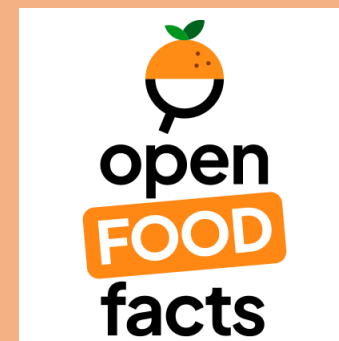




Concevez une application au service
de la santé publique

Recommandation Alimentaire sur la similarité des caractéristiques des produits

Source:



<https://world.openfoodfacts.org/>

Sommaire

- I - Objectif de l'application
- II - Nettoyage des données
- III - Analyse des données
- IV - Conception de l'application
- V - Conclusion

Sommaire

I - Objectif de l'application

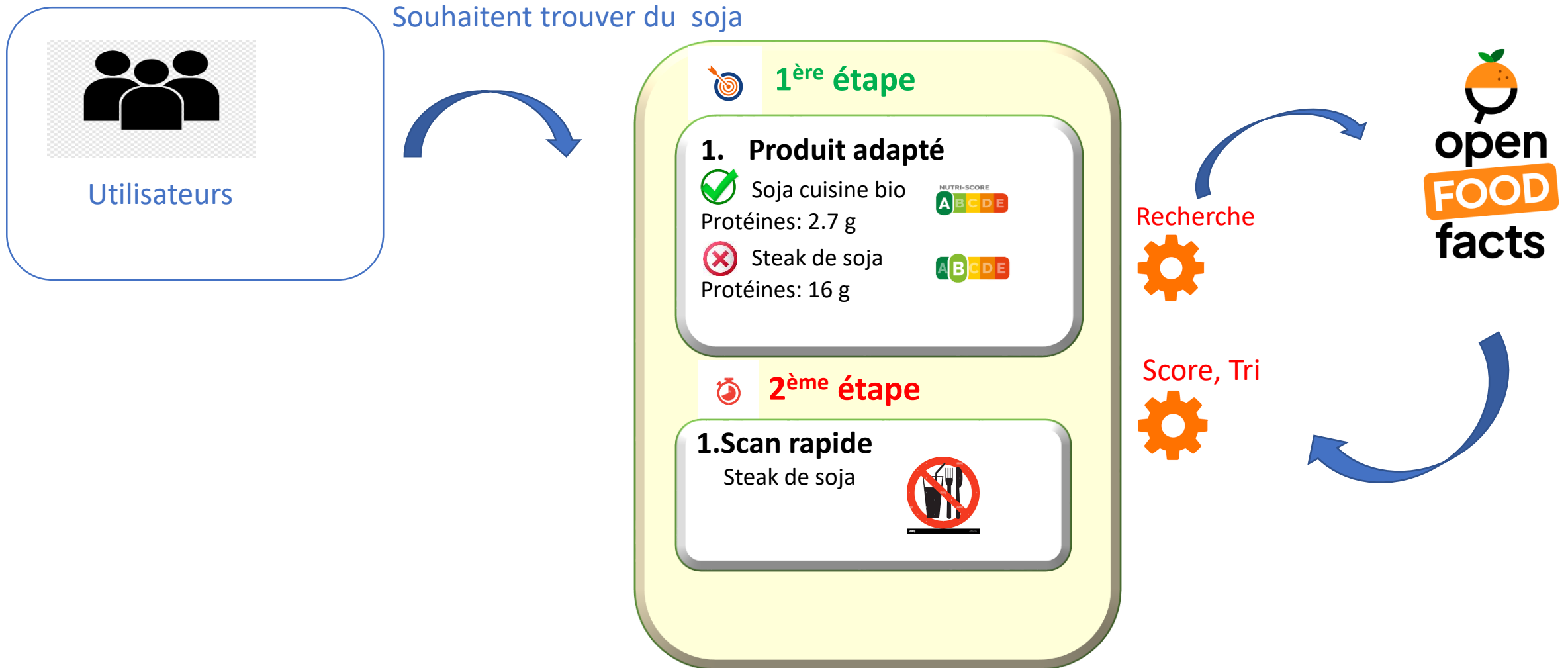
II - Nettoyage des données

III - Analyse des données

IV - Conception de l'application

V - Conclusion

I - Objectif de l'application



I - Objectif de l'application – Alimentation ?

Alimentation, Saine, et Équilibrée

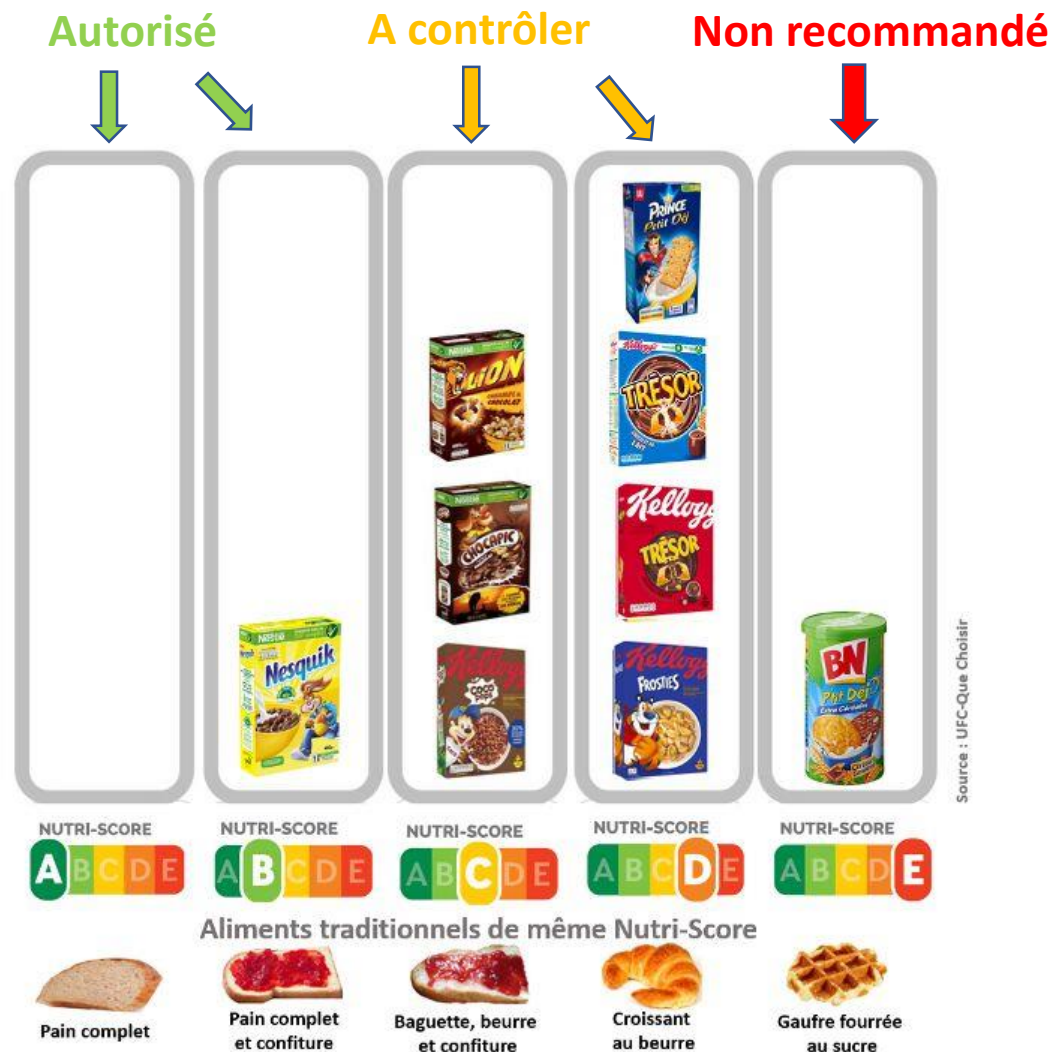


Plus favorable

Qualité nutritionnelle

Moins favorable

Classes	Bornes du score	Couleurs
A	Min à -1	Vert foncé
B	0 à 2	Vert clair
C	3 à 10	Orange clair
D	11 à 18	Orangé moyen
E	19 à Max	Orange foncé



I - Objectif de l'application – Informations retenues

Alimentation saine

- **Nutriscore faible**
- **Peu de sel**
- **Peu de sucre**
- **Peu calorique**

Alimentation équilibrée

- **Macro-nutriments**
 - **Lipides**
 - **Glucides**
- **Micro-nutriments**
 - **Vitamines**
 - **Minéraux**

Produits Disponibles

- **Nom du produit**
- **Photo**
- **Catégorie**
- **Vendus en France**

Sommaire

I - Objectif de l'application

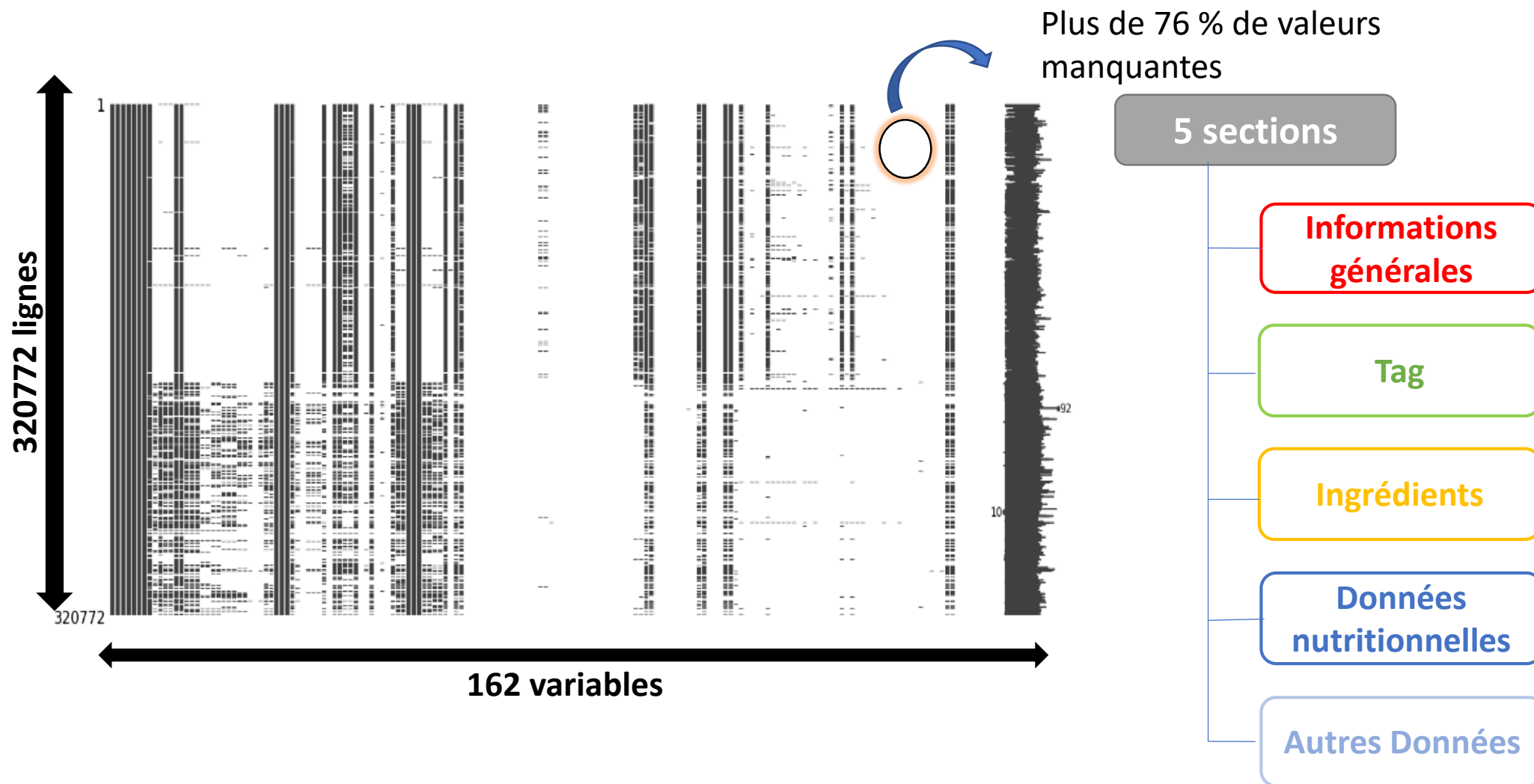
II - Nettoyage des données

III - Analyse des données

IV - Conception de l'application

V - Conclusion

Dataset



Nettoyage des données: Dataset – Variables Utiles

Alimentation saine

- nutrition_score_fr_100g
- nutrition_grade_fr
- energy_100g
- sugars_100g
- proteins_100g
- salt_100g

Alimentation équilibrée

- fat_100g
- carbohydrates_100g
- saturated-fat_100g
- fiber_100g

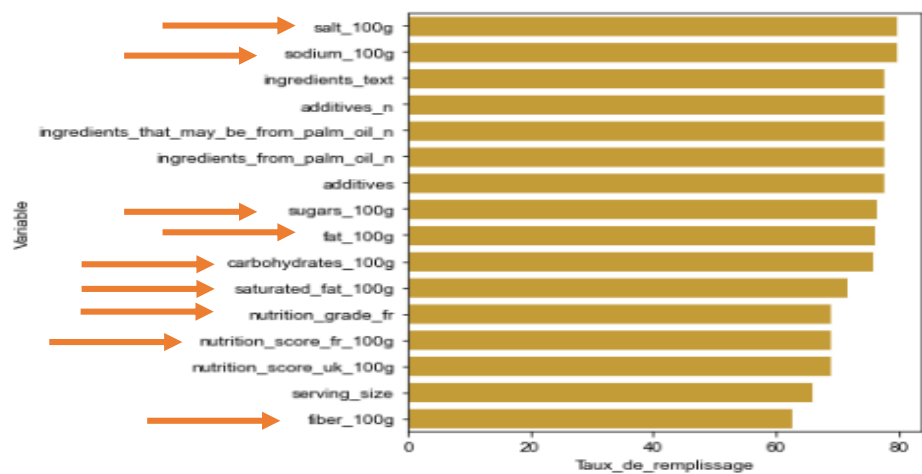
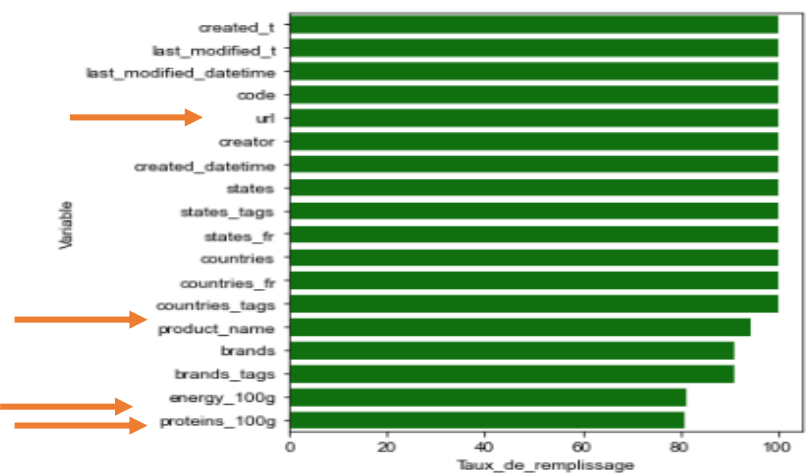
Produits disponibles

- url
- product_name
- image_small_url
- main_category_fr
- categories_tags

Légende:

- *en vert*: <= 20 % NaN
- *en bleu*: <= 40 % NaN
- *en rouge*: <= 80 % NaN

Nettoyage des données: Dataset – Variables Utiles



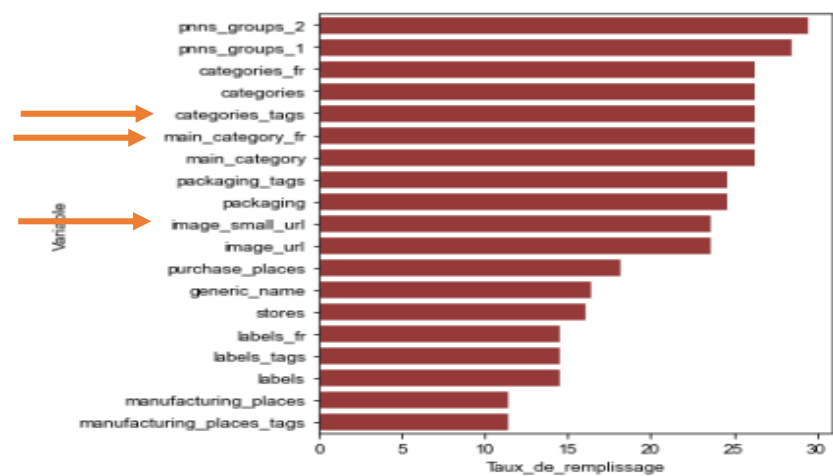
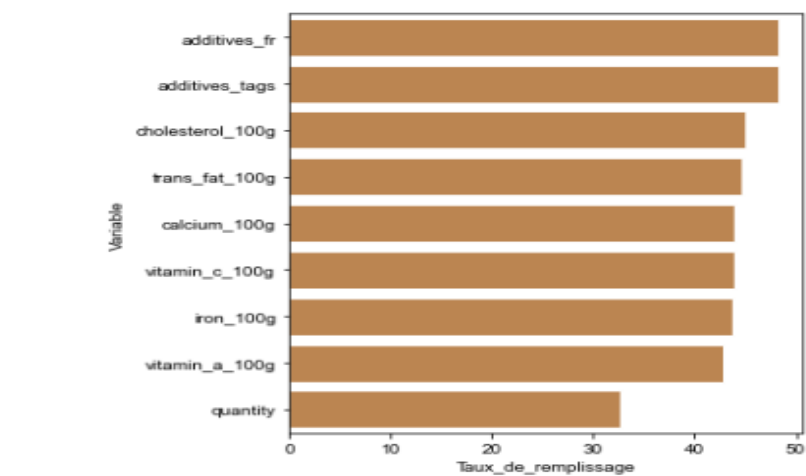
Alimentation saine



Alimentation équilibrée

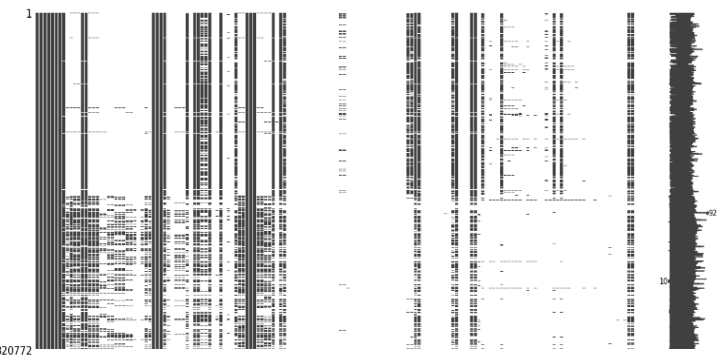


Produits disponibles

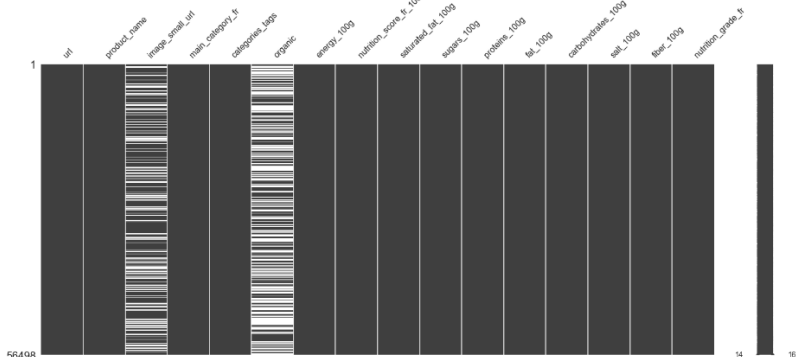


Nettoyage des données: Réduction du dataset

Données initiales: **320772 lignes, 162 variables**

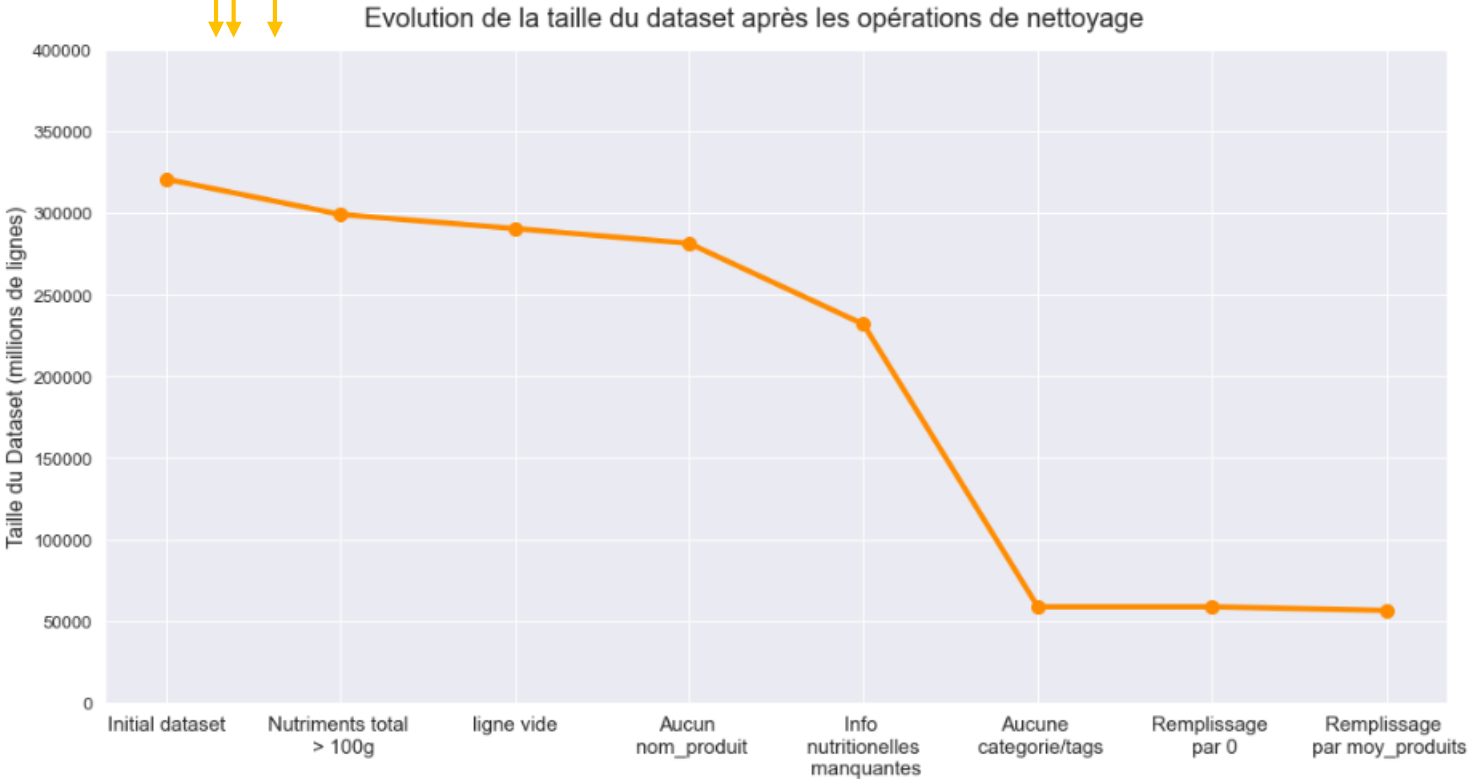


Données finales: **58812 lignes, 16 variables**



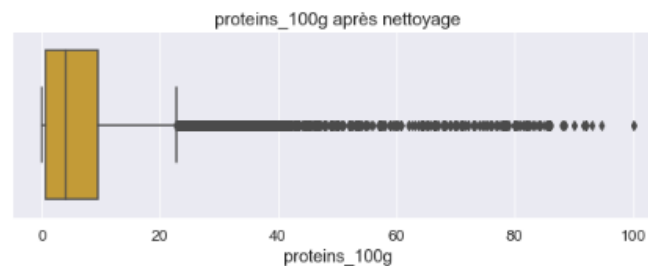
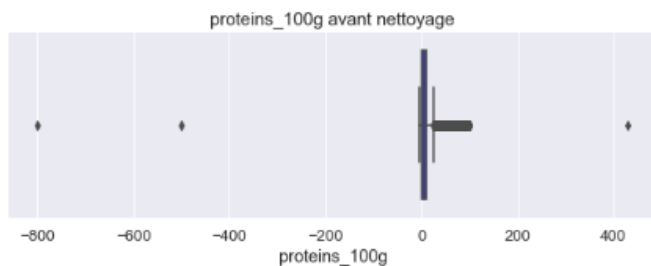
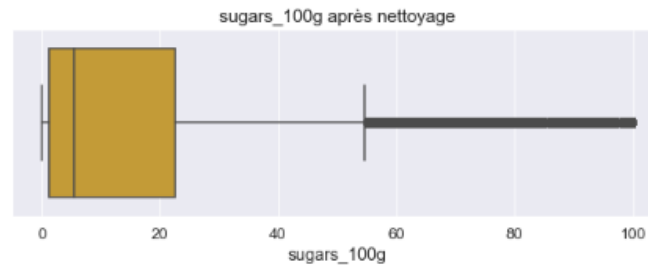
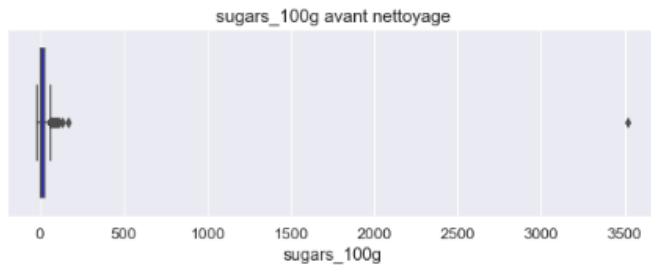
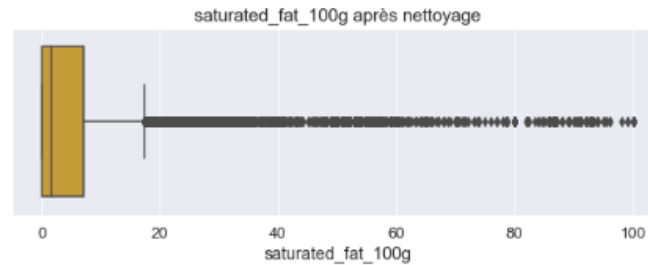
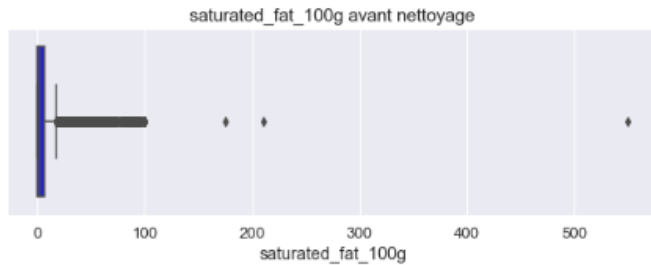
Nb Lignes	Nb. Var
320772	146
320669	146
320669	146

- Suppression des 16 variables vides
- Formatage des ' - ' en ' _ ' dans le nom des variables
- Suppression des 2 doublons sur le code produit
- Sélection des variables pertinentes



Nettoyage des données: Valeurs aberrantes

Evolution des données nutritionnelles avant et après nettoyage



Outliers pour 100g

saturated_fat_100g > 100g

sugars_100g > 100 g

proteins_100g > 100g

fat_100g > 100g

carbohydrates_100g > 100g

salt_100g > 100g

fiber_100g > 100g

energy_100g > 100g

nutrition_score_fr_100g > 100g

Sommaire

I - Objectif de l'application

II - Nettoyage des données

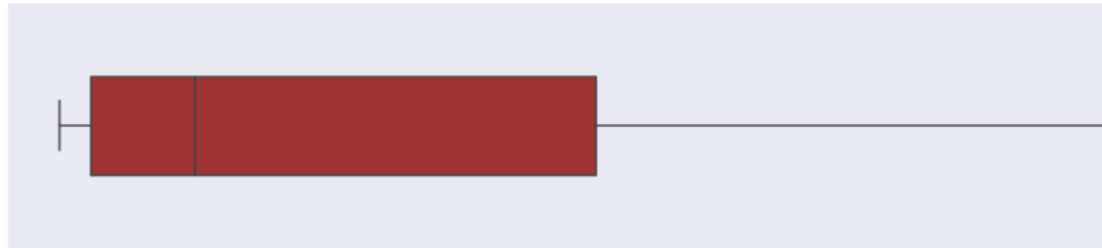
III - Analyse des données

IV - Conception de l'application

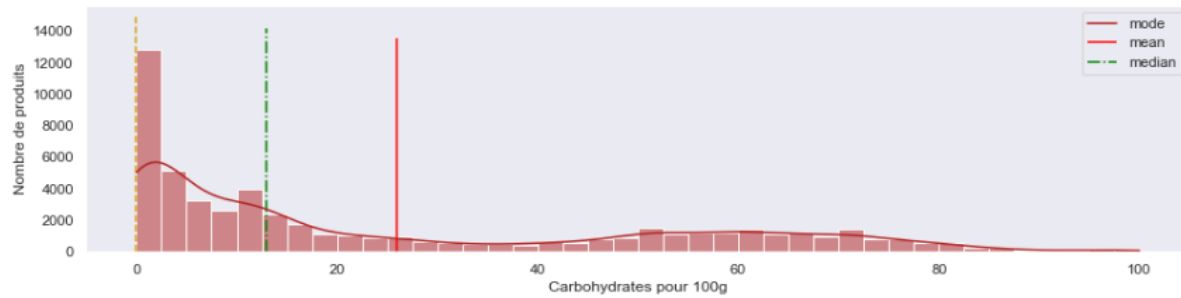
V - Conclusion

Analyse univariée: les nutriments

carbohydrates_100g



carbohydrates_100g



3

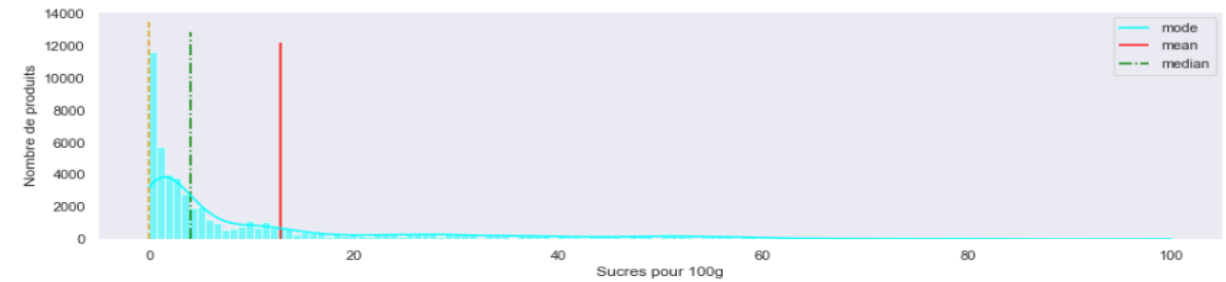
6

Variable name	sugars_100g	carbohydrates_100g
Mean	12.842952	25.94162
Median	4.0	12.9
Skew	1.947357	0.786519
Kurtosis	3.497939	-0.811039
Variance	339.713675	729.821426
Stdev	18.431323	27.015207

sugars_100g



sugars_100g



Analyse univariée: les nutriments

Gras saturés pour 100g

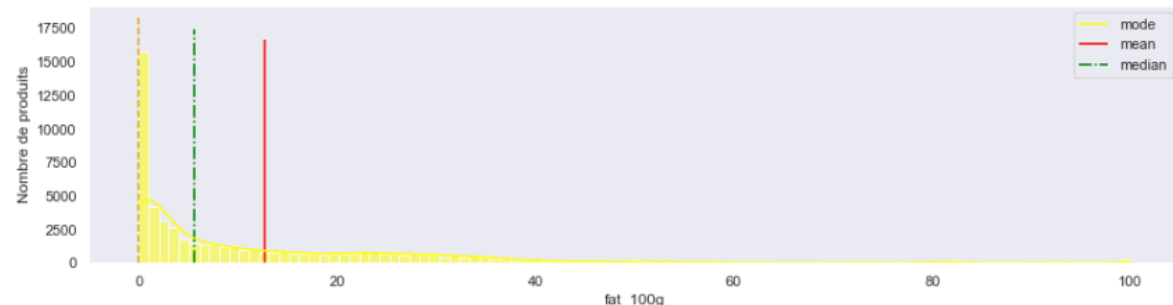
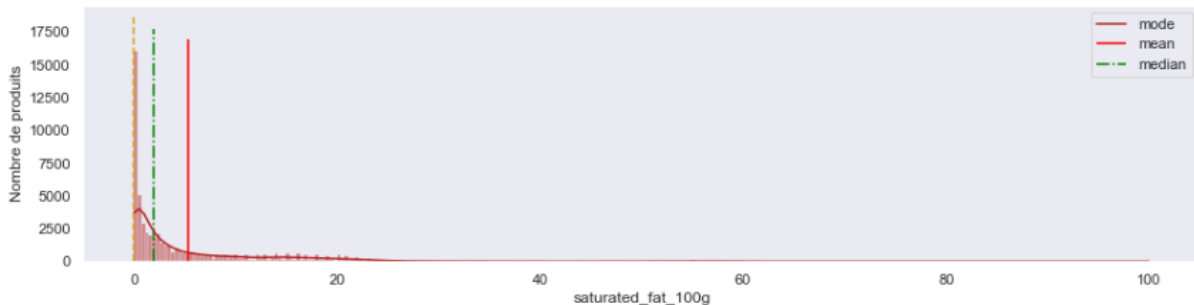


saturated_fat_100g

Gras pour 100g



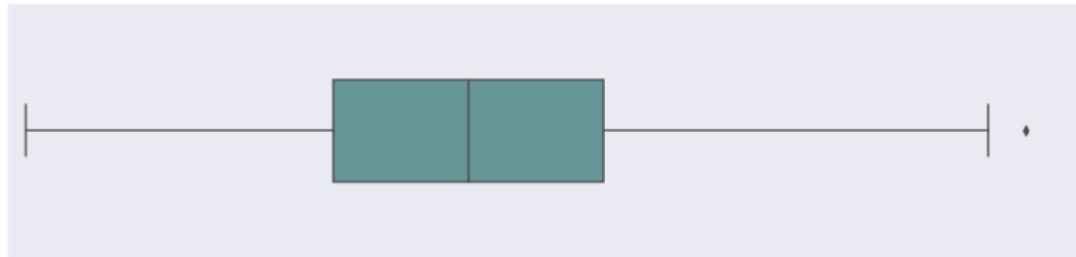
fat_100g



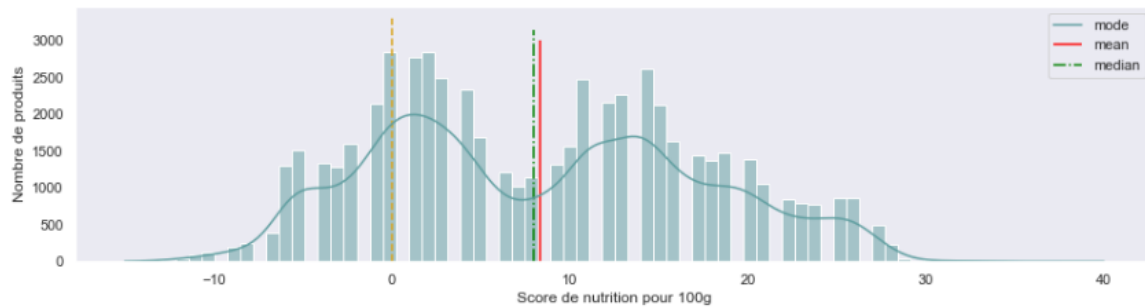
	2	5
Variable name	saturated_fat_100g	fat_100g
Mean	5.291925	12.636447
Median	1.87	5.6
Skew	3.263248	2.234466
Kurtosis	17.630671	6.497715
Variance	67.61702	281.457428
Stdev	8.222957	16.776693

Analyse univariée: distribution des produits

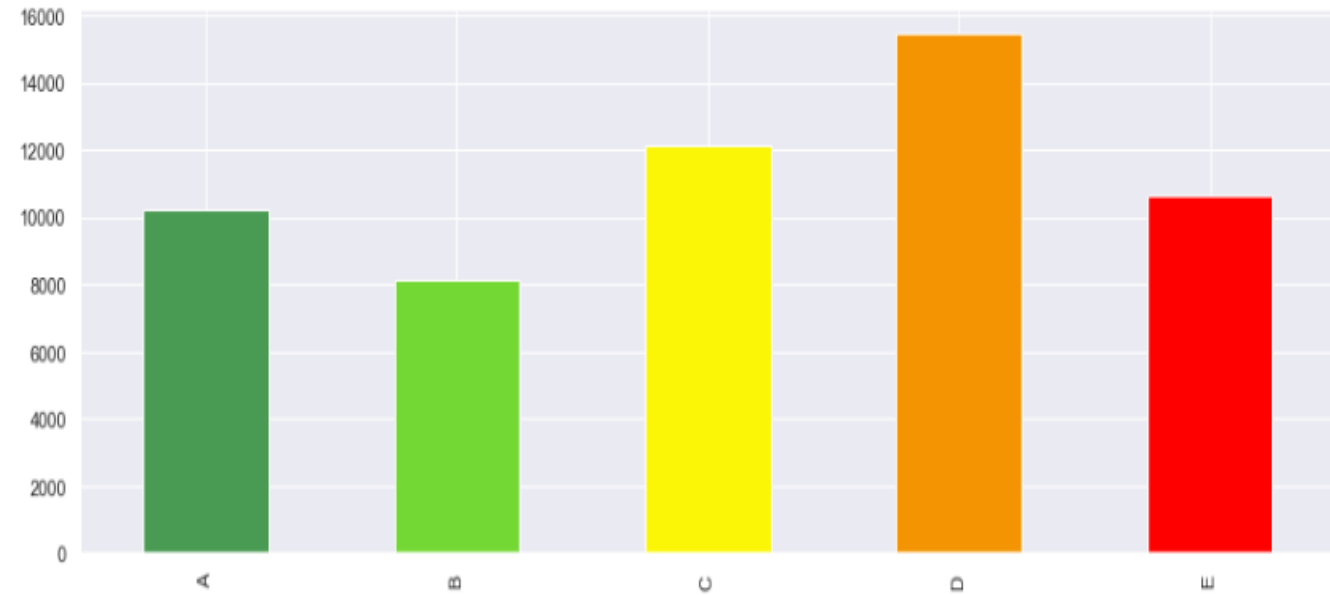
nutrition_score_fr_100g



nutrition_score_fr_100g

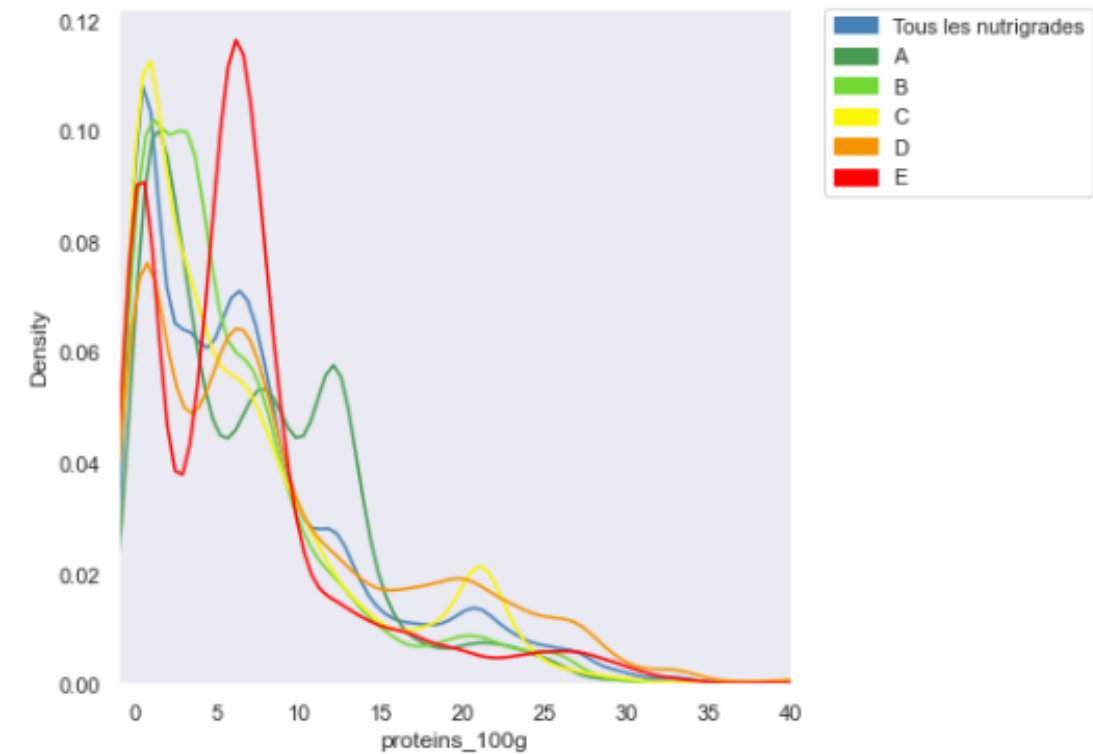


Distribution du score de nutrition

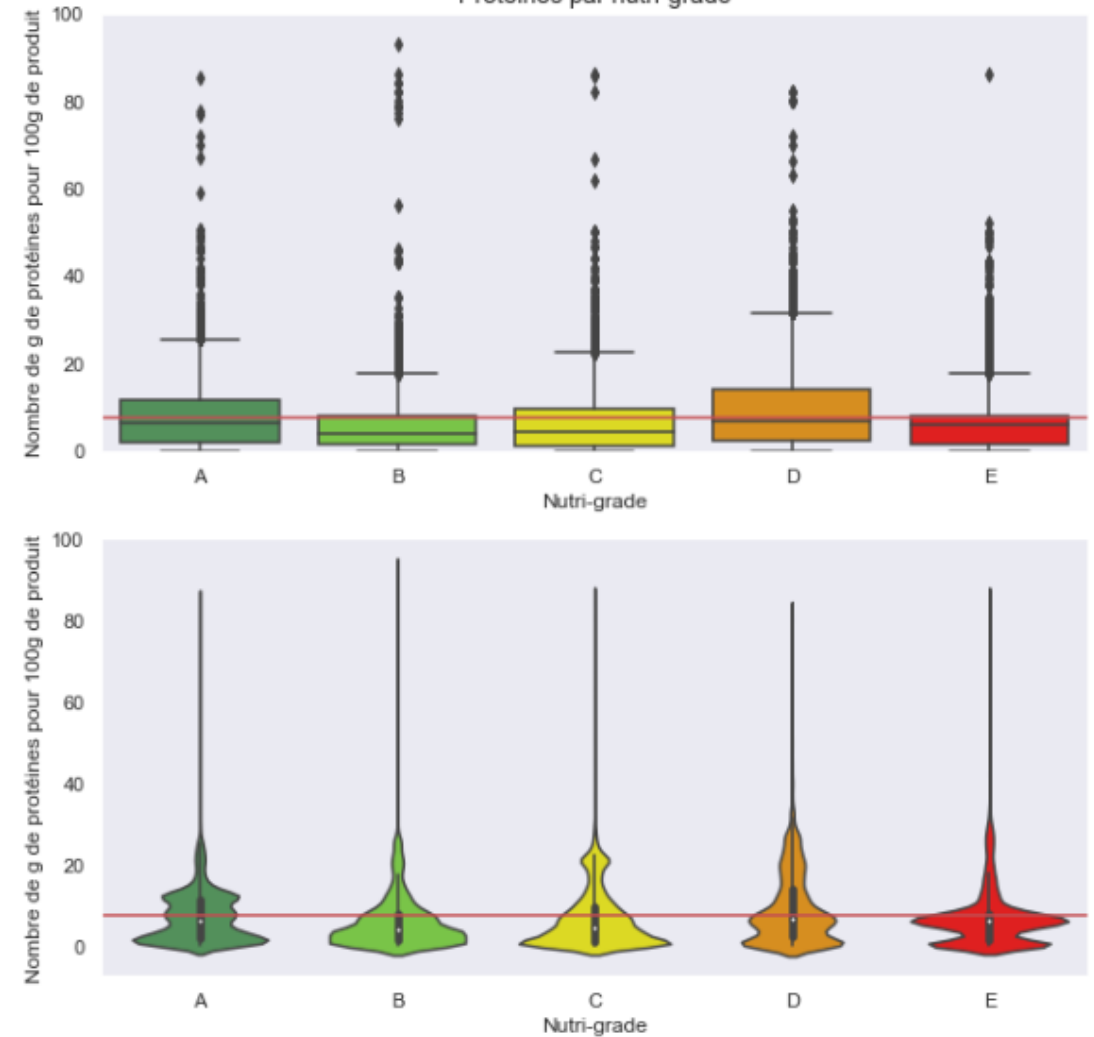


Analyse bivariable: les protéines

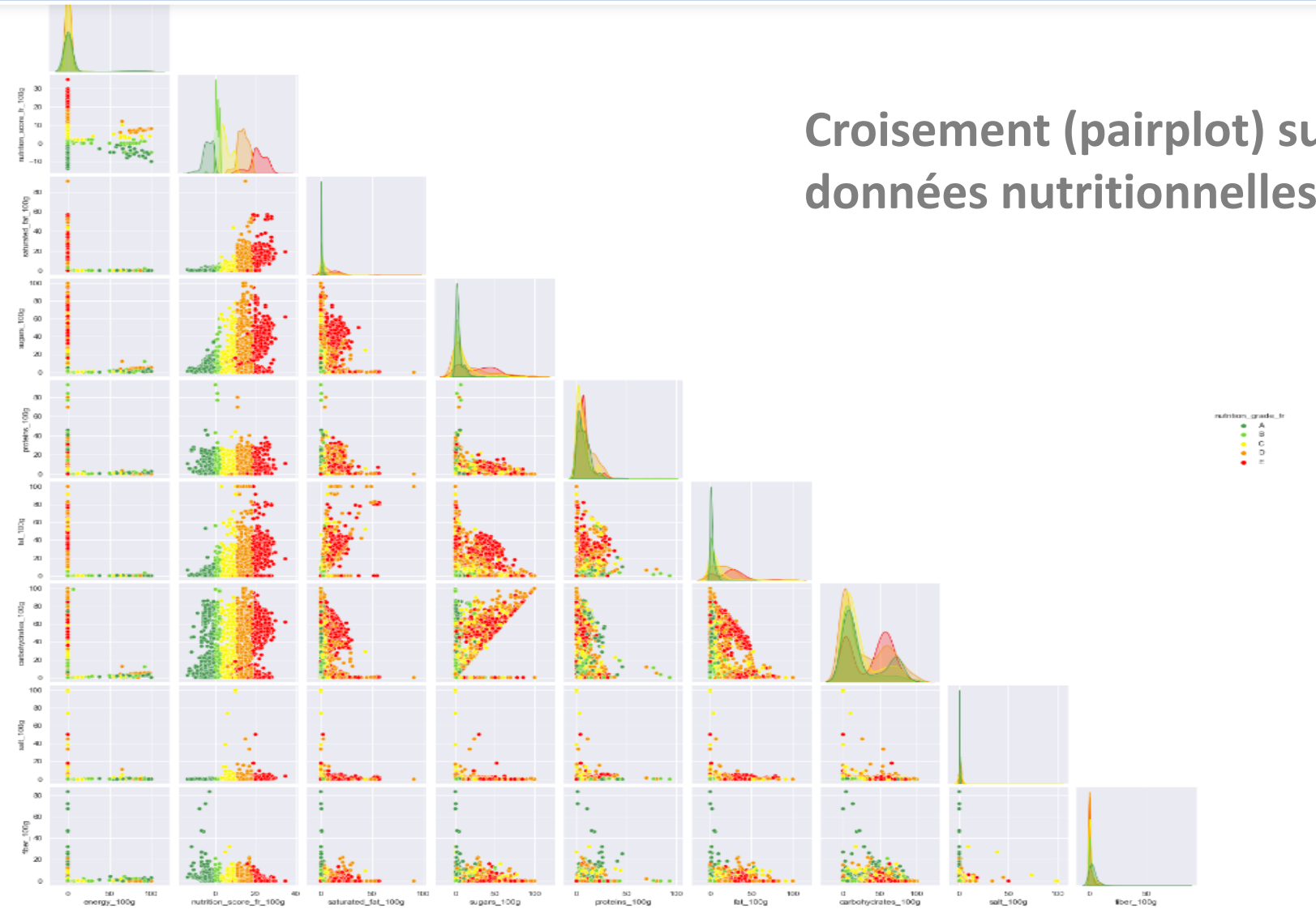
Distribution des protéines



Protéines par nutri-grade



Analyse multivariée: les données nutritionnelles



Analyse multivariée: ANOVA

3 hypothèses d'application du test d'analyse de variance: indépendance, normalité, homoscédasticité

Rejet sur l'hypothèse de normalité de la distribution des protéines, sur la variable catégorique des nutriscores

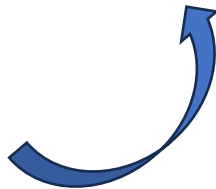


Test de similarité des variances (Test de Levene): peu sensible à la non-normalité

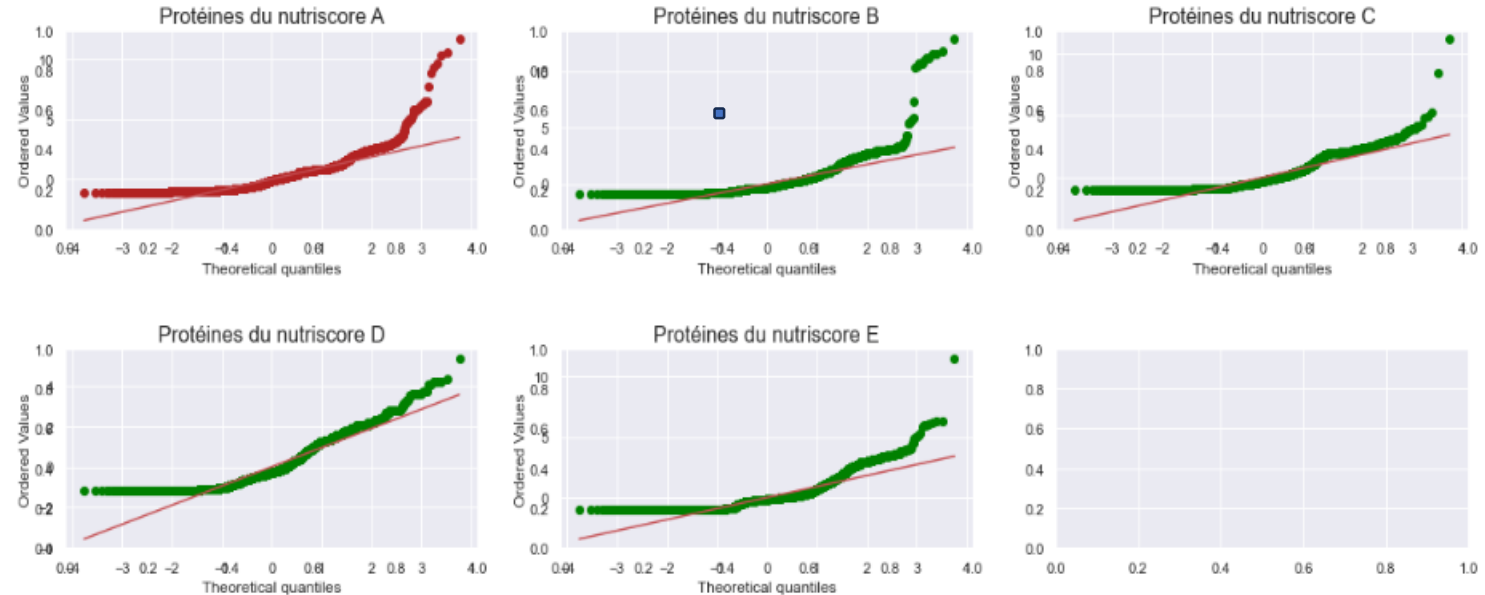


Hétéroscédasticité de la variance des protéines sur les catégories de nutriscores

Test de Kruskal – Wallis (non paramétrique): ANOVA à un facteur



Q-Q plots des produits protéiniques en fonction du nutriscore avec la distribution Gaussienne

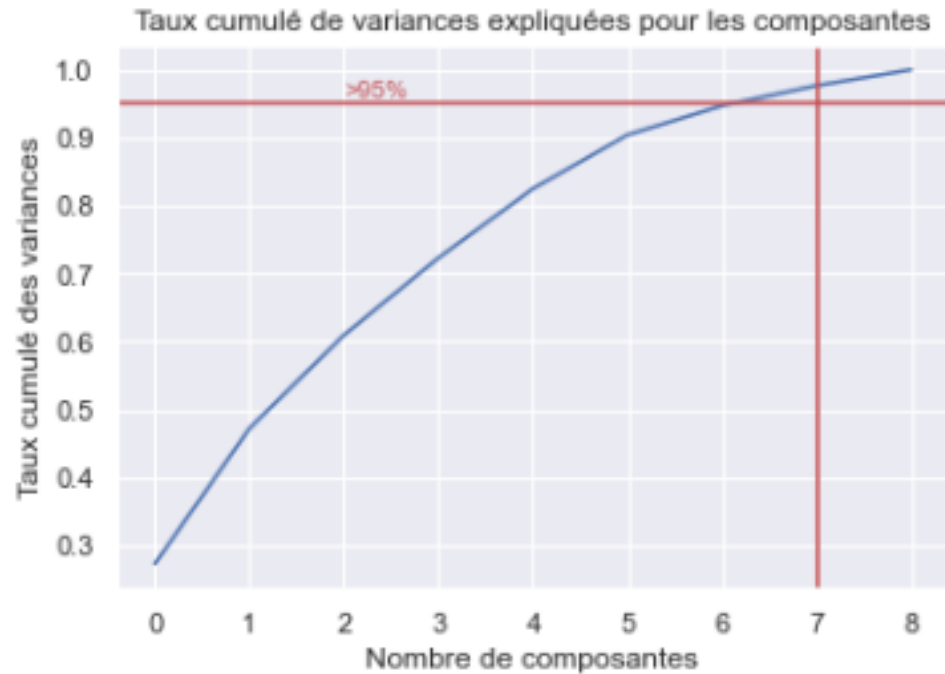


**H = 918,21, et p=0:
Différence médiane/moyenne de la distribution représentative.**

Cependant, du fait de la non-connaissance de la zone de différence entre les catégories de nutriscore:

Accepte (H0): Aucune différence significative des moyennes des protéines sur les différentes catégories de nutriscores

Analyse multivariée: Analyse en Composantes Principales(ACP)



Contribution expliquée à plus de 95 % sur les 7 premières composantes principales

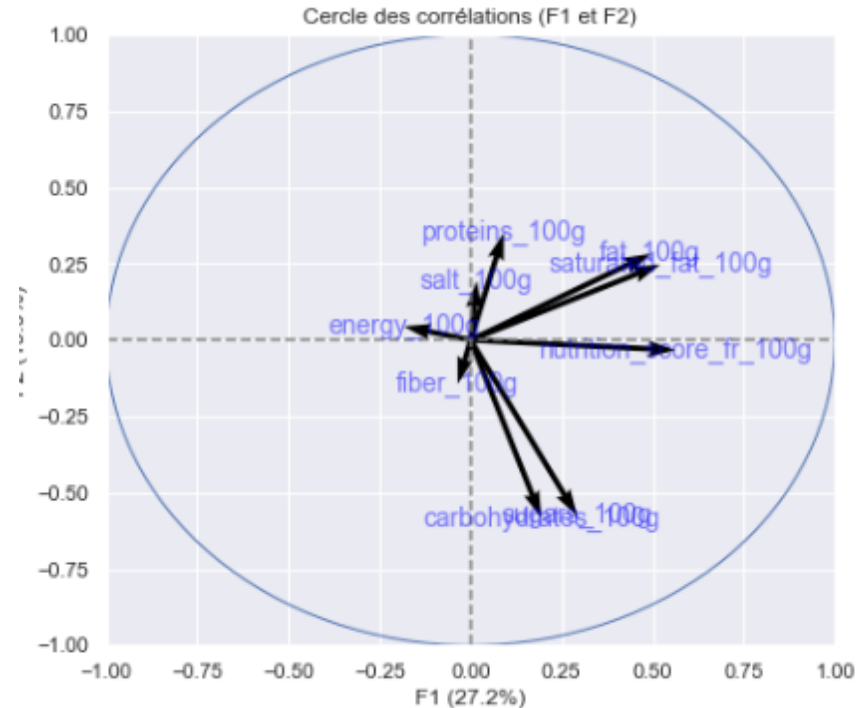


Faible intérêt d'une recherche de réductions de dimensions

Cercle des corrélations



Produits peu gras et
peu caloriques



Sommaire

I - Objectif de l'application

II - Nettoyage des données

III - Analyse des données

IV - Conception de l'application

V - Conclusion

Conception de l'application

Alimentation saine

- nutrition_score_fr_100g
- nutrition_grade_fr
- energy_100g
- sugars_100g
- proteins_100g
- salt_100g

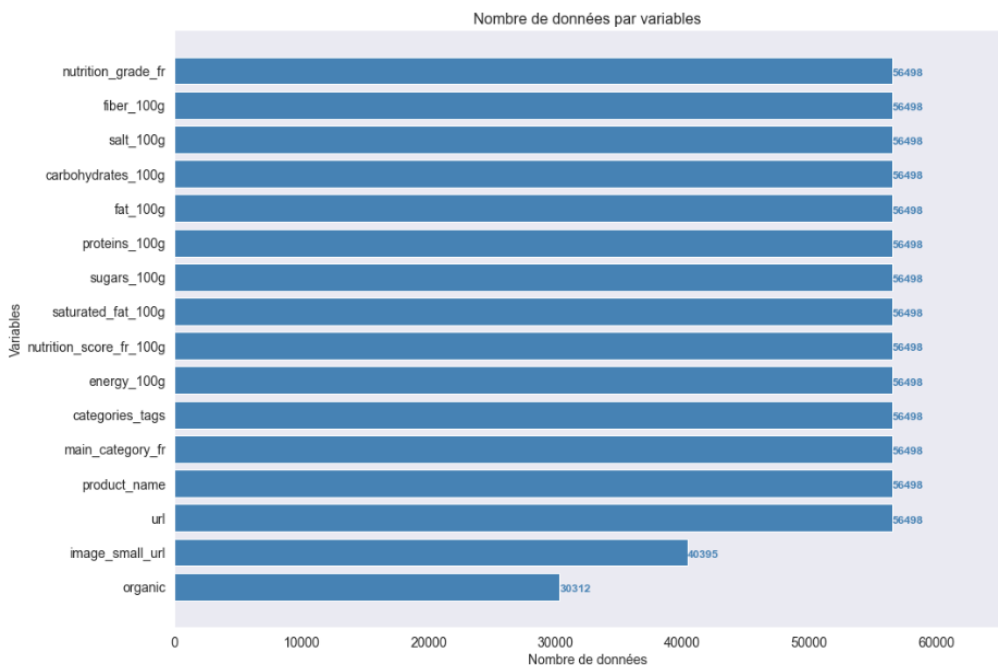
Alimentation équilibrée

- fat_100g
- carbohydrates_100g
- saturated-fat_100g
- fiber_100g

Produits disponibles

- url
- product_name
- image_small_url
- main_category_fr
- categories_tags

Le nettoyage des données a permis d'exploiter les variables d'intérêts, avec le plus d'information possible et de bonne qualité.



Moteur de recommandation réalisable

Variables de scoring



Moteur de recommandation: Score

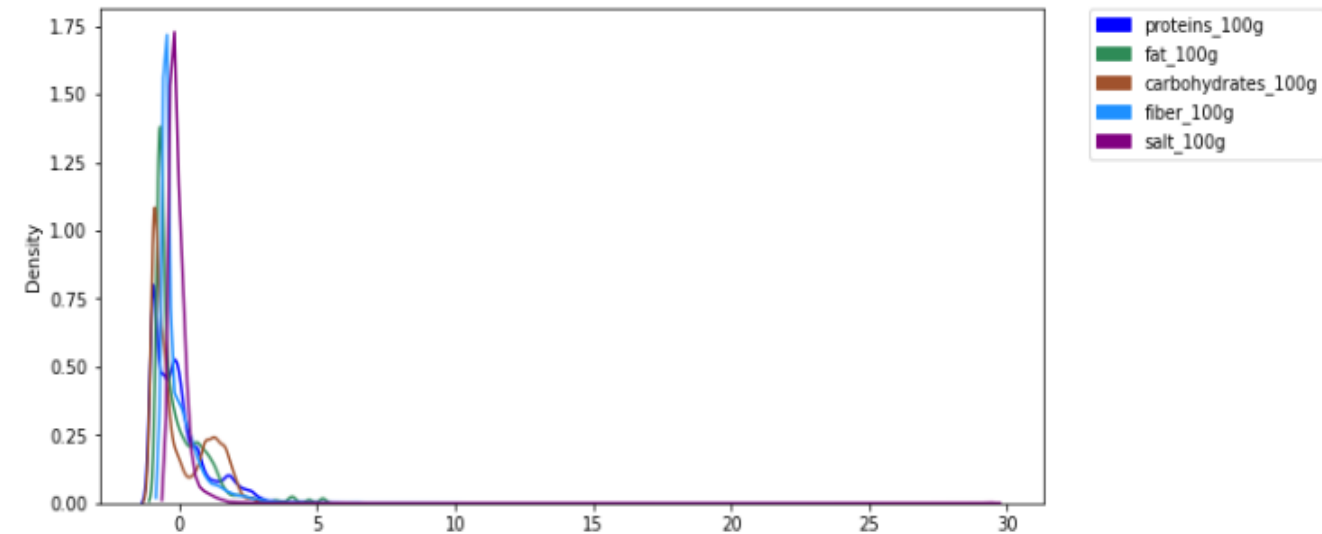
Sélection des variables de scoring: proteins_100g, fat_100g, carbohydrates_100g, fiber_100g, salt_100g



Mise à l'échelle: StandardScaler()



Scoring de pondération: Protéines: 50, gras: 2, carbohydrates: 3, fibres: 5, sel: 7



Moteur de recommandation

Pré-traitement des noms de produit:

Pré-processing: clean() de textthero

Suppression de stopwords français: NLTK



Vectorisation:

TfidfVectorizer de scikit-learn



Dataset de comparaison:

Renommage des variables, ajout de variables, ajout de variables L_G_Su_Se



Moteur de recommandation:


















Vérification de la saisie, suppression des ponctuations, suppression des stopwords, similarité des cosinus par KNN (distance des caractéristiques nutritionnelles des produits), tri de la similarité des recommandations des n produits

```
get_reco_by_features('Soja',15)
```

Produit sélectionné: Soja Index: 27720

Recherche de recommandations.....

Remarque : L_G_Su_Se : lipides - Glucides - Sucre - Sel

	Produit	Photo	protéines/100g prod	Note_nutri_score	Nutri_score	g de matières grasses/100g prod	carbohydrates/100g prod	g de sucres/100g prod	g de sel/100g prod	L_G_Su_Se	Score
30208	Spiruline		67.0		-1.0	6.80	14.0	0.001	0.0000	6.8-14.0-0.001-0.0	36.538814
27720	Soja		32.0		-3.0	16.20	36.2	0.000	0.0000	16.2-36.2-0.0-0.0	19.338679
40086	Knäckebrot Backmischung		16.0		-3.0	21.00	42.0	0.500	0.0400	21.0-42.0-0.5-0.04	10.835711
1334	Casino filet de merlan pané		13.6		-2.0	8.80	20.0	0.880	0.3450	8.8-20.0-0.88-0.345	8.168124
37011	4 Tranches Panées de Cabillaud		13.0		-1.0	6.70	16.0	0.500	0.5500	6.7-16.0-0.5-0.55	7.674240
37013	100 % filet cabillaud		13.0		-1.0	6.50	16.0	0.300	0.5080	6.5-16.0-0.3-0.508	7.667300
30465	Petits poissons panés		11.9		-2.0	8.30	20.1	0.600	0.6000	8.3-20.1-0.6-0.6	7.270549
37015	Colin Lieu ou d'Alaska, Surgelés		12.0		-1.0	6.50	16.0	0.300	0.6350	6.5-16.0-0.3-0.635	7.139133
30464	Filets de Colin d'Alaska Panés		11.6		-2.0	8.90	20.6	0.800	0.4600	8.9-20.6-0.8-0.46	7.120918
1165	Fish Sticks		11.4		-1.0	7.89	17.5	0.000	0.5350	7.89-17.5-0.0-0.535	6.892813

Sommaire

I - Objectif de l'application

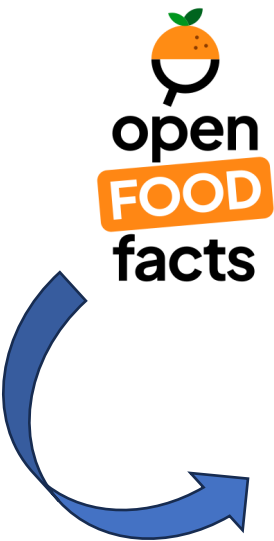
II - Nettoyage des données

III - Analyse des données

IV - Conception de l'application

V - Conclusion

Conclusion



Données
imputées,
et
analysées

Santé

get_rec_by_features('Sole',15)												
Produit sélectionné: Sole Index: 27720												
Historique de recommandations.....												
Recommander : L_s_sole, Se : Lipides - Glucides - Sucre - Sel												
	Produit	Photo	protéines%kg	protéines%kg	protéines%kg	protéines%kg	protéines%kg	protéines%kg	protéines%kg	protéines%kg	protéines%kg	Score
30205	Sardines		27.0	27.0	-1.0	6.00	14.0	0.001	0.0000	0.0010	0.0010	36.53824
27720	Sole		22.0	22.0	-1.0	16.20	36.2	0.000	0.0000	16.2362	0.0010	19.23629
40065	Kristallrot Backmischung		10.0	10.0	-1.0	21.00	42.0	0.000	0.0000	21.0420	0.0010	10.03071
1234	Casino fil de saumon fumé		12.0	12.0	-1.0	0.00	20.0	0.000	0.0000	0.0000	0.0010	0.10012
32011	4 Tranches Pâtée de Caille		12.0	12.0	-1.0	6.70	16.0	0.000	0.0000	6.7160	0.0010	7.07426
27012	100 % Sel		10.0	10.0	-1.0	6.00	16.0	0.000	0.0000	0.0010	0.0010	7.06730
34465	Petits pains		11.0	11.0	-1.0	6.00	20.0	0.000	0.0000	0.0010	0.0010	7.27048
27015	Cake Lait au Chocolat		12.0	12.0	-1.0	6.00	16.0	0.000	0.0000	0.0010	0.0010	7.13913
34464	Fils de Coton d'Alaska		11.0	11.0	-1.0	6.00	20.0	0.000	0.0000	0.0010	0.0010	7.13913
1165	Fils de Coton		11.0	11.0	-1.0	7.00	17.0	0.000	0.0000	0.0010	0.0010	6.00000

Basées sur les caractéristiques
nutritionnelles: protéines, sucres,
matières grasses, carbohydrates,
nutri-score

Limites

Pondération des caractéristiques nutritionnelles
Moteur de recommandation: fiabilité de la similarité sur l'ensemble des produits

Amélioration

Ajout des liens photo manquants
Produits vendus hors de France

Prolongement

Scan pour l'identification des produits, afin de détecter les produits interdits, ou non consommables pour les mineures.
Bilan protéinique journalier/mensuel, calcul des parts