

The background of the slide features a complex, abstract network diagram. It consists of numerous nodes of varying sizes, some colored in dark blue, light blue, and grey, connected by a web of thin, light grey lines. The nodes are distributed across the frame, with some appearing as large, prominent circles and others as smaller dots. The overall aesthetic is clean and modern, suggesting a theme of connectivity, data, or a network structure.

PRÉPAREZ DES DONNÉES POUR UN ORGANISME DE SANTÉ PUBLIQUE

Voahangy Joan ALEONARD – 04/11/2020

AGENDA DU JOUR



PRÉSENTATION DU
PROJET



PRÉPARATION DES
DONNÉES



ANALYSE DES
DONNÉES



DÉMO



PRÉSENTATION DU PROJET

L'APPEL À PROJET



L'agence « **Santé publique France** » est un établissement public sous tutelle du Ministère de la Santé et qui pour mission d'améliorer et de protéger la santé de la population française dans son ensemble.

Elle lance un **appel à projets** afin de **faciliter l'exploitation des données de santé par ses agents**.



Open Food Facts est une base de données publique (open data) portant sur **les produits alimentaires proposés à l'achat aux consommateurs par pays**.

Cette base de données est gérée par une équipe de volontaires à travers le monde entier, et elle se donne pour objectif de **permettre au consommateur de faire des choix informés**.

LES AXES D'EXPLORATION

LA MISSION

- Explorer la base de données des produits alimentaires d'**Open Food Facts** ;
- Les préparer afin que les agents de l'agence puissent s'appuyer sur les résultats pour les exploiter.

LIVRABLE

Un prototype interactif de l'exploration des données sous forme de page web.

AXES D'EXPLORATION

Notre analyse portera donc la recherche de mesures fiables nous permettant de **démontrer la qualité nutritionnelle d'un produit alimentaire donné.**

CRITERES DE QUALIFICATION D'UNE ALIMENTATION DE QUALITE

Un aliment **moins gras / sucré / salé**, avec **le moins d'additifs** possible,
provenant de l'(agri)culture biologique ou le **moins transformé**, et **sans huile de palme.**



JEU DE DONNÉES : PRÉPARATION

DESCRIPTION DU JEU DE DONNÉES

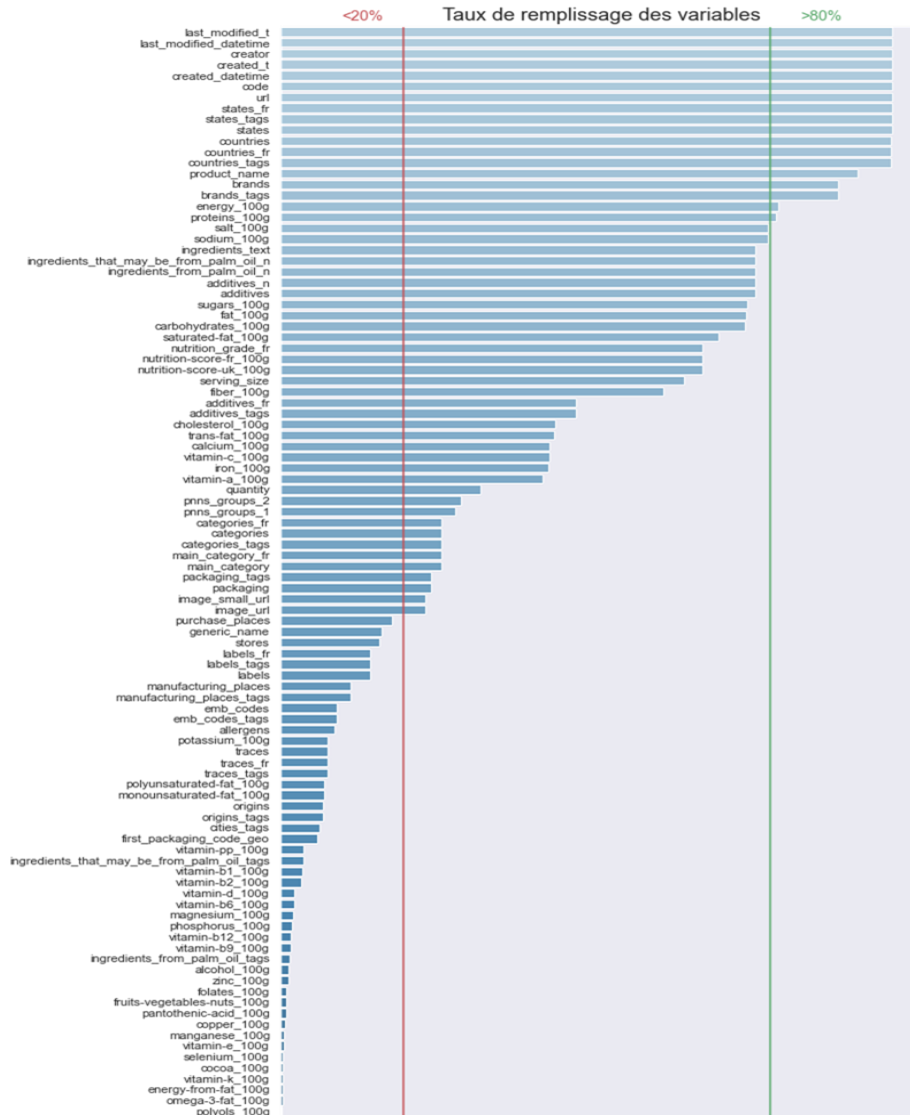
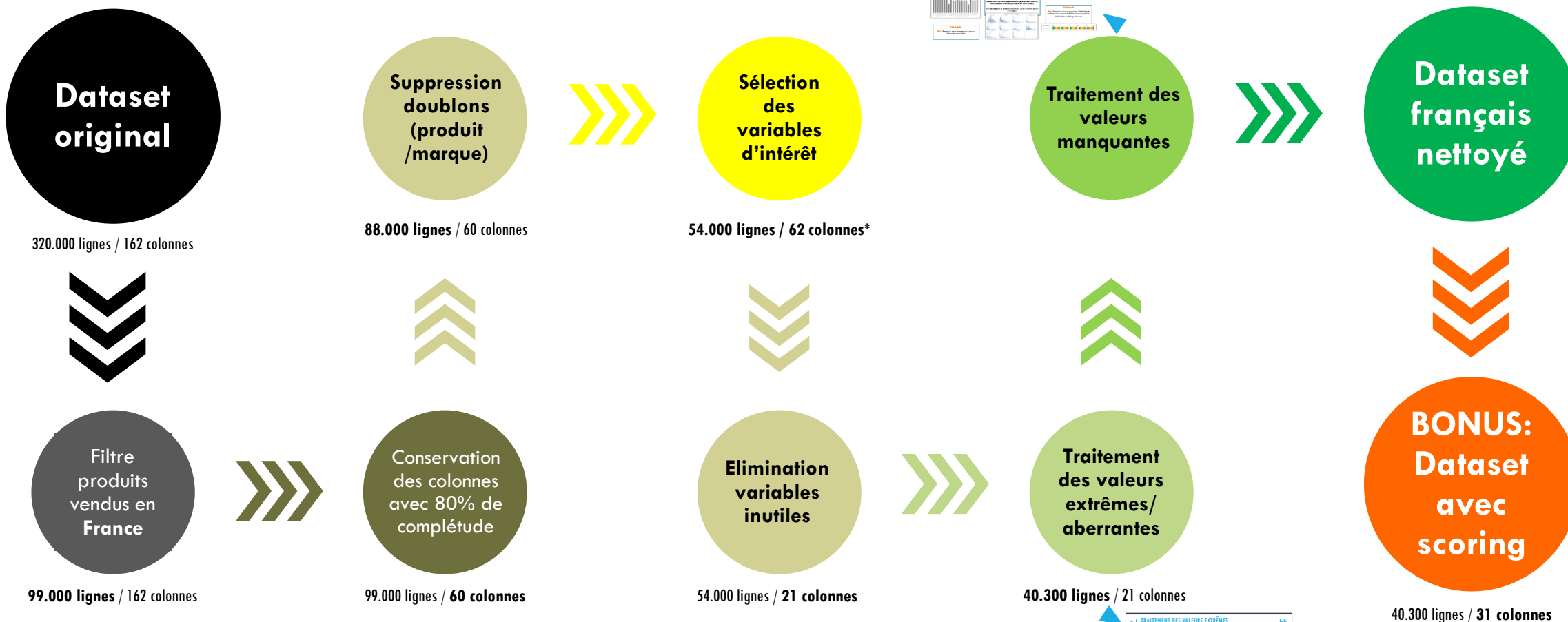


Tableau de synthèse

Indicateurs	Valeurs
Nombre de colonnes	162
Nombre de lignes	320.772
Nombre de variables qualitatives	56
Nombre de variables quantitatives	106
Données manquantes	108 colonnes ont plus de 80% de données manquantes

DÉMARCHE DE PRÉPARATION DES DONNÉES



* Création des variables *produit_bio* grâce aux labels, et *huile_palme* (fusion des 2 variables initiales)

DATASET NETTOYÉ

40.300 lignes et 31 colonnes

10 variables catégorielles

code
nom_produit
marques
categorie_produit
sous_categorie
produit_bio
ingredients
additifs
huile_palme
nutrition_grade

11 variables quantitatives

nutrition_score
energie_kJ_100g
lipides_100g
dt_gras_satures_100g
glucides_100g
dt_sucres_100g
dt_fibres_100g
proteines_100g
sel_100g
sodium_100g
nb_additifs

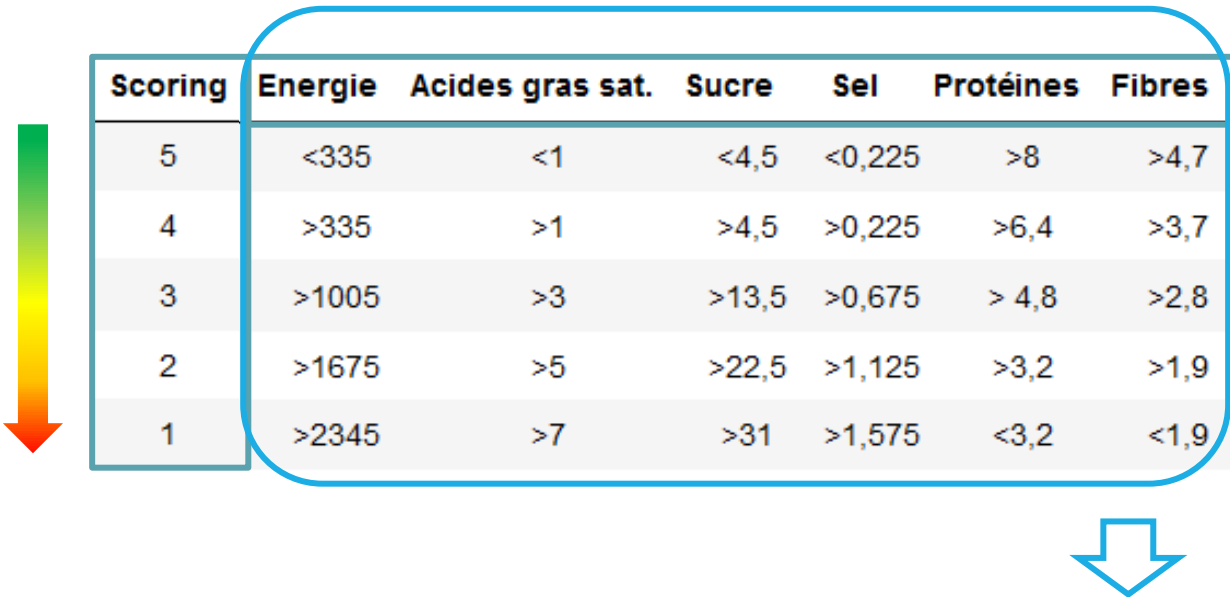
10 variables du scoring

energie_kJ_100g_scoring
dt_gras_satures_100g_scoring
dt_sucres_100g_scoring
sel_100g_scoring
proteines_100g_scoring
dt_fibres_100g_scoring
nutrition_scoring
bio_scoring
palme_scoring
additifs_scoring

DESCRIPTION DU SCORING

Pour aider les agents de Santé publique France à analyser les produits alimentaires à leur disposition, le moyen le plus facile est de leur permettre de comparer ces produits sur la base d'indicateurs communs.

Le principe du **scoring** est d'attribuer une note allant de 1 à 5 à un indicateur donné afin de signaler une appréciation : 5 étant la meilleure appréciation jusqu'à 1, la moins bonne appréciation



Scoring	Energie	Acides gras sat.	Sucre	Sel	Protéines	Fibres	Nutri-score	Bio	Palme	Additifs
5	<335	<1	<4,5	<0,225	>8	>4,7	a	Bio	Sans palme	0
4	>335	>1	>4,5	>0,225	>6,4	>3,7	b	-	Dérivés possibles	>1
3	>1005	>3	>13,5	>0,675	> 4,8	>2,8	c	-	Non spécifié	Non spécifié
2	>1675	>5	>22,5	>1,125	>3,2	>1,9	d	-	Avec palme	>5
1	>2345	>7	>31	>1,575	<3,2	<1,9	e	Non spécifié	Avec palme et dérivés	>10

*cf. **Règlement d'usage du logo "Nutri-score"** de Santé publique France - Version 21 du 16 juin 2020 (page 18 et 19, disponible [ICI](#))

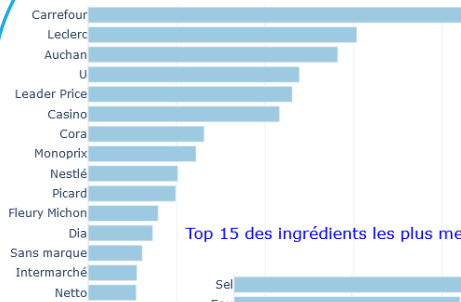


JEU DE DONNÉES : ANALYSE

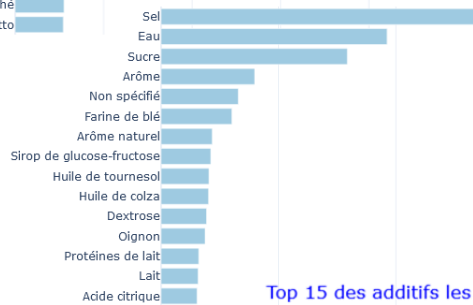
SYNTHÈSE DES VARIABLES CATÉGORIELLES

Les produits

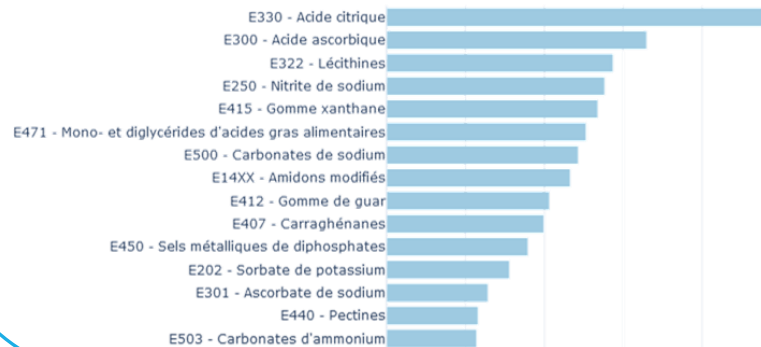
Top 15 des ~10.000 marques



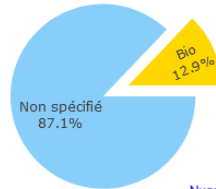
Top 15 des ingrédients les plus mentionnés



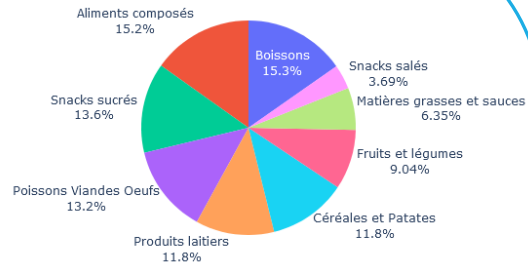
Top 15 des additifs les plus utilisés



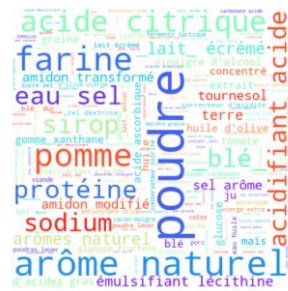
Proportion de produits BIO



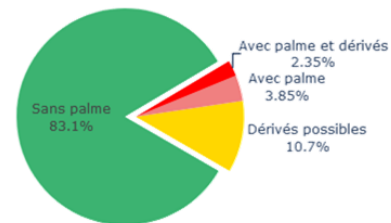
Distribution des 9 catégories de produits



Nuage de mots des ingrédients par ligne de données



Présence d'huile de palme ou ses dérivés



Près de **18.900 produits** disponibles,

répertoriés parmi près de **10.000 marques**,

répartis en **9 catégories principales** et **34 sous-catégories**.

Nous notons que **le sel et le sucre** font partie du top 3 des ingrédients les plus mentionnés lorsqu'ils sont pris isolément.

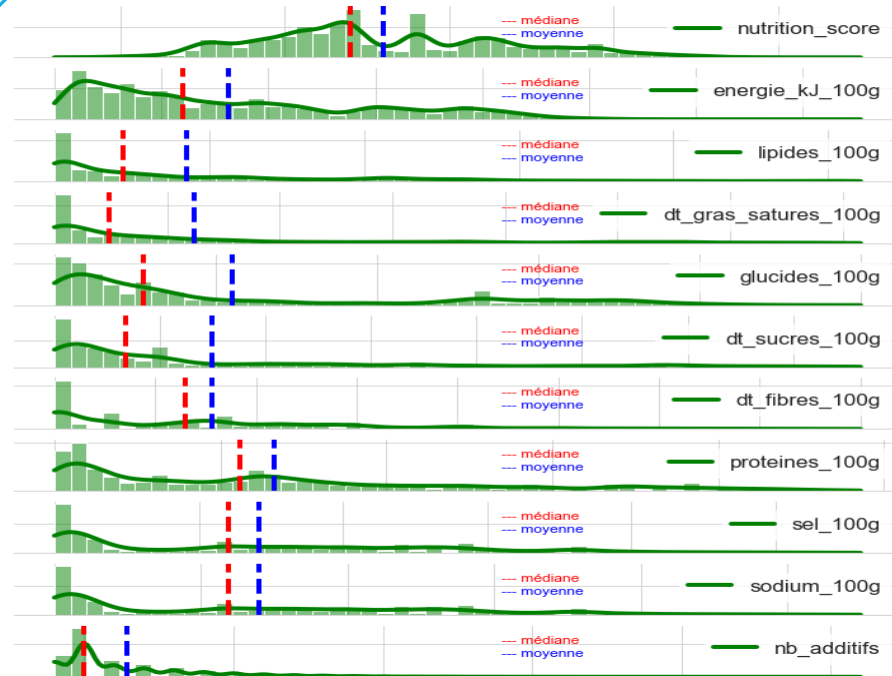
Nous pouvons étudier les **additifs utilisés**,

ou la **présence d'huile de palme** dans les produits lorsque l'information est disponible.

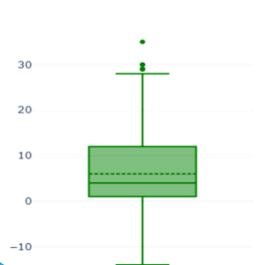
DISTRIBUTION DES VARIABLES QUANTITATIVES

Analyse quantitative

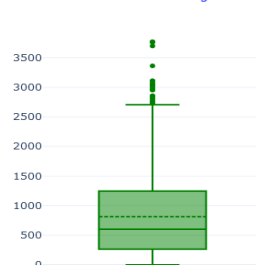
Distribution des données quantitatives



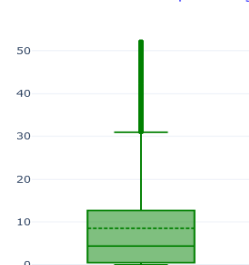
Boîte à moustache : nutrition score



Boîte à moustache : energie kJ 100g



Boîte à moustache : lipides 100g



Pour chacune de ces variables, nous pouvons observer :

- grâce à l'**histogramme**, la **forme** des données ;
- et grâce à la **boîte à moustache**, la **dispersion** des données.

Pour la forme des données, nous notons que :

- la majorité des **distributions sont étalées à droite** ;
- Certaines sont **bimodales**, voire **plurimodales**.

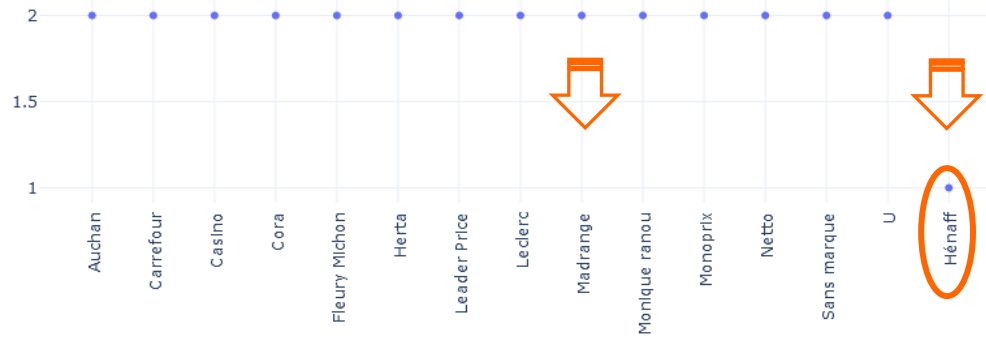
Parmi les mesures de dispersion, nous avons :

- la **médiane** : valeur qui partage en 2 l'étendue des valeurs disponibles (autant de valeurs supérieures qu'inférieures par rapport à elle) ;
- le **quartile Q1** : limite haute des 25% des valeurs les plus petites ;
- le **quartile Q3** : limite basse des 25% des valeurs les plus grandes ;
- les **valeurs minimum et maximum** ;
- de même que les **outliers** (valeurs atypiques / extrêmes) représentées par les points, souvent situés au-delà d'une limite supérieure (*l'upper fence* en anglais).

UTILISATION CONCRÈTE DU SCORING

Comparatif produit

Taux moyen de sel par marque dans les Charcuteries

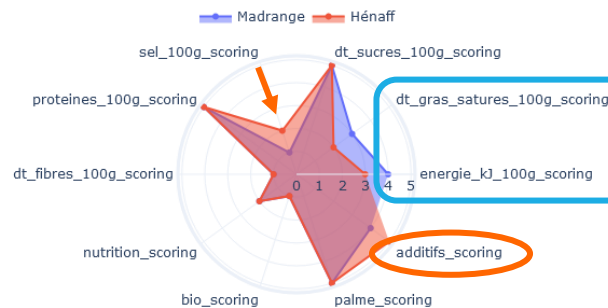


Nous partons d'un constat : parmi le **top 15 des marques de produits de Charcuterie**, la marque **Hénaff** se démarque singulièrement des autres sur le **taux moyen de sel**.

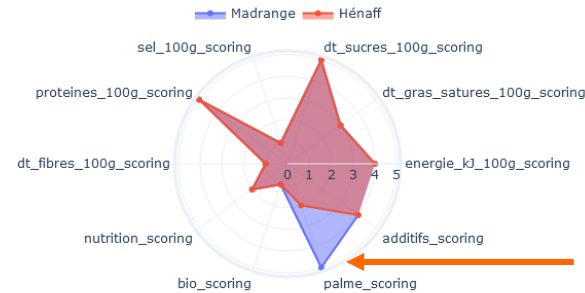
Nous comparons la marque **Hénaff** à une autre — **Madrance**, sur **l'ensemble des indicateurs du scoring** :

- **Hénaff** se démarque toujours sur son score **sel** et on note qu'il n'utilise pas d'additifs ;
- **Madrance** affiche néanmoins de meilleurs scores sur les acides gras saturés et l'apport énergétique.

Comparaison de 2 marques dans Charcuteries



Comparaison de 2 marques dans Charcuteries Mousse de Foie



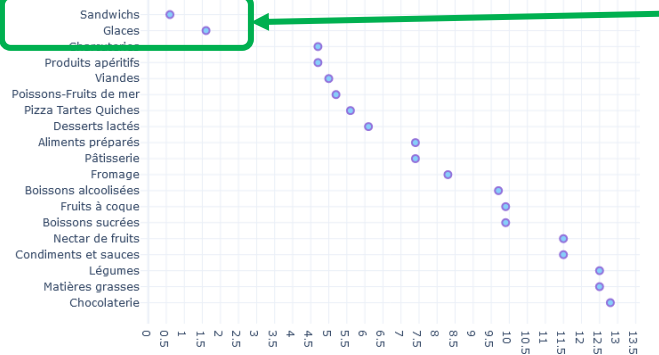
Au final, sur le produit **Mousse de foie**, la marque **Madrance** l'emporte sur un indicateur : le score de l'huile de palme.

Le scoring peut être utilisé comme **outil comparatif des produits** car il permet d'avoir une vision plus fine.

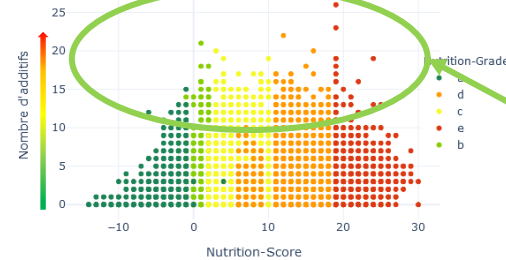
DES DONNÉES À APPROFONDIR

Des sujets à creuser ?

Sous-catégorie avec une proportion de produit bio inférieure à la moyenne



Le Nutrition-score en présence d'additifs

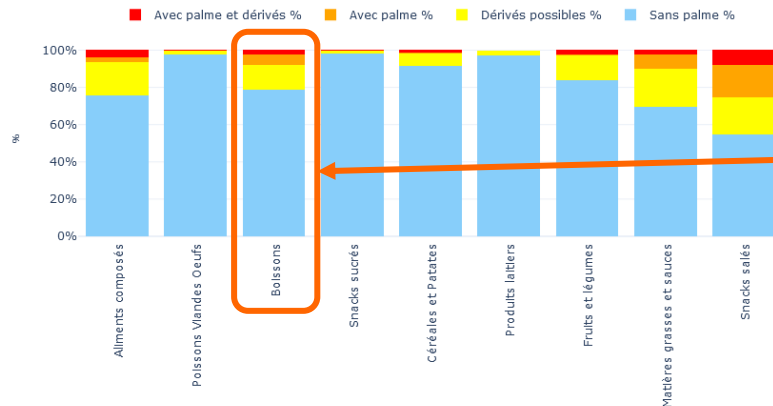


Pourquoi spécifiquement sur les SANDWICHS et les GLACES, il y a moins de produits bio, alors que les composants essentiels peuvent l'être ?

Le nombre très élevé d'additifs (sans notion de nocivité) n'influence pas le nutri-score : un nouveau projet (logo) à mettre en place ?

**Economie.gouv.fr : conditions et modalités d'utilisation des additifs alimentaires*

Absence / Présente d'huile de palme par catégories de produits



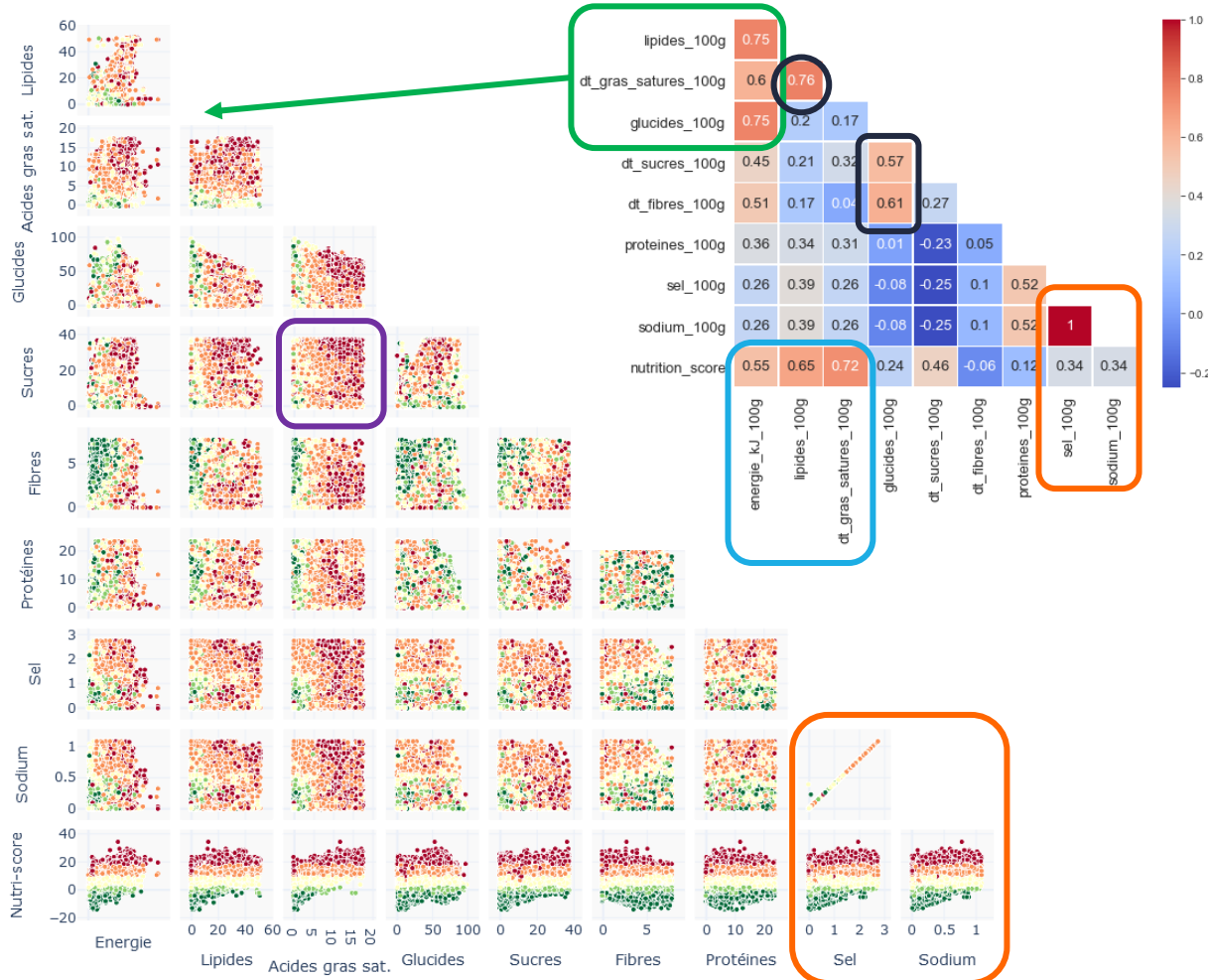
Est-il normal de trouver de l'huile de palme dans les boissons ?

Note : en orange et rouge, le mot palme est mentionné

CORRÉLATION ENTRE LES VARIABLES QUANTI

Matrice de corrélation de Pearson

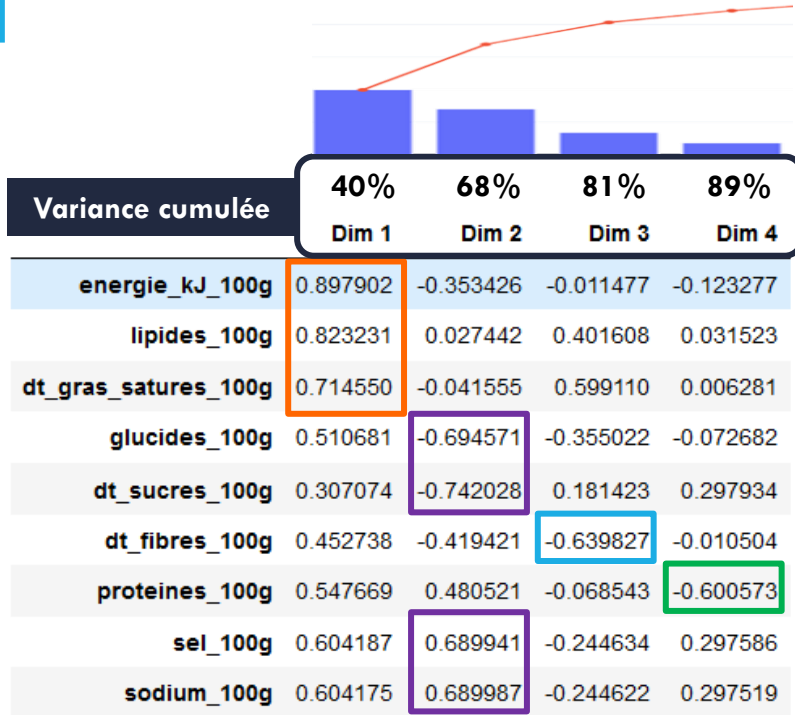
Analyse des variables quantitatives



En se basant sur la **Carte de chaleur** et la **Matrice de nuage de points**, on peut dire que :

- **Le sel et le sodium** pourraient être synthétisés en une seule variable ;
- **L'énergie** est fortement corrélée aux **lipides, acides gras saturés et glucides** ;
- Sans grande surprise, les macronutriments sont corrélés aux nutriments enfants : les **acides gras saturés** aux **lipides**, tout comme le **sucres** et les **fibres** aux **glucides**.
- **Le nutri-score** est corrélé à **l'énergie, les lipides et les acides gras saturés**
- La présence conjointe d'**acides gras saturés** et de **sucres** semble influencer de manière négative le **nutri-grade**.

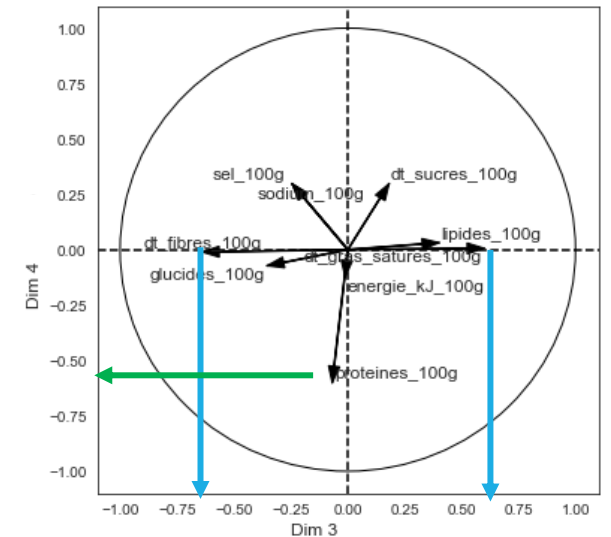
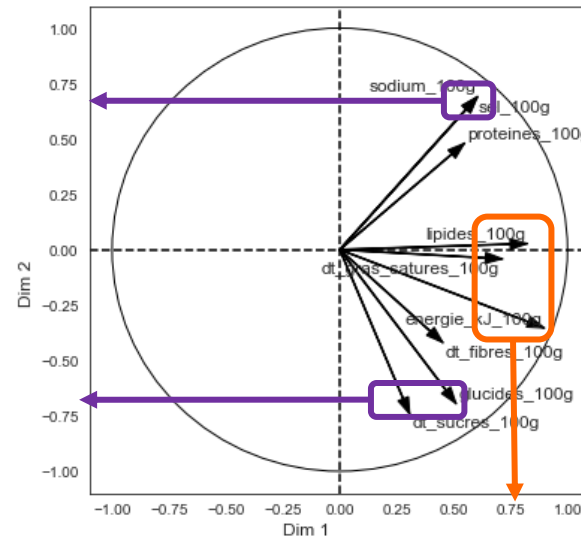
ANALYSE EN COMPOSANTES PRINCIPALES (ACP)



Représentativité :

Les 4 premières composantes (ou dimensions) capturent **89% de la variance des 9 variables initiales**.

Cercles de corrélation – Dimensions 1 à 4



Projection sur les axes des cercles de corrélations :

Voici les variables les mieux représentées par composante principale:

Dim 1 : énergie, lipides et acides gras saturés

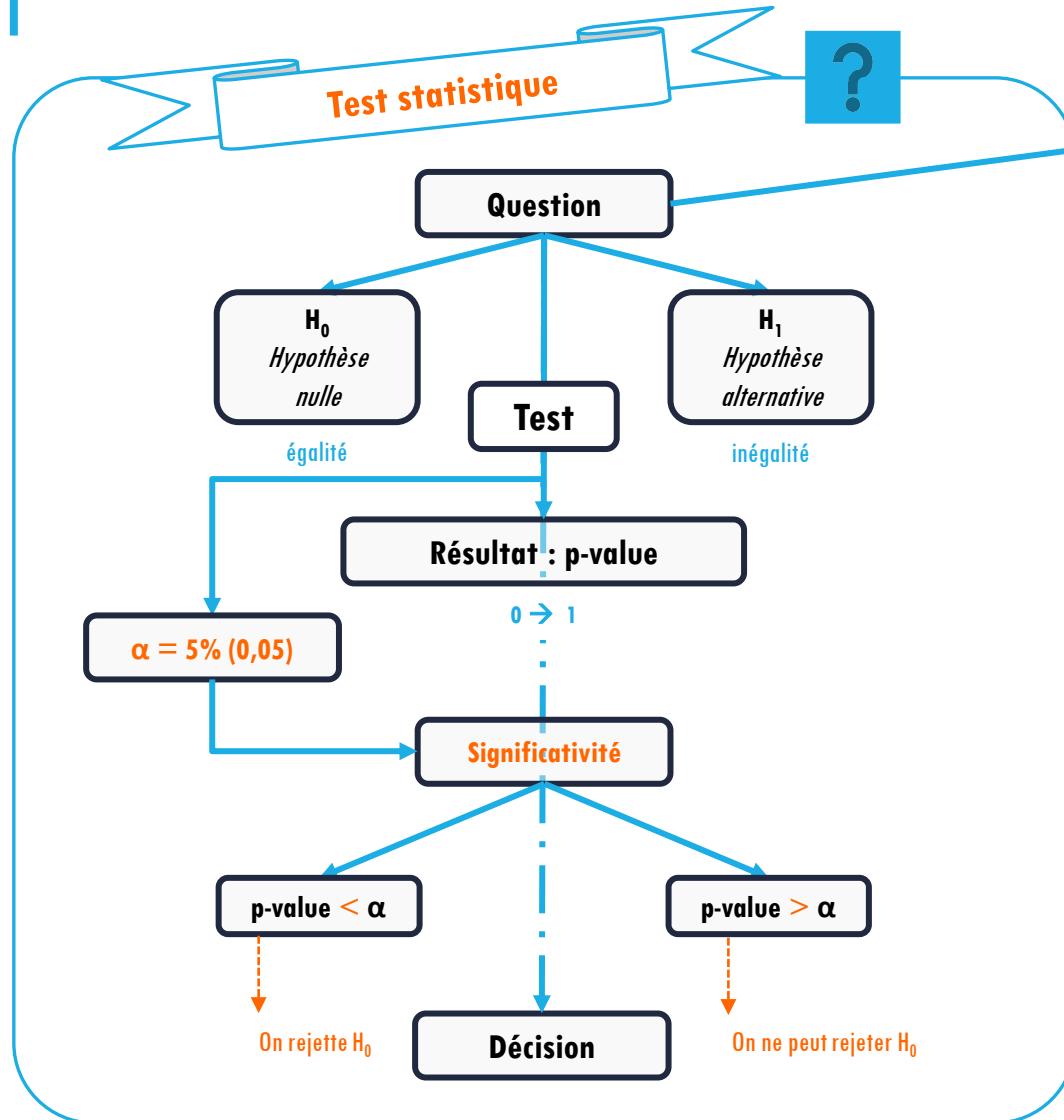
Dim 2 : glucides / sucre ET sel / sodium

Dim 3 : fibres ET acides gras saturés

Dim 4 : protéines

89% de la variance des 9 variables initiales capturée sur 4 nouveaux axes d'analyse (97,7% sur 6 nouveaux axes).

VÉRIFICATION D'HYPOTHÈSES



Peut-on affirmer qu'en moyenne les fruits sont plus sains que les jus de fruits ?

H₀

Il n'y a pas de différence entre fruits et jus de fruits.

H₁

Les fruits sont différents des jus de fruits.

α

Le seuil de significativité est fixé à **5%**

Test utilisé

Test de STUDENT : comparaison des moyennes des nutri-scores sur les fruits et jus de fruits

Test de Student :

Variance Fruits : 1.803

Variance Jus de fruits : 6.475

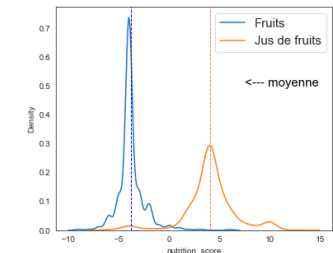
On a une p-value égale à **0.000** (Stat=-87.852)

L'hypothèse nulle est rejetée ; probablement des distributions différentes

p-value < α

La p-value est à 0, H₀ est rejetée.
En moyenne, les fruits sont **significativement** différents des jus de fruits

En moyenne, les fruits sont **significativement plus sains que** les jus de fruits





DÉMO : PAGE WEB

UNE PAGE WEB POUR PARTAGER LES ANALYSES

Auteur : Voahangy Joan Aléonard
 Date dernière version : 01/11/2020
 Librairies utilisées : Pandas, Numpy, Matplotlib, Seaborn, Plotly, Scipy, Skikit-learn, mxtend, ipywidgets, worldcloud, Collections



Santé publique France et Open Food Facts Analyse exploratoire des données

Dans le cadre des politiques de santé publique française, l'agence **Santé publique France** a pour mission de **"diminuer les risques liés aux problématiques nutritionnelles"** en :

- surveillant l'évolution des comportements et fournissant des informations sur la situation nutritionnelle en France ;
- promouvant les comportements favorables à la santé en matière d'alimentation.

Parmi les principes d'une alimentation de qualité, nous retrouvons :

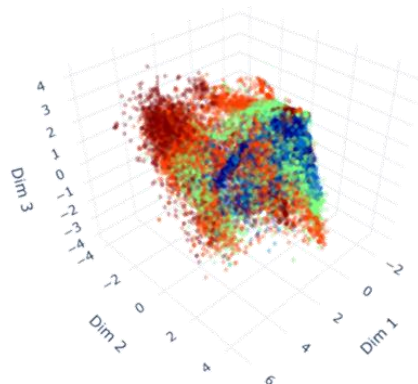
- principalement, la réduction de la consommation de sel, de sucre et de gras ;
- la consommation de produits le moins transformé possible ;
- la consommation de produits bio, avec le moins d'additifs possibles ;
- etc.

Ce notebook a pour objectif de **mettre en exergue les éléments se conformant à ces principes**.

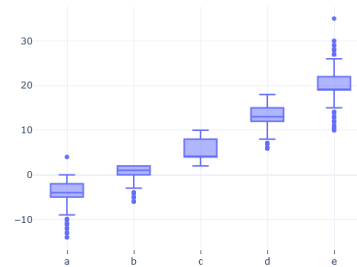
Dans cette analyse exploratoire des données, nous allons :

- visualiser graphiquement le jeu de données nettoyé d'Open Food Facts, afin de mieux l'appréhender ;
- rechercher les relations entre les variables (ou définir des hypothèses que l'on va vérifier) ;
- et enfin, donner des pistes de réflexion à Santé publique France sur les actions pouvant être menées, dans le cas

Importation des librairies et chargement du Dataset nettoyé.



Nutri-Score par Nutri-grade



Dans cette partie, nous analysons principalement les variables 2 à 2, voire un peu plus.

2.1 - Le Nutri-score respecte-t-il les différents seuils de calcul de Santé publique France?

De manière générale, les seuils du nutrition-score se présentent comme suit :



Partie 2 - Analyse bivariée et multivariée

Les produits 'malsains' du grade A :

nom_produit	categorie_produit	sous_categorie	nutrition_score
Sémillante arômes naturels Citron	Boissons	Boissons sucrées	-4.0
Eau de Source Sagueno	Boissons	Boissons non sucrées	0.0
Eau de source de la Doye	Boissons	Boissons non sucrées	0.0
eau minérale naturelle des Vosges du nord - légère	Boissons	Boissons non sucrées	0.0
Eau minérale naturelle gazeuse	Boissons	Boissons non sucrées	0.0

Les produits 'sains' du grade B :

nom_produit	categorie_produit	sous_categorie	nutrition_score
Eau de coco	Boissons	Jus de fruits	-6.0
Jus de citron	Boissons	Jus de fruits	-6.0
Citrons fraîchement pressés	Boissons	Jus de fruits	-6.0
Jus de tomate 100% pur fruit pressé Sélection	Boissons	Boissons non sucrées	-5.0
100 % Pur Jus Tomate	Boissons	Boissons non sucrées	-5.0

Les produits 'sains' du grade D :

nom_produit	categorie_produit	sous_categorie	nutrition_score
Jus de pommes et de mangues pressé non à base de concentré	Boissons	Jus de fruits	6.0
100% Juice Whispers Of Summer	Boissons	Jus de fruits	6.0
Pur Jus Clémentine	Boissons	Jus de fruits	6.0
Pur Jus Orange Clémentine Raisin	Boissons	Jus de fruits	6.0
Concentré de Citron vert pour boisson et assaisonnement	Boissons	Boissons non sucrées	6.0





SYNTHÈSE

QUE POUVONS-NOUS EN CONCLURE ?

Cette analyse exploratoire touche à sa fin.

Nous avons pu constater qu'il y a **moult** **possibilités d'exploration**, avec des **données aussi riches** que la base de produits alimentaires d'Open Food Facts.

Néanmoins, toutes ces analyses doivent répondre à un ou des **besoins métier exprimés par les agents de Santé publique France**, afin que les enseignements tirés soient pertinents.

Un **recueil des besoins** devrait donc être mené auprès des futurs utilisateurs de ces données.

The background of the slide is a complex network diagram. It consists of numerous small grey dots connected by thin grey lines, forming a web-like structure. Several larger circles are also present: a large white circle with a dark blue center at the top center, a large blue circle at the bottom left, and a large grey circle at the bottom center. There are also smaller blue and dark blue circles scattered throughout the network.

QUESTIONS / RÉPONSES





ANNEXES

TRAITEMENT DES VALEURS EXTRÊMES (VOIRE ABERRANTES)



Nettoyage métier

= suppression des valeurs aberrantes

Nutriments = substances composant les aliments pour 100g (ou 100ml) de portion comestible (il en existe des centaines)

Macronutriments = nutriments contribuant à l'énergie (lipides, glucides, protéines)

Nutriment enfant = nutriment constituant les macronutriments : les **acides gras** pour les lipides, le **sucre** pour les glucides

Règle 1 : Valeur d'un nutriment > 0 et ≤ 100

Règle 2 : Energie ≤ 3.800 kJ

Règle 3 : \sum macronutriments ≤ 100

Règle 4 : valeur nutriment enfant \leq valeur macronutriment

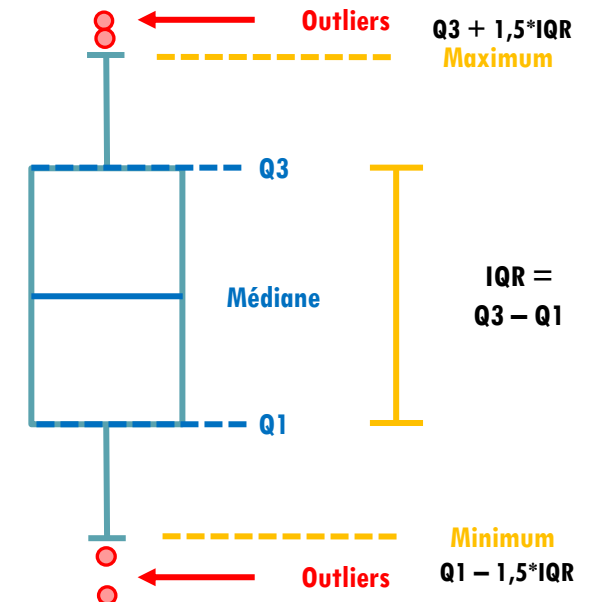
Règle 5 : sel $\leq 2,5$ x sodium

Nettoyage statistique

= exclusion des valeurs extrêmes

** Appliqué uniquement sur les nutriments*

Méthode utilisée = Ecart interquartile (IQR)



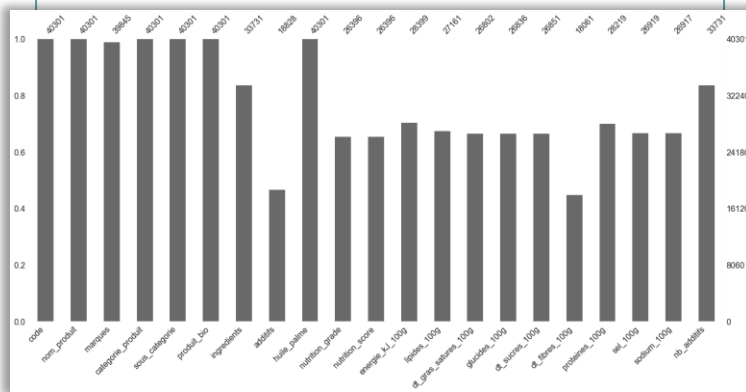
TRAITEMENT DES VALEURS MANQUANTES : IMPUTATION



1) Variables catégorielles

marques, catégorie produit, sous-catégorie, ingrédients et additifs

Règle : Remplacer les valeurs manquantes par « **Non spécifié** »



2) Nb d'additifs

Règle : Remplacer les valeurs manquantes par -1 pour les distinguer des valeurs à zéro

3) Nutriments et nutri-score

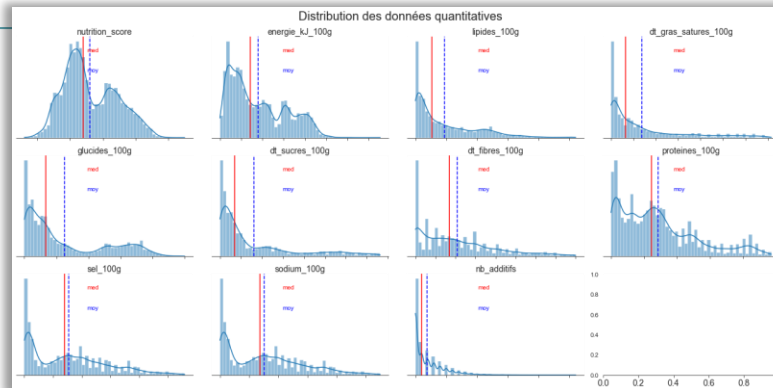
lipides/gras saturés, glucides, sucre, fibre, protéines, sel/sodium, nutri-score

Règle : Remplacer les valeurs manquantes par la **médiane****, pondérée par les « sous-catégories »

Explication :

Médiane car distributions pas normales, désaxées (majoritairement étalées vers la droite), moyenne > médiane (donc sensible aux valeurs extrêmes)

Par sous-catégorie, car agrégation très différente ; par ex. les matières grasses et les légumes



4) Energie

Règle 1 : Remplacer les valeurs à zéro par « **NaN** »

Règle 2 : Remplacer les valeurs manquantes par calcul avec les macronutriments

$$E = 38 * \text{lipides} + 17 * (\text{glucides} + \text{protéines})$$

5) Nutri-grade

Règle : Remplacer les valeurs manquantes par l'application de conditions sur les valeurs du Nutri-score en respectant les limites fixées sur l'image ci-dessous.





Ce document a été produit dans le cadre de la soutenance du projet n°3 du parcours Ingénieur IA d'OpenClassrooms :
« Préparez des données pour un organisme de santé publique »

Mentor : Thierno DIOP
Evalueur : Moussa CAMARA

