Soutenance Projet 5

Pierre Schifflers

Agenda

- Problématique
- Nettoyage des données et analyse exploratoire
- Modélisation et évaluation des clusters
- Interprétation des clusters
- Conclusions
- Q&A

Environnement

- Python 3.7.6
- JupyterLab 1.2.6
- Scikit Learn 0.22.1
- Numpy 1.18.1
- Seaborn 0.10.0
- Matplotlib 3.1.3
- Fuzzy Wuzzy 0.17.0
- Tableau 2020.1

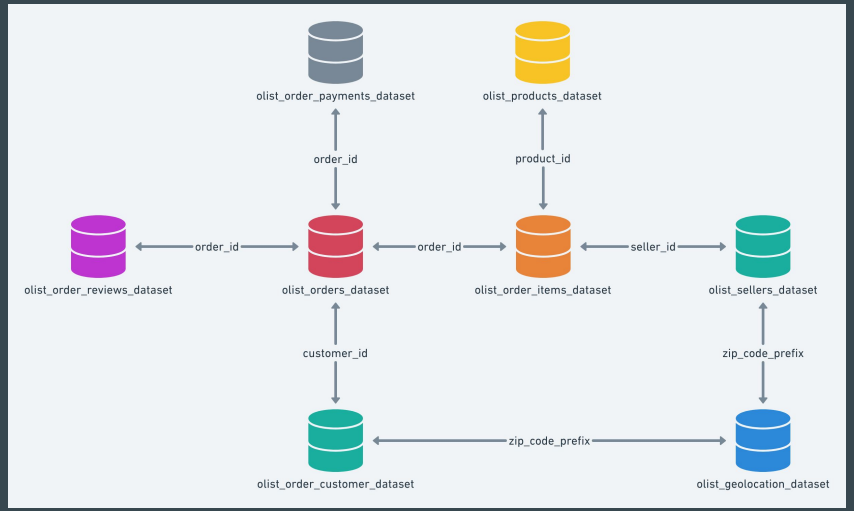
Problématique

olist

- Site e-commerce en ligne Brésilien
- Créer une segmentation de clients pour les campagnes de communication
- Comprendre différent types d'utilisateurs
- Fournir description actionable des clusters
- Proposition de contrat de maintenance

Données

- 8 fichiers CSV + 1 fichier de traduction des catégories vers l'anglais
- Assemblage du jeu de données sur base du schéma fourni
- Vérification des données dupliquées sur base des clés uniques
- Modification du fichier Geolocation → médiane des coordonnées géographiques par code postal

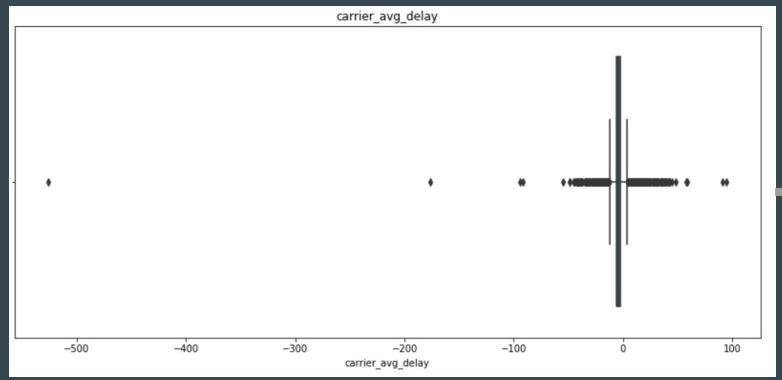


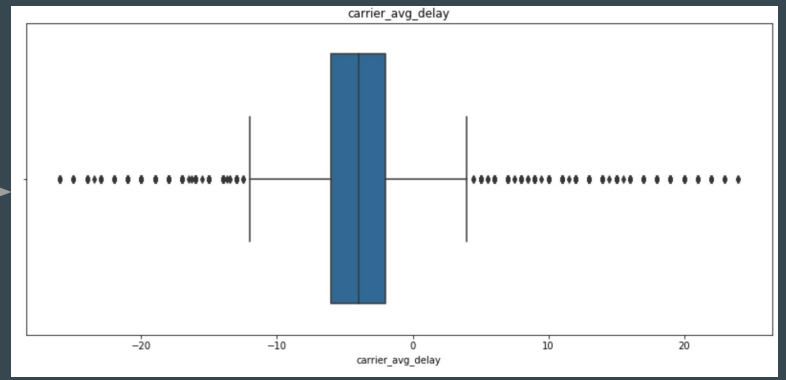
Feature Engineering

- Reconstruction des données avec Customer Unique ID comme granularité
- Construction de features:
  - # de commandes par client
  - # de produits commandés par client
  - Prix payé par client
  - Score moyen des critiques
  - Temps de livraison moyen
  - Retard moyen comparé à la date de livraison prévue
  - Catégorie avec le plus de dépense (one-hot encoding)
  - Moyen de paiement le plus utilisé (one-hot encoding)
- Regroupement de catégories de produits en 13 groupes plus génériques

Nettoyage des données

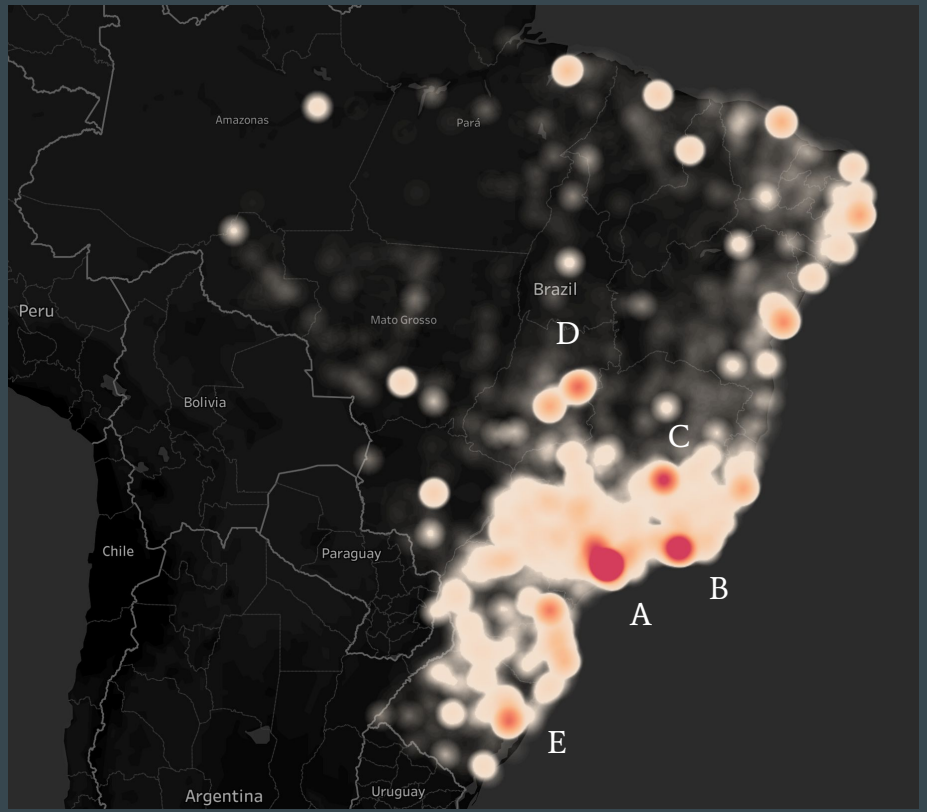
- Suppression de toutes commandes non complètes
- Suppression des données plus basses que le 0.1ème quantile et plus élevées que le 99.9ème quantile dans chaque feature (3% des données)
- Vérification des valeurs lexicales et de la cohérence des valeurs aberrantes restantes.





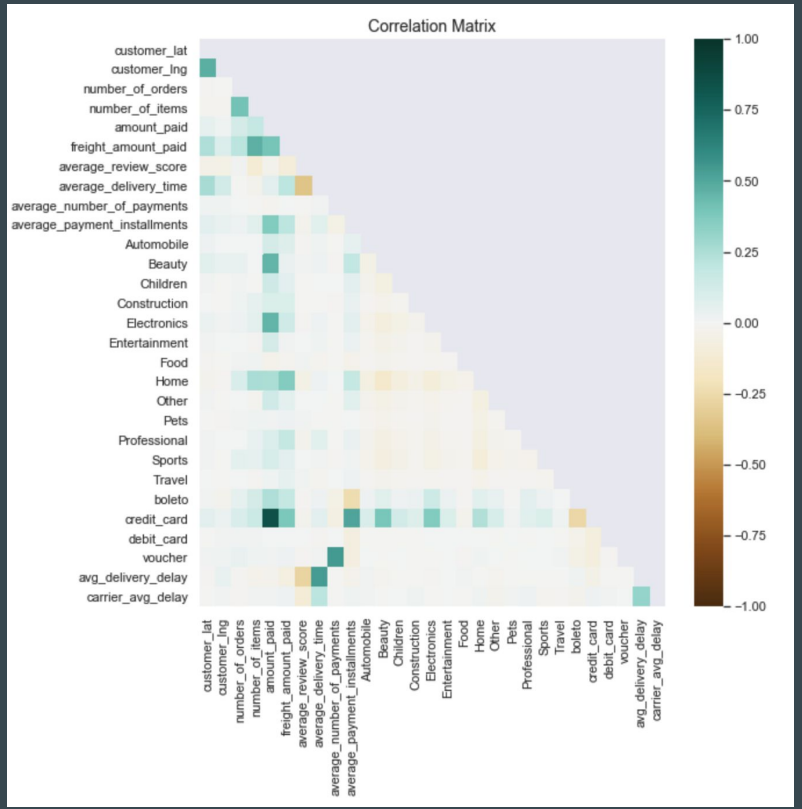
Exploration

- Concentration des clients dans les grandes villes:
  a. São Paulo
  b. Rio de Janeiro
  c. Belo Horizonte
  d. Brasilia
  e. Porto Alegre



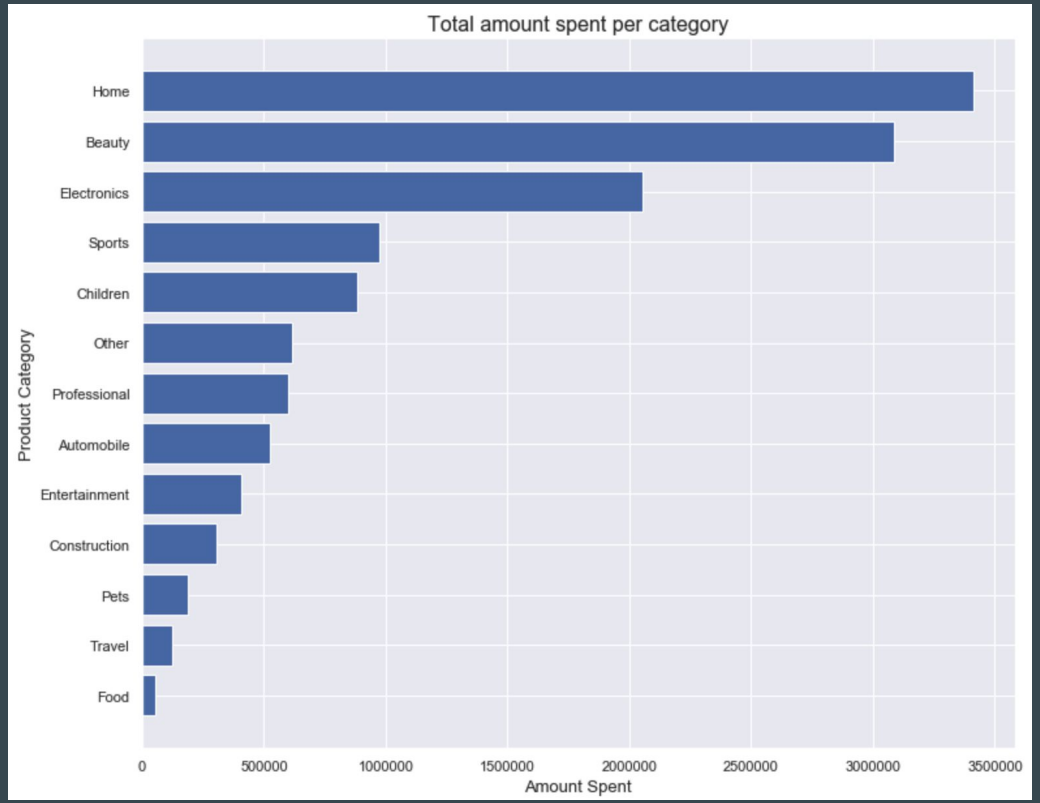
Exploration

- Corrélation entre les montants payés et le montant de fret payé
- Corrélation négative entre la critique et le temps de livraison/retard
- Corrélation négative entre Boleto et carte de crédit → Alternatives
- Corrélation entre le moyen de paiement et le nombre de paiements



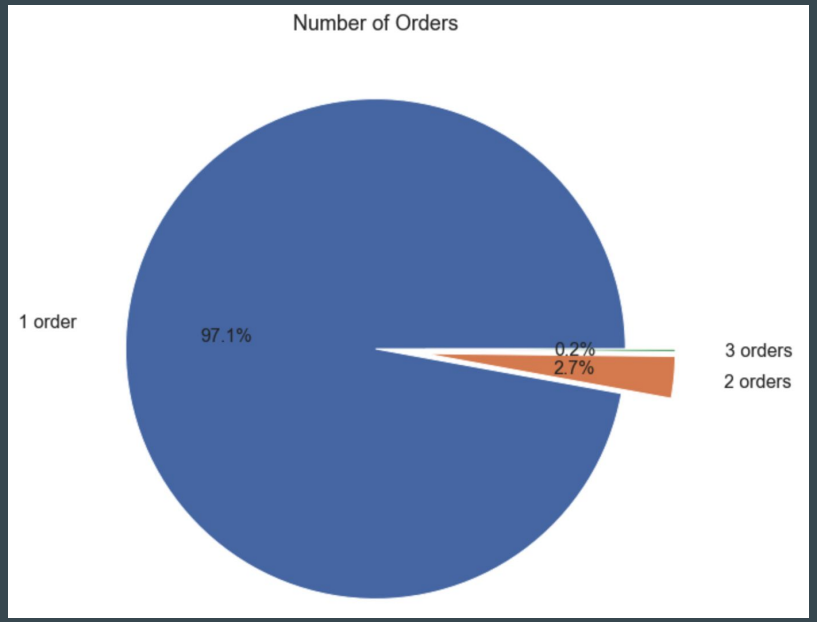
Exploration

- Les 3 plus grandes catégories sont Home, Beauty et Electronics.
- Ces 3 catégories représentent 65% des dépenses totales



Exploration

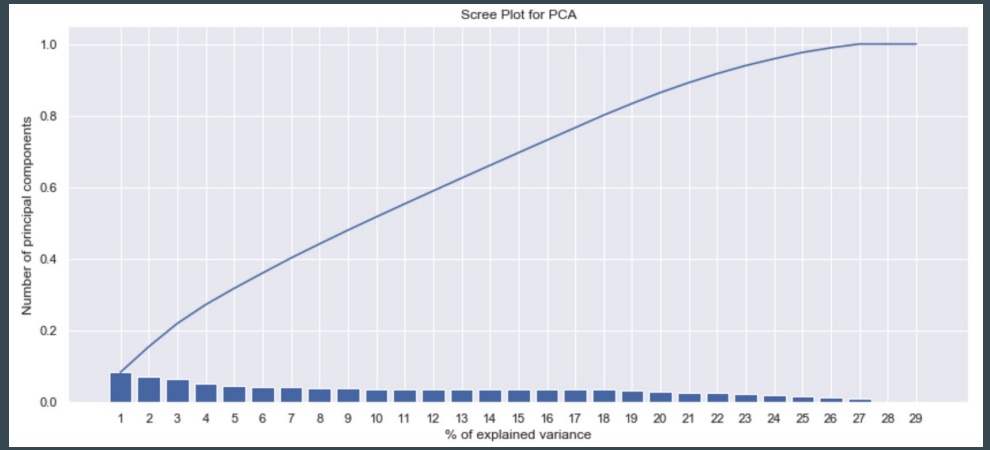
- 97% des clients n'ont placé qu'une seule commande (86,525 observations sur 89,202)
- Limitation pour déterminer les caractéristiques à long-terme des clients



Modélisation

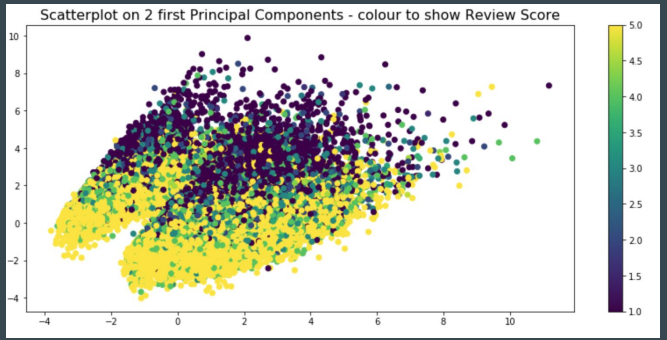
Principal Component Analysis

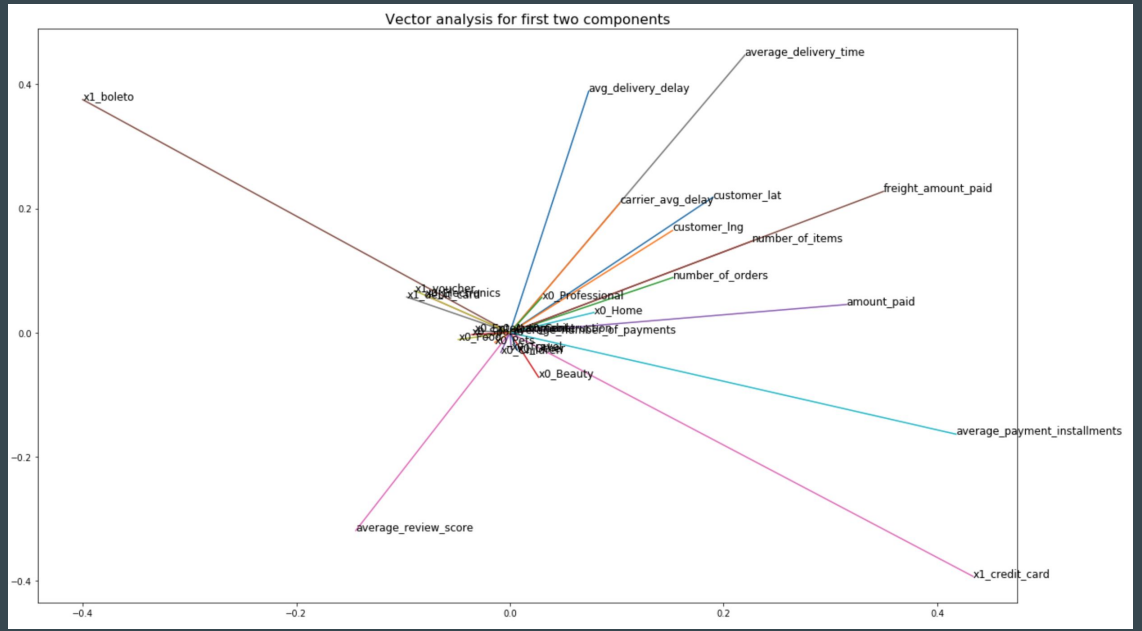
- Pas de ‘coude’ bien précis, un petit coude à 3 PCA mais seulement 22% de la variance expliquée
- Pas de réduction de composantes utilisées dans le cadre de ce modèle



Modélisation

- 15% de la variance expliquée
- PC1: Montant payé par les clients
- PC2: Score de critique et délai de livraison

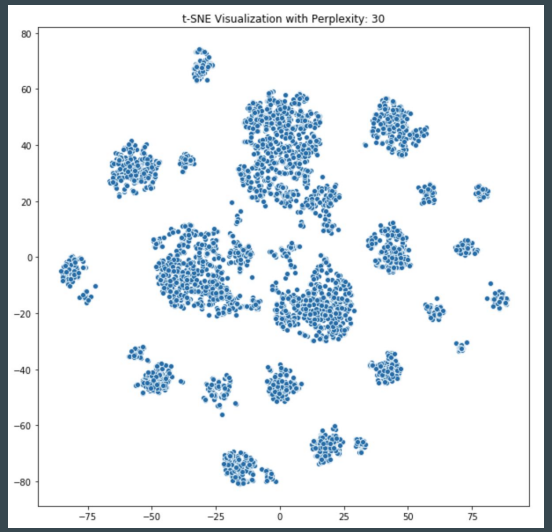


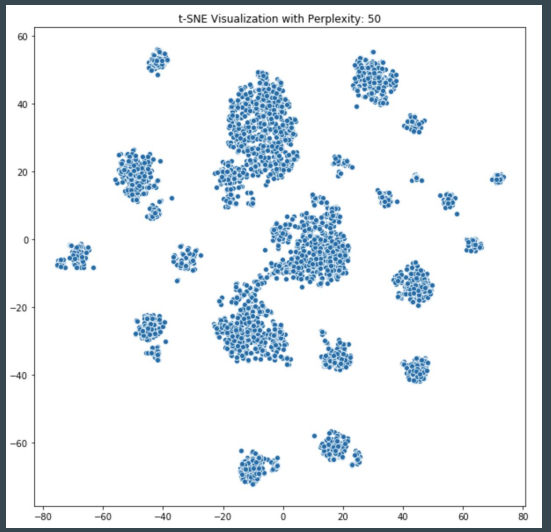


Modélisation

Visualisation avec t-SNE

- On voit déjà apparaître quelques clusters assez bien définis à partir d'une valeur de perplexité de 30
- ! le t-SNE est un algorithme volatile

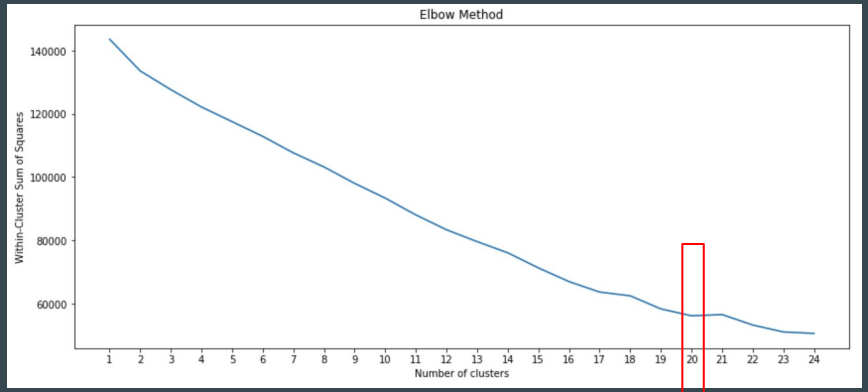


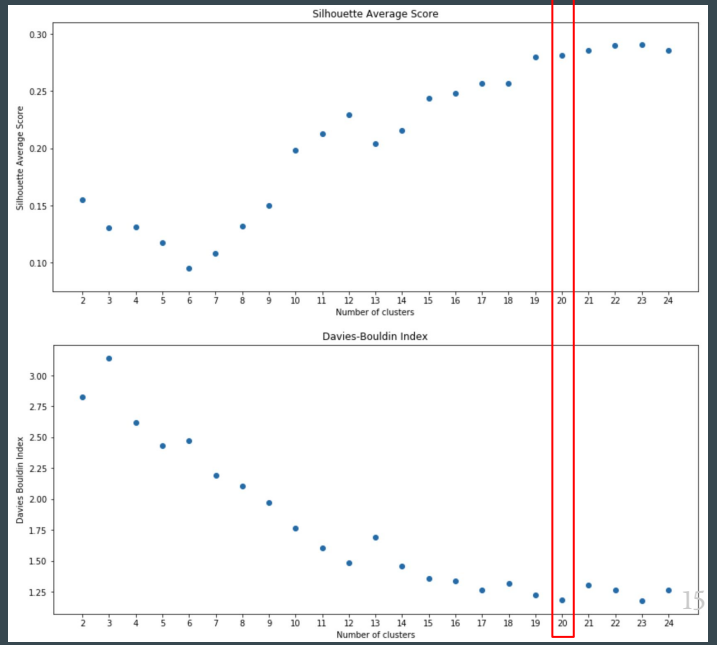


Modélisation

K-Means

- Test de l'algorithme avec plusieurs clusters sur un échantillon de 5,000 observations
- 20 clusters semble donner un bon résultat

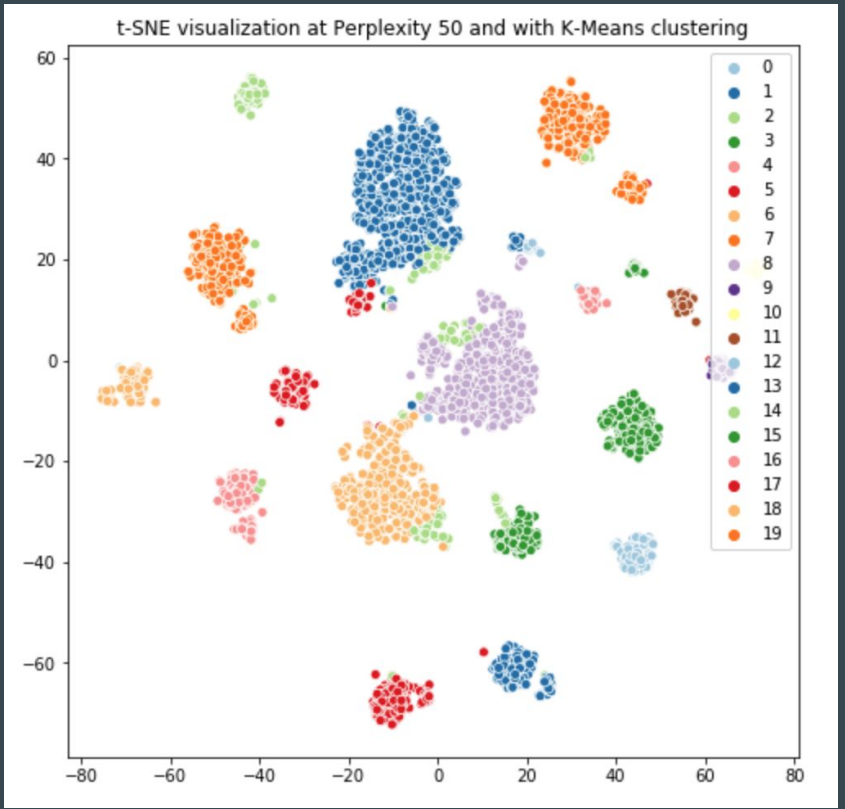




Modélisation

K-Means

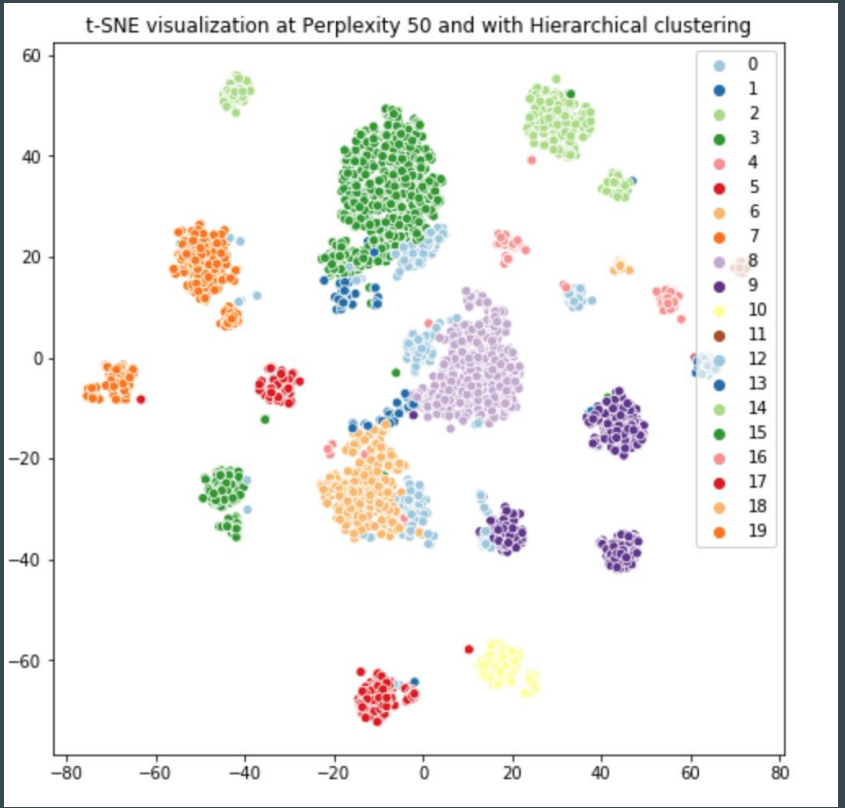
- Visualisation des clusters définis par l'algorithme K-Means sur la visualisation t-SNE.
- Les clusters ne se chevauchent pas beaucoup
- Silhouette Score: 0.28
- Davies Bouldin index: 1.18



Modélisation

Hierarchical Clustering

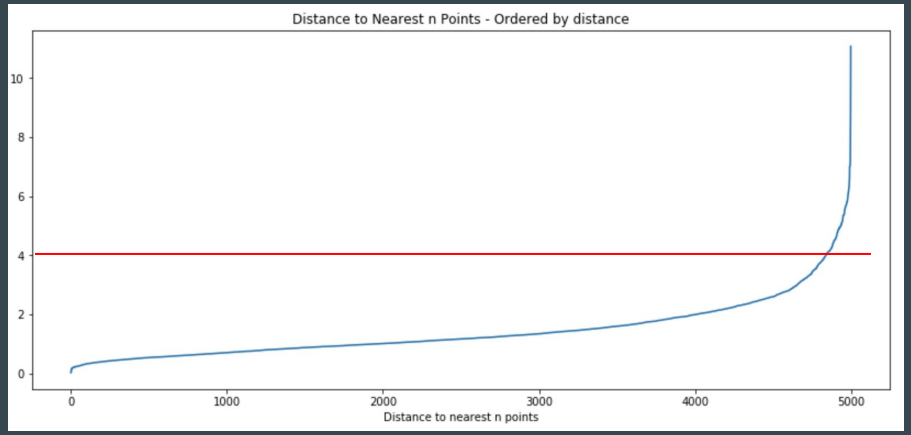
- Visualisation des clusters définis par l'algorithme Hierarchical Clustering sur la visualisation t-SNE avec 20 clusters.
- Silhouette Score: 0.26
- Index Davies Bouldin: 1.39

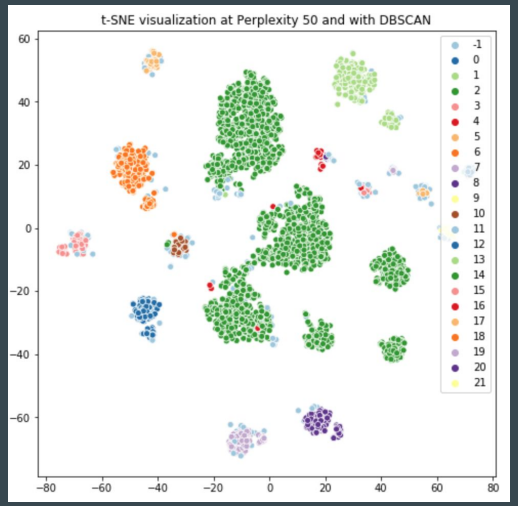


Modélisation

DBSCAN

- Visualisation des clusters définis par l'algorithme DBSCAN sur la visualisation t-SNE avec 22 clusters et Epsilon = 4
- Silhouette Score: 0.22
- Indice Davies-Bouldin: 1.88





Modélisation

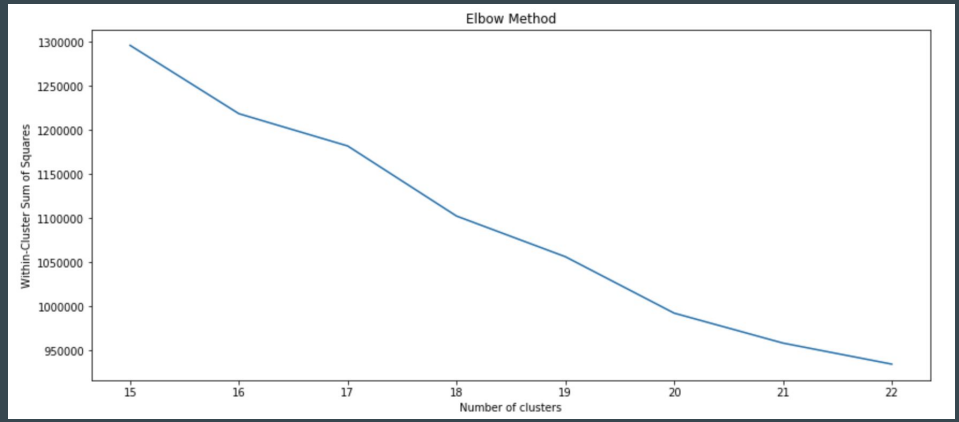
Comparaison des Algorithmes

	K-Means	Hierarchical Clustering	DBSCAN
Nombre de Clusters	20	20	22
Silhouette Score	0.28	0.26	0.22
Indice Davies-Bouldin	1.18	1.39	1.88

Modélisation

K-Means sur le jeu de données complet

- Coude à 20 clusters
- Silhouette Score: 0.28
- Indice Davies-Bouldin: 1.18
- Avec 3 itérations, les scores changent très peu (+-5%) et restent meilleurs qu'avec un autre nombre de clusters. Le nombre de clusters sélectionné est donc stable.



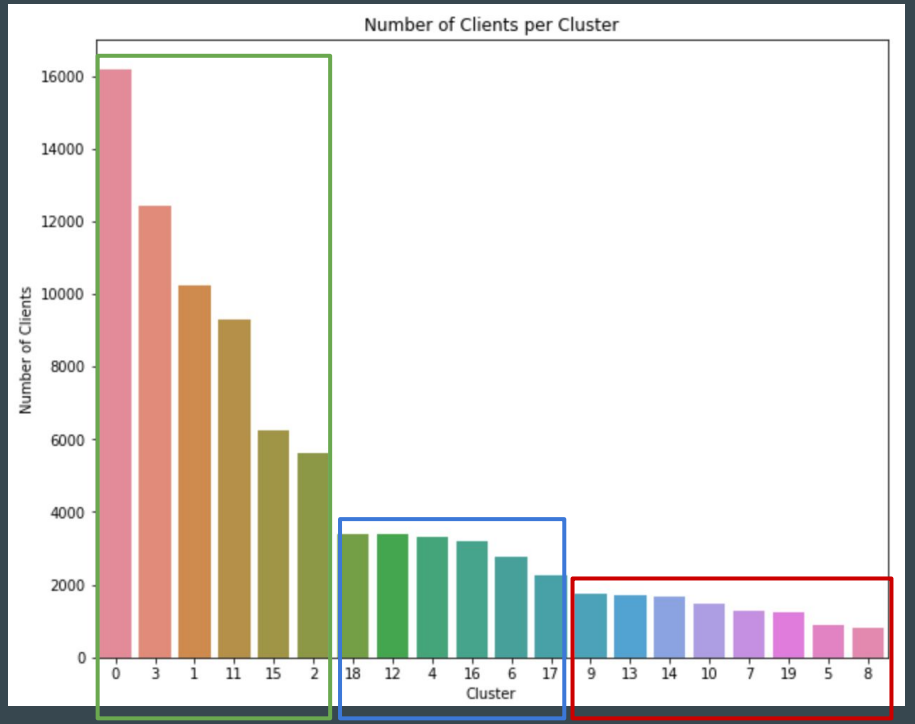
Modélisation

Stabilité dans le temps

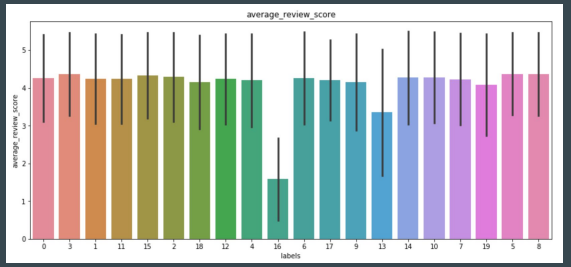
- Evaluation de la stabilité des clusters dans le temps avec les mesures ARI et AMI.
- Comparaison des clusters détectés en enlevant le dernier mois de données à chaque itération
- L'indice tombe de +-0.10 points après **1 mois.**

Interprétation

- Les clusters ne sont pas disproportionnés, il n'y a pas un grand cluster qui est beaucoup plus large que les autres.
- 6 clusters principaux se démarquent (encadré vert) et représentent 67.3% des clients.
- 6 clusters secondaires (encadré bleu) représentent 20.5% des clients
- Les 8 autres clusters (encadré rouge) représentent 12.2% des clients



Interprétation



Clusters Principaux

Cluster	Nom de Cluster	% de clients	Description
0	Home	18%	Clients dont la majorité des dépenses est dans la catégorie “Home”
3	Beauty	14%	Clients dont la majorité des dépenses est dans la catégorie “Beauty”
1	Boleto	11.5%	Clients qui paient en un seul versement par Boleto
11	Electronics	10.5%	Clients dont la majorité des dépenses est dans la catégorie “Electronics”
15	Sports	7%	Clients dont la majorité des dépenses est dans la catégorie “Sports”
2	Children	6%	Clients dont la majorité des dépenses est dans la catégorie “Children”

Interprétation

Clusters Secondaires

Cluster	Nom de Cluster	% de clients	Description
18	Professional	3.8%	Clients dont la majorité des dépenses est dans la catégorie “Professional”
12	Other	3.8%	Clients dont la majorité des dépenses est dans la catégorie “Other”
4	Automobile	3.7%	Clients dont la majorité des dépenses est dans la catégorie “Automobile”
16	Dissatisfied	3.6%	Clients dont les critiques sont négatives et qui sont victimes de livraisons retardées. Ces clients habitent généralement plus au nord-ouest du pays.
6	Entertainment	3.1%	Clients dont la majorité des dépenses est dans la catégorie “Entertainment”
17	Returning Customers	2.6%	Clients qui ont passé plus d’une commande chez Olist

Interprétation

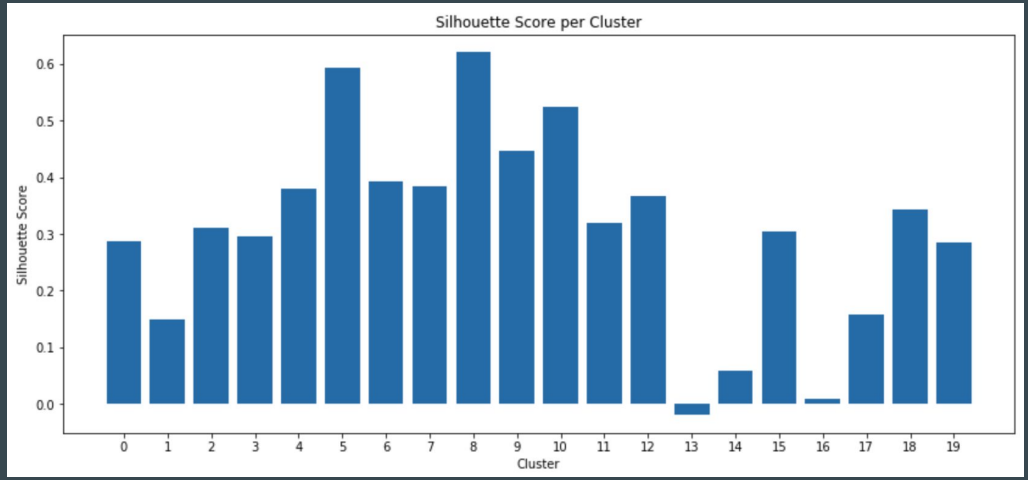
Clusters Tertiaires

Cluster	Nom de Cluster	% de clients	Description
9	Construction	2%	Clients dont la majorité des dépenses est dans la catégorie “Construction”
13	Large Orders	1.9%	Clients qui achètent plus de produits par commande que la moyenne. Ils paient plus de frais de transport et leur critique est plus négative que la moyenne (⅗ vs ⅘)
14	Large Payments	1.9%	Clients dont les commandes sont plus chères que la moyenne. Ils paient en plusieurs versements
10	Pets	1.6%	Clients dont la majorité des dépenses est dans la catégorie “Pets”
7	Debit Card	1.4%	Clients qui paient par carte de débit et en un seul versement
19	Vouchers	1.4%	Clients qui paient par vouchers, typiquement en plusieurs versements et dépensent de plus petits montants
5	Travel	1%	Clients dont la majorité des dépenses est dans la catégorie “Travel”
8	Food	0.9%	Clients dont la majorité des dépenses est dans la catégorie “Food”

Interprétation

Score silhouette par cluster

- Les clusters liés à une catégorie de produit ont un meilleur score Silhouette que la moyenne.
- Les clusters liés à d'autres features que la catégorie d'achat la plus importante ont des scores plus bas (overlap avec les clusters liés aux catégories de produits)



Contrat de Maintenance

- Sur base de l'analyse de stabilité temporelle, algorithme à ré-entraîner 1x par mois en prenant en compte les nouvelles données.
- Evolution de la quantité de nouvelles données reçues par mois à suivre également

Conclusions

Clusters

- Définition de 20 clusters pour les clients de Olist grâce à l'algorithme K-Means
- Ces clusters donnent des informations sur les préférences des clients, leurs habitudes de consommation et leur attitude vis-à-vis du service de Olist
- Certains clusters donnent des nouvelles informations sur les clients

Limitations

- La majorité des clients n'ont effectué qu'une commande chez Olist, il serait intéressant d'évaluer les habitudes des clients sur plusieurs commandes.
- Les catégories de produits définies lors de notre exercice pourraient avoir une influence sur le clustering, la qualité du clustering dépendrait donc partiellement de notre classification.
- Une partie des clusters n'apporte pas de nouvelles observations ou corrélations

Perspectives

Pour aller plus loin

- Tester l'algorithme sur plus de données, notamment plus de clients avec des commandes multiples
- Utiliser plus d'information sur les clients (âge, sexe, plateforme utilisée pour l'achat)
- Utiliser plus d'information sur les produits achetés (nouveaux produits?)
- Création de features additionnelles (par exemple analyse de tags du contenu des critiques écrites, regarder les marques achetées)
- Classification additionnelle des clients via une analyse RFM
- Analyser les caractéristiques des clients qui reviennent pour comprendre comment fidéliser la clientèle