PROJET RAKUTEN

#PRODUIT

#CLASSIFICATION

#MACHINE LEARNING

#VISUALISATION

Equipe projet :

- Louis VALENTIN
- Souleymane TOURE
- Ouissam GOUNI
- Abdel YEZZA

Date : août 2024

Version : 0.1

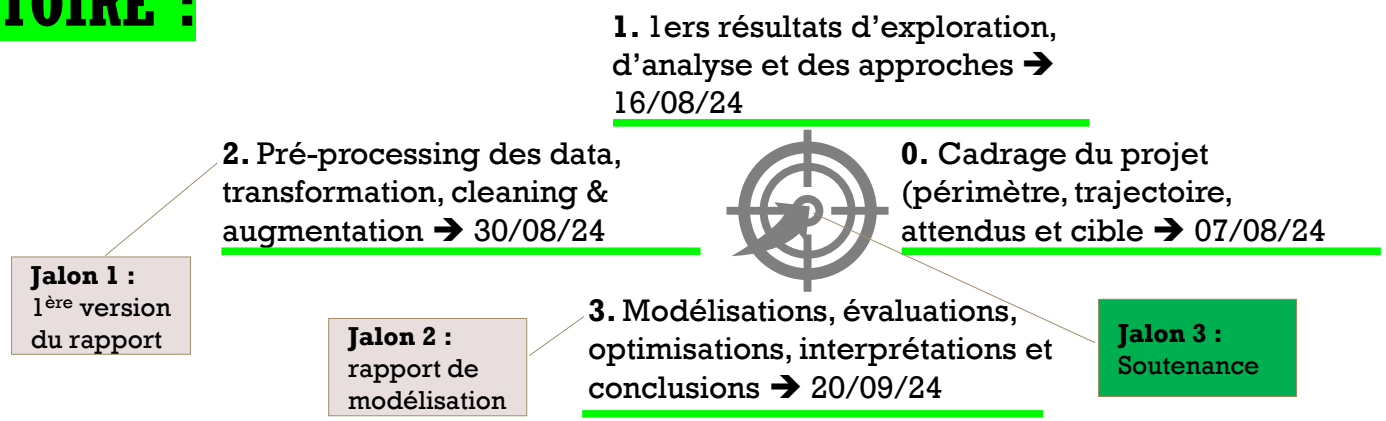
RÉSUMÉ DU PROJET RAKUTEN

RAKUTEN

CONTEXTE : Ce projet s'inscrit dans le cadre de la promotion de formation DataScientist de juillet 2024. Il porte sur des données provenant de **RAKUTEN Institut of technology** et a fait l'objet d'un challenge avec une publication des résultats des benchmarks

OBJECTIF : Sur la base des datasets fournis, construire un ou plusieurs modèles en mesure de catégoriser tout produit à partir de données descriptives et/ou des images selon le catalogue de référence RAKUTEN.

TRAJECTOIRE :



DONNEES SOURCES (1/4)
LISTE DES FEATURES X_TRAIN, X_TEST (VARIABLES HORS IMAGES)

Datasets fournis :	Description
X_train_update.csv : - **Dimensions :** (84916, 4) - Les articles les plus récents référencés, datent de 2019 ! Les données n'ont pas été mises à jour depuis.	Ensemble des données d'entrainement ayant été utilisé dans le cadre du du concourt **challenge Rakuten France Multimodal Product Data Classification**. Il est constitué des colonnes (features) : 1. **index numérique incrémental (1ère colonne) ne représentant pas une variable** 2. **designation** : désignation textuelle du produit, donnée par RAKUTEN sur son site e-commerce. *Les identifiants ne sont pas codés sur une même longueur* (longueurs : [10, 9, 8, 7, 6]) . 3. **description** : Description détaillée du produit ayant comme source un contenu HTML 4. **productid** : identifiant unique du produit référencé par RAKUTEN. *Les identifiants ne sont pas codés sur une même longueur* (longueurs : [10, 9, 8, 6, 7, 5]). 5. **Imageid** : un identifiant unique de l'image qui représente l'article, il fait partie du nom du fichier matérialisant l'image avec le **productid** en plus pour faire le lien avec le produit illustré par l'image (voir Diapositive 6)
X_test_update.csv : - **Dimensions :** (13812, 4) - 14% de l'ensemble d'entrainement et test	Il est constitué des mêmes colonnes que l'ensemble d'entrainement avec une 1ère dimension moindre évidemment **Remarques :** 1. La proportion de **14%** que constitue l'ensemble de test peut ne pas être suffisante pour traiter la plupart des cas possibles et ainsi donner des résultats satisfaisants lors des prédictions et validations. La constance du contenu de l'ensemble de test implique celle de l'entrainement. Cela risque de mener à un « **overfitting** » et ainsi donner des résultats décevants lors de la phase de validation sur des cas que le modèle n'est pas en mesure de bien les situer et donc prédire. 2. L'ensemble de test ne devrait pas être constant en étant constitué des mêmes données lors de la phase de test des modèles à entrainer/tester

DONNEES SOURCES (2/4)
DATASETS TRAIN & TEST (X_TRAIN, X_TEST) – INFORMATIONS DE BASE

X_train_update.csv

Shape : (84916, 4)

```
Index: 84916 entries, 0 to 84915
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   designation  84916 non-null  object
 1   description  55116 non-null  object
 2   productid    84916 non-null  int64
 3   imageid      84916 non-null  int64
dtypes: int64(2), object(2)
```

	designation	description	productid	imageid
manquantes	0	29800	0	0
manquantes (%)	0	35.10	0	0
unique (nombre)	82265	47507	84916	84916
uniques (%)	96.88	55.94	100	100

~3% des désignations répétées

~45% des descriptions répétées. S'agit-il d'articles dans la même catégorie ? **(Effet de Copier/Coller !)**

Anomalie dans le fichier **X_train_update.csv (lignes 58871 à 58873** à corriger ! **Supprimer ces lignes**
Article avec **index=58868** n'a ni **productid** ni **imageid** (peut-être une conséquence) ➔ à renseigner ou supprimer

```
58870   58868,Manette Contrôleur Classic Pro Pour Nintendo Wii Wii U - 120 M - Blanc,"Jouez avec une manette classique à votre Nintendo Wii ou Wii U.<br>Se branche sur la Wiimote.<br>Manette avec
58871   50  2118164518  ...  Ttx Tech Manette Pad Joystick Analogique Filaire Usb Pour Playstation 3/Pc Blanc  Accessoires jeux vidéo  Jeux-Video-et-Consoles  Jeux-Video-et-Consoles_Accessoires
58872   1160  2724395348  ...  Nezahal Marée Primordiale - Mtg - Les Combattants D&#39;Ixalan - R - 45/196 Cartes de jeux  Jeux-Video-et-Consoles  Jeux-Video-et-Consoles_Cartes-de-jeux  ...  ...
58873   50  1502606277  ...  Unité Motion Plus Contrôle Manette Wiimote Console Jeu Nintendo Wii + Housse Silicone  Accessoires jeux vidéo  Jeux-Video-et-Consoles  Jeux-Video-et-Consoles_Accessoi
58874   58869,Kit piscine acier aspect bois Gré Sicilia ovale 527 x 327 x 122 m + Bâche à bulles,Une piscine acier imitant le bois à la perfection pour embellir votre jardin. Peu encombrante elle
```

X_test_update.csv

Shape : (13812, 4)

```
Index: 13812 entries, 84916 to 98727
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   designation  13812 non-null  object
 1   description  8926 non-null   object
 2   productid    13812 non-null  int64
 3   imageid      13812 non-null  int64
dtypes: int64(2), object(2)
```

	designation	description	productid	imageid
manquantes	0	4886	0	0
manquantes (%)	0	35.37	0	0
unique (nombre)	13681	8347	13812	13812
uniques (%)	99,05	60,43	100.0	100.0

1% des désignations répétées

~40% des descriptions répétées

DONNEES SOURCES (3/4)
L'ENSEMBLE DES CLASSIFICATIONS CONNUES DES PRODUITS (Y_TRAIN)

Datasets fournis :

Description

Y_train_CVw08PX.csv :

- **Dimensions :** (84916, 1)

```
Index: 84916 entries, 0 to 84915
Data columns (total 1 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   prdtypecode  84916 non-null  int64
dtypes: int64(1)
```

	prdtypecode
count	84916.000000
mean	1773.219900
std	788.179885
min	10.000000
25%	1281.000000
50%	1920.000000
75%	2522.000000
max	2905.000000

Ensemble des données cibles d'entrainement connues ayant été utilisées dans le cadre du concourt **challenge Rakuten France Multimodal Product Data Classification**. Il est constitué des colonnes :

1. **index numérique incrémental (1ère colonne) alignée sur l'index de l'ensemble d'entrainement utilisé (X_train)**

2. **prdtypecode** : représente le code de la catégorie à laquelle le produit appartient. Cette variable est la cible dans laquelle les modèles doivent classer n'importe quel article à partir d'images et/ou texte/mots

Remarques :

1. Aucune donnée n'est manquante !
2. Valeurs manquantes/uniques :

	prdtypecode	
manquantes	0	
manquantes (%)	0	
unique (nombre)	27	Ce n'est pas suffisant
uniques (%)	0.031796	
Liste des codes :	10 40 50 60 1140 1160 1180 1280 1281 1300 1301 1302 1320 1560 1920 1940 2060 2220 2280 2403 2462 2522 2582 2583 2585 2705 2905	

- **Uniquement 84900 lignes dans « X_train_update.csv » possèdent des « prdtypecode » dans « Y_train_CVw08PX.csv » à cause de l'anomalie notée précédemment dans le fichier X_train_update.csv. Par conséquent, on ne garde que 84900 lignes au total (16 de moins).**
- **Convertir tous les champs numériques en type int64 au lieu de float64 à la source**

DONNEES SOURCES (4/4)
LES IMAGES

1. Les images sont réparties au même titre que les datasets d'entrainement et de test, dans deux dossiers distincts, un pour l'entrainement contenant 84916 et l'autre pour le test contenant 13812 en concordance avec les datasets textuels

2. Toutes les images sont au **format JPG**

3. Chaque image porte le nom : **image_xxxxxxx_product_yyyyy.jpg**

   Représente la donnée sous la colonne **imageid** des dataset textuels avec une longueur allant de 5 à 10

   Représente la donnée sous la colonne **productid des dataset textuels avec une longueur allant de 6 à 10**

4. Elles sont toutes de la taille **500x500** avec des couleurs **RGB ([0, 255], [0, 255], [0, 255])**. Par conséquent, les numériser en couleurs d'origine dans un dataset peut être couteux en traitement et en espace de stockage.

5. Beaucoup d'images contiennent du texte, ce qui peut révéler le type de l'article à partir des mots que l'on peut extraire de l'image. Cela peut être renforcé en testant dans les article associés (lignes) la présence de ces mots dans les champs textuels **designation** et **description**

6. Pour chaque article, une seule image est mise à disposition, ce qui peut limiter les informations que l'on peut tirer au travers de la reconnaissance de l'image (OCR) et des mots qu'elle peut contenir si elles sont de bonne qualité

7. Beaucoup d'images ne correspondent plus à l'état actuel du marché; elles représentent des articles qui ne sont pas plus vendus ou ont évolué en image et en texte (modèle, version, auteur, acteur etc.)

QUELLES OPTIONS POUR LES DATASETS ET IMAGES ?

1. DATASETS textuels

✗ Option 1.1

1. Garder les mêmes datasets fournis sans aucune opération de type fusion ou concaténation
2. Analyser séparément les datsets qui feront l'objet d'exploration, d'analyse et de pré-processing

✗ Option 1.2

1. **Concaténer** les sets train et test utilisés (X_train_update.csv et X_test_update.csv) dans un CSV unique
2. **Fusionner** le dataset des classifications connues (Y_train_CVw08PX.csv). Remplacer les « **prdtypecode** » manquants.
3. Le résultat final, disons **rakuten.csv**, contiendra l'ensemble des données texte utilisées qui fera l'objet d'analyse, d'exploration, d'augmentation par des nouvelles variables et de pré-processing

✓ Option 1.3

1. Garder **X_train_update,csv** et **Y_train_CVw08PX**.
2. **X_test_update.csv** doit être écarté du fait qu'il ne possède pas d'infos sur les catégories de produits
3. Les fusionner en **un seul dataset que l'on appellera** rakuten.csv
4. Garder uniquement les images qui correspondent au dataset

Facteurs des 1ers choix

1. Complétude des variables
2. Disponibilité des variables manquantes afin de compléter les informations nécessaires aux modèles
3. Cohérences des variables (texte/nombre/image) avec l'information véhiculée et inter-variables
4. Remplacement possible des variables manquantes n'impactant pas l'existant
5. 1ères statistiques sur la base des variables satisfaisantes
6. Les données sont actuelles pour rendre le classement approprié à l'état du marché d'e-commerce

2. IMAGES

✓ Option 2.1

1. Garder les images telles quelles en couleurs
2. Préserver leurs dimensions

✓ Option 2.2

1. Faire passer les images au mode « **grayscale** »
2. Et/ou Réduire la taille des images

✗ Option 2.3

1. Numériser les images et les faire passer à un dataset. Chaque image occupera 500*500 = 250 000 éléments x [R, G, B] !
2. Numériser les images après le « **grayscaling** » réduisant ainsi les 3 canaux RGB à un seul
3. En plus du « **grayscaling** » réduire la taille des images

Facteurs des 1ers choix

1. Les images sont exploitables
2. Si on opte pour un « **grasacling** », les images demeurent utilisables et fourniront des résultats presque à l'identique au cas des images en couleurs
3. Si on opte pour une numérisation des images, cela peut être couteux en traitement et maintenance ➔ écarté

Options retenues

DATASETS textuels

1. …

IMAGES

1. …

REMARQUES GLOBALES

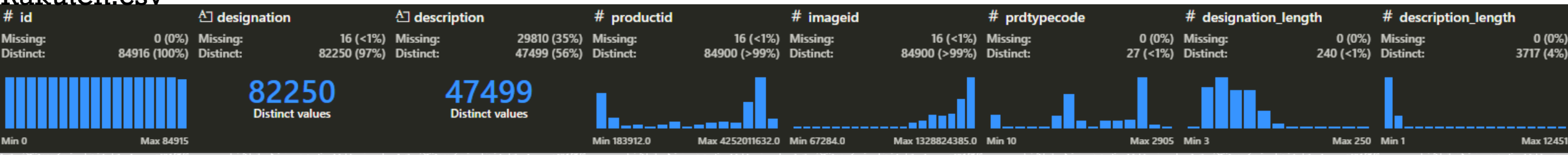
DATASETS textuels

1. Les variables présentes sont loin d'être suffisantes
2. Les données relatives à la désignation des produits ne sont pas nécessairement toutes d'actualité, voire même obsolètes
3. Quant aux descriptifs, ils peuvent être toujours valables s'ils n'ont pas d'adhérence forte avec la désignation du produit, autrement dit s'ils sont toujours d'actualité
4. ~~Le nombre de catégories de produits de 27 n'est pas suffisant pour entrainer le dataset, il y a un véritable déséquilibre réel entre la taille du dataset et les catégories uniques identifiées~~
5. Rajouter une variable descriptive qui représente le nom de chacune des 27 catégories déjà connues par son ID en s'appuyant sur les images qui leurs sont associées (travail manuel !)

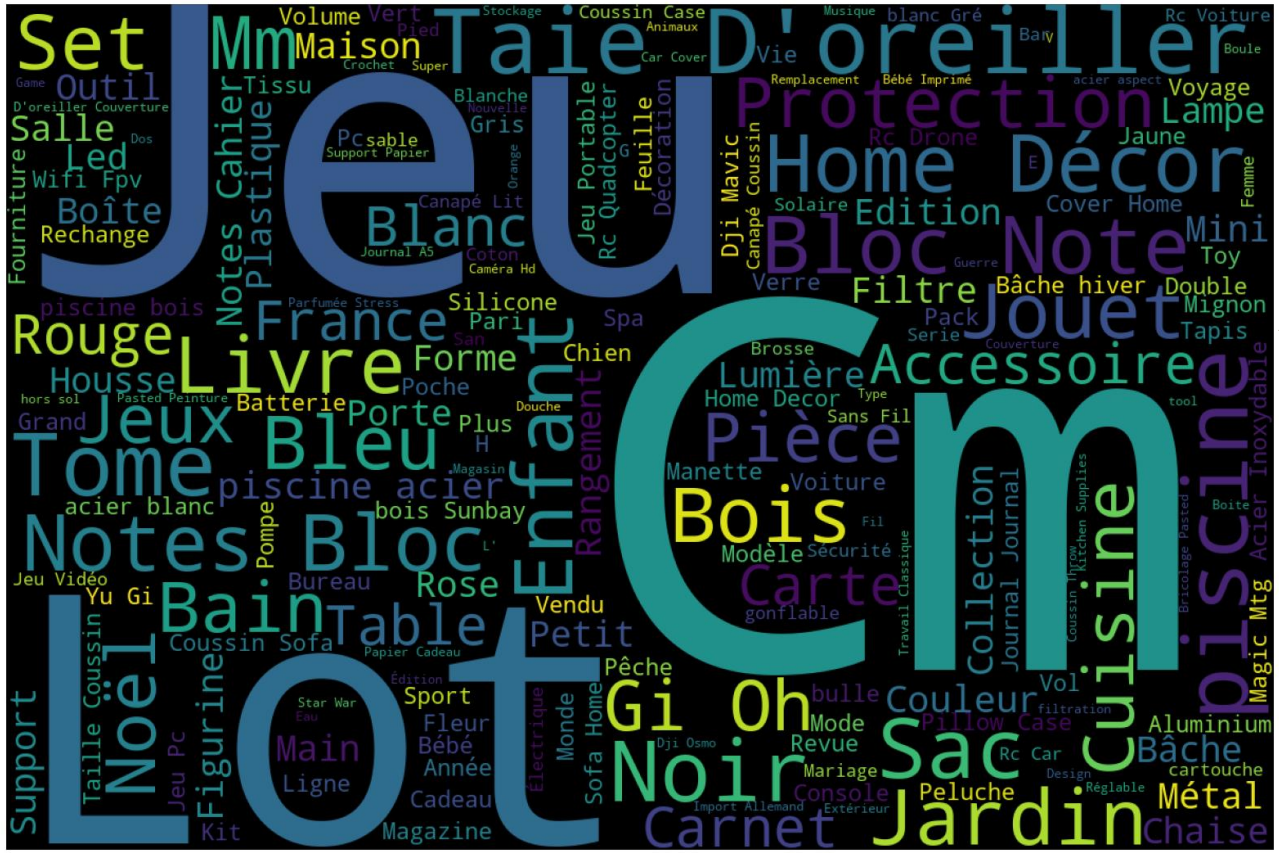
IMAGES

1. …

Rakuten.csv

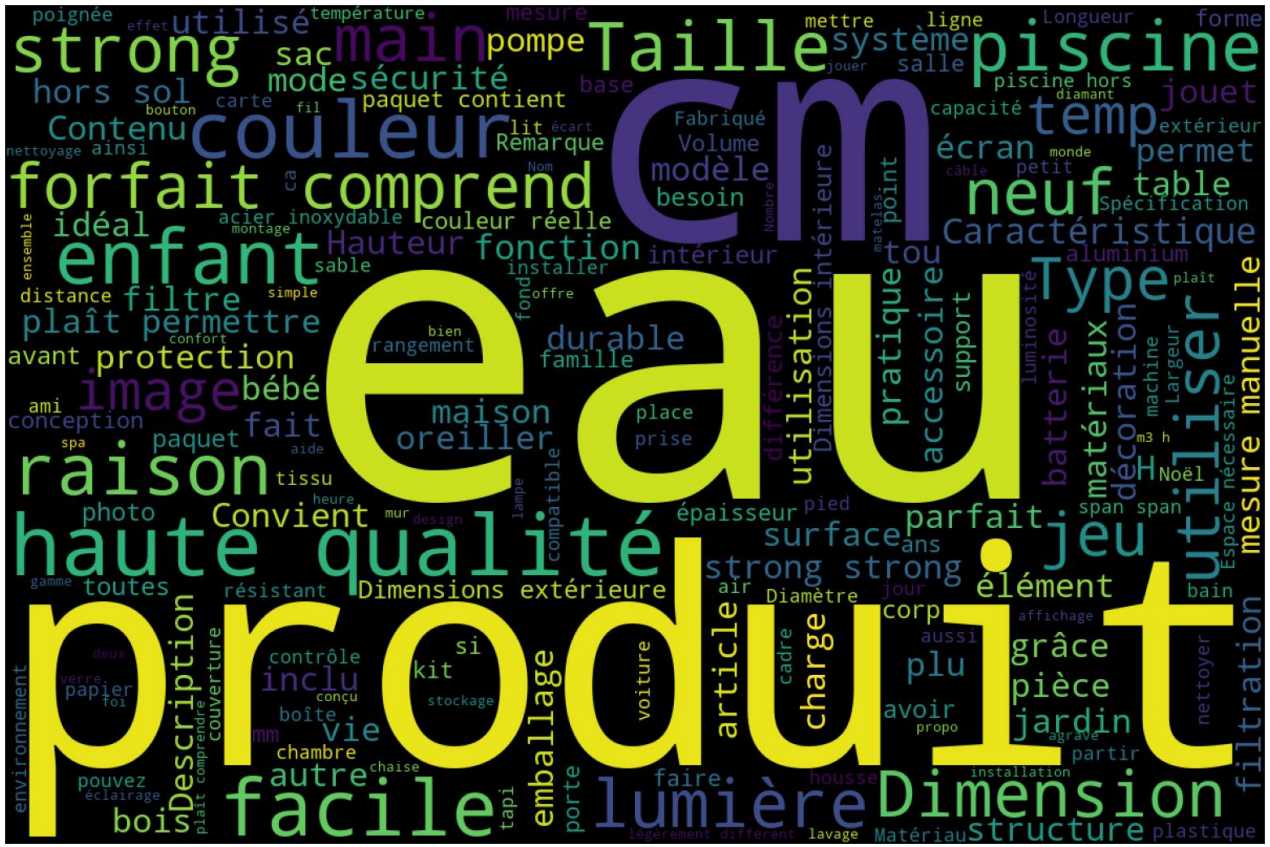


CARTOGRAPHIE DES MOTS UTILISES



Designation

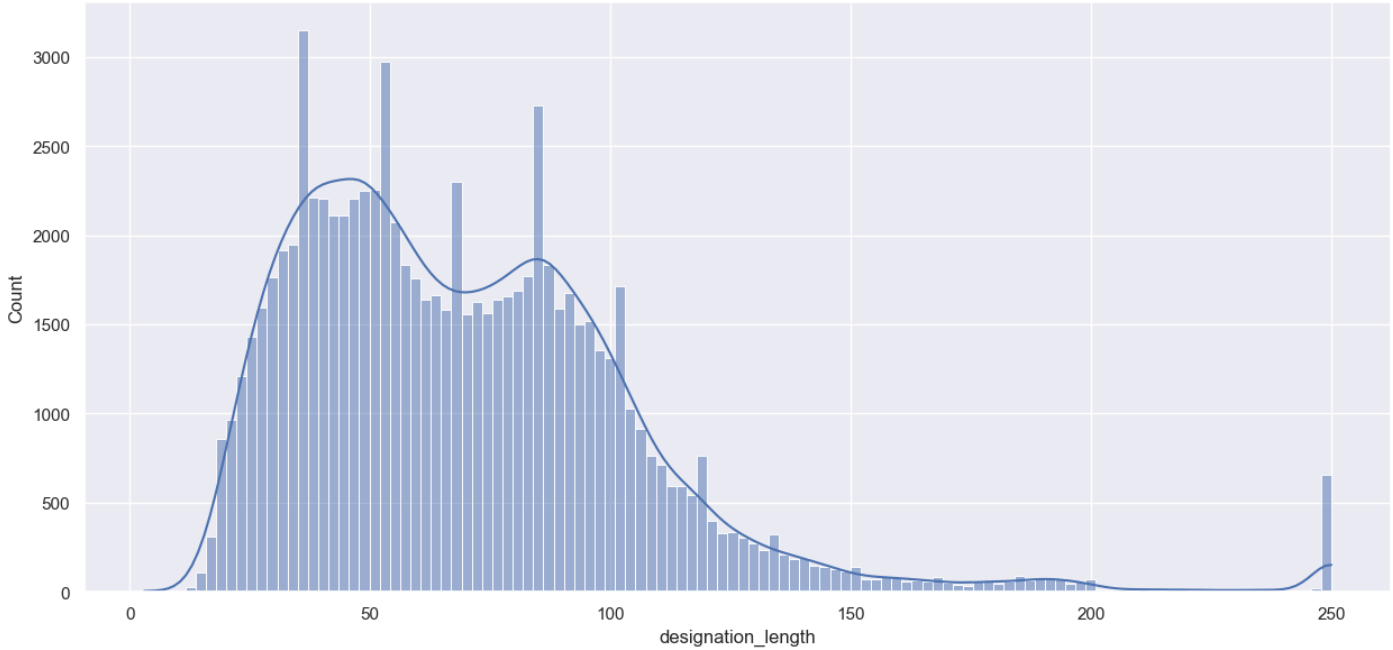
Les mots les plus fréquents révèlent en partie les catégories de la majorité des articles du dataset

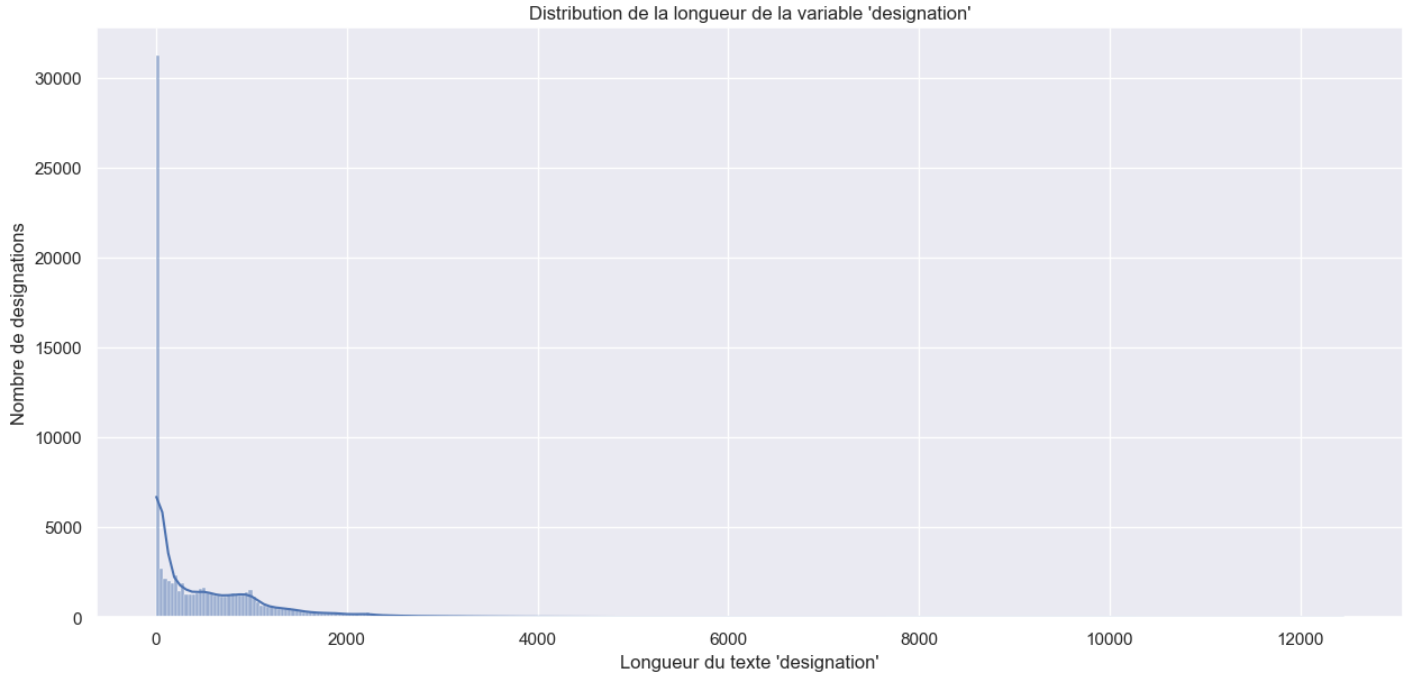


Description

- Tout le code HTML exclu
- Les mots les plus répétés révèlent l'importance accordée aux critères les plus recherchées/mis en avant par les clients/e-commerçants

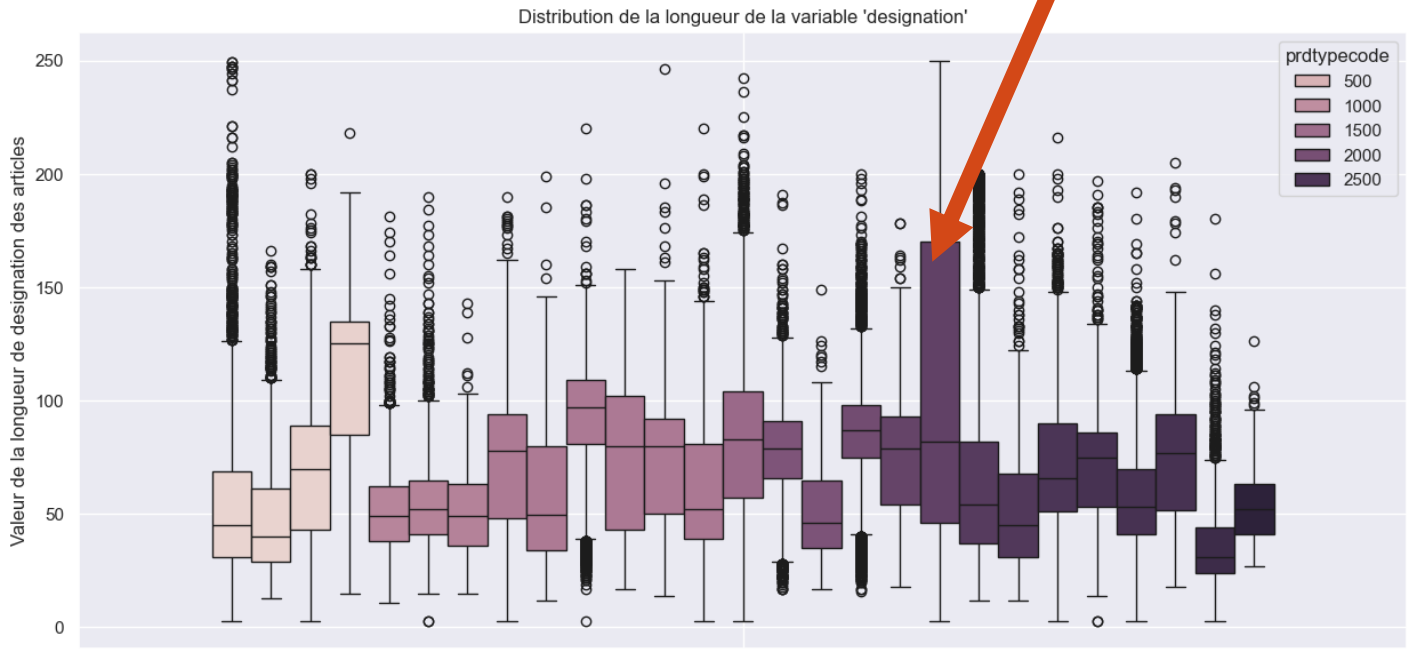
QUELQUES GRAPHES PRELIMINAIRES (1/3)





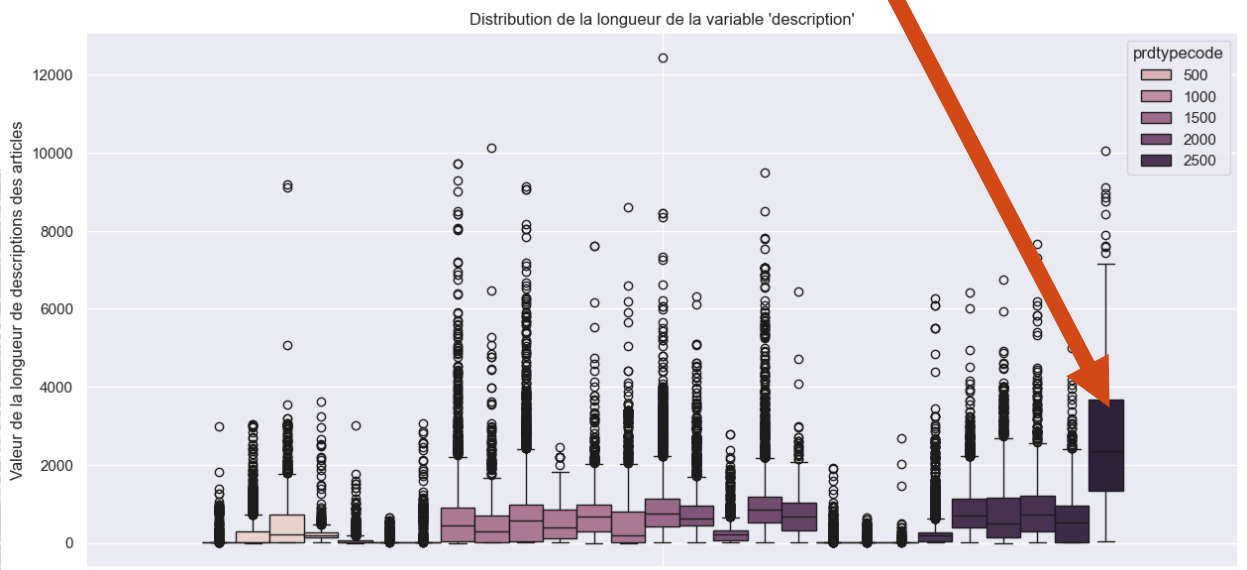
QUELQUES GRAPHES PRÉLIMINAIRES (2/3)

Quelle catégorie qui se démarque des autres ?



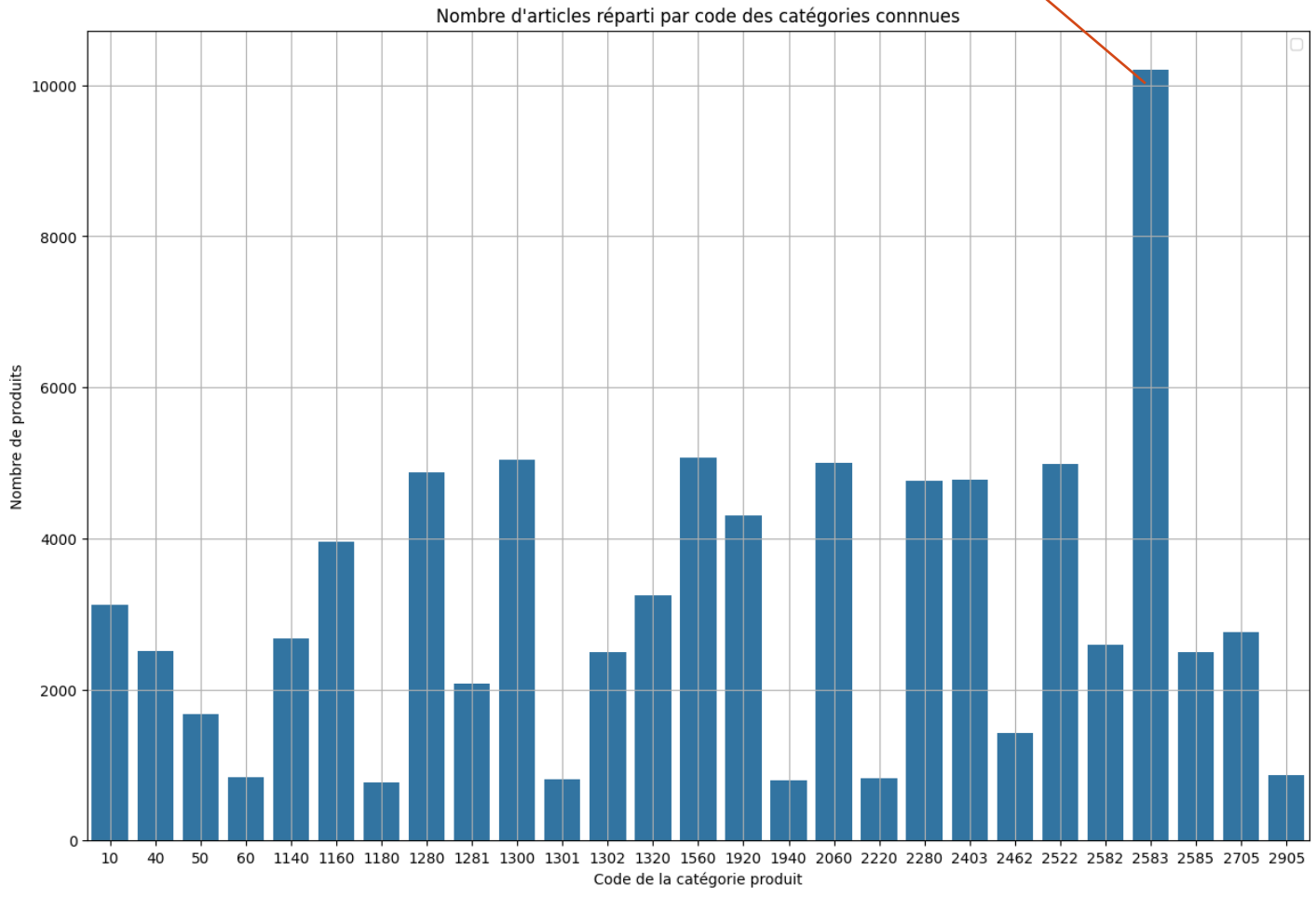
Refaire ce graphe une fois on disposera des noms des catégories des produits plus parlant au lieu des codes

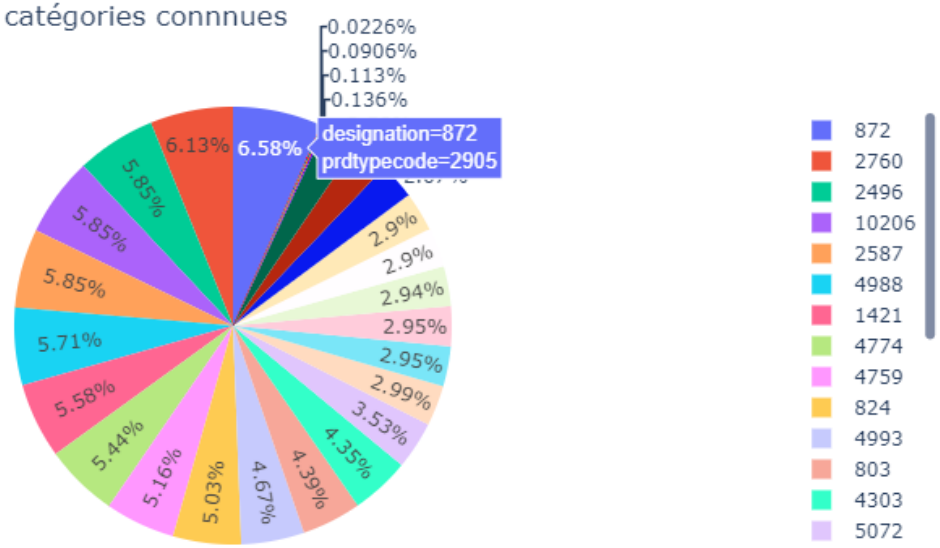
Quelle catégorie qui se démarque des autres ?



QUELQUES GRAPHES PRELIMINAIRES (3/3)

Examiner des méthodes existantes afin d'équilibrer le nombre de catégories dans le dataset





1. Rajouter une variable descriptive de la catégorie (27 en tout)
2. La catégorie **2585** est trop dominante, un équilibrage doit être opéré pour éviter l'overfitting (à voir…)

CAS D'USAGE - EXEMPLES

Use case 1

1. Je visite le site e-commerce en tant que visiteur ou client
2. Je consulte plusieurs articles dans une catégorie spécifique sélectionnée sur le site **CAT1** et référencée explicitement
3. Sur la base de la catégorie **CAT1**, le site peut proposer en plus d'autres articles se rapprochant de celle-ci sur la base des images, des désignations et descriptions des articles, de préférence dans la même catégorie

US 1 (User Story)

1. En tant que **visiteur** du site e-commerce consultant une catégorie d'articles spécifique, **je souhaite** le trouver très rapidement sur la base de mes consultations **afin de** passer le moins de temps

Use case 2

1. Je visite le site e-commerce en tant que visiteur ou client
2. J'effectue une recherche en saisissant plusieurs mots clés faisant référence à sa désignation ou un descriptif de son utilité et usage
3. Sur la base de la saisie de l'utilisateur, le site peut proposer plusieurs catégories d'articles se rapprochant le plus du descriptif de l'utilisateur et de ses désirs
4. Les articles affichés peuvent être classés par ordre de rapprochement décroissant en affichant d'autres articles auxquels l'utilisateurs n'avait pas pensé nécessairement ou ne font pas partie de ses désirs sur le moment

US 2 (User Story)

1. En tant que **visiteur ou client** du site e-commerce **je souhaite** trouver l'article recherché très rapidement en saisissant des mots clés **afin de** passer le moins de temps

Use case 3

1. Je visite le site e-commerce en tant que visiteur ou client
2. Je sélectionne des articles et les rajoutent au panier
3. Sur la base du contenu du panier et selon les catégories auxquelles ils appartiennent, le site peut proposer d'autres articles se rapprochant le plus ces typologies
4. Les articles affichés peuvent être reliés respectivement aux articles du panier classés par ordre de rapprochement décroissant et auxquels l'utilisateurs n'avait pas pensé ou ne font pas partie de ses désirs sur le moment

US 3 (User Story)

1. **En tant que client ou visiteur** du site e-commerce, **je souhaite** que l'on me propose d'autres articles dans la même catégorie de ceux de mon panier et qui pourraient m'intéresser en termes de choix et de prix **afin de** profiter des promotions en cours et disponibilité immédiate

HISTORIQUES-ACTIONS

Date	Sujet	Actions	Auteur	Echéance
14/08/24	Ecarter le dataset de test X_test_update.csv **ET NE GARDER QUE** X_train_update.csv **(DATASET DE BASE) ET** Y_train_CVw08PX.csv qui contient toutes les cibles.			28/08/24
	Rajouter une variable descriptive de la catégorie (27 en tout) et peupler le dataset avec			28/08/24
	La catégorie **2585** est trop dominante, un équilibrage doit être opéré pour éviter l'overfitting (à voir comment…)			28/08/24

- **Résumé des colonnes rakuten.csv :**