

PROJET 5

SEGMENTEZ DES
CLIENTS D'UN SITE
E-COMMERCE



PLAN

- 01 CONTEXTE ET JEU DE DONNÉES
- 02 NETTOYAGE, FEATURES ENGINEERING ET EXPLORATION DES DONNÉES
- 03 MODÉLISATION (SEGMENTATION DES CLIENTS)
- 04 ÉTUDE DE LA STABILITÉ DES SEGMENTS AU COURS DU TEMPS
- 05 CONCLUSION



01

CONTEXTE ET JEU DE DONNÉES

CONTEXTE

🎯 OBTENIR UNE
SEGMENTATION MARKETING
DES CLIENTS

- Comprendre les différents types d'utilisateurs
- Fournir une description marketing actionnable de la segmentation
- Étude de la stabilité des segments au cours du temps

The logo for 'olist' is displayed in a bold, blue, sans-serif font. The letters are lowercase and have a slightly rounded, friendly appearance. The 'o' is a simple circle, the 'l' is a single vertical stroke, the 'i' has a dot, and the 's' is a simple curve. The background is white, and the logo is centered within a white rectangular area that is part of a larger slide.



SCORE DE RÉCENCE



SCORE DE FRÉQUENCE



SCORE DE MONTANTS

SEGMENTATION CLIENTS

- Scores RFM: grandeurs utilisées en marketing traditionnel permettant de faire ressortir les habitudes d'achat des clients
- Notre objectif: apporter une analyse précise des comportements des clients grâce à des algorithmes de Machine Learning non supervisés

FICHIER	NB DE LIGNES	NB DE COLONNES	DESCRIPTION
Customers	99441	5	Données clients
Geolocation	1000163	5	Données Géo localisation
Order items	112650	7	Articles commandés
Order payments	103886	5	Paiement commandes
Order reviews	99224	7	Avis commandes
Orders	99441	8	Commandes
Products	32951	9	Produits
Sellers	3095	4	Vendeurs
Product Category Name Translation	71	2	Traduction des noms de catégorie produit

LES DONNÉES

- ✓ Base de données anonymisée années 2016 à 2018
- ✓ 9 fichiers
- ✓ Types d'informations:
 - Les clients (Customers, Geolocation)
 - Les commandes (Orders, Order_items, Order_payments, Order_reviews)
 - Les vendeurs (Sellers)
 - Les produits (Products, Product_category_name_translation)



02

NETTOYAGE, FEATURES ENGINEERING ET EXPLORATION DES DONNÉES

ENVIRONNEMENT DE DÉVELOPPEMENT

ANACONDA

Installation d'Anaconda:
plateforme de distribution
python la plus populaire

ENVIRONNEMENT VIRTUEL

Mise en place d'un
environnement virtuel dédié
au projet

INSTALLATION DES PAQUETS

Installation des paquets
nécessaires (numpy,
pandas, matplotlib, seaborn,
sklearn, scipy, yellowbrick,
plotly) avec la commande **pip
install**

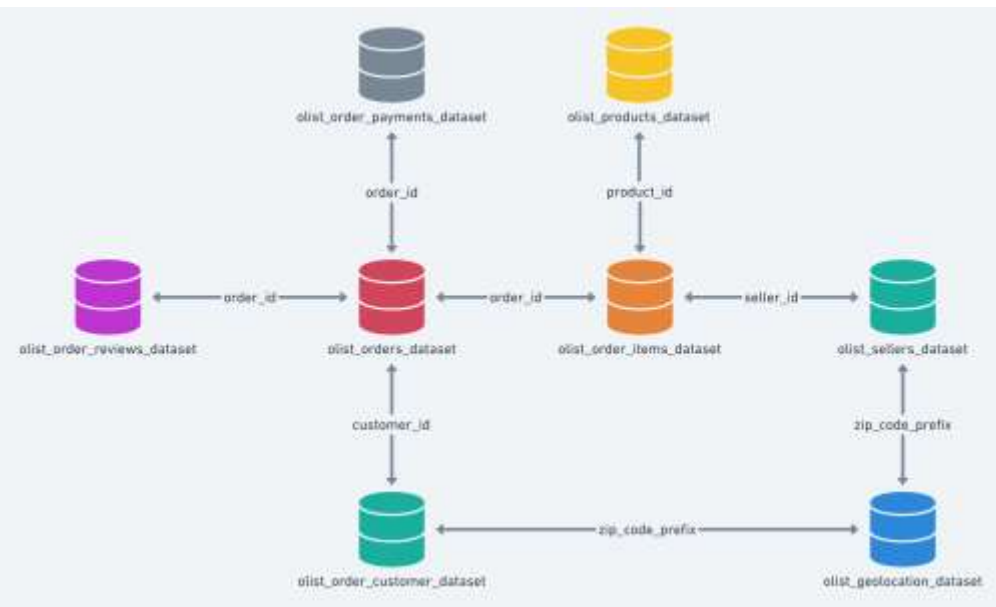
JOINTURES FICHIERS

- ✓ Datasets bien complétés (peu de valeurs nulles)
- ✓ Les fichiers sont reliés entre-eux par des clés primaires (voir image)
- ✓ Nous ne considérons que les commandes annotées « livrées » dans notre analyse

JOINTURE DES FICHIERS

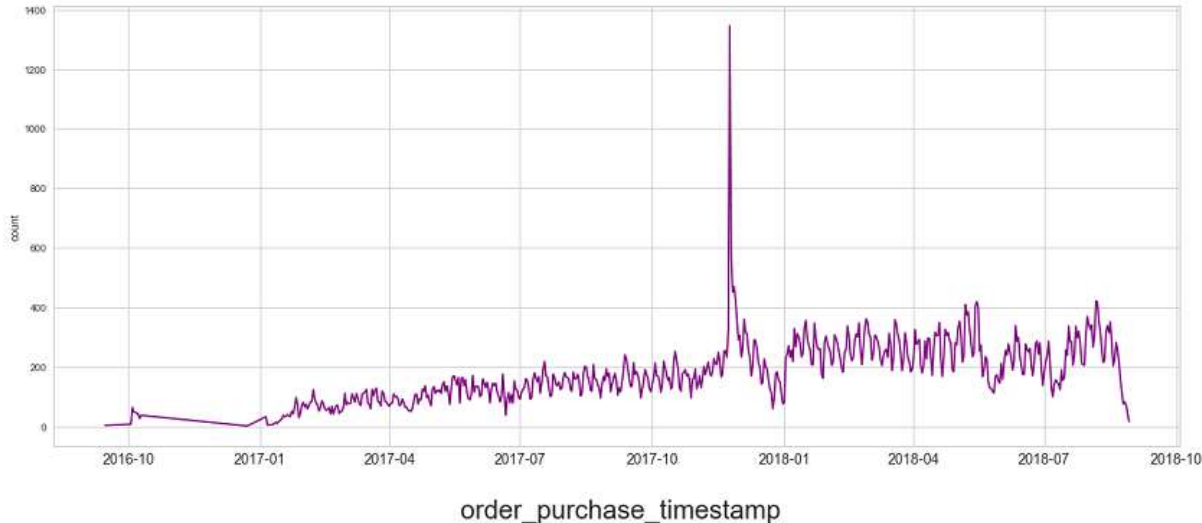
- ✓ Produits et Catégories produits
- ✓ Commandes et Clients

Note: fichier sur les vendeurs non exploité



EVOLUTION DES COMMANDES

Evolution du nombre de commandes journalières

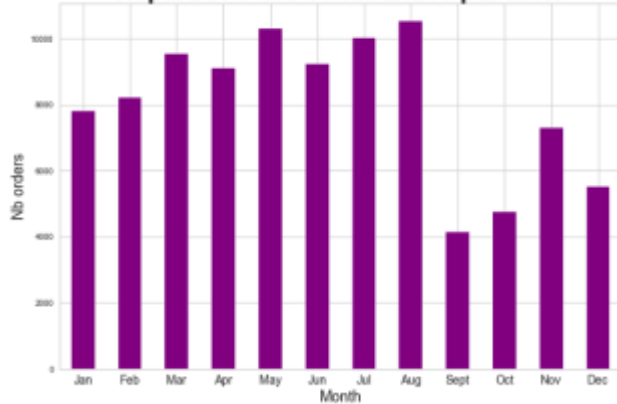


➤ Légère croissance du nombre de commandes journalière depuis 2016

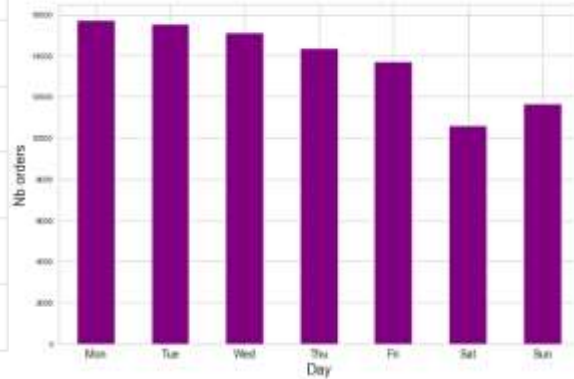
➤ Pic du nombre de commandes en fin 2017 (peut-être dû aux fêtes de fin d'années)

RÉPARTITIONS DES COMMANDES

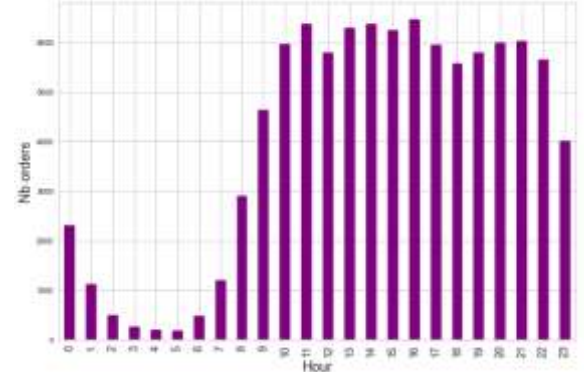
Répartition des commandes par mois



Répartition des commandes par jour de la semaine



Répartition des commandes par heure de la journée

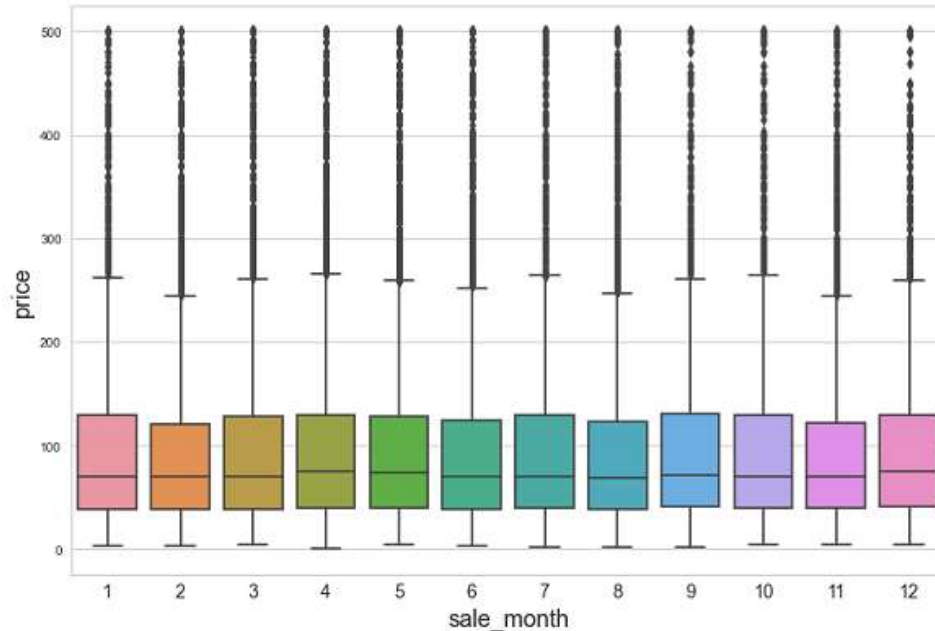


- ✓ Répartition plus ou moins régulière entre Janvier à Août.
- ✓ Baisse du nombre de commandes entre Septembre et Décembre
- ✓ Plus de commandes entre 10h et 22h

- ✓ Les clients ont tendance à plus commander en semaine que le week-end
- ✓ Nous avons décidé de **prendre en considération le mois de la commande comme feature dans notre segmentation**

DISTRIBUTION DU C.A./MOIS

Distribution du C.A. sur les mois de l'année

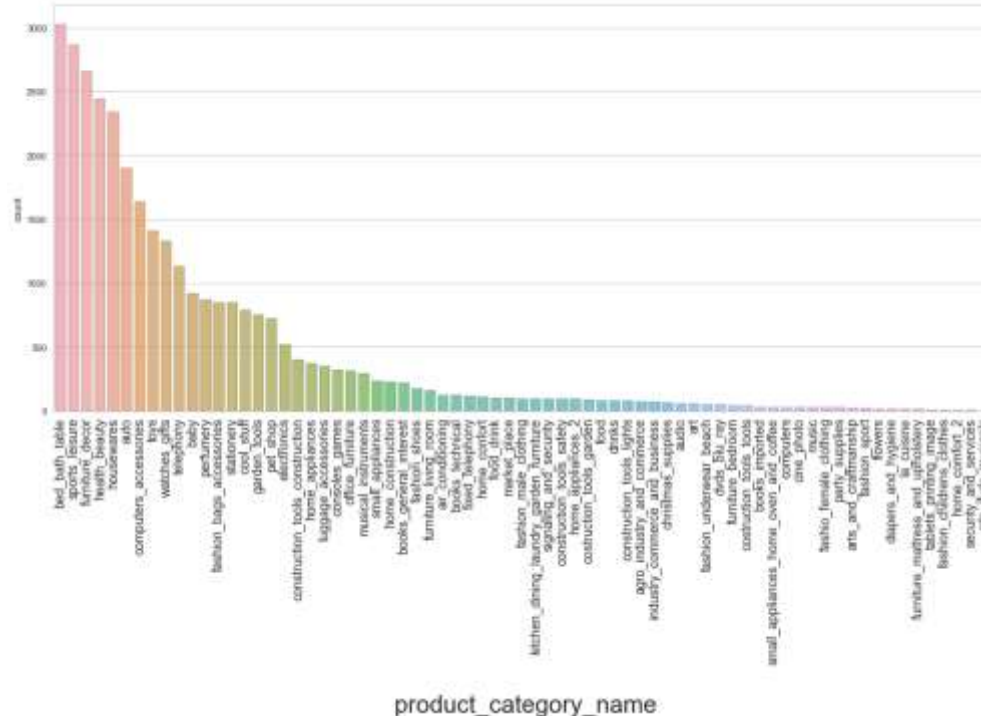


➤ Distribution du chiffre d'affaire assez équitable sur l'ensemble des mois de l'année

➤ Les médianes et variances restent très proches sur les différents mois de l'année

LES CATÉGORIES PRODUIT

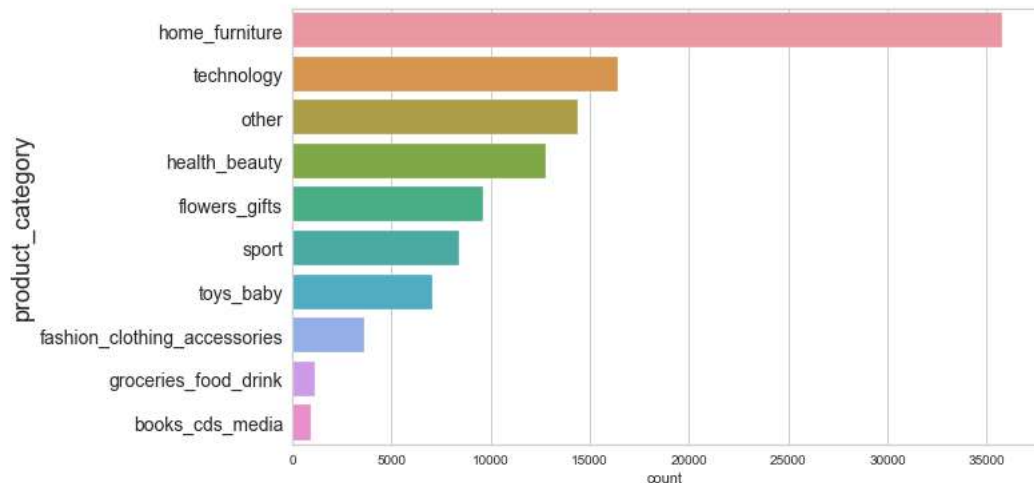
Les catégories produits les plus représentées



- 71 catégories produit
- Catégories produits les plus représentées:
 - bed_bath_table
 - sports_leisure
 - furniture_decor
- Regroupement en des catégories de plus haut niveau pour notre segmentation

REGROUPEMENT EN 10 CATÉGORIES

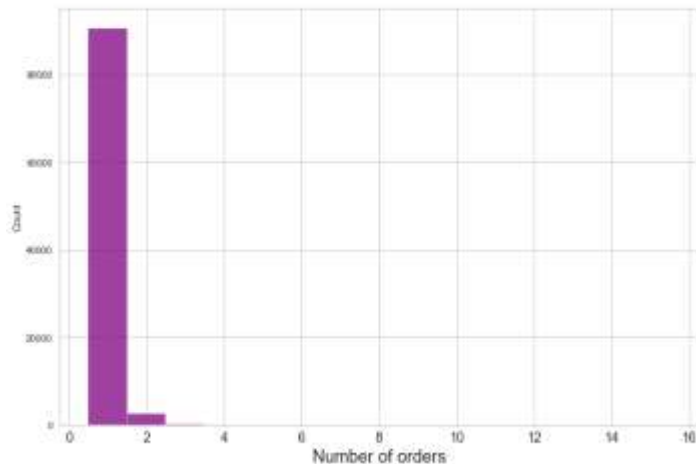
Les nouvelles catégories produits les plus représentées



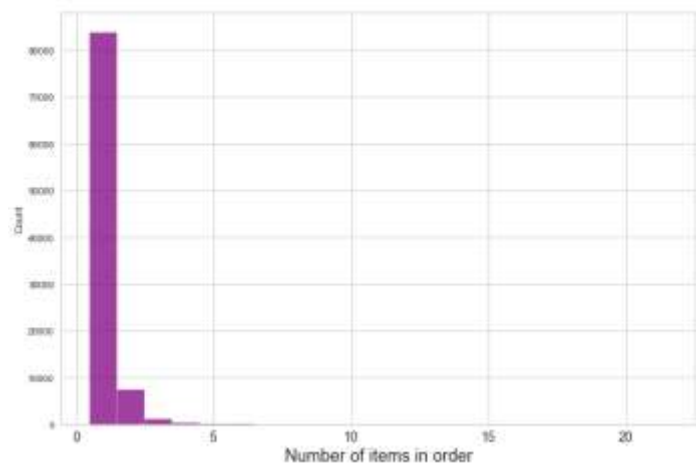
➤ Regroupement inspiré des catégories principales sur les sites de vente en ligne en 2017 (*source: <https://www.statista.com/>*)

➤ La catégorie **home_furniture** est largement la plus représentée

Nombre de commandes par client



Nombre moyen d'articles par commande



LES COMMANDES DES CLIENTS

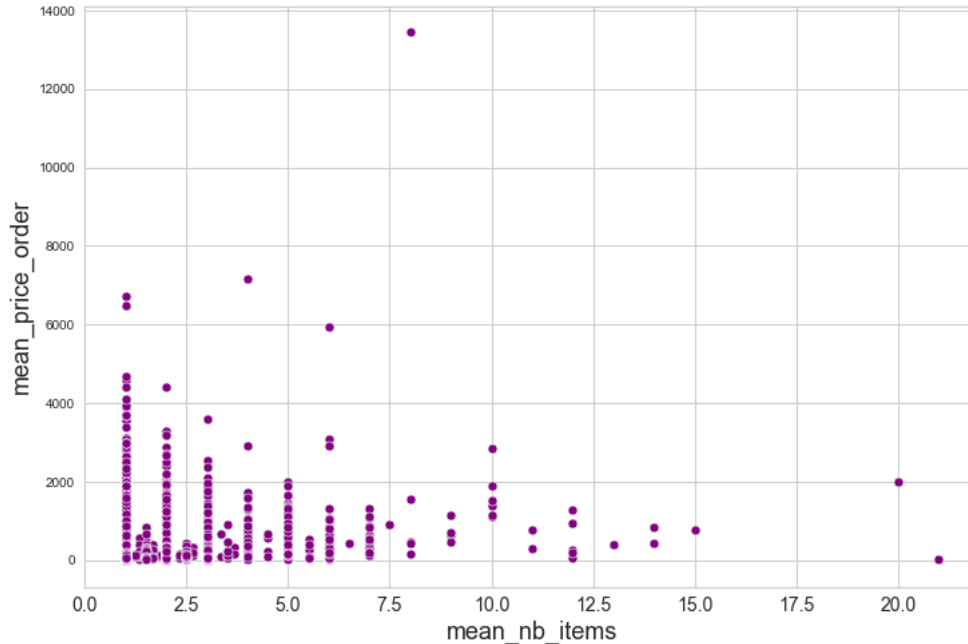
- La plupart des clients n'ont commandé qu'une seule fois
- La plupart des commandes ne contiennent qu'un seul article

CE QUI EN DECOULE

- Une segmentation RFM classique ne serait peut-être pas adaptée (fréquence = 1)
- Avoir une variable qui représenterait la catégorie produit préférée pour chaque client ne serait pas pertinent
- Nous avons créé une variable par catégorie produit représentant le ratio moyen d'articles de la catégorie comparé au nombre total de commandes passées par le client

PRIX MOYEN EN FONCTION DU NOMBRE D'ARTICLES

Répartition des prix moyen de commandes en fonction du nombre d'articles

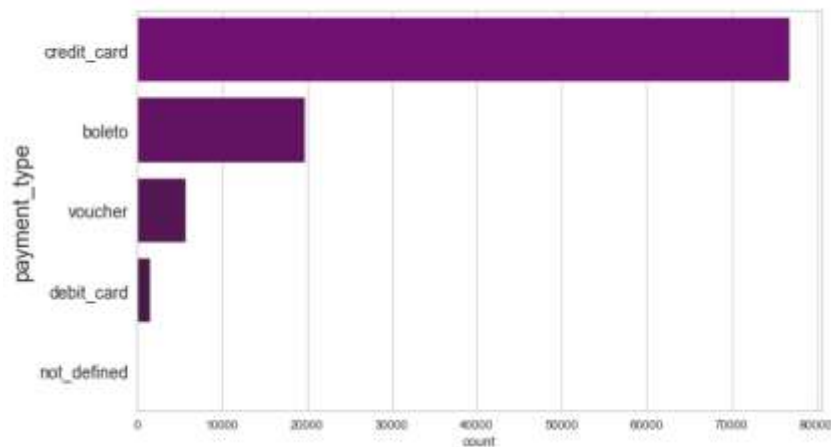


➤ Pas de relation linéaire entre le prix de la commande et le nombre d'articles commandés

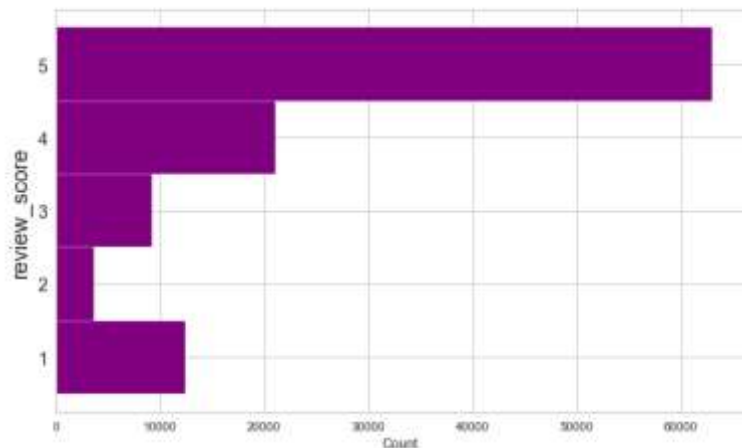
➤ Quelques commandes atypiques (mais pas aberrantes):

- Commande à plus de 13000 reais
- Commande de 20 articles avec un prix très faible

Les moyens de paiement utilisés sur le site



Répartition des notes attribuées aux commandes

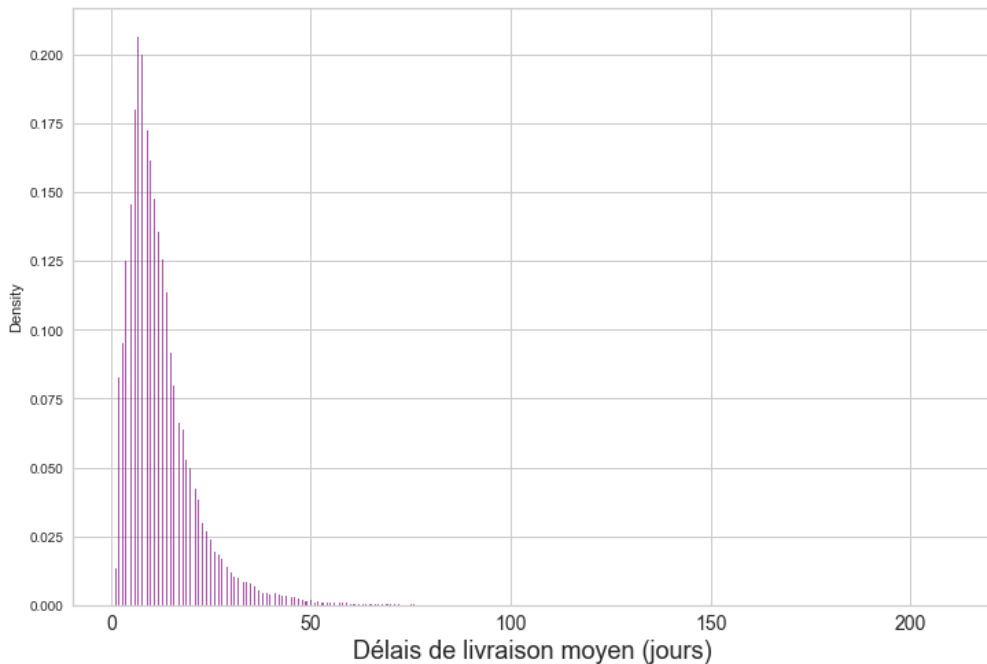


TYPES PAIEMENT ET NOTES SATISFACTION

- Une grande majorité des clients paient par carte de crédit
- Dans l'ensemble les clients sont plutôt satisfaits

LES DÉLAIS DE LIVRAISON

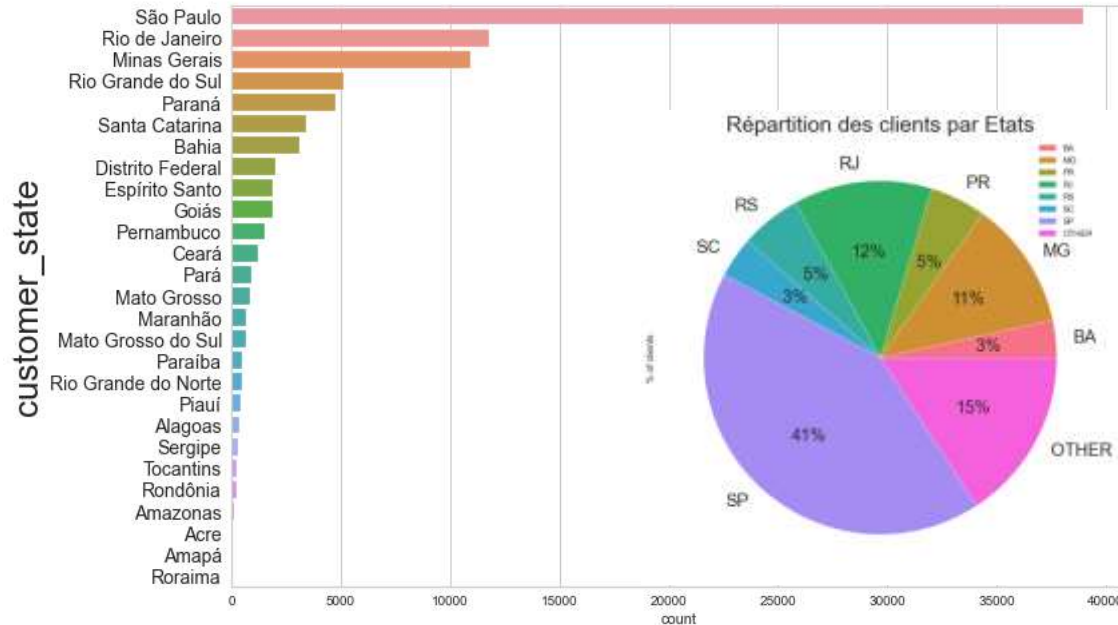
Répartition des délais de livraison moyens



- Majorité des commandes livrées en une quinzaine de jours
- Ajout d'une feature qui représente le délai de livraison moyen pour chaque client

LOCALISATION DES CLIENTS

Les états Brésiliens les plus représentées



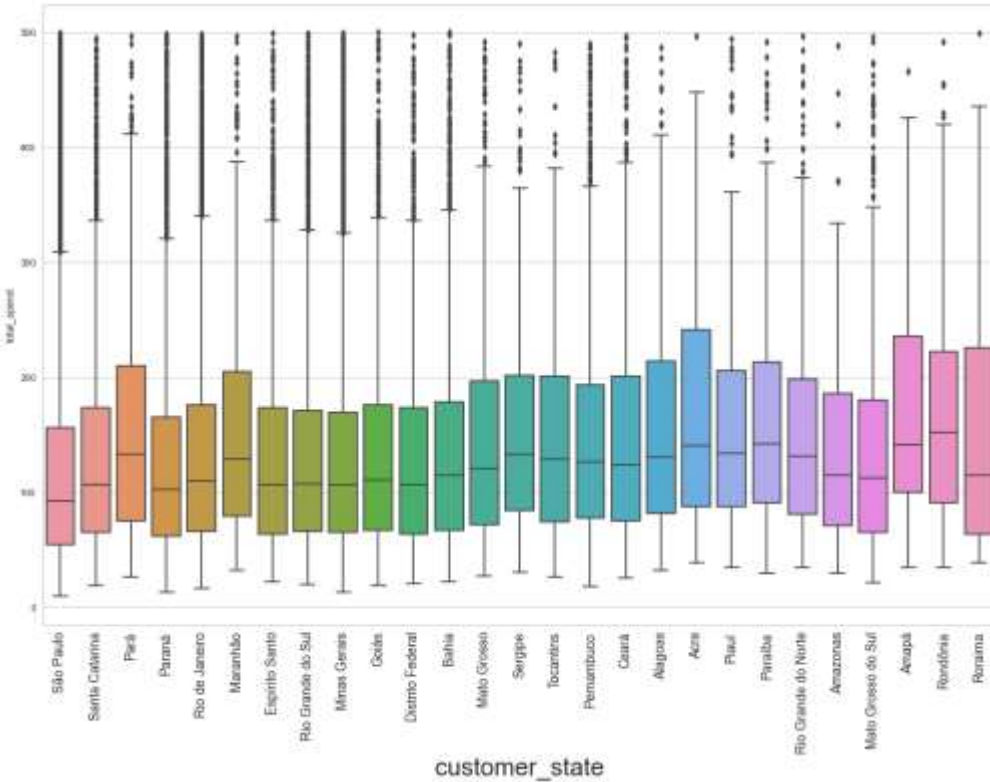
➤ Plus de 80% des clients sont regroupés dans 7 états

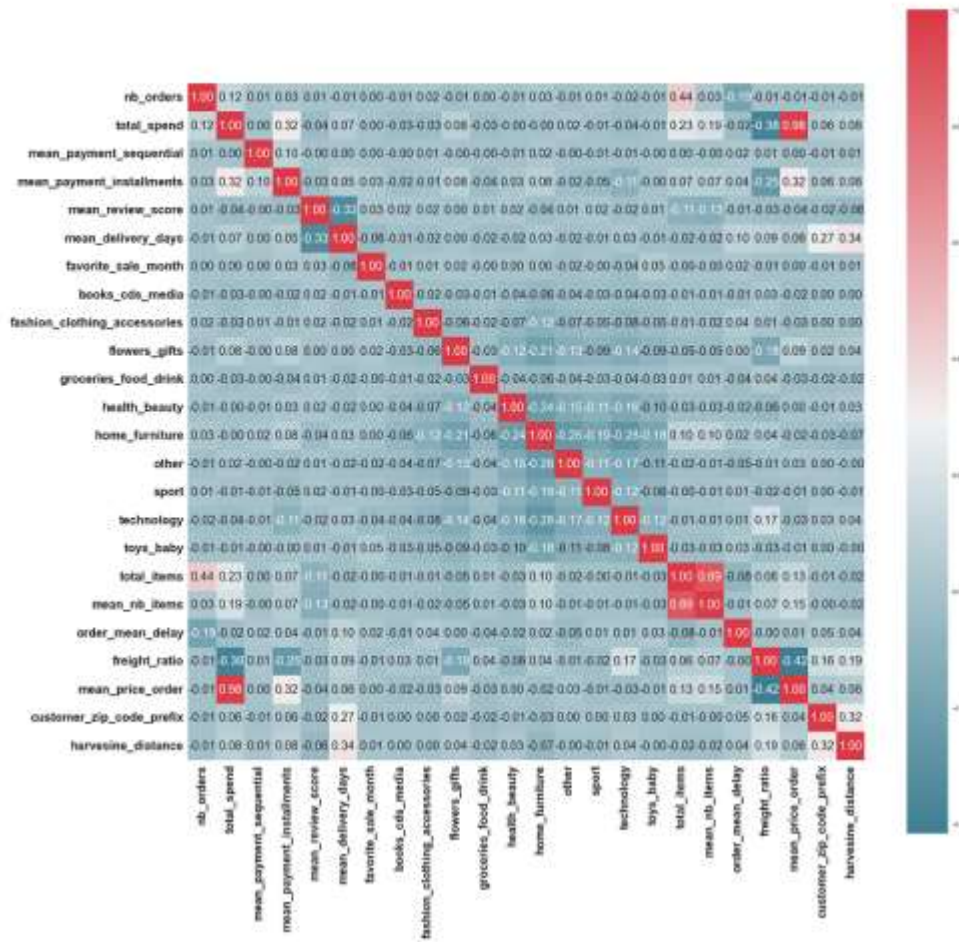
➤ Les clients proviennent principalement de l'Etat de São Paulo

➤ Nous avons créé une variable qui représente la **distance de haversine** entre les coordonnées moyenne de l'Etat du client et celle du siège d'Olist

LOCALISATION CLIENT VS. C.A.

- l'état du client a un impact faible sur les variances et médianes des dépenses sur le site





CORRÉLATIONS LINÉAIRES

On observe des corrélations fortes entre les couples de variables suivants:

- Total dépensé / prix moyen par commande
- Nombre moyen d'articles / nombre total d'articles

Corrélations certainement dues au fait que la majeure partie des clients n'ont commandé qu'une seule fois.

1 variable/2 sera conservée

FEATURES	DESCRIPTION
customer_unique_id	ID client
nb_orders	Nombre moyen de commandes
total_spend	Montant total dépensé
mean_payment_sequential	Nombre moyen de moyens de paiement par commande
mean_payment_installments	Nombre moyen de paiements par commande (nombre d'échéances)
mean_review_score	Note de satisfaction moyenne
delivery_delta_days	Délai moyen de livraison
favorite_month	Le mois favori de commande
Product_categories (10)	Ratios des catégories produit
total_items	Nombre total d'article commandé
order_mean_delay	Délai moyen entre commandes
freight_ratio	Le ratio frais de port / prix total commande
harvesine_distance	La distance Haversine entre l'état du client (moyenne des latitudes et longitudes de l'état) et le siège de Olist

JEU DE DONNEES FINAL

✓ 92755 clients

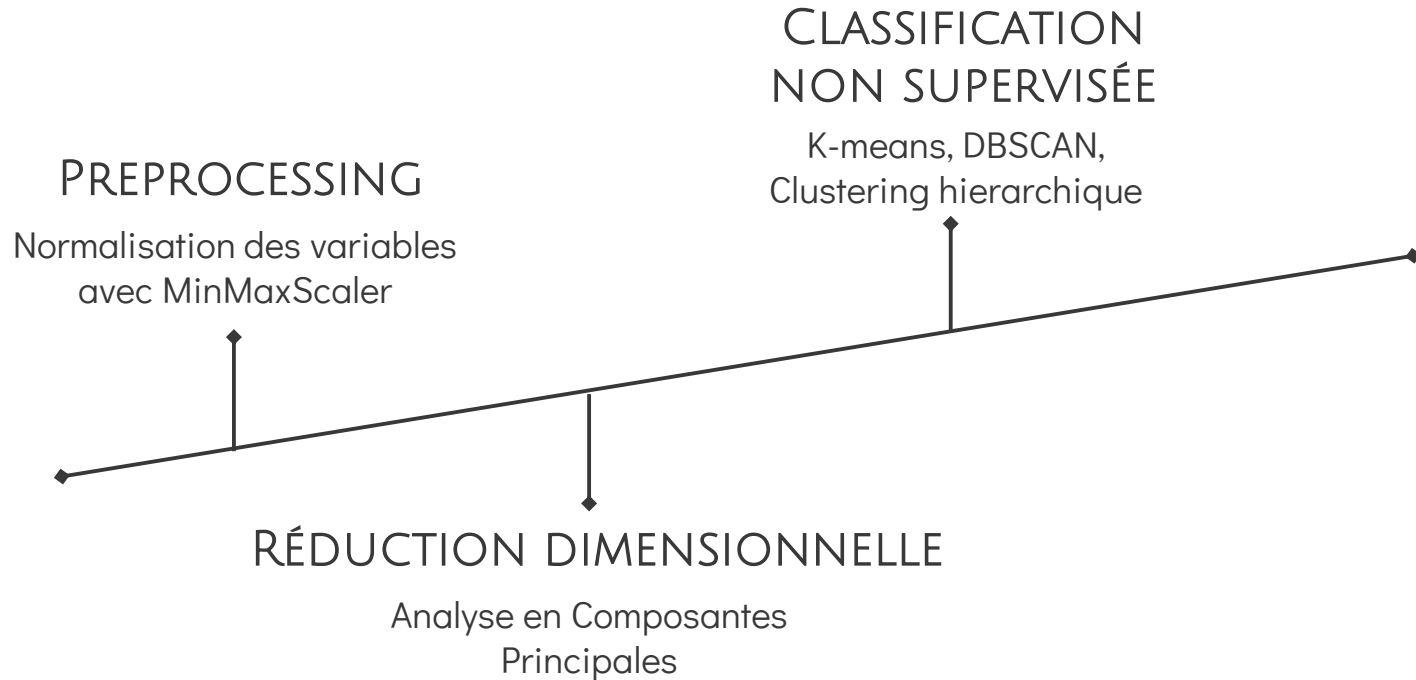
✓ 22 features (tous numériques)

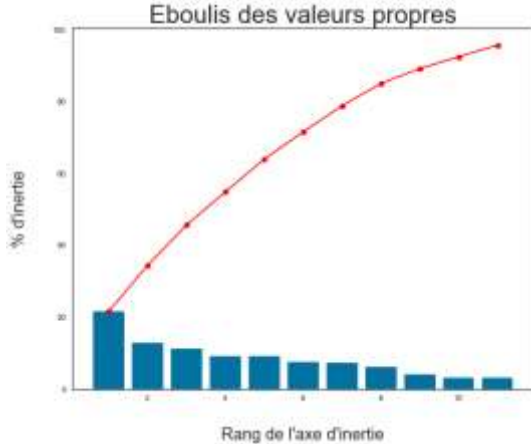


03

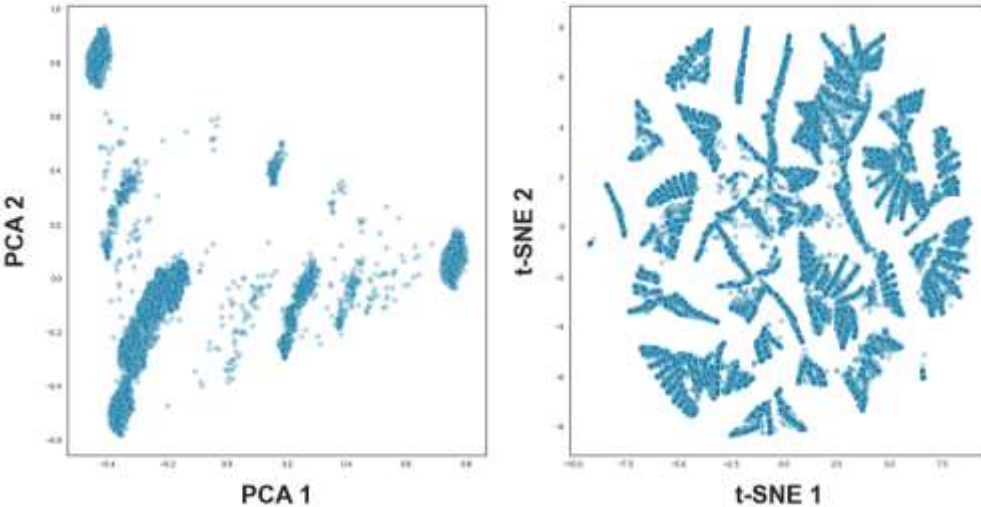
MODÉLISATION (SEGMENTATION DES CLIENTS)

MÉTHODOLOGIE





Projections des observations sur les axes
PCA et t-SNE



PCA+VISUALISATION T-SNE

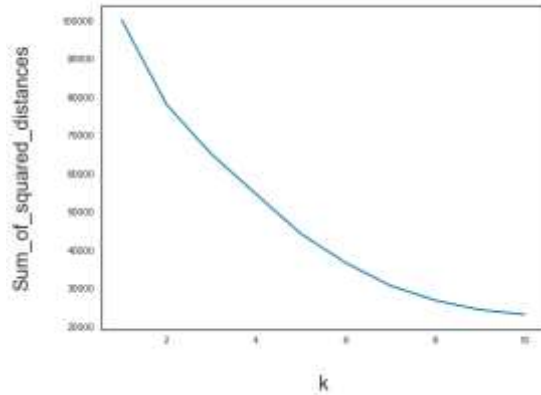
- 11 composantes expliquent 95% de la variance observées dans les données
- On peut observer la présences de sorte de « grappes » sur la visualisation des données

CHOIX DU MODÈLE

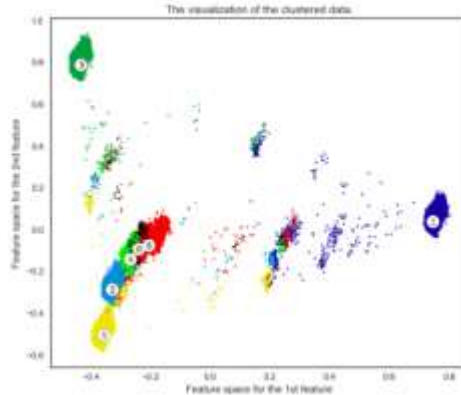
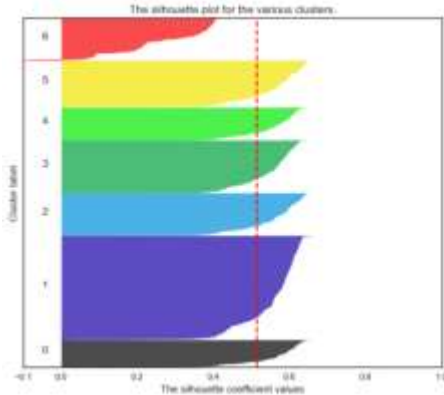


- Test de plusieurs algorithmes de clustering (K-Means, DBSCAN, Clustering hiérarchique) sur les données réduites avec PCA
- Critères de comparaison
 - Coefficient de silhouette [-1;1]
 - Compatibilité avec des connaissances spécifiques au domaine

Elbow Method For Optimal k



Silhouette analysis for KMeans clustering on sample data with n_clusters = 7



Algorithme	silhouette	Nb Clusters	Time
K-Means	0.513339	7	1.516018
K-Means_CV	0.513339	7	2.086654

K-MEANS

- La méthode du coude suggère que le nombre de clusters optimal autour de 7
- La méthode du coefficient moyen de Silhouette confirme que le nombre de clusters optimal se situe bien à 7
- Coefficient de silhouette : 0.51
- Tentative d'amélioration des performances avec une gridsearch (params: n_init, max_iter, tol, random_search) mais sans succès

	Algorithme	silhouette	Nb Clusters	Time
0	DBSCAN	0.052683	133	24.05119

```

eps value is 0.1
For eps value = 0.1 The avearge silhouette_score is : 0.05268325575854354
eps value is 0.2
For eps value = 0.2 The avearge silhouette_score is : 0.13150462632758542
eps value is 0.3
For eps value = 0.3 The avearge silhouette_score is : 0.0415826174770964
eps value is 0.4
For eps value = 0.4 The avearge silhouette_score is : -0.2300227661255724
3
eps value is 0.5
For eps value = 0.5 The avearge silhouette_score is : -0.0065108714387975
78
eps value is 0.6
For eps value = 0.6 The avearge silhouette_score is : 0.04439617442819740
5

```

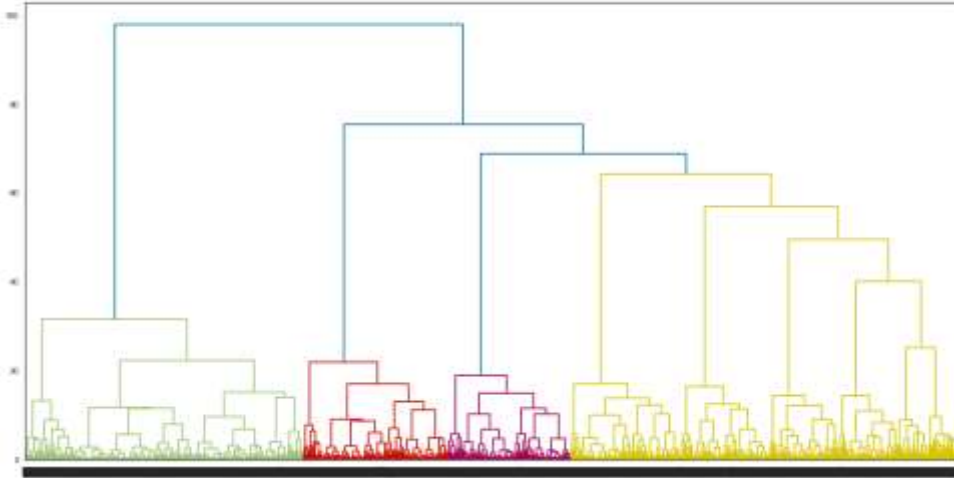
DBSCAN

- Test DBSCAN avec les paramètres suivants: min_sample=2*nb de features, eps=0.1
- DBSCAN se casse complètement la figure pour le clustering des données.
- Il segmente les données en plus de 130 clusters et donne un très mauvais score de silhouette.
- Tentative d'amélioration des performances en faisant varier eps
- Légère amélioration pou eps=0.2 mais le score de silhouette reste assez mauvais

CLUSTERING

HIERARCHIQUE

Observations Dendrograms



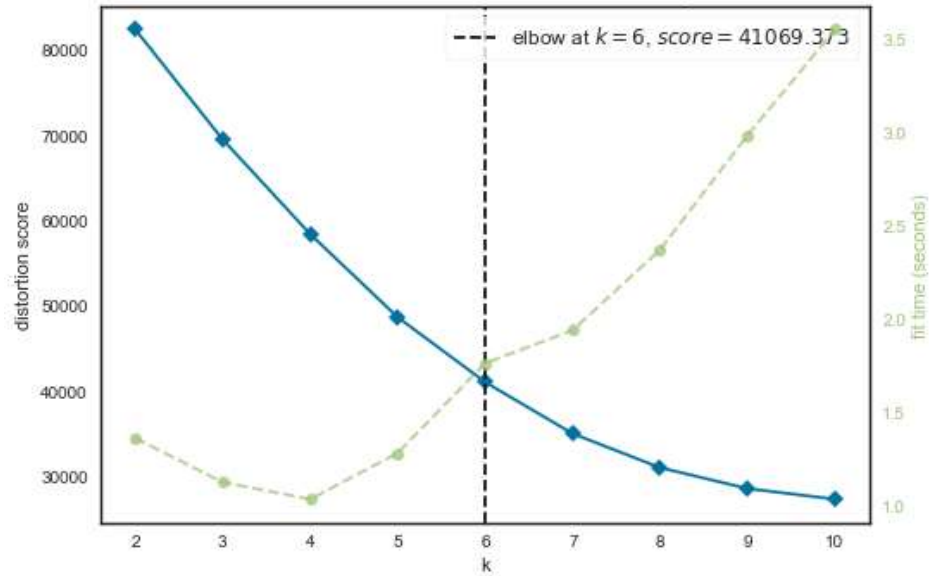
	Algorithme	silhouette	Nb Clusters	Time
0	Hierarchical	0.503076	7	46.283445

- Test fait seulement sur un échantillon réduit des données (soucis d'espace mémoire)
- Performances proches de K-Means
- Temps d'exécution beaucoup plus élevé et requiert beaucoup d'espace mémoire

CONCLUSION CHOIX DU MODÈLE

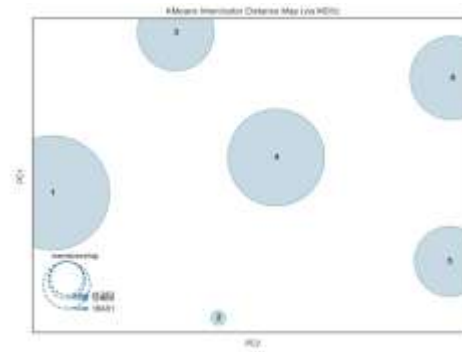
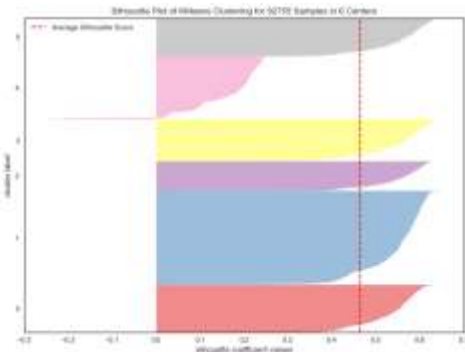
L'algorithme le plus rapide, créant les clusters
les plus nets et possédant le meilleur
coefficient de silhouette est celui du **K-means**

Distortion Score Elbow for KMeans Clustering



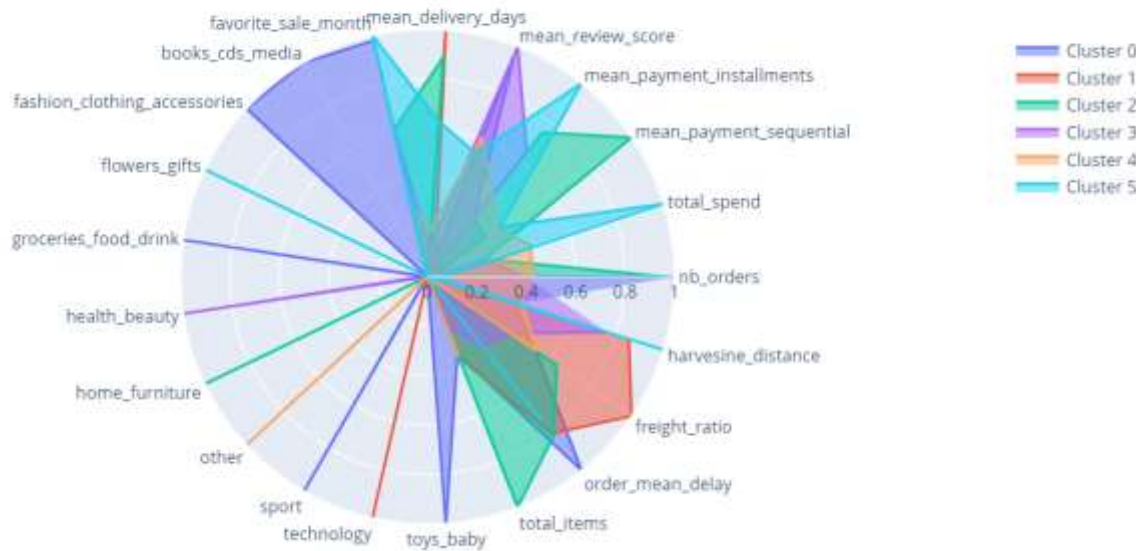
K-MEANS (SANS PCA)

- Nombre optimal de cluster suggéré par la méthode du coude = 6
- Score de silhouette = 0.48
- Clusters relativement bien répartis (un des groupes a une densité supérieure aux autres)
- Clusters bien séparés (projection des clusters sur les 2 premières composantes de la MDS (Multi Dimentional Scaling))



CARACTÉRISATION DES CLUSTERS

Projection des moyennes par variable des clusters

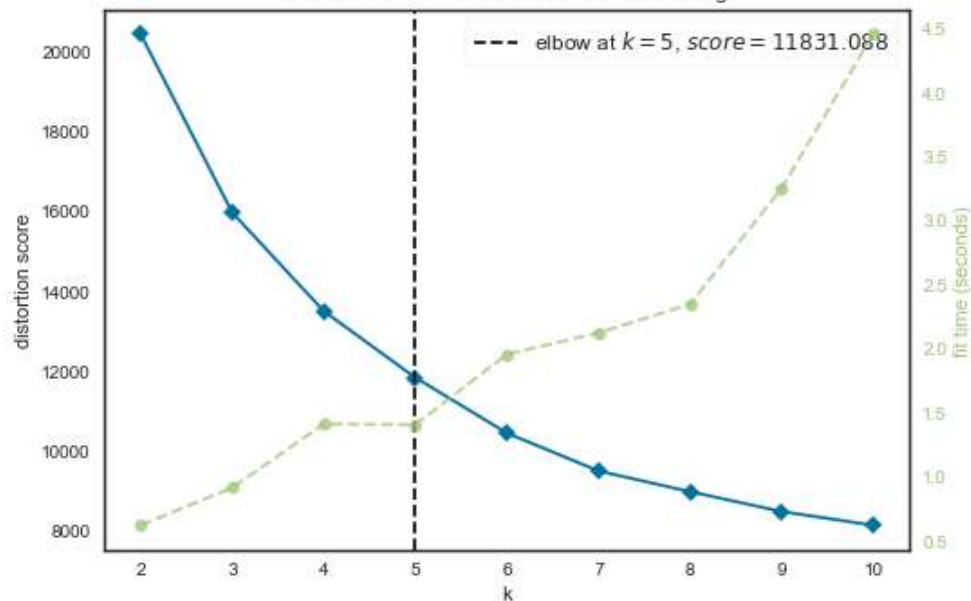


➤ Malheureusement ici, la segmentation se base principalement sur les catégories de produit achetées.

➤ Le poids de ces features masquent les autres axes de catégorisation

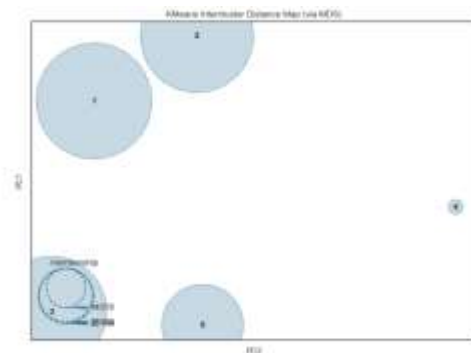
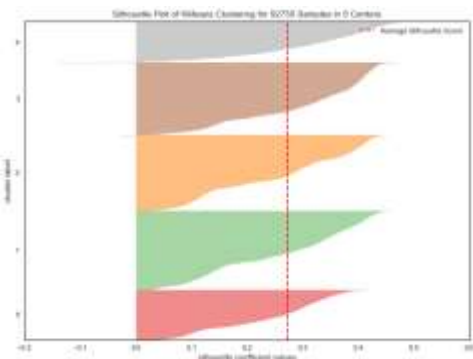
➤ Nous allons donc réaliser un nouveau K-Means en supprimant ces variables 32

Distortion Score Elbow for KMeans Clustering



K-MEANS SANS LES CATÉGORIES PRODUIT

- La méthode du coude propose un clustering en 5 clusters
- Clusters plus homogènes
- Clusters bien séparés
- Score de silhouette en baisse (0.28)



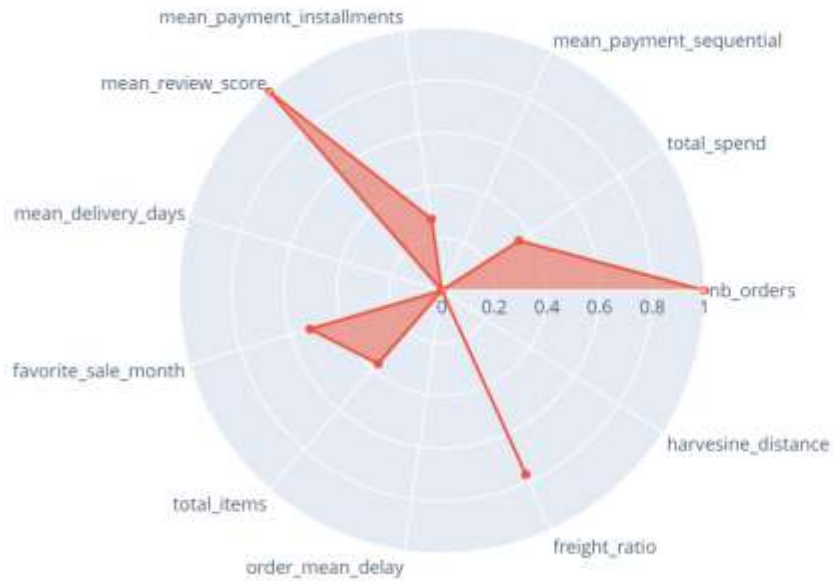
CARACTÉRISATION DES CLUSTERS



Cluster 0

- 14578 clients
- Utilisent plusieurs moyens de paiement et un nombre important d'échéances
- Pas très éloignés du siège d'Olist
- Ont tendance à espacer les délais entre 2 commandes
- Ont très peu commandé et donc très peu dépensés
- Très satisfaits

CARACTÉRISATION DES CLUSTERS



Cluster 1

- 23035 clients
- Ont passés un nombre important de commandes
- Proches géographiquement du siège d'Olist avec de courts délais de livraison
- Paient comptant pour un montant moyen de commande
- Très satisfaits

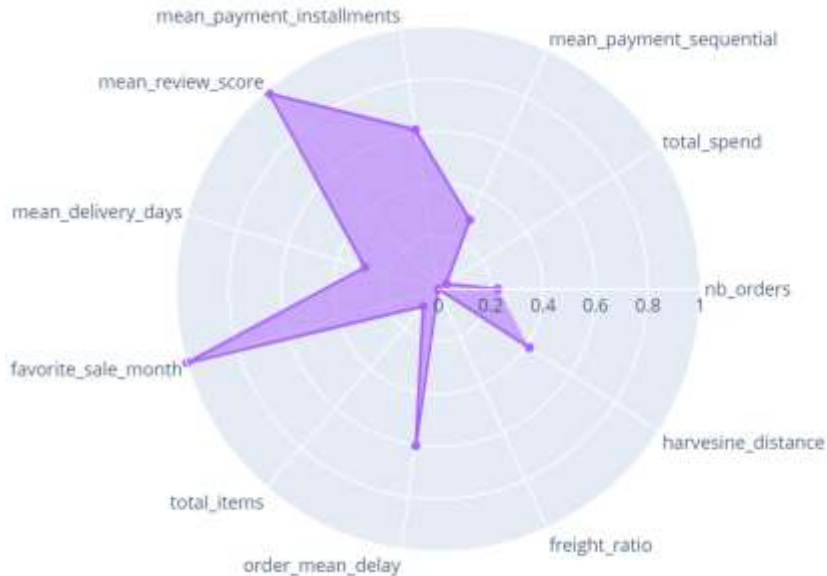
CARACTÉRISATION DES CLUSTERS



Cluster 2

- 22106 clients
- Proches géographiquement du siège d'Olist avec de courts délais de livraison
- Commandent principalement en début d'année pour des montants faibles
- Paient avec 1 type de moyen de paiement et avec un nombre faible d'échéances
- Très satisfaits

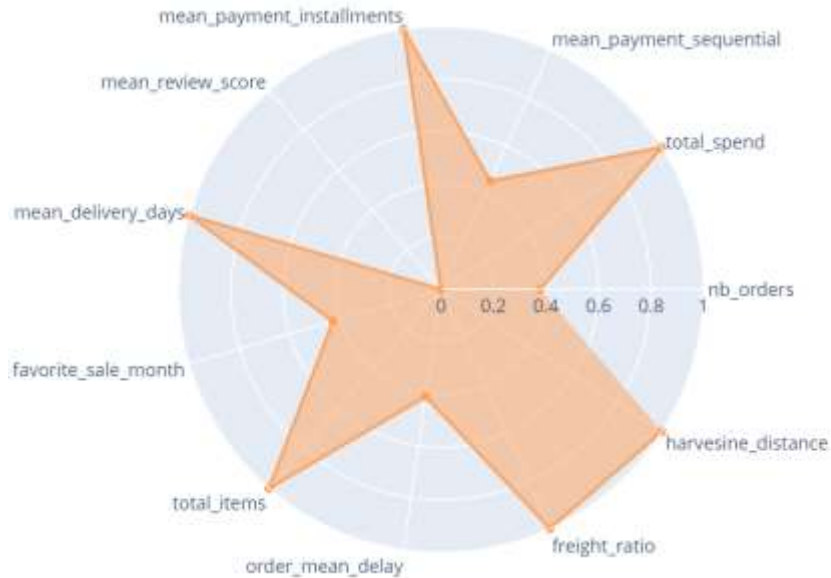
CARACTÉRISATION DES CLUSTERS



Cluster 3

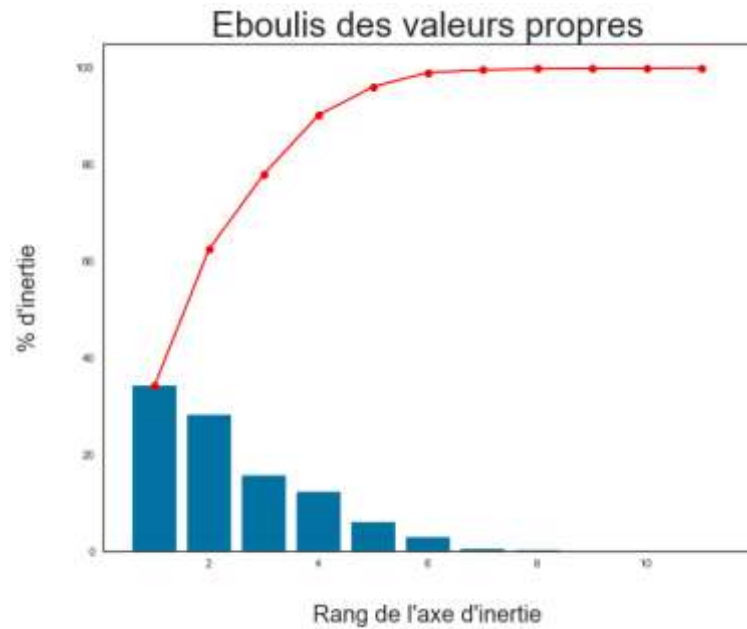
- 21132 clients
- Clients de fin d'année
- Géographiquement peu éloignés avec des délais de livraison moyens
- Paient en plusieurs échéances avec plusieurs moyens de paiement pour des montants faibles
- Très satisfaits

CARACTÉRISATION DES CLUSTERS



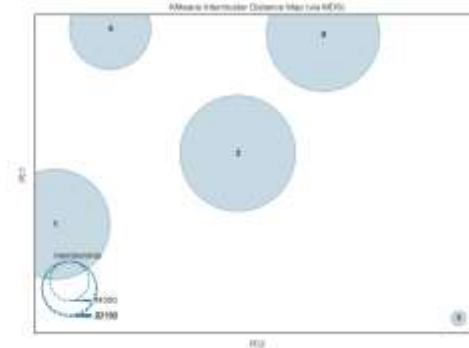
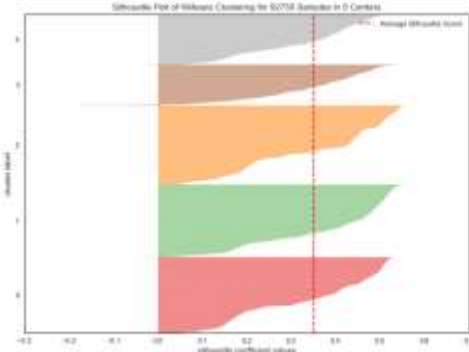
Cluster 4

- 11904 clients
- Éloignés géographiquement du siège d'Olist avec de frais et délais de livraison élevés
- Ont dépensés le plus et commandé un grand nombre d'articles et passés un nombre de commande moyen
- Paient en plusieurs échéances avec plusieurs moyens de paiement
- Très insatisfaits



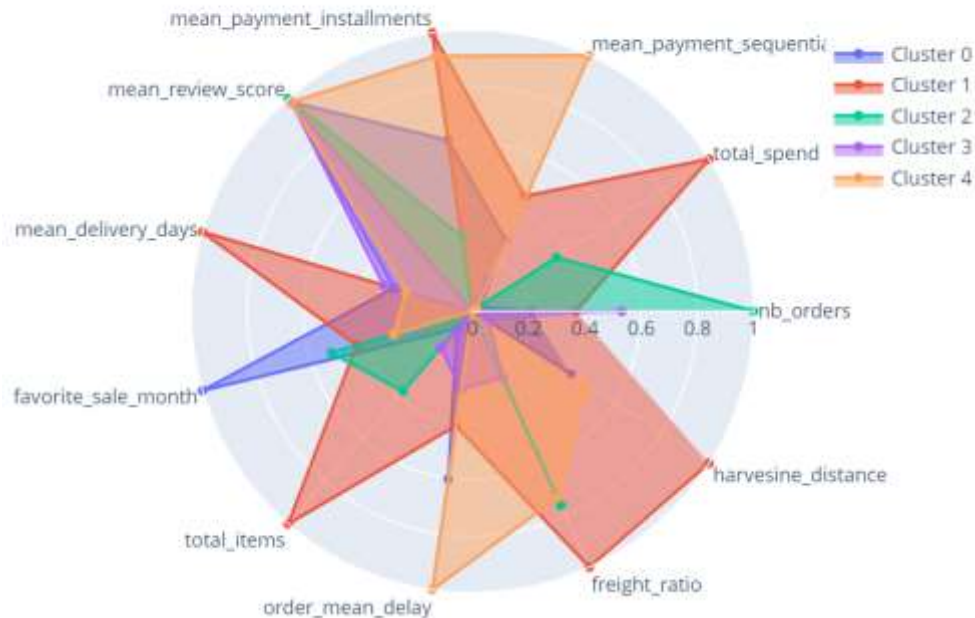
K-MEANS (AVEC PCA)

- PCA à nouveau après suppression des catégories produit
- 5 composantes expliquent 95% de la variance des données
- Score de silhouette = 0.35
- Densité des clusters assez bien répartie et clusters bien séparés



CARACTÉRISATION DES CLUSTERS

Projection des moyennes par variable des clusters



➤ On retrouve les mêmes axes de segmentation après réduction de dimension (les segments clients sont conservés)

➤ le score de silhouette moyen est meilleur comparé à celui obtenu avec les données brutes


	Iteration	silhouette	ARI	Time
0	0.0	0.341223	0.999950	0.269085
1	1.0	0.341223	0.999925	0.177918
2	2.0	0.316205	0.566571	0.201203
3	3.0	0.341223	0.999919	0.149117
4	4.0	0.336672	0.601328	0.216980
5	5.0	0.315773	0.566268	0.175997
6	6.0	0.342655	0.607031	0.178920
7	7.0	0.341223	1.000000	0.175065
8	8.0	0.342410	0.606032	0.165730
9	9.0	0.341223	0.999925	0.224489

STABILITÉ DE K-MEANS

- Objectif : tester la stabilité à l'initialisation de K-Means
- Entraînement de K-Means plusieurs fois sans fixer le `random_state` et avec `n_init=1`

Les différentes itérations montrent:

- des scores de silhouette proches
- des scores ARI proches de 1
- Mise en évidence de la stabilité à l'initialisation de K-Means



04

ETUDE DE LA STABILITÉ DES SEGMENTS AU COURS DU TEMPS

periode		ARI
0	3	0.960803
1	6	0.762966
2	9	0.579155

STABILITÉ DES SEGMENTS

Principe

- Sélectionner 12 mois de données initialement
- itérer le K-Means sur toute la période restante avec des deltas de 3 mois et calculer le score ARI, en prenant garde à bien comparer les mêmes clients (ceux des 12 mois initiaux)

Observation : Après ajout de 6 mois de données sur les 12 mois de données initiales, on constate une dégradation plus ou moins importante du score ARI.

Suggestion: mettre à jour la segmentation tous les 6 mois



05

CONCLUSION ET PERSPECTIVES

CONCLUSION

- L'algorithme K-Means est adapté pour la segmentation des clients d'Olist
- De meilleurs segments marketing seraient certainement obtenus si on a un jeu de donnée plus enrichi (où les clients ont fait plusieurs commandes)
- L'étude de la stabilité des segments au cours du temps a révélé qu'une mise jour de la segmentation serait nécessaire tous les 6 mois.

MERCI

Des questions ?

