

Projet 3 : Concevez une application au service de la santé publique





Sommaire

1. Idée d'application
2. Nettoyage des données
3. Analyse exploratoire
4. Conclusion



Idée d'application

Application pour le choix de meilleurs produits:

- Pour chaque produit scanné, calcul d'un nutriscore et d'une recommandation (pondération entre local et bio).
- L'application propose un produit appartenant à la même catégorie mais ayant un meilleur nutriscore et Bio si possible.

Idée d'application

Exemple de proposition :

Produit Scanné:



Proposition:



Nettoyage des données

Suppression des variables à 100 % de valeurs manquantes

Choix des variables :

```
listeVariableGardées = ['product_name',  
                        'countries',  
                        'categories',  
                        'additives_n',  
                        'nutriscore_grade',  
                        'energy_100g',  
                        'energy-kj_100g',  
                        'energy-kcal_100g',  
                        'proteins_100g',  
                        'carbohydrates_100g',  
                        'sugars_100g',  
                        'fat_100g',  
                        'saturated-fat_100g',  
                        'fiber_100g',  
                        'sodium_100g',  
                        'salt_100g',  
                        'nutrition-score-fr_100g']
```

Nettoyage des données

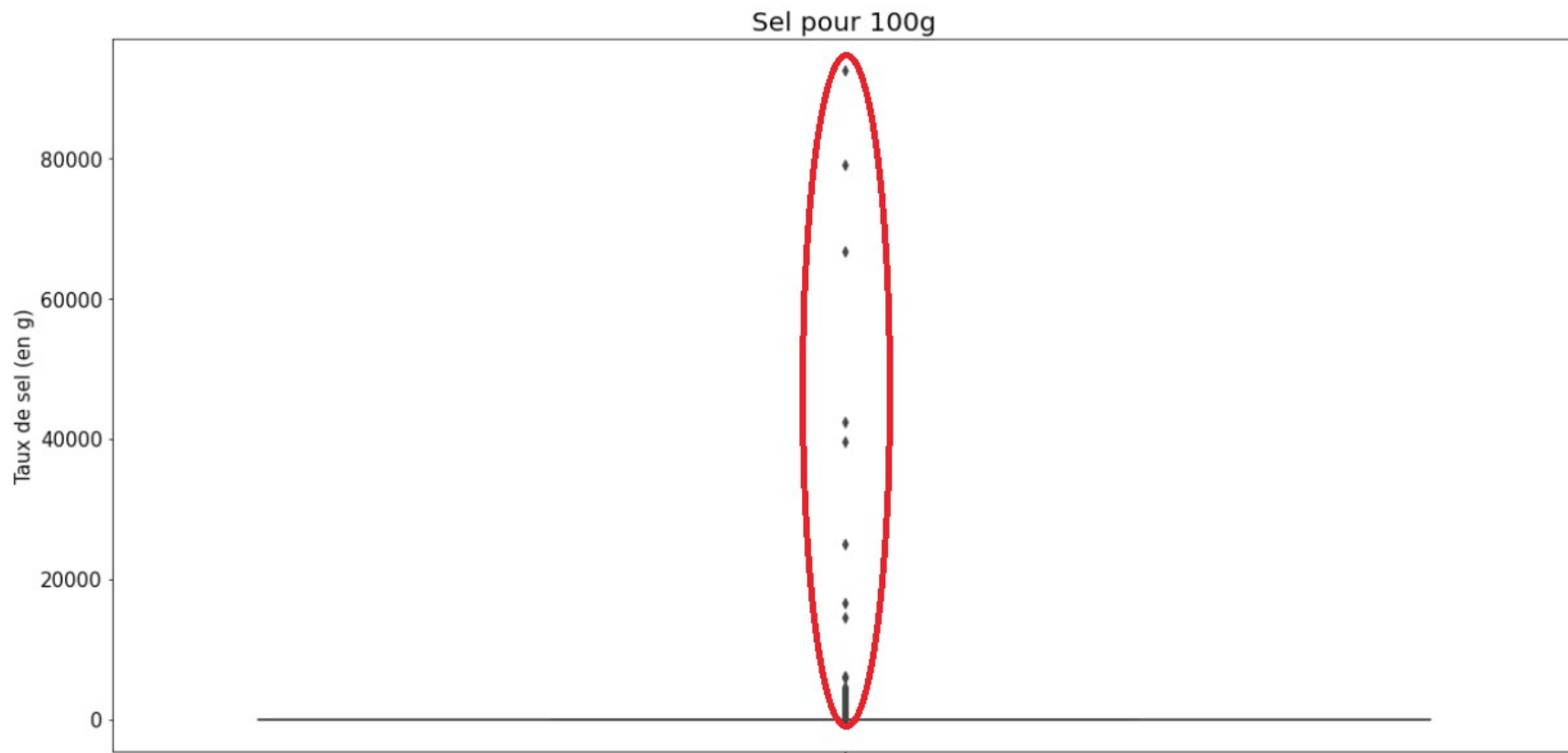
Remplacement des valeurs aberrantes et valeurs manquantes :

Exemple avec le sel (variable salt_100g):

```
salt_100g :  
  0.25      0.07  
  0.50      0.56  
  0.75      1.40  
  1.00  92500.00  
Name: salt_100g, dtype: float64
```

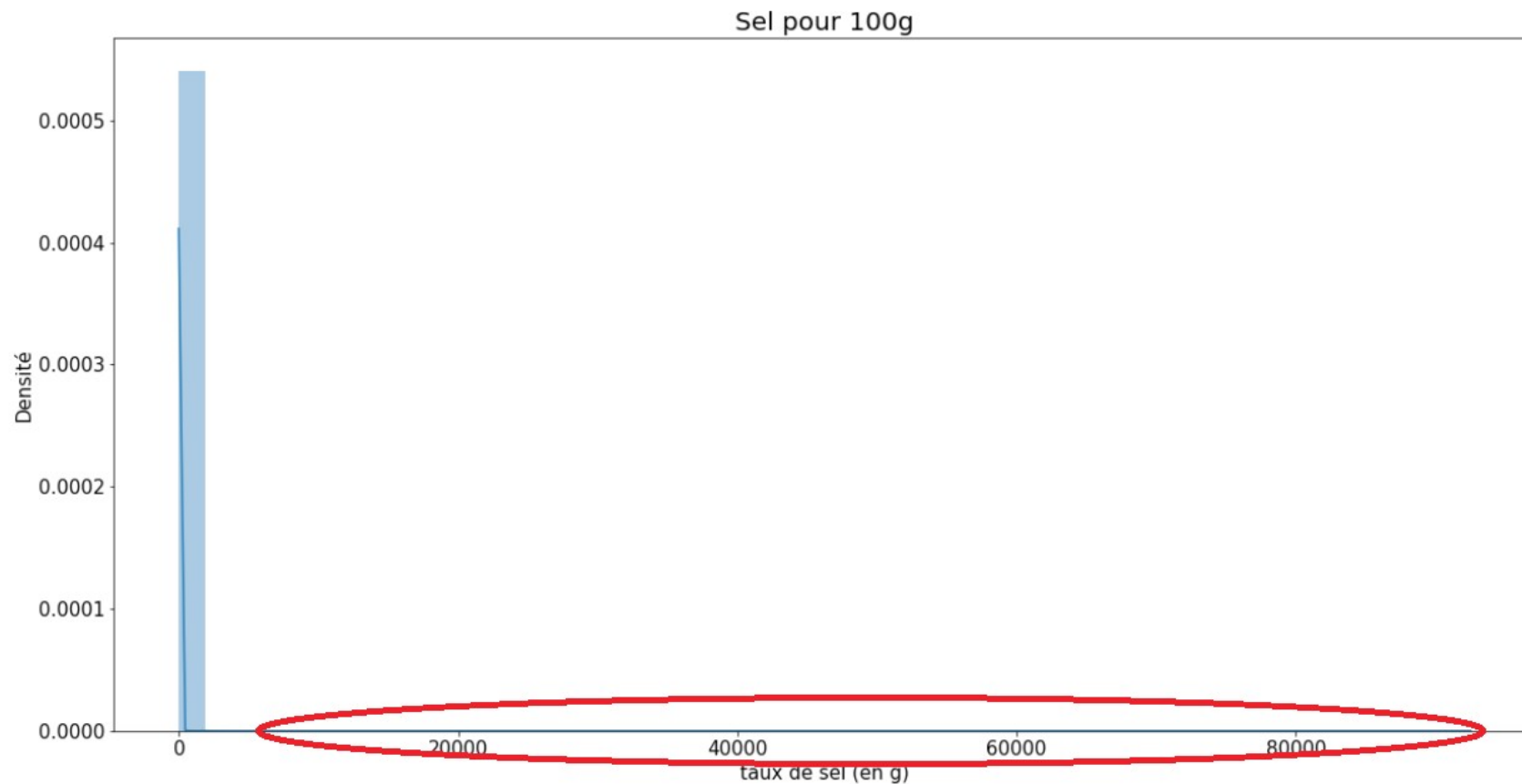
Nettoyage des données

Boxplot du sel pour 100g :



Nettoyage des données

Distribution du sel pour 100g :



Nettoyage des données

Les variables par 100g:

Remplacement des valeurs aberrantes (supérieur à 100 et inférieur à 0).

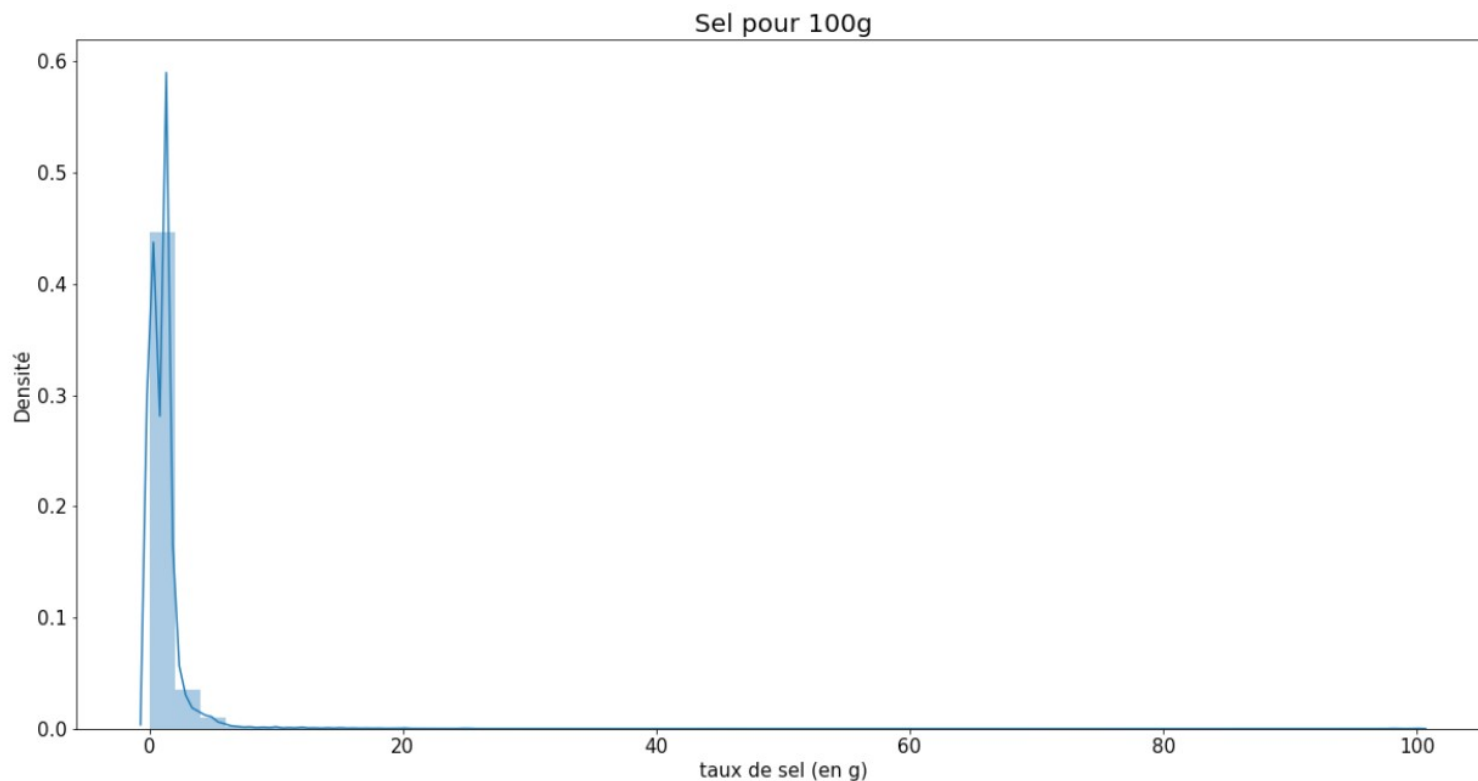
Les autres variables numérique :

Utilisation de borne (écart interquartile) :

```
Q1 = foodDataset[variable].quantile(0.05)
Q3 = foodDataset[variable].quantile(0.95)
borneInf = Q1 - 1.5*(Q3 - Q1)
borneSup = Q3 + 1.5*(Q3 - Q1)
```

Nettoyage des données

Distribution du sel pour 100g après remplacement des valeurs aberrantes et manquantes :





Nettoyage des données

Dataset :

Avant prétraitement:

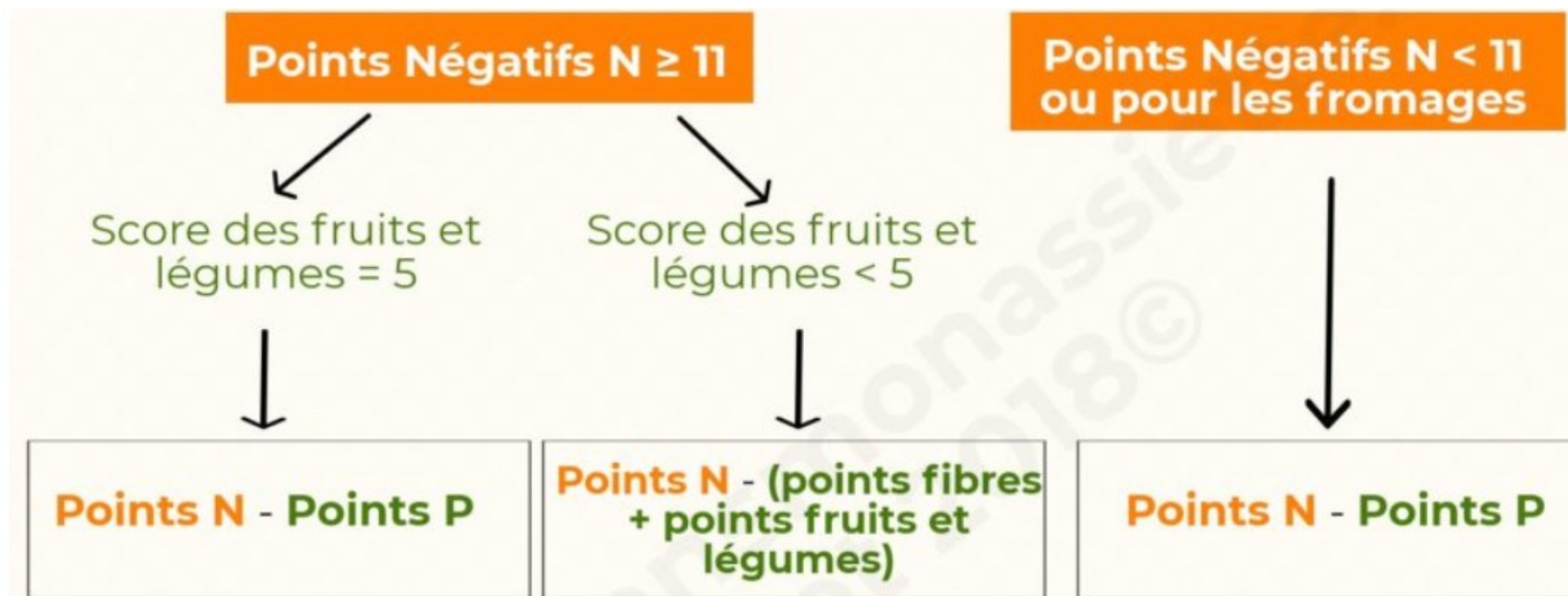
1772411 lignes/ 186 colonnes et 79,57% de valeurs manquantes

Après prétraitement:

1772411 lignes/18 colonnes et 10,97% de valeurs manquantes

Analyse Exploratoire

Calcul du nutriscore:



Analyse Exploratoire

Calcul du nutriscore:

Nutriments à limiter

Points N		Seuils pour les boissons				Seuils pour les matières grasses	
Points	Energie (kJ)	Sucres (g)	Energie (kJ)	Sucres (g)	Graisses saturées (g)	Graisses saturées (%)	Sodium (mg)
0	≤ 335	≤ 4,5	≤ 0	≤ 0	≤ 1	< 10	≤ 90
1	> 335	> 4,5	≤ 30	≤ 1,5	>1	< 16	> 90
2	> 670	> 9	≤ 60	≤ 3	>2	< 22	> 180
3	> 1005	> 13,5	≤ 90	≤ 4,5	>3	< 28	> 270
4	> 1340	> 18	≤ 120	≤ 6	>4	< 34	> 360
5	> 1675	> 22,5	≤ 150	≤ 7,5	>5	< 40	> 450
6	> 2010	> 27	≤ 180	≤ 9	>6	< 46	> 540
7	> 2345	> 31	≤ 210	≤ 10,5	>7	< 52	> 630
8	> 2680	> 36	≤ 240	≤ 12	>8	< 58	> 720
9	> 3015	> 40	≤ 270	≤ 13,5	>9	< 64	> 810
10	> 3350	> 45	> 270	> 13,5	>10	≥ 64	> 900
Gamme (points)	0 à 10	0 à 10	0 à 10	0 à 10	0 à 10	0 à 10	0 à 10
Total		Somme des points pour l'énergie, les sucres, les graisses saturées et le sodium					

Analyse Exploratoire

Calcul du nutriscore:





Nutriments, aliments à encourager

Points P		Seuils pour les boissons		
Points	Fruits, légumes (%)	Fruits, légumes (%)	Fibres (g)	Protéines (g)
0	≤ 40	≤ 40	≤ 0,7	≤ 1,6
1	> 40	-	> 0,7	> 1,6
2	> 60	> 40	> 1,4	> 3,2
3	-	-	> 2,1	> 4,8
4	-	> 60	> 2,8	> 6,4
5	> 80		> 3,5	> 8,0
6	-	-	-	-
7	-	-	-	-
8	-	-	-	-
9	-	-	-	-
10	-	> 80	-	-
Gamme (points)	0 à 5	0 à 10	0 à 5	0 à 5
Total	Somme des points pour les consommations de fruits et légumes, les fibres et les protéines			

Analyse Exploratoire

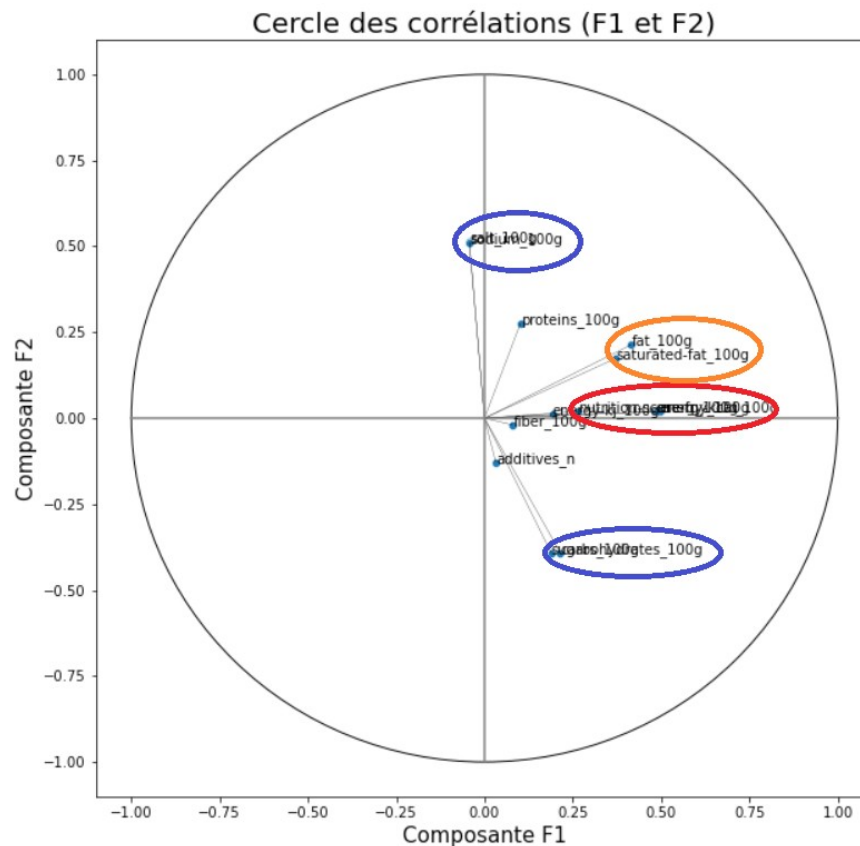
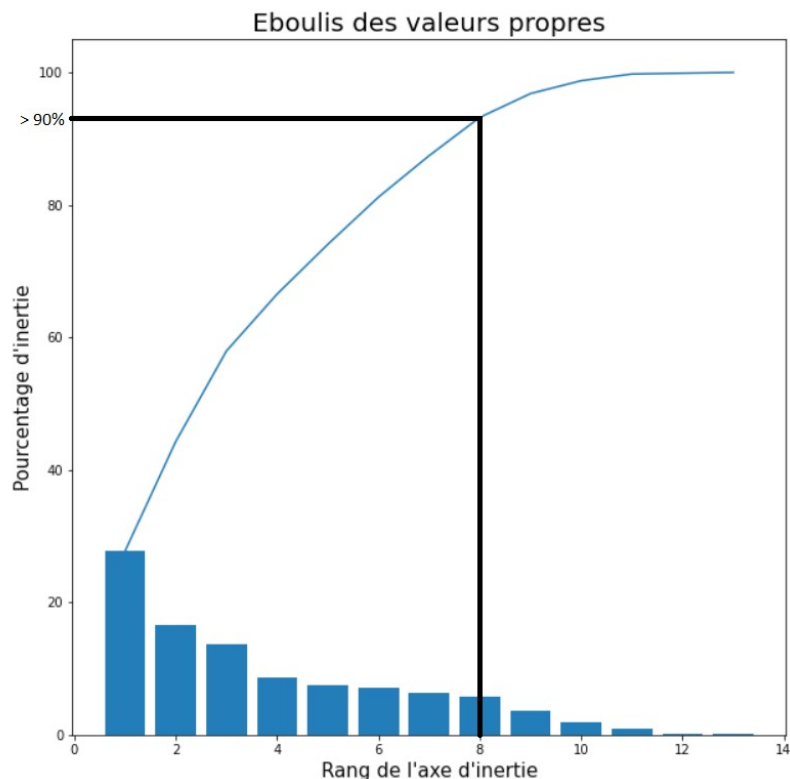
Calcul du nutriscore:

Score final variant de **-15 (qualité nutritionnelle élevée)** à **40 (faible qualité nutritionnelle)**

Aliments solides	Boissons	Logo
Min à -1	Eaux toujours en A	
0 à 2	Min à 1	
3 à 10	2 à 5	
11 à 18	6 à 9	
19 à max	10 à max	

Analyse Exploratoire

Réduction de dimension par ACP et cercle des corrélations :



Analyse Exploratoire

Matrice de Corrélation:

	additives_n	energy-kj_100g	energy-kcal_100g	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	nutrition-score-fr_100g
additives_n	1.0	0.006	0.026	0.027	-0.028	0.0022	0.12	0.13	-0.094	-0.059	-0.017	-0.017	0.11
energy-kj_100g	0.006	1.0	0.17	0.29	0.23	0.18	0.13	0.089	0.068	0.05	-0.0066	-0.0069	0.19
energy-kcal_100g	0.026	0.17	1.0	0.98	0.71	0.51	0.39	0.25	0.13	0.21	-0.046	-0.045	0.31
energy_100g	0.027	0.29	0.98	1.0	0.73	0.53	0.41	0.26	0.14	0.21	-0.046	-0.045	0.33
fat_100g	-0.028	0.23	0.71	0.73	1.0	0.7	-0.074	-0.02	0.059	0.15	-0.017	-0.018	0.27
saturated-fat_100g	0.0022	0.18	0.51	0.53	0.7	1.0	-0.018	0.092	0.0058	0.14	-0.011	-0.011	0.37
carbohydrates_100g	0.12	0.13	0.39	0.41	-0.074	-0.018	1.0	0.68	0.16	-0.18	-0.071	-0.066	0.17
sugars_100g	0.13	0.089	0.25	0.26	-0.02	0.092	0.68	1.0	0.016	-0.25	-0.081	-0.074	0.24
fiber_100g	-0.094	0.068	0.13	0.14	0.059	0.0058	0.16	0.016	1.0	0.1	-0.0093	-0.0095	-0.022
proteins_100g	-0.059	0.05	0.21	0.21	0.15	0.14	-0.18	-0.25	0.1	1.0	0.047	0.041	0.059
salt_100g	-0.017	-0.0066	-0.046	-0.046	-0.017	-0.011	-0.071	-0.081	-0.0093	0.047	1.0	0.9	0.042
sodium_100g	-0.017	-0.0069	-0.045	-0.045	-0.018	-0.011	-0.066	-0.074	-0.0095	0.041	0.9	1.0	0.04
nutrition-score-fr_100g	0.11	0.19	0.31	0.33	0.27	0.37	0.17	0.24	-0.022	0.059	0.042	0.04	1.0

Analyse Exploratoire par KNN

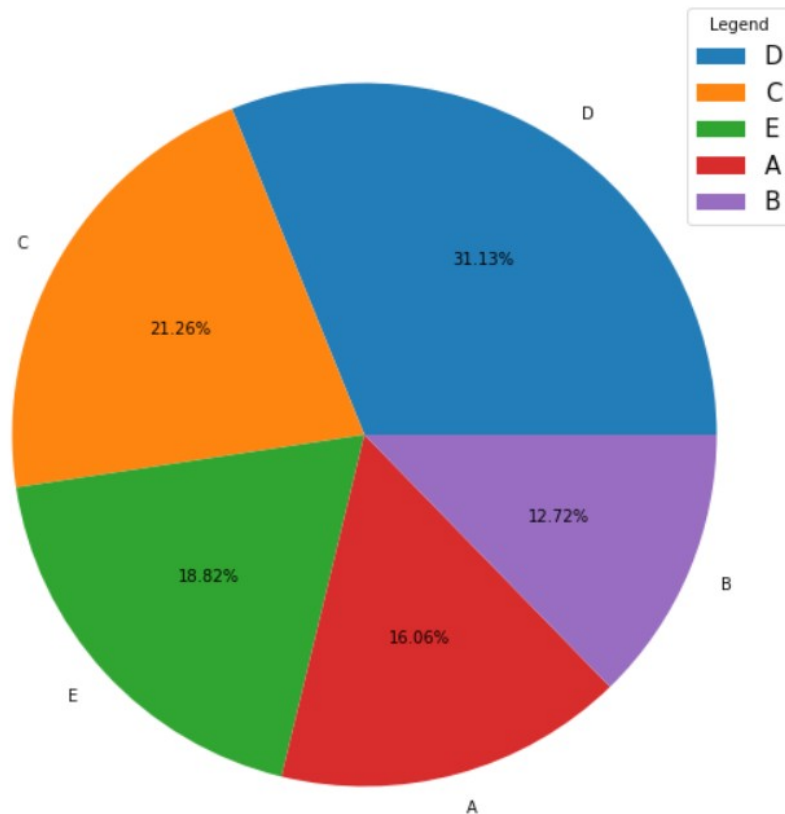
Matrice de Corrélation:

	additives_n	energy_100g	energy-kj_100g	energy-kcal_100g	proteins_100g	carbohydrates_100g	sugars_100g	fat_100g	saturated-fat_100g	fiber_100g	sodium_100g	salt_100g	nutrition-score-fr_100g
additives_n	1.0	-0.011	0.00039	-0.0076	-0.018	0.049	0.11	-0.012	0.0066	-0.15	0.072	0.072	0.19
energy_100g	-0.011	1.0	0.95	1.0	0.32	0.48	0.3	0.69	0.6	0.25	0.11	0.11	0.35
energy-kj_100g	0.00039	0.95	1.0	0.94	0.25	0.47	0.32	0.64	0.56	0.24	0.036	0.036	0.33
energy-kcal_100g	-0.0076	1.0	0.94	1.0	0.32	0.48	0.3	0.69	0.6	0.24	0.11	0.11	0.35
proteins_100g	-0.018	0.32	0.25	0.32	1.0	-0.14	-0.27	0.43	0.41	0.13	0.5	0.5	0.27
carbohydrates_100g	0.049	0.48	0.47	0.48	-0.14	1.0	0.75	0.012	0.033	0.38	-0.19	-0.19	0.024
sugars_100g	0.11	0.3	0.32	0.3	-0.27	0.75	1.0	0.032	0.11	0.25	-0.24	-0.24	0.077
fat_100g	-0.012	0.69	0.64	0.69	0.43	0.012	0.032	1.0	0.89	0.16	0.35	0.35	0.45
saturated-fat_100g	0.0066	0.6	0.56	0.6	0.41	0.033	0.11	0.89	1.0	0.11	0.33	0.33	0.47
fiber_100g	-0.15	0.25	0.24	0.24	0.13	0.38	0.25	0.16	0.11	1.0	0.025	0.025	-0.016
sodium_100g	0.072	0.11	0.036	0.11	0.5	-0.19	-0.24	0.35	0.33	0.025	1.0	1.0	0.47
salt_100g	0.072	0.11	0.036	0.11	0.5	-0.19	-0.24	0.35	0.33	0.025	1.0	1.0	0.47
nutrition-score-fr_100g	0.19	0.35	0.33	0.35	0.27	0.024	0.077	0.45	0.47	-0.016	0.47	0.47	1.0

Analyse Exploratoire

Répartition des nutrigrades:

Répartition des Nutriscores Grades

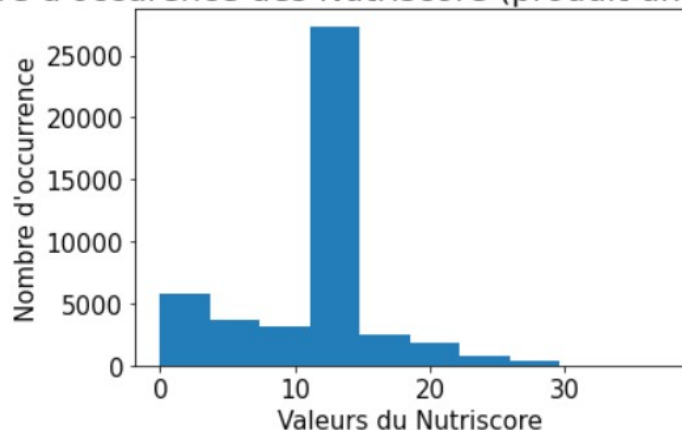


Analyse Exploratoire : Recommandation fourni par application :

Nombre d'occurrence Nutrition score des produits BIO :

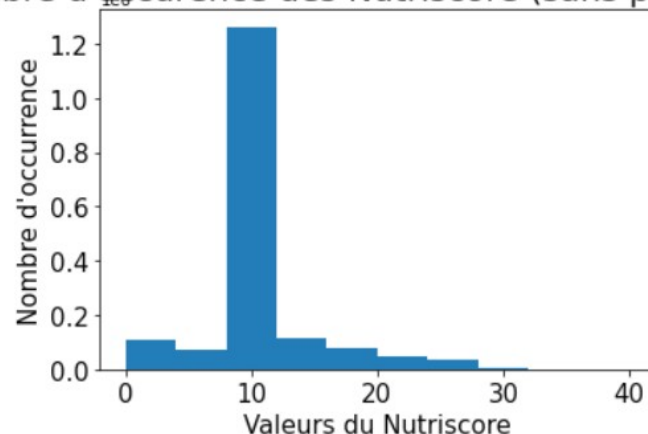
(45465 produits BIO et 1726946 produits non BIO)

Nombre d'occurrence des Nutriscore (produit uniquement BIO)



Il y a 1827 produits ayant un score 0.0 / (Proportion: 4.018475750577368 %)
Il y a 1447 produits ayant un score 1.0 / (Proportion: 3.1826679863631364 %)
Il y a 1344 produits ayant un score 2.0 / (Proportion: 2.956120092378753 %)
Il y a 1146 produits ayant un score 3.0 / (Proportion: 2.520620257340812 %)
Il y a 1086 produits ayant un score 4.0 / (Proportion: 2.3886506103596172 %)
Il y a 913 produits ayant un score 5.0 / (Proportion: 2.008138128230507 %)

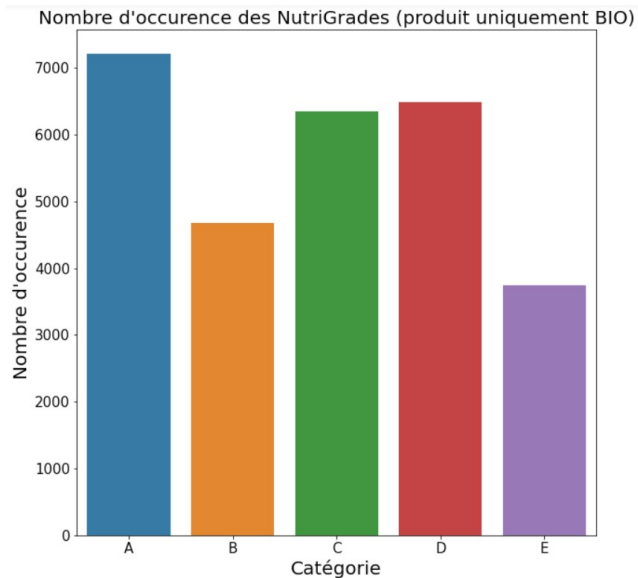
Nombre d'occurrence des Nutriscore (sans produit BIO)



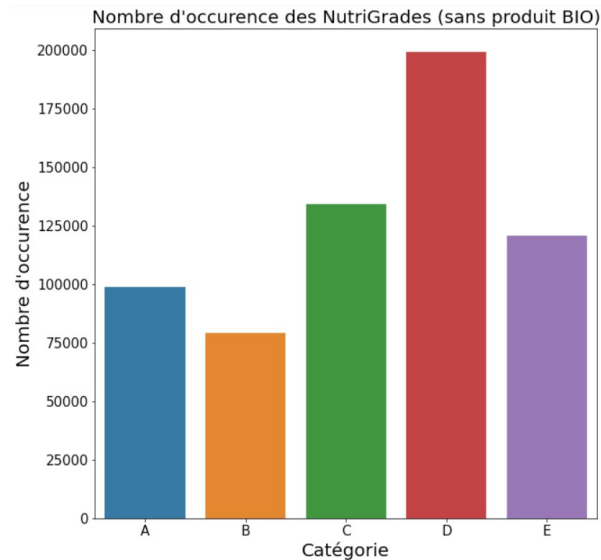
Il y a 31170 produits ayant un score 0.0 / (Proportion: 1.8049203623043222 %)
Il y a 25074 produits ayant un score 1.0 / (Proportion: 1.4519272750856136 %)
Il y a 26357 produits ayant un score 2.0 / (Proportion: 1.5262202755615983 %)
Il y a 25365 produits ayant un score 3.0 / (Proportion: 1.4687778309223334 %)
Il y a 22747 produits ayant un score 4.0 / (Proportion: 1.3171807340820152 %)
Il y a 20143 produits ayant un score 5.0 / (Proportion: 1.1663943169039448 %)

Analyse Exploratoire : Recommandation fourni par application :

Nombre d'occurrence Nutrition grade des produits BIO et des produits non BIO :



Il y a 7208 produits ayant un grade a / (Proportion: 25.34013007558446 %)
Il y a 4670 produits ayant un grade b / (Proportion: 16.417648092810687 %)
Il y a 6480 produits ayant un grade d / (Proportion: 22.780805062401125 %)
Il y a 3737 produits ayant un grade e / (Proportion: 13.13763403058534 %)
Il y a 6350 produits ayant un grade c / (Proportion: 22.323782738618387 %)



Il y a 98858 produits ayant un grade a / (Proportion: 15.645182987141443 %)
Il y a 79325 produits ayant un grade b / (Proportion: 12.553907022749753 %)
Il y a 199097 produits ayant un grade d / (Proportion: 31.508921859545 %)
Il y a 120548 produits ayant un grade e / (Proportion: 19.07782393669634 %)
Il y a 134047 produits ayant un grade c / (Proportion: 21.214164193867457 %)



Conclusion

- Indépendance des variables entre elles
- Calcul du nutriscore simple et taux de complétion des variables du calcul du nutriscore satisfaisante : faisabilité validée
- Validation de la recommandation fournie : démonstration de l'importance du BIO
- Absence de l'indication local dans dataset/ limitation du aux données manquantes
- Quelles perspectives d'améliorations ?

Merci de votre attention