Classifiez automatiquement des biens de consommation

Meltem YILDIRIM

14/09/2021

Sommaire

I. Problématique

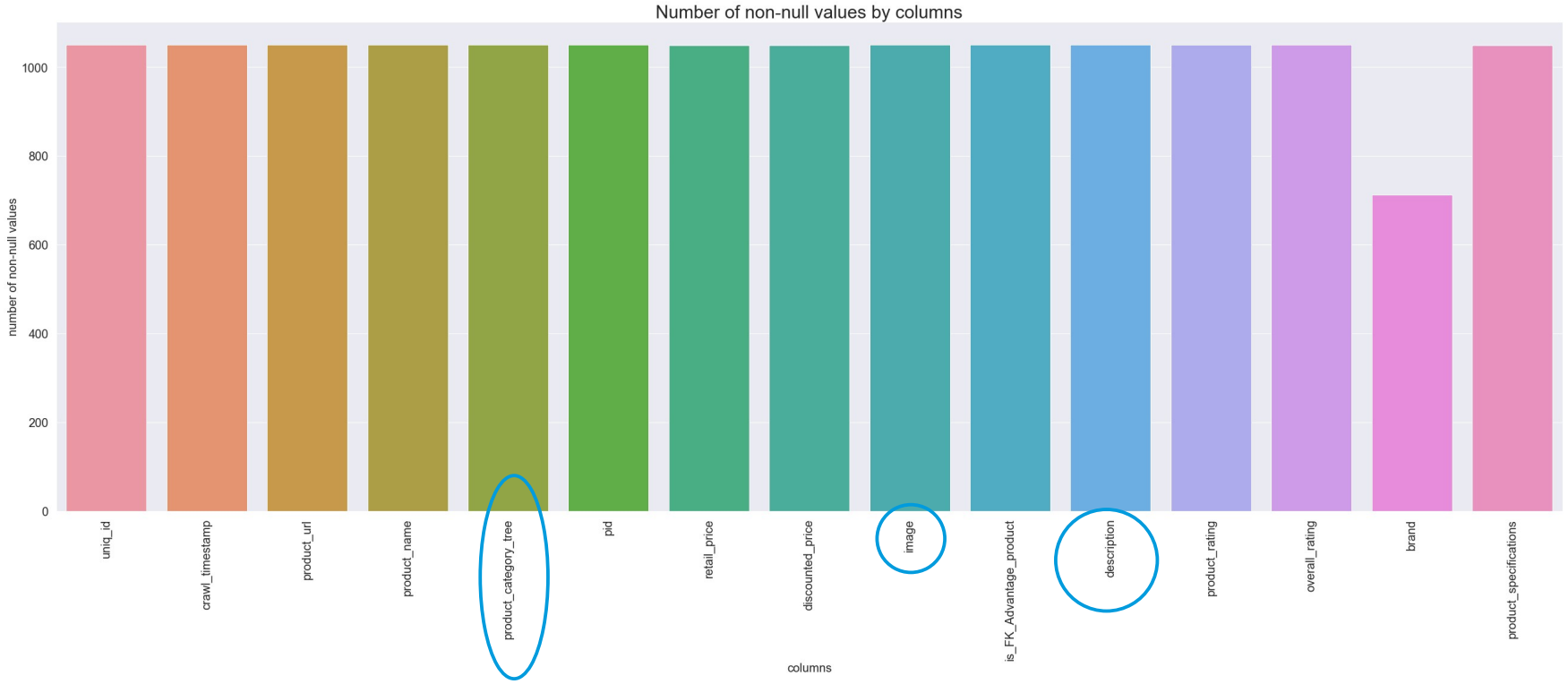
- **Missions :**
  - Prétraitement du jeu de données
  - Réduction de dimension
  - Clustering
  - Réaliser une première étude de faisabilité d’un moteur de classification

II. Interprétation de la problématique

- **Données textes et images :**
  - Réaliser des prétraitements de texte et d’image
  - Classifier ces données
  - Évaluer les clusters obtenus

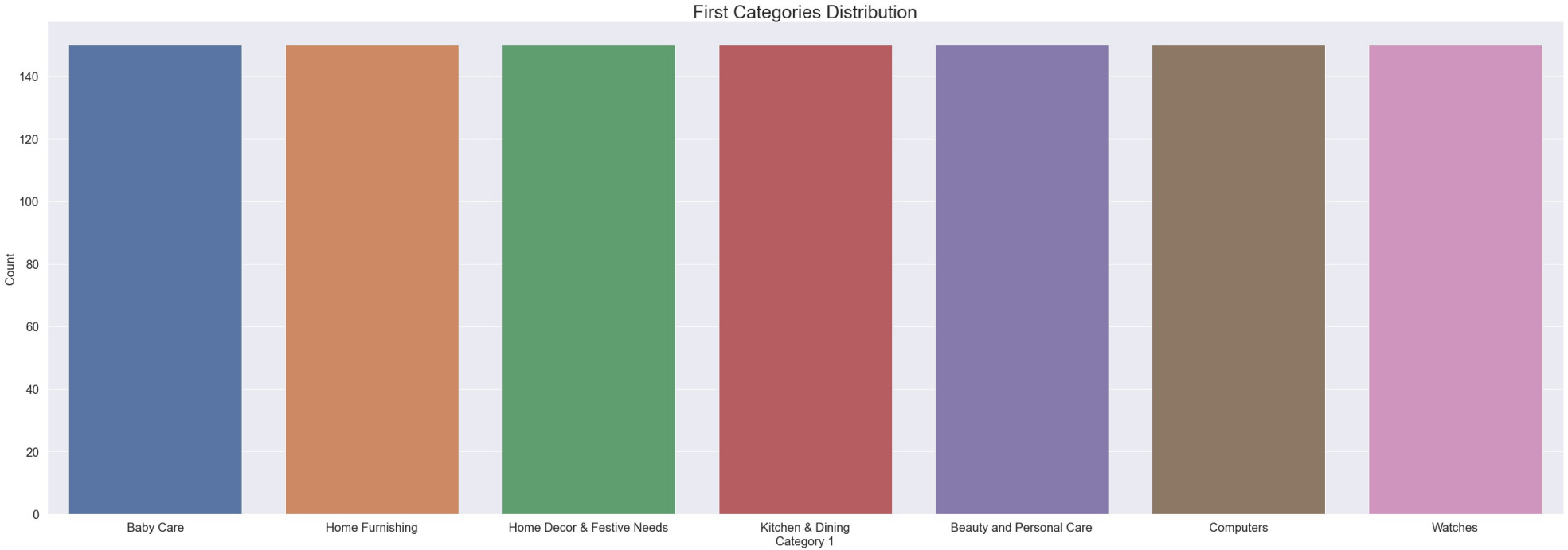
III. Présentation du jeu de données

1050 lignes (articles) et 15 colonnes



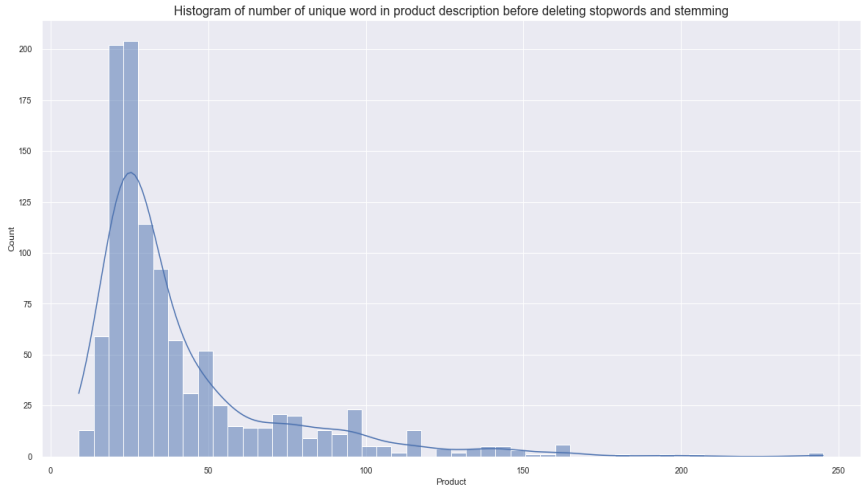
IV. Prétraitements

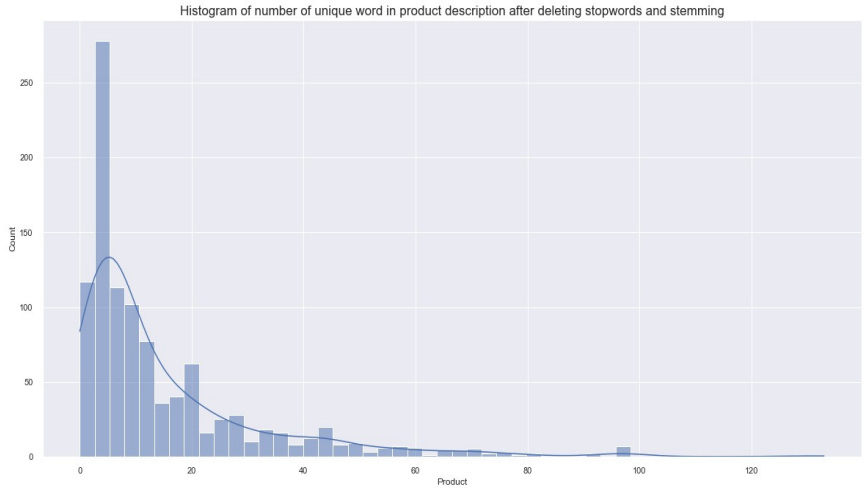
Extraction des catégories principales de la colonne **product_category_tree**



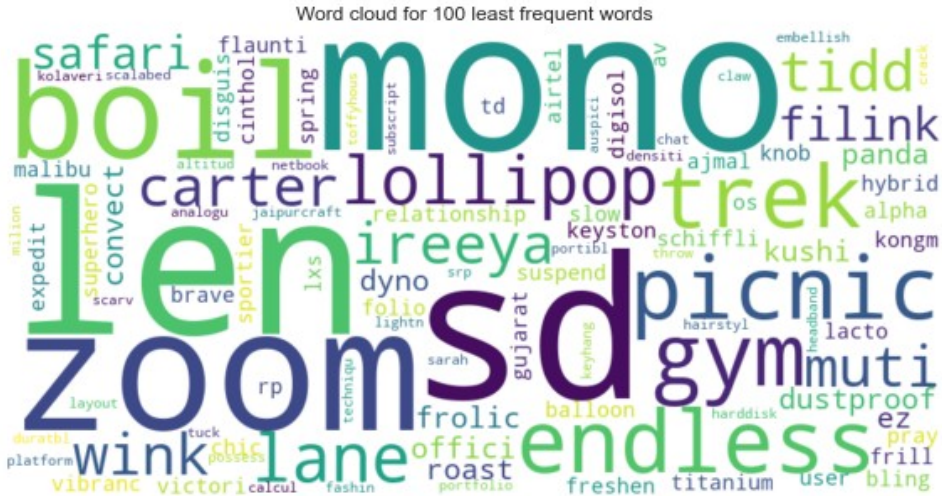
IV. Prétraitements

- **Description :**
  - Minuscule, tokenisation, stopwords, stemming
  - Avant : 5016 mots uniques
  - Après : 2776 mots uniques



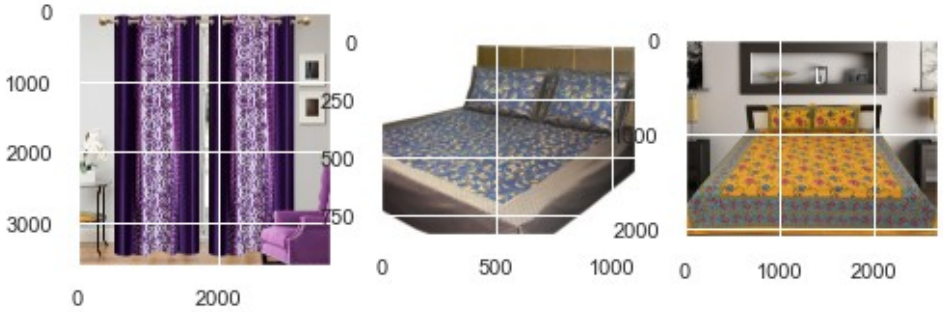


IV. Prétraitements

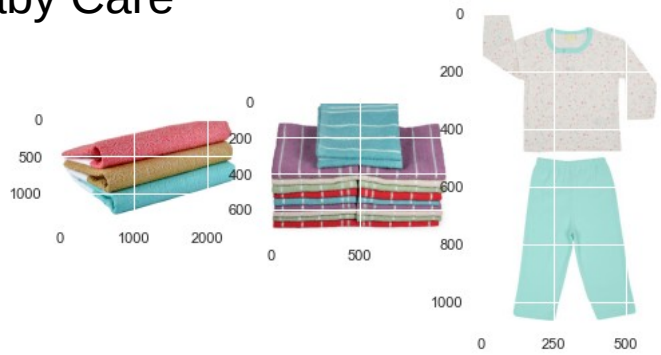


IV. Prétraitements

Home Furnishing



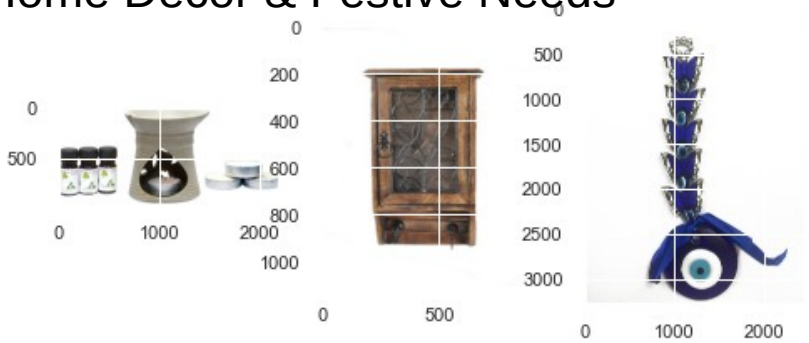
Baby Care



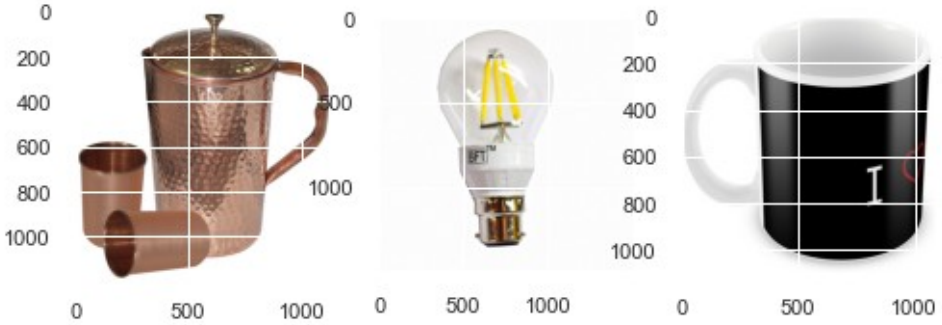
Watches

IV. Prétraitements

Home Decor & Festive Needs

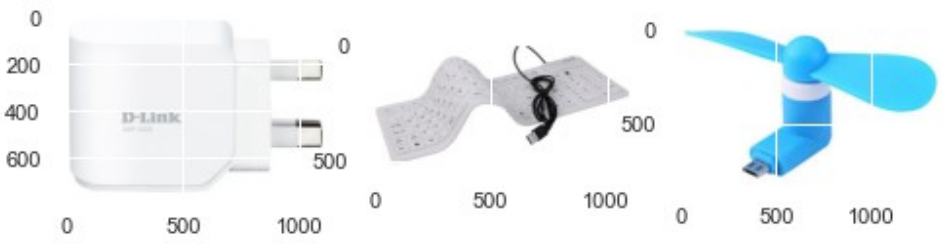


Kitchen & Dining



Beaury and Personal Care

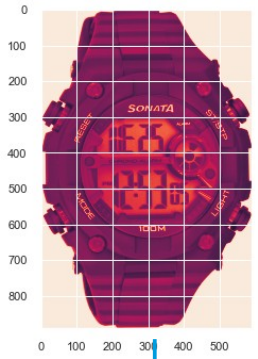
Computers

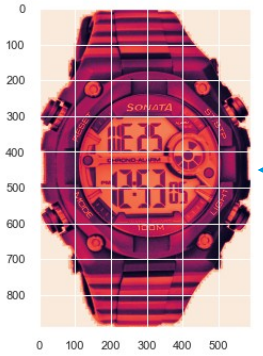


IV. Prétraitements

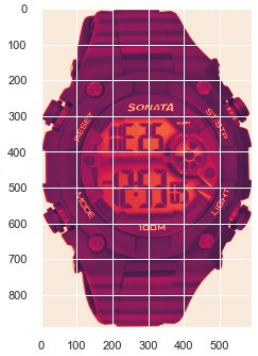
Image Originale

Passage en gris





Égalisation d’histogrammes



Filtrage de bruit

IV. Prétraitements

Identification des descripteurs avec la méthode ORB

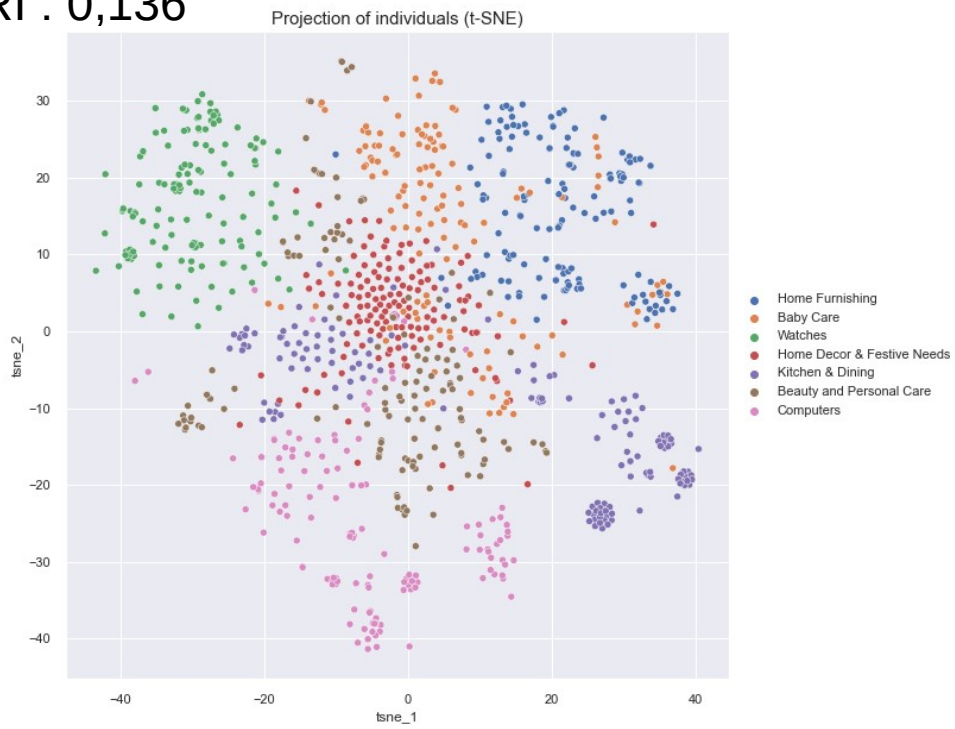
Les images contient 520145 descripteurs et chaque descripteur est un vecteur de longueur 32.

V. Clustering

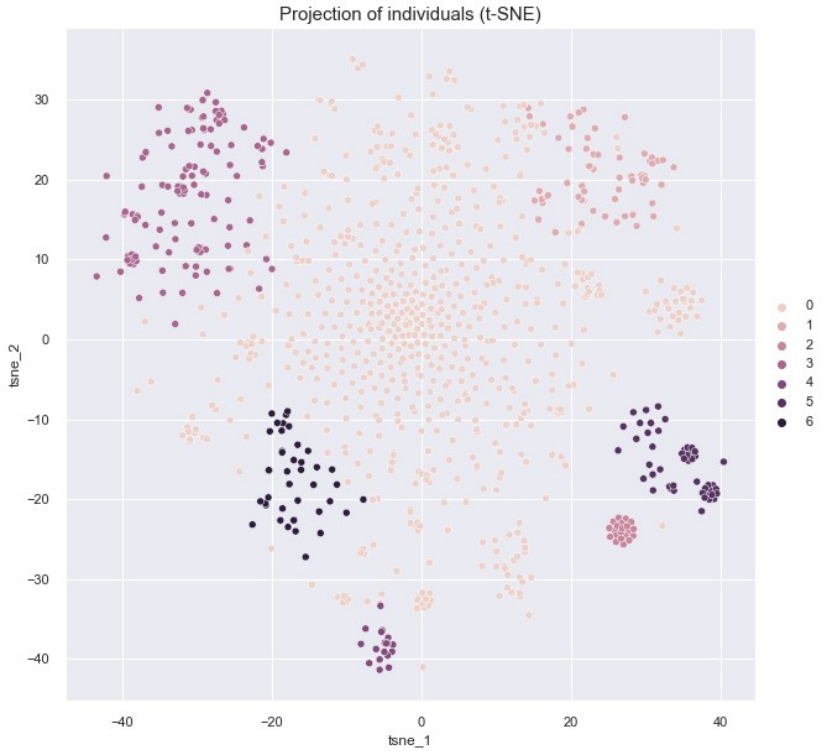
Méthode	ARI Score
TFIDF + Kmeans (Elbow pour le nombre de cluster)	0,122
TFIDF + Kmeans (nombre de cluster égale 7)	0,136
Word2Vec (sans pré-entraînement)	0,045
Word2Vec (avec pré-entraînement)	0,154
Word embeddings pré-entraîné (avec prétraitement de texte)	0,845
Word embeddings pré-entraîné (sans prétraitement de texte)	0,874
TFIDF + SGDClassifier (supervisée)	0,952
ORB + Kmeans (nombre de cluster égale 7)	-0,001
Xception + transfer learning + fine-tuning (sans prétraitement d'image)	0,854
Xception + transfer learning + fine-tuning (avec prétraitement d'image)	0,807

V. Clustering

TFIDF (Term Frequency – Inverse Document Frequency) + Kmeans (nombre de cluster égale 7)
ARI : 0,136



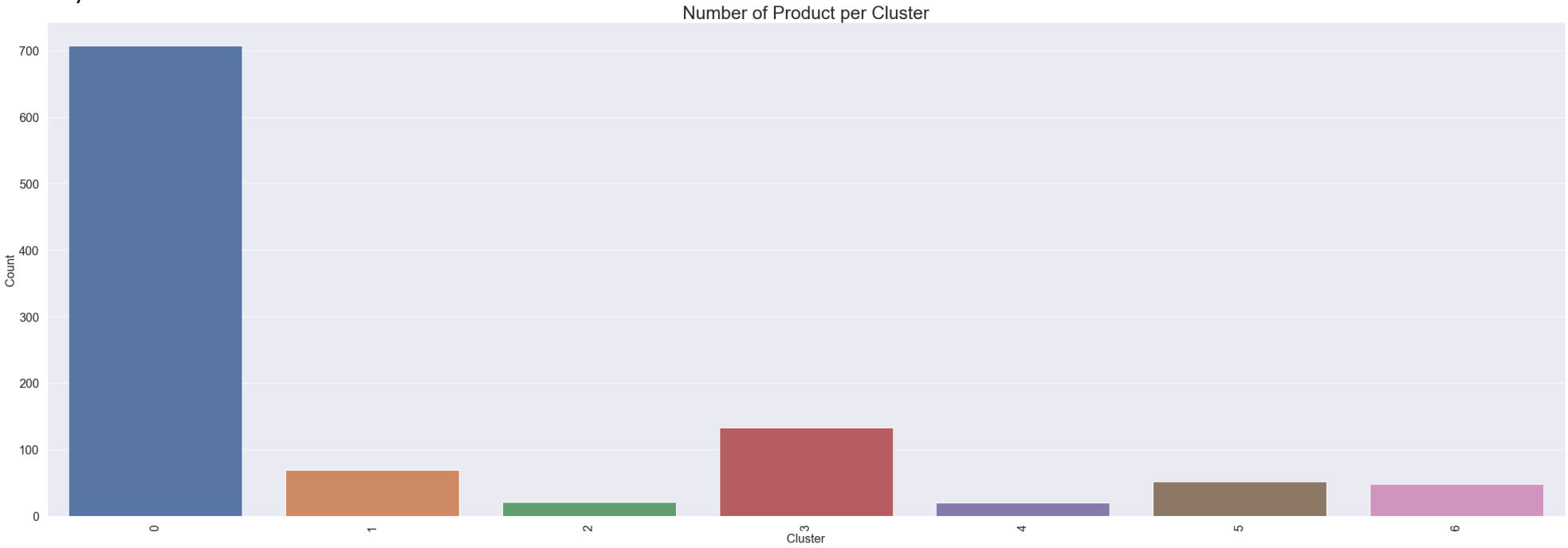
True categories



Predicted categories

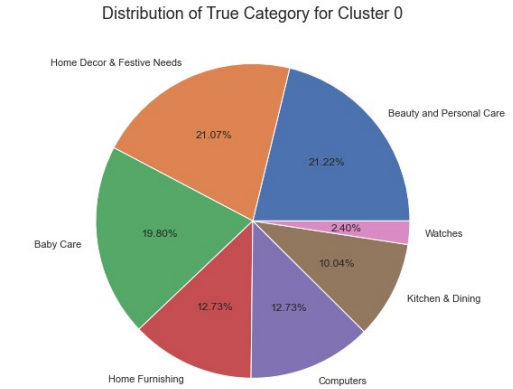
V. Clustering

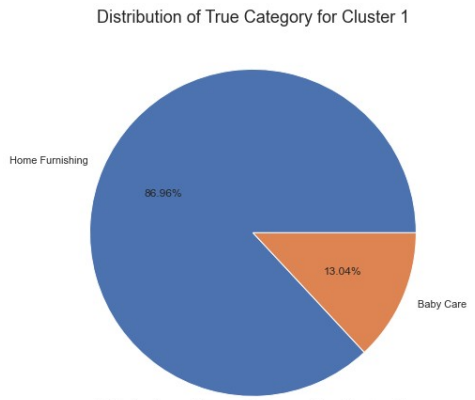
TFIDF (Term Frequency – Inverse Document Frequency) + Kmeans (nombre de cluster égale 7)
ARI : 0,136

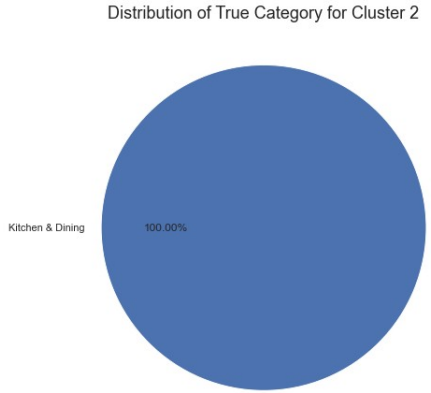


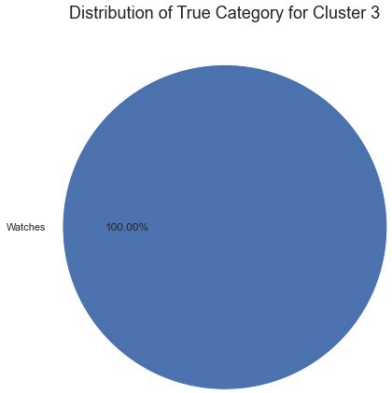
V. Clustering

TFIDF (Term Frequency – Inverse Document Frequency) + Kmeans (nombre de cluster égale 7)
ARI : 0,136





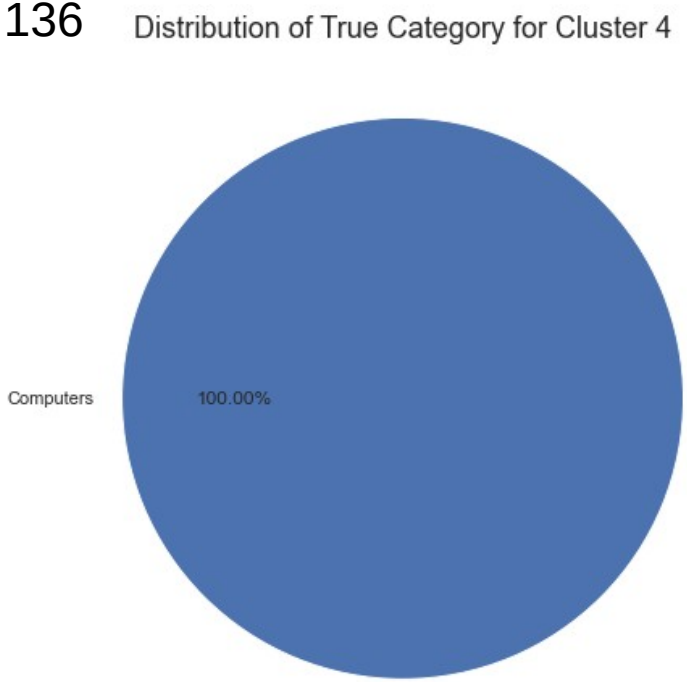


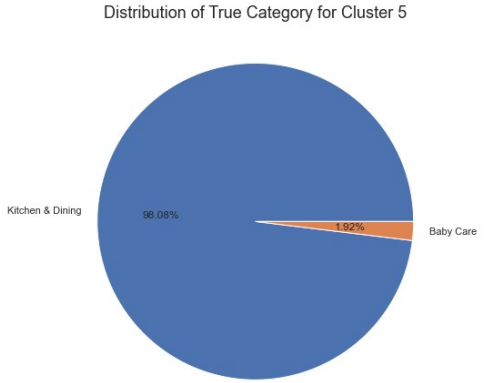


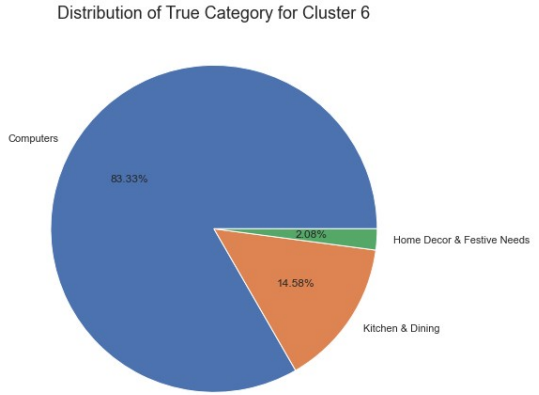
V. Clustering

TFIDF (Term Frequency – Inverse Document Frequency) + Kmeans (nombre de cluster égale 7)

ARI : 0,136



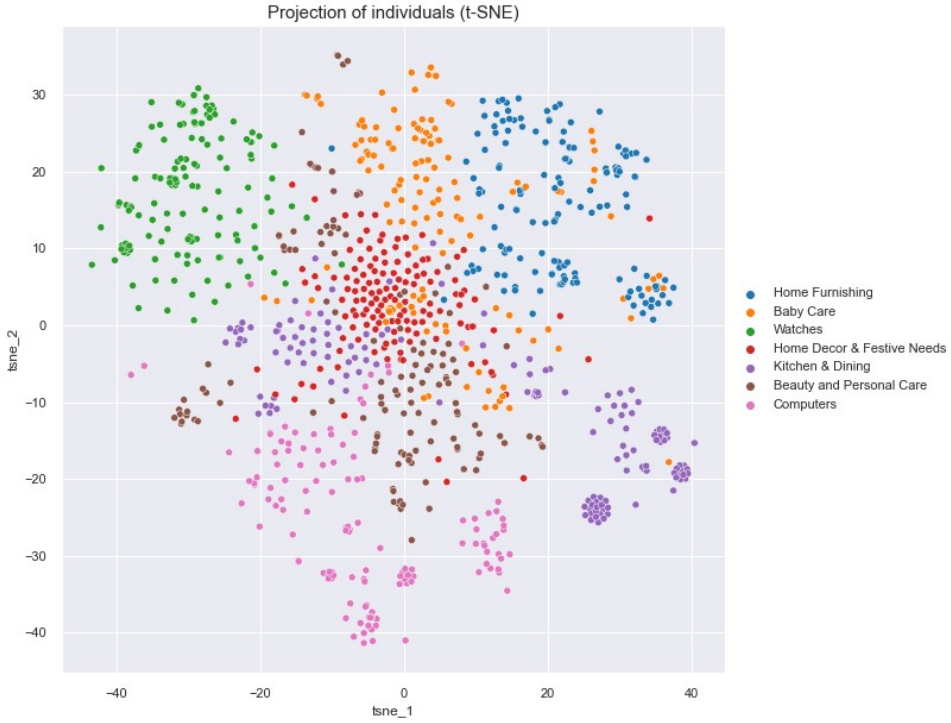




V. Clustering

TFIDF + SGDClassifier (supervisée)
ARI : 0,952

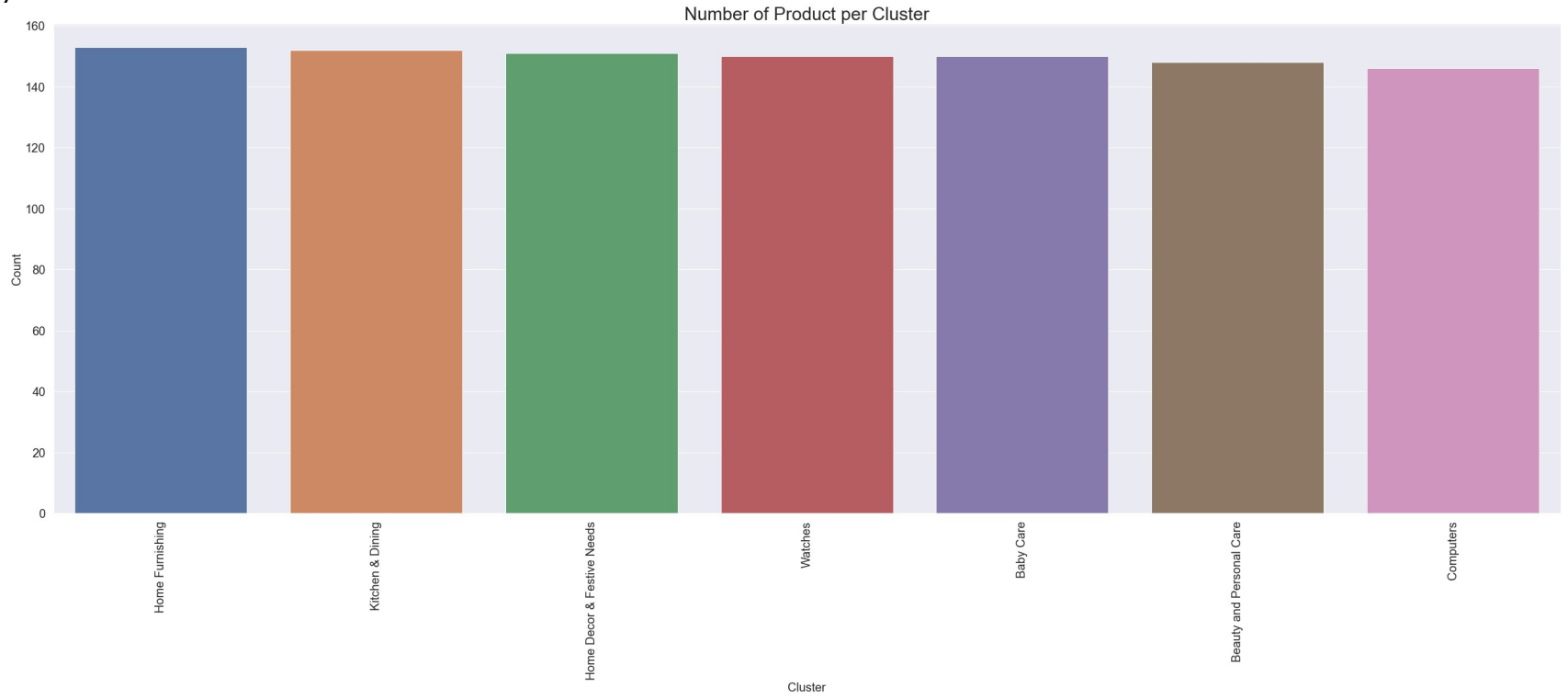
True categories



Predicted categories

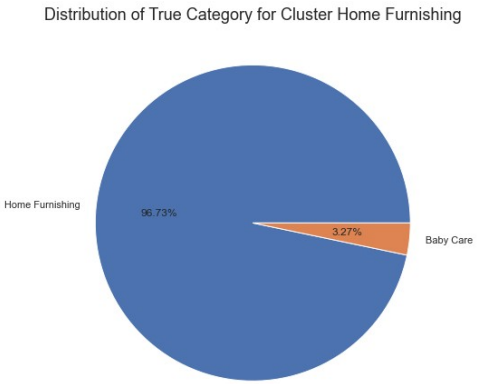
V. Clustering

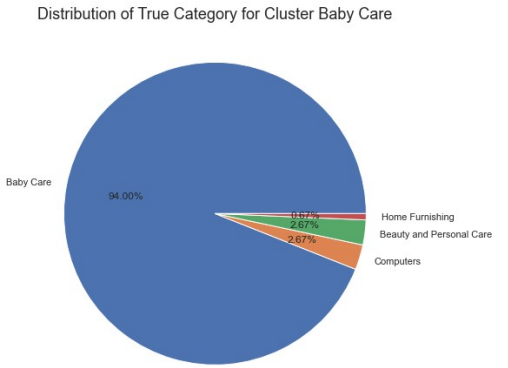
TFIDF + SGDClassifier (supervisée)
ARI : 0,952

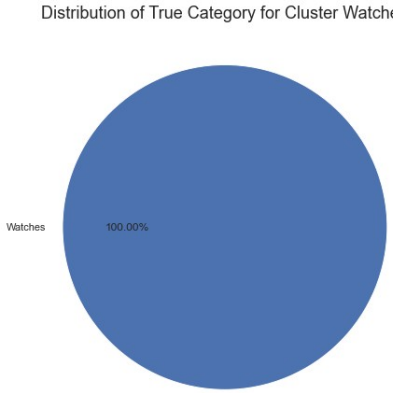


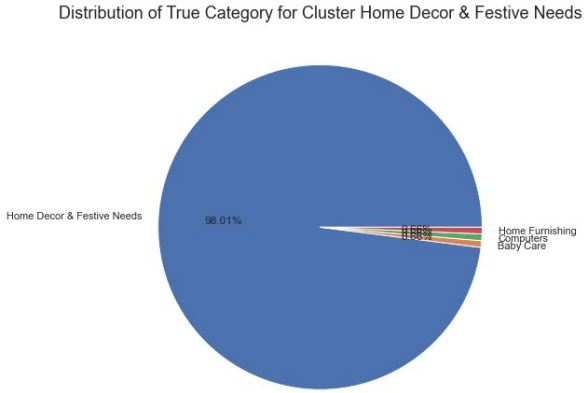
V. Clustering

TFIDF + SGDClassifier (supervisée)
ARI : 0,952



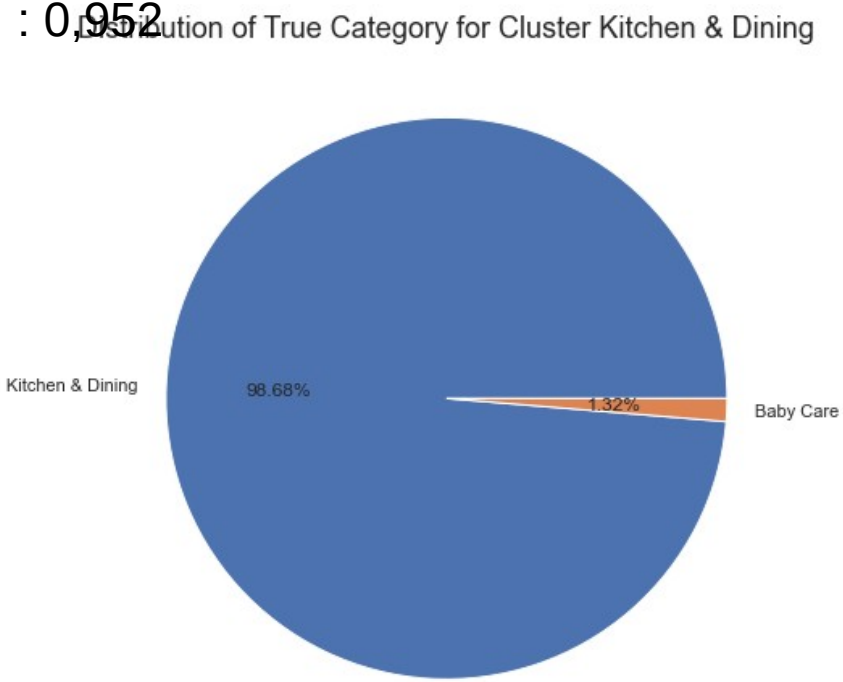


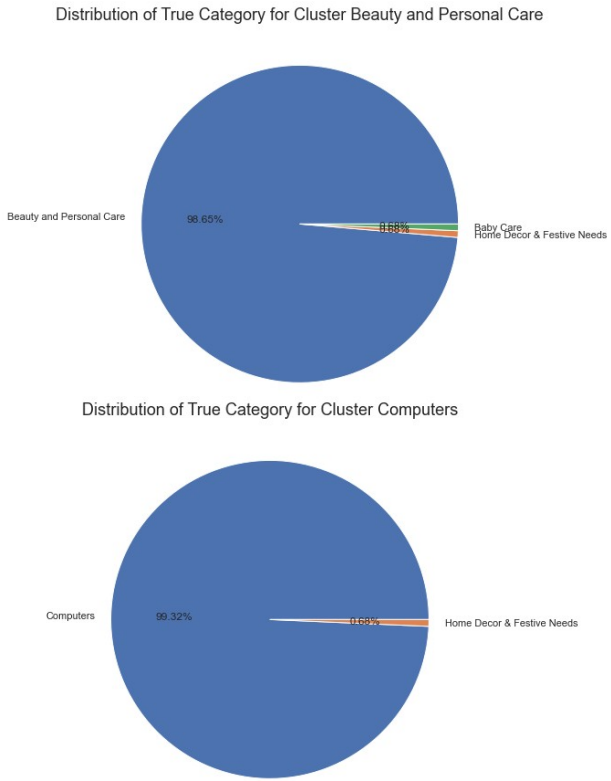




V. Clustering

TFIDF + SGDClassifier (supervisée)
ARI : 0,952

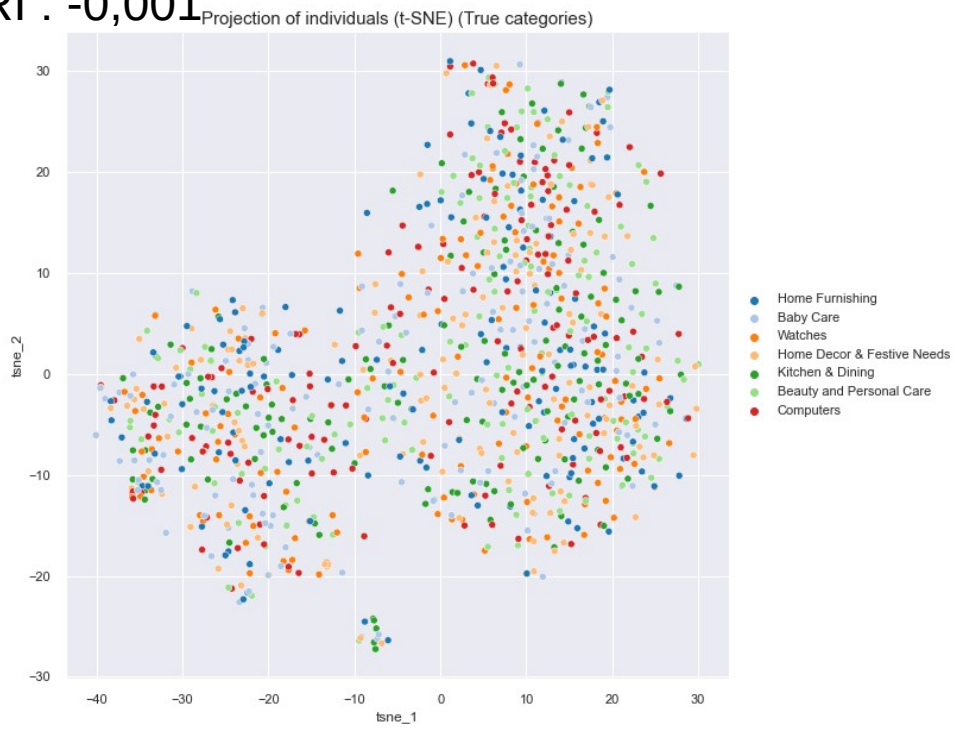




V. Clustering

ORB + Kmeans (nombre de cluster égale 7)

ARI : -0,001

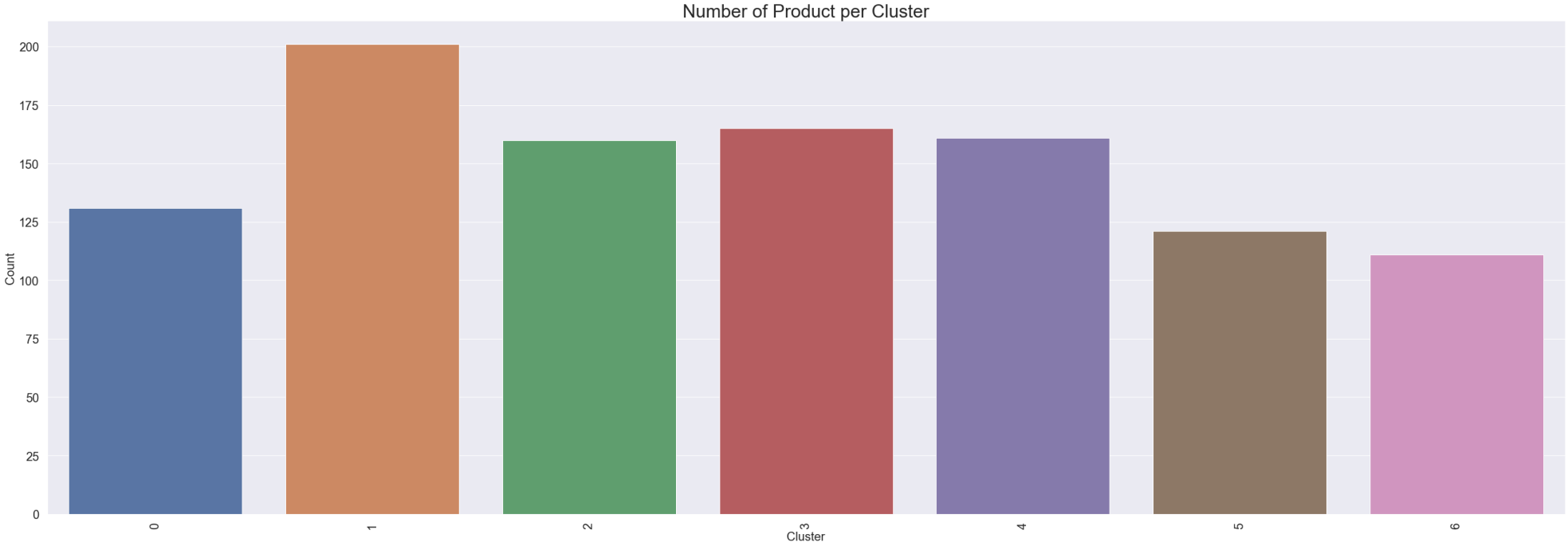


True categories

Predicted categories

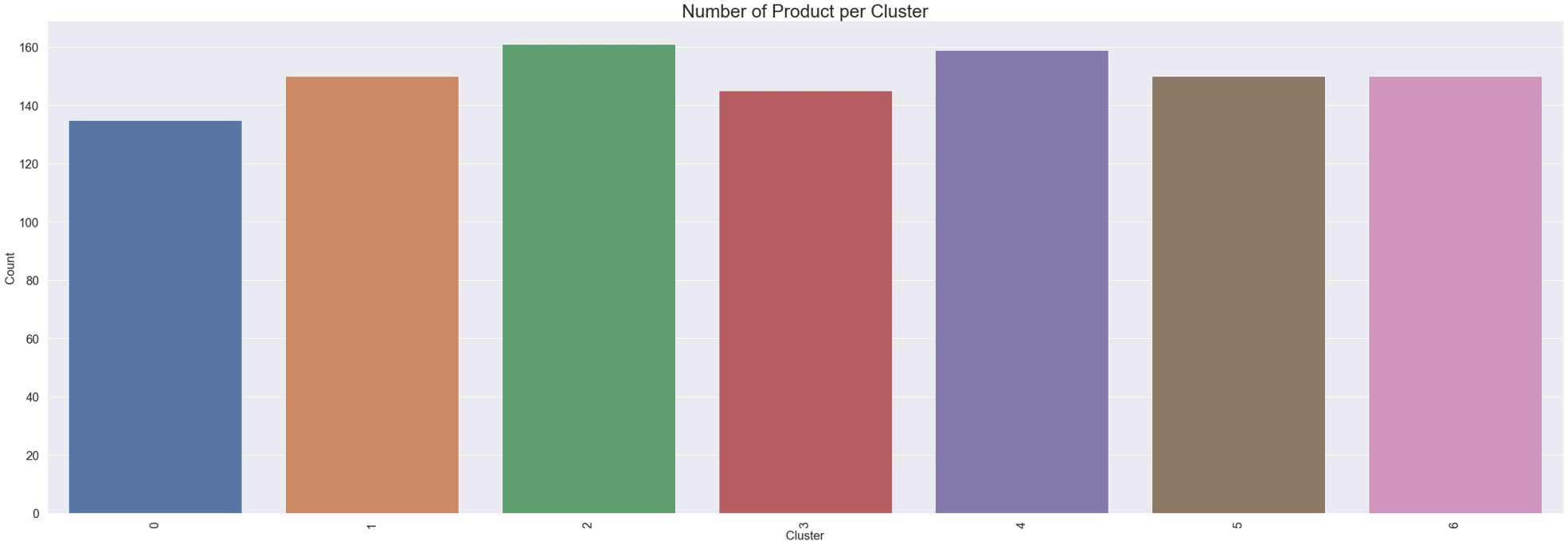
V. Clustering

ORB + Kmeans (nombre de cluster égale 7)
ARI : -0,001



V. Clustering

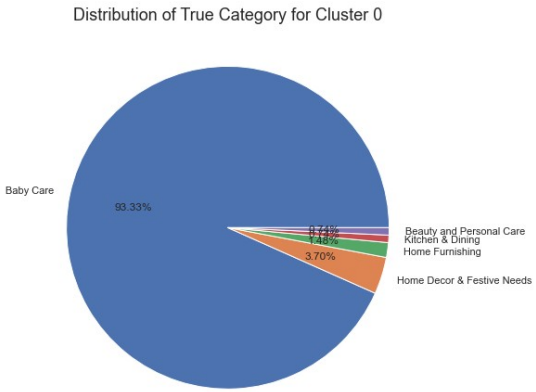
Xception + transfer learning + fine-tuning (sans prétraitement d'image)
ARI : 0,854

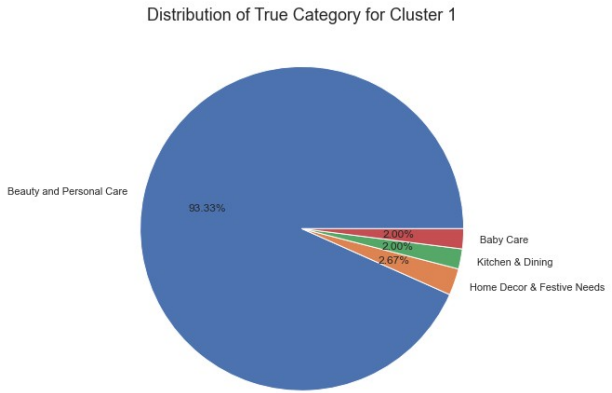


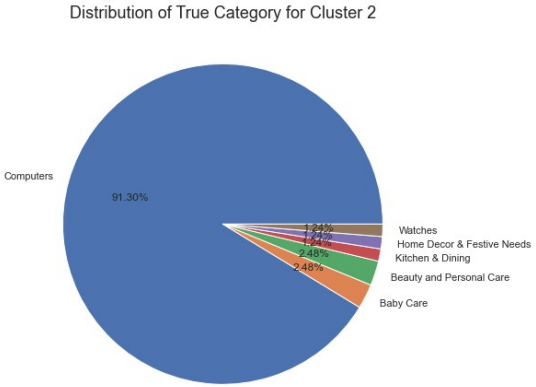
V. Clustering

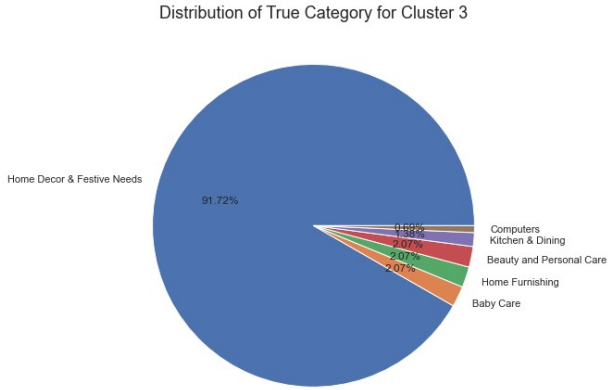
Xception + transfer learning + fine-tuning (sans prétraitement d'image)

ARI : 0,854



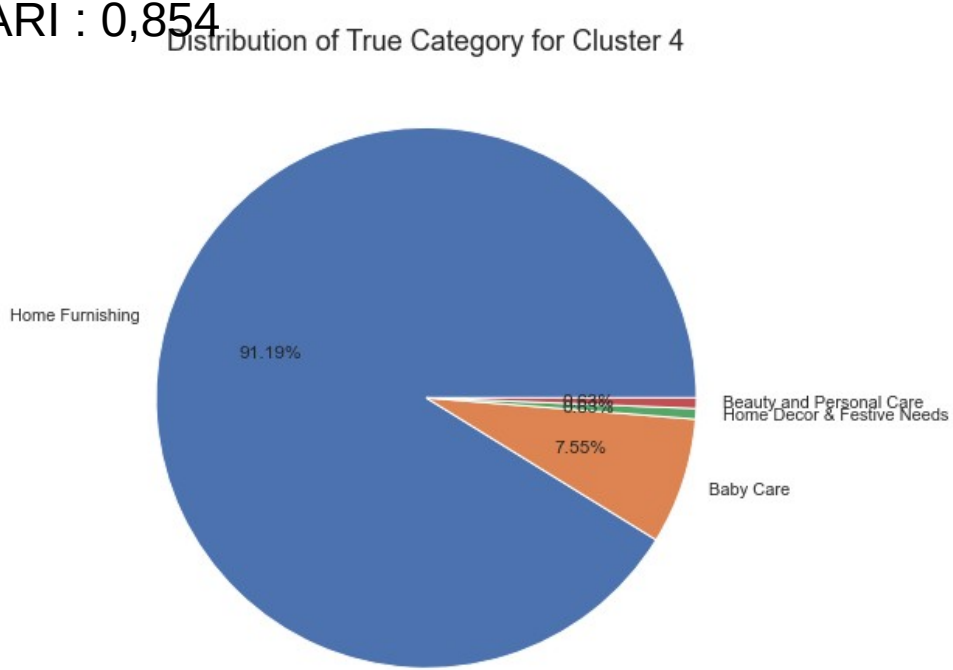


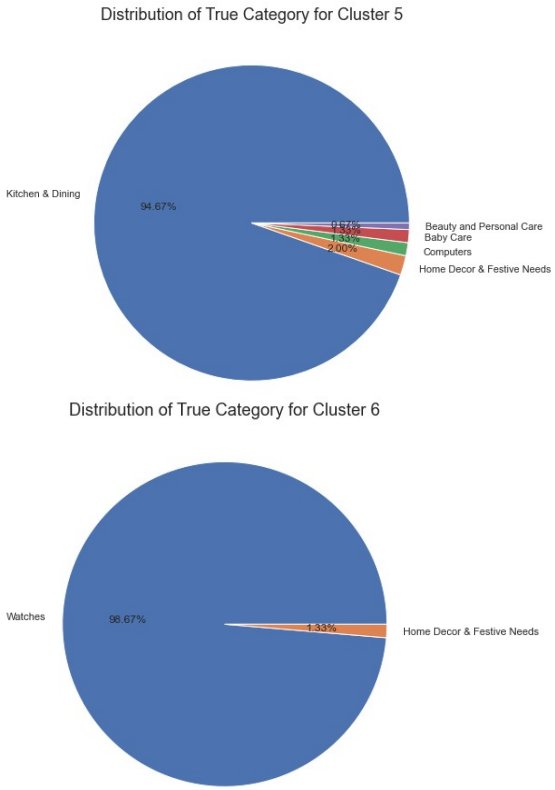




V. Clustering

Xception + transfer learning + fine-tuning (sans prétraitement d'image)
ARI : 0,854





VI. Conclusion

- **Faisabilité du moteur de classification pour données textes :**
  - Résultats mauvais pour les méthodes non supervisées
  - Résultats bons pour les méthodes supervisées
- **Faisabilité du moteur de classification pour données images :**
  - Résultats mauvais pour les méthodes non supervisées
  - Résultats bons pour la méthode de réseaux de neurones (transfer learning)