

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

Présenté par : NAMA NYAM Guy Anthony

Mentor : Julien Hendrick

22 Janvier 2020

**OPENCLASSROOMS**

# Sommaire

- 1 Introduction
- 2 Source de données
- 3 Regroupement des catégories de produits
- 4 Segmentation avec l'analyse RFM
- 5 Segmentation automatique : K-Means
- 6 Maintenance
- 7 Conclusion

# 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

- 1 Introduction
- 2 Source de données
- 3 Regroupement des catégories de produits
- 4 Segmentation avec l'analyse RFM
- 5 Segmentation automatique : K-Means
- 6 Maintenance
- 7 Conclusion

# Description

- Base de données(`olist_customers_dataset.csv`) anonymisée du site d'e-commerce **Olist**, téléchargeable ([https ://www.kaggle.com/olistbr/brazilian-ecommerce](https://www.kaggle.com/olistbr/brazilian-ecommerce)).
- Quatre(4) tables retenues sur les huit(8) initiales pour déduire le comportement RFM de chaque client et regrouper les noms des catégories de produits.
- Les tables retenues : **olist\_customers\_dataset**, **olist\_orders\_dataset**, **olist\_order\_items\_dataset** , **olist\_products\_dataset**

# Nettoyage de la donnée et fusion

- La suppression des transactions non approuvées.

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

168

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

- La suppression des transactions sans catégories produits

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

618

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

- Jointure naturelle entre les lignes des différentes tables.

```
donnees.head()
```

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

```
donnees.shape
```

(111033, 4)



# Sommaire

- 1 Introduction
- 2 Source de données
- 3 Regroupement des catégories de produits**
- 4 Segmentation avec l'analyse RFM
- 5 Segmentation automatique : K-Means
- 6 Maintenance
- 7 Conclusion

# Algorithme de K-Modes

- L'algorithme de **K-Modes** est utilisé pour le clustering des variables catégorielles.
- Cet algorithme est efficace sur des données binaires ; donc nous utilisons le **One-hot-encoding** pour l'encodage des catégorielles.
- Algorithme disponible sous le module tiers **kmodes.kmodes.KModes**

---

## Algorithme 1 Algorithme K-Modes

---

- 1: Choisir aléatoirement K observations dans les données disponibles à considérer comme centroïdes
  - 2: Calculer la mesure de dissimilarité des centroïdes à chaque ligne ou observation.
  - 3: Affecter chaque ligne au centroïde de distance minimale
  - 4: Mise à jour des centroïdes
-

# K-Modes : pratique

- Nous utilisons la variable des noms de catégories produits.

```
print(len(donnees["product_category_name"].unique()))
```

73

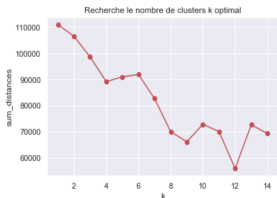
- Encodage : One-hot-encoding

```
OH_cols.head()
```

|   | moveis_escritorio | utilidades_domesticas | casa_conforto | esporte_lazer | informatica_acessorios | brinquedos | moveis_decoracao | automotivo | climatizacao | tele |
|---|-------------------|-----------------------|---------------|---------------|------------------------|------------|------------------|------------|--------------|------|
| 0 | 1.0               | 0.0                   | 0.0           | 0.0           | 0.0                    | 0.0        | 0.0              | 0.0        | 0.0          | 0.0  |
| 1 | 1.0               | 0.0                   | 0.0           | 0.0           | 0.0                    | 0.0        | 0.0              | 0.0        | 0.0          | 0.0  |
| 2 | 1.0               | 0.0                   | 0.0           | 0.0           | 0.0                    | 0.0        | 0.0              | 0.0        | 0.0          | 0.0  |
| 3 | 1.0               | 0.0                   | 0.0           | 0.0           | 0.0                    | 0.0        | 0.0              | 0.0        | 0.0          | 0.0  |
| 4 | 1.0               | 0.0                   | 0.0           | 0.0           | 0.0                    | 0.0        | 0.0              | 0.0        | 0.0          | 0.0  |

5 rows × 11 columns

- Minimiser la somme de distance des points à leurs clusters.



# Résultat et discussions

## Résultat :

|                                      |  |   |                                      |
|--------------------------------------|--|---|--------------------------------------|
| cluster 0:<br>cluster sans catégorie | cluster 6:<br>utilidades_domesticas        | cluster 0:<br>cluster sans catégorie                                | cluster 6:<br>automotivo             |
| cluster 1:<br>ferramentas_jardim     | cluster 7:<br>beleza_saude                 | cluster 1:<br>cama_mesa_banho                                       | cluster 7:<br>utilidades_domesticas  |
| cluster 2:<br>cama_mesa_banho        | cluster 8:<br>perfumaria                   | cluster 2:<br>brinquedos<br>telefonica<br>beleza_saude<br>papeleria | cluster 8:<br>informatica_acessorios |
| cluster 3:<br>bebes                  | cluster 9:<br>moveis_decoracao             | cluster 3:<br>brinquedos  | cluster 9:<br>climatizacao           |
| cluster 4:<br>moveis_escritorio      | cluster 10:<br>cool_stuff                  | cluster 4:<br>relogios_presentes                                    | cluster 10:<br>eletronicos           |
| cluster 5:<br>esporte_lazer          | cluster 11:<br>casa_conforto<br>telefonica | cluster 5:<br>eletrodomesticos                                      | cluster 11:<br>moveis_decoracao      |

- Différents regroupements dépendant de l'initialisation.
- 12/73** noms des catégories de produits apparaissent en sortie : ceci s'explique par le fait que la plupart des clients n'ont qu'acheté une seule catégorie de produits.

# Sommaire

- 1 Introduction
- 2 Source de données
- 3 Regroupement des catégories de produits
- 4 Segmentation avec l'analyse RFM**
- 5 Segmentation automatique : K-Means
- 6 Maintenance
- 7 Conclusion

# 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 :

```
##### Calculs : recency, frequency, monetary #####
rfmTable = donnees.groupby('customer_unique_id').agg({"order_purchase_timestamp": lambda x: (NOW - x.max()).days,
                                                       "customer_unique_id": lambda x: len(x), "price": lambda x: sum(x)})

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

rfmTable.head()
```

|                                  | recency | frequency | monetary_value |
|----------------------------------|---------|-----------|----------------|
| customer_unique_id               |         |           |                |
| 0000366f3b9a7992bf8c76cfd3221e2  | 115     | 1         | 129.90         |
| 0000b849f77a49e4a4ce2b2a4ca5be3f | 118     | 1         | 18.90          |
| 0000f46a3911fa3c0805444483337064 | 541     | 1         | 69.00          |
| 0000f6ccb0745a6a4b88665a16c9f078 | 325     | 1         | 25.99          |
| 0004aac84e0df4da2b147fca70cf8255 | 292     | 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
quantiles = rfmTable.quantile(q=[0.25, 0.5, 0.75])
print(quantiles)
```

```
      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', quantiles))
segmented_rfm["f_quartile"] = segmented_rfm["frequency"].apply(FMScore, args=('frequency', quantiles))
segmented_rfm["m_quartile"] = segmented_rfm["monetary_value"].apply(FMScore, args=('monetary_value', quantiles))

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

segmented_rfm.head()
```

|                                  | recency | frequency | monetary_value | r_quartile | f_quartile | m_quartile | RFMScore |
|----------------------------------|---------|-----------|----------------|------------|------------|------------|----------|
| customer_unique_id               |         |           |                |            |            |            |          |
| 0000366f3b9a7992bfc76cfd3221e2   | 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 intervalles 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"]

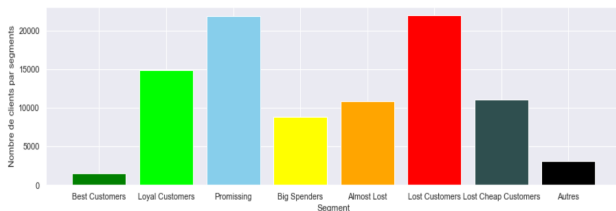
Lost Customers : ["3-4-3", "3-4-4", "4-1-1", "4-4-1", "4-4-2"]

Lost Cheap Customers : ["4-4-4", "4-4-3"]

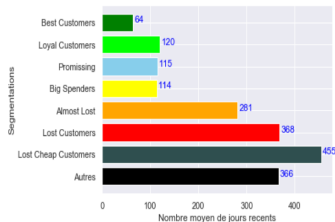
|   | segment              | nombre de clients | description   | marketing  |
|---|----------------------|-------------------|---|--|
| 0 | Best Customers       | 1482              | Les clients qui ont acheté le plus récemment, le plus souvent et qui dépensent le plus.       | Pas d'incitation aux prix, 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 intervalles pour le réduire si important                               |

# Visualisation des segments

- Distribution de clients par segment.

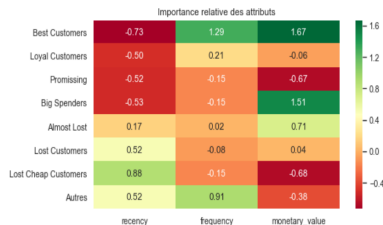


- Nombre moyen de jours écoulé du dernier achat par segment.



# Importance relative des variables aux segments

- Compréhension rapide des segments avec cette visualisation.



- Le résultat est assez satisfaisant dans l'ensemble pour tous les segments hormis le segment **Loyal Customers** biaisé par le comportement de la variable **fréquency**.

# Sommaire

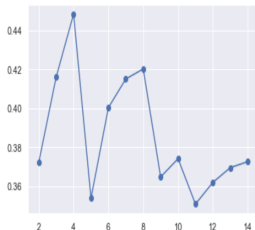
- 1 Introduction
- 2 Source de données
- 3 Regroupement des catégories de produits
- 4 Segmentation avec l'analyse RFM
- 5 Segmentation automatique : K-Means**
- 6 Maintenance
- 7 Conclusion

# Cadre théorique

- Le jeu de données est constitué des clients et les variables R, F, M précédentes.
- Recherche du nombre optimal K de clusters ; nous utilisons un critère de forme : le coefficient de silhouette.

Le nombre de clusters optimal est : 4

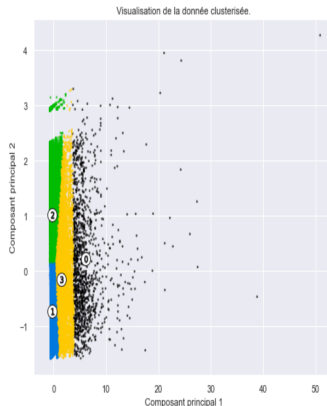
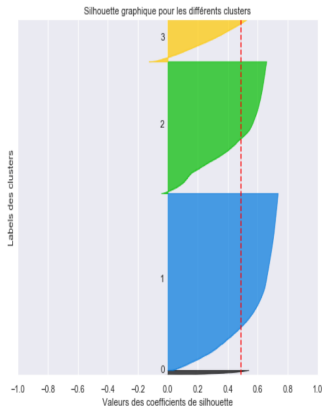
```
[<matplotlib.lines.Line2D at 0x1d9b98ce198>]
```



# Visualisation des clusters avec PCA

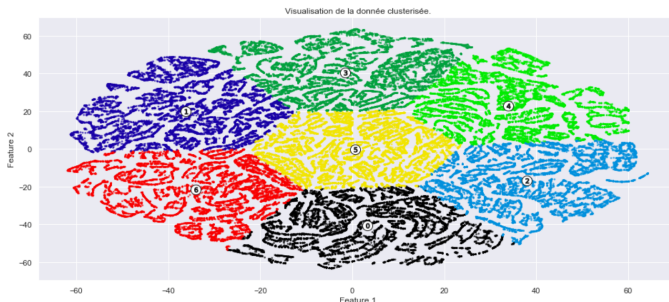
- Visualisation proportionnée des clusters par rapport à leurs distributions.

Analyse de la silhouette pour l'algorithme de clusterisation K-means pour les données RFM avec  $n\_clusters = 4$



# Algorithme K-Means avec TSNE : $K = 7$

- L'objectif est de comparer les segments obtenus dans l'analyse RFM et les segments recouverts automatiquement.



- L'étape suivante consistera à faire la correspondance des clusters obtenus aux segments de l'analyse RFM.

# Algorithme K-Means avec TSNE : $K = 7$

- Statistiques sur les clusters obtenus :

|                      | recency    | frequency | monetary_value | size  |
|----------------------|------------|-----------|----------------|-------|
|                      | mean       | mean      | mean           |       |
| cluster_appartenance |            |           |                |       |
| 0                    | 130.911588 | 1.000000  | 49.279659      | 13505 |
| 1                    | 250.551502 | 1.379904  | 386.299155     | 13485 |
| 2                    | 249.884004 | 1.167034  | 97.353859      | 13147 |
| 3                    | 359.584852 | 1.466329  | 160.210115     | 12622 |
| 4                    | 218.111238 | 1.305476  | 137.190502     | 12217 |
| 5                    | 230.364647 | 1.000000  | 78.423570      | 15744 |
| 6                    | 260.536334 | 1.000000  | 98.899655      | 13376 |

- Intersection clients entre les segments du K-Means et RFM.

|   | Best Customers | Loyal Customers | Promissing | Big Spenders | Almost Lost | Lost Customers | Lost Cheap Customers | Autres |
|---|----------------|-----------------|------------|--------------|-------------|----------------|----------------------|--------|
| 0 | 236            | 611             | 3519       | 848          | 2254        | 1679           | 4002                 | 32     |
| 1 | 855            | 1108            | 0          | 3919         | 2942        | 3273           | 1213                 | 0      |
| 2 | 10             | 2791            | 8246       | 154          | 859         | 87             | 0                    | 664    |
| 3 | 244            | 3114            | 0          | 2761         | 989         | 4700           | 1211                 | 172    |
| 4 | 0              | 3404            | 3497       | 0            | 1760        | 4286           | 0                    | 77     |
| 5 | 137            | 557             | 3062       | 666          | 1428        | 3969           | 2839                 | 325    |
| 6 | 0              | 3303            | 3559       | 499          | 637         | 3964           | 1785                 | 1849   |



# Sommaire

- 1 Introduction
- 2 Source de données
- 3 Regroupement des catégories de produits
- 4 Segmentation avec l'analyse RFM
- 5 Segmentation automatique : K-Means
- 6 Maintenance**
- 7 Conclusion

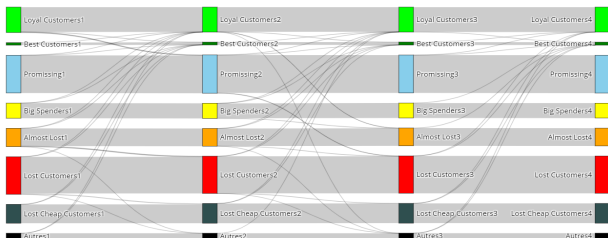
# Diagramme de flux graphique

- 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 de clients entre segments



# Diagramme de flux en chiffres

|                                 | Best Customers<br>(2018, 7, 3) | Loyal Customers<br>(2018, 7, 3) | Promissing<br>(2018, 7, 3) | Big Spenders<br>(2018, 7, 3) | Almost Lost<br>(2018, 7, 3) | Lost Customers<br>(2018, 7, 3) | Lost Cheap<br>(2018, 7, 3) | Autres<br>(2018, 7, 3) |
|---------------------------------|--------------------------------|---------------------------------|----------------------------|------------------------------|-----------------------------|--------------------------------|----------------------------|------------------------|
| Best Customers<br>(2018, 6, 3)  | 1203                           | 2                               | 0                          | 0                            | 0                           | 0                              | 0                          | 0                      |
| Loyal Customers<br>(2018, 6, 3) | 32                             | 11717                           | 286                        | 0                            | 0                           | 0                              | 0                          | 0                      |
| Promissing (2018,<br>6, 3)      | 18                             | 33                              | 17825                      | 0                            | 0                           | 0                              | 0                          | 0                      |
| Big Spenders<br>(2018, 6, 3)    | 19                             | 22                              | 0                          | 7155                         | 0                           | 0                              | 0                          | 0                      |
| Almost Lost (2018,<br>6, 3)     | 12                             | 2                               | 0                          | 0                            | 8373                        | 300                            | 0                          | 3                      |
| Lost Customers<br>(2018, 6, 3)  | 19                             | 11                              | 0                          | 0                            | 0                           | 17313                          | 441                        | 3                      |
| Lost Cheap (2018,<br>6, 3)      | 4                              | 7                               | 0                          | 0                            | 0                           | 0                              | 8684                       | 0                      |
| Autres (2018, 6, 3)             | 6                              | 4                               | 0                          | 0                            | 0                           | 0                              | 0                          | 2567                   |
|                                 | Best Customers<br>(2018, 8, 3) | Loyal Customers<br>(2018, 8, 3) | Promissing<br>(2018, 8, 3) | Big Spenders<br>(2018, 8, 3) | Almost Lost<br>(2018, 8, 3) | Lost Customers<br>(2018, 8, 3) | Lost Cheap<br>(2018, 8, 3) | Autres<br>(2018, 8, 3) |
| Best Customers<br>(2018, 7, 3)  | 1306                           | 7                               | 0                          | 0                            | 0                           | 0                              | 0                          | 0                      |
| Loyal Customers<br>(2018, 7, 3) | 19                             | 11683                           | 0                          | 0                            | 82                          | 0                              | 0                          | 14                     |
| Promissing (2018,<br>7, 3)      | 14                             | 28                              | 17950                      | 0                            | 0                           | 121                            | 0                          | 0                      |
| Big Spenders<br>(2018, 7, 3)    | 12                             | 33                              | 0                          | 7064                         | 46                          | 0                              | 0                          | 0                      |
| Almost Lost (2018,<br>7, 3)     | 13                             | 1                               | 0                          | 0                            | 8313                        | 33                             | 0                          | 13                     |
| Lost Customers<br>(2018, 7, 3)  | 13                             | 10                              | 0                          | 0                            | 0                           | 17540                          | 43                         | 7                      |
| Lost Cheap (2018,<br>7, 3)      | 2                              | 6                               | 0                          | 0                            | 0                           | 0                              | 9317                       | 0                      |
| Autres (2018, 7, 3)             | 0                              | 4                               | 0                          | 0                            | 0                           | 0                              | 0                          | 2569                   |
|                                 | Best Customers<br>(2018, 9, 3) | Loyal Customers<br>(2018, 9, 3) | Promissing<br>(2018, 9, 3) | Big Spenders<br>(2018, 9, 3) | Almost Lost<br>(2018, 9, 3) | Lost Customers<br>(2018, 9, 3) | Lost Cheap<br>(2018, 9, 3) | Autres<br>(2018, 9, 3) |
| Best Customers<br>(2018, 8, 3)  | 1366                           | 13                              | 0                          | 0                            | 0                           | 0                              | 0                          | 0                      |
| Loyal Customers<br>(2018, 8, 3) | 12                             | 11760                           | 0                          | 0                            | 0                           | 0                              | 0                          | 0                      |
| Promissing (2018,<br>8, 3)      | 6                              | 20                              | 17924                      | 0                            | 0                           | 0                              | 0                          | 0                      |
| Big Spenders<br>(2018, 8, 3)    | 9                              | 0                               | 0                          | 7055                         | 0                           | 0                              | 0                          | 0                      |
| Almost Lost (2018,<br>8, 3)     | 14                             | 0                               | 0                          | 0                            | 8427                        | 0                              | 0                          | 0                      |
| Lost Customers<br>(2018, 8, 3)  | 14                             | 4                               | 0                          | 0                            | 0                           | 17676                          | 0                          | 0                      |
| Lost Cheap (2018,<br>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<br>(2018, 8, 3) | frequency<br>(2018, 8, 3) | monetary_value<br>(2018, 8, 3) | RFMScore<br>(2018, 8, 3) | recency<br>(2018, 9, 3) | frequency<br>(2018, 9, 3) | monetary_value<br>(2018, 9, 3) | RFMScore<br>(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                    |
| 71a92fd3087501bcbfa6e6e1ef7e8fd7 | 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                    |
| 4702ba5faa8283e0f6b6a545cdfa8a9f | 444                     | 1                         | 59.00                          | 4-4-3                    | 26                      | 2                         | 219.00                         | 1-1-1                    |
| 4cfa5155c7cff8eb15e0b12041d058e  | 363                     | 1                         | 59.90                          | 4-4-3                    | 29                      | 2                         | 169.89                         | 1-1-1                    |

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

# Sommaire

- 1 Introduction
- 2 Source de données
- 3 Regroupement des catégories de produits
- 4 Segmentation avec l'analyse RFM
- 5 Segmentation automatique : K-Means
- 6 Maintenance
- 7 Conclusion

# Observations et discussions

- L'algorithme K-Modes qui permet de regrouper un ensemble de concept à l'exemple de LinkedIn pour identifier le métier associé à un ensemble de compétences.
- Cet algorithme pose des limites lorsqu'on a peu de diversification de concept pour l'ensemble des observations.
- 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 dans ce cas est meilleur que les algorithmes automatiques. 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 pour une meilleure visualisation des difficultés en entreprise.