Projet 4 OC : Segmentation des clients d'un site e-commerce

Présenté par : NAMA NYAM Guy Anthony

Mentor : Julien Hendrick

22 Février 2020

OPENCLASSROOMS

Sommaire

Contexte

- Toute entreprise aspire à un retour sur investissement le plus rapidement possible.
- Pour atteindre cet objectif pour les sites d'e-commerces, les entreprises préconisent la **segmentation** pour réduire les ressources à allouer dans le secteur du marketing.
- La segmentation des clients dans notre contexte consiste à découper **analytiquement** en sous-clients homogènes.
- Ces segments constituent la base des campagnes de communication des équipes de marketing, c'est dire l'importance d'une telle opération.

Problématiques et objectifs

- De diverses difficultés se dressent dans la recherche des segments clients, donc :
- Comprendre les différents types d'utilisateurs grâce à leurs données personnelles.
- Rechercher les critères ou variables de mise en évidence des segments pour une utilisation optimale.

Objectif

Fournir une description actionnable de la segmentation obtenue et proposer une analyse de la stabilité des segments au cours du temps.

Approches

- Pour ce faire, nous utilisons deux approches à savoir :
  1. L'approche analytique avec la technique **RFM**
  2. L'approche automatique avec l'algorithme de **K-Means**
- L'analyse RFM(**Recency, Frequency, Monetary**) combinent ces trois paramètres(provenant de l'historique des transactions) pour segmenter les clients basés sur leurs comportements.
- L'algorithme de K-Means partitionne en K groupes l'ensemble des observations en minimisant la **distance euclidienne** de chaque observation à la moyenne de son cluster d'appartenance.

Sommaire

Description

- Base de données(olist_customers_dataset.csv) anonymisée du site d'e-commerce **Olist**, téléchargeable (https ://www.kaggle.com/olistbr/brazilian-ecommerce).
- Six(6) tables retenues sur les huit(8) initiales pour déduire le comportement RFM de chaque client.
- Les tables retenues : **olist_customers_dataset**, **olist_orders_dataset**, **olist_order_items_dataset** , **olist_products_dataset**, **olist_order_reviews_dataset**, **product_category_name_translation**

Nettoyage de la donnée et fusion

- La suppression des transactions non approuvées.

```
orders_data["order_approved_at"].isna().sum()
```

```
160
```

```
orders_data = orders_data[~orders_data["order_approved_at"].isna()]
```

- La suppression des transactions sans catégories produits

```
products_data["product_category_name"].isna().sum()
```

```
610
```

```
products_data = products_data[~products_data["product_category_name"].isna()]
```

- Jointure naturelle entre les lignes des différentes tables.

	customer_unique_id	customer_state	order_purchase_timestamp	price	product_category_name	review_score
0	861eff4711a542e4b93843c6dd7febb0	SP	2017-05-16 15:05:35	124.99	office_furniture	4
1	9eae34bbd3a474ec5d07949ca7de67c0	PA	2017-11-09 00:50:13	112.99	office_furniture	1
2	9eae34bbd3a474ec5d07949ca7de67c0	PA	2017-11-09 00:50:13	112.99	office_furniture	1
3	3c799d181c34d51f6d44bbbc563024db	RS	2017-05-07 20:11:26	124.99	office_furniture	3
4	23397e992b09769faf5e66f9e171a241	RJ	2018-02-03 19:45:40	106.99	office_furniture	4

```
[72] donnees.shape
```

```
(111696, 14)
```

Sommaire

Comportement clients

- De l'historique de transaction des clients(identifiant client), on déduit les comportements suivants :
  1. R : nombre de jour écoulé du dernier achat
  2. F : nombre d'achat effectué.
  3. M : somme totale dépensée.

- La table RFM obtenue :

```
rfmTable = donnees.groupby('customer_unique_id').\
            agg({"order_purchase_timestamp": lambda x: (NOW - x.max()).days,
                 "order_id": lambda x: x.nunique(),
                 "price": lambda x: sum(x)})

rfmTable['order_purchase_timestamp'] = rfmTable['order_purchase_timestamp'].astype(int)
rfmTable.rename(columns={'order_purchase_timestamp': 'recency',
                         'order_id': 'frequency',
                         'price': 'monetary_value'}, inplace=True)
```

```
rfmTable.head()
```

customer_unique_id	recency	frequency	monetary_value
0000366f3b9a7992bf8c76cfdf3221e2	116	1	129.90
0000b849f77a49e4a4ce2b2a4ca5be3f	119	1	18.90
0000f46a3911fa3c0805444483337064	542	1	69.00
0000f6ccb0745a6a4b88665a16c9f078	326	1	25.99
0004aac84e0df4da2b147fca70cf8255	293	1	180.00

Calcul RFM score

- Pour regrouper les clients, nous avons utilisé les quartiles de leurs différents comportements.

```
####  Calculs des valeurs de quartiles pour notre jeu de données
quartiles = rfmTable.quantile(q=[0.25, 0.5, 0.75])
print(quartiles)
```

```
      recency  frequency  monetary_value
0.25  118.0    1.0        47.9
0.50  222.0    1.0        89.8
0.75  351.0    1.0        155.8
```

- Attribuer les valeurs selon les quartiles.

```
############### Segmentation RFM score  ##################
segmented_rfm['r_quartile'] = segmented_rfm['recency'].apply(RScore, args=('recency',quartiles))
segmented_rfm['f_quartile'] = segmented_rfm['frequency'].apply(FMScore, args=('frequency',quartiles))
segmented_rfm['m_quartile'] = segmented_rfm['monetary_value'].apply(FMScore, args=('monetary_value',quartiles))

segmented_rfm['RFMScore'] = (segmented_rfm["r_quartile"].astype(str) + "-").str.cat(segmented_rfm["f_quartile"].astype(str) + "

segmented_rfm.head()
```

customer_unique_id	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore
0000366f3b9a7992bf8c76cfdf3221e2	115	1	129.90	1	4	2	1-4-2
0000b849f77a49e4a4ce2b2a4ca5be3f	118	1	18.90	1	4	4	1-4-4
0000f46a3911fa3c0805444483337064	541	1	69.00	4	4	3	4-4-3
0000f6ccb0745a6a4b88665a16c9f078	325	1	25.99	3	4	4	3-4-4
0004aac84e0df4da2b147fca70cf8255	292	1	180.00	3	4	1	3-4-1

Segmentation clients

- Dans le premier essai, nous avons utilisé la segmentation la plus présente dans la littérature.

	rfm	segments clients	nombre de clients	activité	Conseil d'action
0	1-1-1	Best Customers	1482	Les clients qui ont acheté le plus récemment, le plus souvent et qui dépensent le plus.	Pas d'incitation aux prix, Nouveaux produits et programmes de fidélité
1	X-1-X	Loyal Customers	4492	Les clients ayant acheté le plus récemment	Vendre des produits de plus grande valeur. Demandez des commentaires. Engagez-les.
2	X-X-1	Big Spenders	8847	Les clients qui dépensent le plus	Commercialisez vos produits les plus chers.
3	3-1-1	Almost Lost	1365	N'ont pas acheté depuis un certain temps, mais ont acheté fréquemment et ont dépensé le plus.	Offrez d'autres produits pertinents et des rabais spéciaux.
4	4-1-1	Lost Customers	1215	N'ont pas acheté depuis longtemps, mais ont acheté fréquemment et ont dépensé le plus.	Incitations à des prix agressifs
5	4-4-4	Lost Cheap Customers	5681	Dernier achat il y a longtemps, acheté peu et dépensé peu.	Ne passez pas trop de temps à essayer de vous ré-acquérir.
6	autres	autres	71014		Ajuster les intervales pour le réduire si important

- Usage de la distribution des inter-quartiles pour équilibrer les segments(satisfaction client)

Recency X Frequency

	1	2	3	4
1	3003	0	0	20779
2	2971	0	0	20347
3	2958	0	0	20605
4	2741	0	0	20692

Recency X Monetary_value

	1	2	3	4
1	5945	6065	5790	5982
2	5826	5892	5634	5966
3	5888	5833	6135	5707
4	5861	5696	5894	5982

Frequency X Monetary_value

	1	2	3	4
1	5504	3113	2022	1034
2	0	0	0	0
3	0	0	0	0
4	18016	20373	21431	22603

Segmentation clients

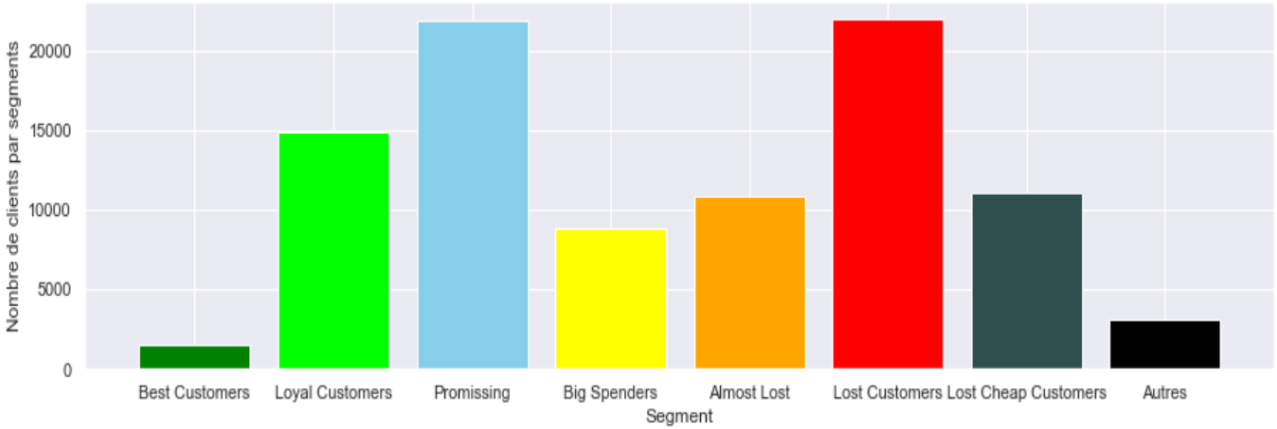
- Nouvelle segmentation avec les RFM score suivants :

  Best customers : ["1-1-1"]
  Loyal Customers : ["1-1-2", "1-1-3", "1-1-4", "2-1-1", "2-1-2", "2-1-3", "2-1-4", "1-4-2", "2-4-2"]
  Promissing : ["1-4-3", "1-4-4", "2-4-3", "2-4-4"]
  Big Spenders : ["1-2-1", "1-3-1", "1-4-1", "2-3-1", "2-4-1"]
  Almost Lost : ["3-1-1", "3-4-1", "3-4-2"]
  Lost Customers : ["3-4-3", "3-4-4", "4-1-1", "4-4-1", "4-4-2"]
  Lost Cheap Customers : ["4-4-4", "4-4-3"]

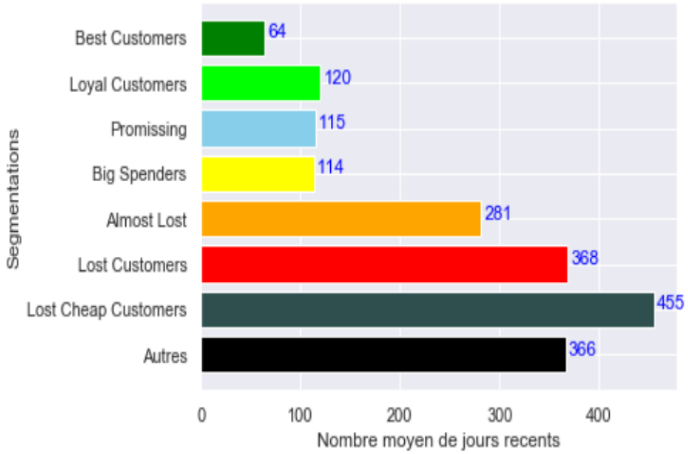
	segment	nombre de clients	description	marketing
0	Best Customers	1482	Les clients qui ont acheté le plus récemment, le plus souvent et qui dépensent le plus.	Pas d'incitation aux prix, Nouveaux produits et programmes de fidélité
1	Loyal Customers	14888	Les clients ayant acheté le plus récemment	Vendre des produits de plus grande valeur. Demandez des commentaires. Engagez-les.
2	Promissing	21883	Des acheteurs récents, mais qui n'ont pas beaucoup dépensé	offrir des essais gratuits
3	Big Spenders	8847	Les clients qui dépensent le plus	Commercialisez vos produits les plus chers.
4	Almost Lost	10869	N'ont pas acheté depuis un certain temps, mais ont acheté fréquemment et ont dépensé le plus.	Offrez d'autres produits pertinents et des rabais spéciaux.
5	Lost Customers	21958	N'ont pas acheté depuis longtemps, mais ont acheté fréquemment et ont dépensé le plus.	Incitations à des prix agressifs
6	Lost Cheap Customers	11050	Dernier achat il y a longtemps, acheté peu et dépensé peu.	Ne passez pas trop de temps à essayer de vous ré-acquérir.
7	Autres	3119		Ajuster les intervales pour le réduire si important

Visualisation des segments

- Distribution de clients par segment.

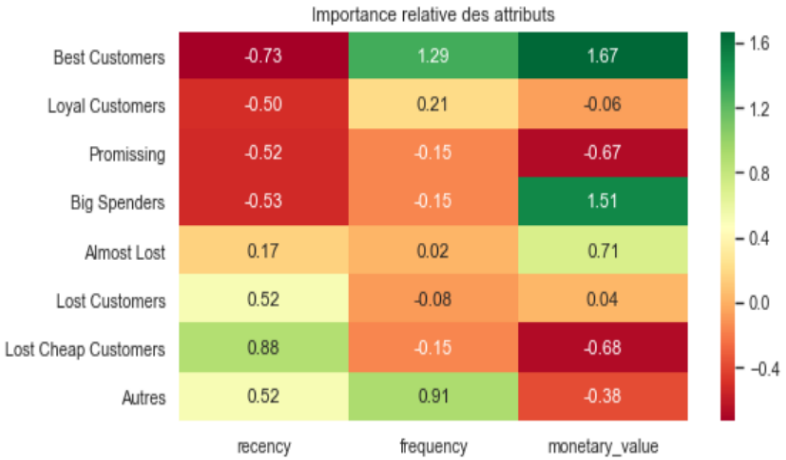


- Nombre moyen de jours écoulé du dernier achat par segment.



Importance relative des variables aux segments

- Compréhension rapide des segments avec cette visualisation.



- Le résultat est assez satisfaisant pour tous les segments hormis les segments **Loyal Customers**, **Almost lost**, **Lost Customers** biaisés par le comportement de la variable **fréquency**.

Sommaire

Algorithme de K-Means

- L'algorithme de **K-Means** est utilisé pour le clustering des variables numériques.
- Nous avons utilisé le **One-hot-encoding** pour l'encodage des catégorielles.

Algorithme 1 Algorithme K-Means

1: Choisir aléatoirement K points(centre de gravité) de l'hyper-plan(dimension correspondant au nombre de features)
2: Calculer la distance euclidienne de chaque observation aux centres de gravité
3: Affecter chaque observation au centroïde de distance minimale
4: Mise à jour des centroïdes par la moyenne des observations de leur cluster

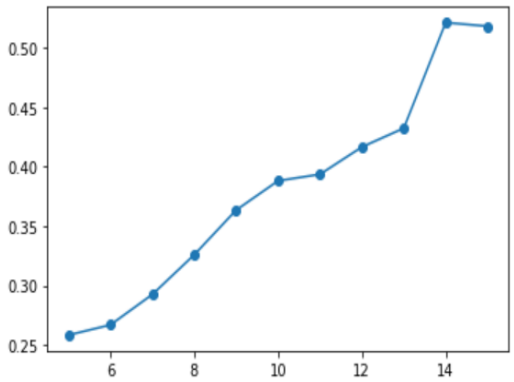
Mise en place

- L'information de la RFM est porté uniquement par la variable M car la majorité des clients ont effectué un seul achat.
- Donc en plus des variables R-F-M(log1p(M)), nous avons ajouté à notre bunch des features les variables : état, densité de population de l'état, catégories de produits, score de revue d'un produit acheté, le nombre de produits achetés par client.
- Les variables retenues qui améliorent la silhouette RFM : état, densité de population de l'état, le nombre de produits achetés par client.

Hyper-paramètre K de clusters

- Recherche du nombre optimal K de clusters ; nous utilisons un critère de forme : le coefficient de silhouette.

```
Le nombre de clusters optimal est : 14
La valeur de la silhouette est de : 0.5214692570237116
[<matplotlib.lines.Line2D at 0x7f1501207208>]
```

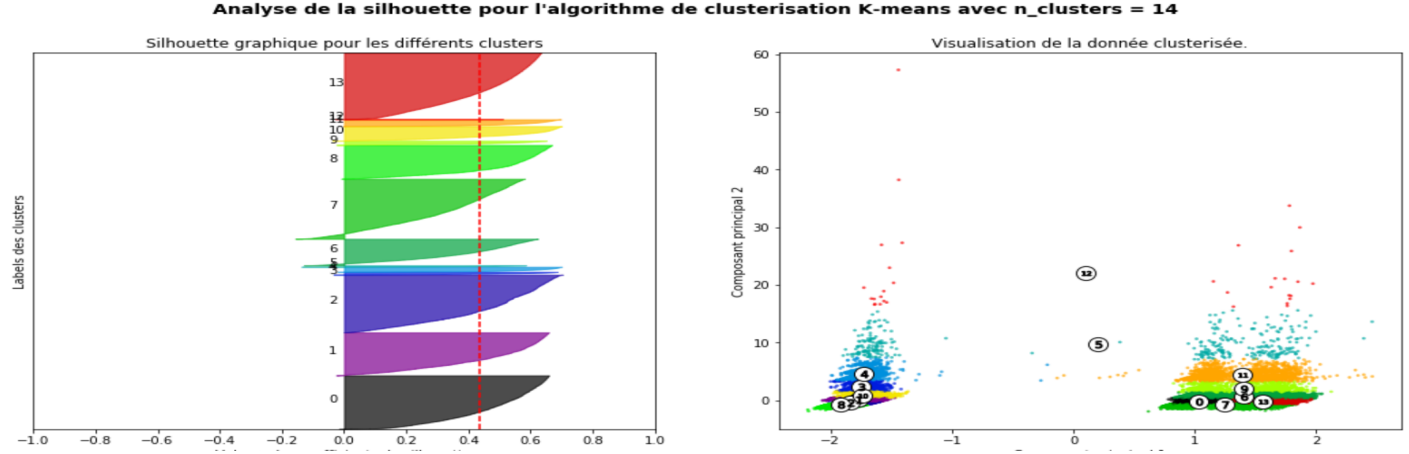


- L'hyperparamètre est K = 14

Visualisation PCA et TSNE

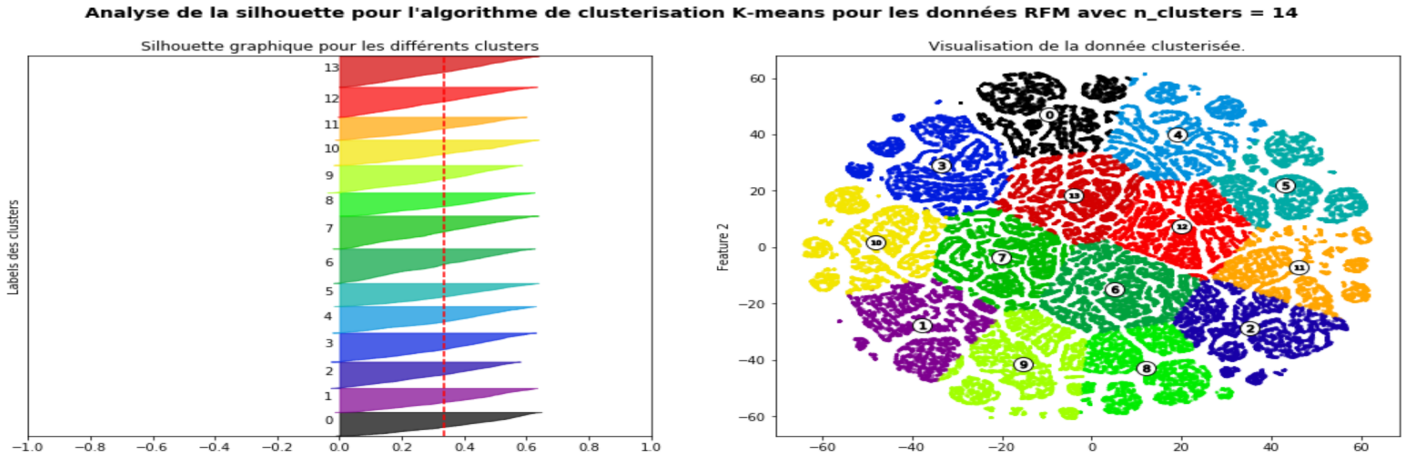
- Visualisation PCA : variance expliquée par 0.126

Pour n_clusters = 14 , Le score de la silhouette moyenne est : 0.4362573911195213



- Visualisation TSNE.

Pour n_clusters = 14 , Le score de la silhouette moyenne est : 0.33455202



Statistiques : centroïdes clusters

- Statistiques sur les clusters obtenus :

	recency	frequency	monetary_value	size
	mean	mean	mean	
cluster_appartenance				
0	0.067151	-0.161146	0.026390	11684.0
1	0.070968	-0.070336	0.013764	3426.0
2	1.068827	-0.161146	-0.199652	4173.0
3	-0.624377	-0.161146	0.121485	6498.0
4	0.089379	-0.145897	-0.004249	5022.0
5	-0.067002	-0.161146	-0.116804	38038.0
6	0.170898	-0.026756	0.216970	926.0
7	0.035055	-0.056346	0.061879	3197.0
8	0.087519	-0.159001	0.172256	6695.0
9	-0.015755	-0.080021	0.027911	4720.0
10	0.013553	-0.085818	-0.041265	4702.0
11	-0.117602	4.774786	0.956548	2688.0
12	0.034184	-0.037275	-0.008281	1932.0
13	0.185986	-0.015738	0.272557	395.0

- Identification des clusters.

	segment	matching_cluster	number of customers	segment description	marketing action
0	Best Customers	11	2688.0	Les clients qui ont acheté le plus récemment, ...	Pas d'incitation aux prix, Nouveaux produits e...
1	Loyal Customers	6	926.0	Les clients ayant acheté le plus récemment	Vendre des produits de plus grande valeur. Dem...
2	Promissing	3	6498.0	Des acheteurs récents, mais qui n'ont pas beau...	offrir des essais gratuits
3	Big Spenders	13	395.0	Les clients qui dépensent le plus	Commercialisez vos produits les plus chers.
4	Almost Lost	0	11684.0	N'ont pas acheté depuis un certain temps, mais...	Offrez d'autres produits pertinents et des rab...
5	Lost Customers	8	6695.0	N'ont pas acheté depuis longtemps, mais ont ac...	Incitations à des prix agressifs
6	Lost Cheap Customers	2	4173.0	Dernier achat il y a longtemps, acheté peu et ...	Ne passez pas trop de temps à essayer de vous ...

Sommaire

Diagramme de flux graphique

- Pour la maintenance, nous avons analysé la stabilité des segments au cours du temps qui consiste à étudier les migrations des clients entre les segments pour une période donnée.
- Nous avons pour ce faire, considéré les trois derniers mois

```
{0: datetime.datetime(2018, 9, 3, 0, 0),
 1: datetime.datetime(2018, 8, 3, 0, 0),
 2: datetime.datetime(2018, 7, 3, 0, 0),
 3: datetime.datetime(2018, 6, 3, 0, 0)}
```

- Les résultats obtenus sont représentés par le diagramme de **Sankey**

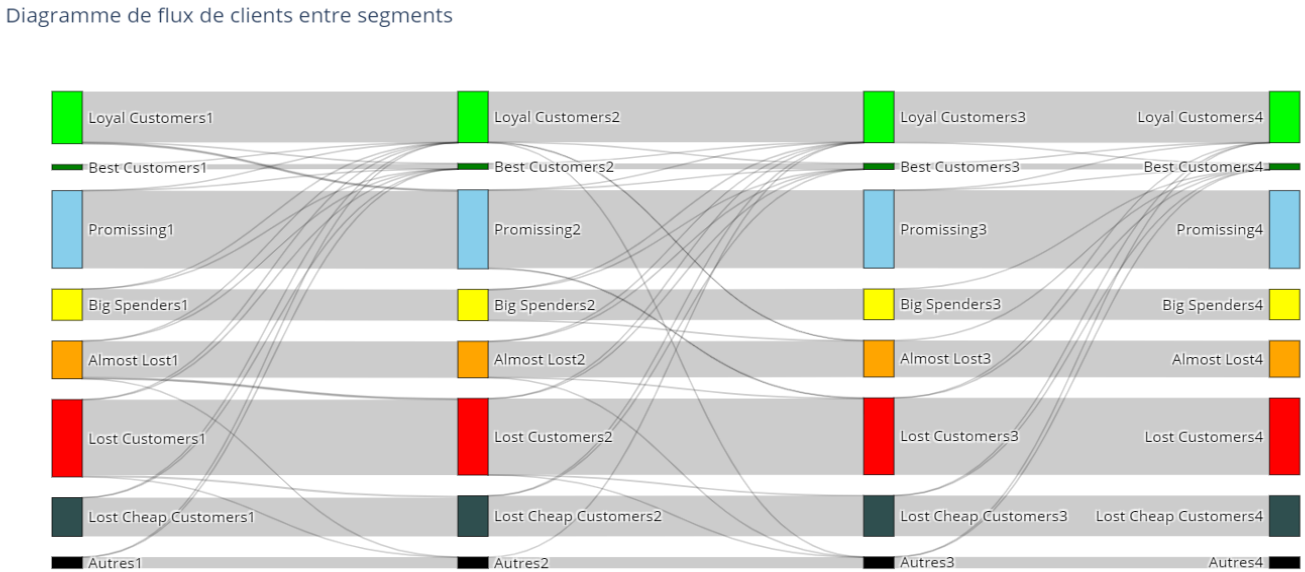


Diagramme de flux en chiffres

	Best Customers (2018, 7, 3)	Loyal Customers (2018, 7, 3)	Promissing (2018, 7, 3)	Big Spenders (2018, 7, 3)	Almost Lost (2018, 7, 3)	Lost Customers (2018, 7, 3)	Lost Cheap (2018, 7, 3)	Autres (2018, 7, 3)
Best Customers (2018, 6, 3)	1203	2	0	0	0	0	0	0
Loyal Customers (2018, 6, 3)	32	11717	288	0	0	0	0	0
Promissing (2018, 6, 3)	18	33	17825	0	0	0	0	0
Big Spenders (2018, 6, 3)	19	22	0	7155	0	0	0	0
Almost Lost (2018, 6, 3)	12	2	0	0	8373	300	0	3
Lost Customers (2018, 6, 3)	19	11	0	0	0	17313	441	3
Lost Cheap (2018, 6, 3)	4	7	0	0	0	0	8884	0
Autres (2018, 6, 3)	6	4	0	0	0	0	0	2567

	Best Customers (2018, 8, 3)	Loyal Customers (2018, 8, 3)	Promissing (2018, 8, 3)	Big Spenders (2018, 8, 3)	Almost Lost (2018, 8, 3)	Lost Customers (2018, 8, 3)	Lost Cheap (2018, 8, 3)	Autres (2018, 8, 3)
Best Customers (2018, 7, 3)	1306	7	0	0	0	0	0	0
Loyal Customers (2018, 7, 3)	19	11683	0	0	82	0	0	14
Promissing (2018, 7, 3)	14	28	17950	0	0	121	0	0
Big Spenders (2018, 7, 3)	12	33	0	7064	46	0	0	0
Almost Lost (2018, 7, 3)	13	1	0	0	8313	33	0	13
Lost Customers (2018, 7, 3)	13	10	0	0	0	17540	43	7
Lost Cheap (2018, 7, 3)	2	6	0	0	0	0	9317	0
Autres (2018, 7, 3)	0	4	0	0	0	0	0	2569

	Best Customers (2018, 9, 3)	Loyal Customers (2018, 9, 3)	Promissing (2018, 9, 3)	Big Spenders (2018, 9, 3)	Almost Lost (2018, 9, 3)	Lost Customers (2018, 9, 3)	Lost Cheap (2018, 9, 3)	Autres (2018, 9, 3)
Best Customers (2018, 8, 3)	1366	13	0	0	0	0	0	0
Loyal Customers (2018, 8, 3)	12	11760	0	0	0	0	0	0
Promissing (2018, 8, 3)	6	20	17924	0	0	0	0	0
Big Spenders (2018, 8, 3)	9	0	0	7055	0	0	0	0
Almost Lost (2018, 8, 3)	14	0	0	0	8427	0	0	0
Lost Customers (2018, 8, 3)	14	4	0	0	0	17676	0	0
Lost Cheap (2018, 8, 3)	5	8	0	0	0	0	9347	0
Autres (2018, 8, 3)	3	1	0	0	0	0	0	2599

Observations sur le spectre de flux

- Il suffit juste d'un nouvel achat pour presque passé dans le lot de "Best Customers".

customer_unique_id	recency (2018, 8, 3)	frequency (2018, 8, 3)	monetary_value (2018, 8, 3)	RFMScore (2018, 8, 3)	recency (2018, 9, 3)	frequency (2018, 9, 3)	monetary_value (2018, 9, 3)	RFMScore (2018, 9, 3)
cd6c68c5fad15e0a5a5c1150546704e0	417	1	572.00	4-4-1	20	2	636.90	1-1-1
46ed126bcf1df6e195dbc63d7c320983	438	1	199.90	4-4-1	12	2	229.80	1-1-1
3fe3e628c6c7a15ae96416826a4c5952	374	1	119.99	4-4-2	15	2	368.99	1-1-1
d08c29302907086e8fe823369542f3ae	383	2	388.98	4-1-1	10	4	418.78	1-1-1
71a92fd3087501bcfba6e6e1ef7e8fd7	449	1	198.00	4-4-1	23	2	223.00	1-1-1
ee04cc9bca4c9198bec5c54c2542dd3b	408	1	79.90	4-4-3	27	2	206.89	1-1-1
5eefb861d4921a3e628bbc65c50a480a	515	1	45.99	4-4-4	28	2	195.89	1-1-1
2e49a3bbeb76297ee0ff49df39c2456c	468	1	72.90	4-4-3	14	2	171.90	1-1-1
4702ba5faa8283e0f6b6a545cdaf8a9f	444	1	59.00	4-4-3	26	2	219.00	1-1-1
4cfa5155cf7cff8eb15e0b12041d058e	363	1	59.90	4-4-3	29	2	189.89	1-1-1

- Cela montre une fois de plus la sensibilité du dataset lorsque lorsqu'on a des fréquences peu variables.

Sommaire

Observations et discussions

- L'analyse RFM nous permet d'identifier les différents clients en particulier les meilleurs(best customer).
- Cette analyse peut être utilisé dans différent types de métiers tel que **call center**, etc.
- La principale difficulté à l'analyse RFM est lorsque les clients ont à majorité un seul achat comme le dataset d'étude.
- L'analyse est fortement liée à la satisfaction client.

Observations et discussions

- Les algorithmes de clustering permettent de regrouper les données proches mais posent néanmoins un problème d'interprétation(par exemple quel est le segment de best customer).
- Ces algorithmes sont meilleurs lorsque le nombre de variables est important et donc les combinaisons moins exhaustives comme RFM.
- L'analyse RFM est meilleur que les algorithmes automatiques pour cette tâche. Ces derniers peuvent néanmoins être combinés pour une analyse encore plus pertinentes.
- Le diagramme de flux donne un bon spectre de la mise à jour de la segmentation RFM.

Perspectives

- Appliquer la segmentation des quintiles telle que réalisée dans l'outil d'analyse RFM de PUTLER et comparer avec nos résultats.
- Appliquer l'analyse RFM sur un dataset plus réel(fréquence plus représentative) pour une meilleure visualisation des difficultés en entreprise.