Classifiez automatiquement des biens de consommation

Jeu de données :

https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/Parcours_data_scientist/Projet+-+Textimage+DAS+V2/Dataset+projet+prétraitement+textes+images.zip

Présentation par Hortense Monnard

Problématique :

L'entreprise **Place de marché** lance une marketplace e-commerce sur laquelle des vendeurs proposent des articles à des acheteurs en postant une photo et une description.

A l'heure actuelle, il y a peu d'articles disponibles et l'attribution d'une catégorie à chaque article est effectuée manuellement par les vendeurs et est donc peu fiable.

→ Besoin d'une classification automatique lorsque vendeurs entre une description pour un article.

- Faciliter la mise en ligne des articles pour les vendeurs ;
- Faciliter la recherche par les acheteurs .

Objectif :

Création d'un moteur de classification des articles en différentes catégories, avec un niveau de précision suffisant.

PLAN

1. Présentation du jeu de données

2. Analyse Exploratoire

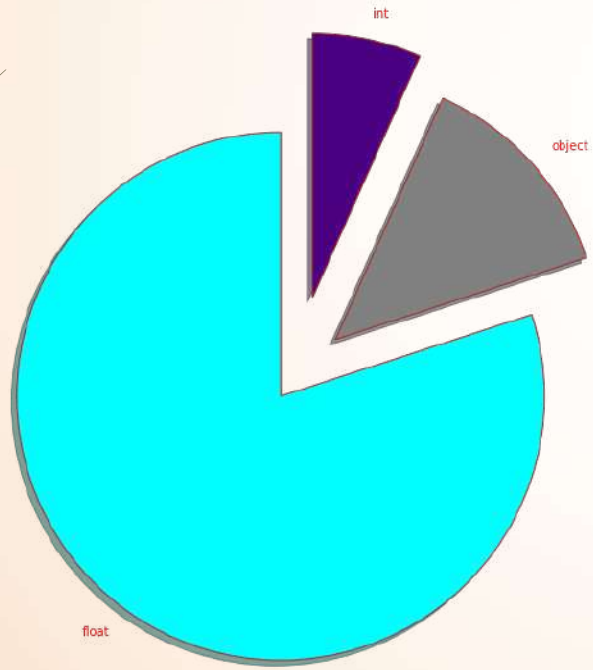
3. Modélisation
   - 3.a. Moteur de classification des descriptions
   - 3.b. Moteur de classification des images
   - 3.c. Moteur de classification des descriptions et des images

4. Conclusions et Perspectives

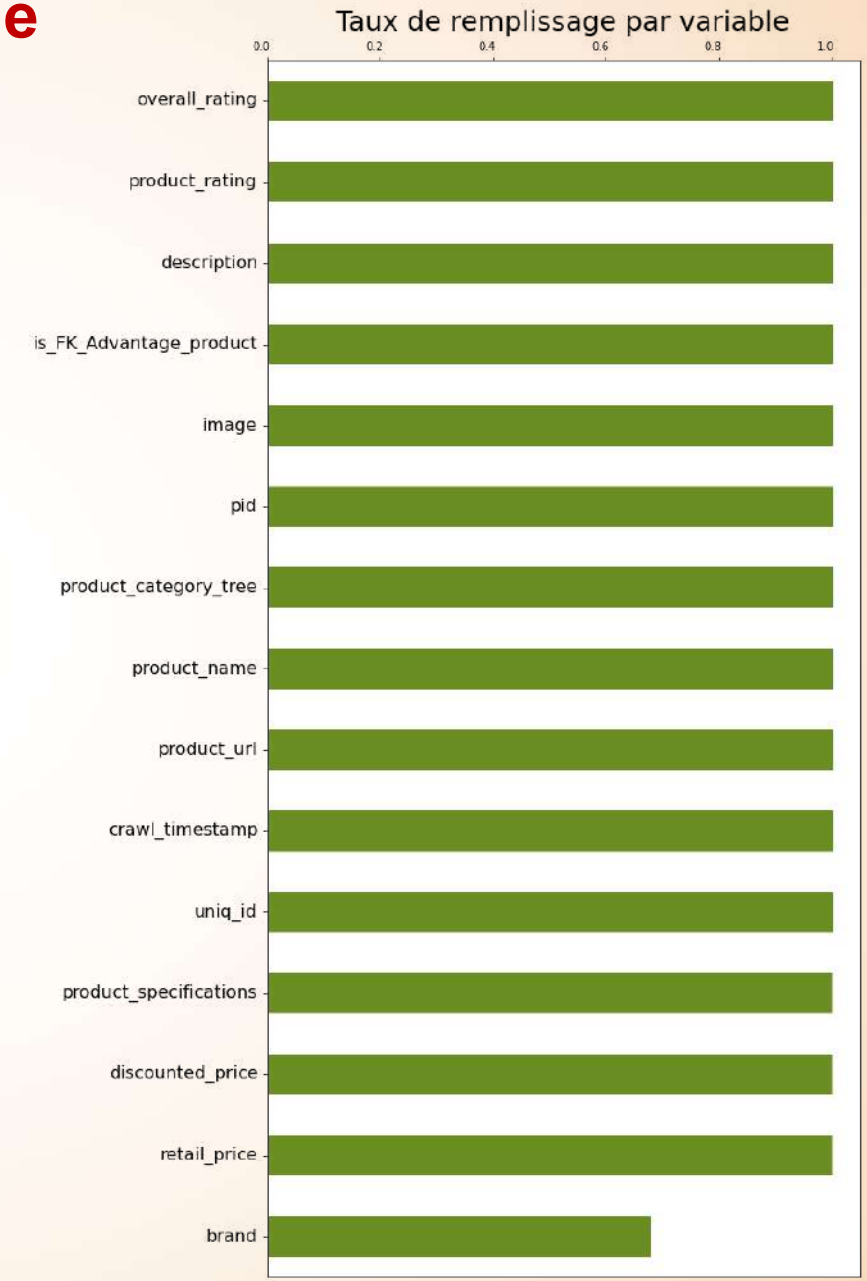
1. Présentation du jeu de données

Avant Nettoyage

Dimension du data set :
1050 lignes et 15 colonnes

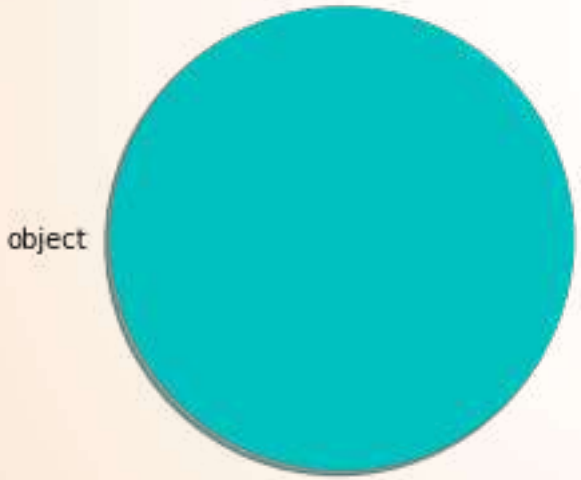


Types de variables

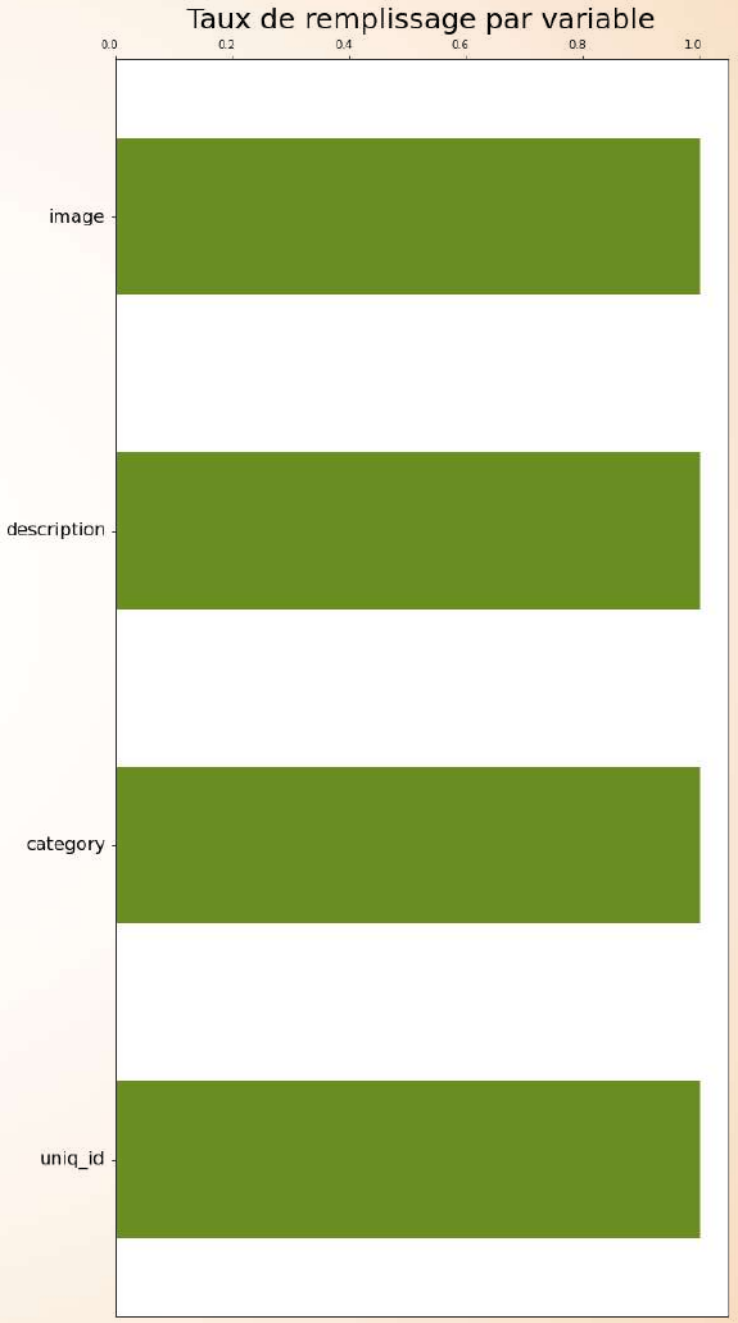


Après Nettoyage

Dimension du data set :
1050 lignes et 4 colonnes



Types de variables



2. Analyse Exploratoire

Analyses Exploratoires

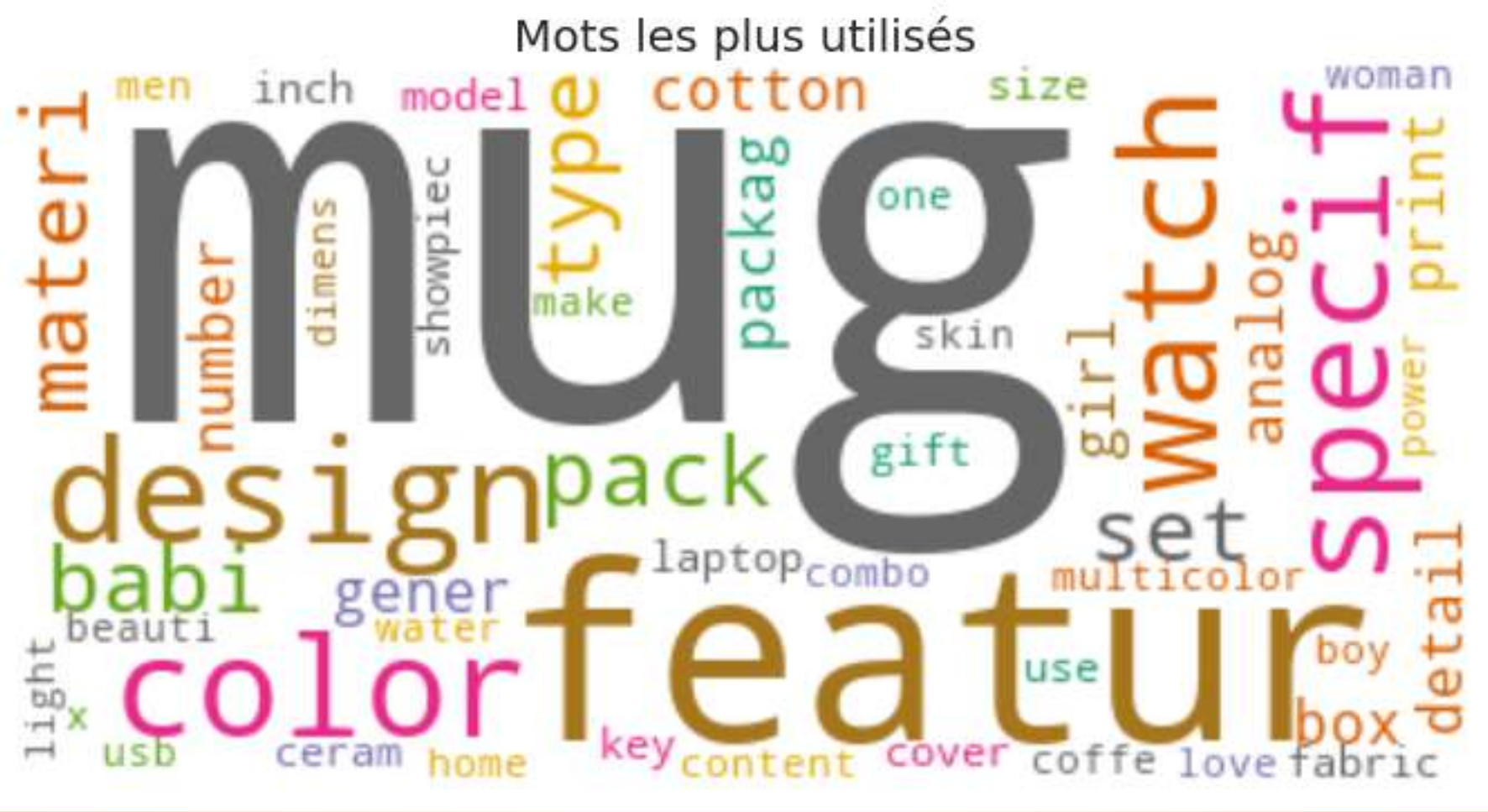
Les 7 catégories de produits proposés par les vendeurs pour les 1500 produits listés:

- Home Furnishing ;
- Baby Care ;
- Watches ;
- Home Decor & Festive Needs ;
- Kitchen & Dining ;
- Beauty and Personal Care ;
- Computers

Analyses Exploratoires

Analyses Exploratoires

Avec un filtrage supplémentaire des mots les plus courants non informatifs sur les produits



Analyses Exploratoires

Méthodes d'Exploration et d'Analyse utilisées

- « **Bag of Words** » (BOW) pour les mots ;

- « **Bag Of Visual Words** » (BOVW) pour les images.

Principe de ces méthodes :
Extraire des racines de mots (lexèmes) ou features d'images afin de permettre une classification.

Analyses Exploratoires - Texte

1. Nettoyage du texte :

- Transformer les contractions ;
- Supprimer les Stop-Words ;
- Supprimer les chiffres ;
- Extraire les racines des mots ;
- Lemmatiser.

Key Features of Prime Printed 4 Seater Table Cover Length 60 inch/152 cm Width 40 inch/101 cm,Prime Printed 4 Seater Table Cover (Multicolor, PVC) Price: Rs. 499 Prime Center Table Cover Printed 4 Seater,Specifications of Prime Printed 4 Seater Table Cover (Multicolor, PVC) In The Box Number of Contents in Sales Package Pack of 1 General Brand Prime Type Table Cover Model Name 0.281 Material PVC Model ID 281 Color Multicolor Dimensions Weight 250 g Length 60 inch / 152 cm Width 40 inch / 101 cm Seating Capacity 4 Seater

key featur prime print seater tabl cover length width prime print seater tabl cover multicolor pvc prime center tabl cover print seater specif prime print seater tabl cover multicolor pvc box number content sale packag pack gener brand prime type tabl cover model name materi pvc model id color multicolor dimens weight g length inch width inch seat capac seater

2. Création d'une matrice avec les fréquences d'apparition des lexèmes pour chaque produit (dimension : 1049, 254).

Analyses Exploratoires - Images

1. Prétraitement des images :

- Appliquer une échelle de gris ;
- Egaliser les histogrammes des vecteurs descripteurs.

2. Création d'une matrice avec les fréquences d'apparition des features extraites ou « Visual Words » pour chaque image

(dimension : 1049 lignes, 70 colonnes).

Analyses Exploratoires

« Bag of Words » (BOW) pour le texte

Matrice avec 254 lexèmes (racines de mots) : 1050, 254

	abstract	add	addit	addit featur	addit style	analog	analog watch	art	babi	babi boy	babi girl	bath	batteri	beauti	black	blanket
220	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.126033	0.159077	0.000000	0.0	0.0	0.000000	0.0	0.0
698	0.753046	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.0	0.0
870	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.0	0.0
967	0.000000	0.0	0.101376	0.121577	0.0	0.0	0.0	0.0	0.199051	0.125619	0.114846	0.0	0.0	0.109366	0.0	0.0
1027	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.0	0.0

« Bag of Visual Words » (BOVW) pour les images

Matrice avec 70 features (Visual Words) : 1049, 70

	1	2	3	4	5	6	7	8	9	10	11	12
0	0.000261	0.000763	0.000182	0.000207	0.000719	0.000338	0.000335	0.001097	0.000236	0.000640	0.000541	0.000283
1	0.000596	0.000872	0.000221	0.000266	0.000231	0.000676	0.000527	0.000981	0.000636	0.000615	0.000162	0.000142
2	0.000596	0.000599	0.000154	0.000207	0.000603	0.000620	0.000718	0.000115	0.000236	0.000418	0.000180	0.000425
3	0.001341	0.000136	0.000202	0.000207	0.000115	0.002141	0.000072	0.001501	0.000177	0.000959	0.000162	0.000261
4	0.000596	0.000844	0.000154	0.000355	0.000231	0.000282	0.000814	0.000462	0.000310	0.000664	0.000198	0.000163

3. Modélisation

3.a. Moteur de classification des descriptions

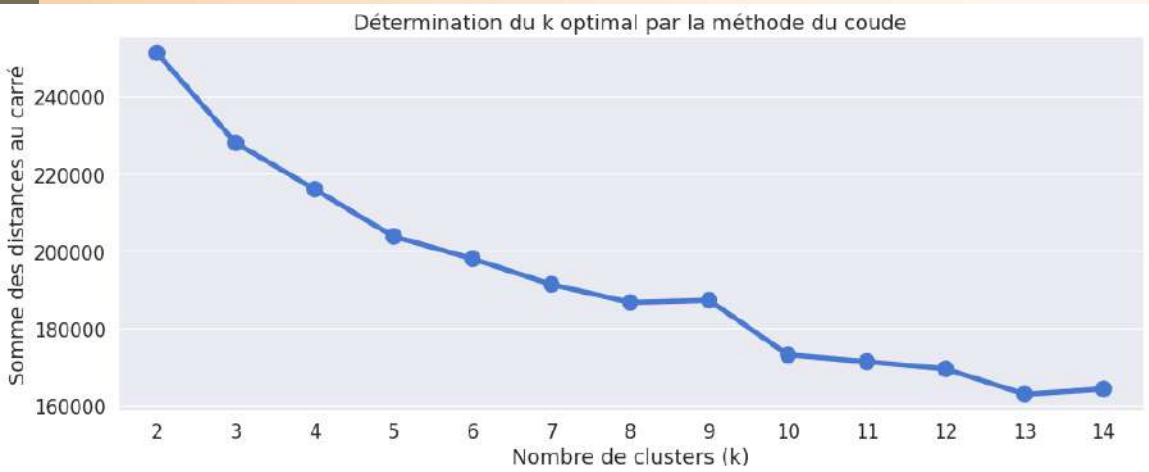
Méthode BOW

1. Nettoyage du texte :

- Transformer les contractions ;
- Supprimer les Stop-Words ;
- Supprimer les chiffres ;
- Extraire les racines des mots ;
- Lemmatiser.

2. Création d'une matrice avec les fréquences d'apparition des lexèmes pour chaque produit (dimension : 1049, 254).

Méthode BOW



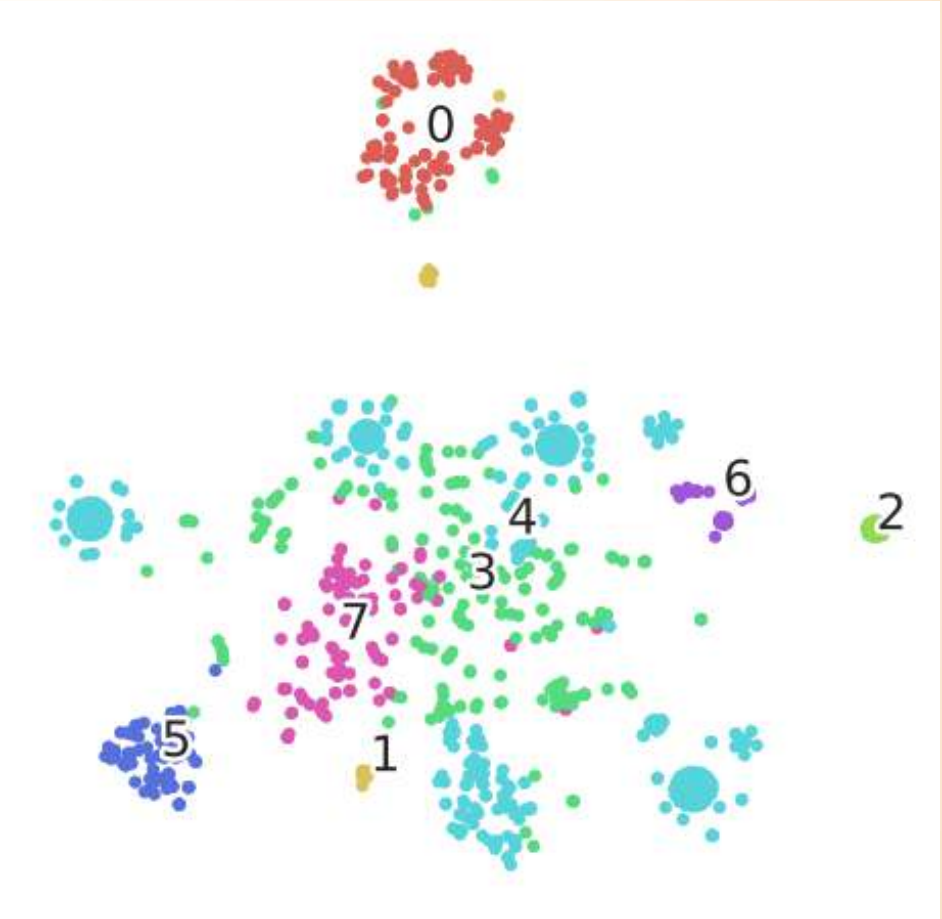

```
Pour n_clusters = 2 , le silhouette_score moyen est : 0.15609238717141216
Pour n_clusters = 3 , le silhouette_score moyen est : 0.09672760482313428
Pour n_clusters = 4 , le silhouette_score moyen est : 0.11551970278493738
Pour n_clusters = 5 , le silhouette_score moyen est : 0.08999862826276057
Pour n_clusters = 6 , le silhouette_score moyen est : 0.10933482528206241
Pour n_clusters = 7 , le silhouette_score moyen est : 0.12585273639674563
Pour n_clusters = 8 , le silhouette_score moyen est : 0.13903252710756708
Pour n_clusters = 9 , le silhouette_score moyen est : 0.09710509205785732
```

Méthode BOW

Effectifs des clusters

4 : 428
3 : 221
0 : 134
7 : 114
5 : 75
6 : 36
2 : 21
1 : 21

Visualisation des clusters



Les clusters manquent de définition.

3.b. Moteur de classification des images

Méthode ORB et BOVW

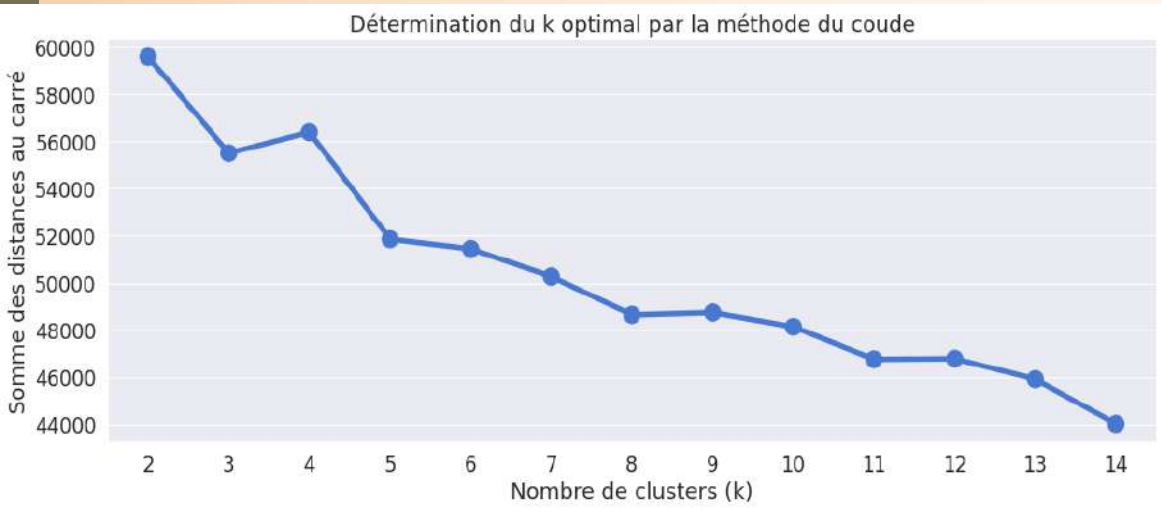
- **Extraction des vecteurs descripteurs** des points d'intérêt ;

- Création de **clusters avec les 70 features** les plus pertinentes ;

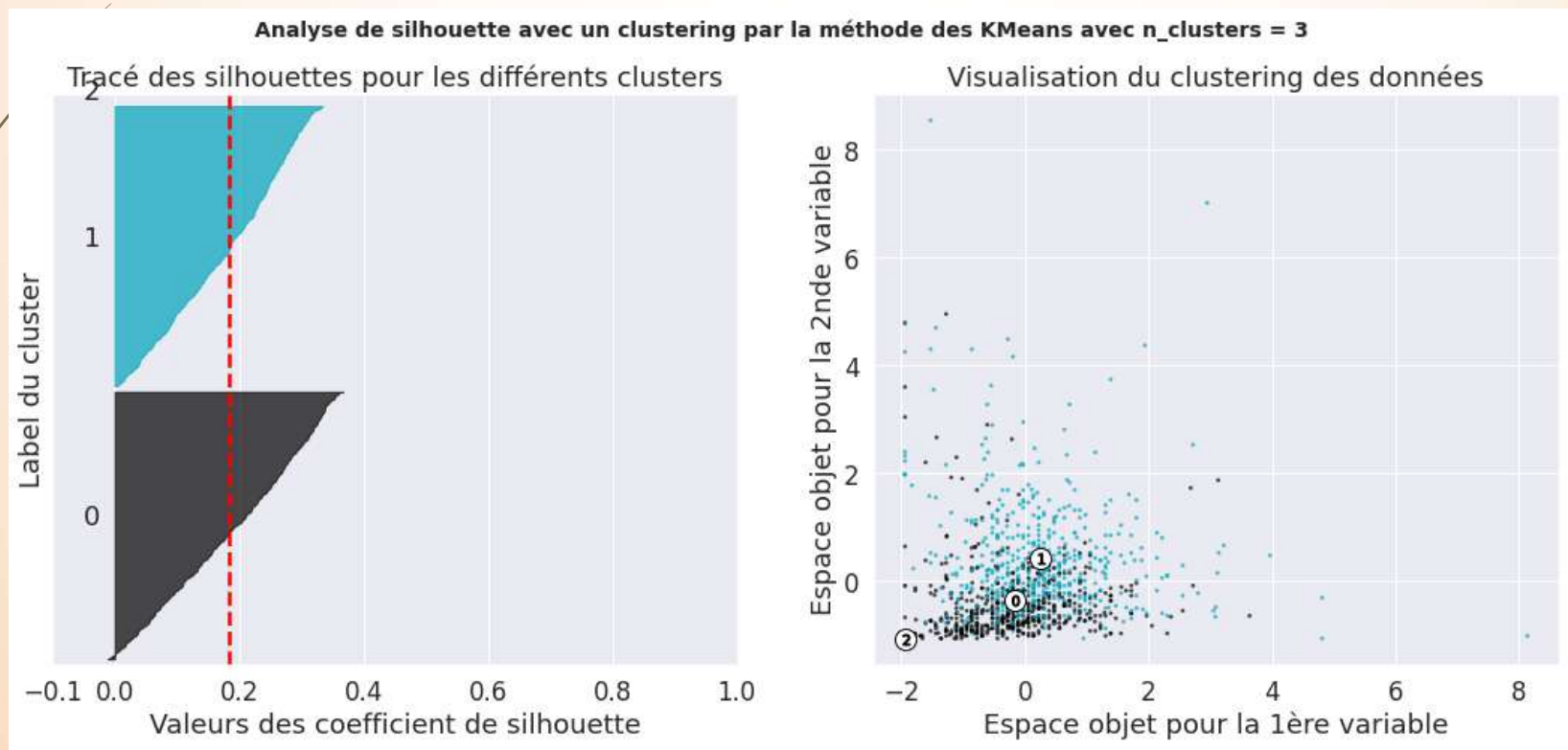
- Création d'une **matrice avec les fréquences d'apparition des features** extraites ou « Visual Words » pour chaque image (dimension : 1049, 70).

NB : Testé avec ou sans l'application d'un traitement préalable des images.

Méthode ORB et BOVW



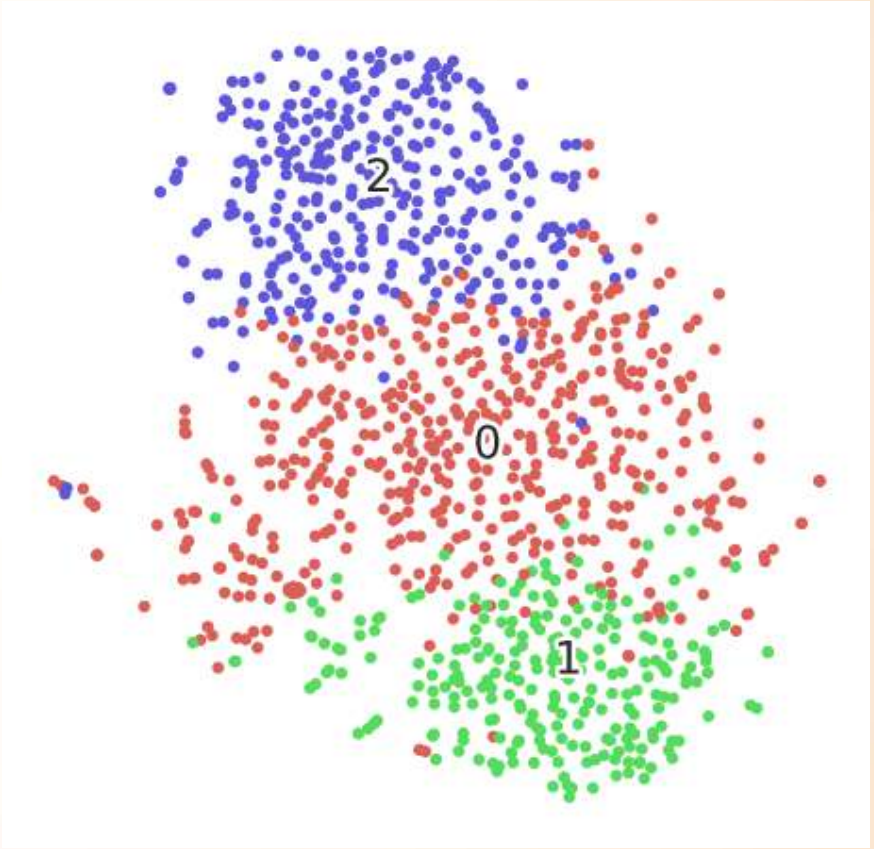

```
Pour n_clusters = 2 , le silhouette_score moyen est : 0.18287058328109226
Pour n_clusters = 3 , le silhouette_score moyen est : 0.1828568840147146
Pour n_clusters = 4 , le silhouette_score moyen est : 0.11545338410707923
Pour n_clusters = 5 , le silhouette_score moyen est : 0.10413636337892691
Pour n_clusters = 6 , le silhouette_score moyen est : 0.09008654460716525
Pour n_clusters = 7 , le silhouette_score moyen est : 0.0646126351663406
Pour n_clusters = 8 , le silhouette_score moyen est : 0.06927523424771669
Pour n_clusters = 9 , le silhouette_score moyen est : 0.06167963611423534
```



Méthode ORB et BOVW

Effectifs des clusters
0 : 479
2 : 324
1 : 246

Visualisation des clusters



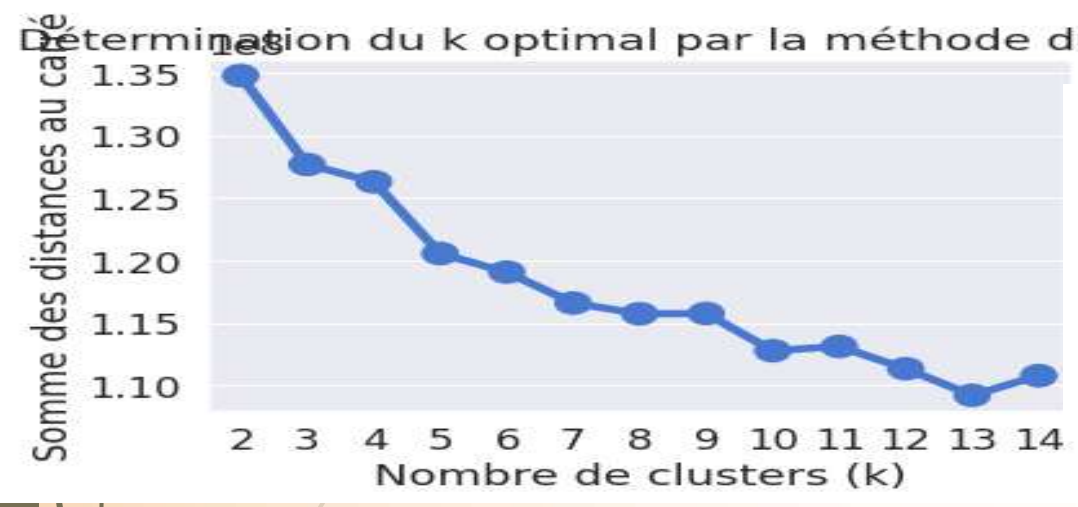
Les clusters semblent bien définis mais il n'y a que 3 clusters.
L'application du prétraitement choisi donne des clusters moins définis.

Méthode CNN

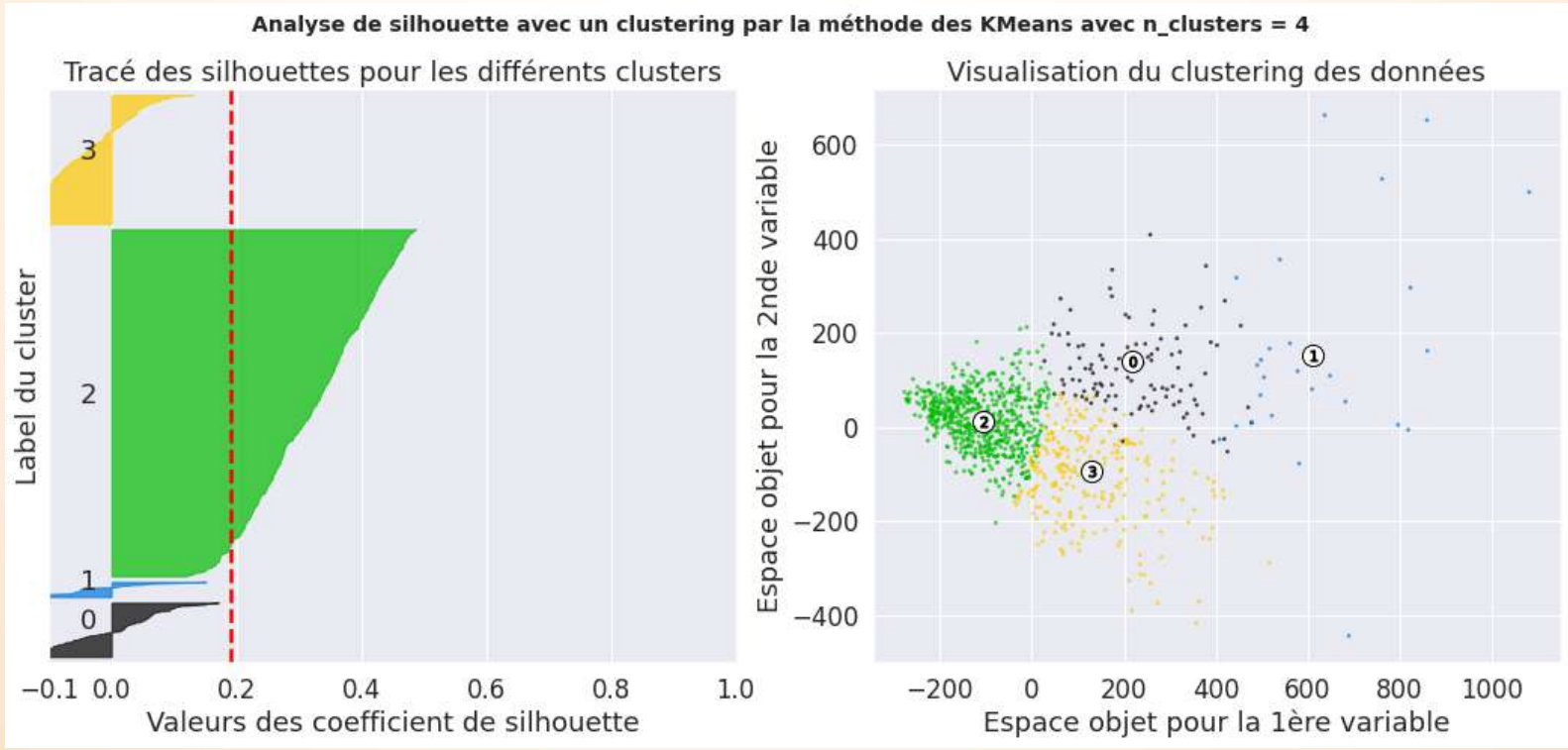
Utilisation du **réseau convolutif de neurones (CNN)** : ResNet 50

- Création d'une matrice avec 150528 **features reconnues** (dimension : 1050, 150528) ;
- **Transformation au log** des données ;
- **Réduction par PCA** à 1049 features.

Méthode CNN



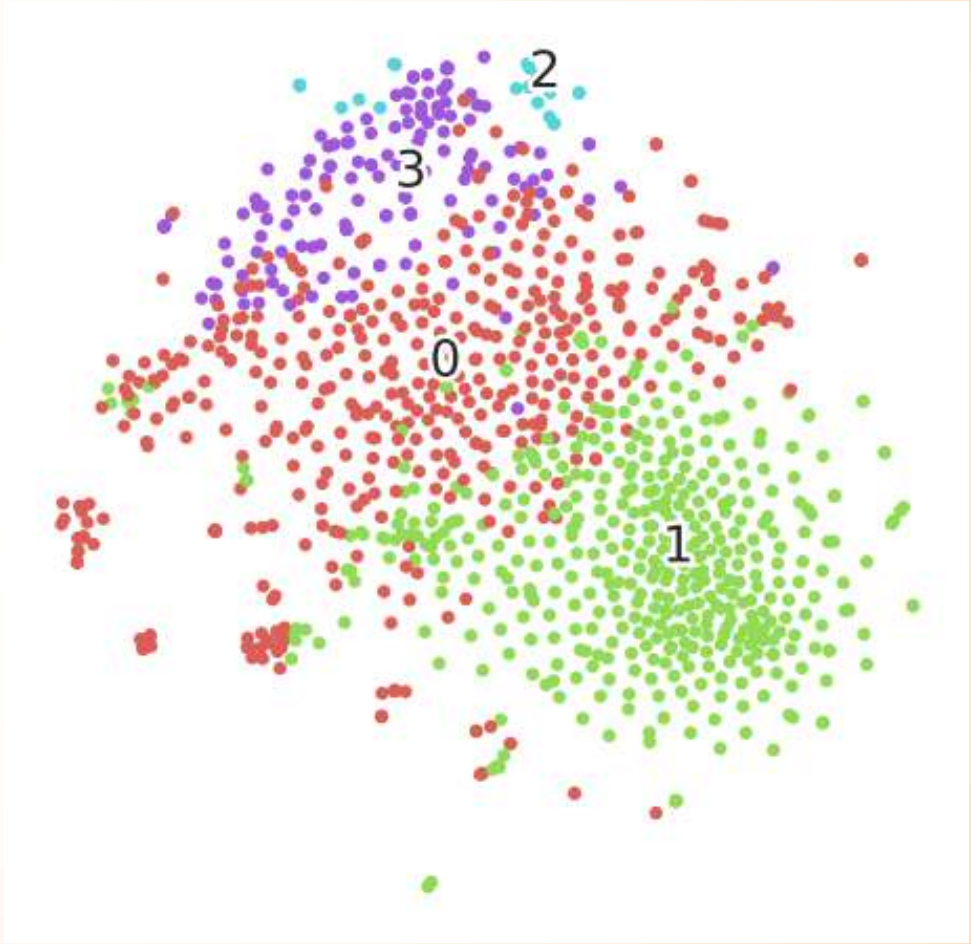

```
Pour n_clusters = 2 , le silhouette_score moyen est : 0.2618638877142364
Pour n_clusters = 3 , le silhouette_score moyen est : 0.15176464017470373
Pour n_clusters = 4 , le silhouette_score moyen est : 0.19103441219725162
Pour n_clusters = 5 , le silhouette_score moyen est : 0.1288427734638383
Pour n_clusters = 6 , le silhouette_score moyen est : 0.13578828641579463
Pour n_clusters = 7 , le silhouette_score moyen est : 0.09769350340081369
Pour n_clusters = 8 , le silhouette_score moyen est : 0.099046949314981
Pour n_clusters = 9 , le silhouette_score moyen est : 0.09523693117625523
```



Méthode CNN

Effectifs des clusters
1 : 447
0 : 443
3 : 137
2 : 23

Visualisation des clusters



Il y a 4 clusters de détecter et toutes les images prises sont acceptés par le modèle.

3.c. Moteur de classification des descriptions et des images

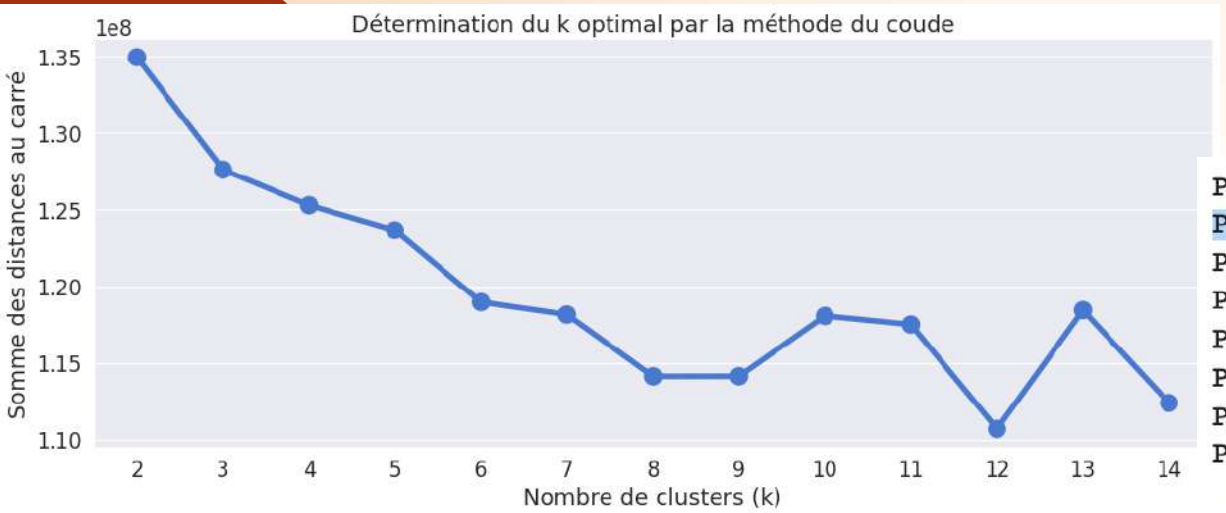
Moteur de classification des descriptions et des images

- Création d'une matrice avec 152533 **features reconnues** (dimension : 1050, 152533) ;

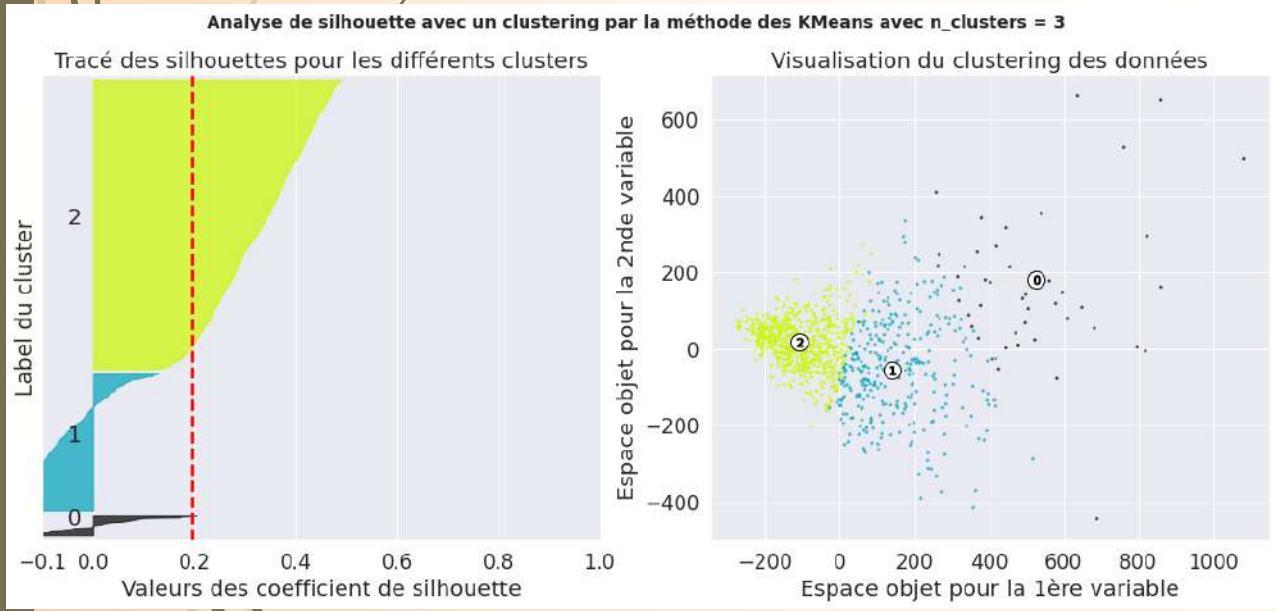
- **Transformation au log** des données visuelles issues du réseau convolutif de neurones (CNN) ResNet 50 ;

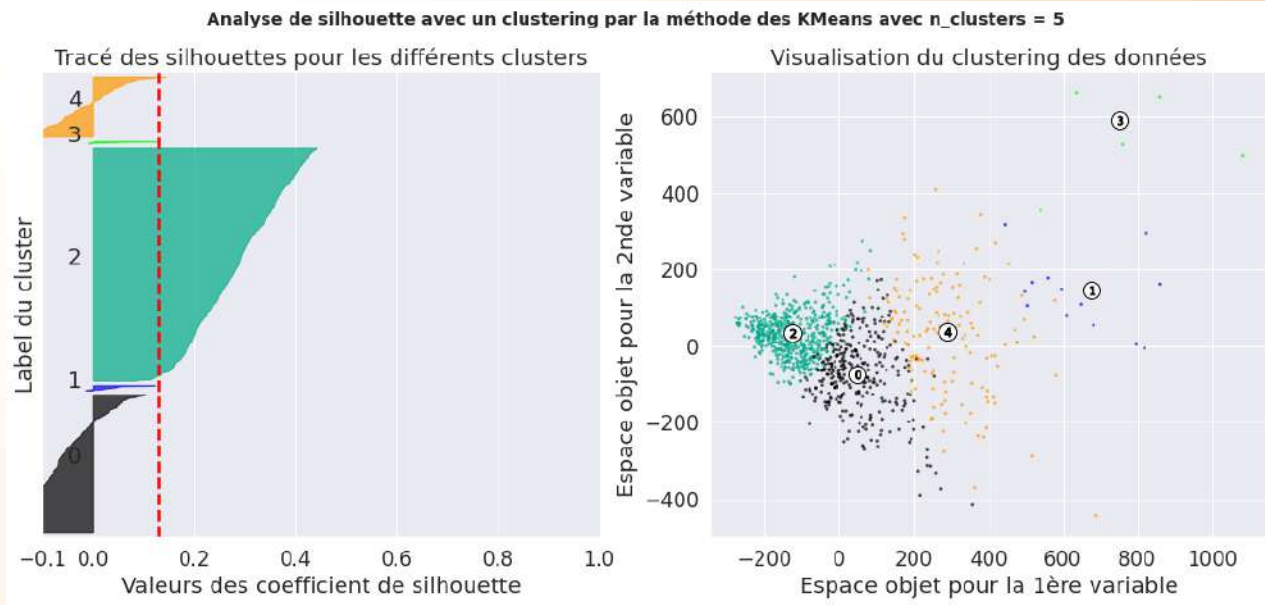
- **Réduction par PCA** à 1050 features.

Moteur de classification des descriptions et des images




```
Pour n_clusters = 2 , le silhouette_score moyen est : 0.2621275948042942
Pour n_clusters = 3 , le silhouette_score moyen est : 0.19486657456164347
Pour n_clusters = 4 , le silhouette_score moyen est : 0.08830422621716384
Pour n_clusters = 5 , le silhouette_score moyen est : 0.12880249510396727
Pour n_clusters = 6 , le silhouette_score moyen est : 0.07624491126278421
Pour n_clusters = 7 , le silhouette_score moyen est : 0.09340844367605422
Pour n_clusters = 8 , le silhouette_score moyen est : 0.07603714206206348
Pour n_clusters = 9 , le silhouette_score moyen est : -0.05367381443366244
```



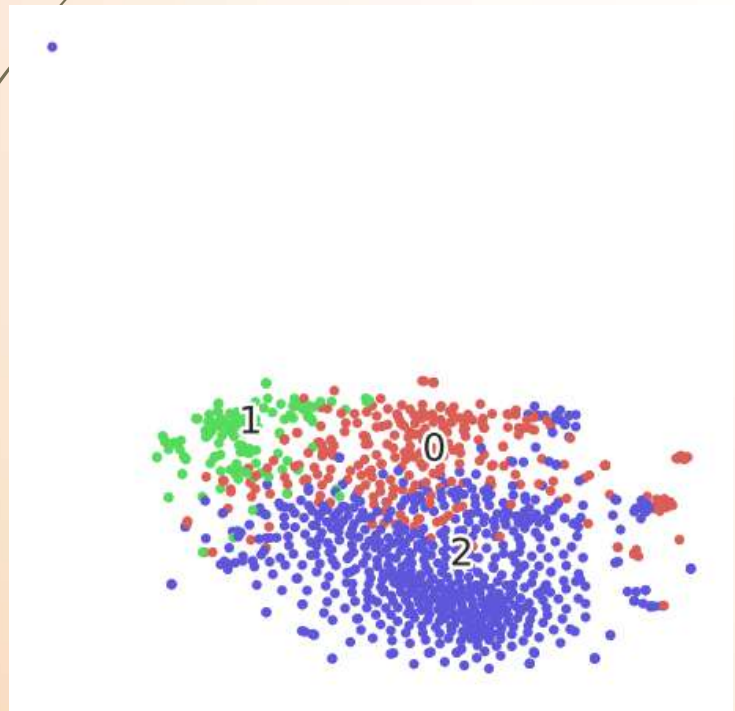


Moteur de classification des descriptions et des images

Effectifs des clusters

2 : 517
1 : 407
0 : 126

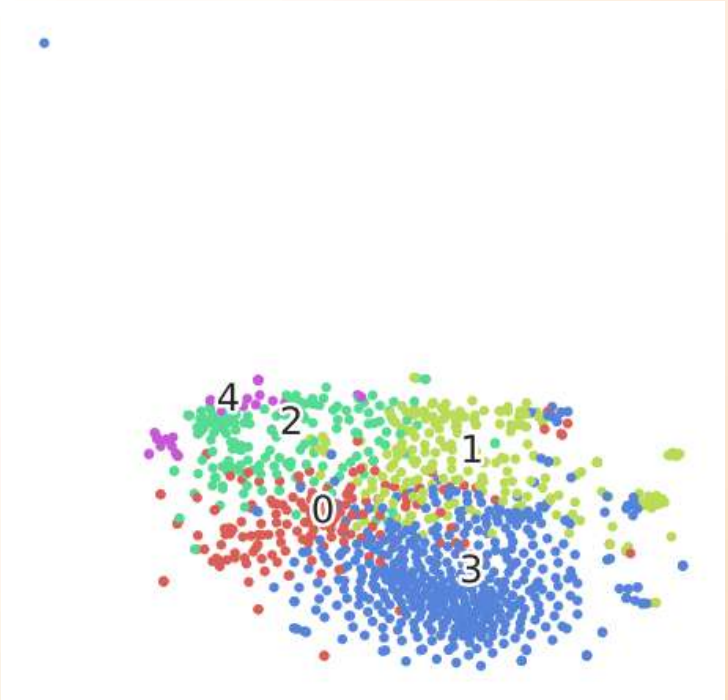
Visualisation des clusters



Effectifs des clusters

3 : 475
1 : 217
2 : 163
0 : 163
4 : 32

Visualisation des clusters



4. Conclusions et Perspectives

Conclusions :

- Les catégories désignées par les vendeurs ne semblent pas être les plus évidentes et certaines pourraient être rassemblées (« Home Furnishing », « Home Decor & Festive Needs » et « Kitchen & Dining ») ;
- Comparée à l'utilisation de ORB et Bag Of Visual Words, la méthode du réseau de neurones convolutifs (CNN) permet d'analyser toutes les images.

Perspectives :

Améliorer le prétraitement des images pour obtenir des modèles plus performants.

MERCI !