Projet Ecommerce

Efkan TUREDI

Intro

Notre client, Olist, souhaite une segmentation client pour à des fins de marketings ciblés. Les livrables attendus par le clients sont:

- Une description actionable de notre segmentation et de sa logique
- Contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps.

Pour faire cette analyse, nous disposons d'un nombre élevé de fichiers .csv qu'il faudra nettoyer et analyser avec l'aide d'algorithme non-supervisés.

Le nettoyage

Nous avons des problèmes courants

- Certaines données sont en portugais
- Erreurs d’entrées dans les données (ex: string dans une colonne float)
- NaNs dans certaines colonnes
- Les catégories de produits sont trop granulaires, nous les avons simplifiées (par exemple catégories de produits)

Quelques lignes de codes pour illustration

```
master_df['product_category_name'].replace(['Unknown','portateis_cozinha_e_preparadores_de_alimentos'],'Other',inplace=True)
```

```
dict_categories = {
#home
'furniture_living_room' : 'Home',
'furniture_mattress_and_upholstery' : 'Home',
'furniture_bedroom' : 'Home',
'furniture_decor' : 'Home',
'bed_bath_table' : 'Home',
'kitchen_dining_laundry_garden_furniture' : 'Home',
'la_cuisine' : 'Home',
'home_confort' : 'Home',
'home_comfort_2' : 'Home',
'christmas_supplies' : 'Home',
....
#appliances
'small_appliances' : 'Appliances',
'small_appliances_home_oven_and_coffee' : 'Appliances',
'home_appliances_2' : 'Appliances',
'home_appliances' : 'Appliances',
'housewares' : 'Appliances',
....
```

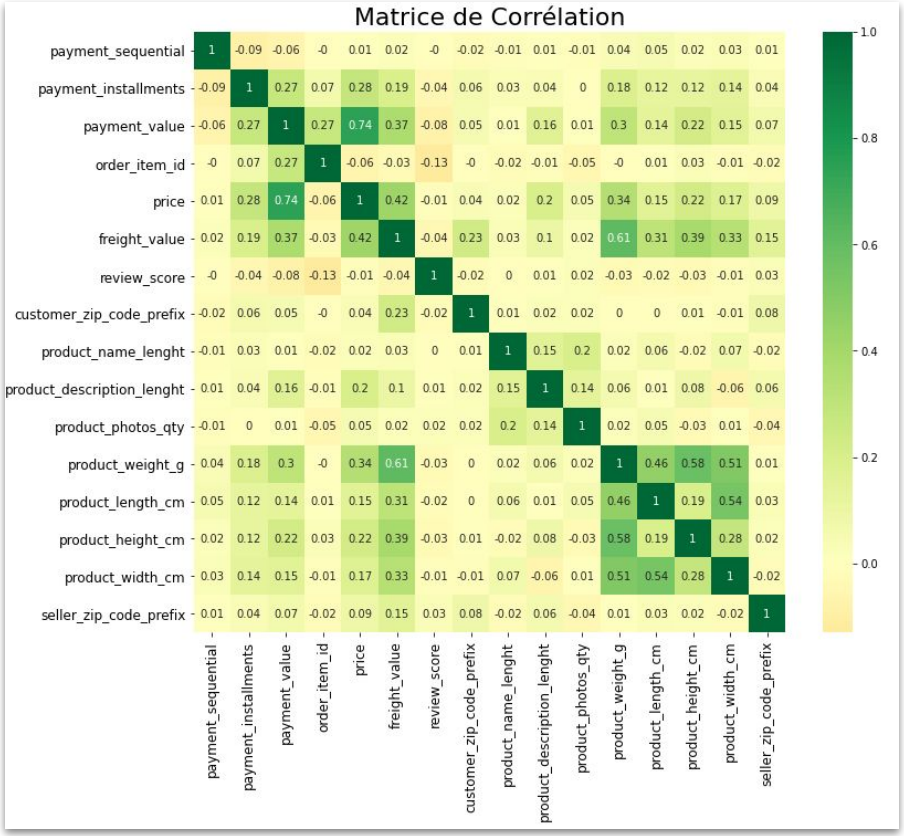
```
master_df
```

	order_id	payment_sequential	payment_type	payment_installments	payment_value
0	b81ef226f3fe1789b1e8b2acac839d17	1	credit_card	8	99.33
1	a9810da82917af2d9aefd1278f1dcfa0	1	credit_card	1	24.39
2	25e8ea4e93396b6fa0d3dd708e76c1bd	1	credit_card	1	65.71
3	ba78997921bbcdc1373bb41e913ab953	1	credit_card	8	107.78
4	ba78997921bbcdc1373bb41e913ab953	1	credit_card	8	107.78
...	...	...	...	...	...
119143	0406037ad97740d563a178ecc7a2075c	1	boleto	1	363.31
119144	7b905861d7c825891d6347454ea7863f	1	credit_card	2	96.80
119145	32609bbb3dd69b3c066a6860554a77bf	1	credit_card	1	47.77
119146	b8b61059626efa996a60be9bb9320e10	1	credit_card	5	369.54
119147	28bbae6599b09d39ca406b747b6632b1	1	boleto	1	191.58

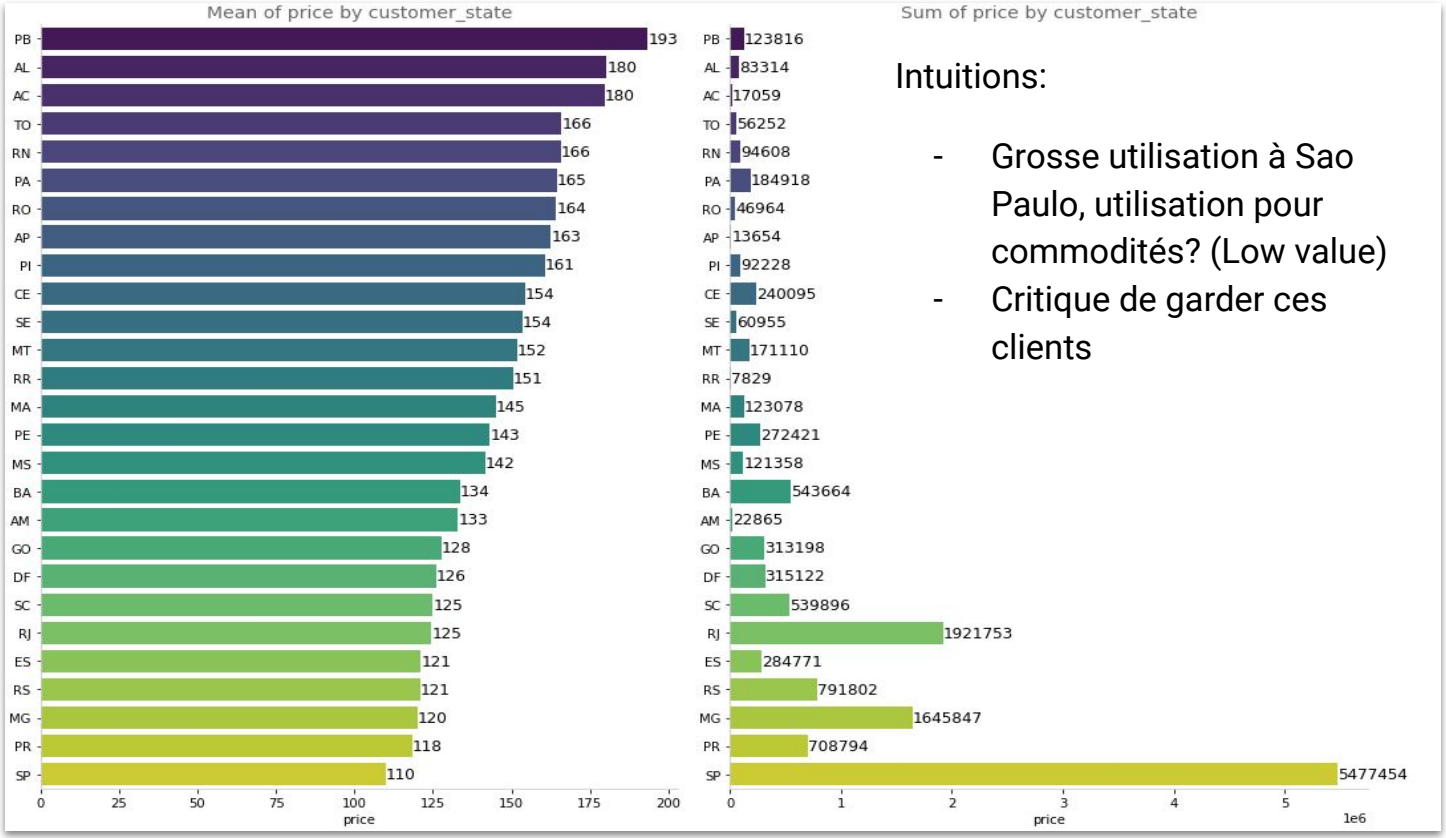
118315 rows × 39 columns

Analyse exploratoire

Matrice de correlation



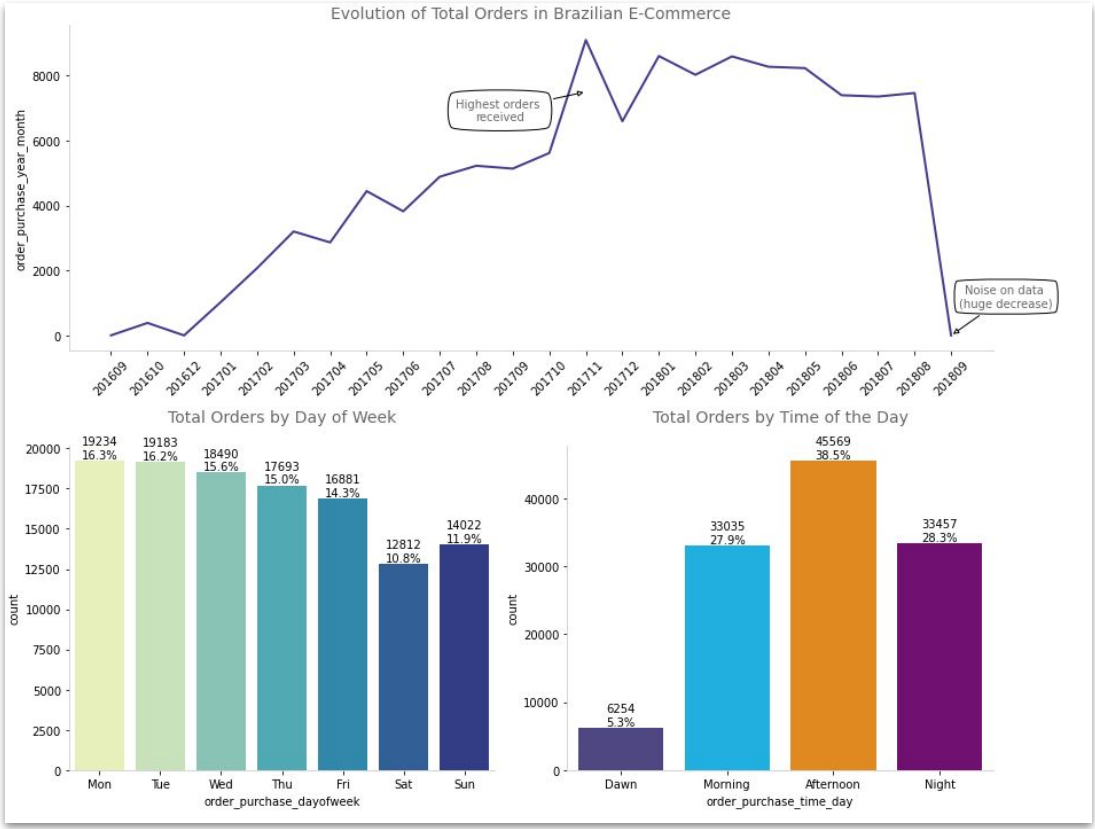
Quelques données sur les clients d'Olist



Intuitions:

- Grosse utilisation à Sao Paulo, utilisation pour commodités? (Low value)
- Critique de garder ces clients

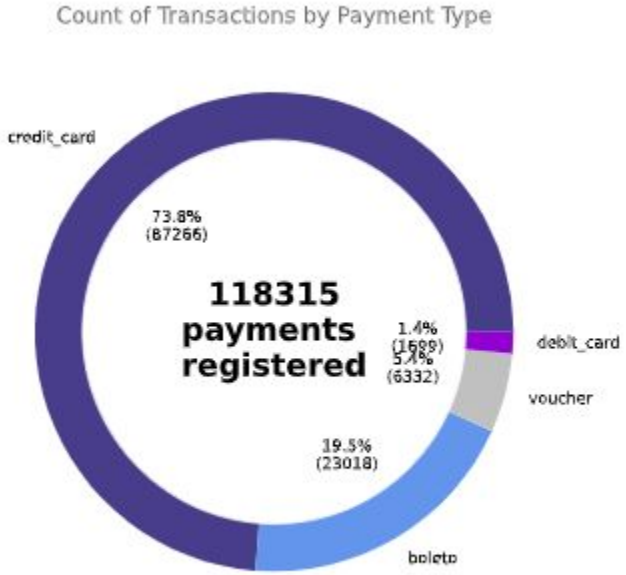
Quelques données sur les clients d'Olist

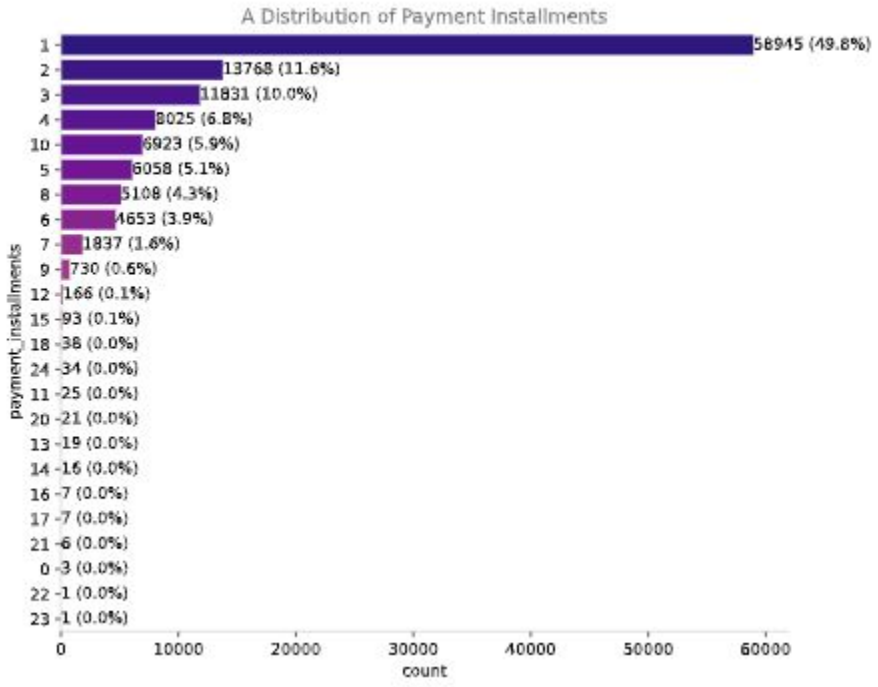


Intuitions:

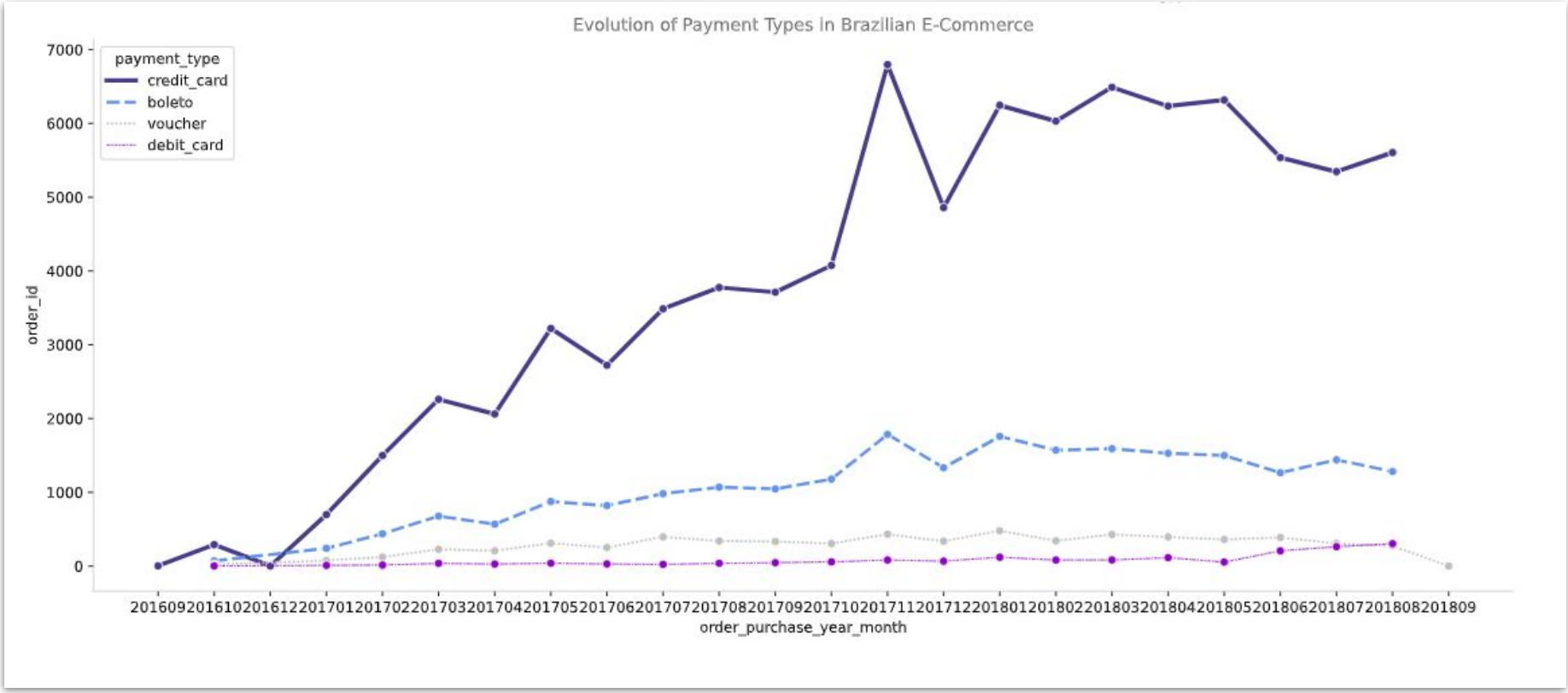
- Nombre total de commandes drivés essentiellement par l'augmentation du nombre de clients plutot qu'une aug. du # de commandes par clients
- Le plus gros des utilisations a lieu en début de semaine, dans l'après midi

Quelques données sur les clients d'Olist

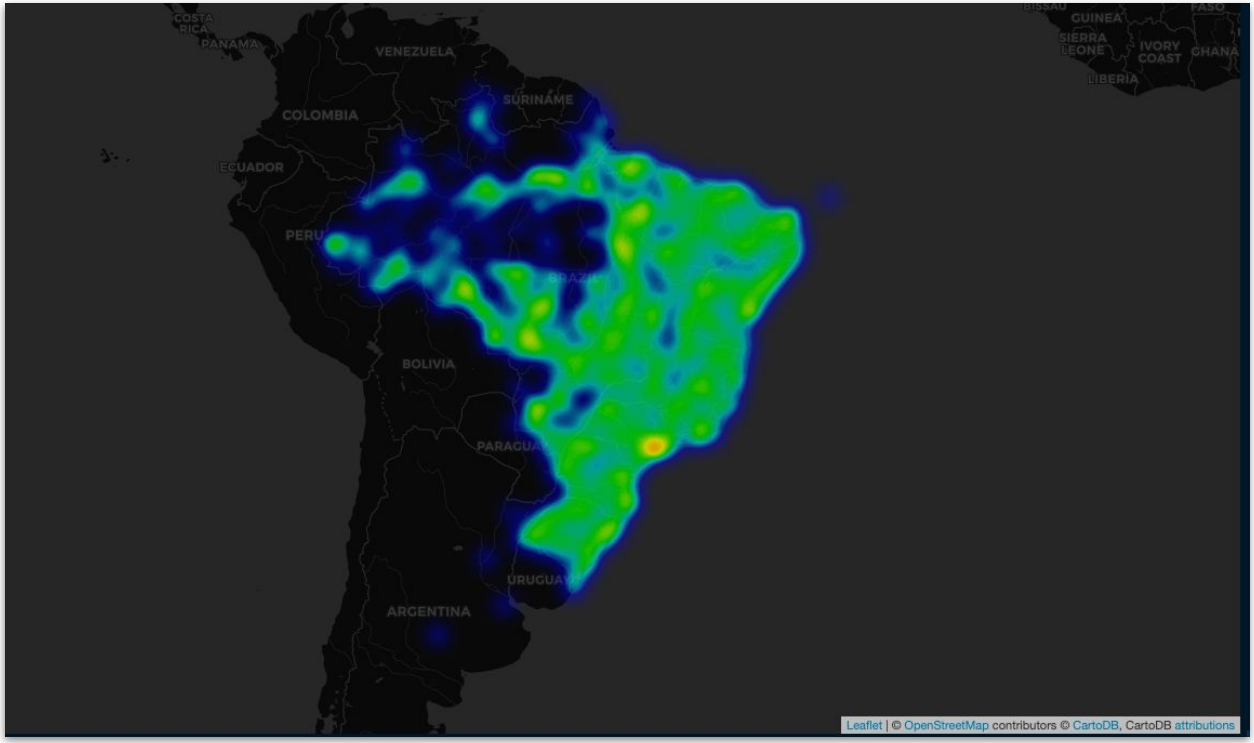




Quelques données sur les clients d'Olist



Quelques données sur les clients d'Olist



Intuitions:

- Clients largement concentré sur les côtes, très urbanisés
- Le reste du pays est plus compliqué notamment au niveau logistique

Analyse RFM

Feature engineering nécéssaire

Nous approchons cette exercice avec une approche RFM (Recency, Frequency, Monetary). Nous avons donc un travail de feature engineering avant de pouvoir utiliser nos modèles ML. Nous avons donc utilisé les features suivants:

- Total_price: somme des prix des articles achetés (CA total provenant de X)
- Order_item_id: Nombre d'articles acheté par le client
- Average_price: Prix moyen de l'article acheté par le client
- Recency: # de jours depuis dernier achat

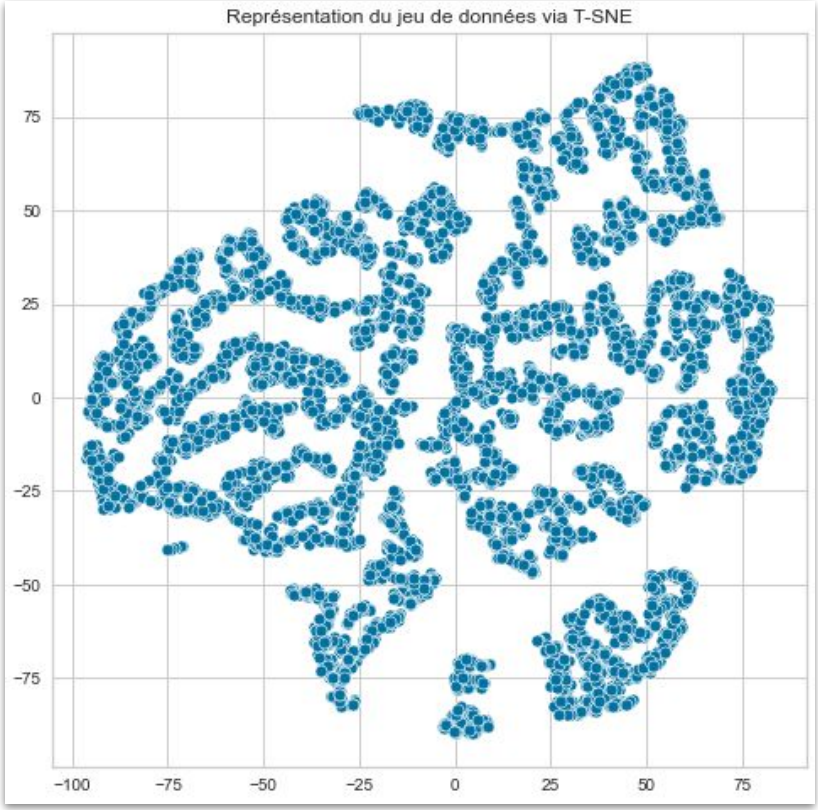
Méthodologies d’optimisation des hyperparamètres

Nous avons utilisé le package Yellowbrick pour voir les paramètre optimaux dans le cas d’un K-means. Nous avons regardé plusieurs indicateurs y compris le silhouette score.

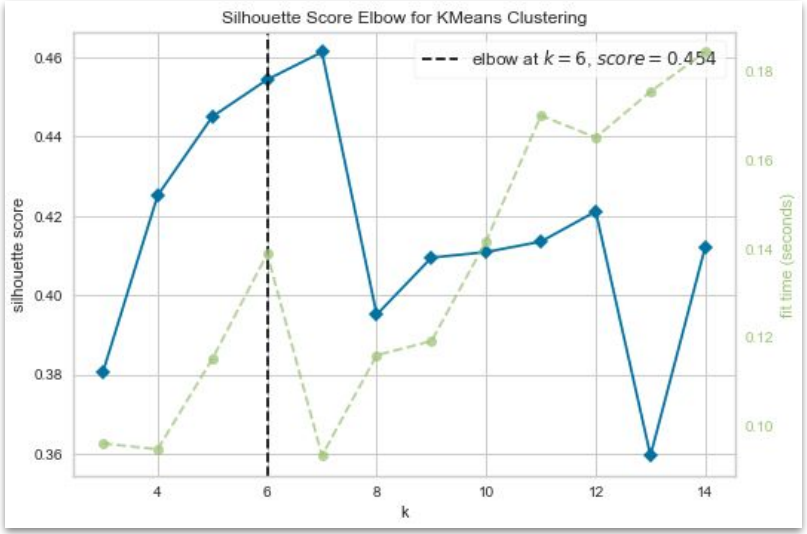
Dans le cas d’un DBSCAN, nous utilisons une méthode plus artisanale consistant à essayer différente valeurs via une boucle “for” (Chère en temps de calcul!) et d’ évaluer le/les meilleurs hyperparamètres via silhouette score.

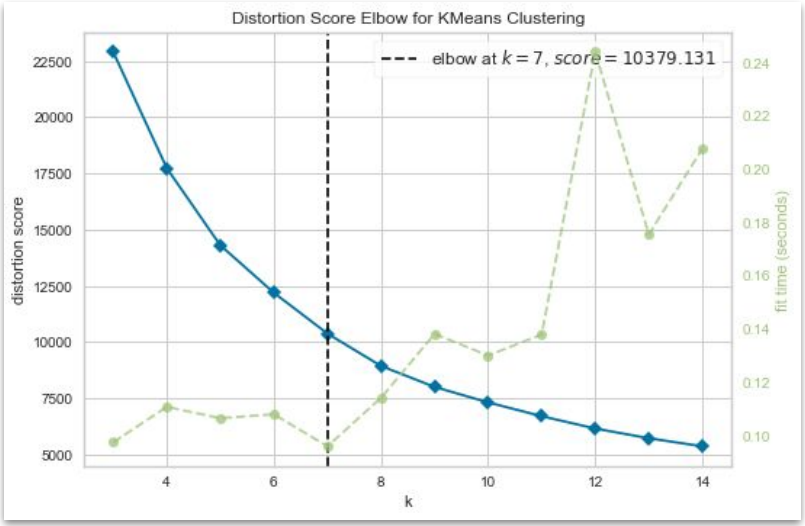
Globalement, les paramètres optimaux restent stables au fil et à mesure des essais avec des résultats entre 6 et 8.

Visualisation T-SNE

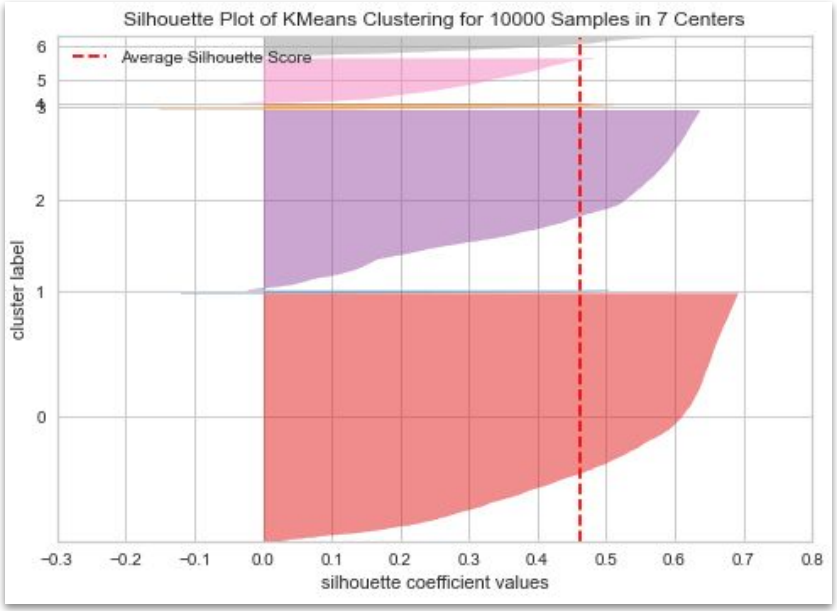


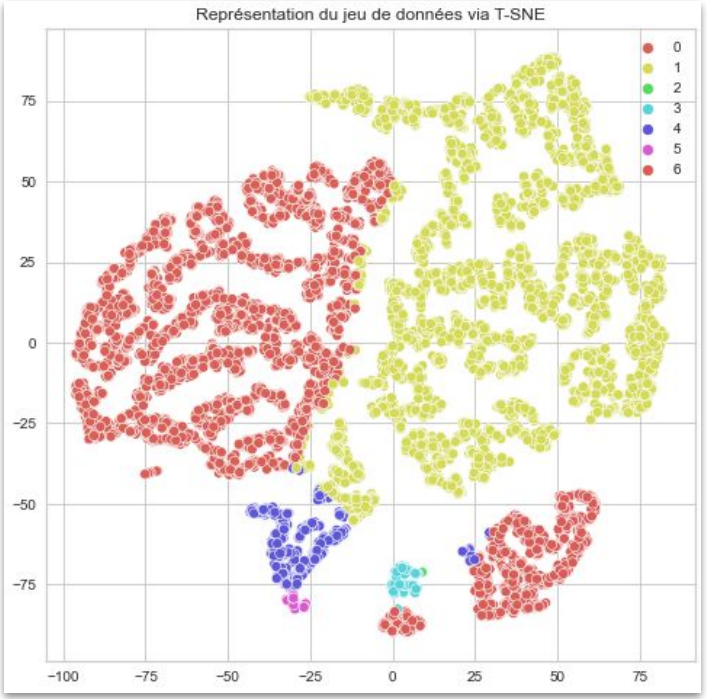
K-Means


Silhouette Score Elbow for KMeans Clustering
elbow at $k = 6$, $score = 0.454$
silhouette score
fit time (seconds)
k
0.46
0.44
0.42
0.40
0.38
0.36
0.18
0.16
0.14
0.12
0.10
4
6
8
10
12
14

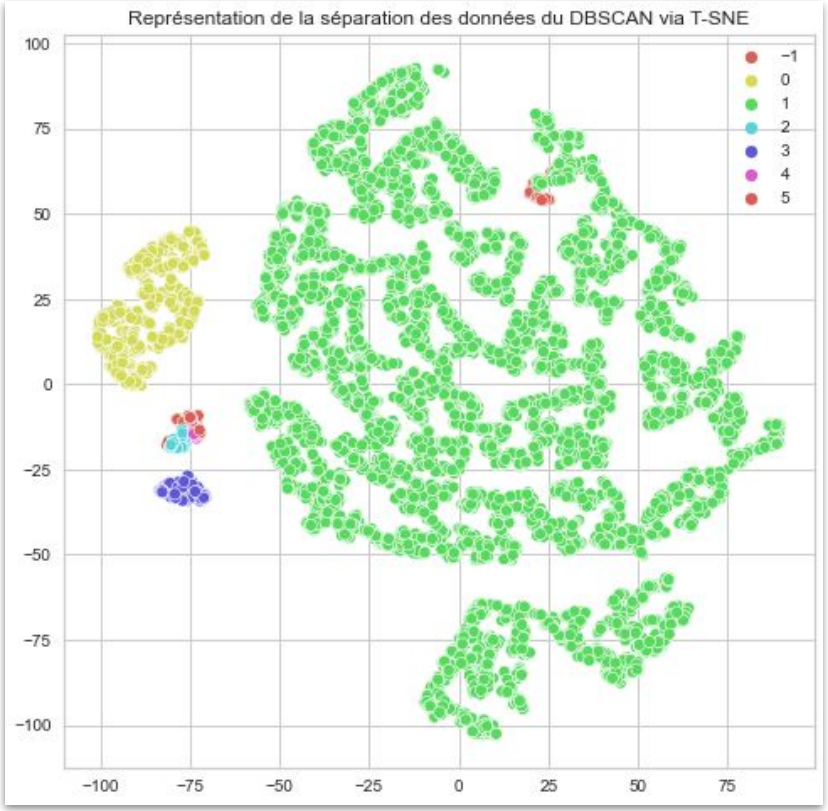

Distortion Score Elbow for KMeans Clustering
elbow at $k = 7$, $score = 10379.131$
distortion score
fit time (seconds)
k
22500
20000
17500
15000
12500
10000
7500
5000
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
4
6
8
10
12
14

K-Means


Silhouette Plot of KMeans Clustering for 10000 Samples in 7 Centers
Average Silhouette Score
cluster label
silhouette coefficient values
6
5
4
3
2
1
0
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


Représentation du jeu de données via T-SNE
0
1
2
3
4
5
6
75
50
25
0
−25
−50
−75
−100
−75
−50
−25
0
25
50
75

DBSCAN


Représentation de la séparation des données du DBSCAN via T-SNE
100
75
50
25
0
−25
−50
−75
−100
−100
−75
−50
−25
0
25
50
75
−1
0
1
2
3
4
5

Ajouter un tableau avec les caractéristiques par label

Conclusions:

(i) 3 clusters prédominants: Low-Value & Récurrents // Low-Value & Rare // Gros dépensier & Rare

(ii) Programme d'abonnement type Amazon pour les deux premiers groupes.

(iii) Offres Combo pour le Gros dépensier

Améliorations possibles pour update?

- Utilisation de NLP pour les commentaires afin d'évaluer les tendances de celles-ci, en fonction du nombre de commandes, prix moyen, dépenses total, etc…
- Essayer un clustering hierarchique?
- Essayer un clustering "manuel" pour comparer avec les résultats de nos algorithmes?
- Un cluster regroupe la grosse majorité des clients: peu d'apports des algorithmes? Nouvelles features nécessaires?

Proposition commerciale contrat de maintenance

Semble raisonnable de faire une mise à jour annuelle de cette base de données, afin d'évaluer l'évolution de la base de clients Olist.

- RDV pour l'année prochaine concernant la mise à jour majeure
- Contrat de maintenance basé sur un forfait à la journée
- Tout travail en dehors de ce périmètre devra être discuté upfront avec le client afin d'éviter des mauvaises surprises avec la facturation

Merci de votre attention!