Préparez des données pour un organisme de santé publique

Soutenance Projet 3 : Habibatou BA 10/08/2023

Sommaire

- Problématique
- Analyser les données de Open Food Facts
- Description des données
- Pré-traitement des données
- Nettoyage des données
- Analyse exploratoire des données
- Les grandes principes du **RGPD**
- Conclusion
- Suite du Projet

Problématique

Santé publique France

L'agence Santé publique France souhaite **améliorer sa base de données Open Food Facts** et fait appel aux services de l'entreprise pour la création d'un système de suggestion ou d'auto-complétion pour aider les usagers à remplir plus efficacement la base de données.

Problématique

- Ces données peuvent-elles être mis à disposition à des particuliers et d'organisations afin de leur permettre de connaître la qualité nutritionnelle de produits. ?

- Déterminer la faisabilité de cette idée d'application de Santé publique France?

Analyser les données de Open Food Facts

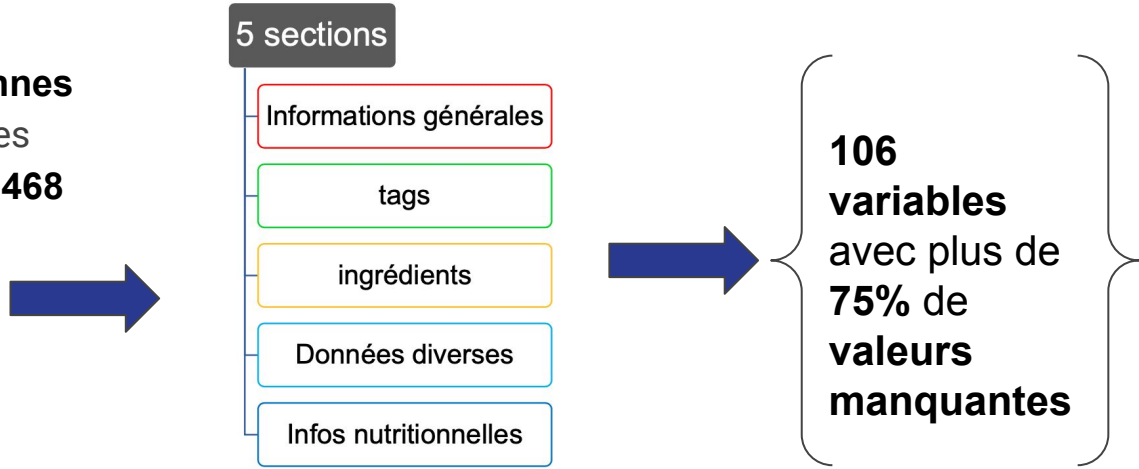
- Repérer des variables pertinentes
- Mettre en évidence les éventuelles valeurs manquantes parmis les variables pertinentes sélectionnées, avec au moins 3 méthodes de traitement adaptées aux variables concernées
- Identifier et traiter les éventuelles valeurs aberrantes de chaque variable.
- Sélectionner / créer des variables à l'aide d'une analyse multivariée.
- Effectuer les tests statistiques appropriés pour vérifier la significativité des résultats.

Santé publique France

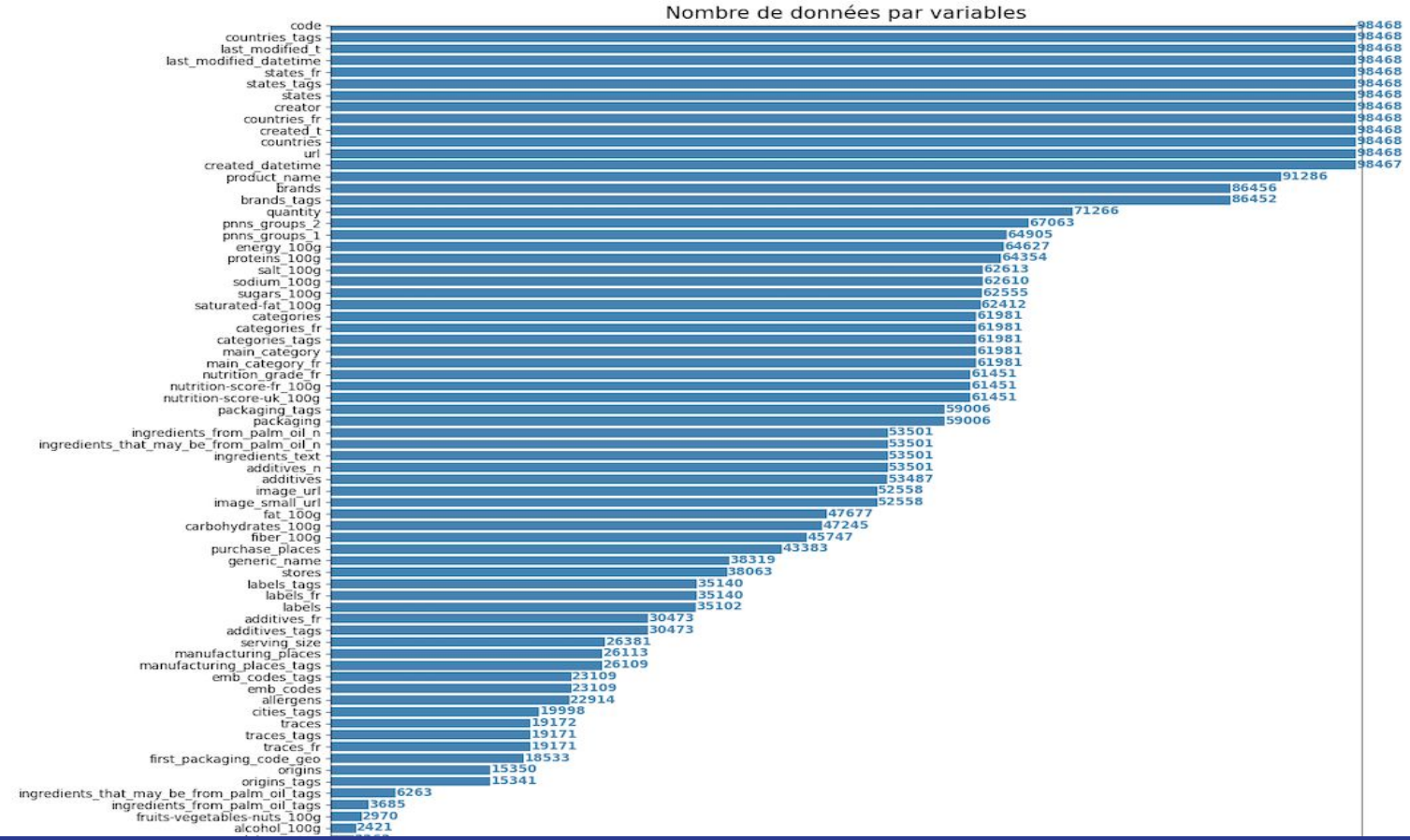
open FOOD facts

Descriptions des données

- Données disponibles à l'adresse : le site officiel
- Constitué de **320772 lignes** et **162 colonnes**
- Après filtrage uniquement sur les données de la **France**. Nous avons travaillé sur **98468 lignes**



Descriptions des données : valeurs utiles



Santé publique France

open FOOD facts

Pré-traitements des données : valeurs utiles

Santé publique France

code
product_name
pnns_groups_1
pnns_groups_2
categories
additives
nutrition_grade_fr
brands
energy_100g
fat_100g
saturated_fat_100g
carbohydrates_100g
sugars_100g
'fiber_100g
proteins_100g
salt_100g
nutrition_score_fr_100g
additives_n
ingredients_from_palm_oil

Alimentation Saine

Alimentation équilibrée

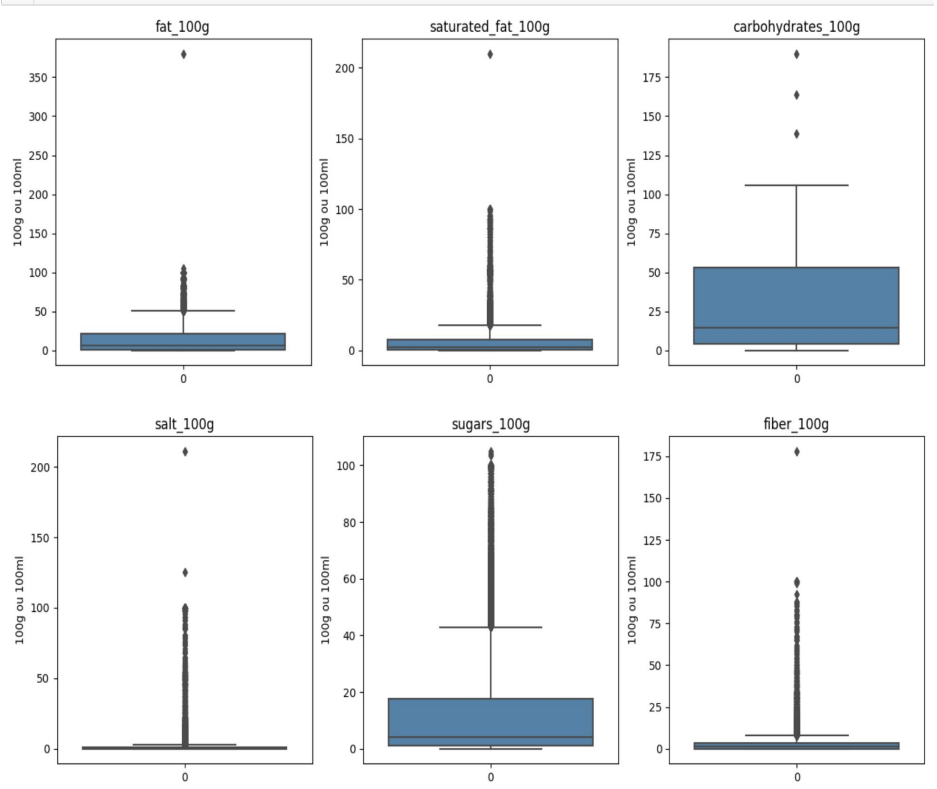
Produits disponibles

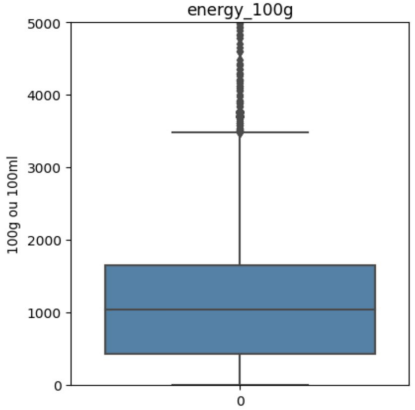
open FOOD facts

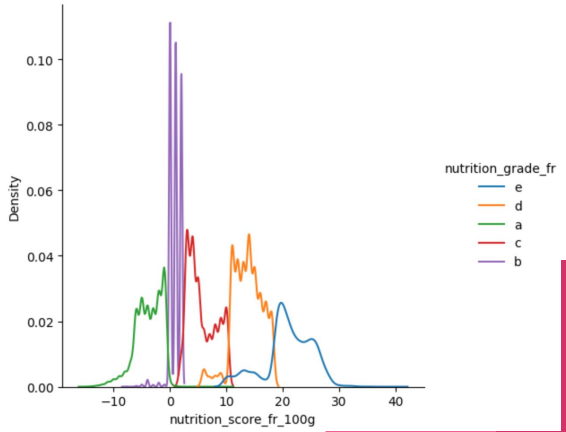
Nettoyage : réductions des données

- Chargement des données => 320772 lignes et 162 colonnes
- Conserver uniquement les produits français => 98468 lignes et 162 colonnes
- Suppression des variables avec %NaN > 80% => 98468 lignes et 60 colonnes
- Conserver les valeurs nutritionnelles utiles => 98468 lignes et 20 colonnes
- Suppressions des produits non renseignés et des noms de produits de moins de 3 caractères => 91265 lignes et 20 colonnes
- Gestions des valeurs aberrantes => 91166 lignes et 20 colonnes
- Gestions des valeurs manquantes => 41469 lignes et 20 colonnes

Nettoyage : valeurs aberrantes







Nettoyage : valeurs aberrantes

Outliers pour 100g ou ml de produits
énergie nutritionnelle > 3766 kJ
nutriments_100g compris entre 0 et 100g
Le score du nutri-score au dessus de 40

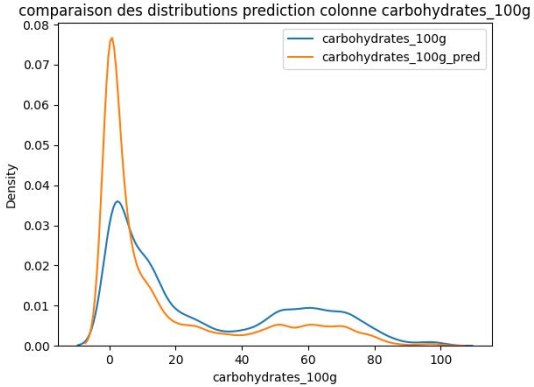
Macro-nutriment et sous-groupe de nutriment
Si la masse totale des glucides inférieure à la masse de sucre
Si la masse totale des lipide est inférieure à la masse des acides gras saturés
Suppressions des chevauchements entre les notes de nutri-score

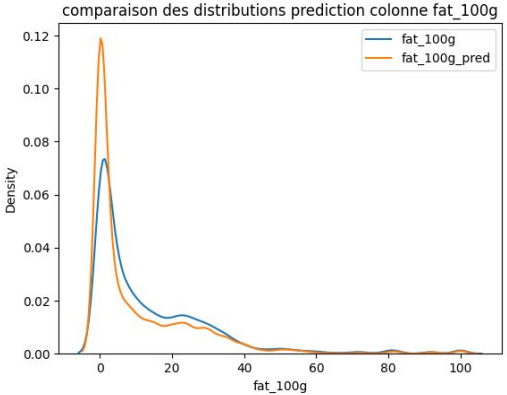
Nettoyage : valeurs manquantes : Méthodes d'Imputations

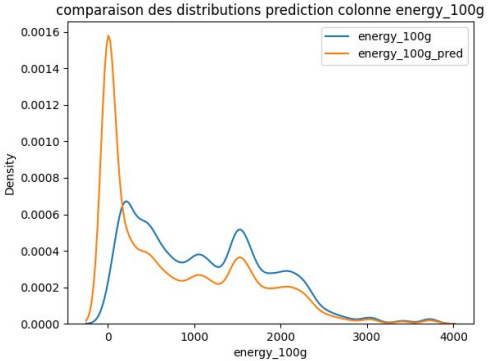
- IMputation par mise à '0' pour certaines variables
- IMputation par la moyenne
- IMputation par KNeighborsRegressor

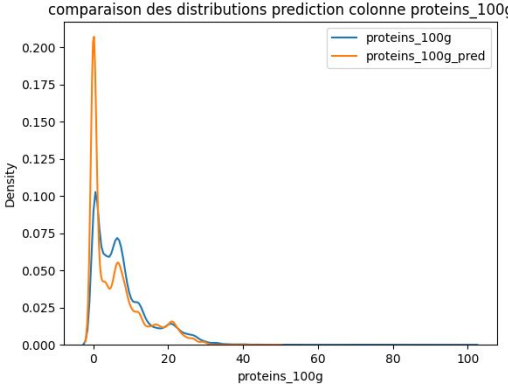
Nettoyage : Imputation des valeurs manquantes

Santé publique France





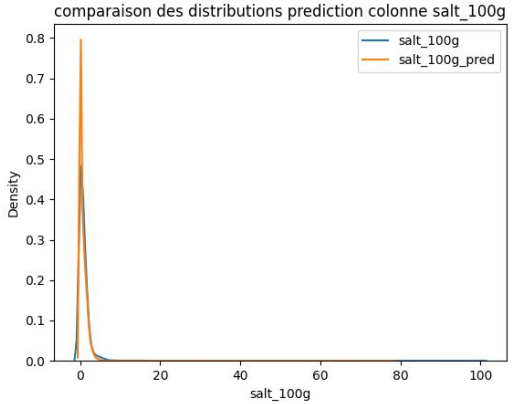


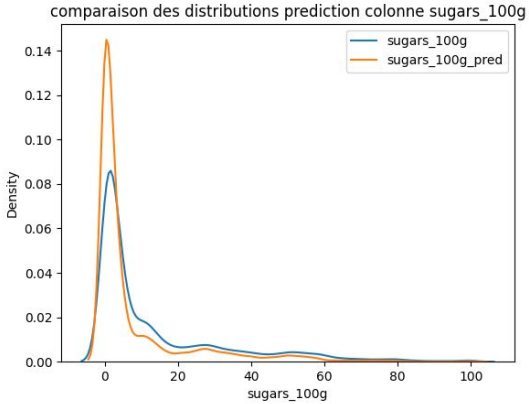


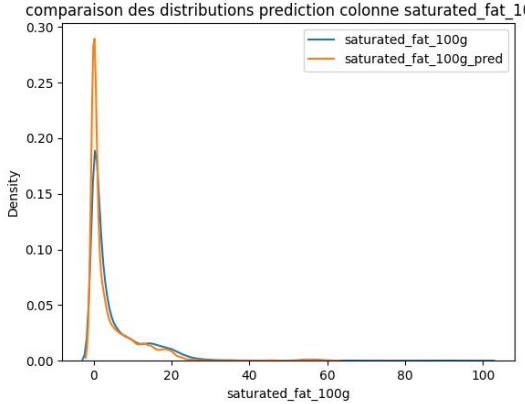
open FOOD facts

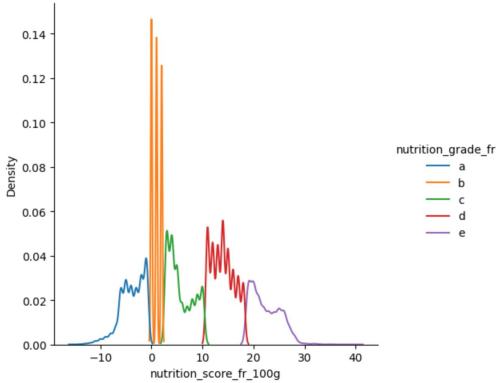
Nettoyage : Imputation des valeurs manquantes

Santé publique France





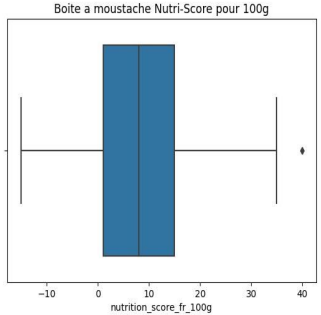


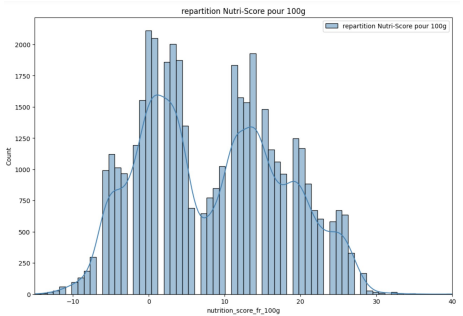


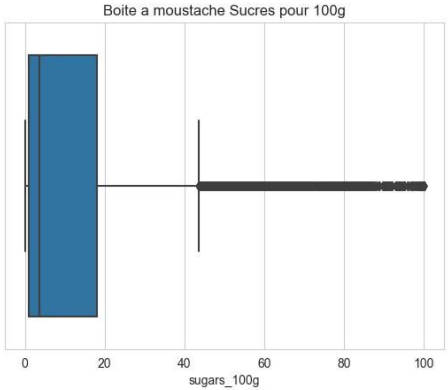
open FOOD facts

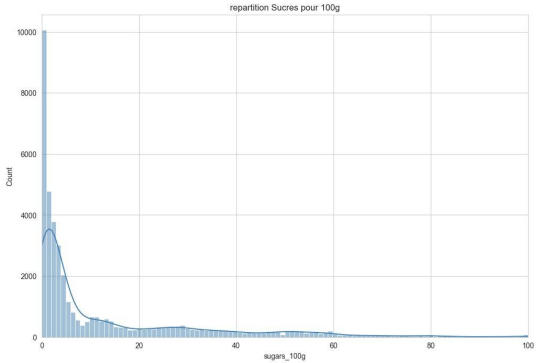
Analyse exploratoire des données : analyse univariée, nutriments

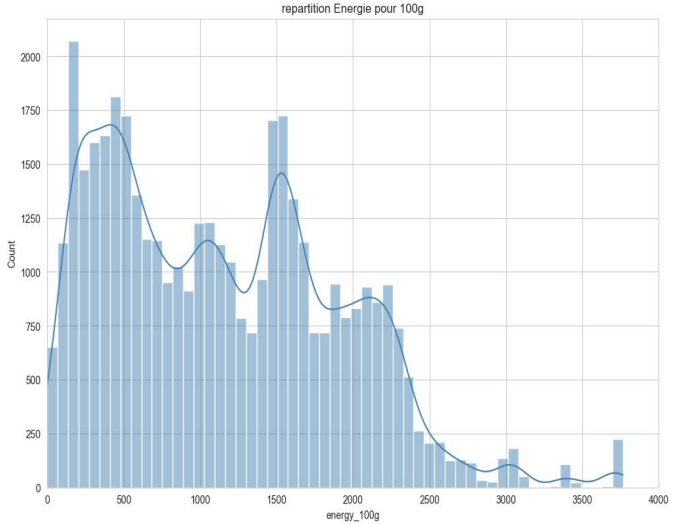
Santé publique France



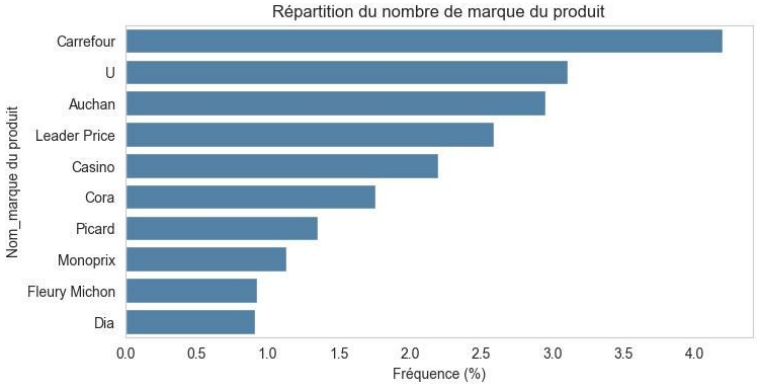


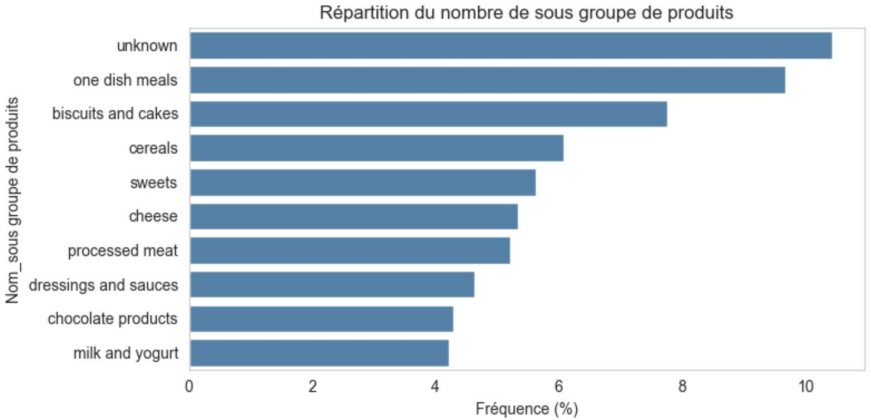


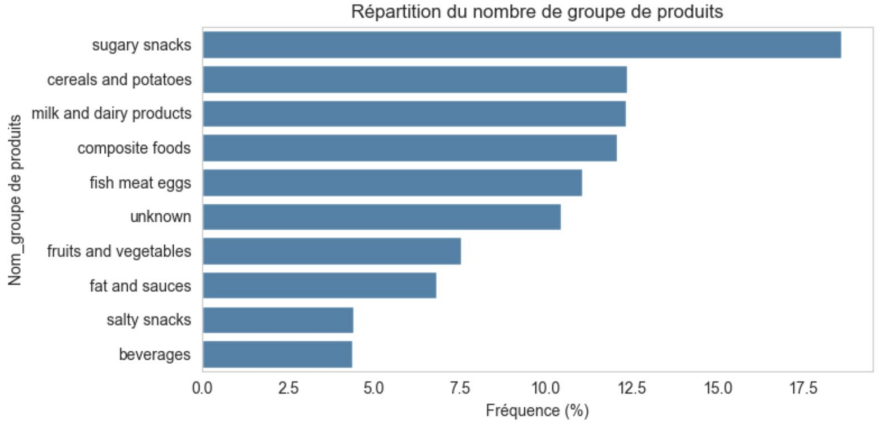


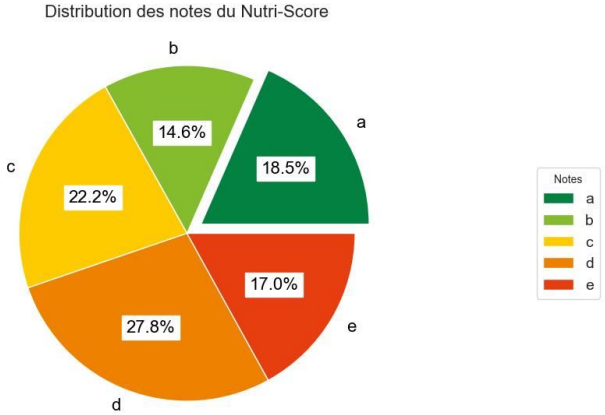


Analyse exploratoire des données : analyse univariée, les produits







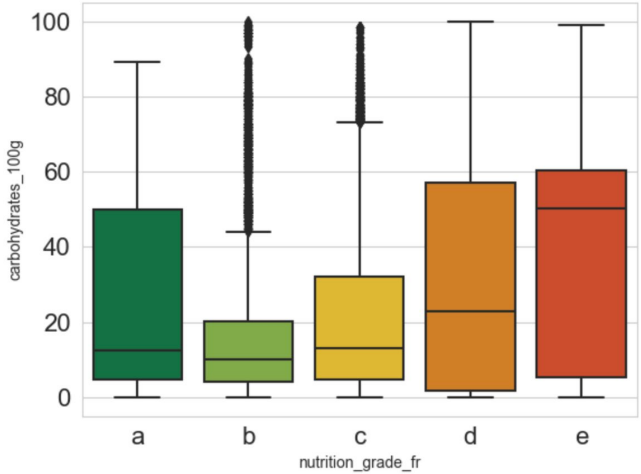


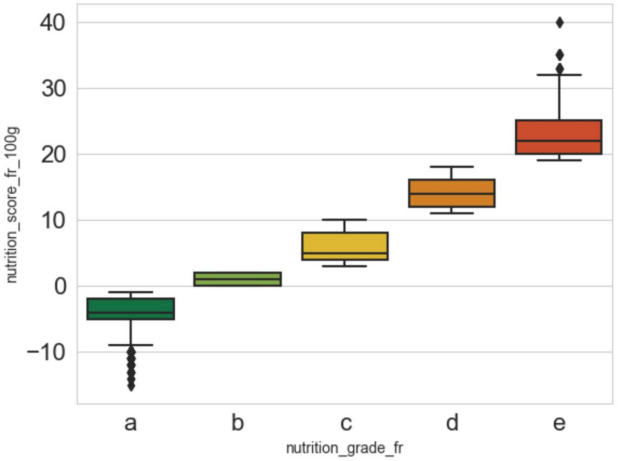
Analyse exploratoire des données : analyse univariée, descriptions des données

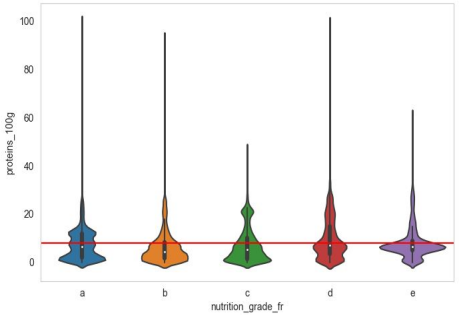
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
code	41530	41530	00013628	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
product_name	41530	34463	Huile d'olive vierge extra	38	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pnns_groups_1	41530	10	sugary snacks	7721	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pnns_groups_2	41530	35	unknown	4336	NaN	NaN	NaN	NaN	NaN	NaN	NaN
categories	41530	18540	Snacks sucrés,Biscuits et gâteaux,Biscuits	260	NaN	NaN	NaN	NaN	NaN	NaN	NaN
additives	41530.0	32637.0	0.0	4163.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
nutrition_grade_fr	41530	5	d	11532	NaN	NaN	NaN	NaN	NaN	NaN	NaN
brands	41530	8284	Carrefour	1743	NaN	NaN	NaN	NaN	NaN	NaN	NaN
energy_100g	41530.0	NaN	NaN	NaN	1151.817749	759.691406	0.0	479.0	1071.0	1676.75	3766.0
fat_100g	41530.0	NaN	NaN	NaN	14.297374	17.027325	0.0	1.9	8.0	22.5	100.0
saturated_fat_100g	41530.0	NaN	NaN	NaN	5.609253	8.299082	0.0	0.4	2.2	8.0	100.0
carbohydrates_100g	41530.0	NaN	NaN	NaN	27.658638	27.266119	0.0	3.7525	15.0	53.0	100.0
sugars_100g	41530.0	NaN	NaN	NaN	13.081664	18.896763	0.0	1.0	3.6	18.0	100.0
fiber_100g	41530.0	NaN	NaN	NaN	2.691402	3.094872	0.0	1.2	2.556959	2.556959	100.0
proteins_100g	41530.0	NaN	NaN	NaN	7.963834	7.311999	0.0	2.5	6.3	11.0	100.0
salt_100g	41530.0	NaN	NaN	NaN	1.042831	3.234858	0.0	0.1	0.64	1.3	100.0
nutrition_score_fr_100g	41530.0	NaN	NaN	NaN	8.407344	9.194624	-15.0	1.0	8.0	15.0	40.0
additives_n	41530.0	NaN	NaN	NaN	1.871466	2.590535	0.0	0.0	1.0	3.0	31.0
ingredients_from_palm_oil	41530.0	NaN	NaN	NaN	0.232723	0.572549	0.0	0.0	0.0	0.0	5.0

Analyse exploratoire des données : analyse bivariée

Santé publique France

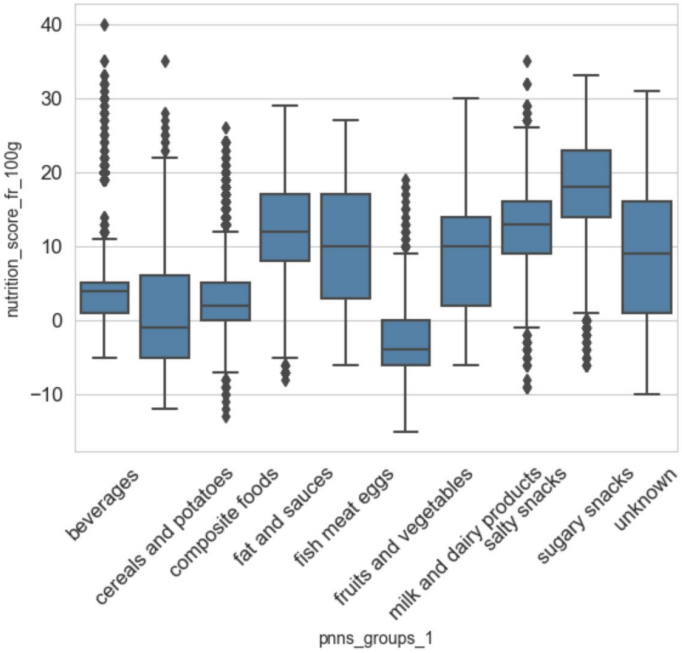


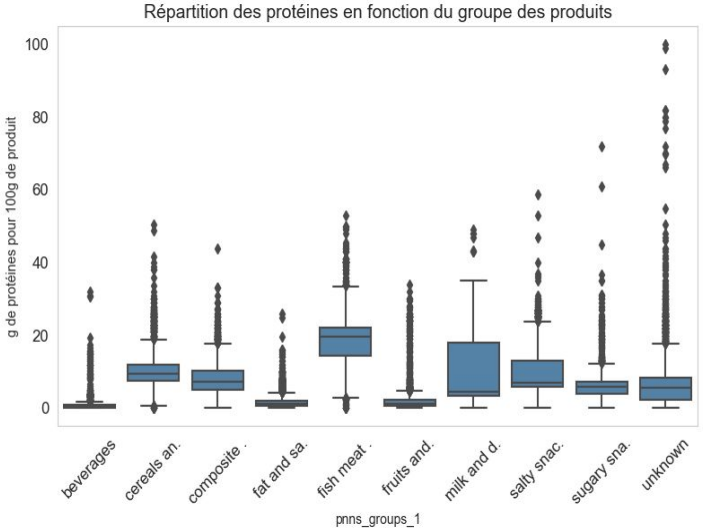




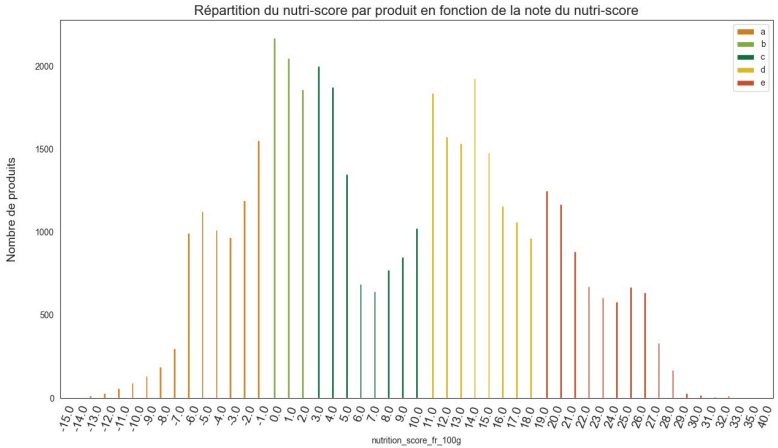
open FOOD facts

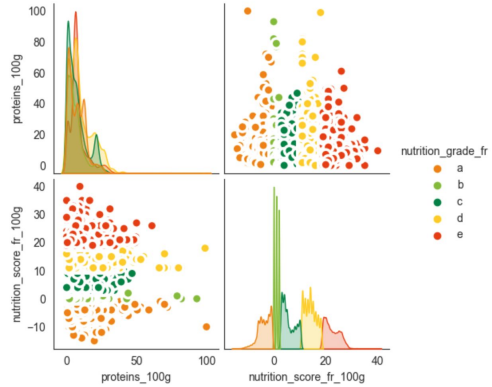
Analyse exploratoire des données : analyse bivariée

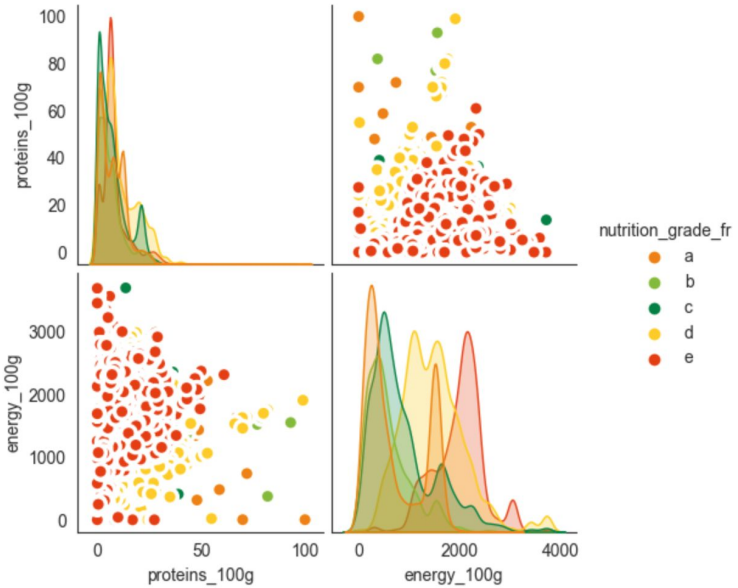




Analyse exploratoire des données : analyse bivariée

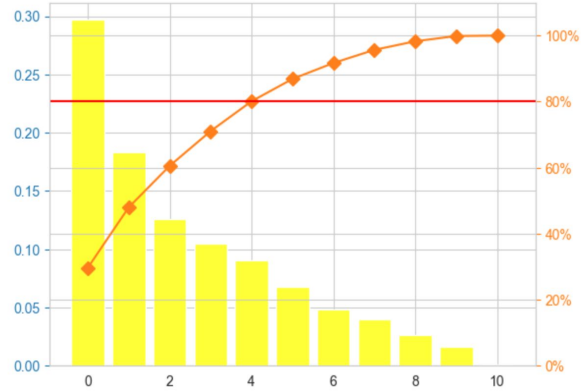




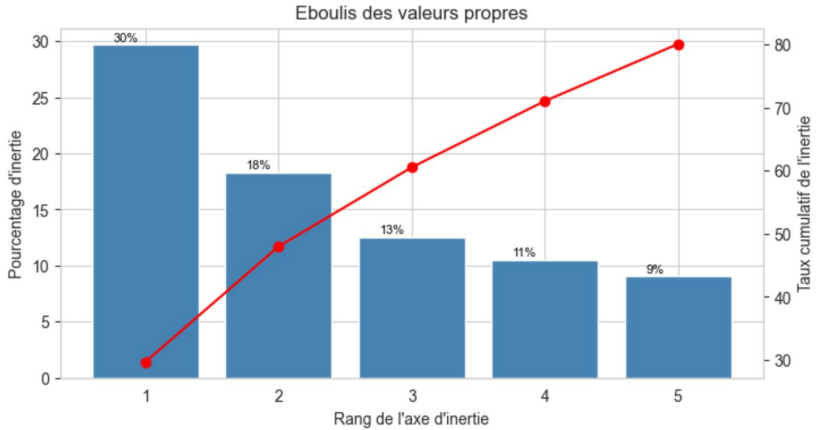


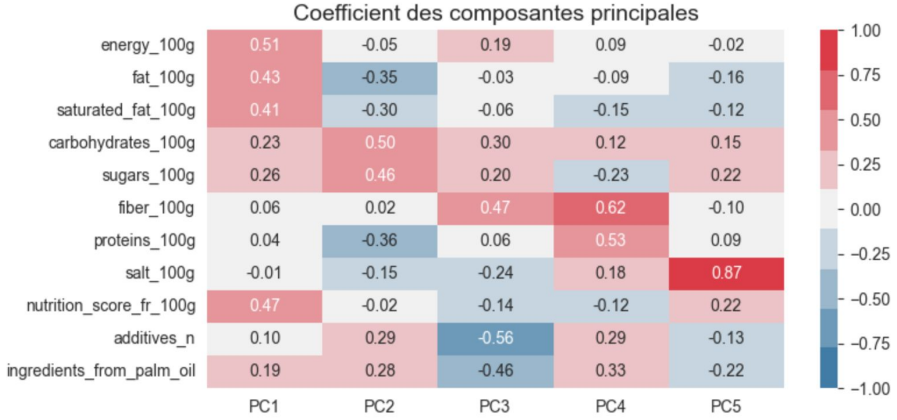
Analyse exploratoire des données : analyse multivariée, ACP

Santé publique France



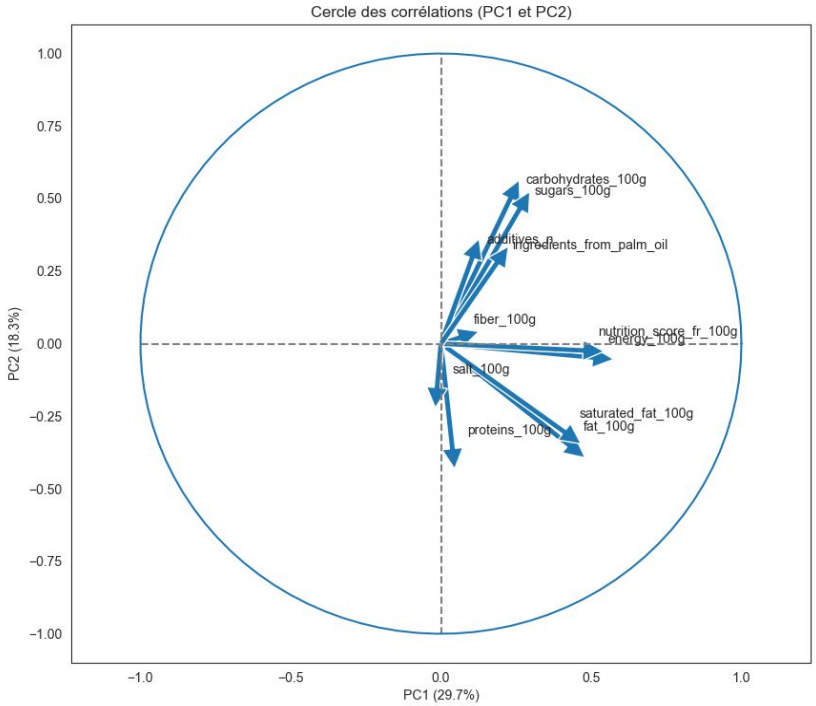
Critère de Kaiser

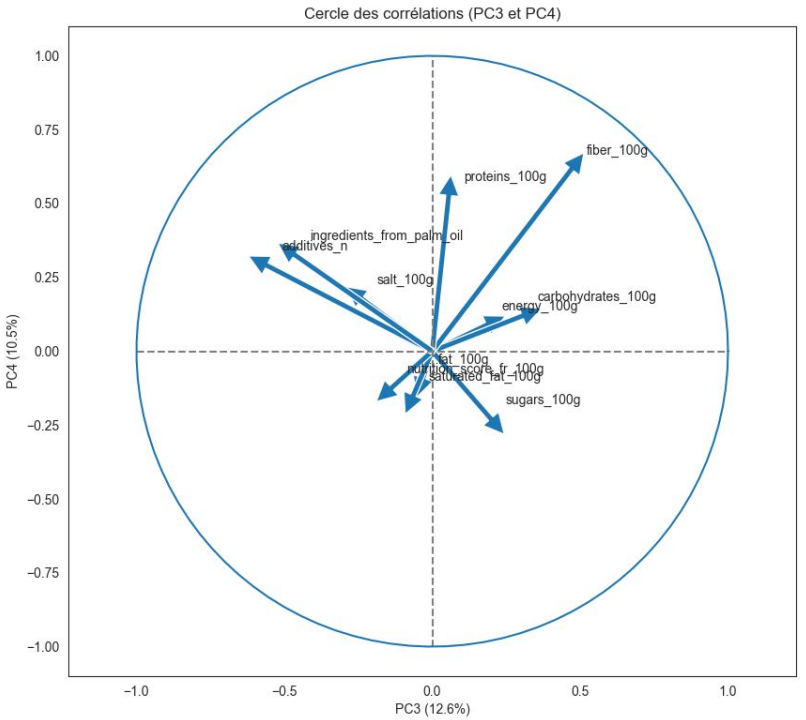




Analyse exploratoire des données : analyse multivariée, cercle de corrélations

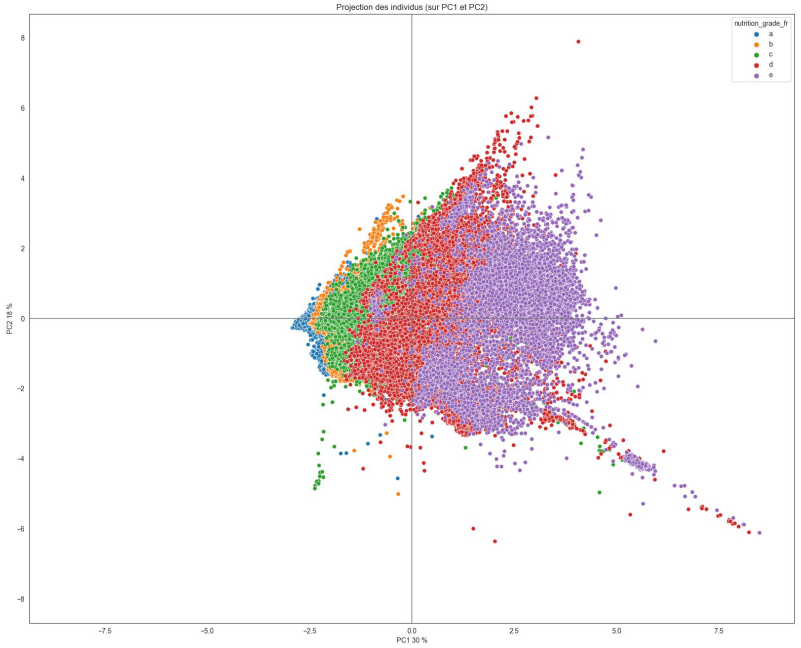
Santé publique France

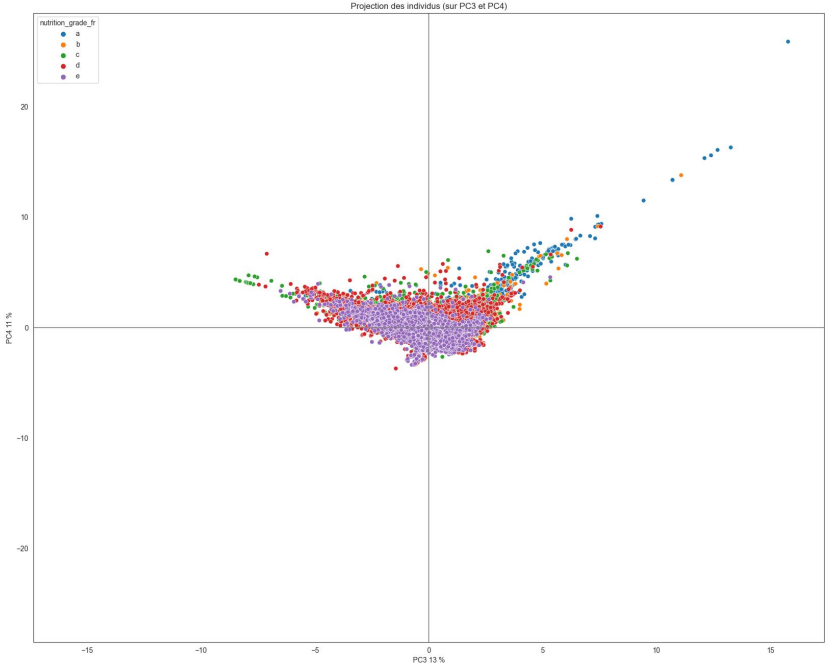




Analyse exploratoire des données : analyse multivariée, projections des individus

Santé publique France





Les grands principes du RGPD

Santé publique France

- **Le principe de finalité** : le responsable d'un fichier ne peut enregistrer et utiliser des informations sur des personnes physiques que dans un but bien précis, légal et légitime ;
- **Le principe de proportionnalité et de pertinence** : les informations enregistrées doivent être pertinentes et strictement nécessaires au regard de la finalité du fichier ;
- **Le principe d'une durée de conservation limitée** : il n'est pas possible de conserver des informations sur des personnes physiques dans un fichier pour une durée indéfinie. Une durée de conservation précise doit être fixée, en fonction du type d'information enregistrée et de la finalité du fichier ;
- **Le principe de sécurité et de confidentialité** : le responsable du fichier doit garantir la sécurité des informations qu'il détient. Il doit en particulier veiller à ce que seules les personnes autorisées aient accès à ces informations ;
- **Les droits des personnes.**

Conclusion

Le jeu de données contient toutes les données nécessaires à notre idée d'application de moteur de recommandations pour les utilisateurs.

- protéines
- nutri-score
- les glucides
- lipides
- sel
- additifs
- fibres

Suite du projet

- Produits vendus hors de France (vacances, voyages scolaires)
- Ajouter le groupe NOVA pour les produits peu-transformés
- Effectuer un modèle de Scoring pour la recommandation des produits
- Mettre en place un prototype d'application de recommandations