EA Stephen
Data scientist

Concevez une application au service de la santé publique

Nutri frozen

Sommaire

1. La présentation de l'application Nutri Frozen (recommandations des produits surgelés)
2. Les opérations de nettoyage effectuées en lien avec l'application Nutri Frozen
3. Analyse exploratoire
4. 3 observations solidement étayées (graphes et/ou tests statistiques à l'appui au besoin) évaluant la pertinence et la faisabilité de mon application.
5. Tester les fonctionnalités du système de recommandation
6. La synthèse des différentes conclusions sur la faisabilité du projet.

La mission

L'agence "Santé publique France" a lancé **un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation.** Je souhaite y participer et proposer une idée d'application.

Le besoin pour les produits surgelés

Monde

220 Mrd $
de ventes générées dans le monde entier

+ 5,1 %
de croissance prévue pour les cinq prochaines années

Les légumes
représentent la majorité des ventes mondiales

France

98,6 %
des ménages achètent des aliments congelés

9 Mrd €
de chiffre d'affaires réalisés en France

2 M de tonnes
de marchandises par an

L’idée d’application

L'application Nutri frozen (France) a pour objectif d'indiquer les apports nutritionnels des produits surgelés, en les scanants avec son smartphone. L'utilisateur pourra s'appuyer sur une mesure nutri score grade qui indiquera la qualité nutritive de l'article, ensuite l'application va recommander des produits similaires au produit scanné avec un meilleur score, ou une catégorie d'article le plus ressemblant à ce dernier. Il aura aussi une information sur la localité afin qu'il puisse trouver ce produit, dans les magasins des grandes distributions alimentaires. Une autre fonctionnalitée sera de fournir des recettes de plats et de recommander les produits en liens avec la recette suggérée par l’application, tout en respectant la suggestion des meilleurs produits Nutri Score Grade.

Des idées de développement sont à l'étude, pour améliorer le système de recommandation qui permettra de proposer des produits en similarité avec d’autres utilisateurs qui auront les mêmes habitudes d'achats et habitude culinaire. Le système permettra aussi que les utilisateurs puissent évaluer les produits qu’ils ont acheté.

Fonctionnement de l'application Nutri Frozen

Scan grade C

User

Recommandation recette et produits

Produit similaire

Localité : Picard

Recommandation user (nutri frozen grade A)

Description du dataset Openfoodsfact

- Les dimensions du dataset: (1985074, 186)
- Le types des colonnes du dataset :
  dtypes: float64(124), int64(2), object(60)
- memory usage avant: 2.8+ GB
  => après 1.5+ GB
- 79.82 % NaN dans le dataset soit (294 719 332 NaN)

Description variables

- Url, contributeurs, noms des produits, pays, marques, magasins, catégories, Nutriscore des produits , apports nutritionnels des produits 100g

Les Opérations de nettoyage effectuées pour l'application

Filtrer les produits vendus en France

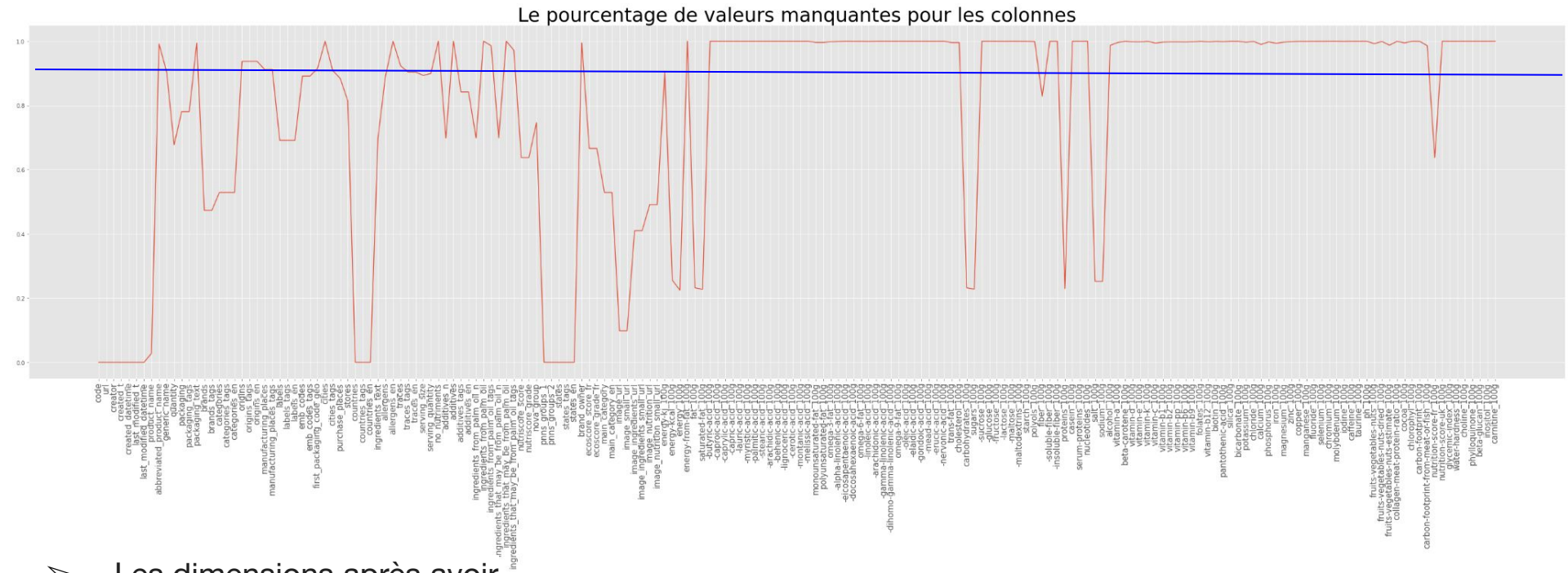
- 'France,'France,United States','Belgium,France', ...,
  'Australia,Belgium,France,New Zealand,Spain',
  'Australia,Belgium,France,United Kingdom,United States',
  'Australia,Belgium,France,Japan,New Zealand,United Kingdom,United States'

- 1 doublon supprimé

- Il y a 844 136 produits et 186 variables qui décrivent ces produits

Nutri frozen

Supprimer les informations selon le taux de remplissage

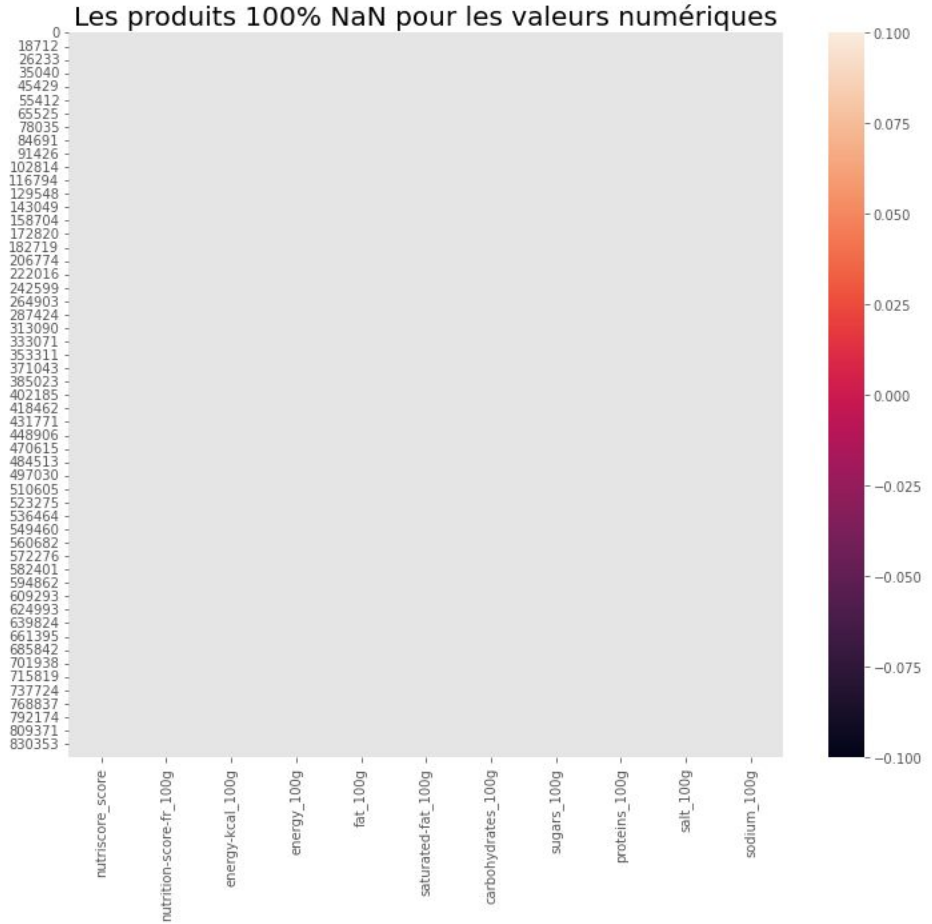


- Les dimensions après avoir supprimés les colonnes avec 90% valeurs manquantes (844136, 64)

- Supprimer 122 colonnes

Supprimer les produits sans attributs

- Supprimer les produits contenant 100% de valeurs manquantes (Floats)
- 175 200 produits supprimés
- Après suppressions, il y a 668 936 produits et 24 colonnes



Sélectionner les colonnes pertinentes pour l'application

- Les colonnes conservées (24 colonnes) :

  Informer le consommateur sur la recommandation et elles ont permis de nettoyer les outliers

- **Vérifier cohérence de la data (packaging du produit)**

  url

- **Détection outliers, création fonctionnalitée du système de recommandation**

  countries, product_name , brands , stores, categories, categories_tag, categories_en, main_category, pnns_groups_1, pnns_groups_2, quantity,

- **Détection des apports nutritionnels mal renseignés, détection des nutriscore faussés par le mauvais renseignement des contributeurs, fonctionnalité du système de recommandation**

  nutriscore_grade, nutriscore_score, energy_100g, saturated-fat_100g, sugars_100g, proteins_100g, fat_100g, salt_100, energy-kcal_100g, sodium_100g, carbohydrates_100g

Supprimer les produits sans (noms, brands, stores)

Product_name

➔ Ne souhaite pas de produits sans description du produit

Brands

➔ Ne souhaite pas de produits sans marques

Stores

➔ Ne souhaite pas de produits sans pouvoir localiser le magasin

➢ Les dimensions après le nettoyage

(134 155,24)

Filtrer les catégories contenant les produits surgelés

1 Filtrer les catégories

- Notnull

2 Normaliser le texte

Objectif : faire correspondre les catégories

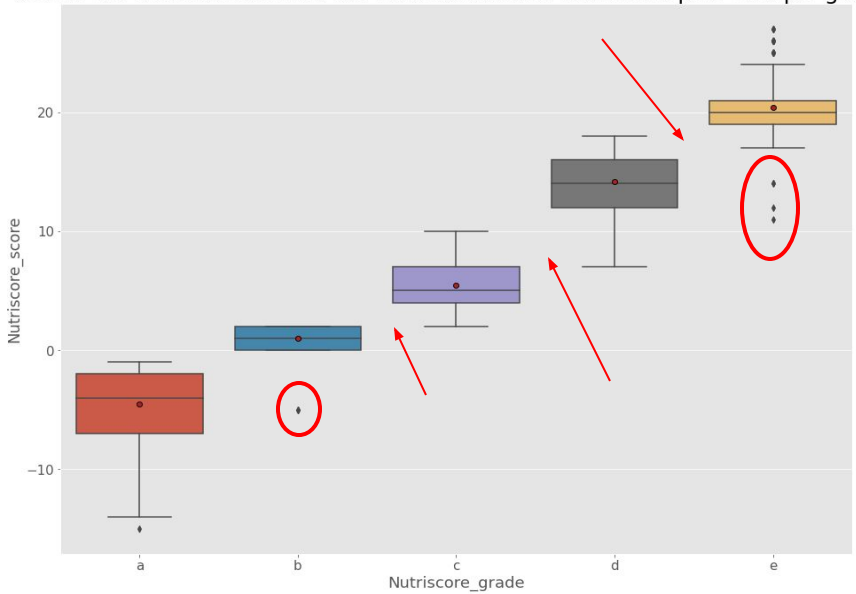
- Minuscule
- Lemmatisation (forme canonique, infinitif, singulier)

3 Sélections expression dans une chaîne

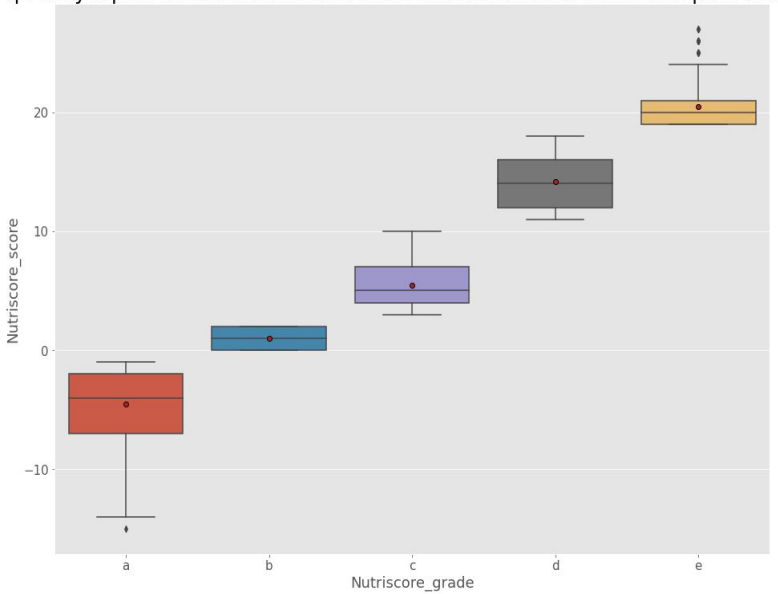
'frozen food|frozen meat|frozen meal|frozen ready meal|frozen vegetable|frozen soup|frozen vegetable soup|frozen meal|frozen steak|frozen plant-based food|frozen cake and pastry|ice cream and sorbet'

Délimiter les groupes nutriscore grade (Local factor outliers)

Vérifier les chevauchements sur les distributions nutriscore pour chaque groupe



Vérifier qu'il n'y a plus de chevauchement entre les distributions nutriscore pour chaque groupe



nutriscore_grade	nutriscore_score min	nutriscore_score max
a	-15.0	-1.0
b	-5.0	2.0
c	2.0	10.0
d	7.0	18.0
e	11.0	27.0

Local factor outlier

[-5], 3 produits supprimés (boissons jus de fruit)
[2.0], 2 produits supprimés (boissons jus de fruit)
[7.0, 9.0], 4 produits supprimés (boissons jus de fruit)
[11.0, 12.0, 14.0, 17.0] 5 produits supprimés (café, coca-cola, jus fruit)

nutriscore_grade	nutriscore_score min	nutriscore_score max
a	-15.000000	-1.000000
b	0.000000	2.000000
c	3.000000	10.000000
d	11.000000	18.000000
e	19.000000	27.000000

Supprimer les outliers pour les apports nutritionnels

Supprimer outliers aberrants mal renseignés

224 produits aberrants supprimés

- produits mal catégorisés par les contributeurs
- Contributeurs mal renseignés sur les apports nutritionnels qui fausse l'attribution de la note à un produit)

Les dimensions après nettoyage

- 7506 produits et 26 descripteurs

Remplacer valeurs mal renseignées

15 valeurs remplacer

- **Le contributeur a inversé les valeurs entre kcal_100g energie_100g (kj)**

Features extractions CV

- Détecter les valeurs mal renseignées par les contributeurs (packaging)
- Informer correctement les utilisateurs

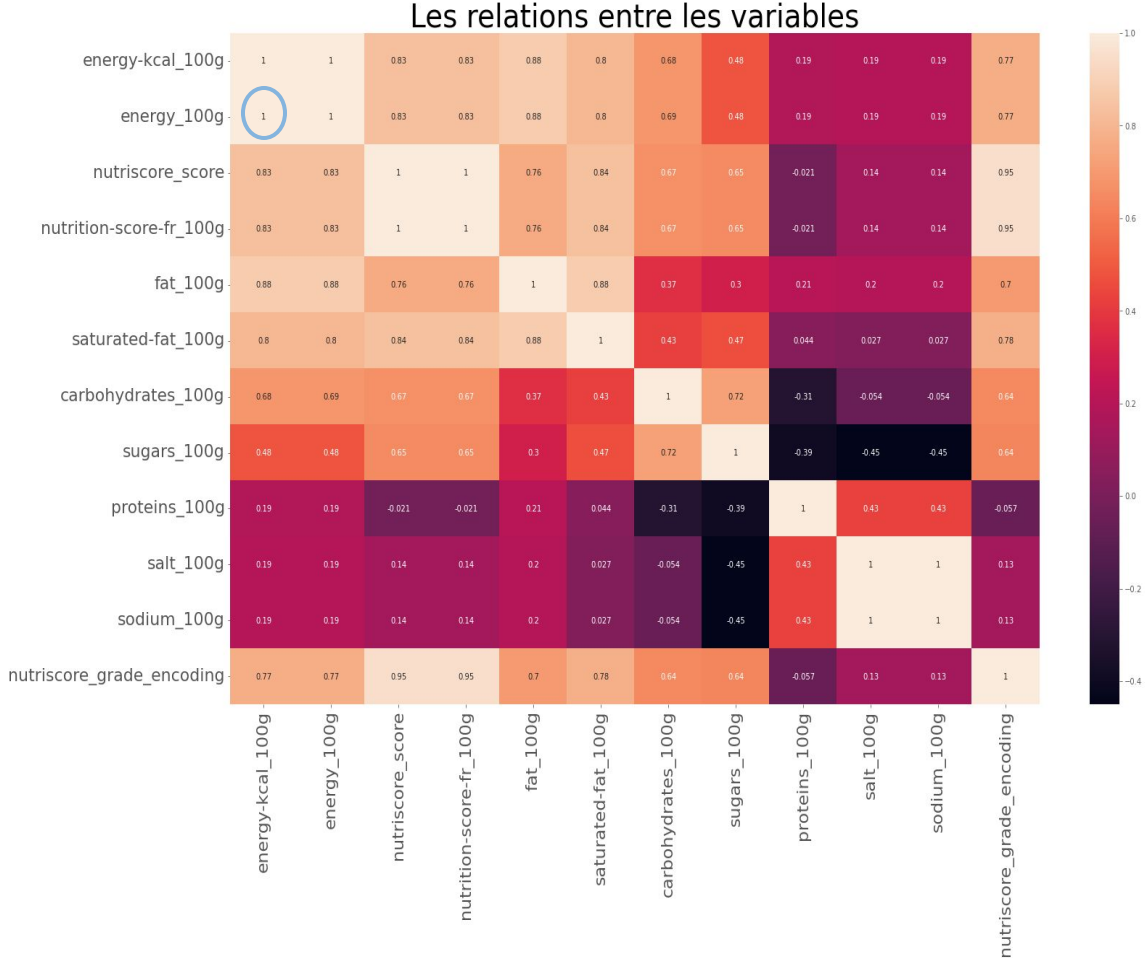
Imputation KNN imputer

1210 valeurs à imputer

	energy-kcal_100g	energy_100g
5884	nan	700.000000
8035	nan	450.000000
8044	181.000000	757.000000
10000	nan	549.000000
11143	87.700000	367.000000
14259	132.000000	550.000000
15922	140.000000	585.000000
16535	57.000000	201.000000
17750	131.580000	551.000000
19784	46.000000	192.000000

	energy-kcal_100g	energy_100g
0	168.0	700.0
1	108.0	450.0
2	181.0	757.0
3	132.0	549.0
4	87.7	367.0
...	...	...
7501	180.0	757.0
7502	398.0	1659.0
7503	324.0	1356.0
7504	226.0	948.0
7505	371.0	1546.0

7506 rows x 2 columns

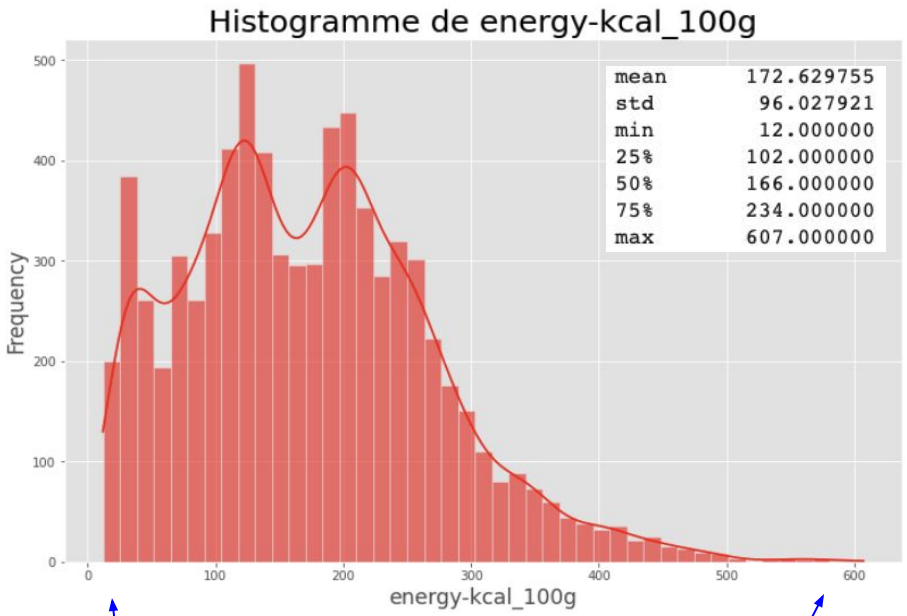


Statistiques descriptives

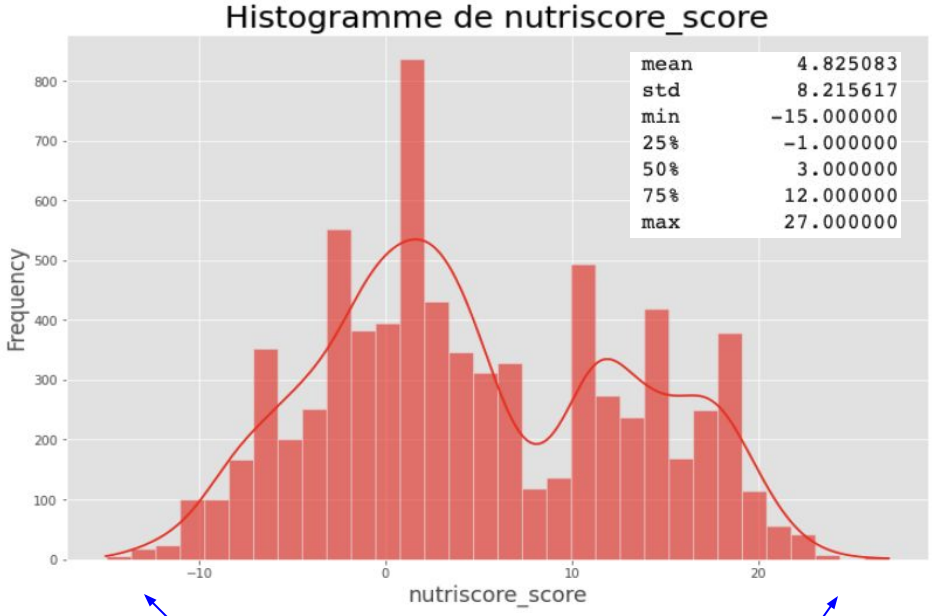
nutriscore_grade	nutriscore_score			energy-kcal_100g			energy_100g			saturated-fat_100g		
	mean	std	median	mean	std	median	mean	std	median	mean	std	median
a	-4.572292	2.991189	-4.000000	85.652157	56.338540	75.000000	358.884519	236.121132	312.000000	0.627536	0.802096	0.300000
b	1.020309	0.813750	1.000000	145.038590	54.230414	138.000000	606.983420	227.353141	573.000000	1.872575	1.358615	1.700000
c	5.476374	2.256258	5.000000	167.503331	59.090896	166.000000	702.118407	246.937715	695.000000	3.002887	2.230008	3.100000
d	14.202604	2.408318	14.000000	256.420133	61.481610	251.000000	1072.345313	256.958291	1049.000000	8.278505	3.680619	7.740000
e	20.126005	1.333076	20.000000	358.973190	60.655375	345.000000	1497.790885	254.836368	1438.000000	13.660912	3.522587	13.500000

nutriscore_grade	sugars_100g			proteins_100g			fat_100g			salt_100g			carbohydrates_100g			sodium_100g		
	mean	std	median	mean	std	median	mean	std	median	mean	std	median	mean	std	median	mean	std	median
a	2.250204	2.662363	1.600000	6.212283	6.542433	2.800000	2.773309	3.369786	1.300000	0.301869	0.301924	0.180000	7.762515	7.710689	4.900000	0.120748	0.120770	0.072000
b	2.828744	4.055698	1.700000	7.120260	5.173717	6.100000	5.811357	3.494266	5.430000	0.677660	0.325984	0.710000	15.074377	9.388572	15.000000	0.271064	0.130393	0.284000
c	10.917746	10.369247	4.700000	5.297740	5.149595	3.700000	6.844230	5.017737	6.700000	0.550247	0.512832	0.495000	20.601017	10.155256	23.000000	0.220098	0.205130	0.198000
d	16.823276	11.170034	21.000000	5.913110	4.284190	4.200000	14.013757	5.994612	13.000000	0.541001	0.572017	0.200000	26.373515	10.643106	27.000000	0.216402	0.228809	0.080000
e	26.893592	9.569098	29.000000	5.237003	3.336947	4.400000	21.952654	6.052804	21.000000	0.385678	0.444614	0.220000	34.587882	10.629854	35.000000	0.154301	0.177942	0.088000

Analyse univariée (Kcal, Nutriscore)

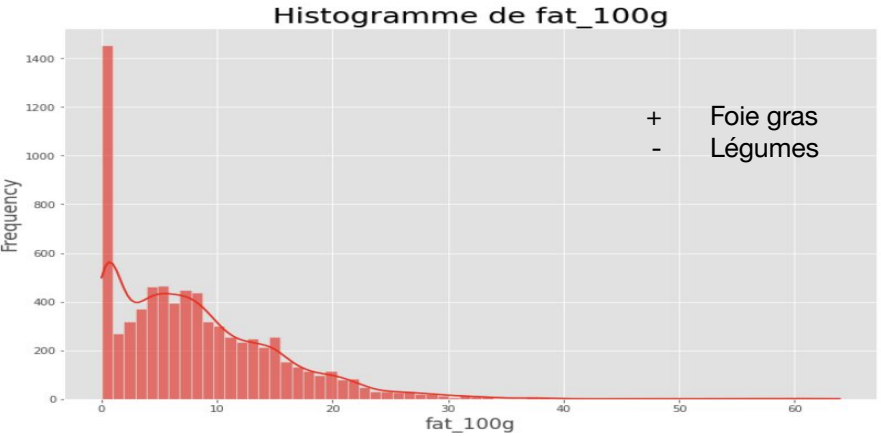


- Légumes (Champignons, choux fleurs, brocolis)
- Foie gras
- pâtisserie

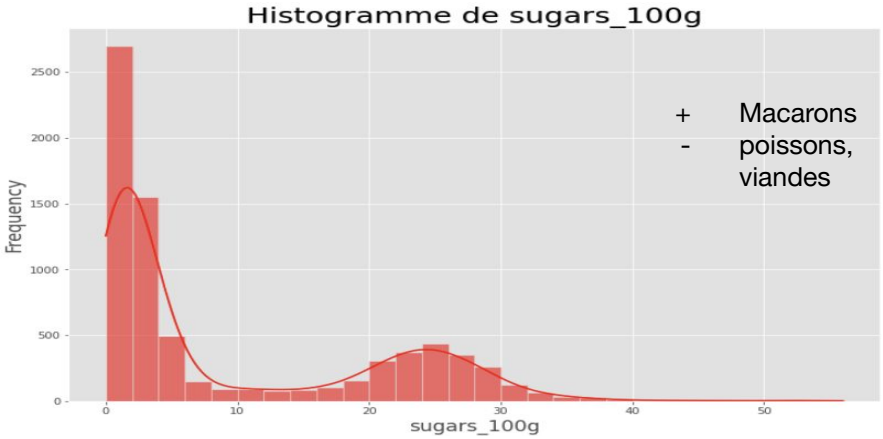


- Légumes verts
- Glaces, patisseries

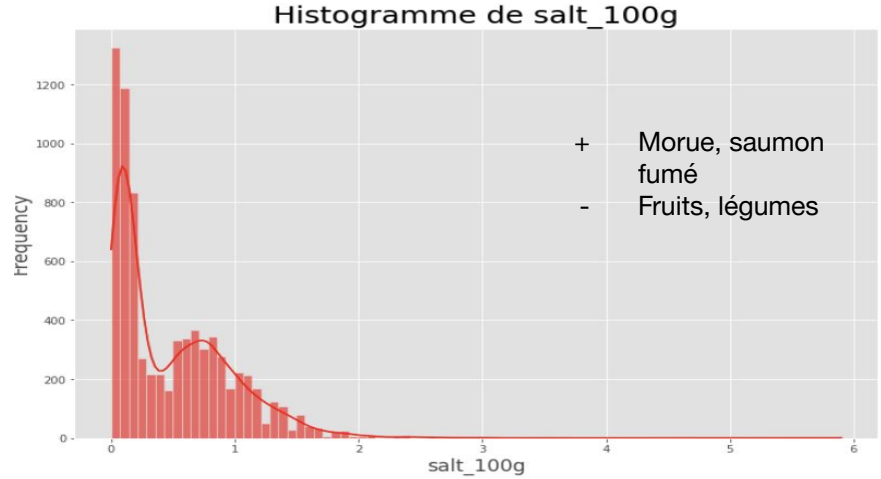
Analyse univariée apports nutritionnels



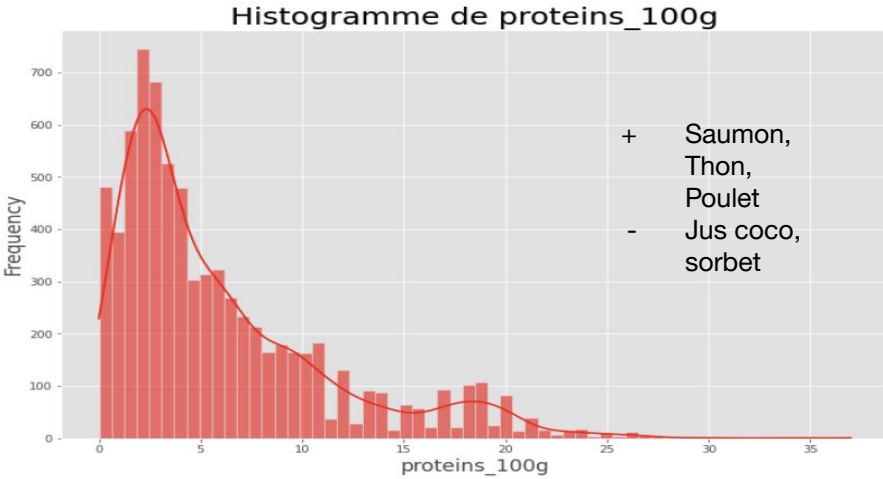
\+ Foie gras
\- Légumes



\+ Macarons
\- poissons, viandes



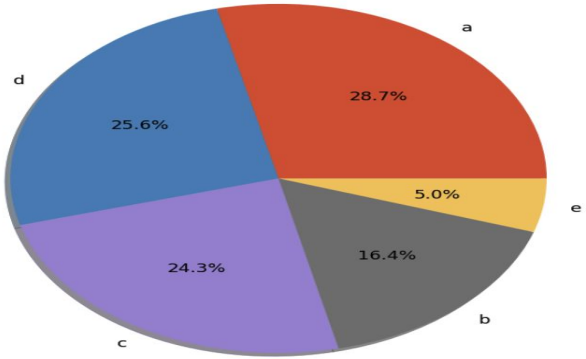
\+ Morue, saumon fumé
\- Fruits, légumes



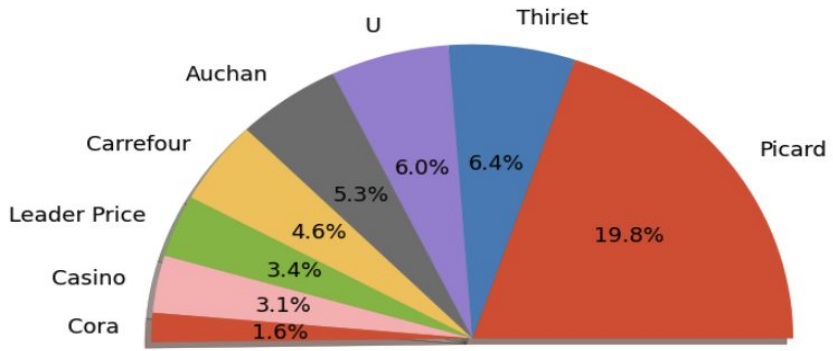
\+ Saumon, Thon, Poulet
\- Jus coco, sorbet

Répartition des produits

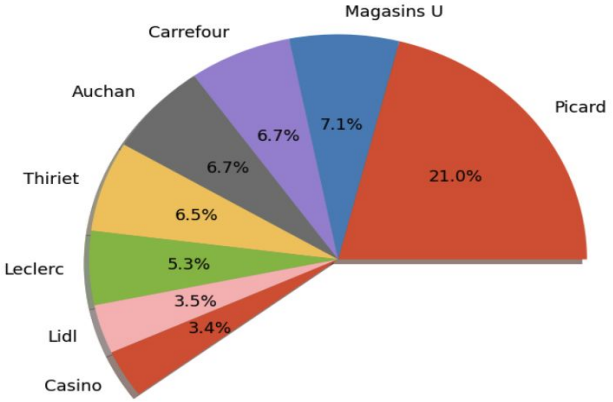
Répartition des nutriscore_grade les plus représentées



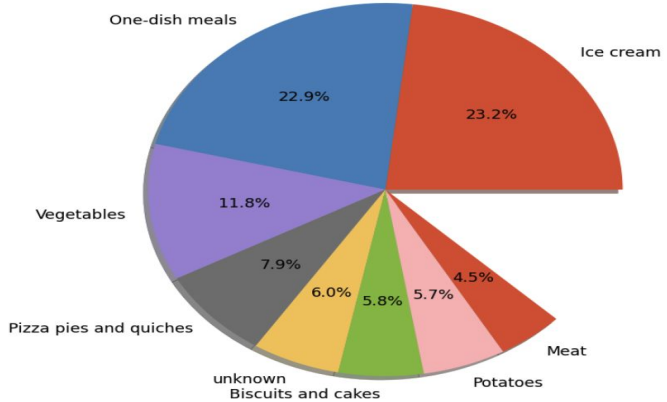
Répartition des brands les plus représentées



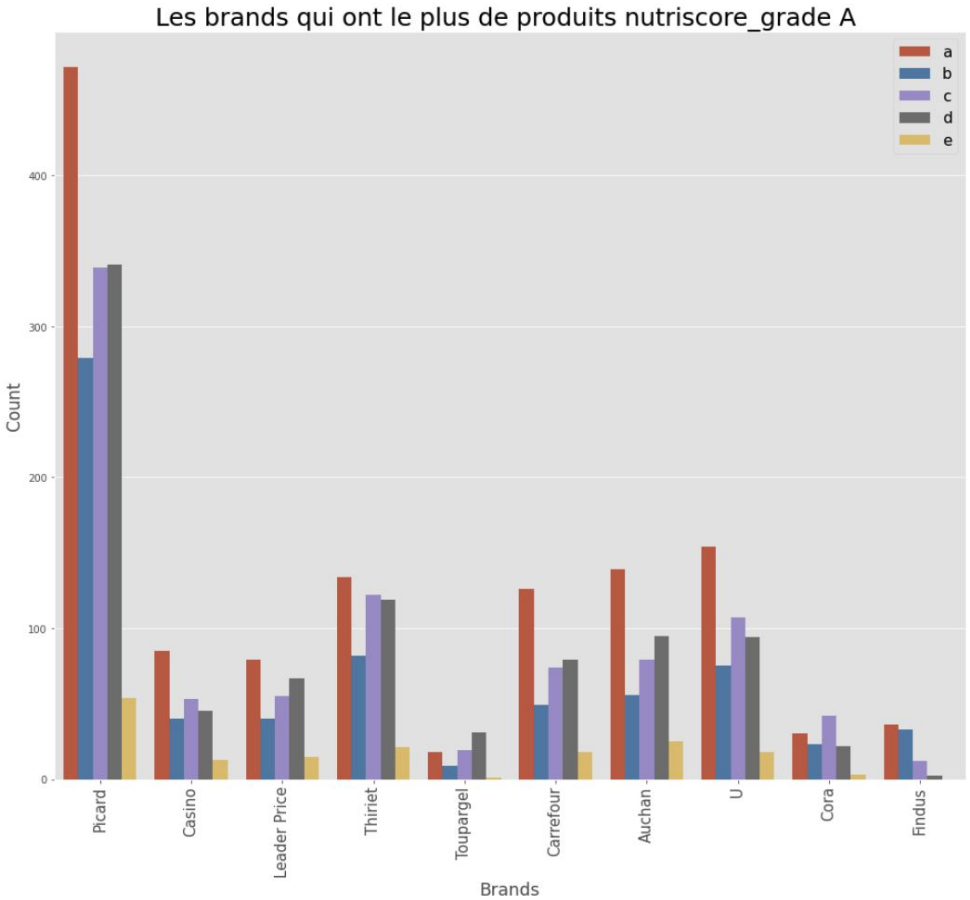
Répartition des stores les plus représentées

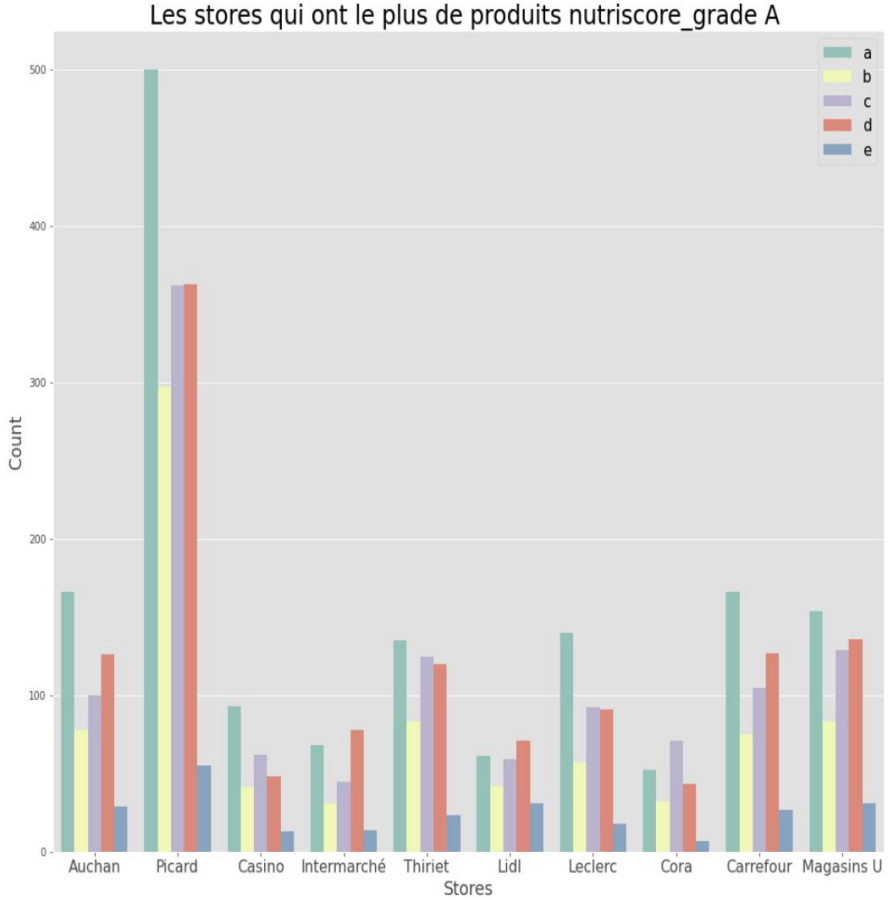


Répartition des pnns_groups_2 les plus représentées

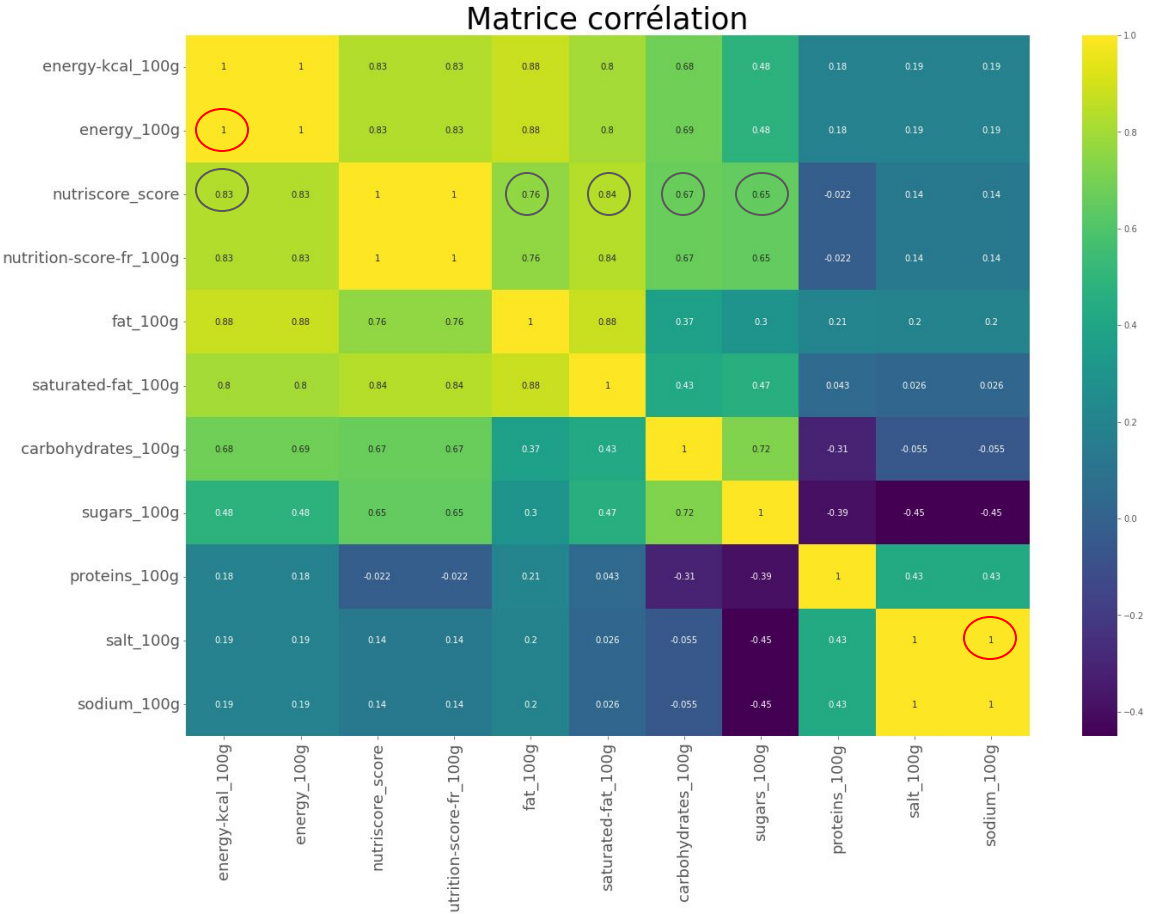


Répartition des produits nutriscore grade A





Les relations entre les attributs des produits



Il y a des corrélations linéaire (pearson) à 1

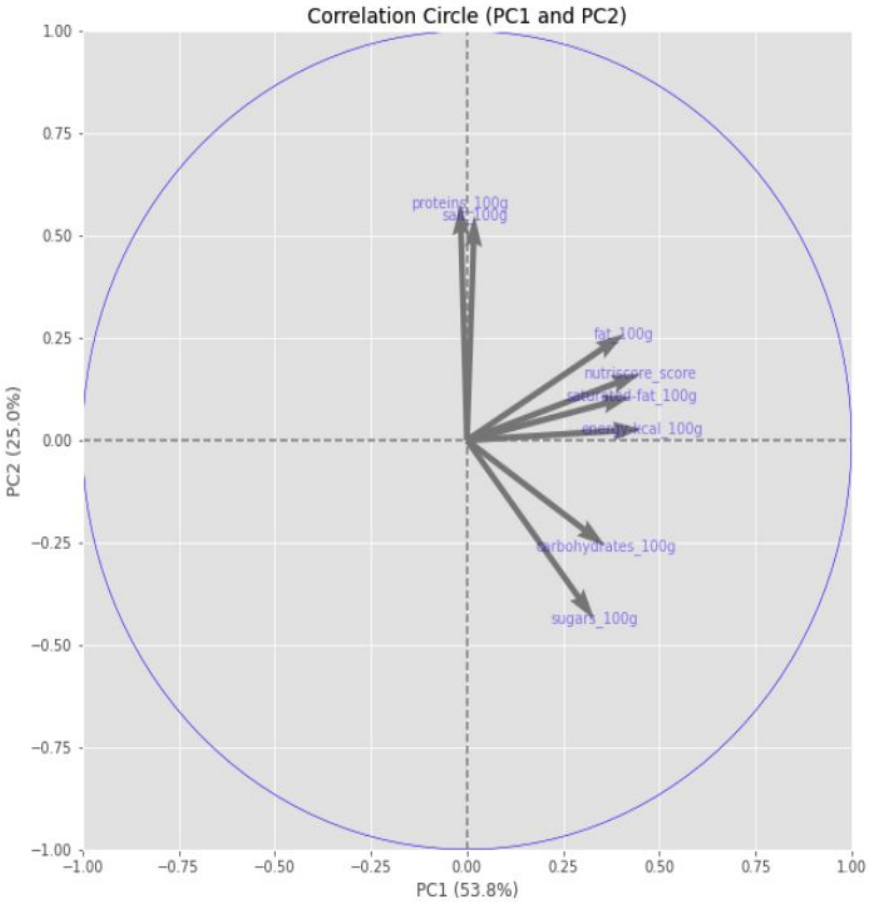
> energie Kcal_100g et Energie_100g (KJ)(mesure physique)

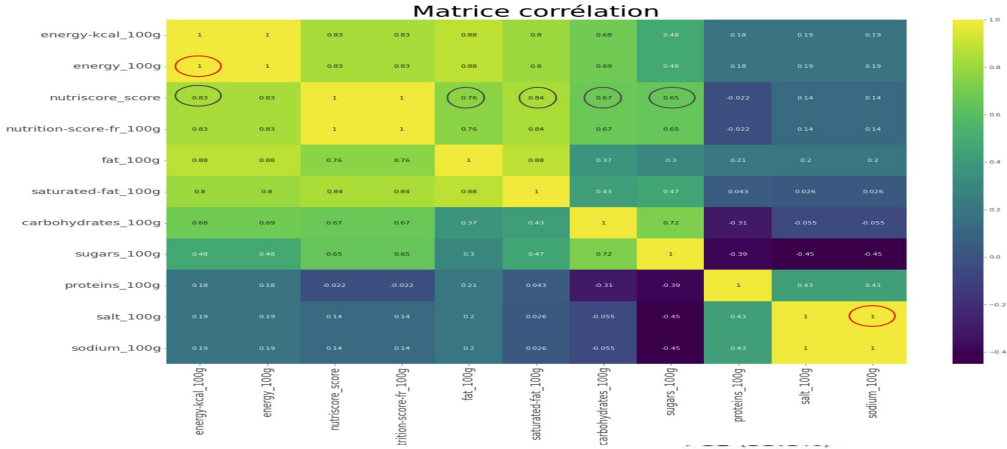
> Salt et Sodium (mesure physique)

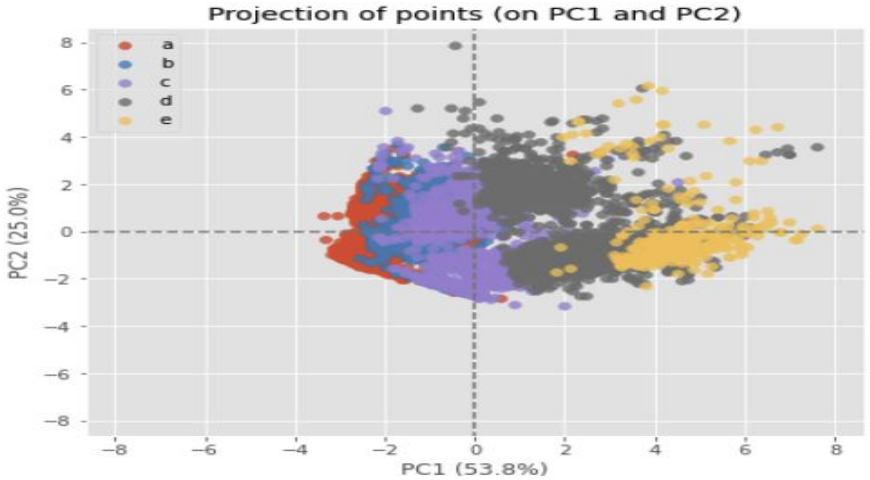
Nutriscore Grade

> + nutriscore_score est élevé et + (energy_kcal_100g, Fat_100g, fat_saturated, sugars, carbohydrates)

ACP


Correlation Circle (PC1 and PC2)
proteins_100g
salt_100g
fat_100g
nutriscore_score
saturated-fat_100g
energy-kcal_100g
carbohydrates_100g
sugars_100g
PC1 (53.8%)
PC2 (25.0%)


Matrice corrélation
energy-kcal_100g
energy_100g
nutriscore_score
nutrition-score-fr_100g
fat_100g
saturated-fat_100g
carbohydrates_100g
sugars_100g
proteins_100g
salt_100g
sodium_100g


Projection of points (on PC1 and PC2)
a
b
c
d
e
PC1 (53.8%)
PC2 (25.0%)

Déterminer si les groupes nutriscore grade sont différents

Hypothèses des test paramétriques non respectées
Variance et normalité

Kruskal-wallis Test 5%

L'hypothèse nulle (H_0) : La médiane est égale dans tous les groupes.

L'hypothèse alternative : (Ha) : La médiane n'est *pas* égale dans tous les groupes.

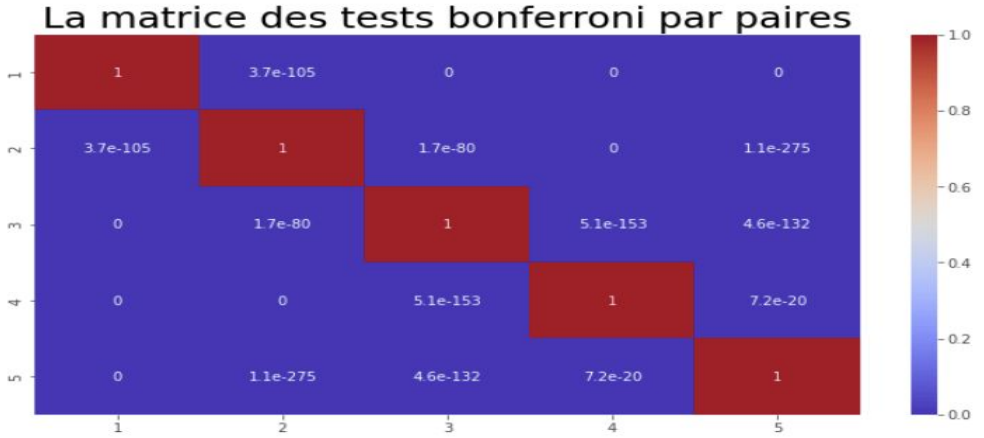
stat=21181.244, p=0.000

Il y a au moins un groupe nutriscore_grade significativement différent des autres groupes

Test Dunn par paires 5%

Si les résultats d'un test de Kruskal-Wallis sont statistiquement significatifs, il est alors approprié d'effectuer le test de Dunn pour déterminer exactement quels groupes sont différents.

rejet HO pour chaque paires (p-adj ≈ 0)
Tous les groupes par paires sont significativement différents



Test paramétrique

Anova à un facteur 5%

H0 : les moyennes des échantillons sont égales.

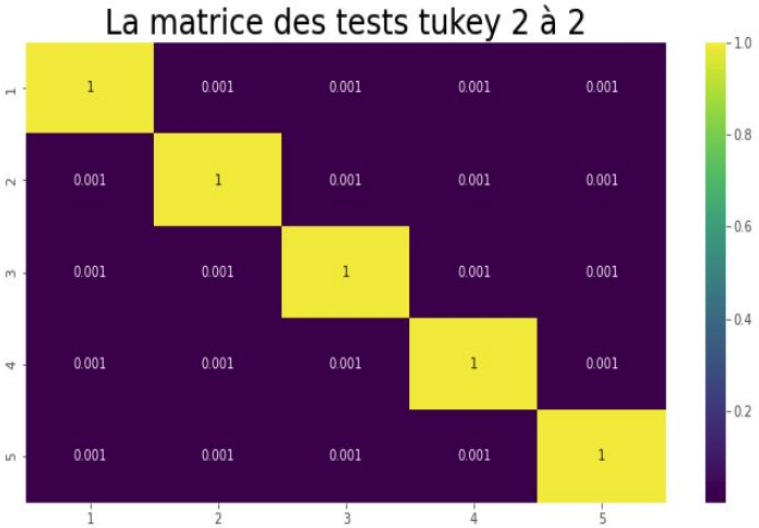
H1 : une ou plusieurs moyennes des échantillons sont inégales.

stat=21181.244, p=0.000

Rejet H0 : P value = 0

Il y au moins un groupe Nutriscore Grade qui diffère des autres

Test Tukey par paires 5%



Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
a	b	5.5926	0.001	5.3642	5.821	True
a	c	10.0487	0.001	9.8451	10.2522	True
a	d	18.7749	0.001	18.5742	18.9756	True
a	e	24.6983	0.001	24.3398	25.0568	True
b	c	4.4561	0.001	4.2202	4.6919	True
b	d	13.1823	0.001	12.9489	13.4157	True
b	e	19.1057	0.001	18.7279	19.4835	True
c	d	8.7262	0.001	8.5171	8.9353	True
c	e	14.6496	0.001	14.2864	15.0129	True
d	e	5.9234	0.001	5.5617	6.2851	True

Tester fonctionnalité de l'app Nutri Frozen

Scan produit: 2258
Nutriscore B

Recommandations 10 produits similarités avec 2258

Nutriscore A

	product_name	nutriscore_grade	brands	stores	energy-kcal_100g	fat_100g	saturated-fat_100g	sugars_100g	proteins_100g	salt_100g	sodium_100g	2258
888	8 biftecks hachés pur bœuf	a	Picard	Picard	155.0	7.8	3.9	0.2	21.0	0.170	0.068	0.987961
886	biftecks haches	a	Picard	Picard	155.0	7.8	3.9	0.2	20.7	0.170	0.068	0.987521
889	6 biftecks hachés de charolais pur boeuf	a	Picard	Picard	144.0	6.8	3.5	0.4	19.9	0.220	0.088	0.978688
1844	steak haché pur bœuf	a	Auchan	Simply market,Auchan	145.0	5.0	2.0	1.0	25.0	0.200	0.080	0.970557
1995	steaks hachés 5% mat. gr.	a	Chantegril, Marque Repère, Férial	Leclerc	132.0	5.0	2.5	0.0	21.0	0.200	0.080	0.960756
240	10 steaks hachés pur boeuf - 5% mat. gr.	a	Casino	Casino	129.0	5.0	2.2	0.0	21.0	0.200	0.080	0.957632
277	10 steaks hachés pur boeuf 5% mg	a	Casino	Casino	129.0	5.0	2.2	0.0	21.0	0.150	0.060	0.956894
2092	10 steaks hachés pur bœuf 5% m.g.	a	Casino	Casino	129.0	5.0	2.2	0.0	21.0	0.150	0.060	0.956769
281	le pur bœuf 5% de m.g.	a	Charal	Franprix,Magasins U	128.0	5.0	2.1	0.0	20.5	0.175	0.070	0.953935
1279	steacks hachés le pur boeuf	a	Carrefour	Carrefour Market, Carrefour	125.0	5.0	2.2	0.0	20.0	0.250	0.100	0.951627

Méthode pour similarités des produits

1 Normaliser texte categories
2 Cosine similarity
3 Robust scaling
4 Filtrer nutriscore_grade A
5 Trier best similarity avec produit 2258

Conclusion faisabilité de l'app Nutri Frozen

1. Les groupes nutri score grade ont été définis par la suppression des outliers.

2. Les tests indiquent que les groupes nutri score grade sont significativement différents.

3. Les stores et les marques ont plus de produits nutri score grade A.

4. Des erreurs de saisies fournis par les contributeurs ont été supprimé ou remplacé.

5. Il faut contrôler les attributs dans les tendances centrales afin de mieux informer l'utilisateur sur les mesures (Computer vision packaging).

6. Des fonctionnalités ont été testé pour application de recommandation.