olist

Segmentez des clients d'un site e-commerce

Soutenance Projet 4

Présenté par Kokou Sitsopé SEKPONA

Introduction

Je suis consultant consultant pour Olist, une entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne
L'entreprise souhaite comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.

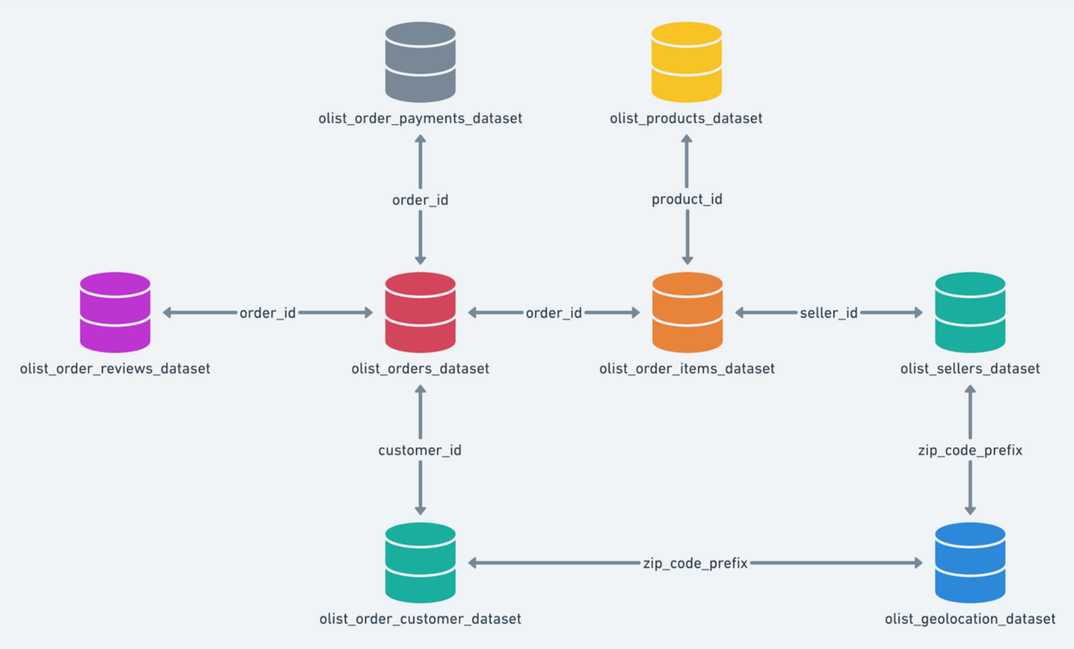
Nous allons donc par les algorithmes de clustering, essayer de segmenter ces clients en differents groupes.
La dataset comprends beaucoup d'informations non utiles, donc nous allons proceder au nettoyage puis à l'analyse exploratoire.
Ensuite utiliser les differents algorithmes de clustering sur le jeu de donnée pour retenir celui qui est le plus performant .
Et finalement lui proposer la frequence de renouvellement du modele.

1. **Nettoyage et Analyse exploratoire**

Choix du meilleur model de prediction par test de differents models

Determination de la Fréquence de renouvellement du modele

Nettoyage



C'est l'architecture de notre base de données,
Nous n'allons garder que les datasets utiles à la segmentation: Nous gardons olist_order_payments_dataset, olist_order_payments_dataset ,olist_order_reviews_dataset et le olist_orders_dataset

Constitution de la base de données

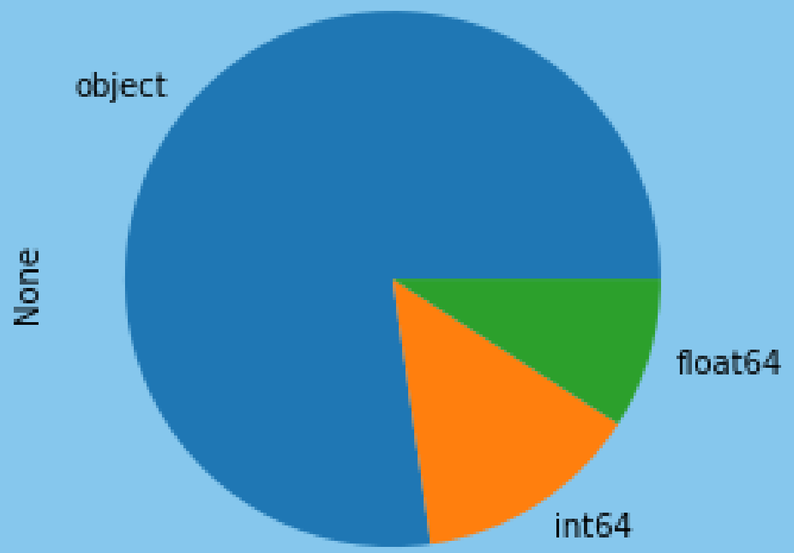
Nous utilisons le merge de pandas pour les jondre suivant les colonnes qui leur sont communes: "order_id" et customer_id

Nous obtenons une base de données de 104477 lignes et 22 colonnes

order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at
89b1e8b2acac839d17	0a8556ac6be836b46b3e89920d59291c	delivered	2018-04-25 22:01:49	2018-04-25 22:15:09
af2d9aefd1278f1dcfa0	f2c7fc58a9de810828715166c672f10a	delivered	2018-06-26 11:01:38	2018-06-26 11:18:58
6fa0d3dd708e76c1bd	25b14b69de0b6e184ae6fe2755e478f9	delivered	2017-12-12 11:19:55	2017-12-14 09:52:34
c1373bb41e913ab953	7a5d8efaaa1081f800628c30d2b0728f	delivered	2017-12-06 12:04:06	2017-12-06 12:13:20
c1373bb41e913ab953	7a5d8efaaa1081f800628c30d2b0728f	delivered	2017-12-06 12:04:06	2017-12-06 12:13:20

Nettoyage

Les differents types de données dans la dataset



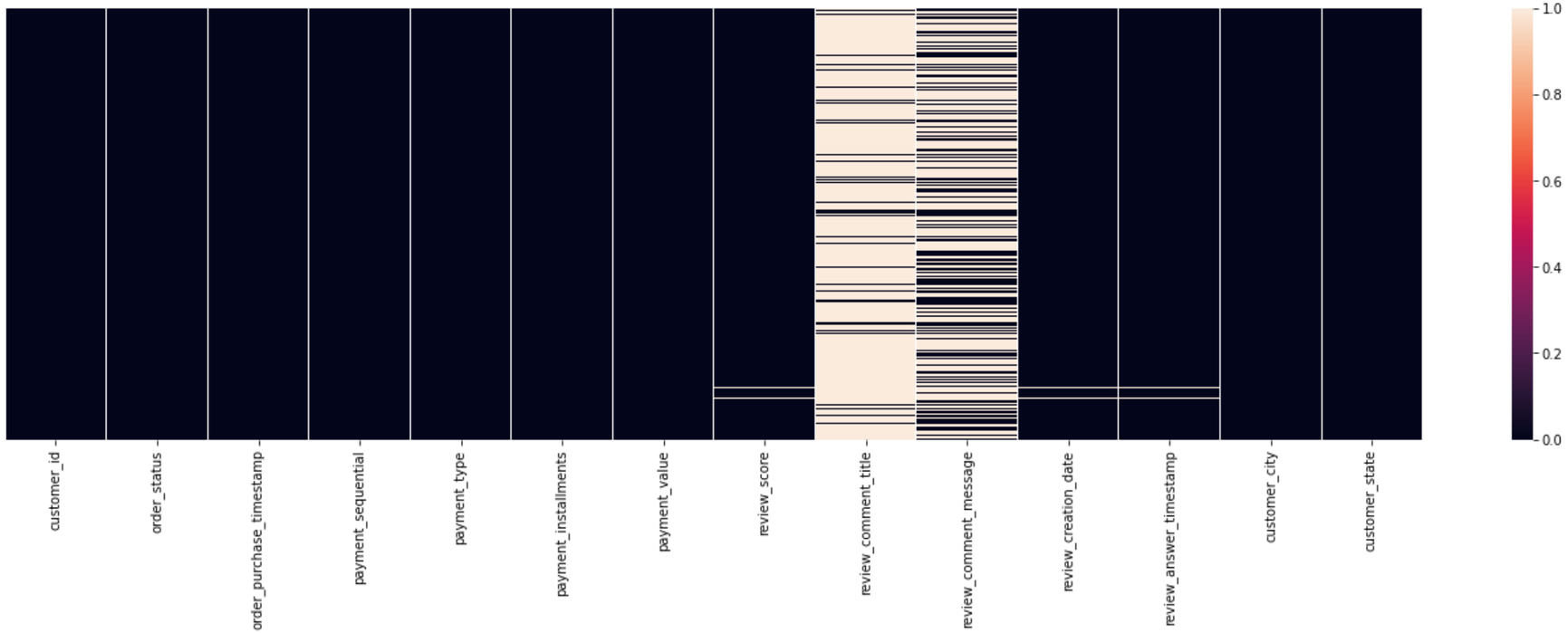
- Majoritairement des Objects
- float64
- int64

Supression des colonnes

Les colonnes 'order_id', 'review_id','order_delivered_carrier_date','customer_id', 'order_delivered_customer_date', 'order_estimated_delivery_date', 'customer_zip_code_prefix', 'order_approved_at' ne sont pas utiles pour la segmentation

Valeurs manquantes

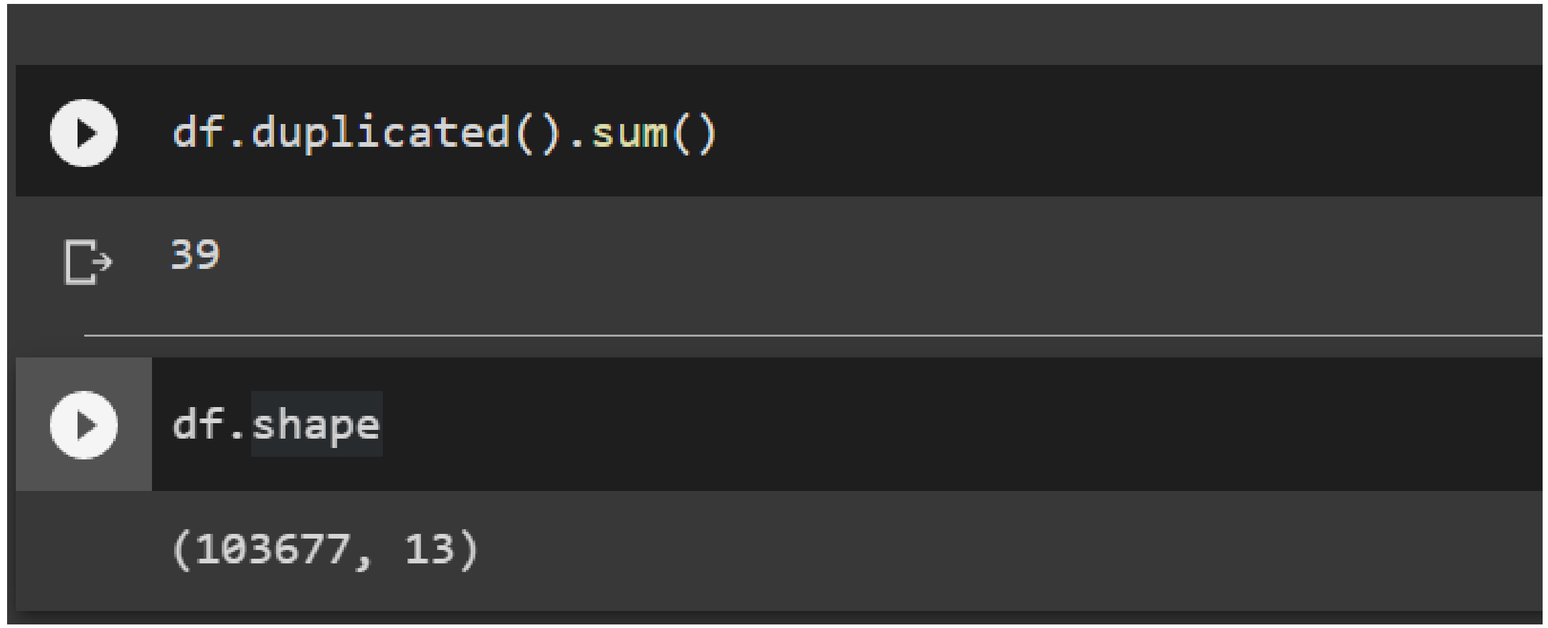
Nous allons joindre le message avec son titre et supprimer là ou il n'y a que les NaN. Cela reduira le nombre de NaN



	percent	columns
8	88.518047	review_comment_title
9	59.019689	review_comment_message
7	0.765719	review_score
10	0.765719	review_creation_date
11	0.765719	review_answer_timestamp

Doublons

Ce ne sot pas de doublons car le client peux payer plusieurs produits en meme temps.



Feature Engineering

Creation de la colonne Paid

Cette colonne est binaire; 1 si la personne a reçu la commande(delivered), 0 sinon.

Conversion de la colonne order_status en variable numérique

Cette colonne change les 8 categories qui existent sur 1 à 8 suivant l'ordre d'importance.

```
def status(x):
  if x=='delivered':
    x=8
  elif x=='invoiced':
    x=7
  elif x=='shipped':
    x=6
  elif x=='approved':
    x=5
  elif x=='processing':
    x=4
  elif x=='created':
    x=3
  elif x=='unavailable':
    x=2
  elif x=='canceled':
    x=1
  return x
```

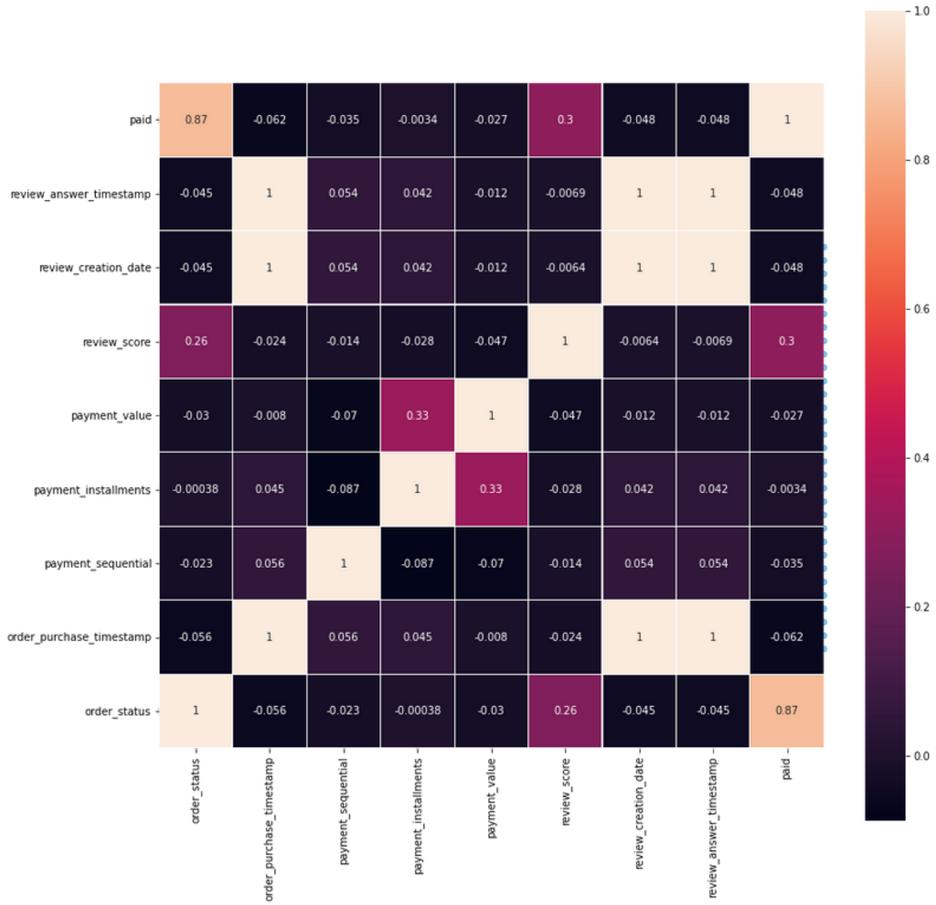
Formatage des dates

Les variables avec dates vont etre transformation nombre de jours séparant chaque date et le 31 decembre 2018, une date future à toutes les dates de la base de donnée.

	order_purchase_timestamp	review_creation_date	review_answer_timestamp
0	189	174	169
1	127	123	120
2	323	316	313
3	329	314	314
4	329	314	313
...	...	...	...
104472	237	228	225

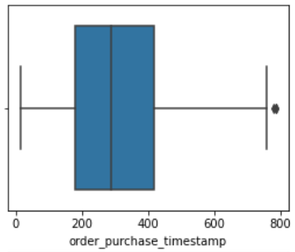
Les variables
order_purchase_timestampreview_creation_date
review_answer_timestamp sont donc transformés

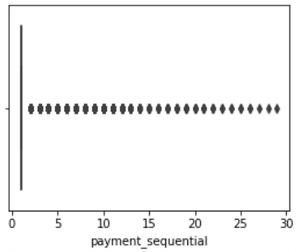
Etude de la correlation entre les variables

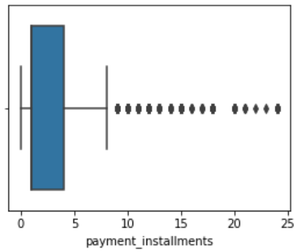


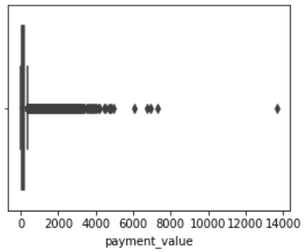
- La variable 'order_purchase_timestamp' est correlé avec (review_creating_date et review_answer_timestamp)
- La variable review_answer_timestamp est tres fortement correlé avec review_creating_date
- Donc on va supprimer ces 2 derniers

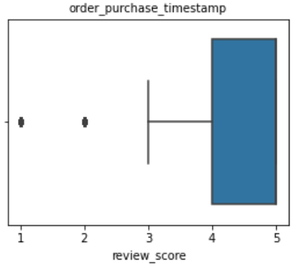
Valeurs aberrantes

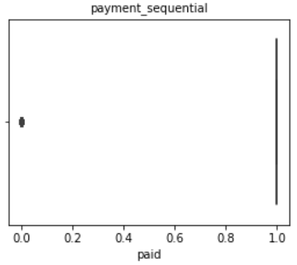


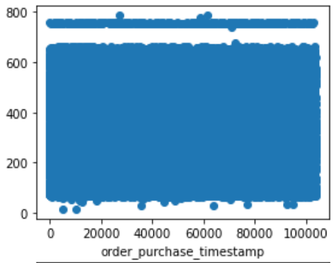


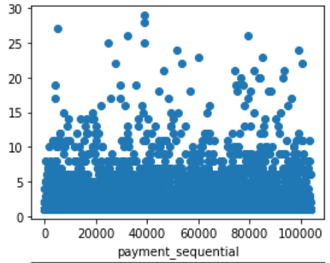


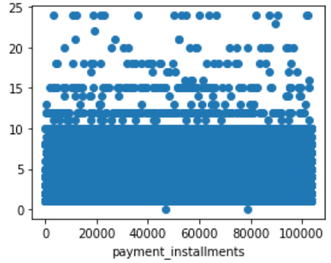


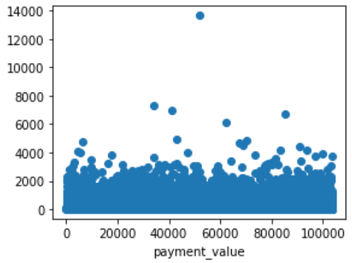


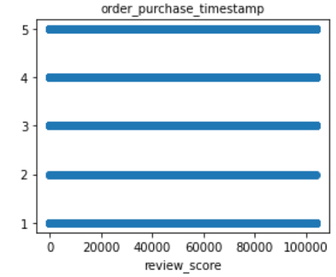


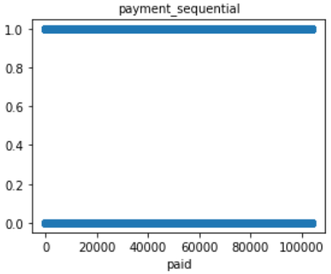












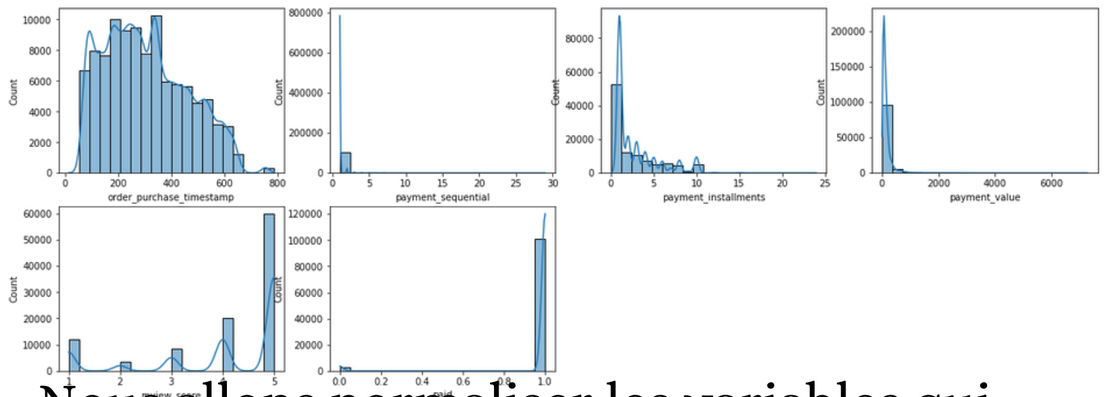
La payement_value et le payement_sequential paraissent avoir des valeurs aberrantes. Analysons les un a un

Apres analyse, ces variables n'ont pas de valeurs aberrantes, car les valeurs qui paraissent aberrantes sont possibles.

Nous allons normaliser les variables qui n'ont pas une distribution presque gausienne. Cela aidera à la précision de notre modèle.

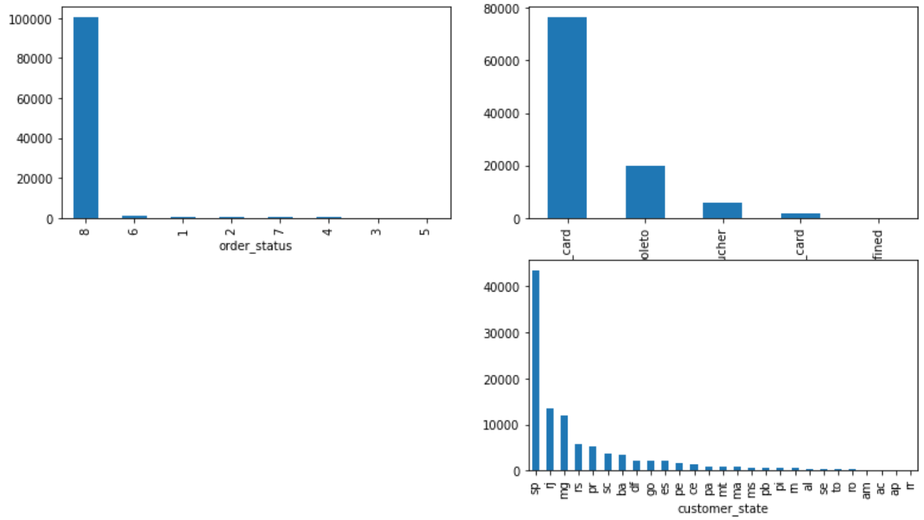
Visualisation

Variables quantitatives



Nous allons normaliser les variables qui n'ont pas une distribution presque gausienne. Cela aidera à la précision de notre modèle.

Variables qualitatives



- Plus de 40% des clients viennent de Sao Paulo(sp),
- Plus de 13% viennent du rio de janeiro
- Et pres de 12% aussi viennent de Minas Gerais Le reste est minoritaire et dispersé dans tous les provinces du brésil

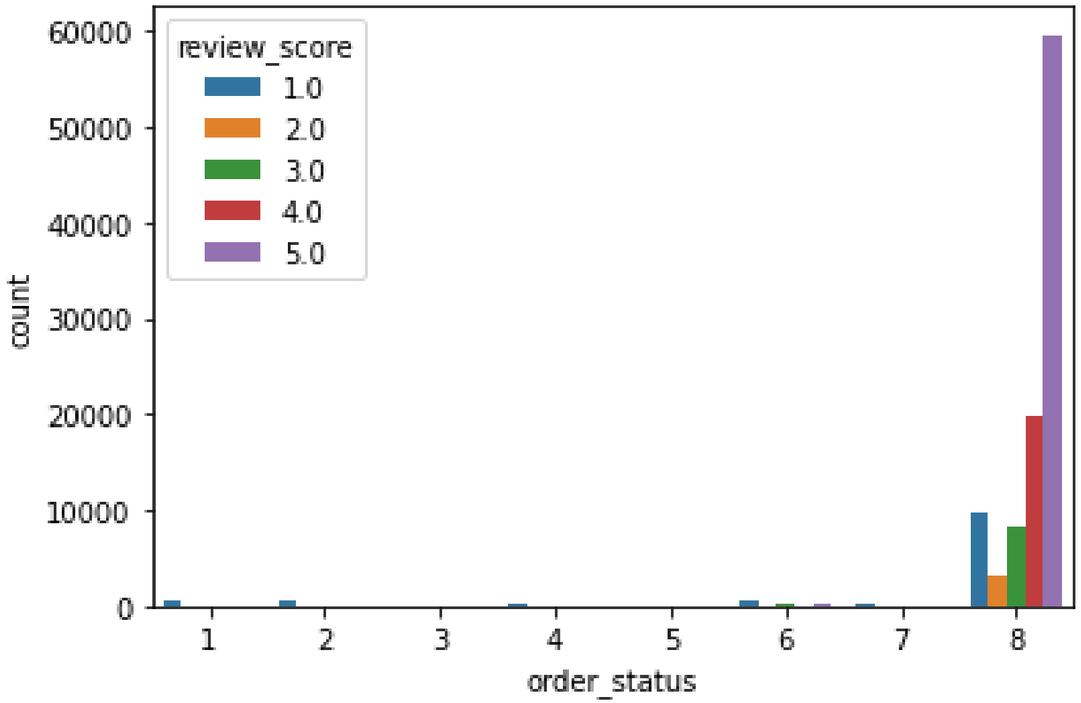
Aussi,

- Plus de 90% de commandes ont été libré et
- Moins de 1% seulement sont ennulés
- Aussi, *La majorité des produits sont payés par carte crédit

Etude de relations les variables

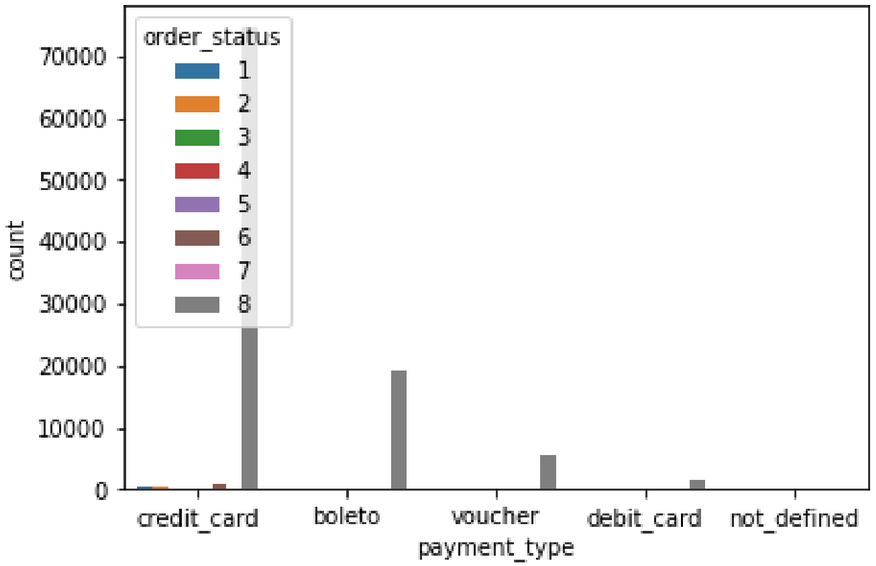
Nous allons vérifier s'il y a un rapport entre :

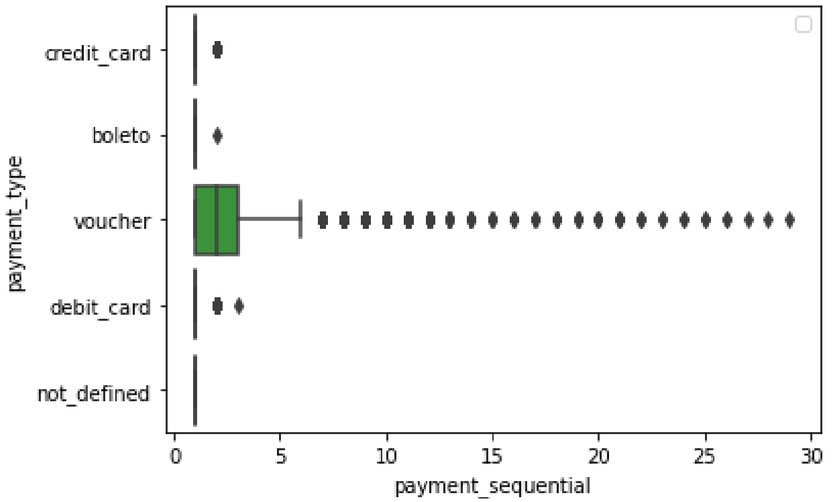
- les commandes annulés et la province ou la ville du client
- S'il y a une relation entre le nombre de fois de fois de payement sequentiel avec le mode de payement d'une part puis de l'autre part la zone de provenance.

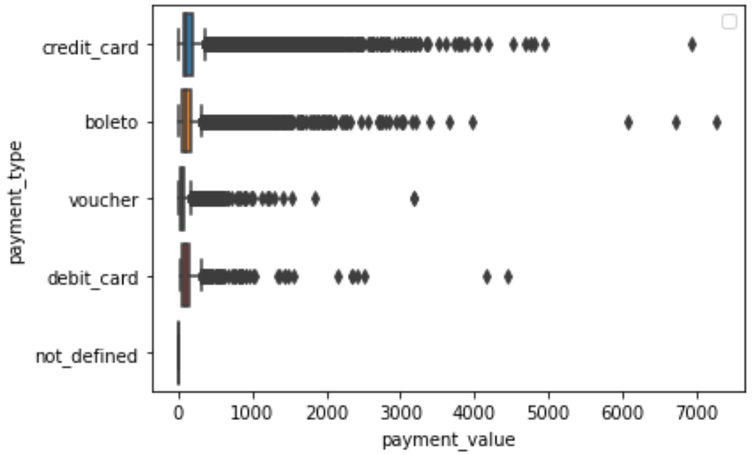


Nous observons ici que mieux les produits sont commentés, plus il sont payés. les produits à commentaires négatifs sont plus ennulés par les utilisateurs et les produits commentaire positifs ou exellents sont plus livrés. Mais il ya peux etre un autre parametre qui l"influence. Le fait que les produit à review_score de 2 sont moins livrés que les produits à review_score de 1

Relation entre certaines variables



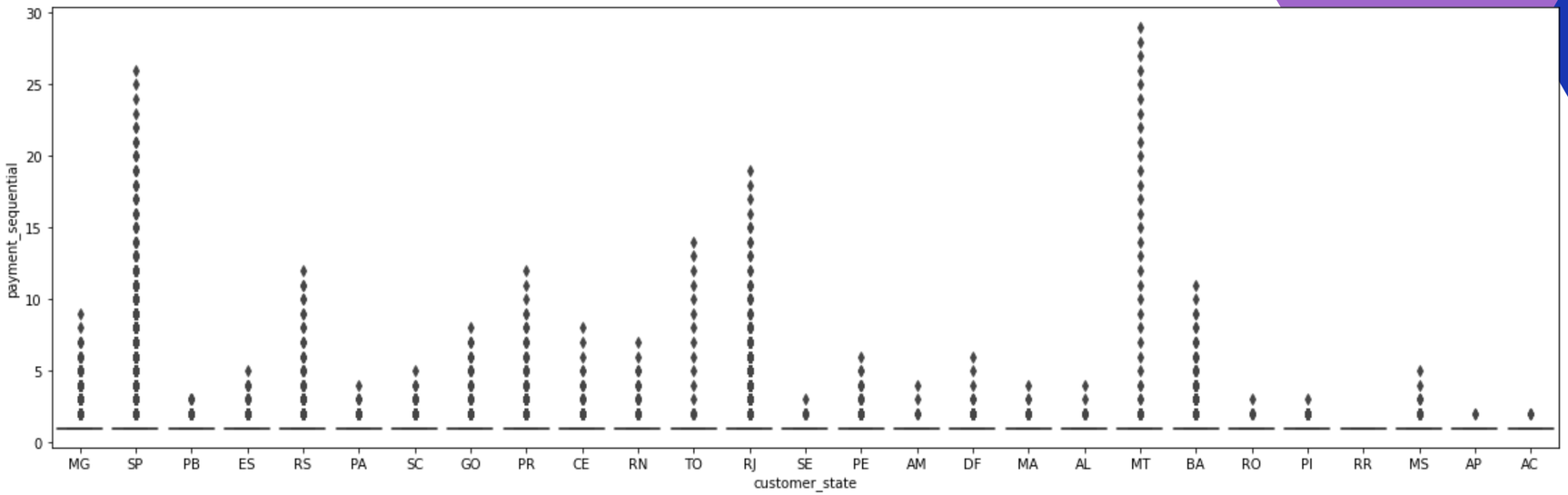




Les produits les plus payés sont livrés par carte credit

Les payements échélonnées sont plus réglés par bon de commande et donc les client payent plus le total d'un coup par carte credit, par carte debit et par boleto.

- Les plus gros payements sont réalisés par carte credit et par et par boleto.
- Les plus petits payements sont réalisés par bon de commandes



Ceux qui sont en Sp, MT et en RJ payent plus de manière échelonné Ceux qui sont en AP, AC payent plus le tout en une, deux ou trois fois au plus.

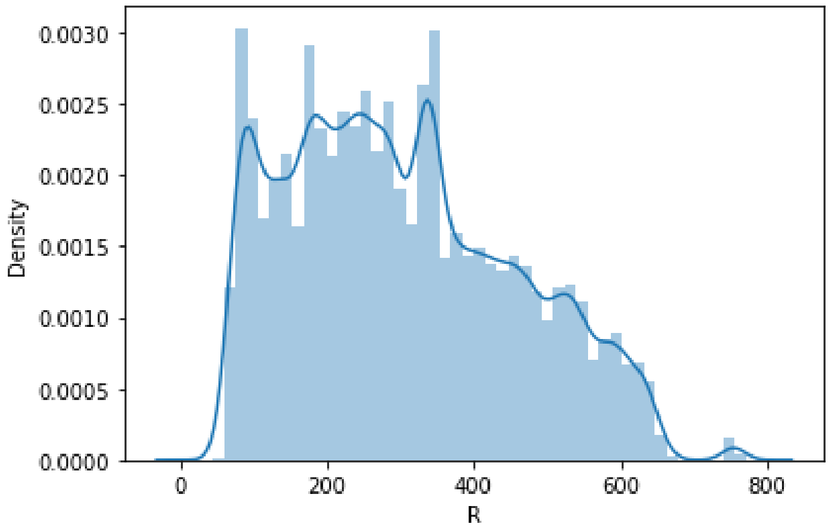
Calcul du RFM

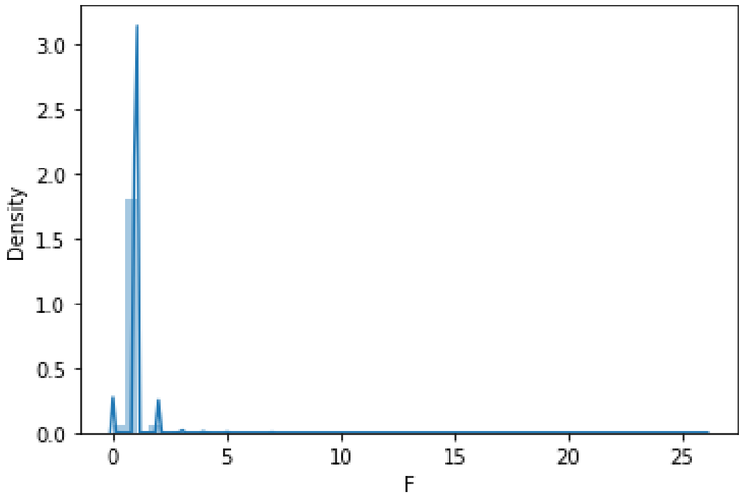
	customer_id	F	M	R	review_score
91274	ec5b2ba62e5743423868716 31fafd3fc	1	7274.88	108	1.0
76921	c6e2731c5b391845f6800c97401a43a9	1	6929.31	626	5.0
24573	3fd6777bbce08a352fddd04e4a7cc8f6	1	6726.66	525	5.0
2051	05455dfa7cd02f13d132aa7a6a9729c6	1	6081.54	341	1.0
86244	df55c14d1476a9a3467f131269c2477f	1	4950.34	578	5.0

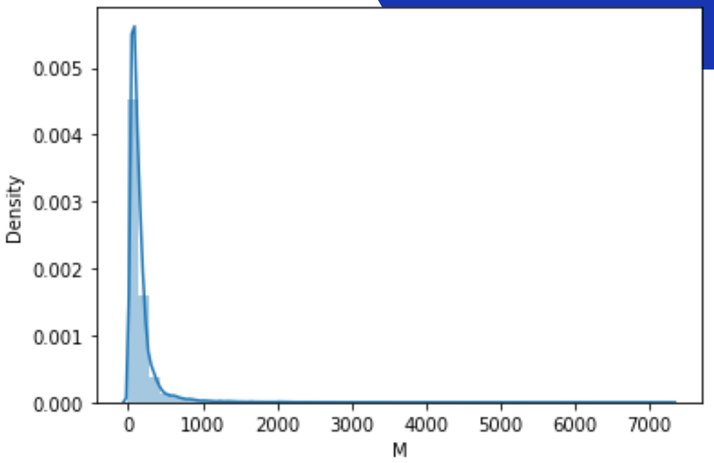
Nous transformons les colonnes quantitatives du jeu de données en Trois colonnes principales:

- Recence qui est le nombre de jours séparant le dernier achat et le 31dec 2018,
- Frequence qui est la frequence d'achat
- Montant qui est le montant total des commandes réalisé par l'utilisateur.

Visualisation du RFM







Laplupart des utilisateurs ont éfectués leurs achat il y a de cela environ 1 an. L'achat le plus recent sur le site date de 14 jours.

En moyenne un utilisateur n'a effectué qu'un payement. La plus grande quantité de produit payés jusqu"a ce jour par individu est de 26.

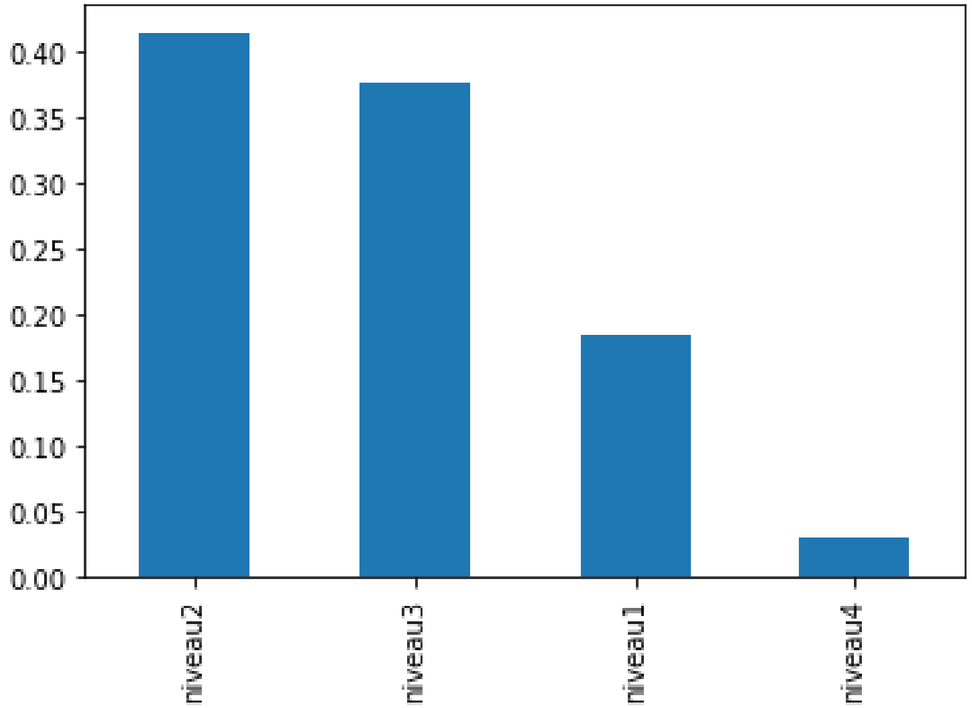
En moyenne les utilisteurs ont payés au total 161.3 euros

Calcul du score RFM pour les segmenter

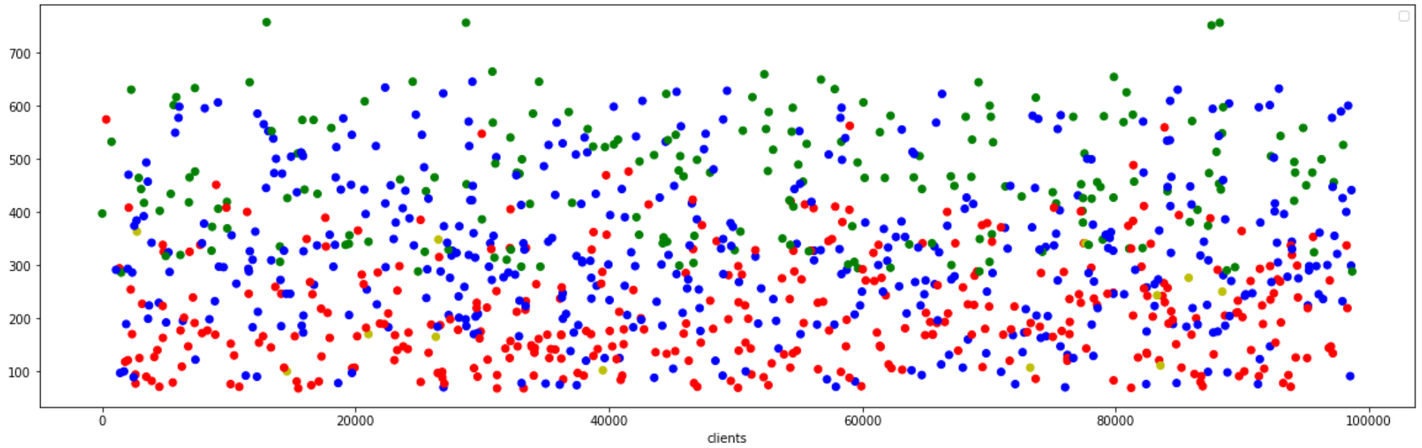
Nous avons au total 4 niveaux de segmention pour les clients selon le score RFM

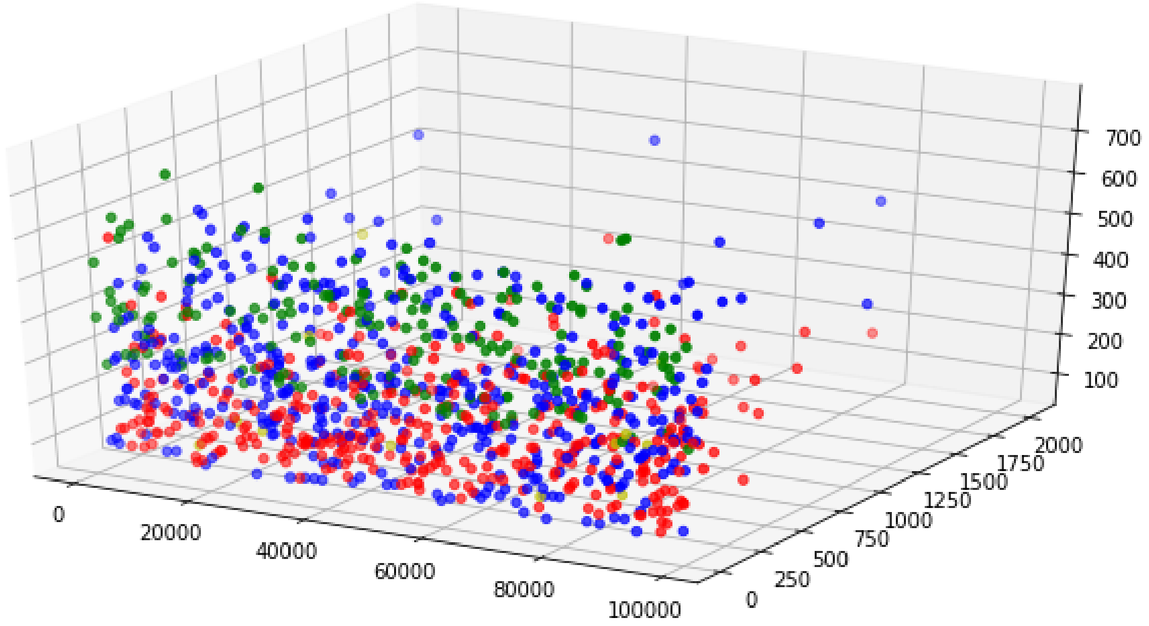
- Niveau 4: Ceux qui ont des scores Supérieur à 10
- Niveau 3: Score supérieur à 7
- Niveau 2: Score supérieur à 4
- Niveau 1: Le reste des clients

	customer_id	F	M	R	review_score	Recence	Frequence	Montant	score	niveau
0	00012a2ce6f8dcda20d059ce98491703	1	114.74	351	1.0	2	1	3	6	niveau2
1	000161a058600d5901f007fab4c27140	1	67.41	472	4.0	1	1	2	4	niveau1
2	0001fd6190edaaf884bcaf3d49edf079	1	195.42	610	5.0	1	1	4	6	niveau2
3	0002414f95344307404f0ace7a26f1d5	1	179.35	441	5.0	1	1	4	6	niveau2
4	000379cdec625522490c315e70c7a9fb	1	107.01	212	4.0	3	1	3	7	niveau3
5	0004164d20a9e969af783496f3408652	1	71.80	567	1.0	1	1	2	4	niveau1
6	000419c5494106c306a97b5635748086	1	49.40	243	1.0	3	1	1	5	niveau2



Segmentation Score RFM: Visualisation





- Niveau 1: 'y'
- Nivau 2: "r",
- Niveau 3: 'b',
- Niveau 4: "g"

Les Rouges et les bleux sont beaucoup plus nombreux:
Les jaunes sont rares: Ce sont les clients importants pour l'entreprise

Approche Non supervisé

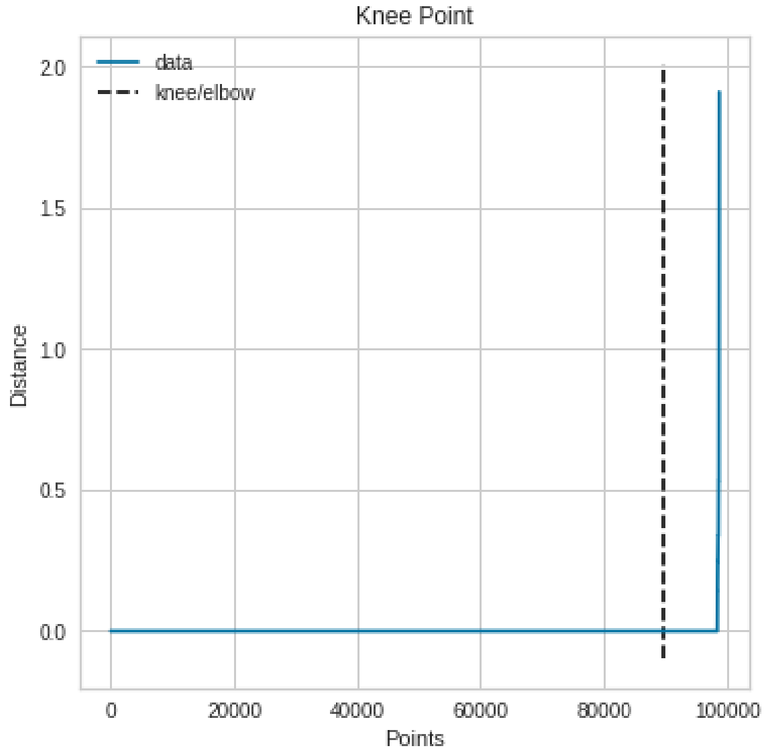
1. **Preprocessing**

Nous passons par la transformation log puis par StandardScaler pour standardiser les donnes

2. Test de differents models et choix du meilleur

3. Recherche de la fréquence de renouvellement du modèle.

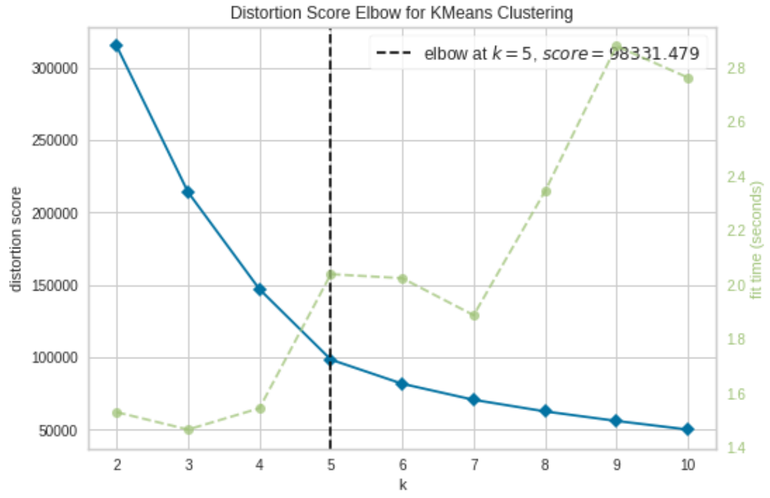
Test de differents models

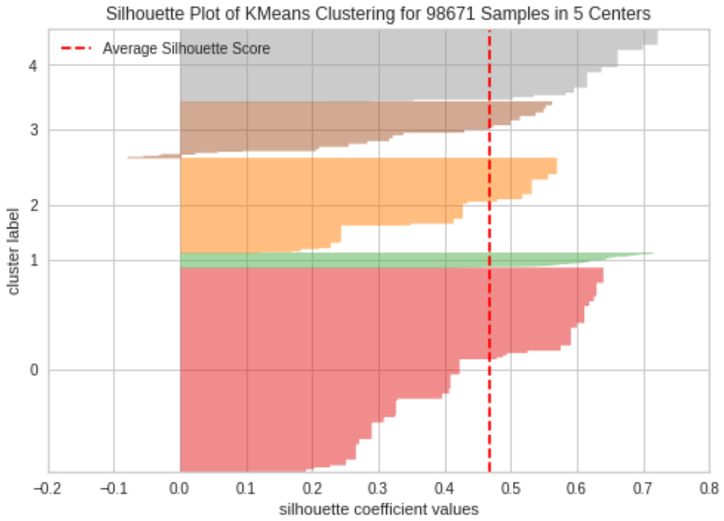


1.DBSCAN

Avec GridSearchCV, nous obtenons les bon hyperparamètres et nous utilisons le model avec les meilleurs parametres pour calculer la silhouette score

silhouette_score= 0.481



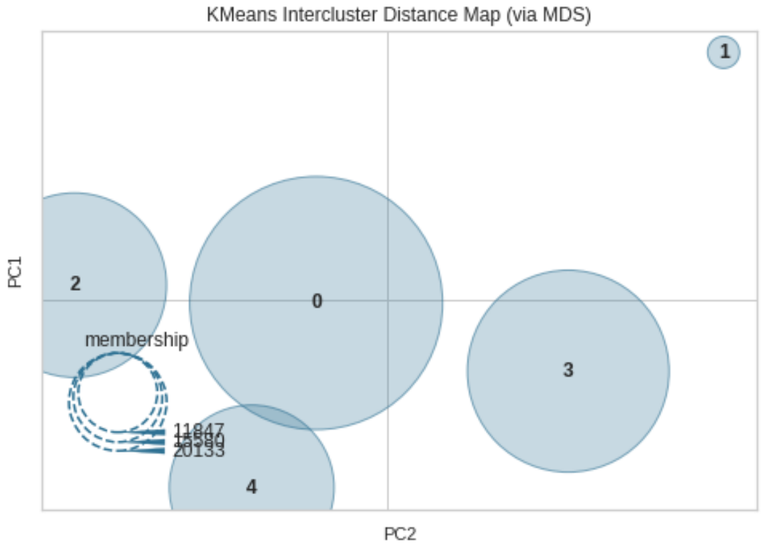


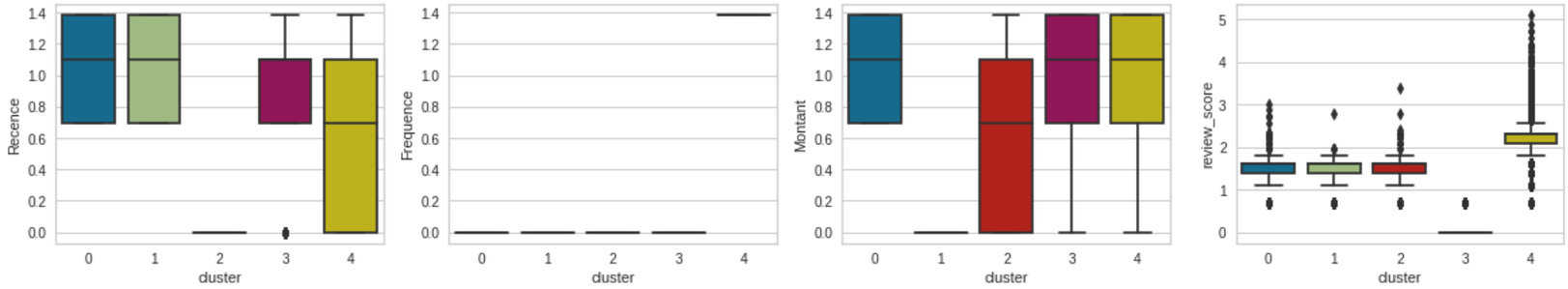
1. KMeans

K= 5 cluster avec la methode des Elbow
Nous obtenons un silhouette score de
Silhouette Score: 0.468

Nous allons garder le KMeans
Car le DBSCAN est très lent à etre entrainé

Visualisation des clusters



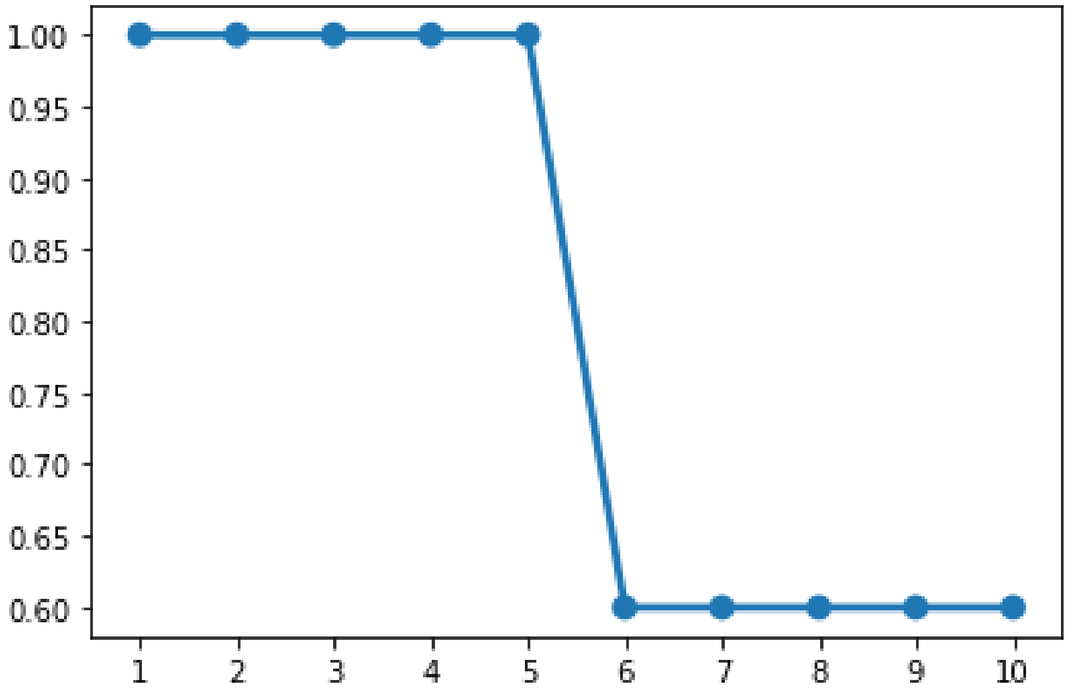


Le premier graphe représente la distance entre les différents clusters.
Le deuxieme boxplot est 4 figures qui representent chaque variable en fonction des differents clusters.

Nous avons au total 5 clusters par ordre d'importance:

- **Cluster 4**
- **Cluster 0**
- **Cluster 3**
- **Cluster 2**
- **Cluster 1**

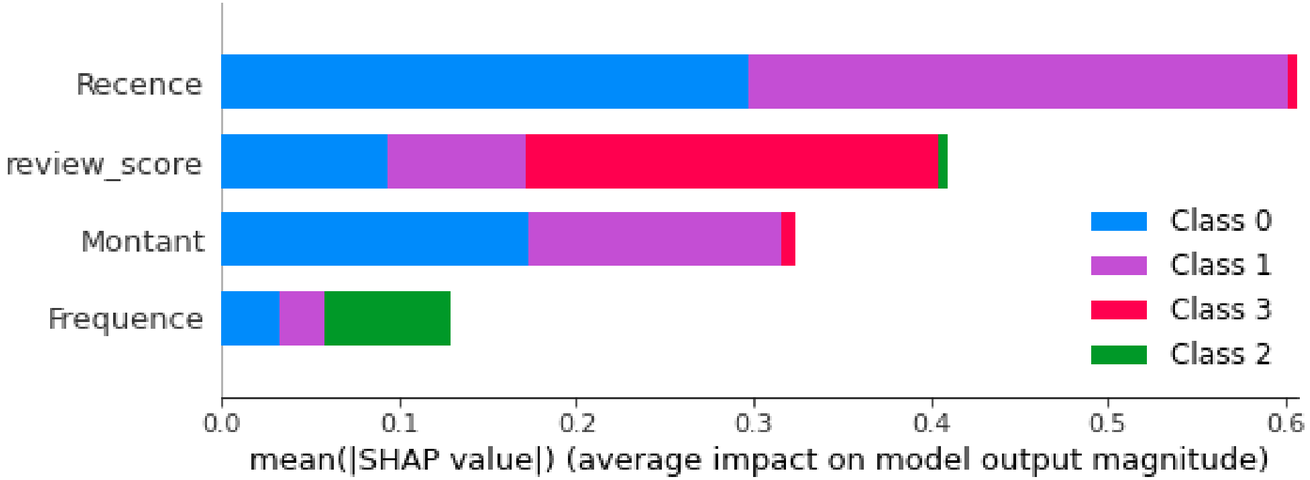
Frequence d'evaluation du modele



Nous entrainons le premier model sur les données du premier mois
et avec Ari score, nous évaluons la différence entre la précision de ce model sur les données au mois i et la précision d'un modèle sur ce même mois mais entrainé sur ce mois

Le modele va donc etre renouvelle tous les 4 mois puisque la précision
change entre le 4eme et le 5eme mois

Interpretation du model



L'importance globale de chaque variable dans le model

Importance dans de chaque variable dans la segmentation d'un client

Merci !

N'hésitez si vous avez des questions.