Formation de Data Science - Openclassrooms

Formation Ouverte et à Distance – FOAD par Pôle Emploi

Solutions 100% à distance

Projet 5 : Segmentez des clients d'un site e-commerce

Étudiant : Maria Daniela Barrios

Mentor : Dan Slama

Avril, 2022

Contexte du problème

- ➔ Olist souhaite que nous fournissons à ses équipes d'e-commerce des informations sur la segmentation des clients qu'elles pourront utiliser pour leurs campagnes de communication quotidiennes.

- ➔ Notre objectif est de comprendre les différents types d'utilisateurs à travers leur comportement et leurs données personnelles. Nous devons fournir à l'équipe marketing une **description exploitable de la segmentation client** et de la logique sous-jacente pour une utilisation optimale, ainsi qu'une **proposition de contrat de maintenance basée sur une analyse de la stabilité du segment dans le temps**.

- ➔ Nous allons donc utiliser des méthodes non supervisées d'apprentissage automatique pour regrouper les clients ayant des profils similaires.

- ➔ Olist fournit une base de données anonymisée contenant des informations sur **l'historique des commandes, les produits achetés, les commentaires de satisfaction et la localisation des clients depuis janvier 2017**.

- ➔ Les données peuvent être téléchargées à : https://www.kaggle.com/olistbr/brazilian-ecommerce

Stratégie pour réaliser la mission

➔ **Phase pré-exploratoire : Analyse générale et découverte des fichiers**
- Décrire les informations contenues dans l'ensemble de données : Nombre de lignes et de colonnes
- Sélection des variables à utiliser dans notre analyse

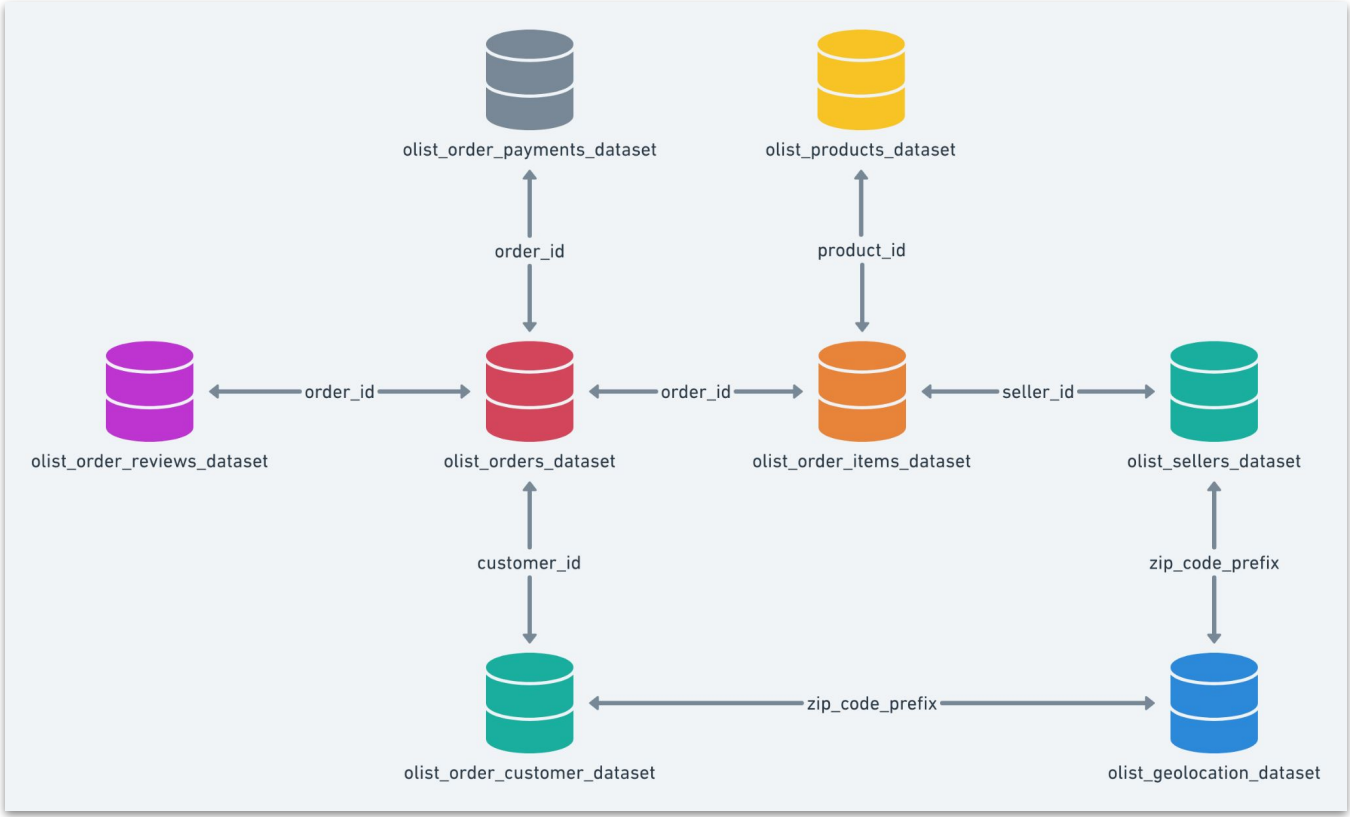
➔ **Analyse exploratoire et nettoyage des données**
- Exploration des valeurs manquantes
- Exploration et traitement des valeurs dupliquée
- Jointure de tables par clés
- Création de nouvelles variables - feature engineering

➔ **Utilisation de la méthode RFM comme référence de segmentation**

➔ **Segmentation des clients à l'aide de techniques de Machine Learning non-supervisés**
- Test des algorithmes de Machine Learning non-supervisés : K-means, DBSCAN
- Réduction de la dimensionnalité par PCA, T-SNE

➔ **Analyse de la maintenance de segmentation - stabilité du segments clients dans le temps**

Phase pré-exploratoire : Analyse générale et découverte des dossiers



- ➔ Le jeu de données est composé de 8 fichiers intégrés chacun dans un tableau
- ➔ Nous pouvons interconnecter les tables en utilisant les clés indiquées
- ➔ Par exemple, **customer_id** est la clé de l'ensemble de données des commandes. Chaque commande a un numéro de client unique

Phase pré-exploratoire : Analyse générale et découverte des dossiers

olist_customers_dataset :

- **99441** lignes et **5** colonnes

olist_geolocation_dataset :

- **1000163** lignes et **5** colonnes

olist_order_items_dataset :

- **112650** lignes et **7** colonnes

olist_order_payments_dataset :

- **103886** lignes et **5** colonnes

olist_order_reviews_dataset :

- **99224** lignes et **7** colonnes

olist_orders_dataset :

- **99441** lignes et **8** colonnes

olist_products_dataset :

- **32951** lignes et **9** colonnes

olist_sellers_dataset :

- **3095** lignes et **4** colonnes

product_category_name_translation :

- **71** lignes et **2** colonnes

Phase pré-exploratoire : Analyse générale et découverte des dossiers

The olist_customers_dataset file contains: 99441 rows and 5 columns

Missing or nan values of:

customer_id	0
customer_unique_id	0
customer_zip_code_prefix	0
customer_city	0
customer_state	0

The olist_order_payments_dataset file contains: 103886 rows and 5 columns

Missing or nan values of:

order_id	0
payment_sequential	0
payment_type	0
payment_installments	0
payment_value	0

dtype: int64

The olist_geolocation_dataset file contains: 1000163 rows and 5 columns

Missing or nan values of:

geolocation_zip_code_prefix	0
geolocation_lat	0
geolocation_lng	0
geolocation_city	0
geolocation_state	0

The olist_order_reviews_dataset file contains: 99224 rows and 7 columns

Missing or nan values of:

review_id	0
order_id	0
review_score	0
review_comment_title	87656
review_comment_message	58247
review_creation_date	0
review_answer_timestamp	0

The olist_order_items_dataset file contains: 112650 rows and 7 columns

Missing or nan values of:

order_id	0
order_item_id	0
product_id	0
seller_id	0
shipping_limit_date	0
price	0
freight_value	0

The olist_orders_dataset file contains: 99441 rows and 8 columns

Missing or nan values of:

order_id	0
customer_id	0
order_status	0
order_purchase_timestamp	0
order_approved_at	160
order_delivered_carrier_date	1783
order_delivered_customer_date	2965
order_estimated_delivery_date	0

The olist_products_dataset file contains: 32951 rows and 9 columns

Missing or nan values of:

product_id	0
product_category_name	610
product_name_lenght	610
product_description_lenght	610
product_photos_qty	610
product_weight_g	2
product_length_cm	2
product_height_cm	2
product_width_cm	2

dtype: int64

Sélection des variables à utiliser dans l'analyse

Variables sélectionnées pour l'analyse

De l'ensemble de données ***olist_customers_dataset***, nous gardons :
- **customer_id** → clé des ensembles de données des commandes.
- **customer_unique_id**.
- **customer_zip_code_prefix**.
- **customer_city**.
- **customer_state**.

A partir de l'ensemble de données ***olist_geolocation_dataset***, nous allons garder :
- **geolocation_zip_code_prefix**.
- **geolocation_lat**.
- **geolocation_lng**.

De l'ensemble de données ***olist_order_items_dataset***, nous allons garder :
- **order_item_id**.
- **order_id**.
- **product_id**.

Dans l'ensemble de données ***olist_order_payments_dataset***, nous gardons :
- **order_id**.
- **payment_type**.
- **payment_value**.

Dans l'ensemble de données des ***olist_order_reviews_dataset*** nous gardons :
- **review_id**.
- **order_id**.
- **review_score**.

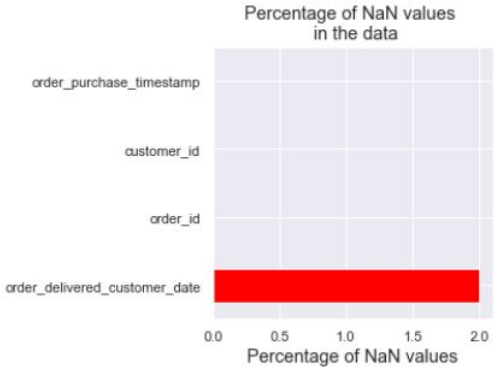
A partir de l'ensemble de données ***olist_orders_dataset***, nous gardons :
- **order_id**.
- **customer_id**.
- **order_purchase_timestamp**.
- **order_delivered_customer_date**.

Dans le jeu de données ***olist_products_dataset***, nous gardons :
- **product_id**.
- **product_category_name**.

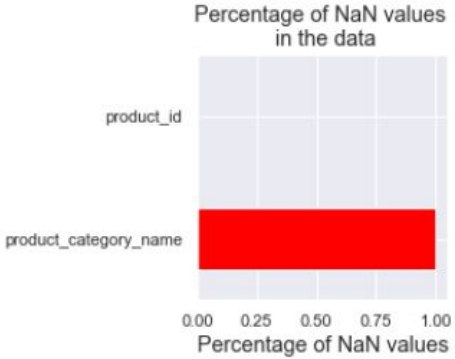
De l'ensemble de données ***product_category_name_translation***, nous gardons :
- **product_category_name**.
- **product_category_name_english**.

Nettoyage des données

Exploration et traitement des valeurs manquantes



Dans l'ensemble de données ***olist_order_items_dataset*** : Il y a **2965** valeurs manquantes dans la variable “order_delivered_customer_date”. Nous les avons supprimées.



Dans le jeu de données ***olist_products_dataset*** : Il y a **610** valeurs manquantes dans la variable “product_category_name”. Nous les avons supprimées.

Exploration et traitement des valeurs dupliquées

Le nombre de valeurs dupliquées dans la variable 'product_id' du jeu de données ***olist_products_dataset*** est : **0**

Le nombre de valeurs dupliquées dans la variable 'order_id' de l'ensemble de données ***olist_order_items_dataset*** est de : **13984**

Le nombre de valeurs dupliquées dans la variable 'product_id' de l'ensemble de données ***olist_order_items_dataset*** est de : **79699**

Une commande peut comporter plusieurs produits ou articles (indiqués dans la variable “order_item_id”).

Nettoyage des données

Exploration et traitement des valeurs dupliquées

Le nombre de valeurs dupliquées dans la variable 'geolocation_zip_code_prefix' du jeu de données ***olist_geolocation_dataset*** est : **981148**

On a éliminé les valeurs dupliquées en regroupant le “zip_code_prefix” et gardé la valeur moyenne de la latitude et de la longitude de chaque client.

Le nombre de valeurs dupliquées dans la variable 'order_id' du jeu de données ***olist_orders_dataset*** est : **0**

Le nombre de valeurs dupliquées dans la variable 'customer_id' du jeu de données ***olist_orders_dataset*** est : **0**

Il n'y a qu'une seule commande associée à chaque identifiant de client dans l'ensemble de données des commandes Olist.

Le nombre de valeurs dupliquées dans la variable ‘customer_unique_id’ du jeu de données ***olist_customers_dataset*** est : **3345**

Le “customer_unique_id” est un identifiant unique pour chaque client. Chaque commande est attribuée à un “customer_unique_id” (le même client obtiendra des identifiants différents pour des commandes différentes alors que le “customer_unique_id” est répété).

Le nombre de valeurs dupliquées dans la variable ‘order_id’ du jeu de données ***olist_order_reviews_dataset*** est : **551**

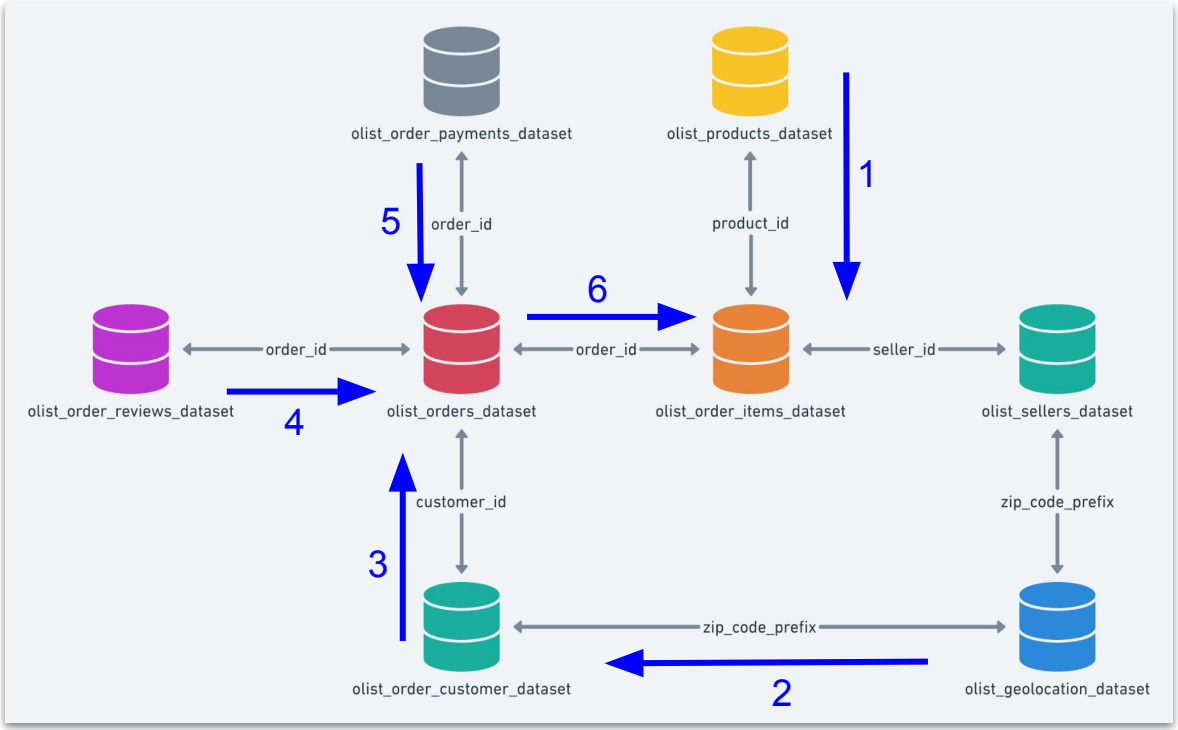
Certains clients peuvent avoir modifié leur évaluation pour une commande.

Le nombre de valeurs dupliquées dans la variable ‘order_id’ du jeu de données ***olist_order_payments_dataset*** est : **4446**

Un client peut payer une commande avec plus d'un mode de paiement.

Jointure de tables par clés

Route pour effectuer les jointures entre les tableaux



Données intégrées dans un seul tableau

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 112912 entries, 0 to 112911
Data columns (total 17 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   product_id                      112912 non-null  object
 1   order_id                        112912 non-null  object
 2   order_item_id                   112912 non-null  int64
 3   review_id                       112912 non-null  object
 4   review_score_for_a_customer_order  112912 non-null  int64
 5   zip_code_prefix                 112912 non-null  int64
 6   customer_geolocation_latitude   112912 non-null  float64
 7   customer_geolocation_longitude  112912 non-null  float64
 8   customer_id                     112912 non-null  object
 9   customer_unique_id              112912 non-null  object
 10  customer_city                   112912 non-null  object
 11  customer_state                  112912 non-null  object
 12  order_purchase_timestamp        112912 non-null  object
 13  order_delivered_customer_date   112912 non-null  object
 14  payment_type                    112912 non-null  object
 15  payment_value_for_a_customer_order  112912 non-null  float64
 16  product_category_name_english   112912 non-null  object
dtypes: float64(3), int64(3), object(11)
memory usage: 15.5+ MB
```

Feature engineering

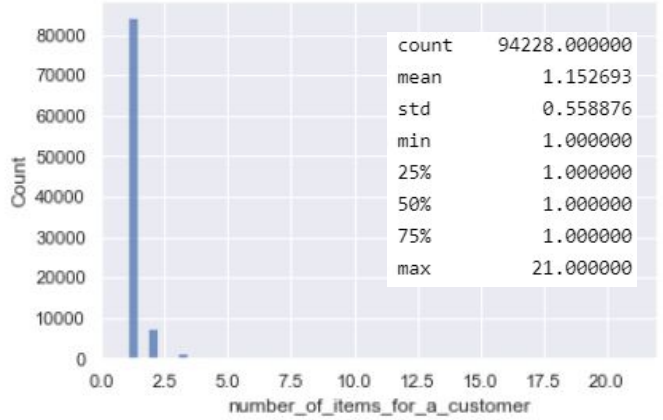
➔ **Création du nombre de commandes d'un client : ‘number_of_orders_for_a_customer’**

Un client (customer_unique_id) a des identifiants de commande (order_id) différents pour chaque commande. Cela signifie que nous pouvons obtenir le nombre de commandes en comptant combien de fois “customer_unique_id” est répété.

6% des clients ont passé plus d'une commande. Puis 0,8% ont passé plus de 2 commandes.

➔ **Création du nombre d'articles d'une commande : ‘number_of_items_for_a_customer’**

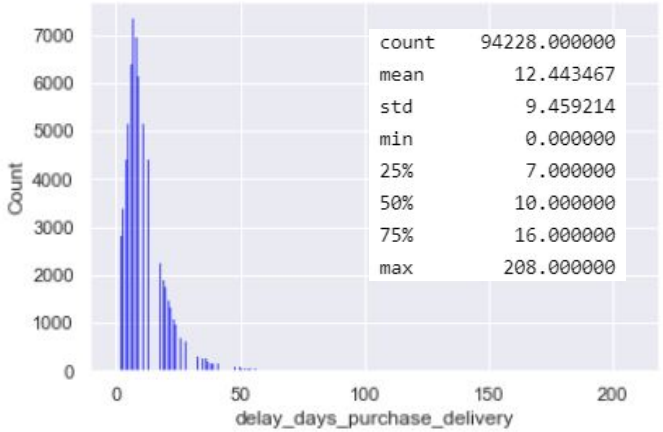
Une commande peut comporter plusieurs produits ou articles (également indiqués dans la variable “order_item_id”).



Feature engineering

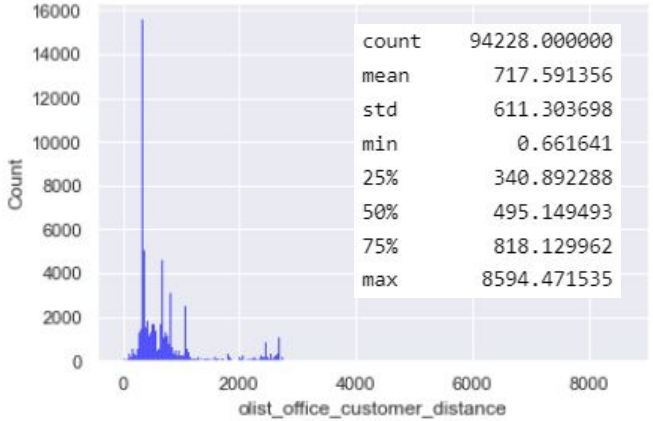
➔ **Création de la variable délai de jours achat-livraison : ‘delay_days_purchase_delivery’**

('order_purchase_timestamp') - ('order_delivered_customer_date')



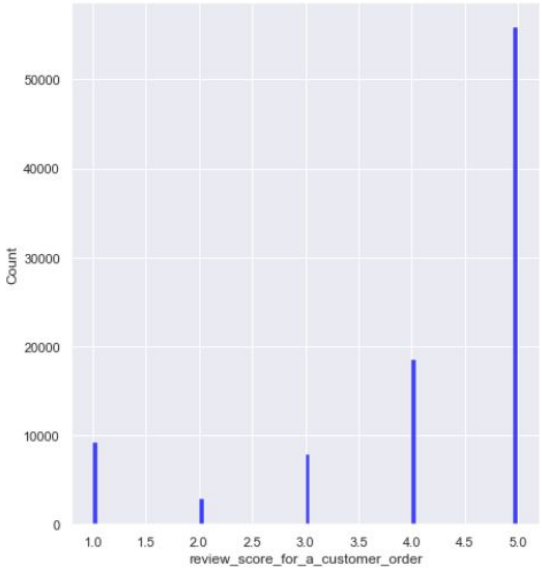
➔ **Création d'une variable avec la distance client-vendeur : ‘olist_office_customer_distance’**

Distance haversine entre le bureau d'Olist. La distance haversine est interprétée en kilomètres.



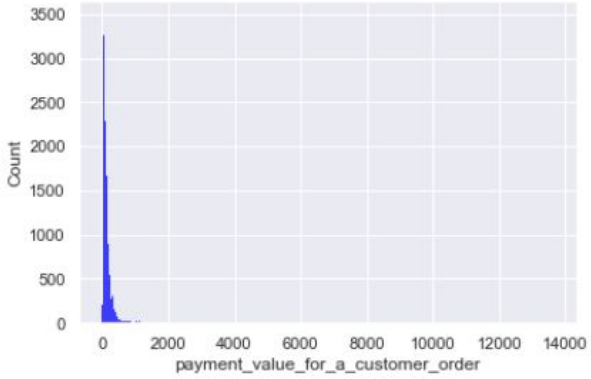
Visualisation de certains indicateurs

Distribution du review score et de l'argent dépensé des customers

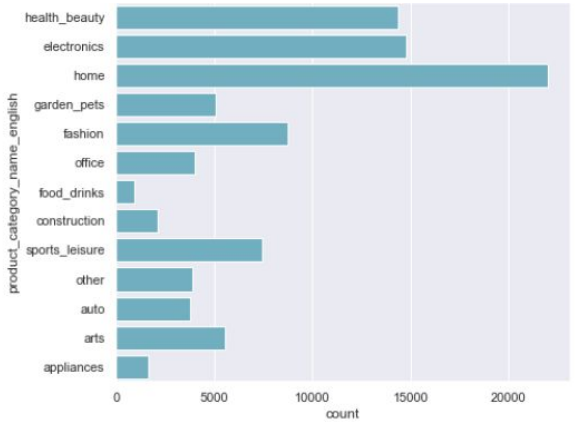


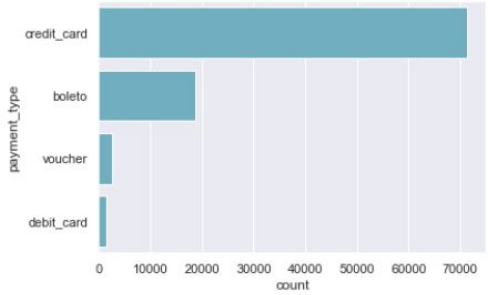
count	94228.000000
mean	4.157448
std	1.283207
min	1.000000
25%	4.000000
50%	5.000000
75%	5.000000
max	5.000000

count	94228.000000
mean	157.112458
std	215.286394
min	0.000000
25%	60.010000
50%	103.005000
75%	174.802500
max	13664.080000



Distribution des produits et modes de paiement utilisés par les clients





Utilisation de la méthode RFM comme référence de segmentation

Pour la segmentation des clients, nous commençons par utiliser une méthode appelée RFM, qui est largement utilisée pour l'analyse des clients dans les études de marché.

Nous allons utiliser la méthode RFM pour comparer la segmentation des clients avec d'autres méthodes de clustering, telles que K-means.

RFM représente les trois dimensions :

Récence - Depuis combien de temps le client a-t-il acheté ? (De la date d'achat à ce jour)
Fréquence - A quelle fréquence achète-t-il ?
Valeur monétaire - Combien dépensent-ils ?

Voir la référence : https://en.wikipedia.org/wiki/RFM_(market_research)

Pour calculer la **Récence**, nous allons calculer le nombre de jours qui se sont écoulés depuis que le client a passé une commande. La **Fréquence** correspond au nombre de commandes passées par un client, et la **Valeur monétaire** est la valeur du paiement de la commande.

Utilisation de la méthode RFM comme référence de segmentation

Dans l'analyse commerciale, nous utilisons souvent ce concept pour diviser les clients en différents segments, comme les clients à forte valeur, les clients à valeur moyenne ou les clients à faible valeur, et bien d'autres encore.

Référence : https://www.geeksforgeeks.org/rfm-analysis-analysis-using-python/

Dans la méthode RFM, nous devons créer un score basé sur des règles heuristiques où R + F + M = Score RFM

- ➔ Il dépend des règles de l'entreprise.
- ➔ Par exemple, une entreprise peut donner plus de priorité à la valeur monétaire.
- ➔ Nous multiplions les termes **R**, **F** et **M** par les constantes **a**, **b** et **c**, pour déterminer en fonction des priorités de l'entreprise.

a(Récence) + b(Fréquence) + c(Valeur monétaire) = Score RFM

Classement du client en fonction du score RFM :

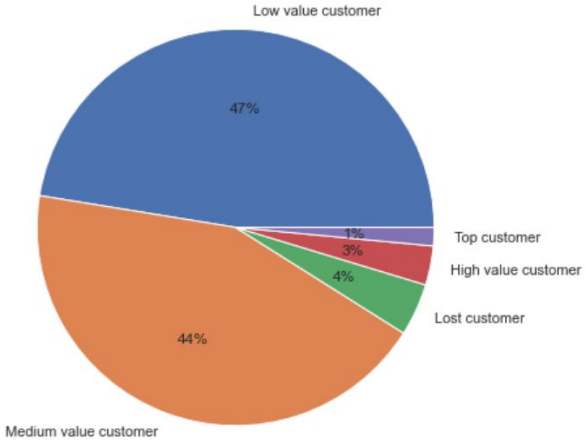
- **Score RFM = 5 : Client de premier ordre**
- **Score RMF = 4 : Client de grande valeur**
- **Score RMF = 3 : Client de valeur moyenne**
- **Score RMF = 2 : Client à faible valeur ajoutée**
- **Score RMF = 1 : Client perdu**

Utilisation de la méthode RFM comme référence de segmentation

Nous allons utiliser la règle suivante :

40 (Récence) + 40 (Fréquence) + 20 (Valeur monétaire) = Score RFM

Ici, nous accordons la même importance à la Récence et à la Fréquence, c'est-à-dire aux clients qui ont récemment passé des commandes et qui ont acheté plus d'un produit par commande. Ensuite, nous accordons moins d'importance à la Valeur monétaire.



	Frequency	Monetary_value	Recency	RFM_Score	size
	mean	mean	mean	mean	
Customer_segment					
High value customer	2.2	141.1	249.6	4.0	3005
Lost customer	1.0	150.8	13.1	1.0	4021
Low value customer	1.0	159.2	131.2	2.0	44698
Medium value customer	1.0	157.2	369.3	3.0	41091
Top customer	2.3	140.2	468.8	5.0	1413

Utilisation de la méthode RFM comme référence de segmentation

Méthode des quintiles RFM

Sur la base d'une bibliographie et d'articles récents sur le modèle RFM :
(https://practicaldatascience.co.uk/data-science/how-to-visualise-rfm-data-using-treemaps)

- Kabaskal, İ., 2020. Customer Segmentation Based On Recency Frequency Monetary Model: A Case Study in E-Retailing. International Journal of InformaticsTechnologies, 13(1).
- Putler Analytics – RFM analysis for successful customer segmentation, https://www.putler.com/rfm-analysis, 26.04.2019.
- Uysal, Ü.C., 2019. RFM-based Customer Analytics in Public Procurement Sector (Doctoral dissertation, Ankara Yıldırım Beyazıt Üniversitesi Sosyal Bilimler Enstitüsü).

➔ Nous avons effectué une deuxième méthode pour la segmentation des clients basée sur une méthode des quintiles. Comme nous le verrons plus loin, cette méthode nous permettra en principe d'obtenir une segmentation plus détaillée.

➔ Nous avons regroupé les clients en quintiles en fonction de leur comportement.

➔ Nous avons calculé les scores de **Récence**, de **Fréquence** et de **Valeur monétaire** pour leurs valeurs relatives aux quintiles.

Utilisation de la méthode RFM comme référence de segmentation

➔ Dans cette approche, nous avons calculé le score de **Récence** et **“R_Score”** et nous avons créé le score **"FM_Score"**, qui est la moyenne arrondie des scores de **Fréquence** et de **Valeur monétaire**.

- **Champions :** Ont acheté récemment, achètent souvent et dépensent le plus.
- **Loyal Customers :** Achètent régulièrement. Réactifs aux promotions.
- **Potential Loyalists :** Clients récents avec une fréquence moyenne.
- **Recent Customers :** Ont acheté très récemment, mais pas souvent.
- **Customers Needing Attention :** Récence, fréquence et valeurs monétaires supérieures à la moyenne. Mais ils n'ont peut-être pas acheté très récemment.
- **At risk :** Achetés souvent mais il y a longtemps. Il faut les faire revenir.
- **Can't Lose Them :** Achetait souvent mais n'est pas revenu depuis longtemps.
- **About To Sleep :** Récence et fréquence inférieures à la moyenne. Je les perdrai si je ne les réactive pas.
- **Hibernating :** Le dernier achat remonte à longtemps et le nombre de commandes est faible.
- **Lost Customers :** Ils ont acheté il y a longtemps et ne sont jamais revenus.

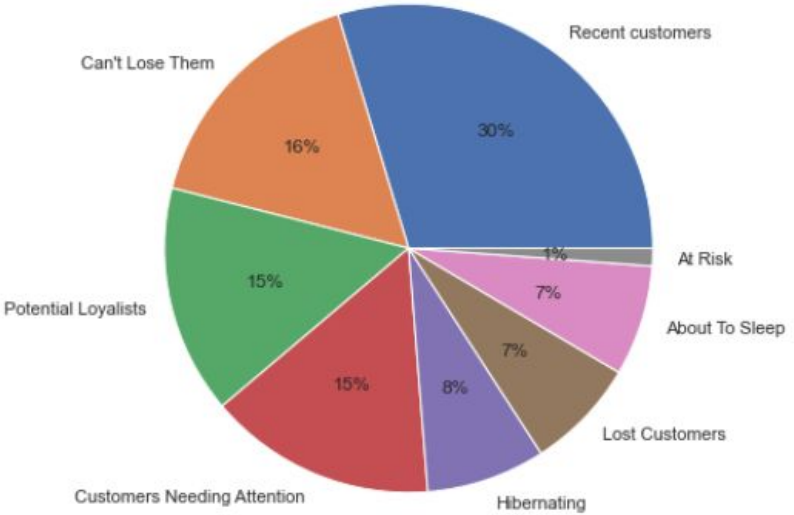
Utilisation de la méthode RFM comme référence de segmentation

➔ Chaque client a une description basée sur son comportement associé à son **R_Score** et **FM_score**.

R_Score, FM_Score	Nom du segment	Description
1,1	**Lost Customers**	Ils ont acheté il y a longtemps et ne sont jamais revenus.
1,2	**Hibernating**	Le dernier achat remonte à longtemps et le nombre de commandes est faible.
2,1	**About To Sleep**	Récence et fréquence inférieures à la moyenne. Je les perdrai si je ne les réactive pas.
1-2, 1-3	**Can't Lose Them**	Achetait souvent mais n'est pas revenu depuis longtemps.
1-2, 3-5	**At risk**	Achetés souvent mais il y a longtemps. Il faut les faire revenir.
2-3, 1-2	**Customers Needing Attention**	Récence, fréquence et valeurs monétaires supérieures à la moyenne. Mais ils n'ont peut-être pas acheté très récemment.
3-5, 1-2	**Recent Customers**	Ont acheté récemment, mais pas souvent.
3-5, 3-5	**Potential Loyalists**	Clients récents avec une fréquence moyenne.
4-5, 4-5	**Loyal Customers**	Achètent régulièrement. Réactifs aux promotions.
5,5	**Champions**	Ont acheté récemment, achètent souvent et dépensent le plus.

Utilisation de la méthode RFM comme référence de segmentation

➔ Chaque client a une description basée sur son comportement associé à son **R_Score** et **FM_score**.



	Frequency	Monetary_value	Recency	R_score	FM_score	size
	mean	mean	mean	mean	mean	
Segment_name						
About To Sleep	1.0	52.0	136.7	2.0	1.0	6894
At Risk	2.2	216.7	95.9	1.5	4.3	1101
Can't Lose Them	1.1	270.1	112.9	1.7	2.5	15422
Customers Needing Attention	1.0	91.2	221.6	3.0	1.5	14206
Hibernating	1.0	131.3	45.9	1.0	2.0	7424
Lost Customers	1.0	51.9	44.4	1.0	1.0	6956
Potential Loyalists	1.3	341.9	341.1	4.0	3.2	14267
Recent customers	1.0	90.6	395.4	4.5	1.5	27958

Segmentation des clients à l'aide de Machine Learning

Données transformées et standardisées

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 102333 entries, 3033 to 121071
Data columns (total 7 columns):
 #   Column                              Non-Null Count   Dtype
---  ------                              --------------   -----
 0   review_score_for_a_customer_order   102333 non-null  float64
 1   number_of_items_for_a_customer      102333 non-null  float64
 2   delay_days_purchase_delivery        102333 non-null  float64
 3   olist_office_customer_distance      102333 non-null  float64
 4   Frequency                           102333 non-null  float64
 5   Monetary_value                      102333 non-null  float64
 6   Recency                             102333 non-null  float64
dtypes: float64(7)
memory usage: 6.2 MB
```

Nous avons maintenant les variables pertinentes pour effectuer la clusterisation en utilisant l'apprentissage automatique

Pourquoi transformer ?

Nous avons aussi transformé les variables qui sont fortement asymétriques en faisant une transformation logarithmique

Pourquoi normaliser ou standardiser ?

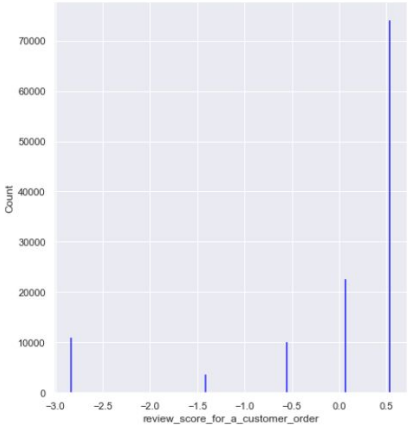
De nombreux algorithmes d'apprentissage automatique tentent de trouver des tendances dans les données en comparant les caractéristiques des points de données. Cependant, un problème se pose lorsque les caractéristiques sont à des échelles radicalement différentes

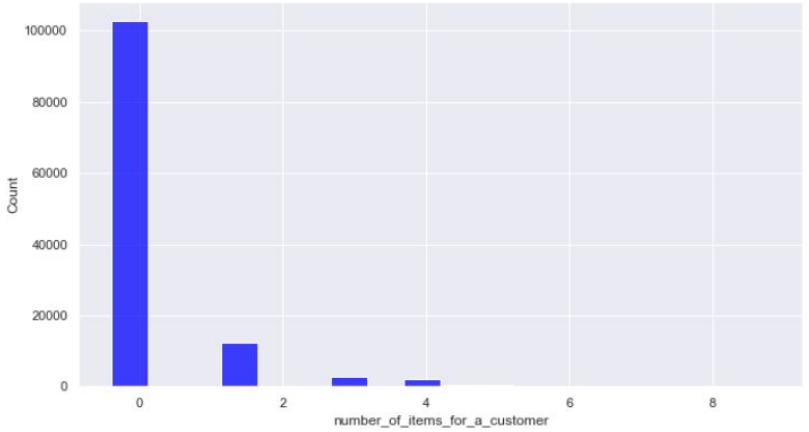
Z-scores pour détecter les valeurs aberrantes

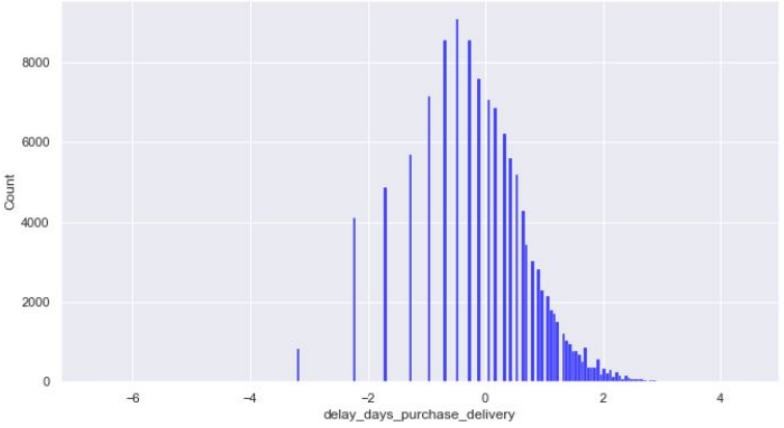
Nous avons utilisé la standardisation des z-scores en considérant une tolérance de +/-3 des z-scores pour supprimer les valeurs aberrantes

Segmentation des clients à l'aide de Machine Learning

Données transformées et standardisées

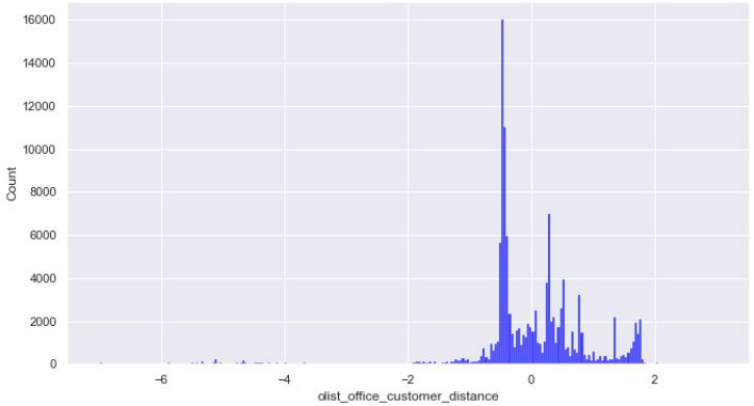


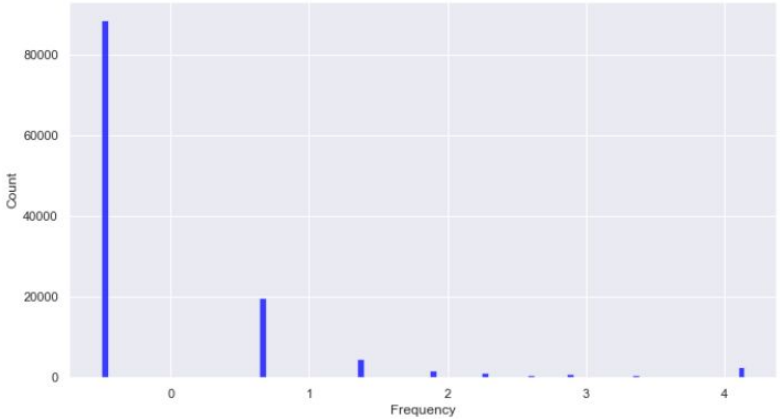


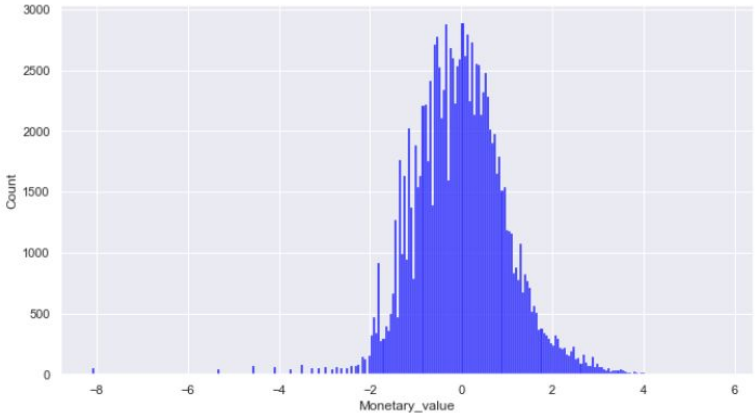


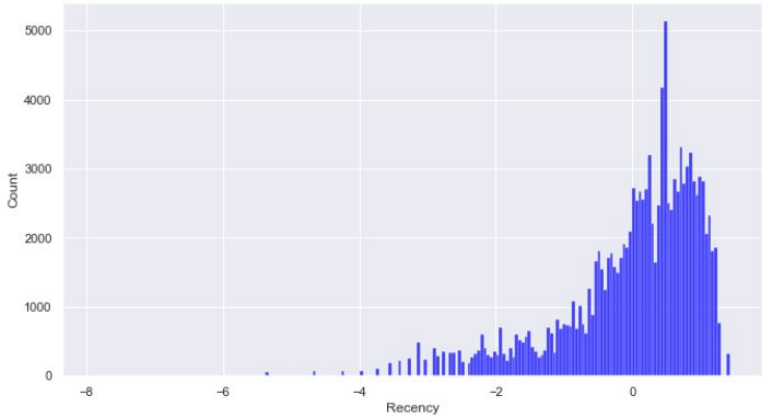
Segmentation des clients à l'aide de Machine Learning

Données transformées et standardisées

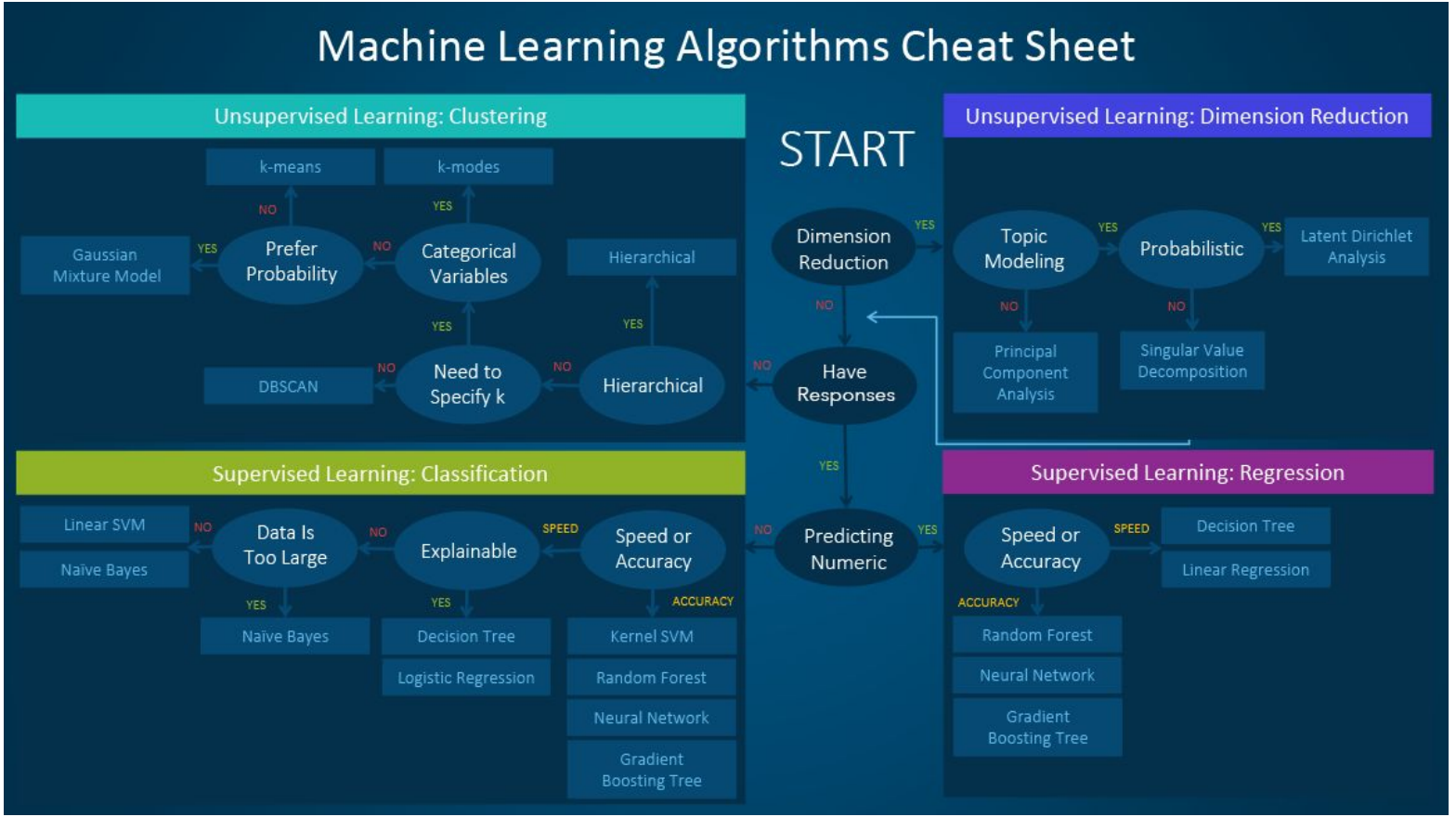








Segmentation des clients à l'aide de Machine Learning

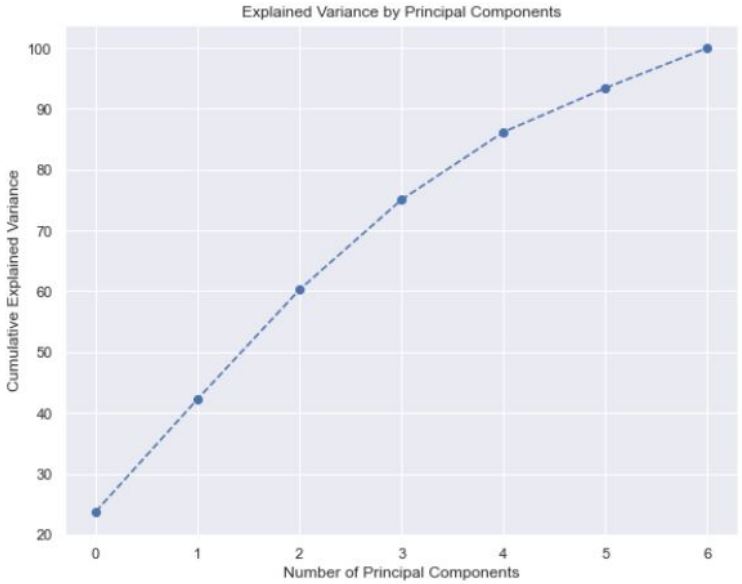


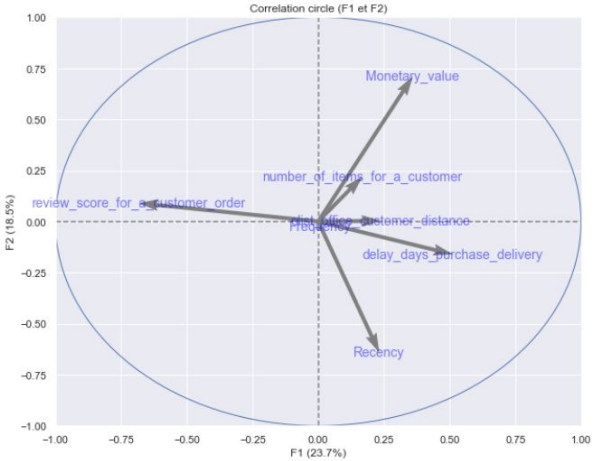
https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/

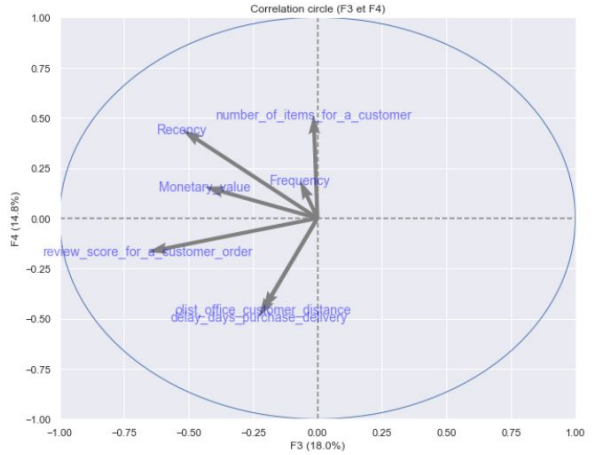
Segmentation des clients à l'aide de Machine Learning

PCA pour la réduction de la dimensionnalité

Nous avons effectué une réduction de la dimensionnalité en utilisant **l'analyse PCA**. C'est-à-dire que nous avons initialement 7 dimensions ou 7 features. L'idée est de visualiser une projection des données dans un plan 2D.



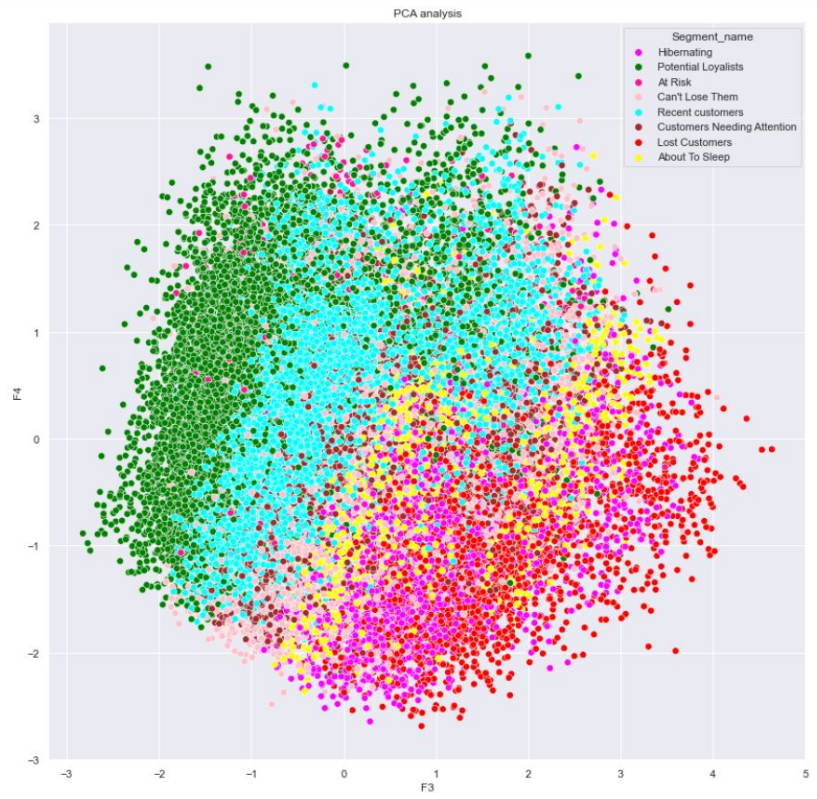




Segmentation des clients à l'aide de Machine Learning

PCA pour la réduction de la dimensionnalité

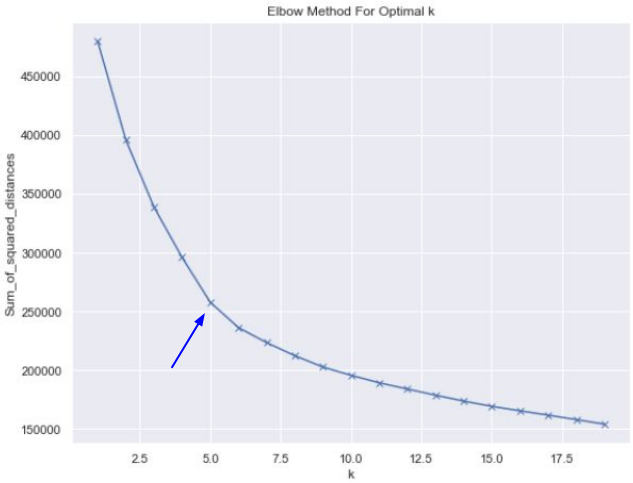
Visualisation des données en composantes principales



Segmentation des clients à l'aide de Machine Learning

Clusterisation K-means

Méthode du coude : pour déterminer le meilleur nombre K de clusters



Lorsque le nombre de clusters augmente, la variance (somme des carrés à l'intérieur des clusters) diminue. Il y a donc un moment où la diminution de la somme des carrés des distances n'est pas significative pour une augmentation de la valeur de K-cluster.

Résultat de K-means

	Frequency	Monetary_value	Recency	review_score_for_a_customer_order
	mean	mean	mean	mean
Segment_k_means_label				
0	2.0	179.0	295.0	5.0
1	2.0	85.0	294.0	5.0
2	2.0	148.0	106.0	5.0
3	2.0	160.0	253.0	2.0
4	3.0	180.0	261.0	4.0

	number_of_items_for_a_customer	delay_days_purchase_delivery	olist_office_customer_distance	size
	mean	mean	mean	
Segment_k_means_label				
0	1.0	16.0	1046.0	116282
1	1.0	10.0	523.0	137433
2	1.0	9.0	615.0	44853
3	1.0	16.0	782.0	31823
4	2.0	12.0	760.0	91532

Segmentation des clients à l'aide de Machine Learning

Clusterisation K-means

Afin d'attribuer des noms de segments de clients aux étiquettes obtenues par la méthode K-means, nous allons d'abord relier la segmentation détaillée obtenue par la méthode RFM aux cinq principaux groupes obtenus auparavant par la méthode heuristique RFM :

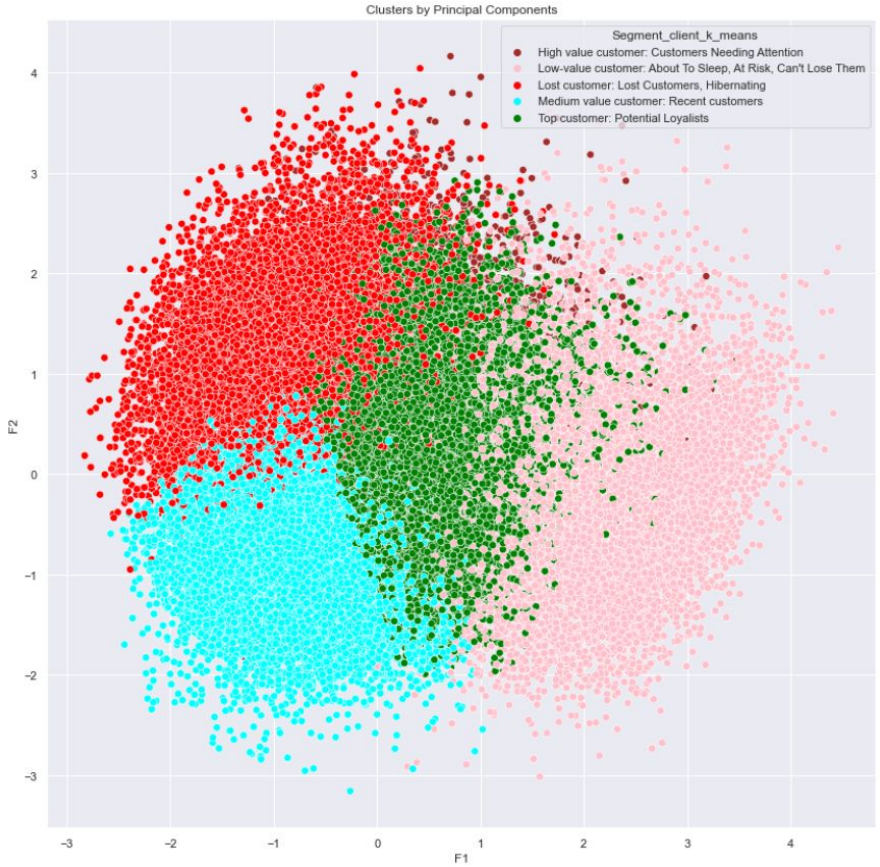
- **Top customer : Potential Loyalists**
- **High value customer : Customers Needing Attention**
- **Medium value customer : Recent customers**
- **Low-value customer : About To Sleep, At Risk, Can't Lose Them**
- **Lost customer : Lost Customers, Hibernating**

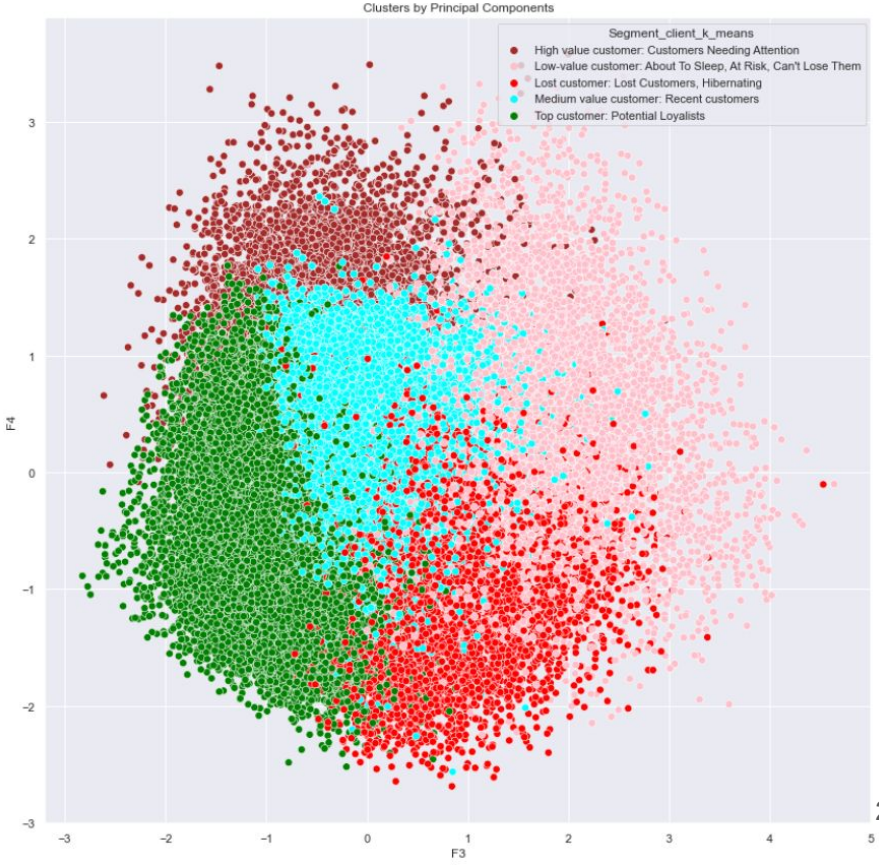
Nous avons le comportement pour chaque sous-segment de clients :

Potential Loyalists : Clients récents avec une fréquence moyenne. Dépensent le plus.
Customers Needing Attention : Récence, fréquence et valeur monétaire supérieures à la moyenne. Ils n'ont peut-être pas acheté très récemment.
Recent customers : Ont acheté récemment, mais pas souvent.
Can't Lose Them: Achetaient fréquemment mais ne sont pas revenus depuis longtemps.
At Risk : Achetés souvent mais il y a longtemps. Besoin de les ramener.
About To Sleep : Récence et fréquence inférieures à la moyenne. Je les perdrai si je ne les réactive pas.
Hibernation : Le dernier achat remonte à longtemps et le nombre de commandes est faible.
Lost Customers : Achetés il y a longtemps et ne sont jamais revenus.

Segmentation des clients à l'aide de Machine Learning

Clusterisation K-means





Segmentation des clients à l'aide de Machine Learning

Clusterisation K-means

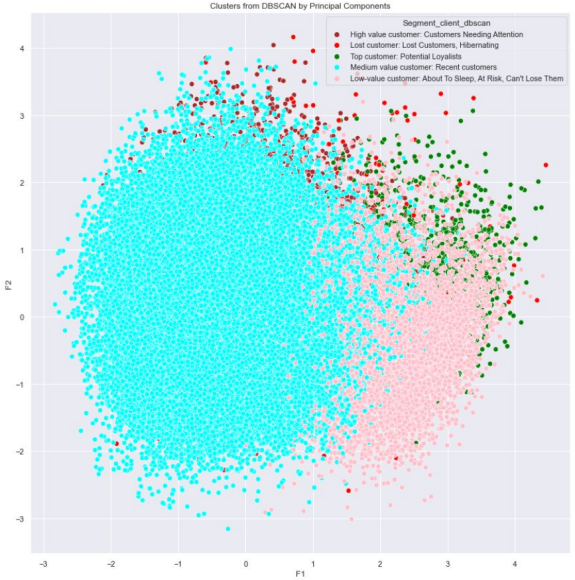
Frequency	Monetary_value	Recency	size	Client_segment	Client_subsegment	Description
mean	mean	mean				
2.0	179.0	295.0	116282	Top customer	Potential Loyalists	Recent customers with average frequency. Spend...
2.0	85.0	294.0	137433	Medium value customer	Recent customers	Bought recently, but not often.
2.0	148.0	106.0	44853	Lost customer	Lost Customers, Hibernating	Purchased long time ago and never came back. A...
2.0	160.0	253.0	31823	Low-value customer	About To Sleep, At Risk, Can't Lose Them	Used to purchase frequently but have not retur...
3.0	180.0	261.0	91532	High value customer	Customers Needing Attention	Above average recency, frequency and monetary ...

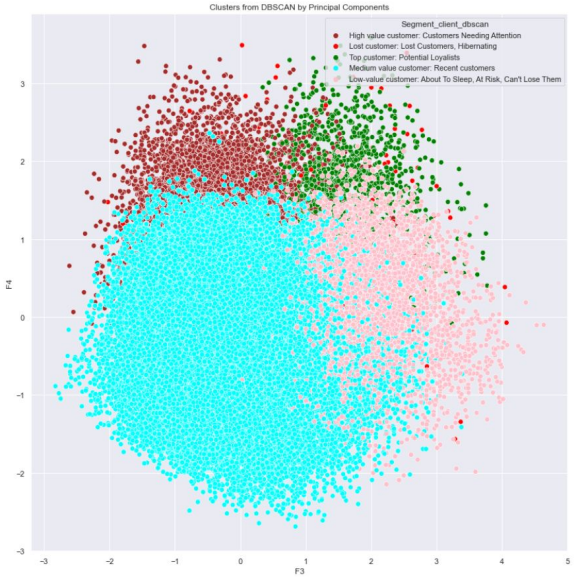
Frequency	Monetary_value	Recency	review_score_for_a_customer_order	number_of_items_for_a_customer	delay_days_purchase_delivery	olist_office_customer_distance	size	Client_segment
mean	mean	mean	mean	mean	mean	mean		
2.0	179.0	295.0	5.0	1.0	16.0	1046.0	116282	Top customer
2.0	85.0	294.0	5.0	1.0	10.0	523.0	137433	Medium value customer
2.0	148.0	106.0	5.0	1.0	9.0	615.0	44853	Lost customer
2.0	160.0	253.0	2.0	1.0	16.0	782.0	31823	Low-value customer
3.0	180.0	261.0	4.0	2.0	12.0	760.0	91532	High value customer

Segmentation des clients à l'aide de Machine Learning

Clusterisation par DBSCAN

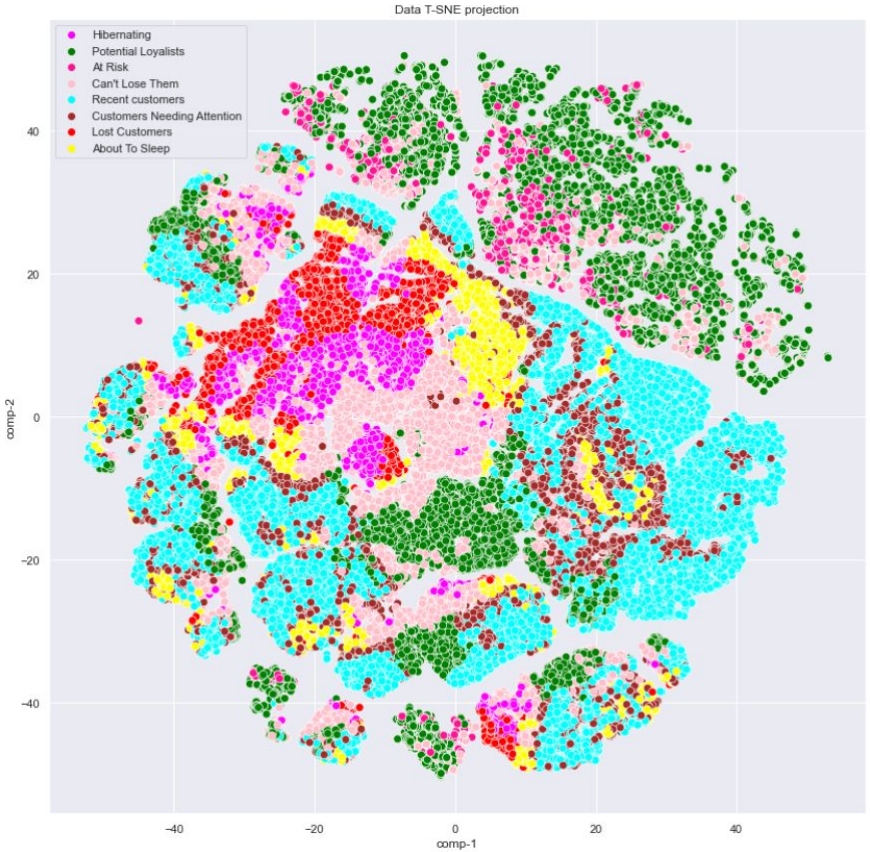
Frequency	Monetary_value	Recency	review_score_for_a_customer_order	number_of_items_for_a_customer	delay_days_purchase_delivery	olist_office_customer_distance	size	Client_segment
mean	mean	mean	mean	mean	mean	mean		
4.0	229.0	214.0	3.0	2.0	16.0	1539.0	8730	Lost customer: Lost Customers, Hibernating
3.0	179.0	259.0	4.0	2.0	12.0	689.0	88543	High value customer: Customers Needing Attention
3.0	194.0	260.0	2.0	2.0	14.0	678.0	10204	Top customer: Potential Loyalists
2.0	129.0	268.0	5.0	1.0	12.0	744.0	298572	Medium value customer: Recent customers
2.0	133.0	253.0	2.0	1.0	18.0	752.0	15874	Low-value customer: About To Sleep, At Risk, C...





Segmentation des clients à l'aide de Machine Learning

T-SNE pour la réduction de la dimensionnalité

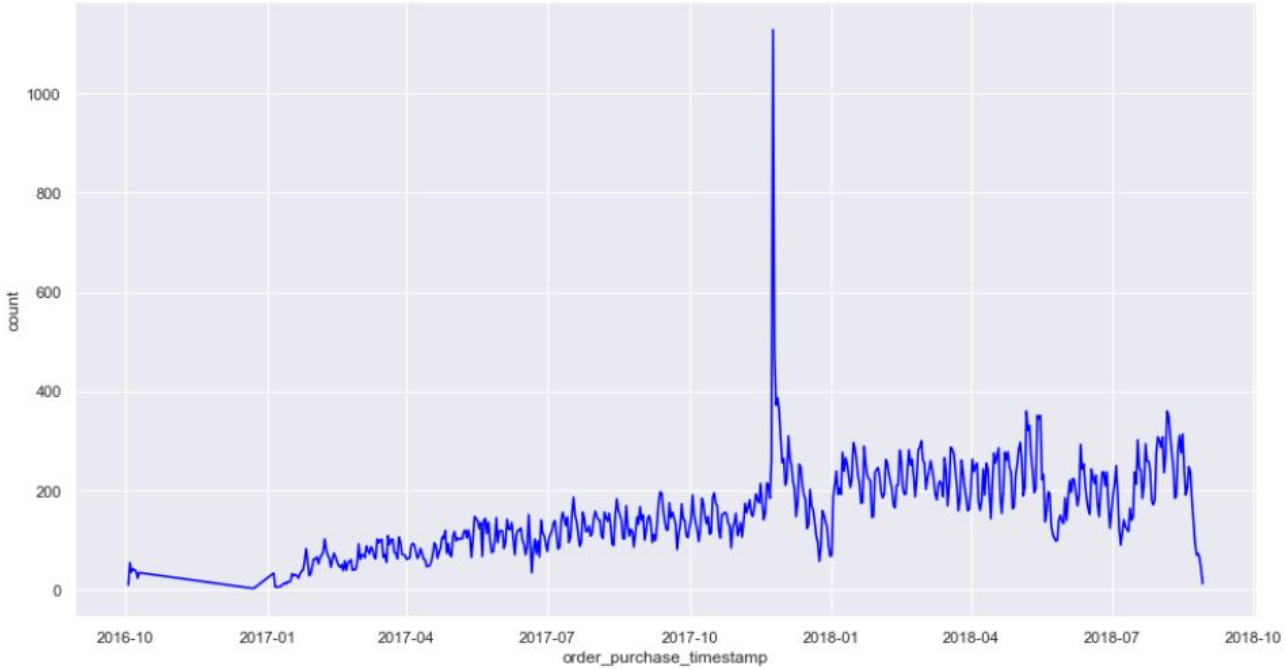


- ➔ T-SNE modélise la distribution de probabilité des voisins autour de chaque point (client) en fonction de leurs caractéristiques.
- ➔ Nous observons que **Lost Customers**, **Hibernating** et **About to Sleep**, ont tendance à être ensemble.
- ➔ Les clients **At Risk** ont tendance à être ensemble avec **Can't Lose Them**.
- ➔ Les **Recent Customers** ont tendance à être proches les groupes **Customers Needing Attention**.

Analyse de la maintenance par segmentation

Maintenant que nous avons associé les étiquettes créées par K-means aux segments de clients, nous allons essayer de déterminer la dynamique de la segmentation des clients dans le temps.

Nous devons d'abord avoir une idée de l'évolution de nos données dans le temps.



Analyse de la maintenance par segmentation

Nous avons défini une période de temps égale à un an. Nous avons utilisé les données de septembre 2017 à septembre 2018.

Nous avons déterminé combien de clients appartiennent à chaque segment de clientèle dans chaque saison de l'année :

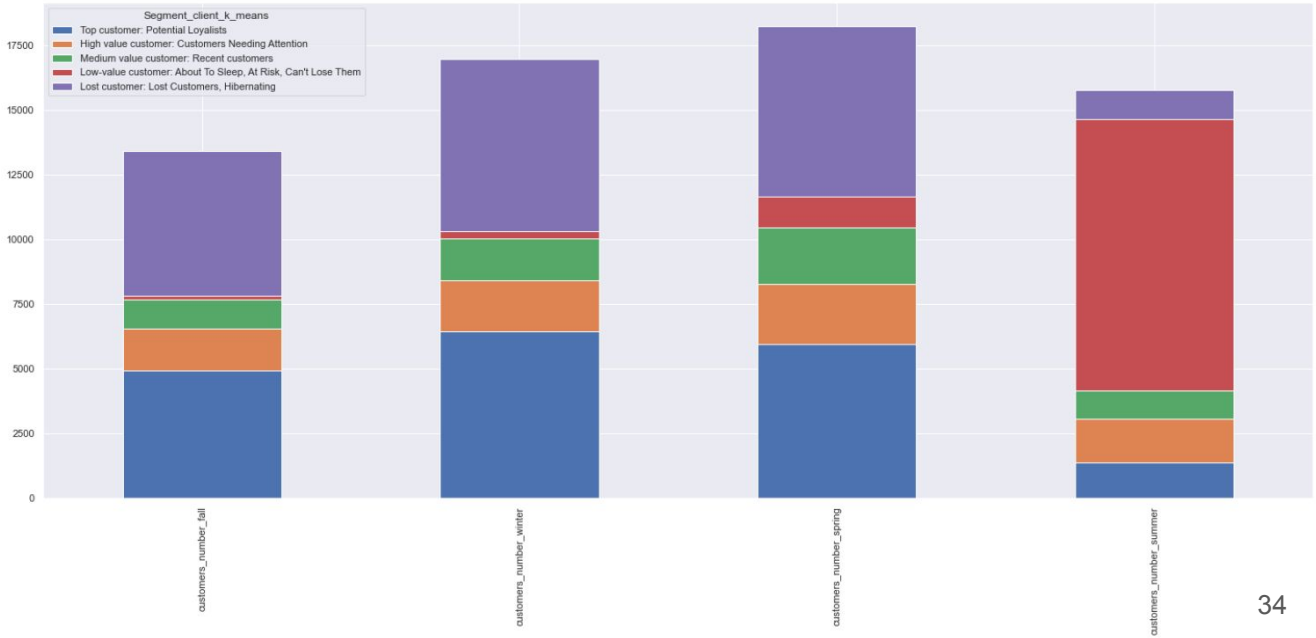
- Septembre, 2017 - Novembre, 2017 → Automne.
- Décembre 2017 - Février 2018 → Hiver
- Mars, 2018 - Mai, 2018 → Printemps
- 2018 - Août 2018 → Été

➔ Il y a un changement dans le nombre de clients appartenant à chaque segment pour chaque saison de l'année.

➔ Au début de la période de l'année (automne, hiver et printemps), il y a plus de Top Clients qu'à la fin de l'année (été).

➔ Nous constatons la perte de bons clients et l'augmentation de mauvais clients au fil de l'année.

➔ L'été commence avec un faible nombre de Top Clients, puis nous voyons qu'ils deviennent des clients à faible valeur ajoutée.



Conclusions

- ➔ Après une sélection appropriée de variables pour notre base de données, nous avons pu tester différentes méthodes de segmentation des clients.

- ➔ Nous avons établi une segmentation de référence en utilisant la méthode RFM et nous avons pu diviser les clients d'Olist en catégories de faible, moyenne et forte valeur.

- ➔ Nous avons ensuite appliqué des algorithmes de Machine Learning non supervisés pour effectuer la segmentation des clients en fonction de leur profil et de leur comportement.
  - ◆ Nous avons observé que les avis des clients, l'argent dépensé et la récence des clients ont une grande influence sur la segmentation.

- ➔ Nous avons testé deux modèles de clustering non supervisé : K-means et DBSCAN, et nous avons obtenu une bonne séparation des clients avec K-means.
  - ◆ Il est important de noter qu'un bon choix des hyperparamètres est critique.

- ➔ Nous avons effectué une analyse de la dynamique des segments de clients dans le temps, dans laquelle nous recommandons d'effectuer la clusterisation tous les trois mois.

- ➔ Nous recommandons fortement d'effectuer l'analyse de segmentation avec une fréquence plus représentative où les clients ont passé un plus grand nombre de commandes.