SEGMENTEZ DES CLIENTS D'UN SITE E-COMMERCE

PROJET 4 - MARTIN VIELVOYE

OPEN CLASSROOM - INGÉNIEUR MACHINE LEARNING

PROBLÉMATIQUE

FOURNISSEZ AU ÉQUIPES D'E-COMMERCE DE OLIST UNE SEGMENTATION DES CLIENTS QU'ELLES POURRONT UTILISER AU QUOTIDIEN POUR LEURS CAMPAGNES DE COMMUNICATION.

9 datasets composants une partie de la base de données clients du site internet Olist. Les bases de données décrivent l'historique de commandes et d'achats avec le site internet des clients.

Data set initial

- 9 bases de données reliées par des clés uniques.
- Plus de 96,000 clients pour presque 110,000 articles passé sous commande.
- Features dont :
    - Les dates (et prix) de commandes, de livraison, statut de livraison
    - Des infos géographiques sur les clients ainsi que leurs évaluation de la commande et si ils ont laissé des commentaires. Leurs paiements et moyens de paiements sont aussi décris.
    - La categories des produits achetés et de leurs vendeurs ainsi qu'une description physique des produits.

Pistes :

- Création de nouvelles variables.
- Clustering de la données une fois compressé a basse dimensions.
- Segmentation RFM.

PRÉPARATION

Nettoyage

- Conservation des colonnes pertinentes pour les modèles ou les manipulations de données. (*17 sur 47*)
- Analyse et nettoyage des dupliquas, des entrées nulles et des valeurs aberrantes. (*0%*)

Feature engineering

- Création de features :
    - Dépense totale, quantité d'articles achetés, quantité de commande, nombre de commentaires, moyenne de scores, moyenne de jours de livraison **(6)**
    - Leurs dernières interaction avec le site internet, nombres d'achat dans la dernière année, leurs dépense dans la dernière année. **(3)**
    - Moyenne d'article acheté dans chaque catégories d'article. **(71)**
- Transformation logarithmique de certaines variables pour améliorer la distribution. **(3)**

Feature engineering :
Base réduite

- Création d'une deuxième bases de données où deux groupes de features sont réduits a travers une PCA pour réduire le nombres de features totale de la base de données.
    - **Groupe 1** : Les cumuls des achats des clients reparties sur les 12 mois de l'années. Réduit a 7 components.
    - **Groupe 2** : Les cumuls des achats des clients reparties sur les 11 categories de produits. Réduit a 8 components.
- Nombres de features totale de 35 à 24.

MODÉLISATIONS

Recap

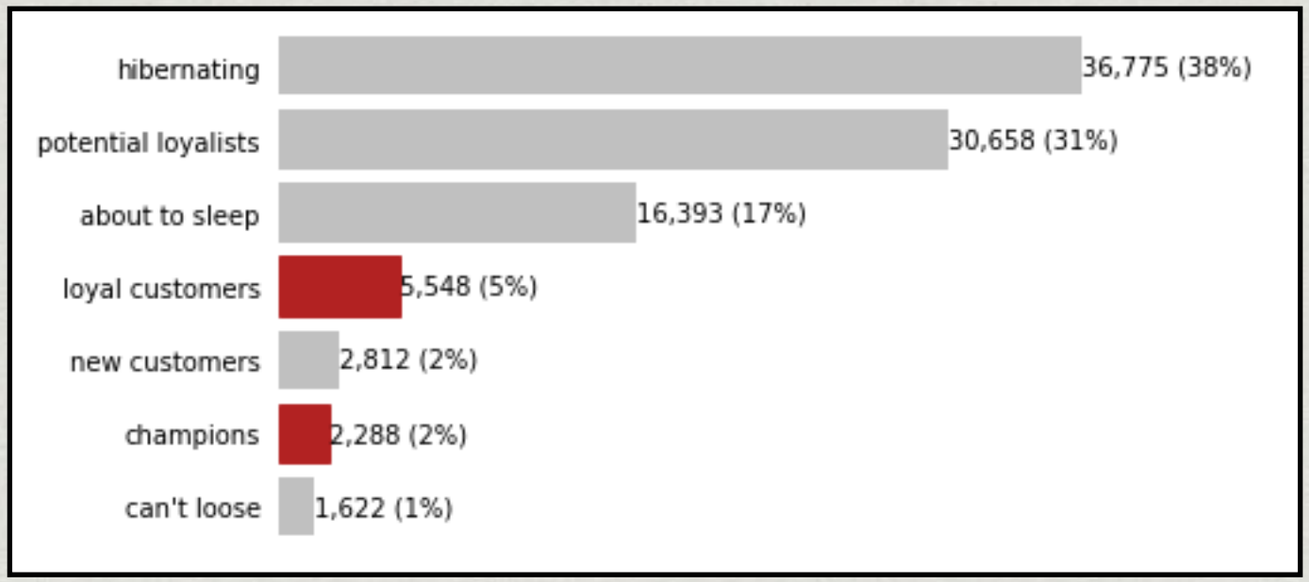
- **Visualisations** :
    - k-PCA
    - LLE
- **Clustering** :
    - kMean
    - HDBSCAN
- **Segmentation** :
    - RFM
    - Group data analysis

RFM on initial data-set

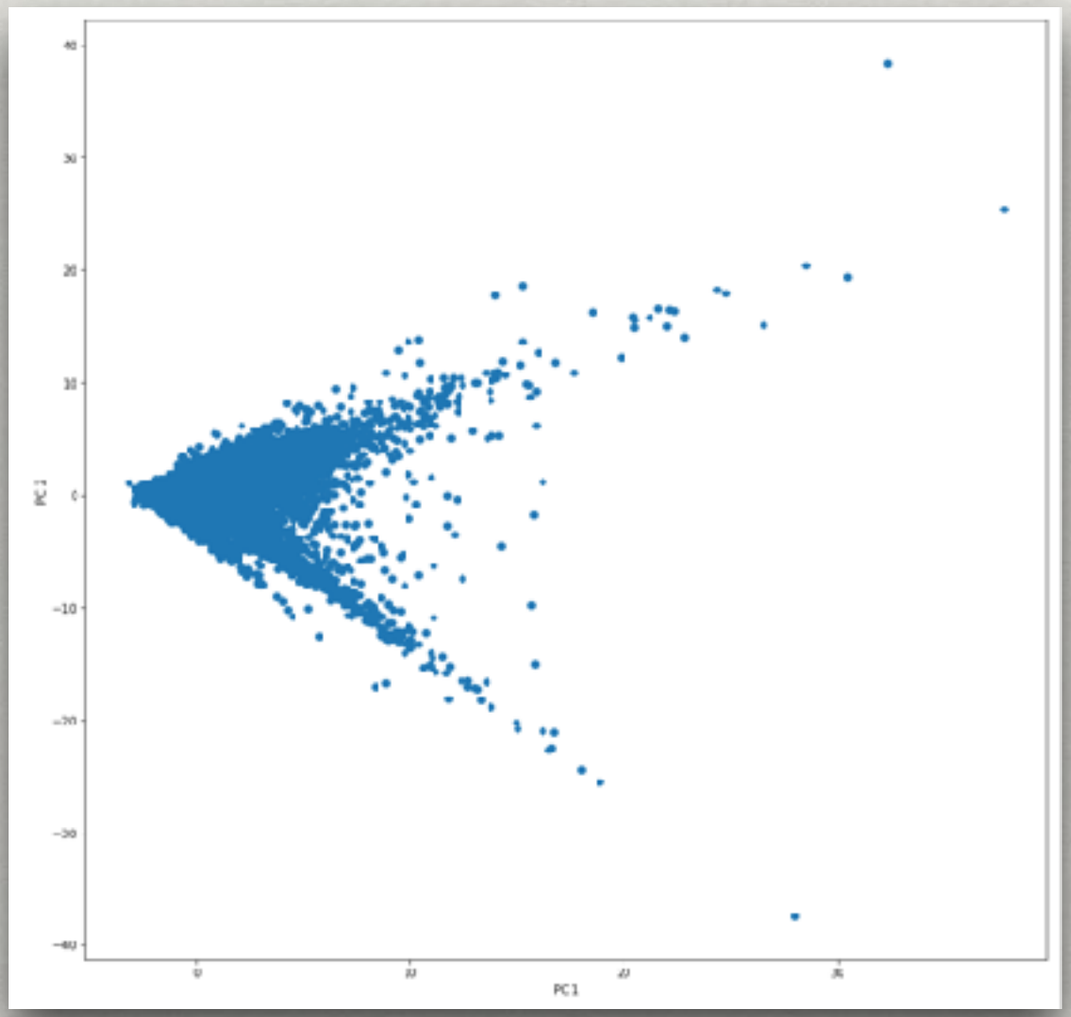
Segment Description

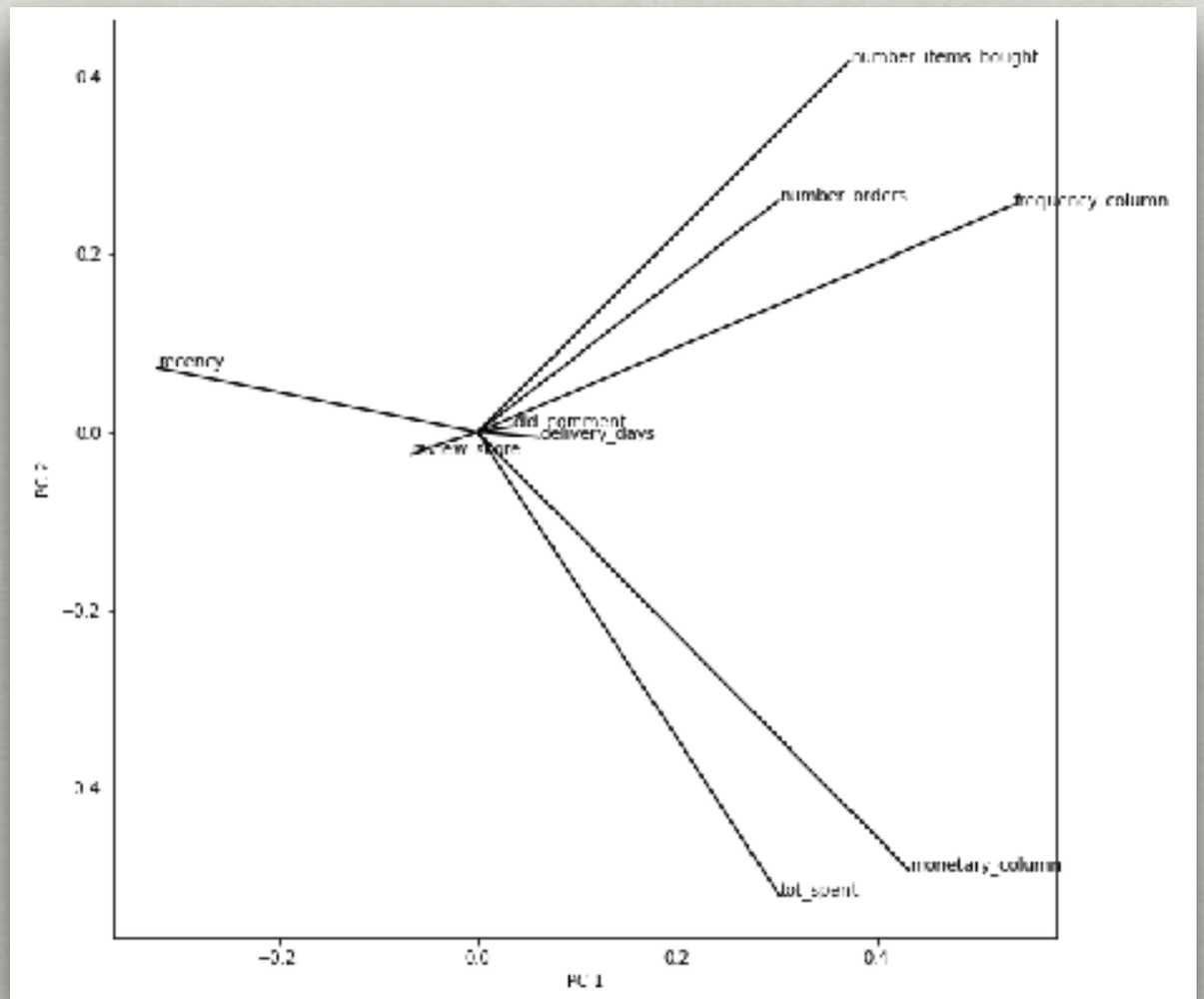
- **Champions**: Bought recently, buy often and spend the most.
- **Loyal Customers**: Buy on a regular basis. Responsive to promotions.
- **Potential Loyalist**: Recent customers with average frequency.
- **Recent Customers**: Bought most recently, but not often.
- **Promising**: Recent shoppers, but haven't spent much.
- **Customers Needing Attention**: Above average recency, frequency and monetary values. May not have bought very recently though.
- **About To Sleep**: Below average recency and frequency. Will lose them if not reactivated.
- **At Risk**: Purchased often but a long time ago. Need to bring them back!
- **Can't Lose Them:** Used to purchase frequently but haven't returned for a long time.
- **Hibernating**: Last purchase was long back and low number of orders. May be lost.

RFM Categories Repartition

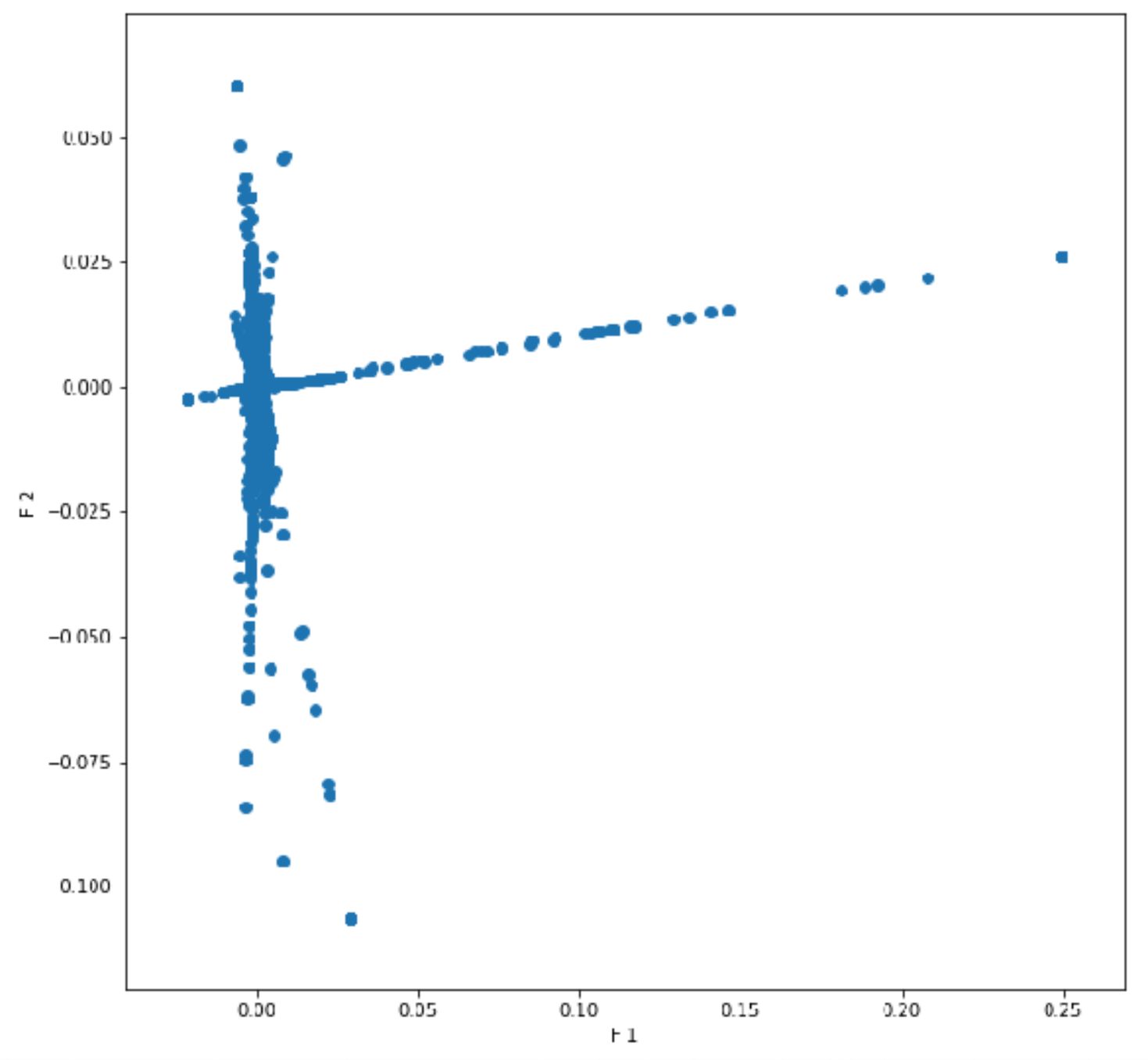


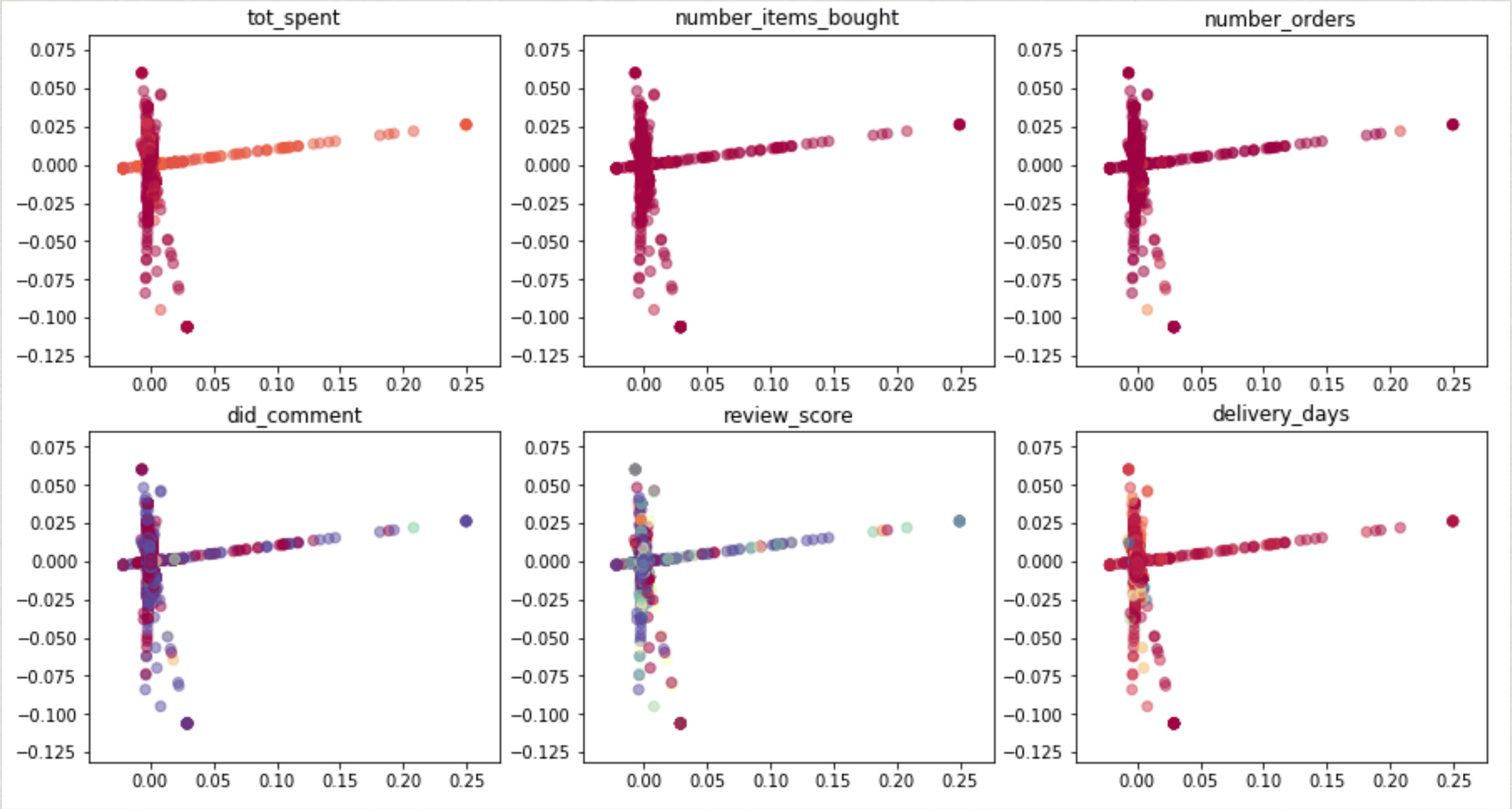
PCA on initial data-set



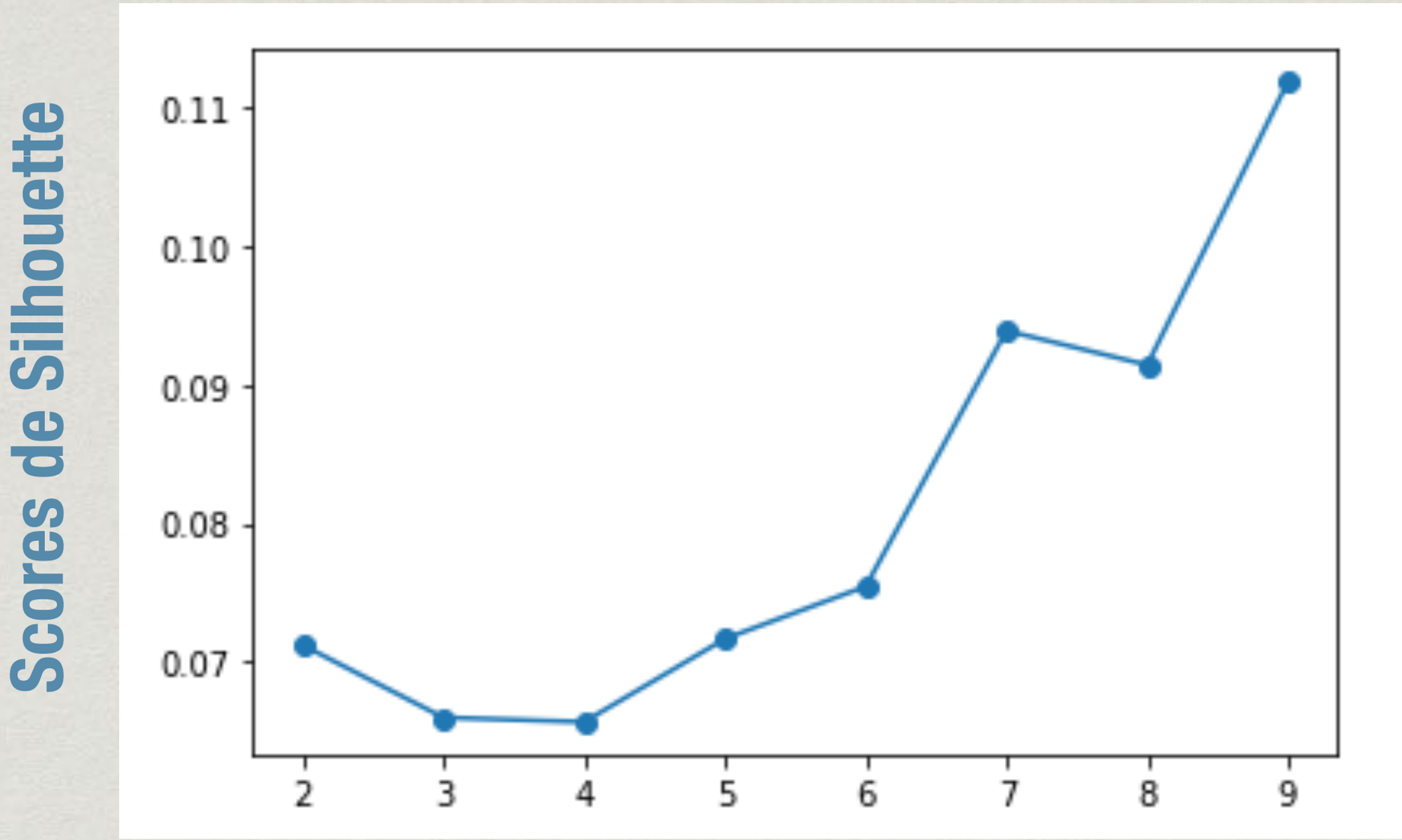


LLE on initial data-set




tot_spent
number_items_bought
number_orders
did_comment
review_score
delivery_days
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
−0.125
0.00
0.05
0.10
0.15
0.20
0.25

KMean


Scores de Silhouette
0.11
0.10
0.09
0.08
0.07
2
3
4
5
6
7
8
9
Nombres de Clusters

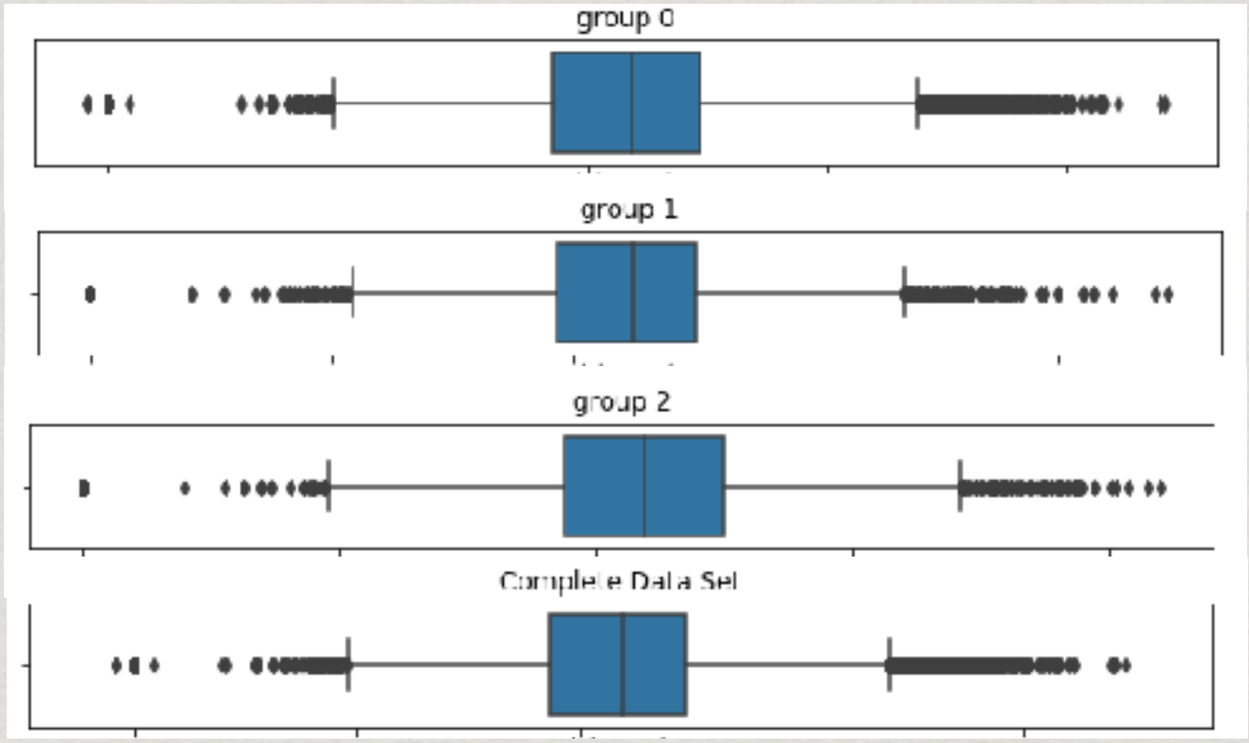
3 GROUPES

- **GROUPE 1** : Client ‘moyen’, groupe le plus general ainsi que le plus peuplé

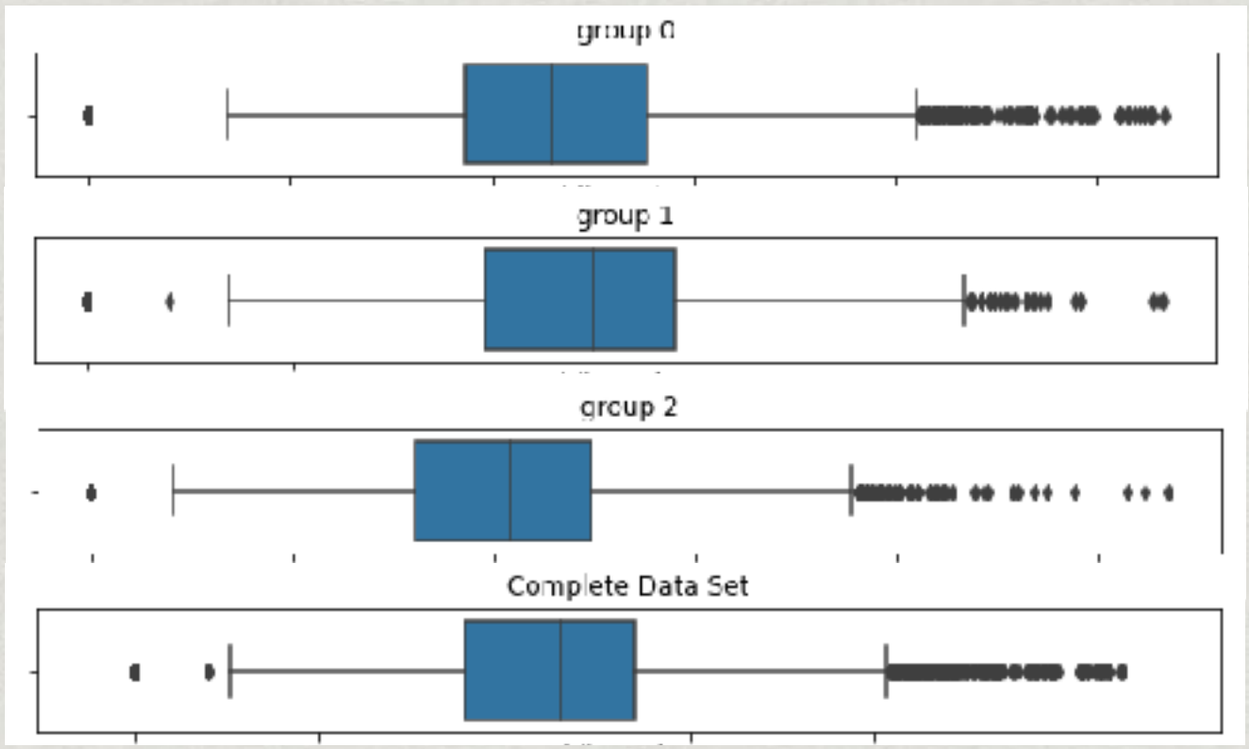
- **GROUPE 2** : Fortes dépenses et nombre d’articles achetés.

- **GROUPE 3** : Livraison rapide, moyenne de satisfaction élevée

Total Money Spent

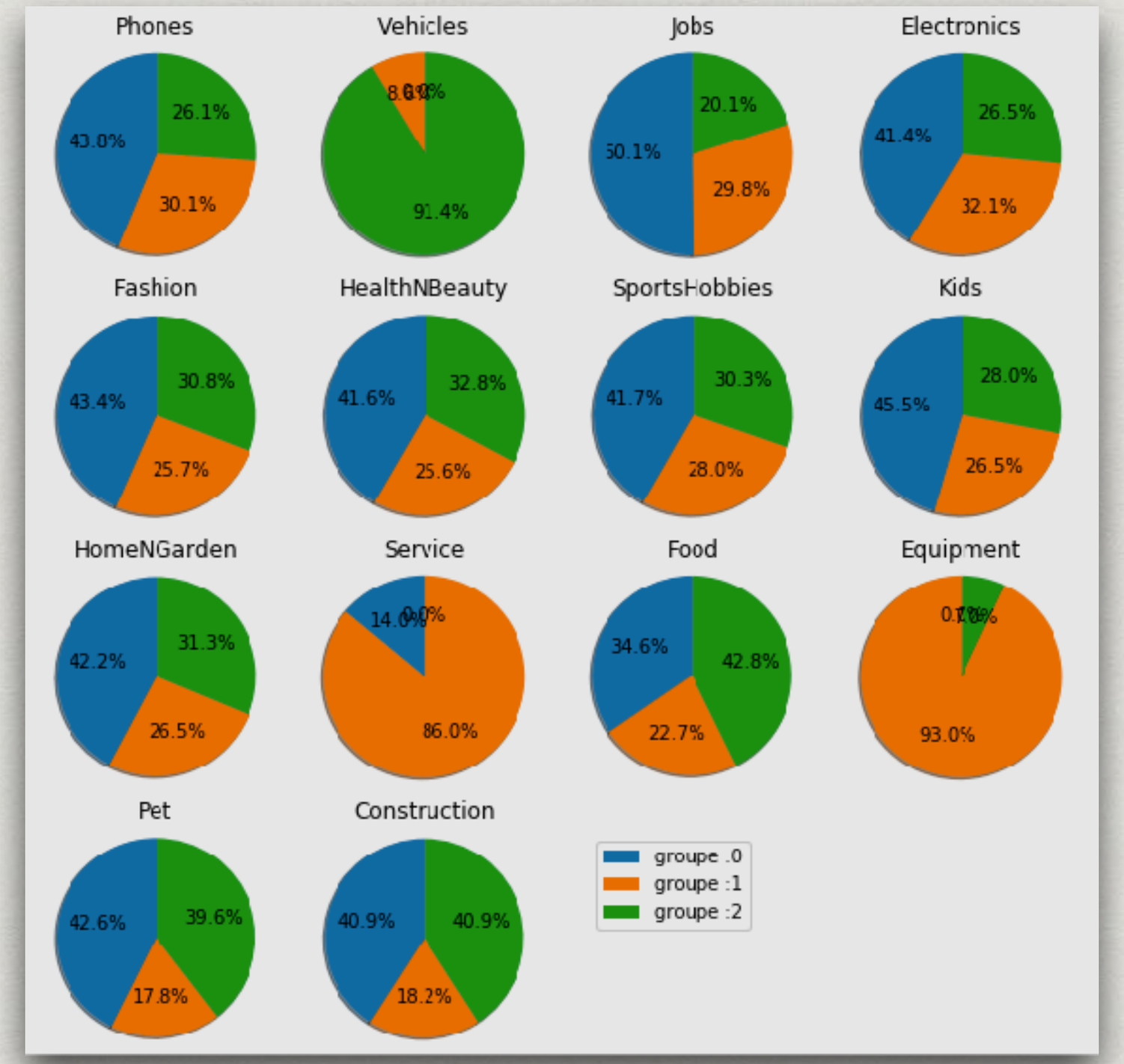


Average Delivery Length (days)

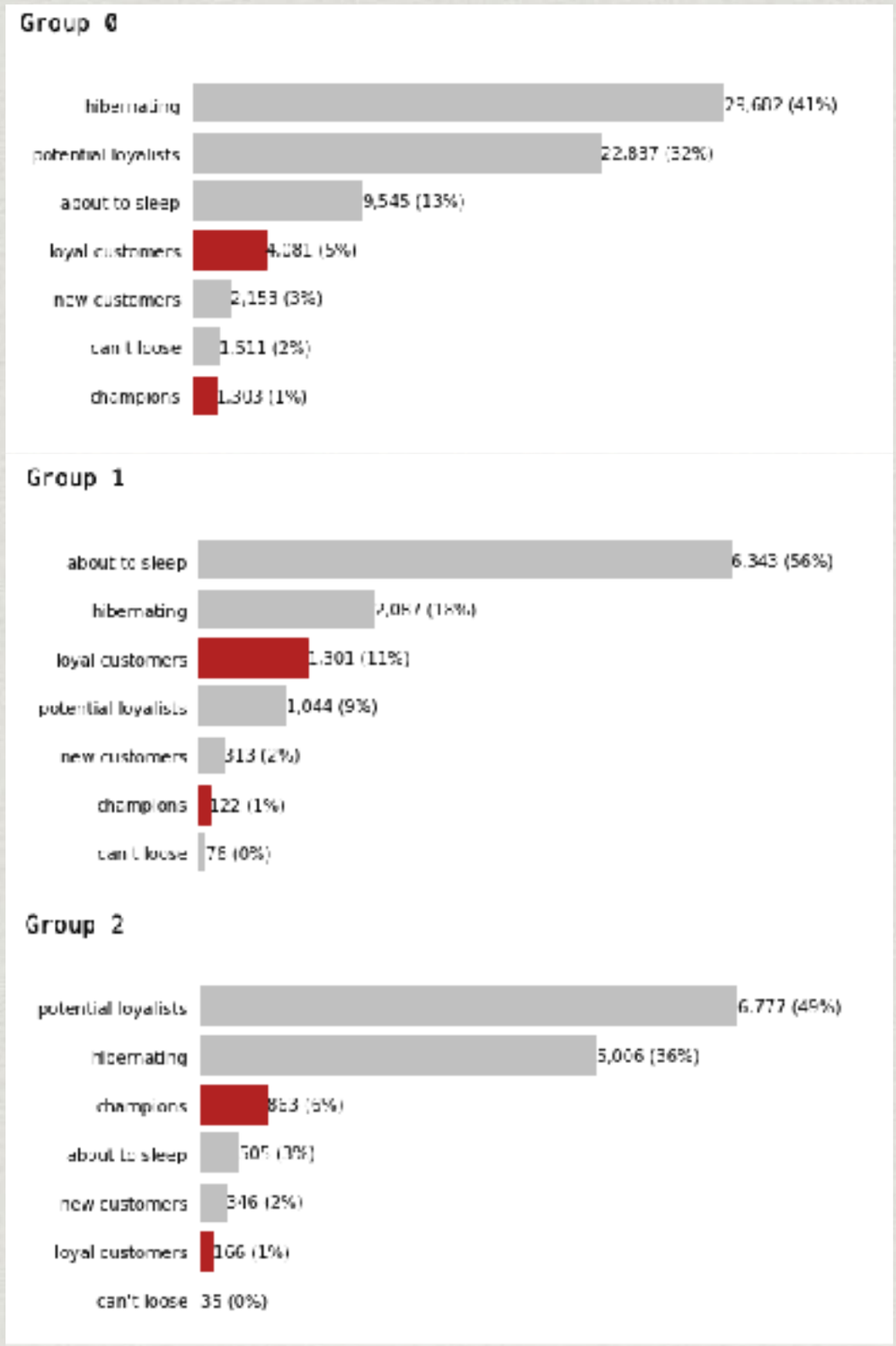


KMean

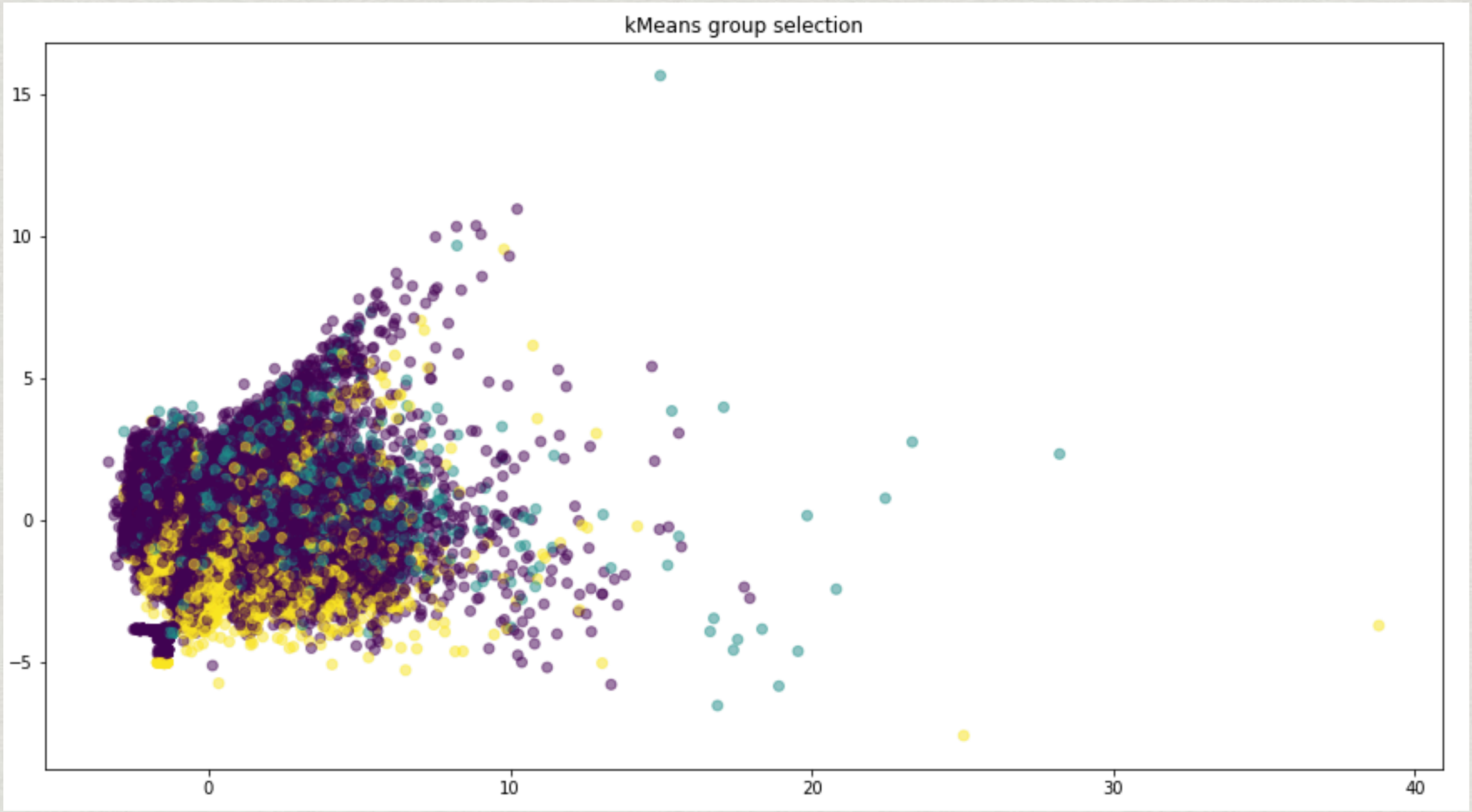
Product Average mean comparison of purchases



Kmean. RFM Analysis



HDBSCAN. PCA



MODÉL FINAL

HDBSCAN

Hierarchical Clustering

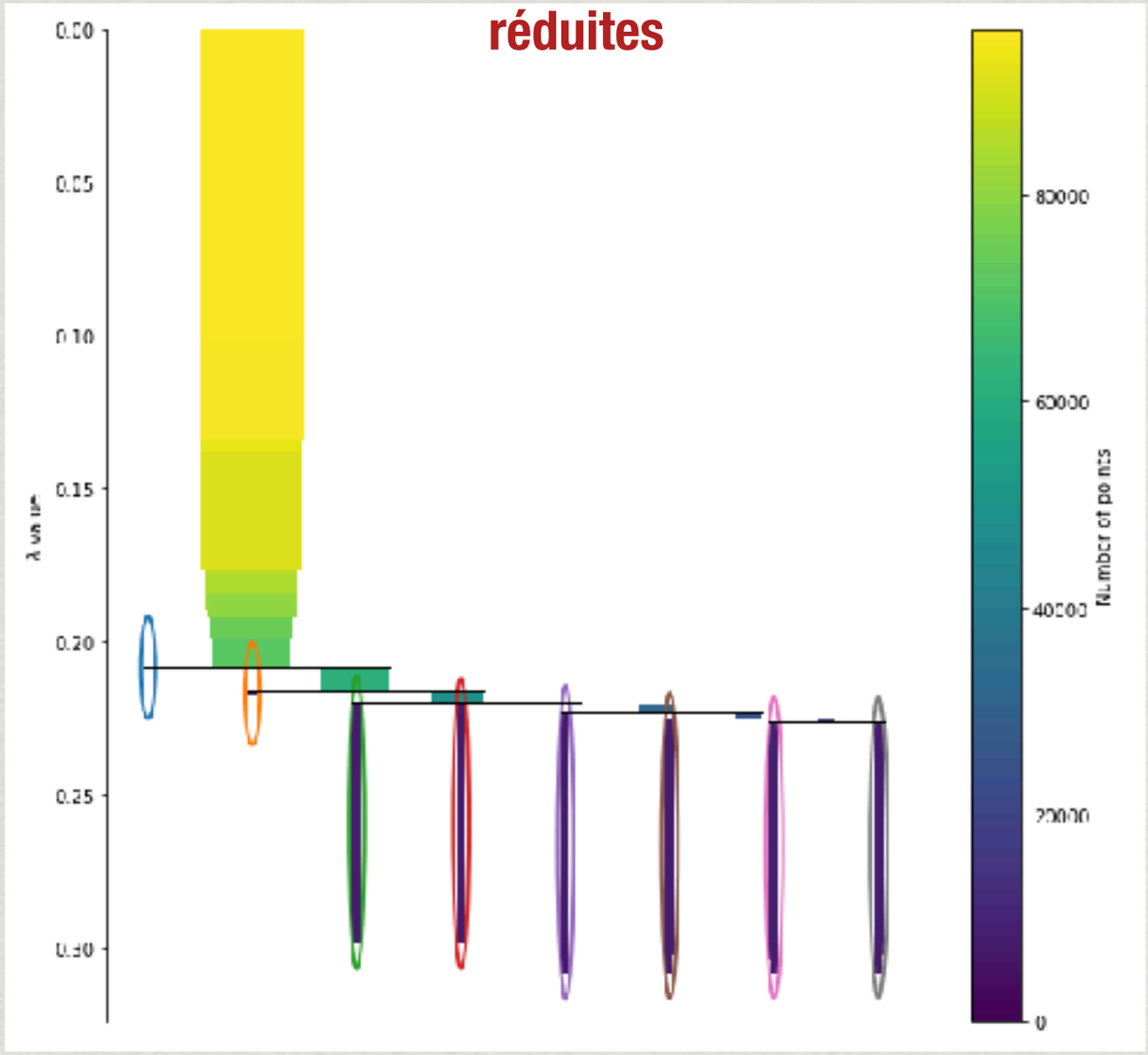
Jeu de donnés complet

- 8 Clusters
- Taille minimale : 7300
- % Outliers : 38.1

Jeu de donnés réduit (PCA)

- 8 Clusters
- Taille minimale : 4350
- % Outliers : 31.1

Structure hiérarchique pour le jeux de données réduites

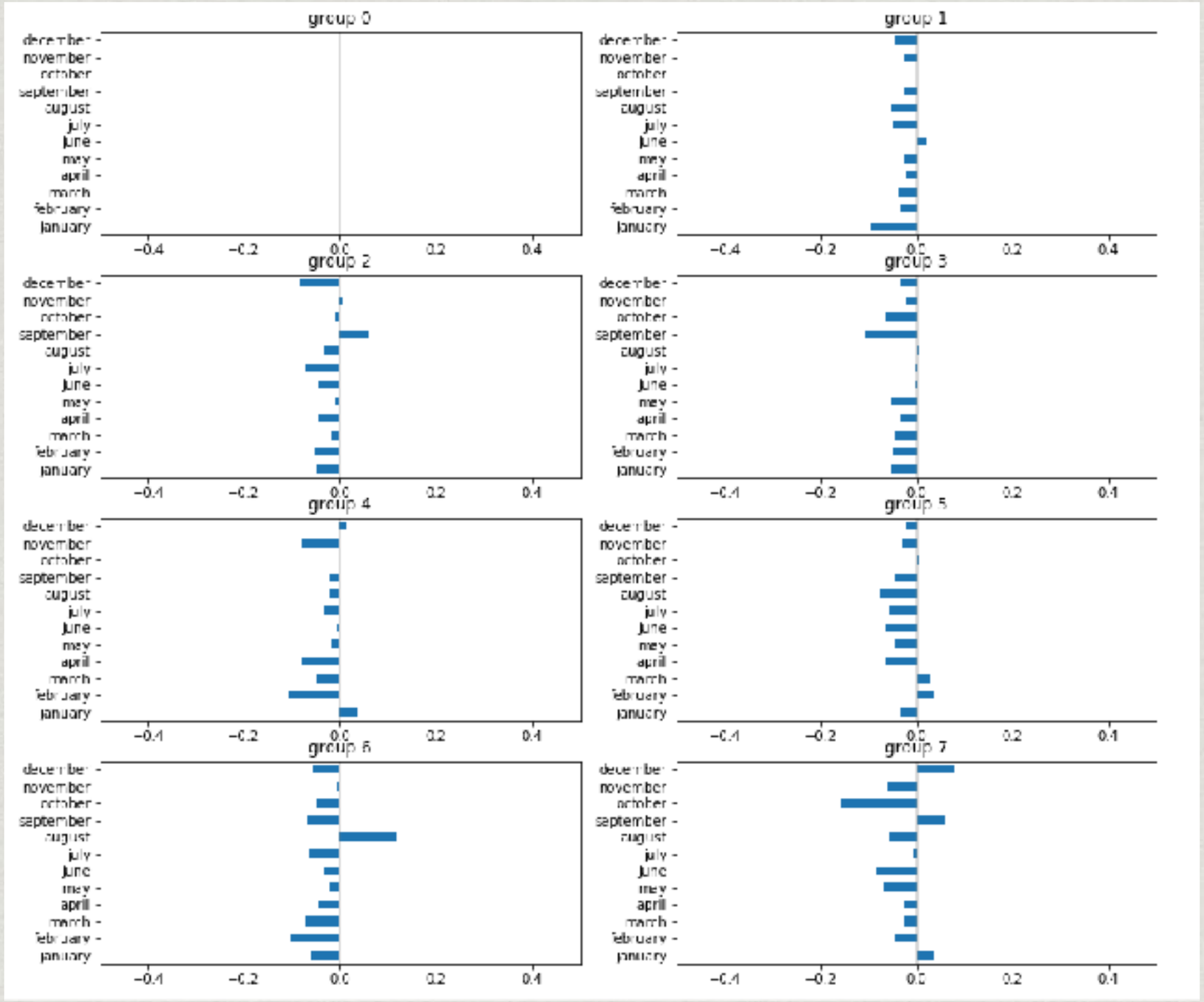


8 GROUPES

- **GROUPES** :
  - Groupes très équilibré dans leurs representation RFM et leurs poids moyen dans les 6 variables créées.
  - Une séparation plus prononcée dans les habitudes des mois d'achats et des catégories de produit acheté.
- **OUTLIERS** : Groupe assez peuplé. Ici résideront ceux qui ont un nombre de commandes et d'articles acheté élevé.

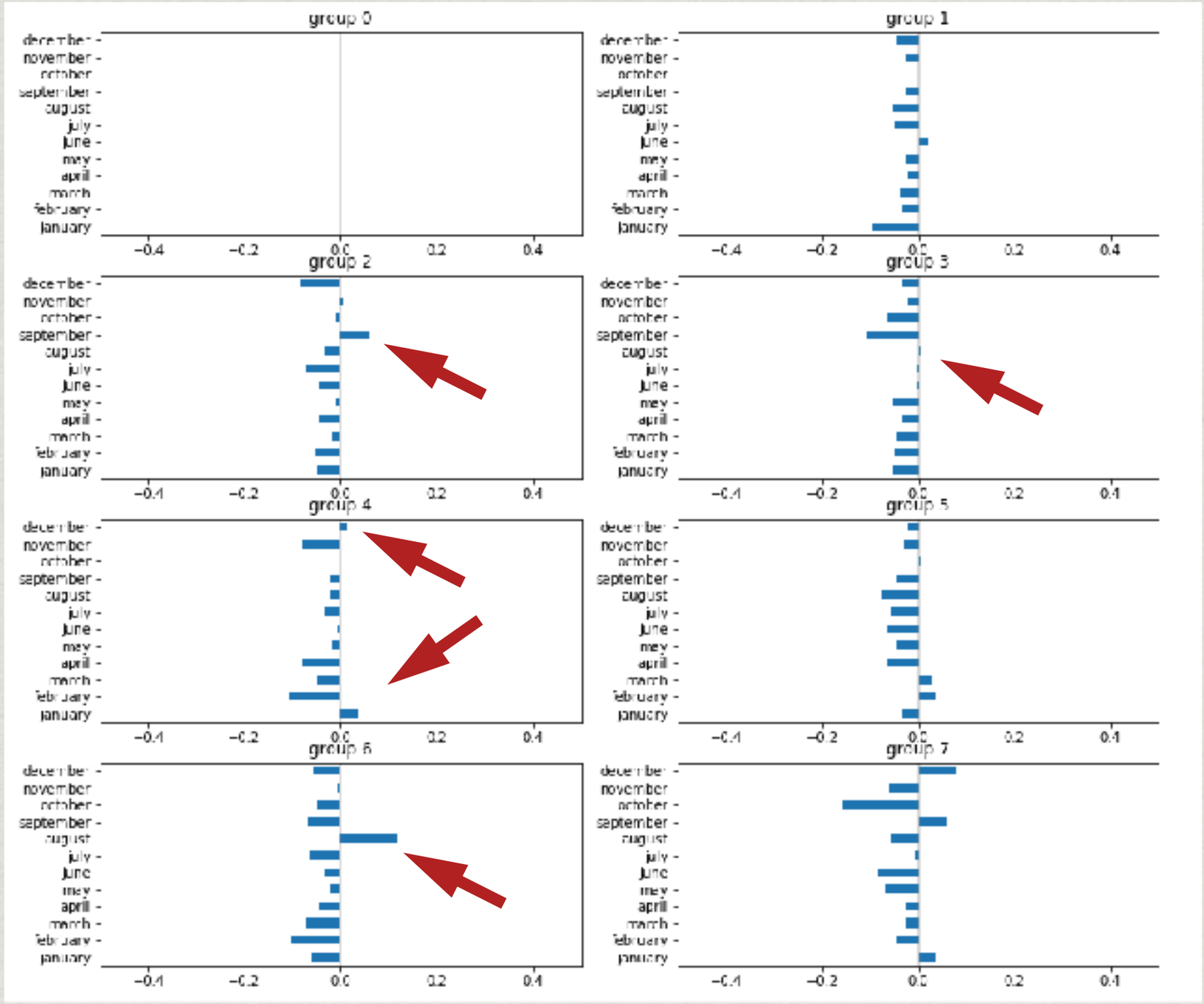
HDBSCAN

Month Average mean comparison of month purchases

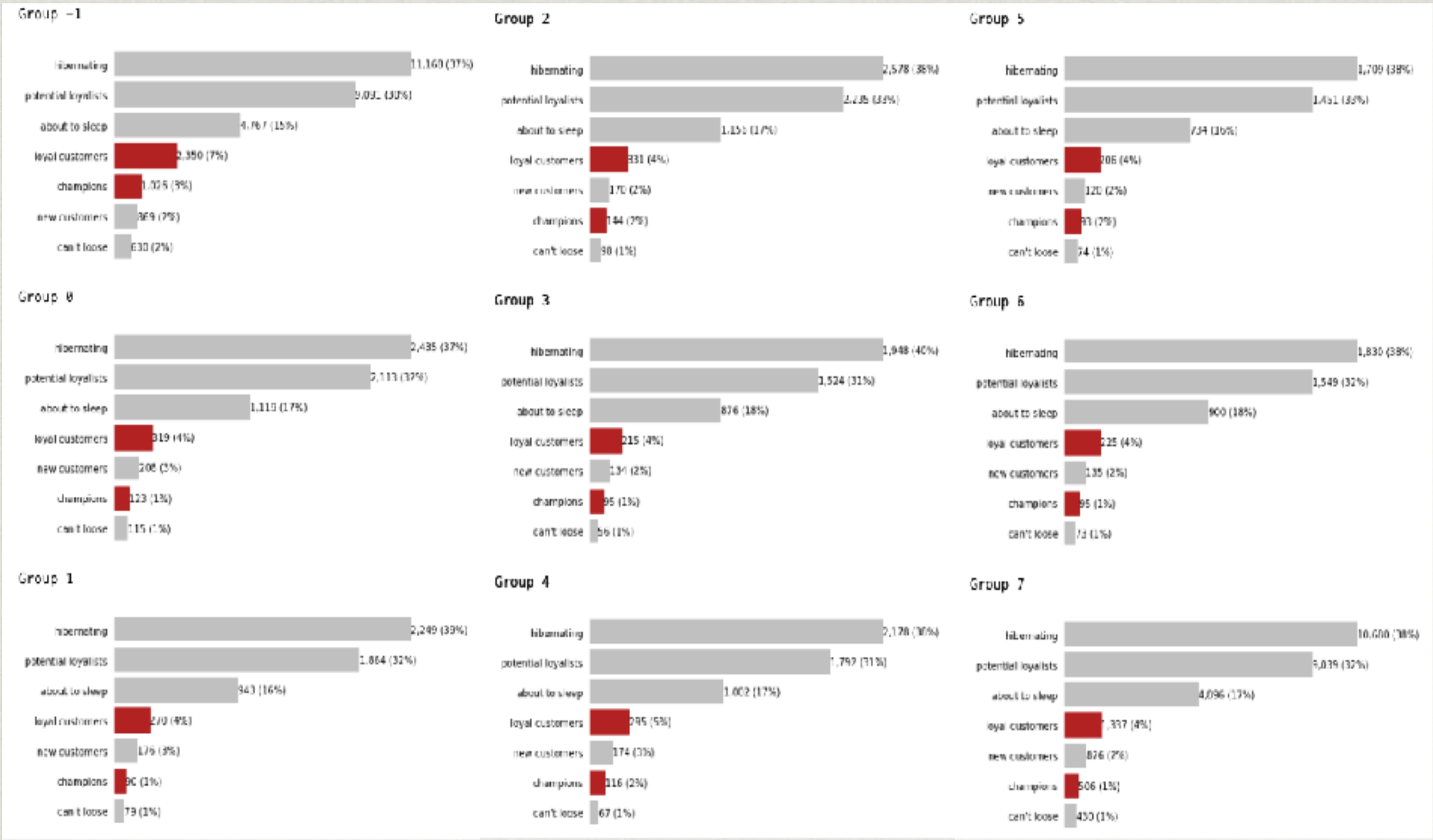


HDBSCAN

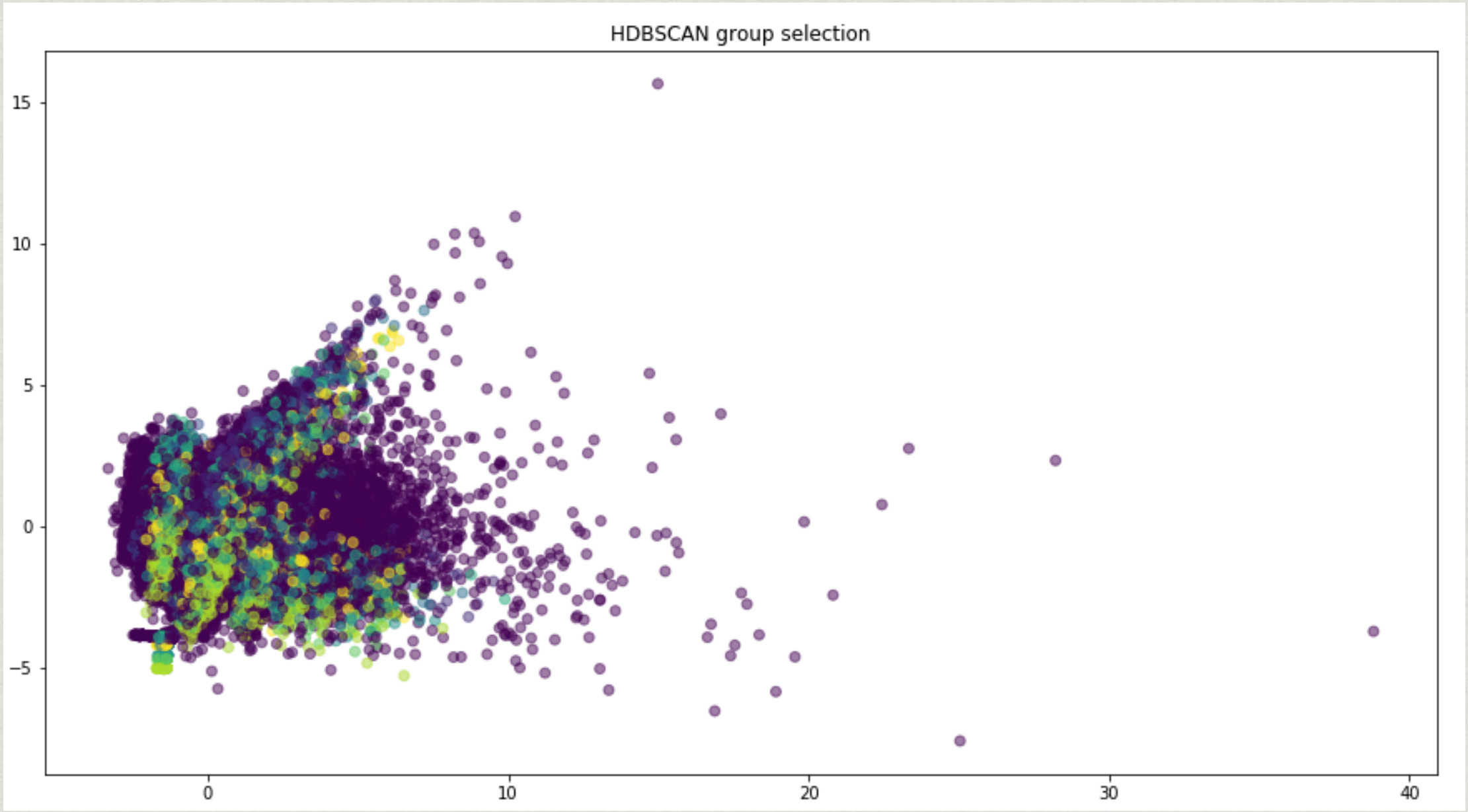
Month Average mean comparison of month purchases



HDBSCAN. RFM Analysis



HDBSCAN. PCA



Stratégies

- **Ajout de nouveaux clients** :
    - Rajout presque « supervisé » car on peut rapidement placer un client et calculer la distance aux clusters déjà présent.
- **Rafraichissement de la segmentation :**
    - La segmentation RFM a un calibre de precision au mois près. Plus sa précision est forte, plus son rafraichissement de calcul est frequent pour être pertinent.
    - La segmentation HDBSCAN ne varie que faiblement a l'ajout de nouveau clients mais un rafraichissement mensuel est maximal afin de garder une segmentation « à jour » et pertinente.