

# T-DAT-901

Groupe 17 :

Lyne ARMAND  
Pierre BOBARD  
Dylan HOCHBERGER  
Quentin SOMMER  
Mathieu SOMMER

# SOMMAIRE

---

## PRÉSENTATION D'ENSEMBLE DU PROJET

### **4** **PRÉSENTATION DE L'ÉQUIPE**

---

### **5** **CONTEXTE DU PROJET**

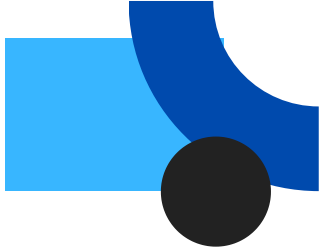
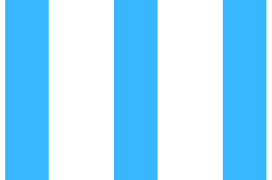
---

### **5** **LANGAGE DE PROGRAMMATION**


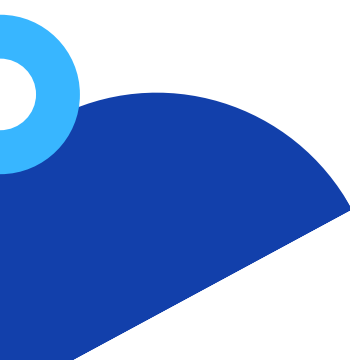
---

### **6** **PRE-PROCESSING**

---



# Présentation de l'ensemble du projet



# PRÉSENTATION DE L'ÉQUIPE

Nous sommes le groupe 17.



**Lyne Armand**

Développeur Frontend



**Pierre BOBARD**

Développeur Backend



**Dylan HOCHBERGER**

Développeur Frontend



**Quentin SOMMER**

Développeur Backend



**Mathieu SOMMER**

Développeur Backend

## CONTEXTE DU PROJET

Une entreprise vous donne accès à KaDo: une base de données contenant des millions d'articles achetés.

Afin de fidéliser les clients, l'entreprise attend de vous :

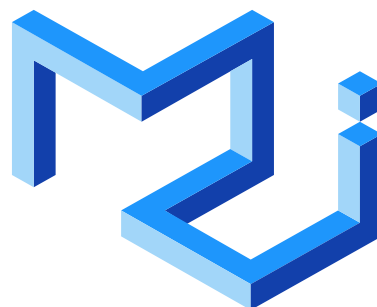
- de segmenter les clients afin d'avoir une vision plus claire de la situation
- d'ajouter des graphiques et des chiffres pour aider cette entreprise à visualiser les profils de ses clients
- de construire un système de recommandation pour offrir un cadeau à chaque client
- d'utiliser différents types de système de recommandation : basé sur l'utilisateur, sur l'article
- rédiger un kickoff pour décrire précisément ce que vous allez faire avec l'ensemble des données

## LANGAGE DE PROGRAMMATION

Pour le backend, on utilise :



Pour le frontend, on utilise :



Bibliothèque MUI

# PRE-PROCESSING

## 1 / Nettoyage

La première étape dans le traitement du dataset est le nettoyage des données que comporte ce dernier.

- **Traitements des données manquantes** : Il est possible que certaines parties de données viennent à manquer ou restent incomplètes dans le dataset, dans ce cas nous les compléterons si c'est possible, ou dans les cas échéant les écarterons pour éviter qu'elles ne viennent fausser nos prochains traitements.
- **Traitements des données non pertinentes** : Certaines données (dans un premier temps) ne peuvent pas avoir d'intérêt ou de signification. Pour améliorer l'efficacité des futurs traitements nous les mettons de côté pour garder des résultats pertinents.

La Dataset étant maintenant plus propice aux traitements des données nous allons chercher à le transformer sous des formes plus adaptées à l'exploration de ses données. Pour cela nous allons utiliser plusieurs outils :

- **Pandas** est un package Python open source fournissant des structures de données rapides, flexibles et expressives conçues pour rendre le travail avec des données « relationnelles » ou « étiquetées » à la fois simple et intuitif.
- **Scikit-learn** est une bibliothèque Python open source et gratuite de machine learning. Elle propose divers algorithmes de classification, de régression et de clustering.

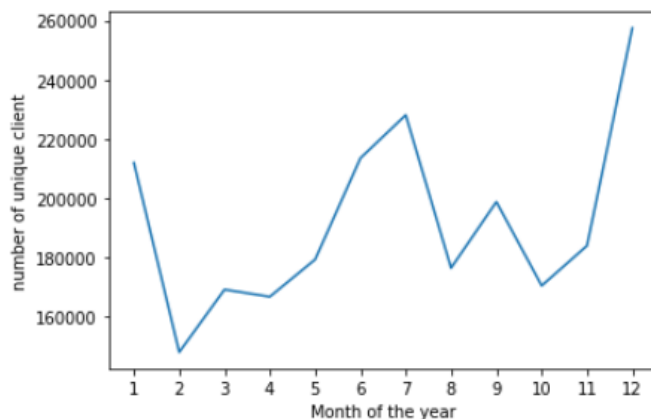


## 2 / Transformation des données

Nous cherchons maintenant à transformer les données afin d'en extraire des statistiques pertinentes et arriver à la suggestion de produits pour les clients.

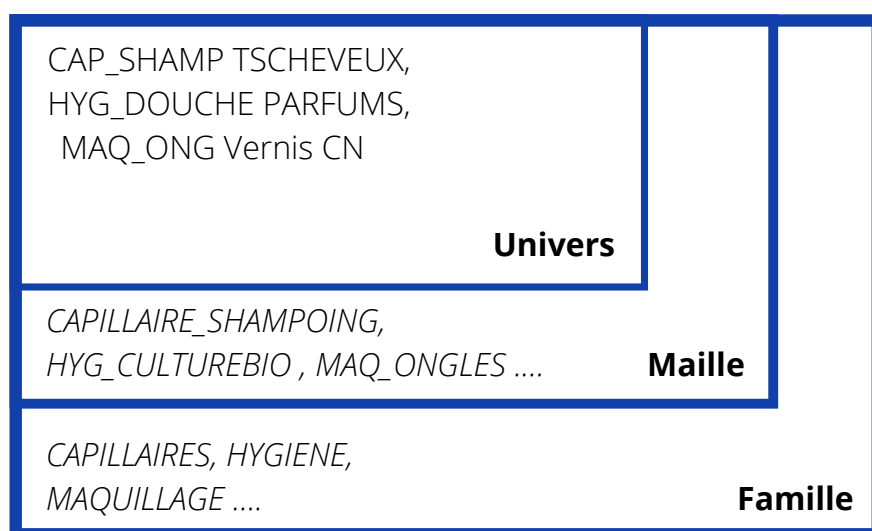
- *Regroupement via Pandas*

Nous effectuons plusieurs regroupements sur les produits et clients afin de connaître un peu plus l'aspect du Dataset, les différentes catégories de produits, les prix, les chiffres des ventes...



MOIS_VENTE	total client	total price
1	211,944.0	3,922,441.38
2	148,009.0	2,845,321.4
3	169,144.0	3,200,551.14
4	166,729.0	3,109,273.54
5	179,313.0	3,769,628.89
6	213,535.0	4,118,695.03
7	228,082.0	4,209,609.93
8	176,442.0	2,951,500.23
9	198,772.0	3,258,067.86
10	170,450.0	2,997,736.24
11	183,951.0	3,485,610.32
12	257,419.0	5,387,936.22
total	2,303,790.0	43,256,372.18

*Nombres de clients selon les mois de l'année*

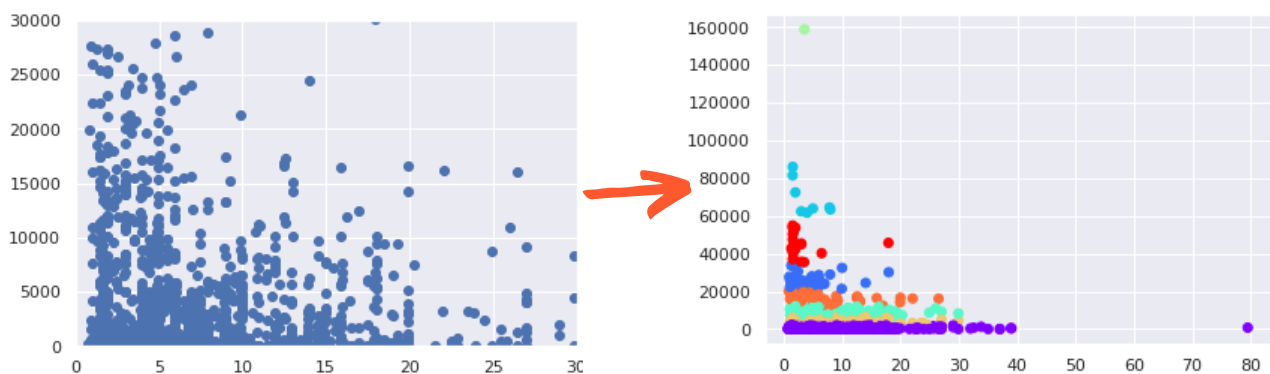


*Arborecences des catégories de produits*

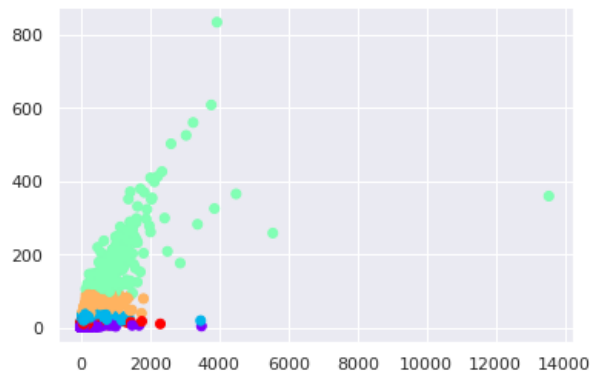
- *k-mean clustering*

Nous allons ensuite utiliser un algorithme de clusterisation des données pour répartir les produits et les clients dans différentes catégories.

Pour les produits nous mettons en corrélations le prix unitaire du produit et son nombre de vente total. L'algorithme détecte lui même le nombre de clusters optimale.



Pour les clients nous mettons en corrélations le nombres de produits achetés et le prix total du paniers afin de nous faire une idée du panier moyen.







# Conclusion

