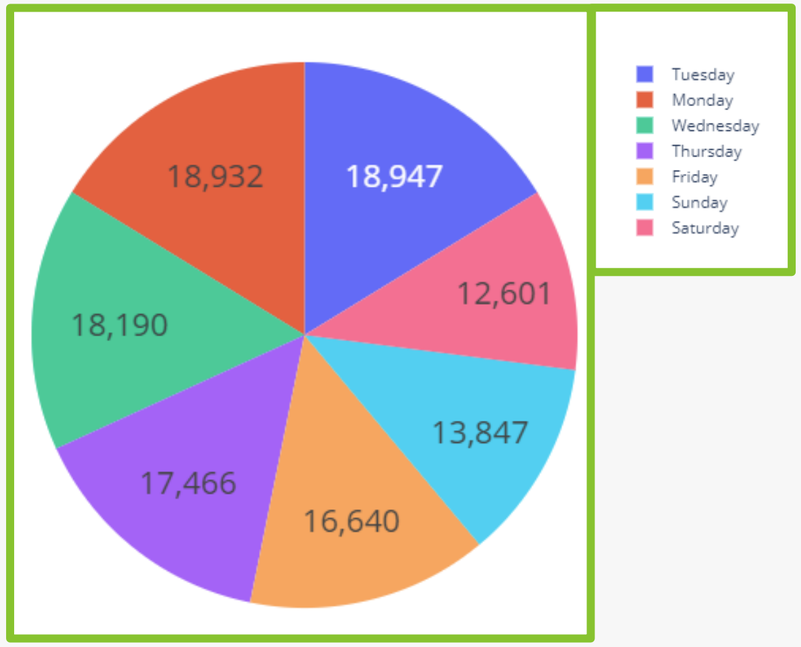
Feature Engineering

1. **Suppression variables inutiles**
   Informations sur les vendeurs | Commandes toujours en cours

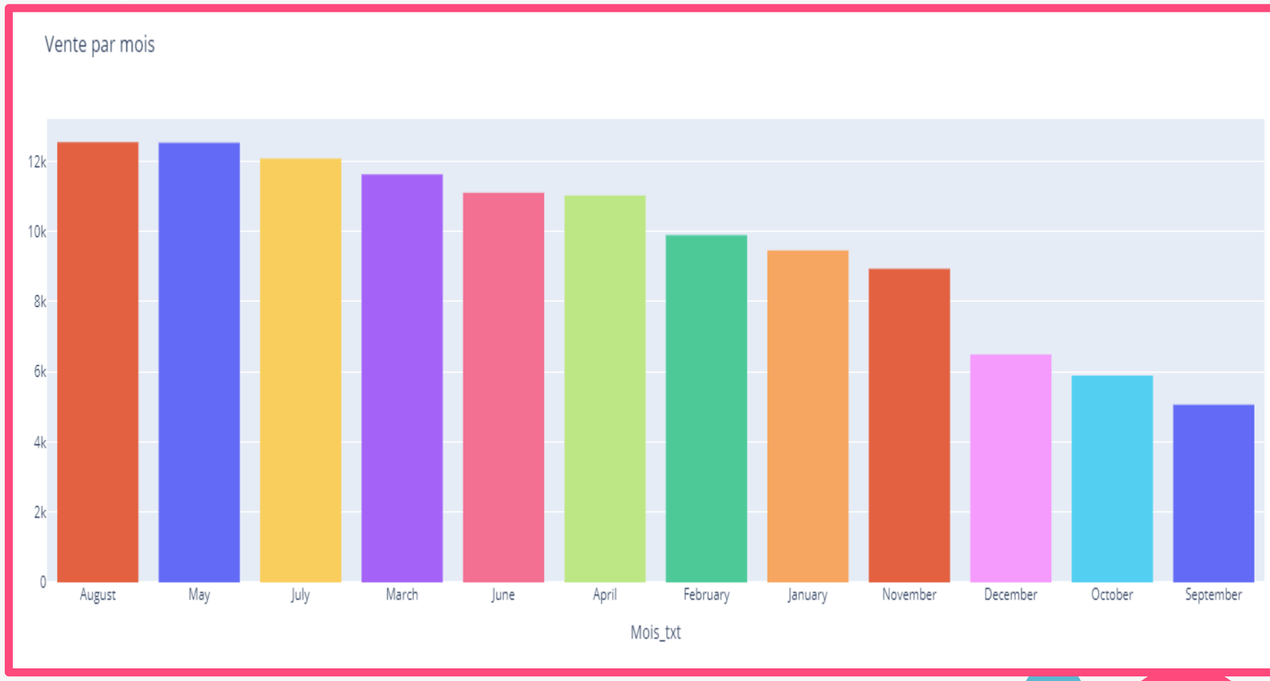
2. **Date de la commande**
   Jour et Mois

Ventes

Jour



Mois



Feature Engineering

1. **Suppression variables inutiles**

   Informations sur les vendeurs | Commandes toujours en cours

2. **Date de la commande**

   Jour et Mois

3. **Volume du produit**

   *Hauteur * largeur * longueur*

4. **Durée entre date d'achat et livraison**

   Date de livraison – Date d'achat

5. **Dernière commande**

   Date d'Achat le plus recent – Date de la dernière commande de chaque client

Dernières commandes

Distribution des dernières commandes (en jours)

Feature Engineering

1. **Suppression variables inutiles**
   Informations sur les vendeurs | Commandes toujours en cours
2. **Date de la commande**
   Jour et Mois
3. **Volume du produit**
   *Hauteur * largeur * longueur*
4. **Durée entre date d'achat et livraison**
   Date de livraison – Date d'achat
5. **Dernière commande**
   Date d'Achat le plus recent – Date de la dernière commande de chaque client
6. **Regroupements**

Regroupement

Prix total

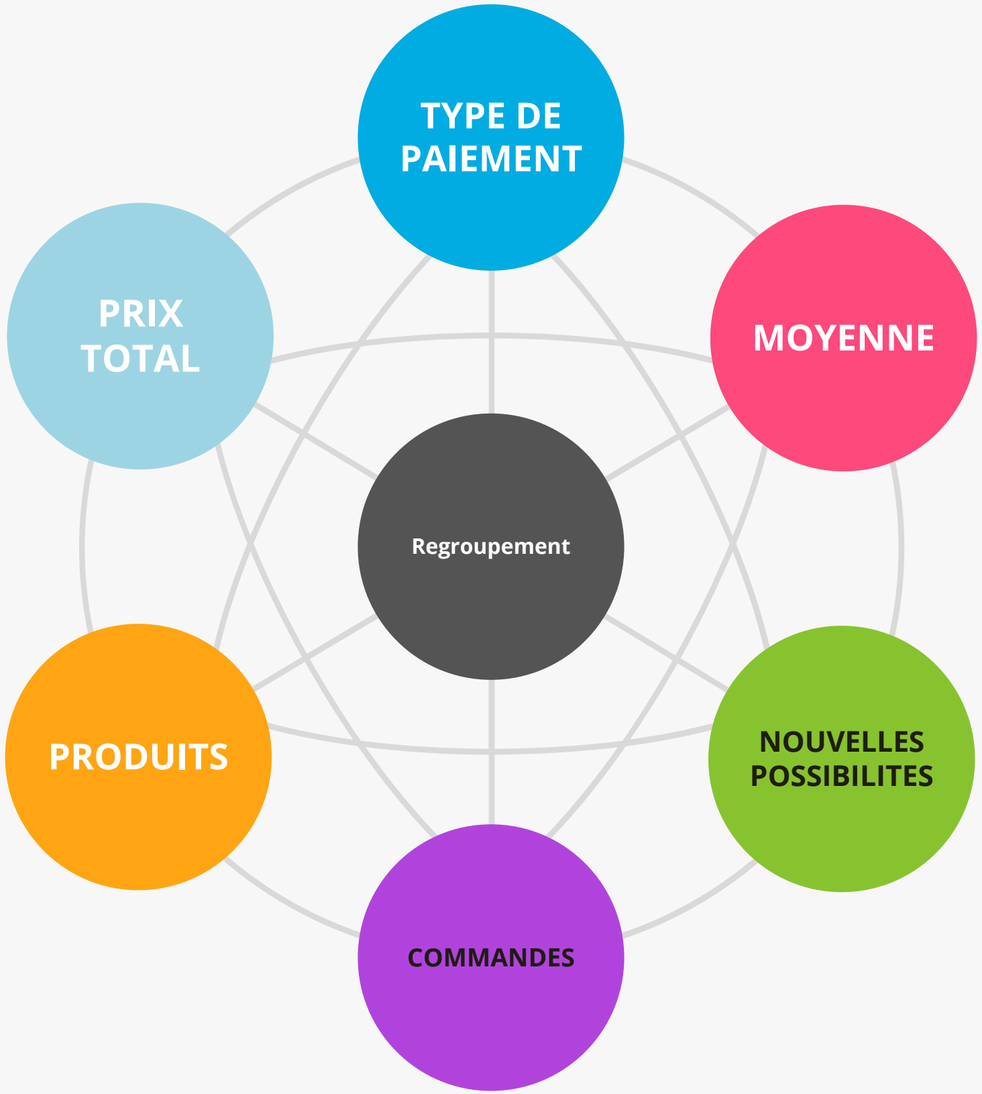
Cumul des achats de chaque client

Nombre de produits

Attention aux échéances…

Nombre de commandes

Affirmation d'Olist correcte



Type de paiement

Nombre de moyens de paiement utilisé

Moyenne

Review score

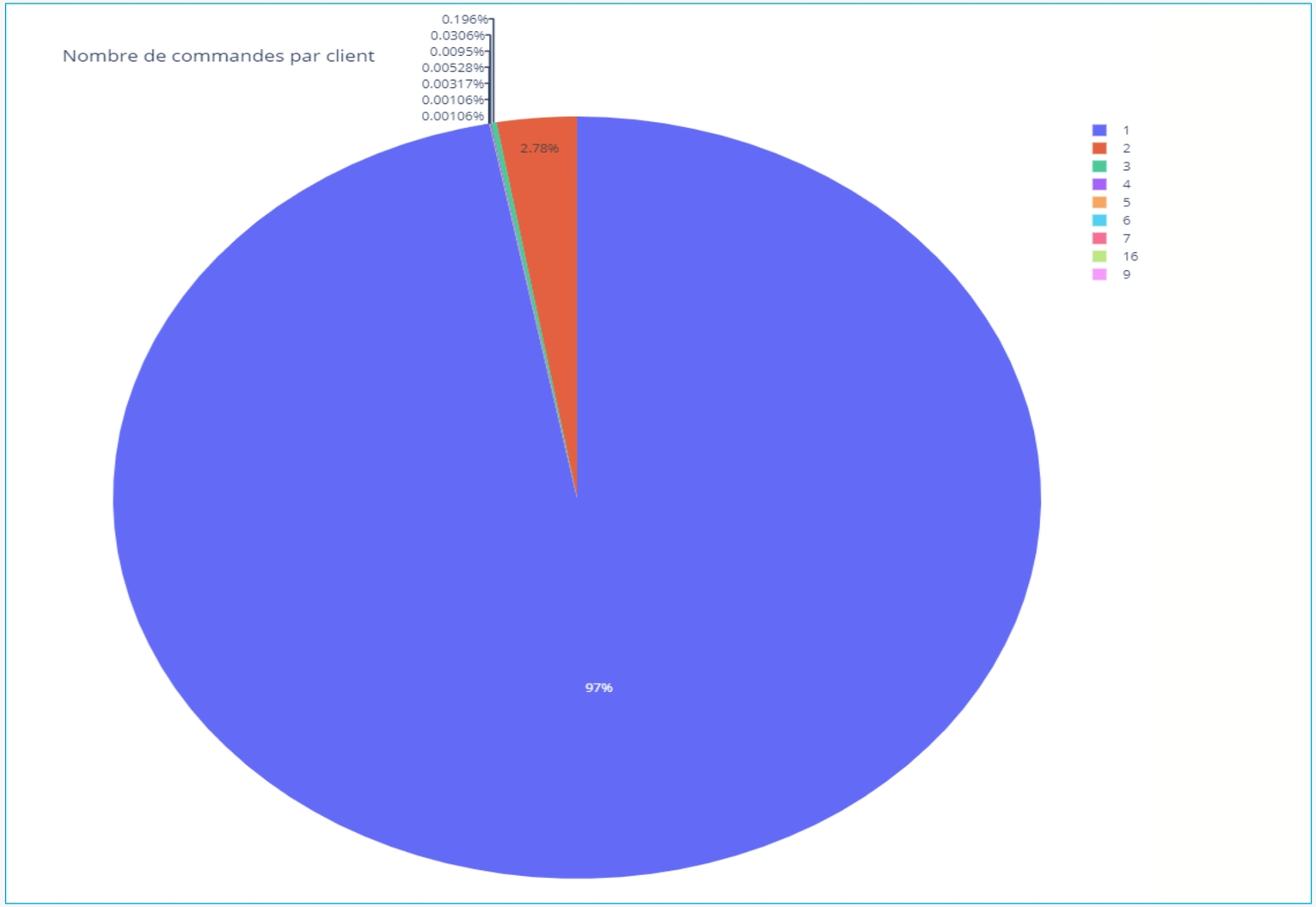
Délai de review moyen

Echances moyennes…

Nouvelles possibilités

Panier moyen

Nombre de commandes par client



Regroupement

Prix total

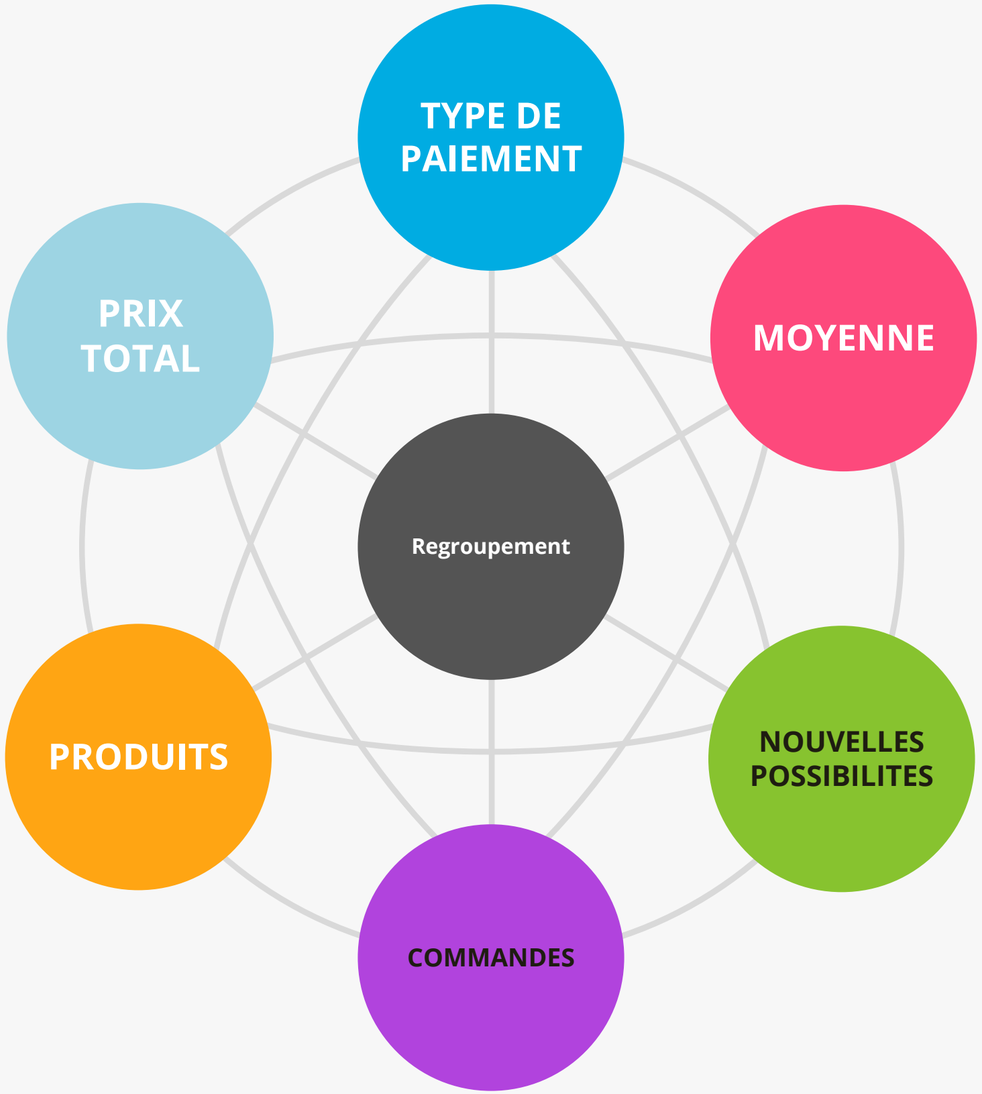
Cumul des achats de chaque client

Nombre de produits

Attention aux échéances…

Nombre de commandes

Affirmation d'Olist correcte



Type de paiement

Nombre de moyens de paiement utilisé

Moyenne

Review score

Délai de review moyen

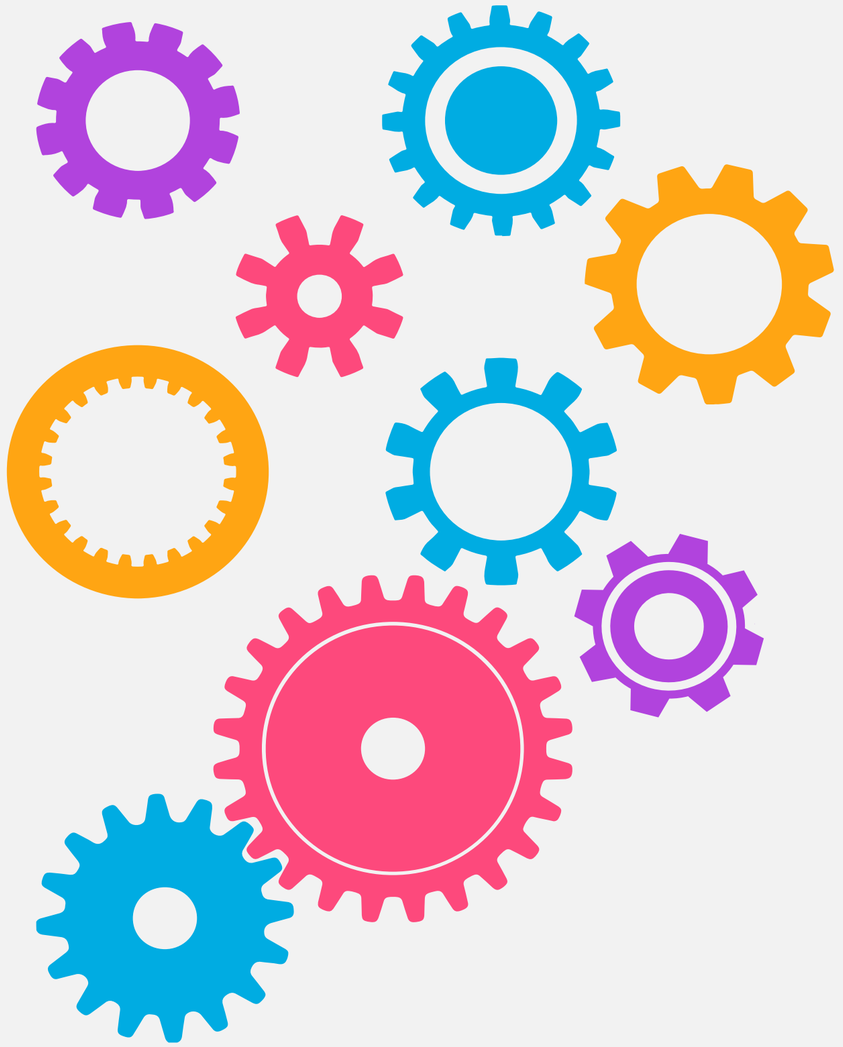
Echances moyennes…

Nouvelles possibilités

Panier moyen

2

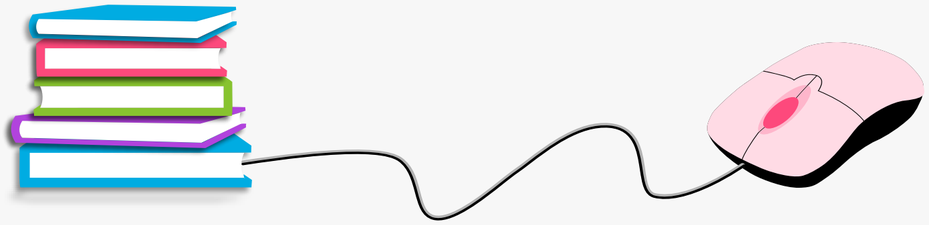
Modélisation



Segmentation RFM

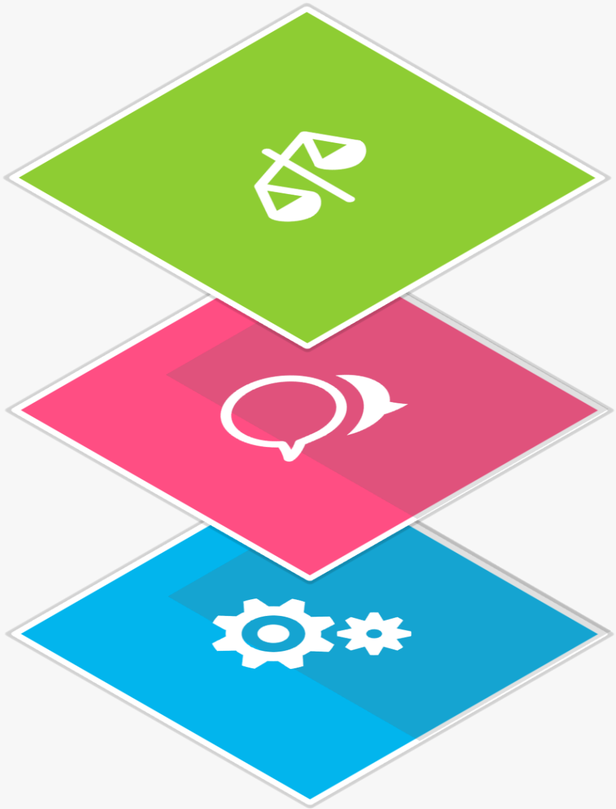
- Dernière commande
- Nombre de commandes
- Prix total payé

Protocole



Apprentissage non-supervisée

La machine apprend sur la structure des données



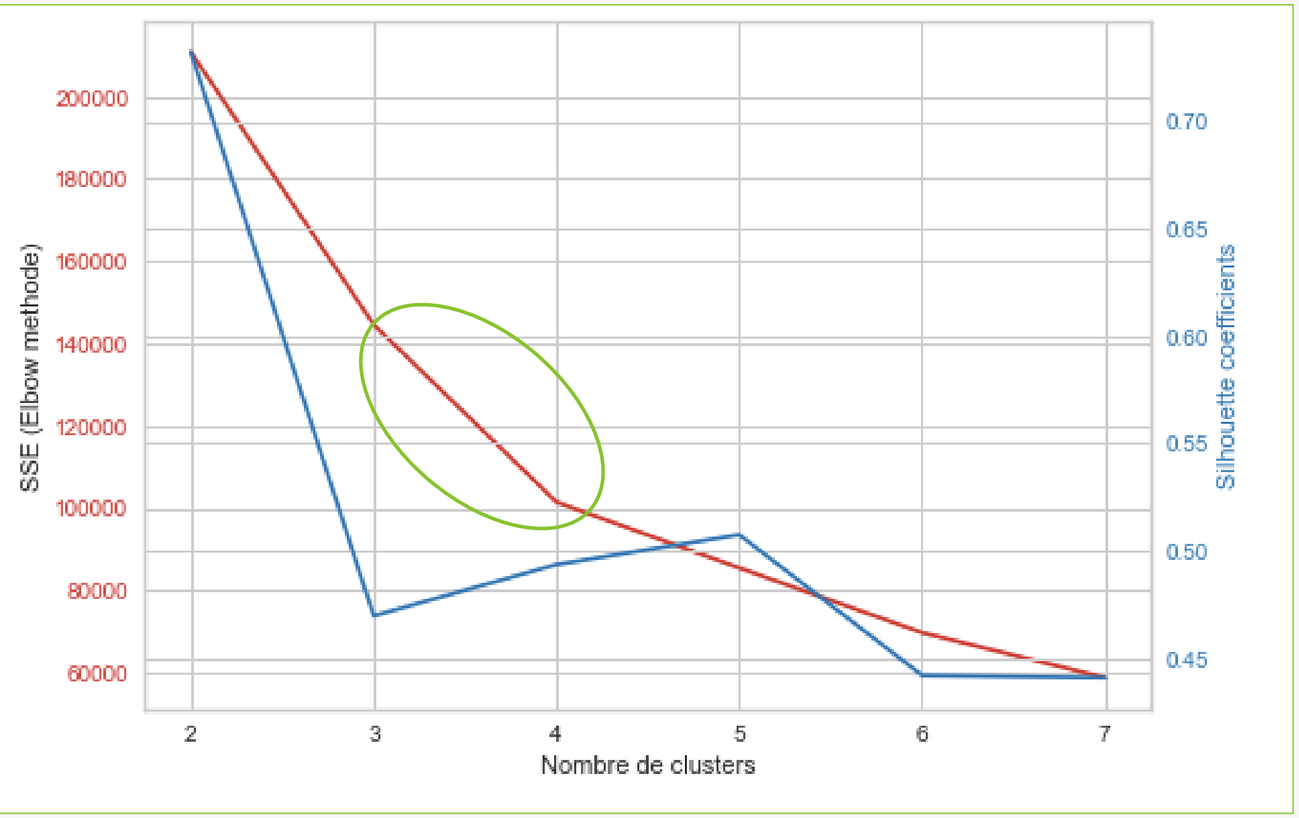
Clustering

Classer des données selon leurs similitudes

K-means

- Centroid
- Chaque point est rattaché à un centroid
- Le centroid est déplacé au milieu du cluster
- Attention aux positions initiales des centroids

Selectionner nos clusters



Elbow Methode

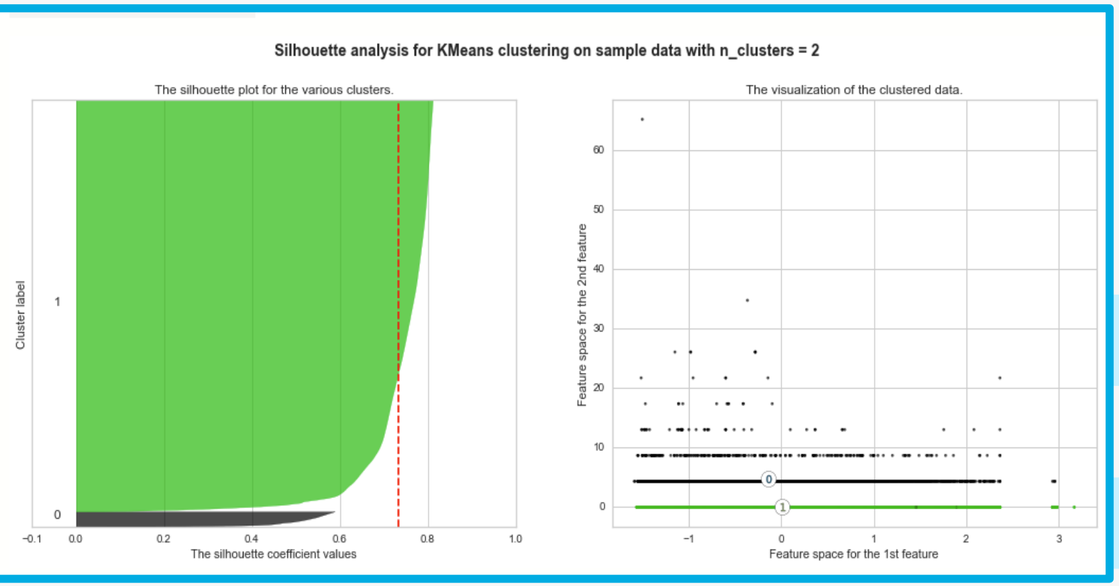
- Détecter une zone de “coude”
- Prendre un nombre de clusters proche du “coude”

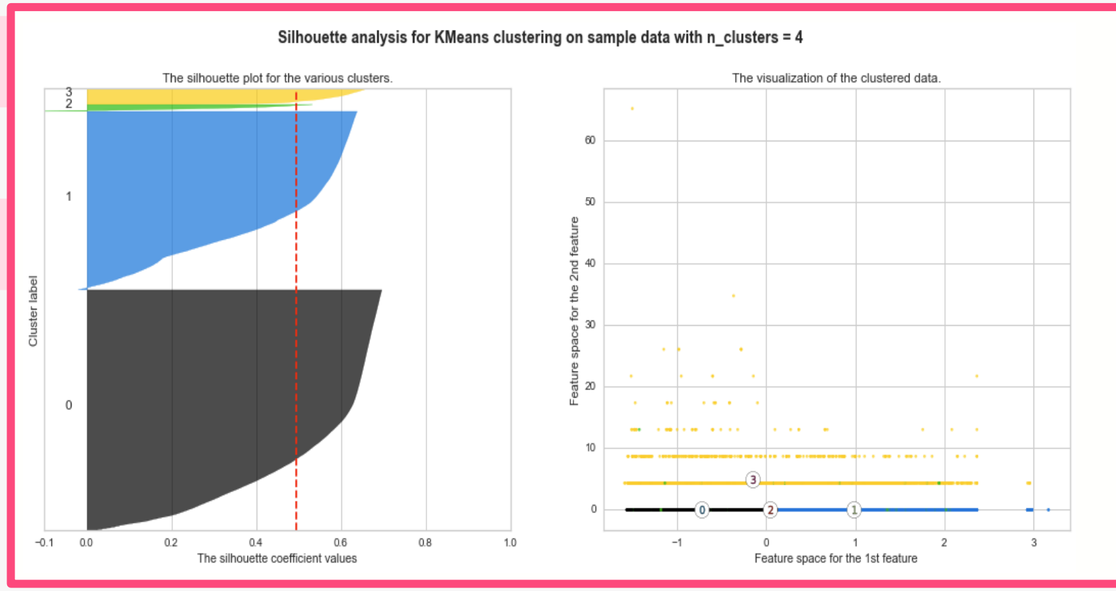
- 3ème ou 4ème cluster

Silhouette Score

- Distance Moyenne entre un point (Cluster A) et un cluster B
- Score le plus proche de 1

Silhouette analyse





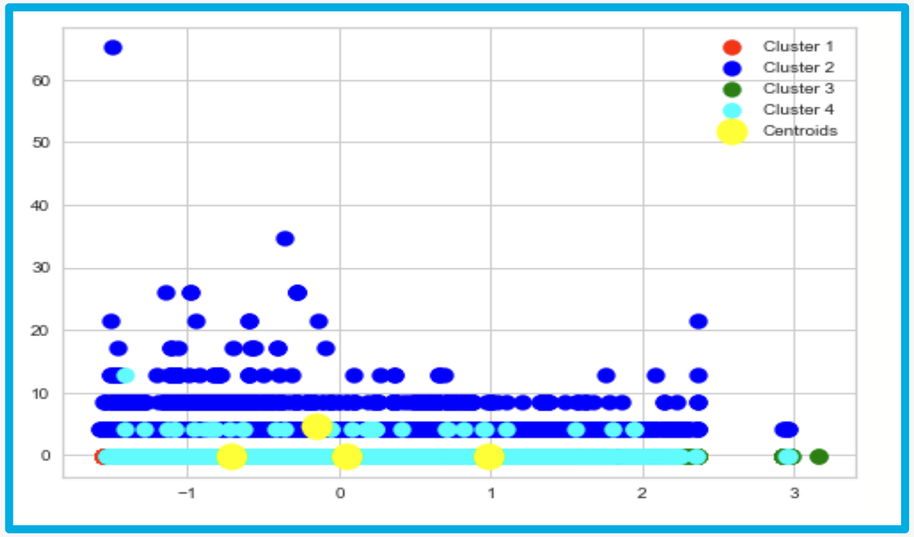
Clusters = 2

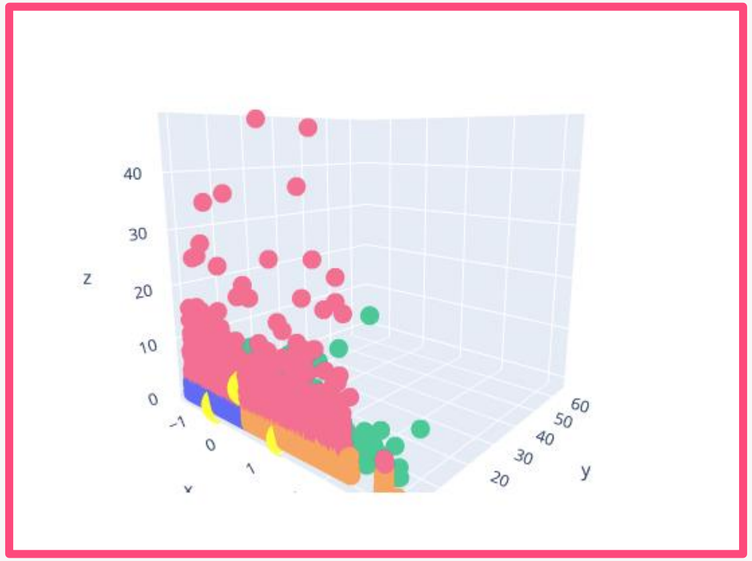
Mauvais choix

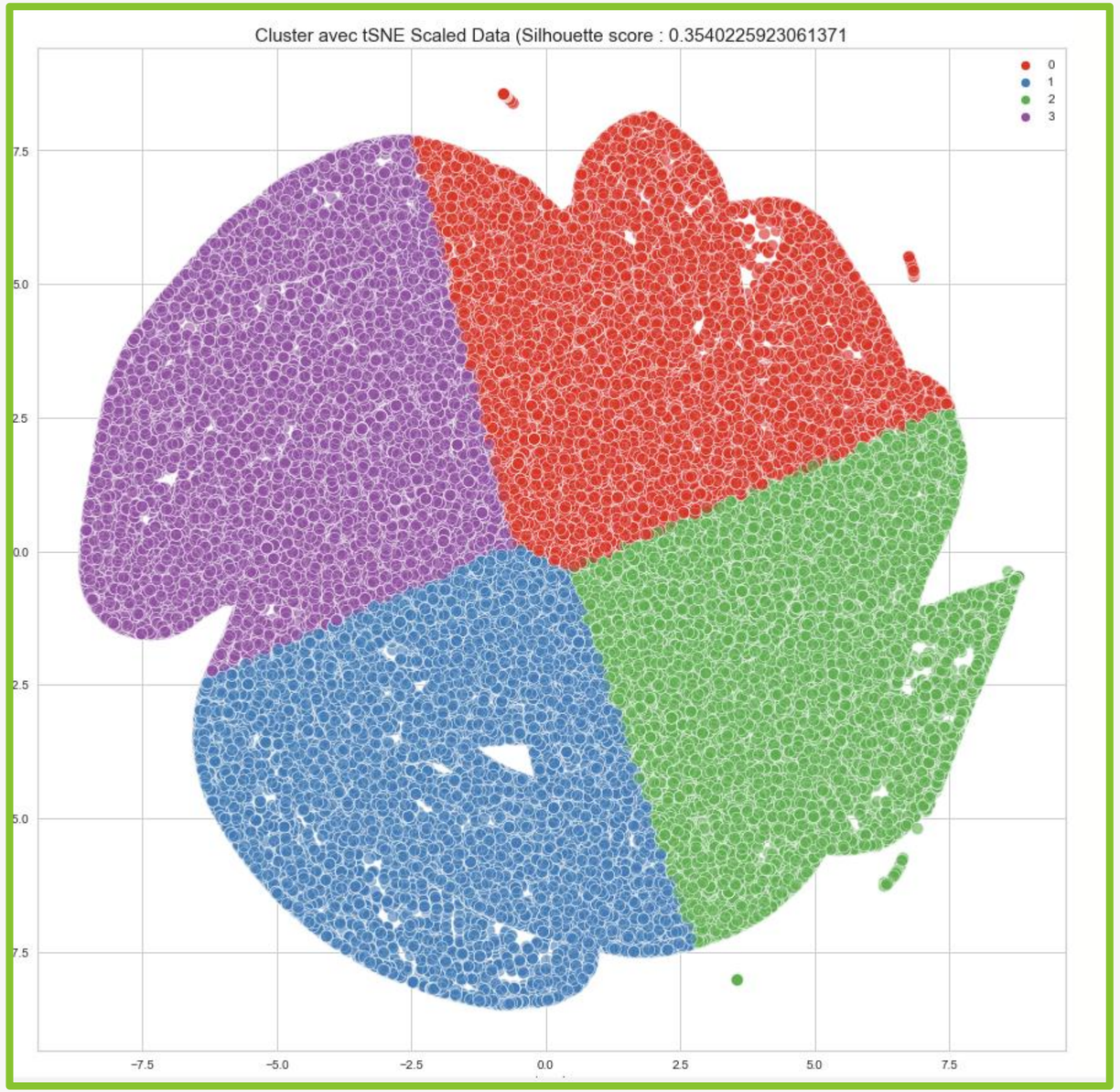
Clusters = 4

Bon choix

Visualisation







Clusters

54%

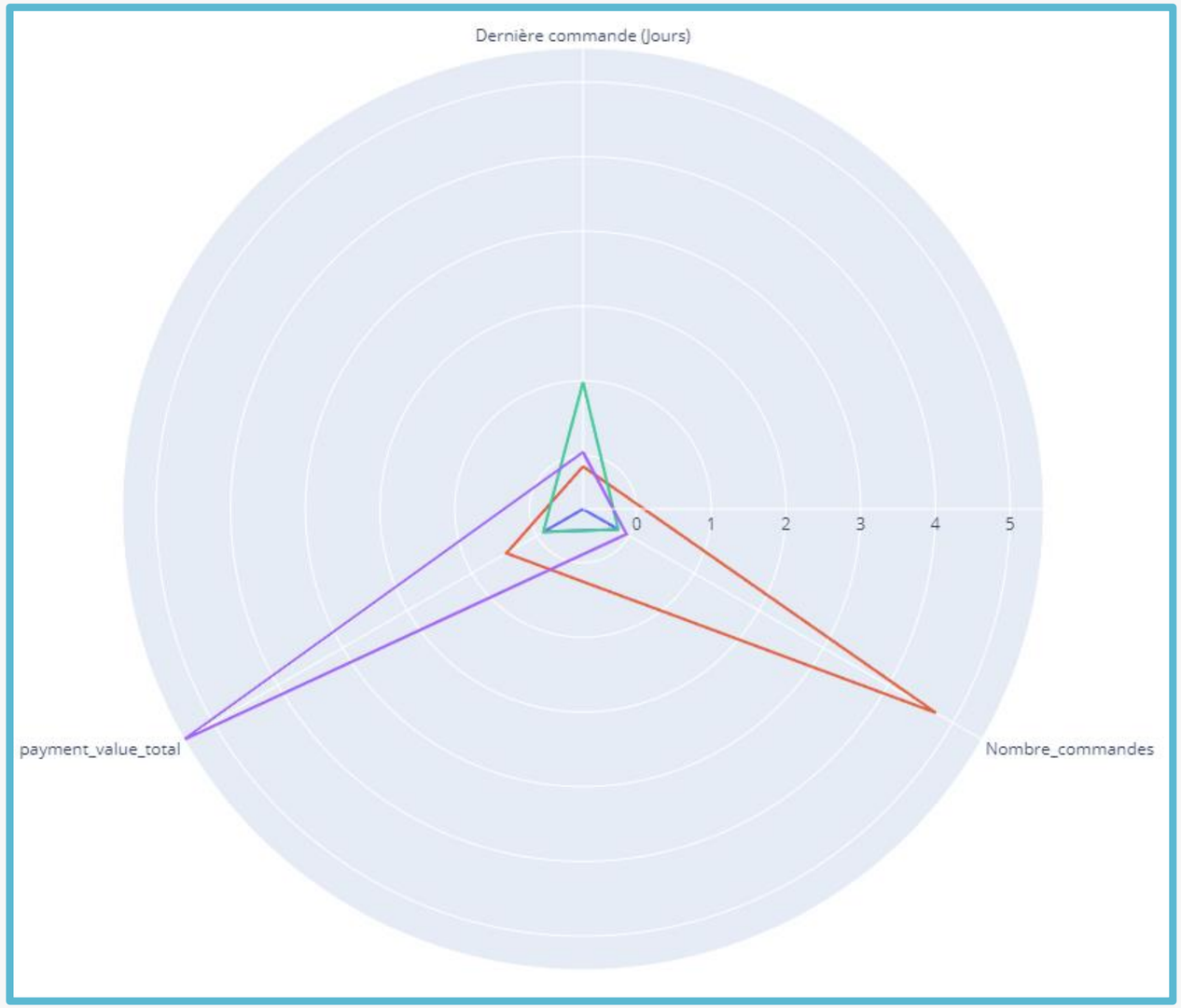
Cluster 1

Clients occasionnels

40%

Cluster 2

Anciens clients



Cluster 3

Clients occasionnels

Haut pouvoir d'achat

Achats spécifiques

1,6%

Cluster 4

Clients fidèles

Faible pouvoir d'achat

3,4%

Segmentation personnalisée

Segmentation RFM
Anciennes variables

Dernière commande

Nombre de commandes

Prix total

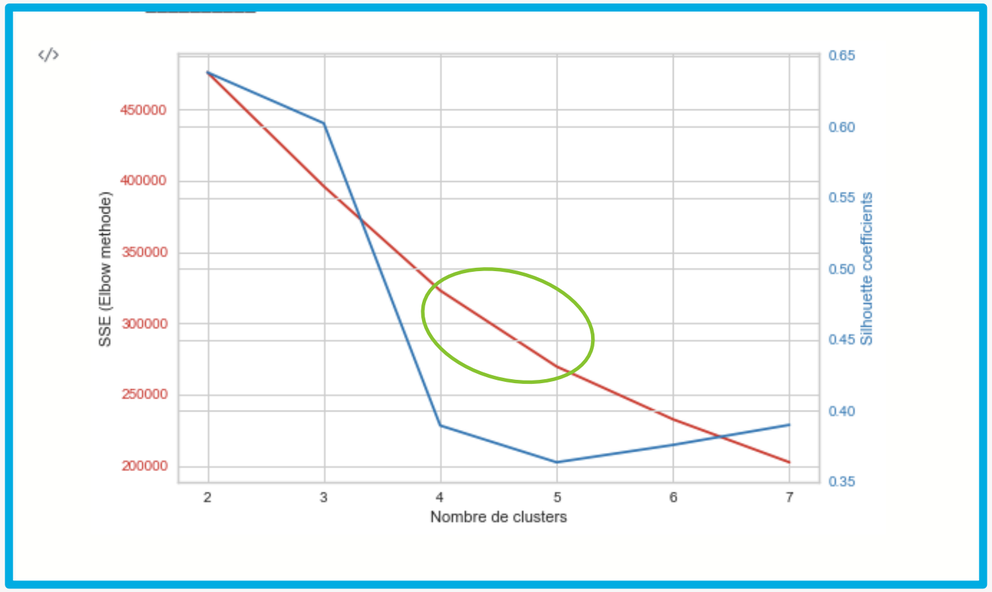
Nombre de moyens de paiement

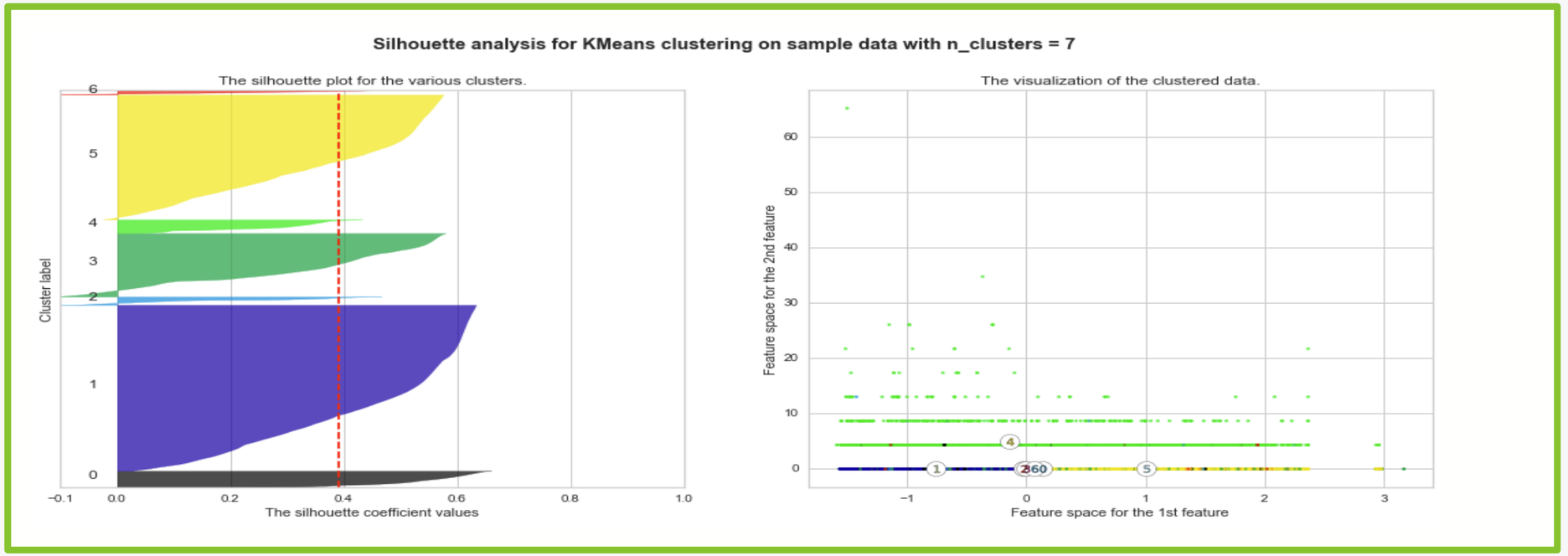
Review score moyen

Nombre de produits achetés

Segmentation personnalisée
Nouvelles variables

Visualisation





Cluster

Scaling

PCA / T-SNE

PCA

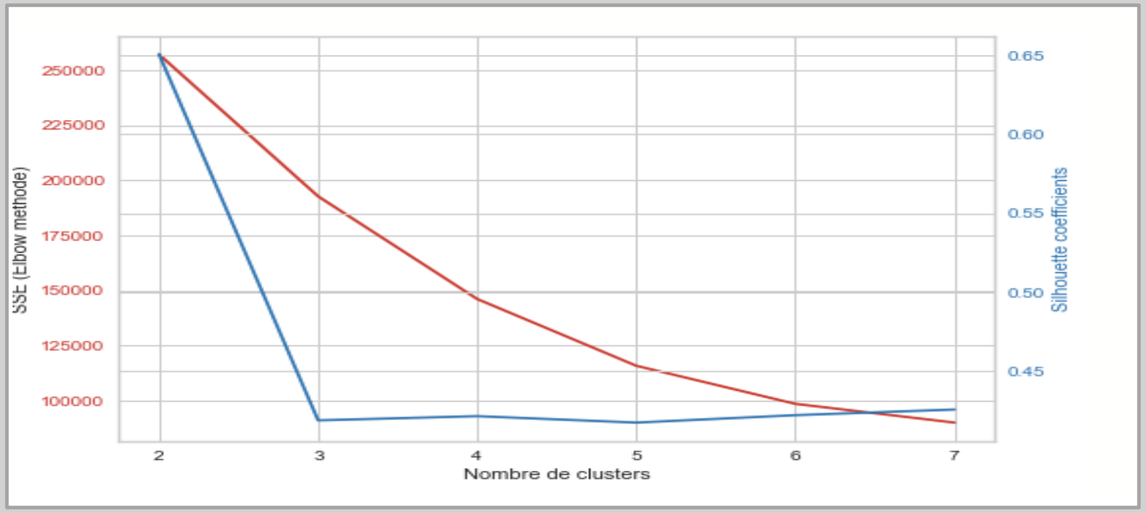
- Linéaire
- Structure globale
  Préserve la variance
- Corrélations
  Affecté par les outliers
- Peu d’hyperparamètres

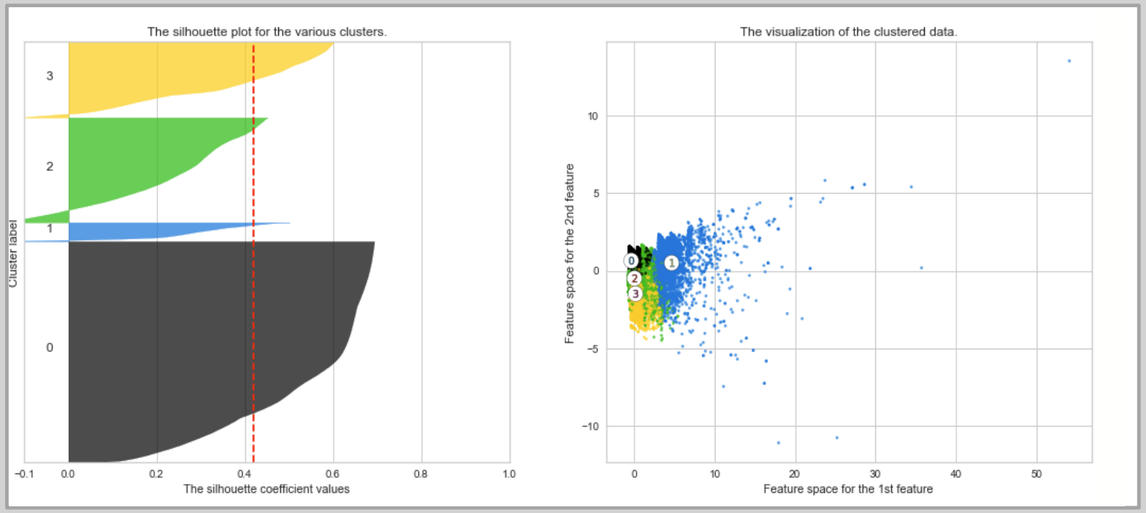
Reduction De dimension

T-SNE

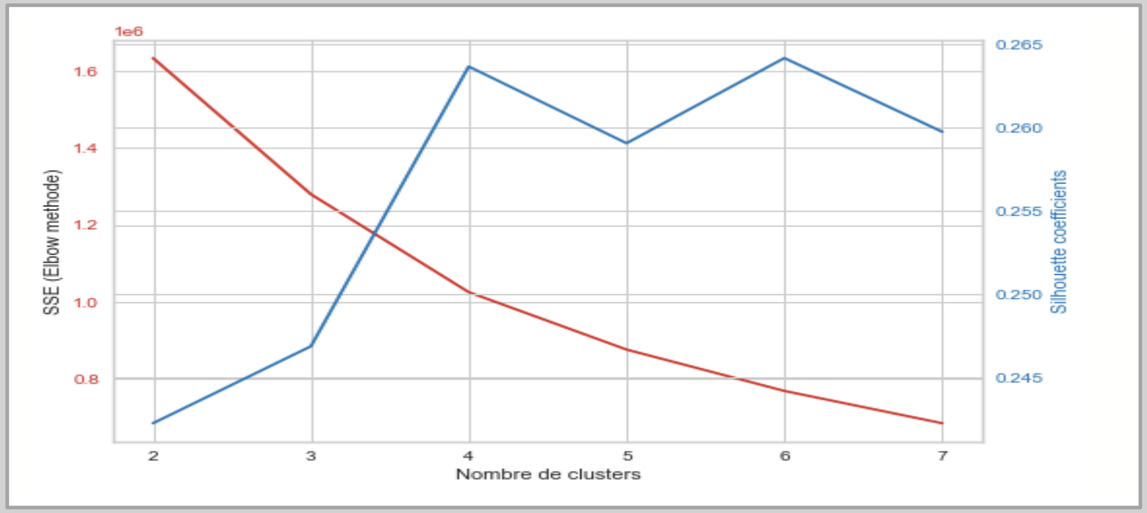
- Non-linéaire
- Structure locale
  Réduit la distance entre les points
- Meilleure visualisation
- Hyperparamètres complets

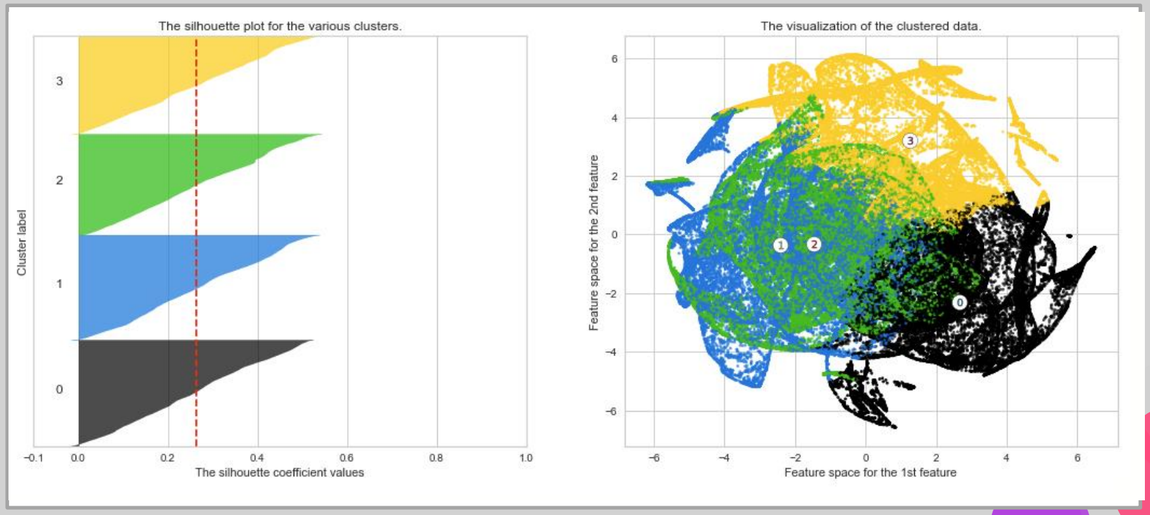
PCA


SSE (Elbow methode)
Silhouette coefficients
Nombre de clusters


The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature

T-SNE


1e6
SSE (Elbow methode)
Silhouette coefficients
Nombre de clusters


The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature

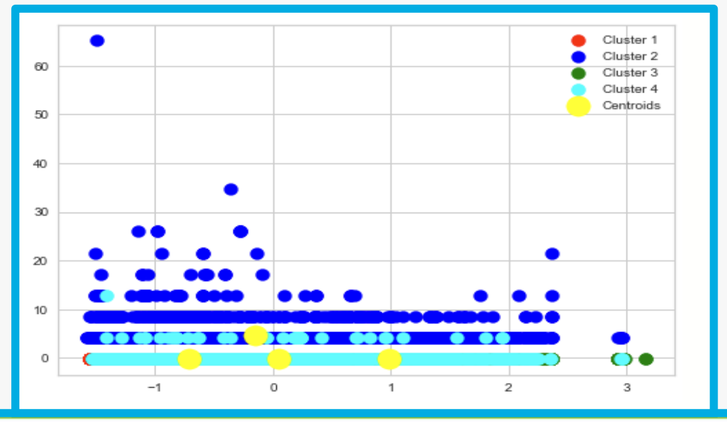
PCA


Dernière commande (Jours)
Nombre_commandes
payment_value_total
Nb de moyens de paiement
review_score_avg
Nombre_produits_achetes
−1
0
1
2
3

T-SNE

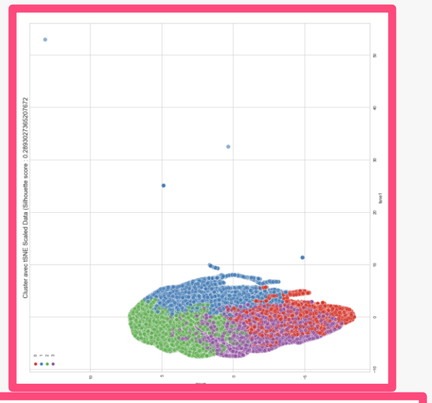

Dernière commande (Jours)
Nombre_commandes
payment_value_total
Nb de moyens de paiement
review_score_avg
Nombre_produits_achetes
−0.4
−0.2
0
0.2
0.4

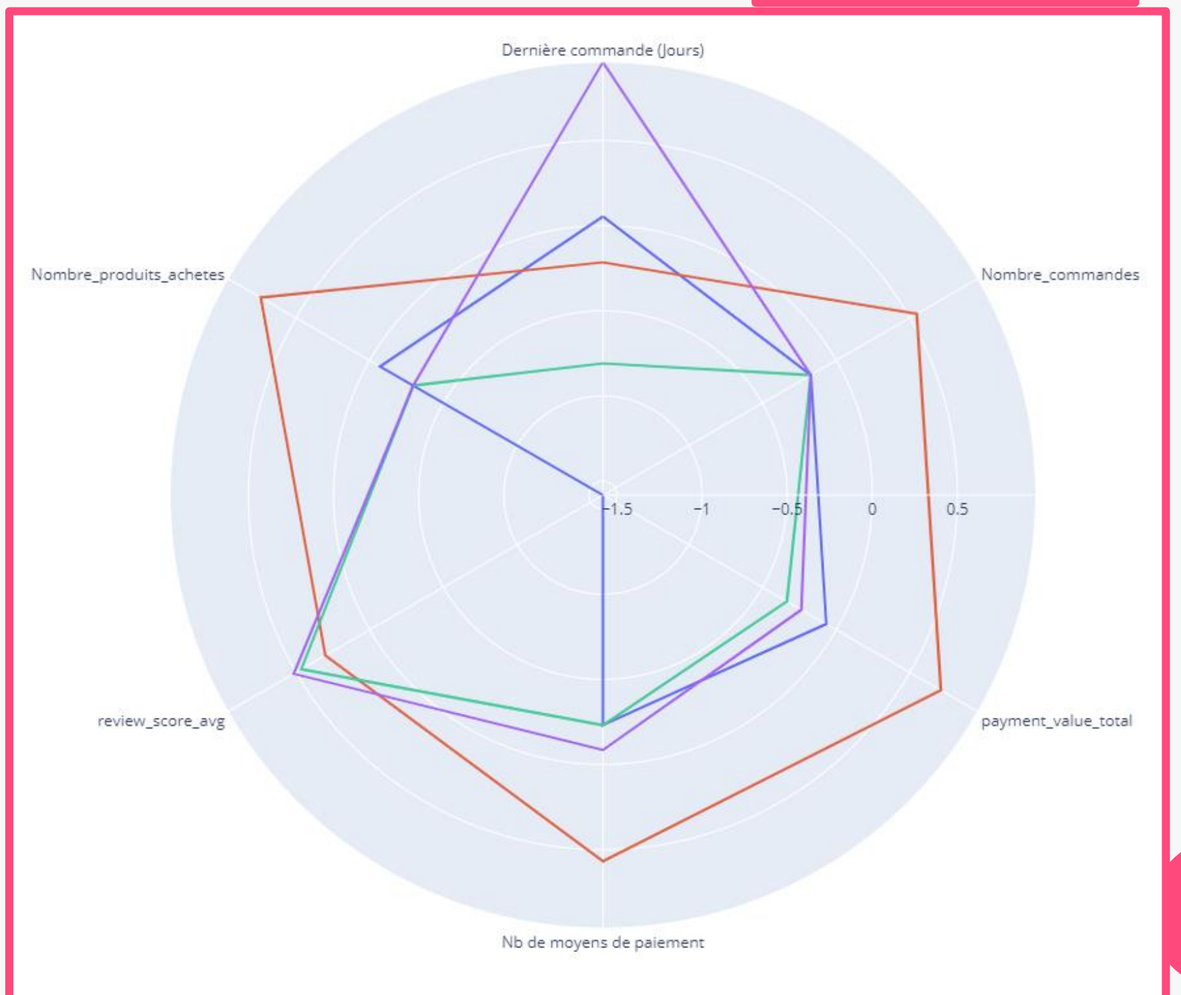
Visualisation



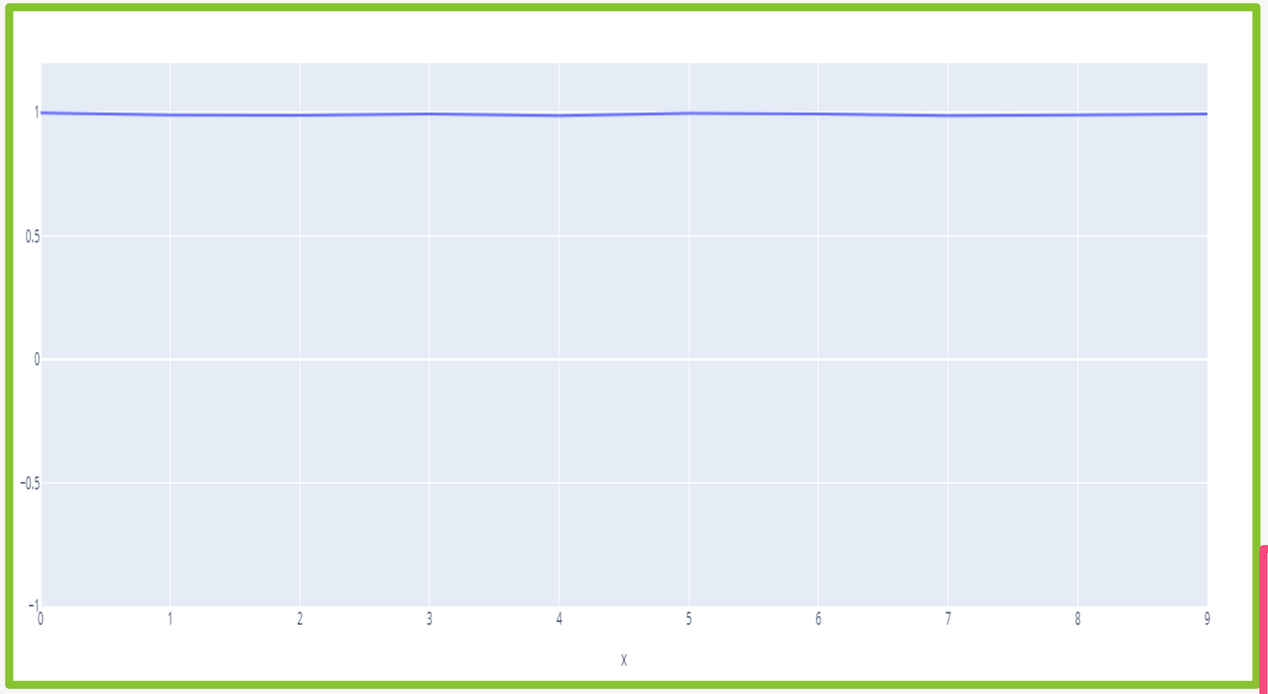
PCA

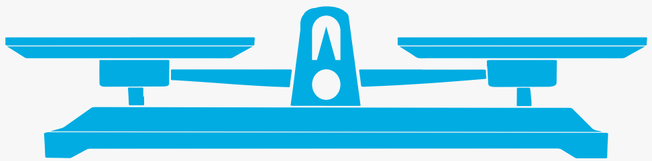
T-SNE (PCA)



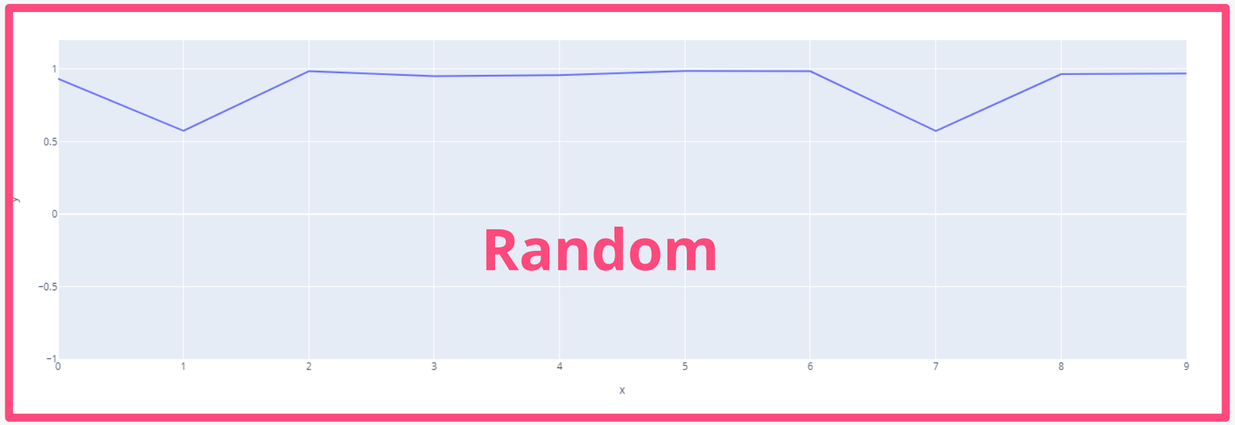


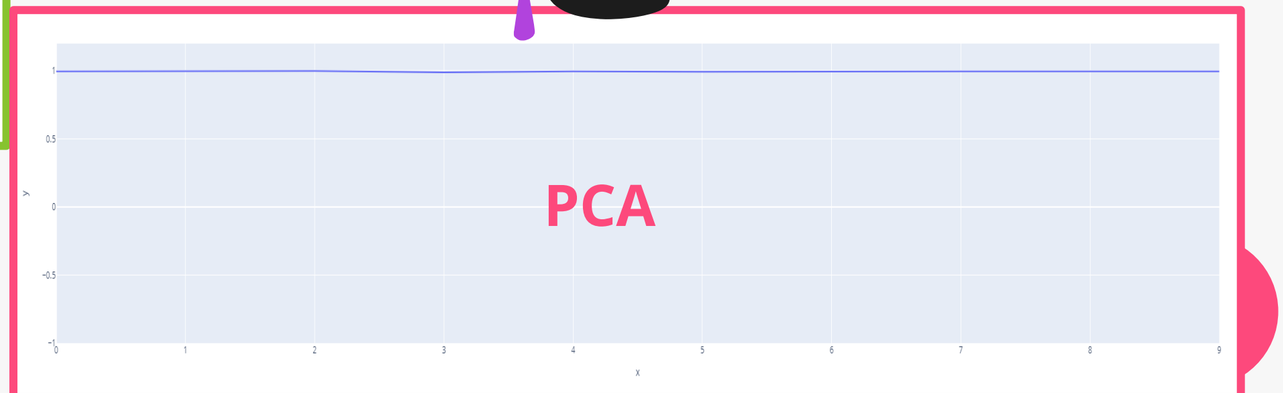
PCA


1
0.5
0
−0.5
−1
0
1
2
3
4
5
6
7
8
9
x

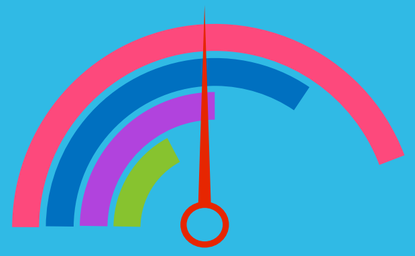


T-SNE


1
0.5
0
−0.5
−1
y
0
1
2
3
4
5
6
7
8
9
x
Random


1
0.5
0
−0.5
−1
y
0
1
2
3
4
5
6
7
8
9
x
PCA

3



Contrat de maintenance

Simulations

Score

Stabilité à partir d’un score supérieur à 0,8

Création des périodes

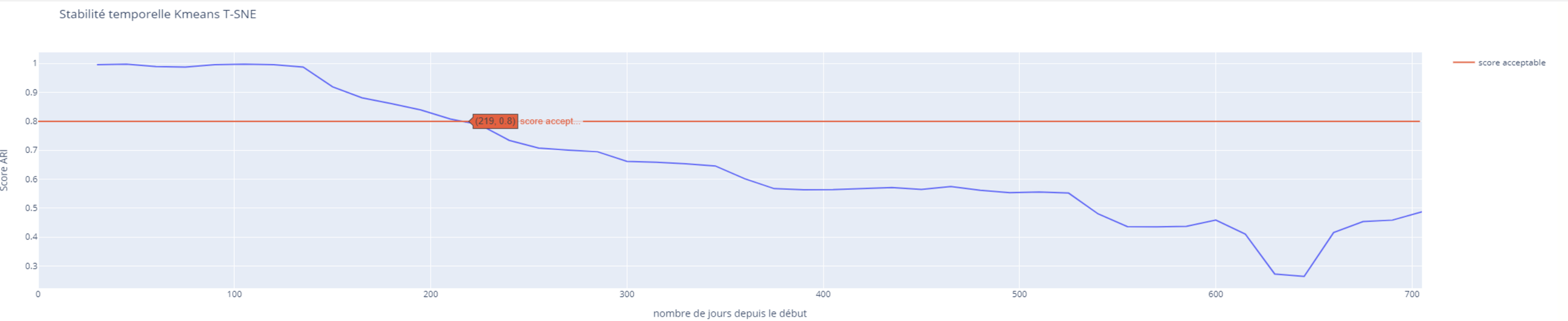
Commande la plus récente, auquel on recule de 15 jours pendant 2 ans

Kmeans

Je garde l’ancien déjà entrainé

Entrainement d’un nouveau kmeans

Stabiliité temporelle Kmeans



219 jours