

Projet 3 :

**Concevez une application
au service de la santé
publique**

Nom : TRABIS

Prénom : Mohamed

Table des matières

1. Introduction

2. Préparation des données

a) Évaluation et découverte

b) Nettoyage et validation

3. Analyse univariée

4. Analyse bivariée et multivariée

5. Application

6. Conclusions

Introduction

Introduction

● Contexte :

- L'agence "[Santé publique France](#)" a lancé un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation.
- Pour y participer il faut proposer une idée d'application en utilisant Le jeu de données Open Food Fact qui est disponible sur [le site officiel](#).



Introduction

● Mission :

Le projet consiste à effectuer les différentes étapes ci-dessous :

- 1) Traiter le jeu de données afin de repérer des variables pertinentes pour les traitements à venir. Automatiser ces traitements pour éviter de répéter ces opérations.
- 2) Tout au long de l'analyse, produire des visualisations afin de mieux comprendre les données. Effectuer une analyse univariée pour chaque variable intéressante, afin de synthétiser son comportement.
- 3) Confirmer ou infirmer les hypothèses à l'aide d'une analyse multi variée. Effectuer les tests statistiques appropriés pour vérifier la significativité des résultats.
- 4) Élaborer une idée d'application. Identifier des arguments justifiant la faisabilité (ou non) de l'application à partir des données Open Food Facts.

Préparation des données

Préparation des données - Évaluation et découverte des données

- La base de données Open Food Facts est disponible sous licence Open Database Licence.
- Les données pour tous les produits, ou une sélection de produits, peuvent être téléchargés au format CSV :

```
en.openfoodfacts.org.products.csv
```

- Le fichier CSV est volumineux (plus de 4 Go), il contient :
 - 1 951 229 Lignes et 186 colonnes (Chaque ligne correspond à un produit)
- Le fichier est enrichi par les utilisateurs via l'application **OpenFoodFacts** ([Android](#) et [iOS](#)). L'application permet aux utilisateurs de scanner le code-barre d'un produit pour accéder à ses informations mais aussi de compléter les informations et les photos pour des produits manquants.

Préparation des données – Nettoyage et validation des données

- Les étapes effectuées pour nettoyer les données :

- Supprimer les colonnes avec 80% de valeurs manquantes.

- Conserver que les produits :

- *Vendus en « France ».*

- *Avec le « Nutri-Score » renseigné.*

- *Avec le « Groupe Nova » renseigné.*

- **Remarque :**

Suite à ce nettoyage nous avons 128 685 lignes au lieu de 1 951 229.

Préparation des données – Nettoyage et validation des données

- Détecter les valeurs aberrantes :

Une valeur aberrante est une valeur extrême, anormalement différente de la distribution d'une variable. En d'autres termes, la valeur de cette observation diffère grandement des autres valeurs de la même variable.

Ci-dessous un exemple :



	nova_group	carbohydrates_100g	energy_kcal_100g	energy_100g	sugars_100g	fat_100g	saturated_fat_100g	salt_100g
count	130371.00	129964.00	112063.00	130034.00	130024.00	130024.00	130027.00	130066.00
mean	3.38	27.38	282.85	1178.79	13.28	14.28	5.63	1.07
std	0.97	533.16	4066.19	22223.26	77.10	82.12	9.97	38.89
min	1.00	-0.50	0.00	0.00	-0.50	0.00	0.00	0.00
25%	3.00	3.20	114.00	469.00	0.90	1.50	0.30	0.09
50%	4.00	14.00	248.00	1020.00	3.70	8.00	2.20	0.60
75%	4.00	50.00	396.00	1644.00	18.50	22.00	8.00	1.27
max	4.00	192000.00	1360000.00	8010000.00	27000.00	29000.00	2000.00	14000.00

- **Remarque :** Par exemple nous constatons que la valeur maximale des calories pour un produit est de 1 360 000 Kcal pour 100g, ce qui est aberrant.

Préparation des données – Nettoyage et validation des données

- Supprimer les valeurs aberrantes de la trame des données des colonnes suivantes :

- « *energy_100g* » : Supprimer les valeurs qui sont inférieure au dernier centile.
- Pour les colonnes qui contiennent des valeurs nutritives pour 100g, nous conserverons les valeurs entre 0g et 100g.
- Remplacer les valeurs nutritives « nan » par 0.

- **Remarque :** Nous constatons que le taux de valeurs manquantes de la colonne «*energy-kcal_100g* » est de 13.87 %.

En effet la colonne «*energy_100g* » a un taux de valeurs manquantes de 0%, en sachant que **1 kJ = 0,2388 kcal**, nous recalculons les valeurs de cette colonne pour éviter toutes incohérences :

```
df_food_fr['energy_kcal_100g'] = np.around(df_food_fr['energy_100g']*0.2388)
```

Préparation des données – Nettoyage et validation des données

- Ci-dessous le résultat suite à ce nettoyage :

	nova_group	carbohydrates_100g	energy_kcal_100g	energy_100g	sugars_100g	fat_100g	saturated_fat_100g	salt_100g
count	128685.00	128685.00	110840.00	128685.00	128685.00	128685.00	128685.00	128685.00
mean	3.40	26.11	263.73	1089.48	13.19	13.32	5.44	0.97
std	0.96	26.26	174.51	720.62	18.54	15.00	7.70	2.13
min	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	3.00	3.40	112.00	465.00	0.90	1.40	0.30	0.10
50%	4.00	14.00	245.00	1008.00	3.70	7.90	2.10	0.60
75%	4.00	50.00	392.00	1625.00	19.00	21.80	7.70	1.28
max	4.00	100.00	2733.00	3109.00	100.00	100.00	81.70	100.00

Analyse univariée

Analyse univariée

- L'exploration de données est une étape importante dans le processus de l'analyse des données.

L'analyse univariée permet d'explorer des différentes variables importantes à la fois. Cette analyse se base sur les statistiques descriptives. Ces dernières permettent de tirer des indications concises sur une caractéristique donnée. Parmi ces indicateurs, on retrouve la moyenne, la médiane ainsi que les mesures de dispersion de données.

- L'analyse univariée va nous permettre d'explorer les variables suivantes :
 - ✓ *Nutri-score*
 - ✓ *Eco-score*
 - ✓ Classification NOVA
 - ✓ *Sucre*

Analyse univariée - Nutri-Score

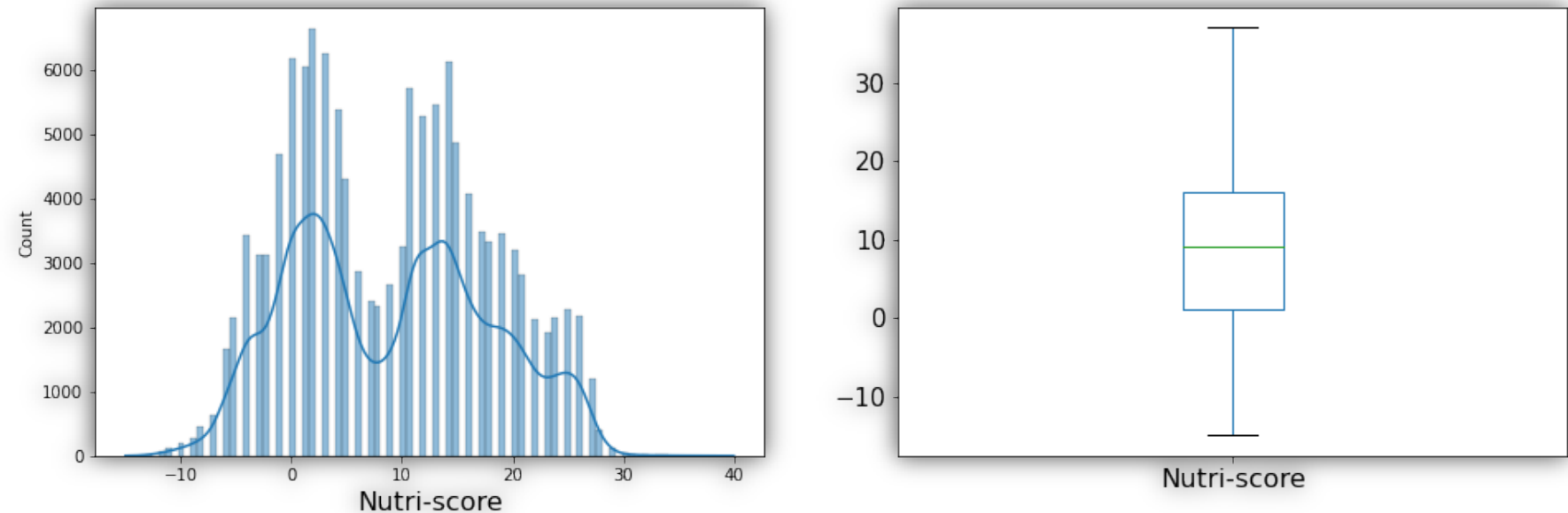
- **Le Nutri-Score** : Il permet sur la base de la composition de l'aliment de donner une valeur unique d'estimation de la valeur nutritionnelle de l'aliment, sur une échelle ordinaire continue allant de -15 (meilleure qualité nutritionnelle) à +40 (moins bonne qualité nutritionnelle).
- Le Nutri-Score est une échelle graphique qui scinde le score nutritionnel en 5 classes (A, B, C, D, E), exprimées par une couleur associée à une lettre, et vise à faciliter la visibilité, la lisibilité, et la compréhension de la valeur nutritionnelle par le consommateur, ci-dessous les 5 logos adaptés à la qualité nutritionnelle de chaque produit :



Analyse univariée - Nutri-Score

- Ci-dessous le graphique représentant le Nutri-score des aliments :

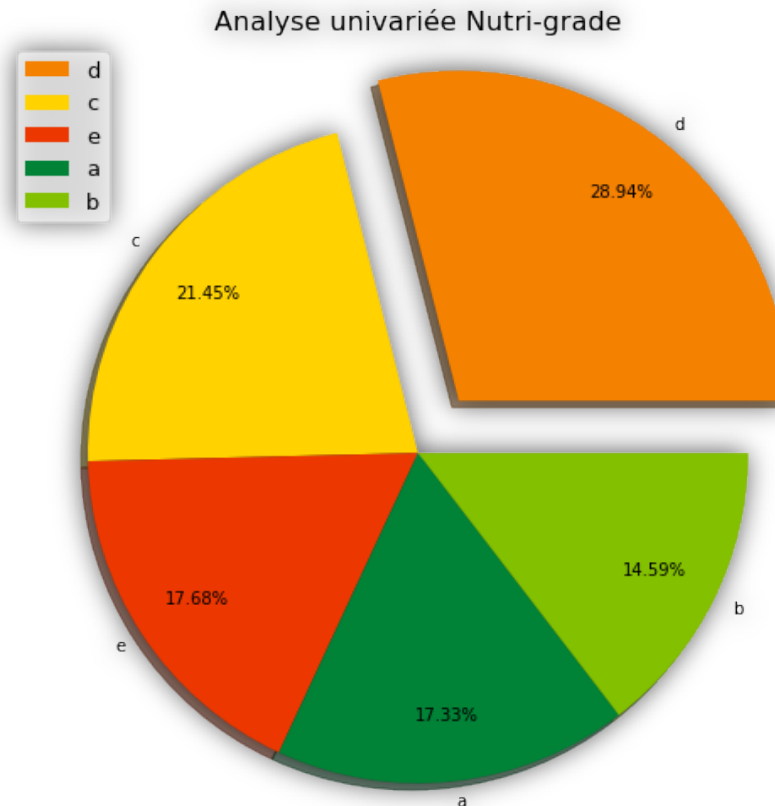
Analyse univariée Nutri-score



- Remarque : La valeur médiane du « Nutri-score » est d'environ 9, la majorité des produits ont un score entre -10 et 28.

Analyse univariée - Nutri-grade

- Ci-dessous le graphique camembert représentant le Nutri-grade :



- Remarque : La majorité des produits ont un Nutri-score « d », « c » et « e », ce qui représente environ 70% des aliments, et les aliments avec un Nutri-score « a » ne représentent que 15%.

Analyse univariée - Eco-score

- **L'Eco-score :**

L'Eco-score est un indicateur représentant l'impact environnemental des produits alimentaires. Il classe les produits en 5 catégories (A, B, C, D, E), de l'impact le plus faible, à l'impact le plus élevé.

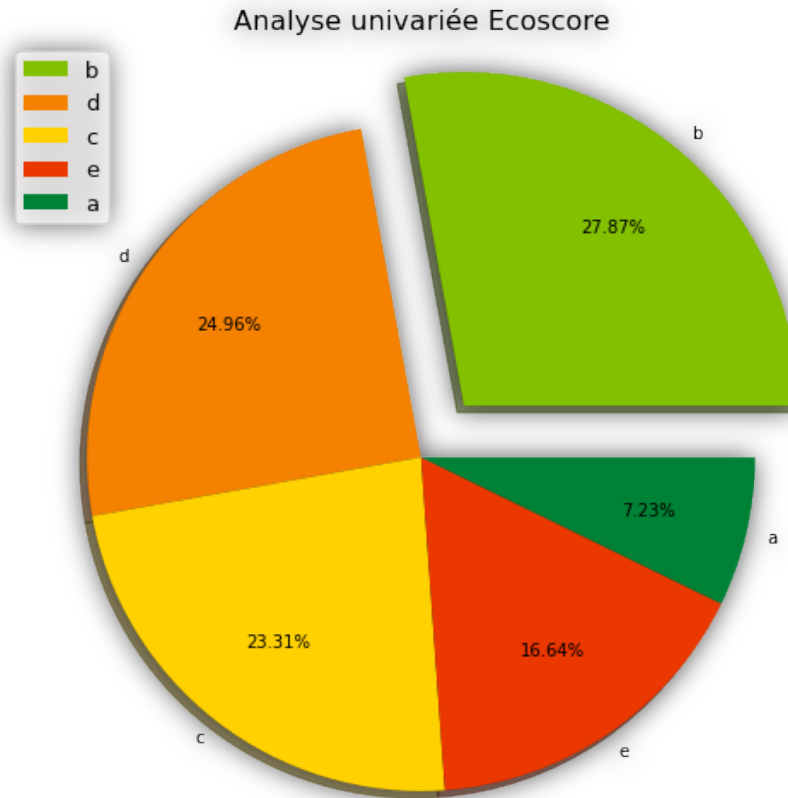


- **Objectif de la démarche :**

Le point de départ de cette démarche est d'avoir une information éclairée sur les impacts environnementaux des produits consommés. L'ambition de l'Eco-score est d'être un outil d'aide à la décision afin de guider nos choix alimentaires vers un mode de consommation plus durable.

Analyse univariée - Eco-score

- Ci-dessous le graphique camembert représentant l'Eco-score :



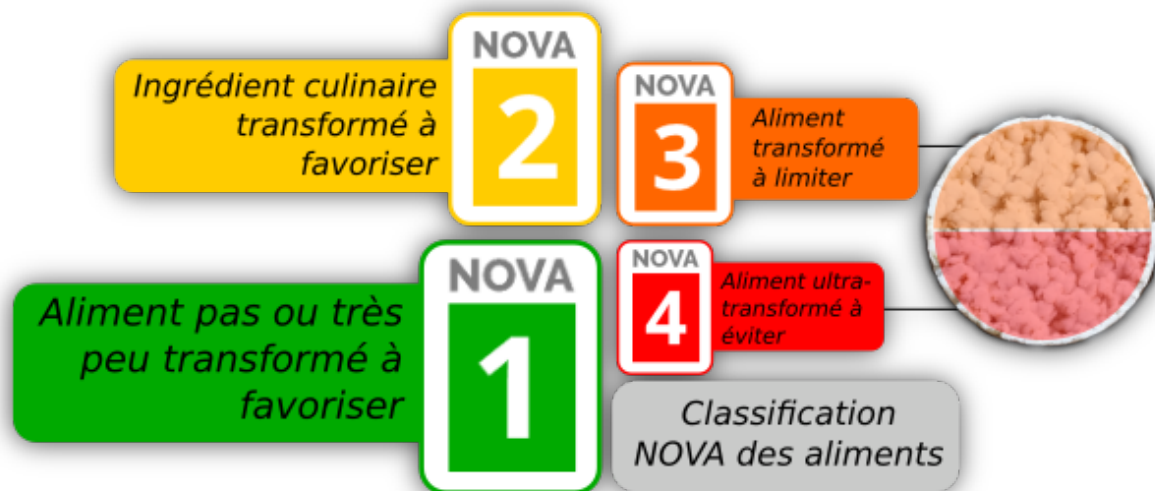
- Remarque : L'eco-score le plus répondu est le « b » (28.11%), par contre l'ecos-core « a » ne représente que 7.25%.

Analyse univariée - Classification NOVA

● La Classification NOVA :

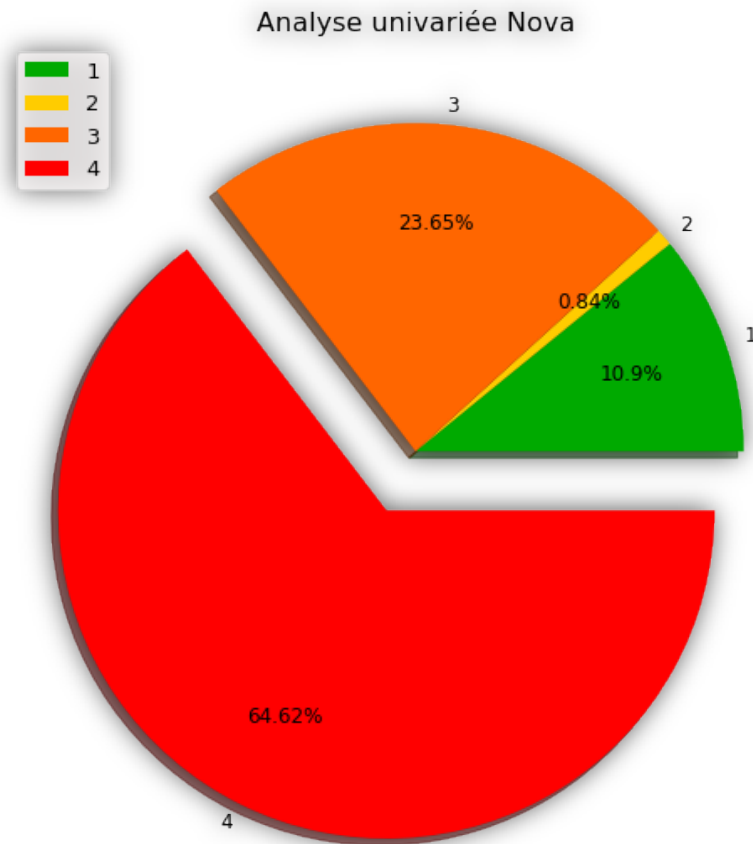
La classification NOVA est une répartition des aliments en quatre groupes en fonction du degré de transformation des matières dont ils sont constitués :

- Groupe 1 : Aliments peu ou non transformés.
- Groupe 2 : Ingrédients culinaires transformés.
- Groupe 3 : Aliments transformés.
- Groupe 4 : Aliments ultra transformés.



Analyse univariée - Classification NOVA

- Ci-dessous le graphique représentant la classification NOVA des aliments :

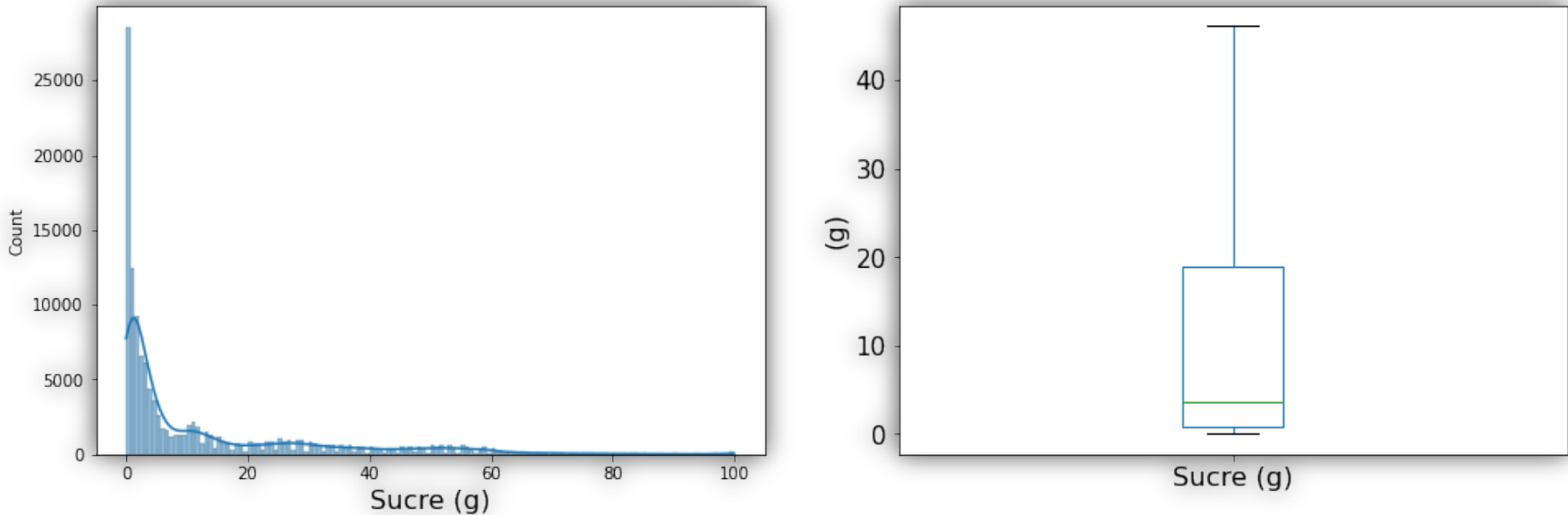


- Remarque : On constate que les aliments ultra transformés (Groupe 4) représentent la majorité des aliments (62.83%), suivi des aliments transformés (Groupe 3).

Analyse univariée - Sucre

- Ci-dessous le graphique représentant la quantité du sucre dans les aliments (Pour 100g):

Analyse univariée - Sucre

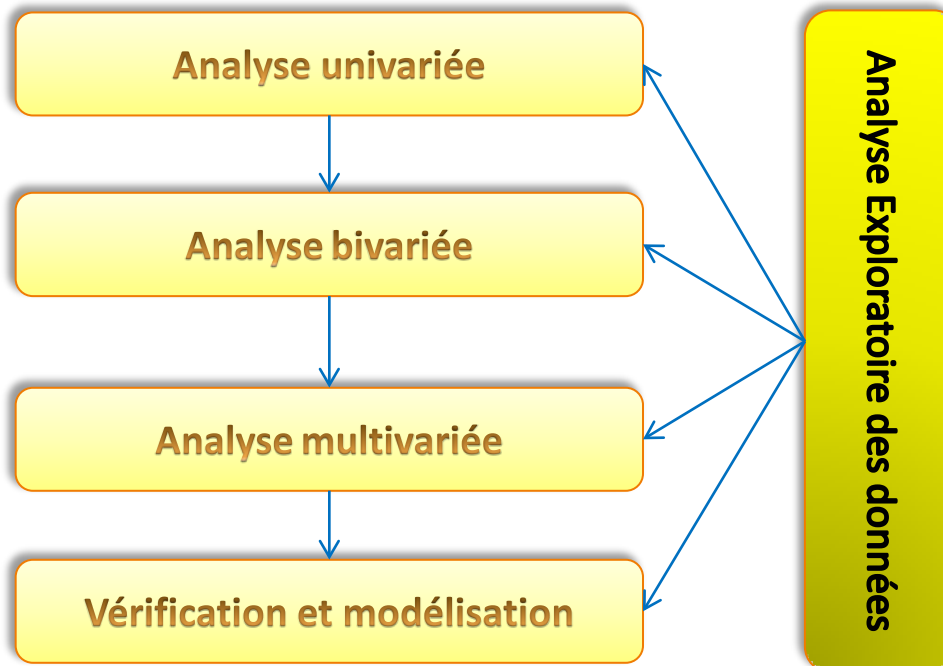


- Remarque : La valeur médiane de la quantité du sucre dans les aliments est d'environ 4g, et la majorité des aliments ont un taux de sucre entre 0 et 10g.

Analyse bivariée et multivariée

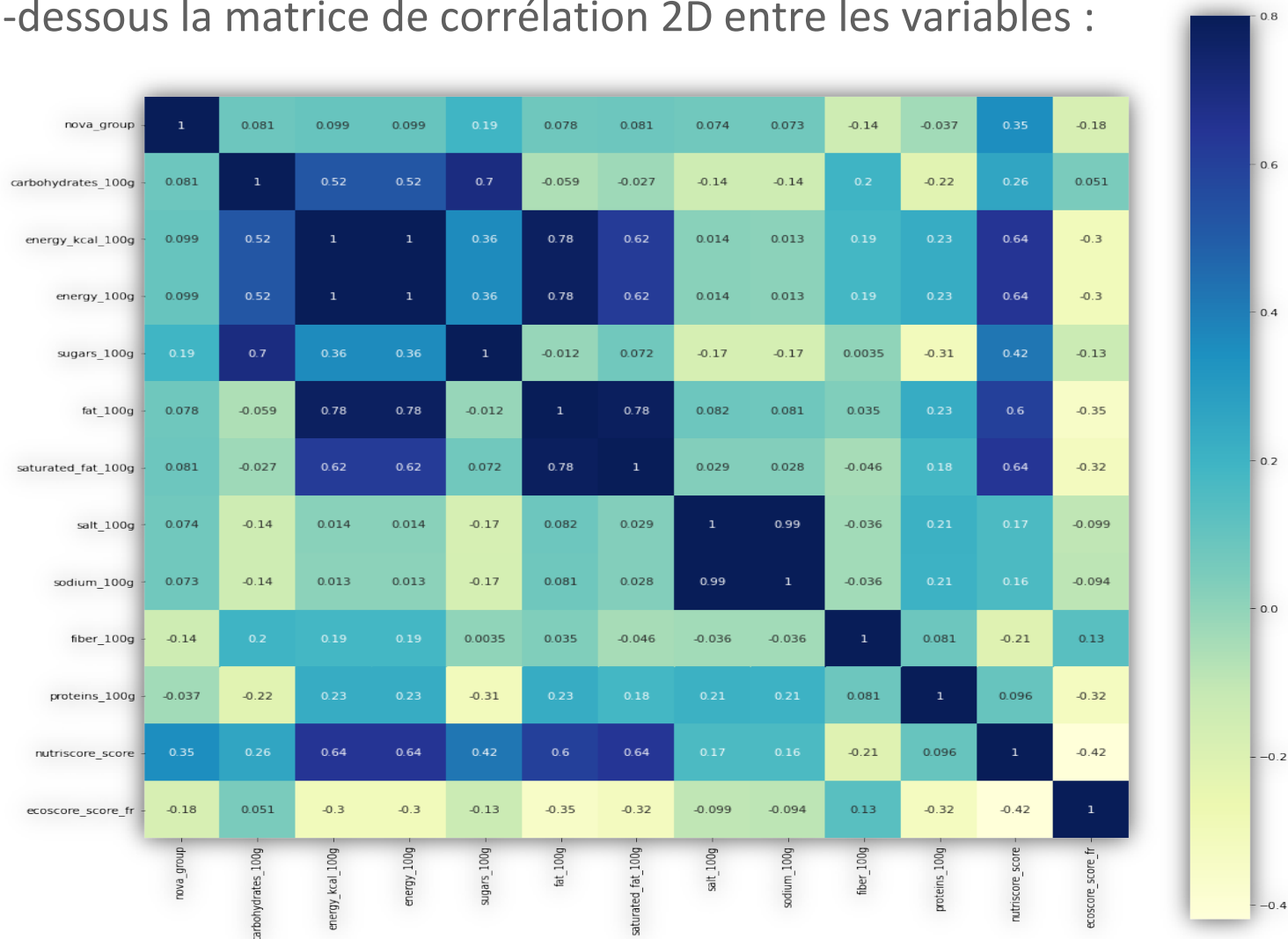
Analyse bivariée et multivariée

- **L'analyse bivariée** permet de comprendre les relations entre deux variables et à quel degré ces dernières agissent sur le phénomène à modéliser.
- **L'analyse multivariée** permet de chercher à établir un lien statistique entre plusieurs variables.



Analyse bivariée et multivariée – Corrélation

- Ci-dessous la matrice de corrélation 2D entre les variables :

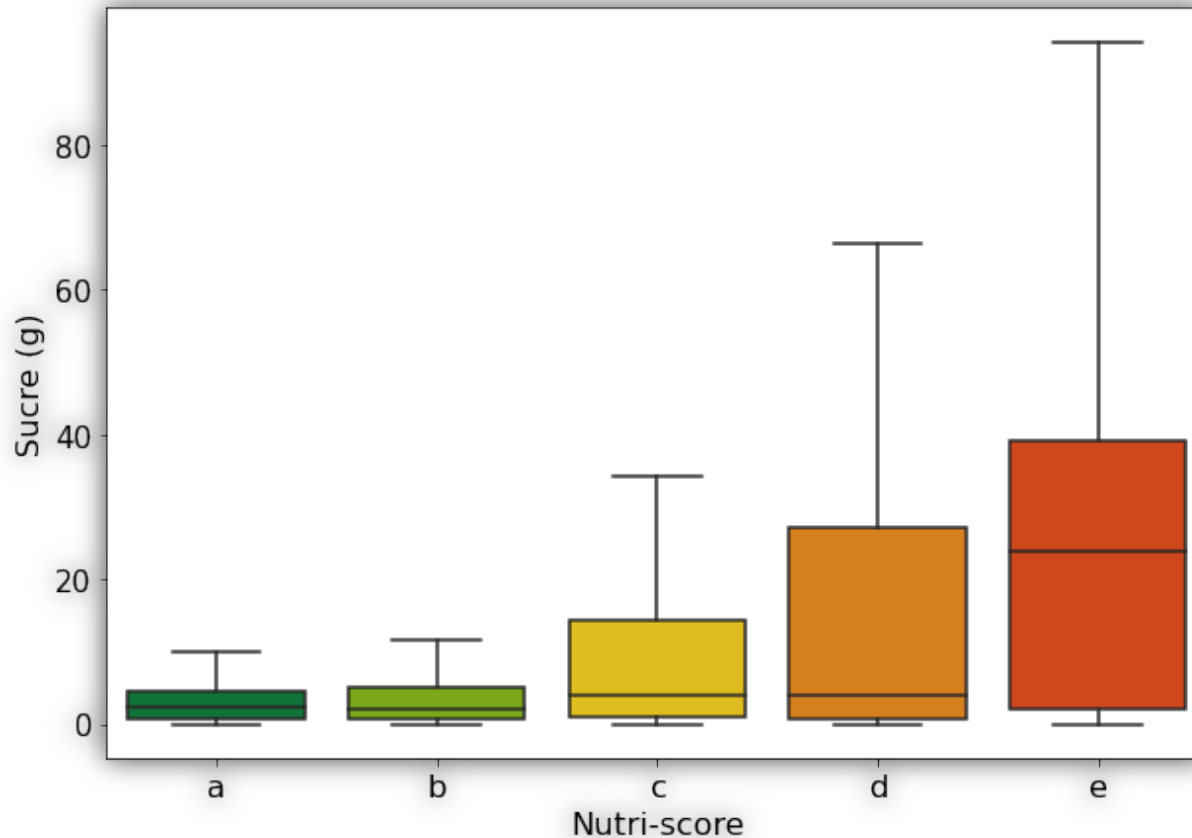


Analyse bivariée et multivariée – Corrélation

- La matrice de corrélation de la page précédente, montre un lien fort entre les variables ci-dessous :
 - Le sucre et les glucides (0.7).
 - L'énergie et les lipides (0.78).
 - Le nutri-score et le gras saturé (0.64)
 - Le nutri-score et l'énergie (0.64)
 - Le sel et le sodium (0.99)

Analyse bivariée

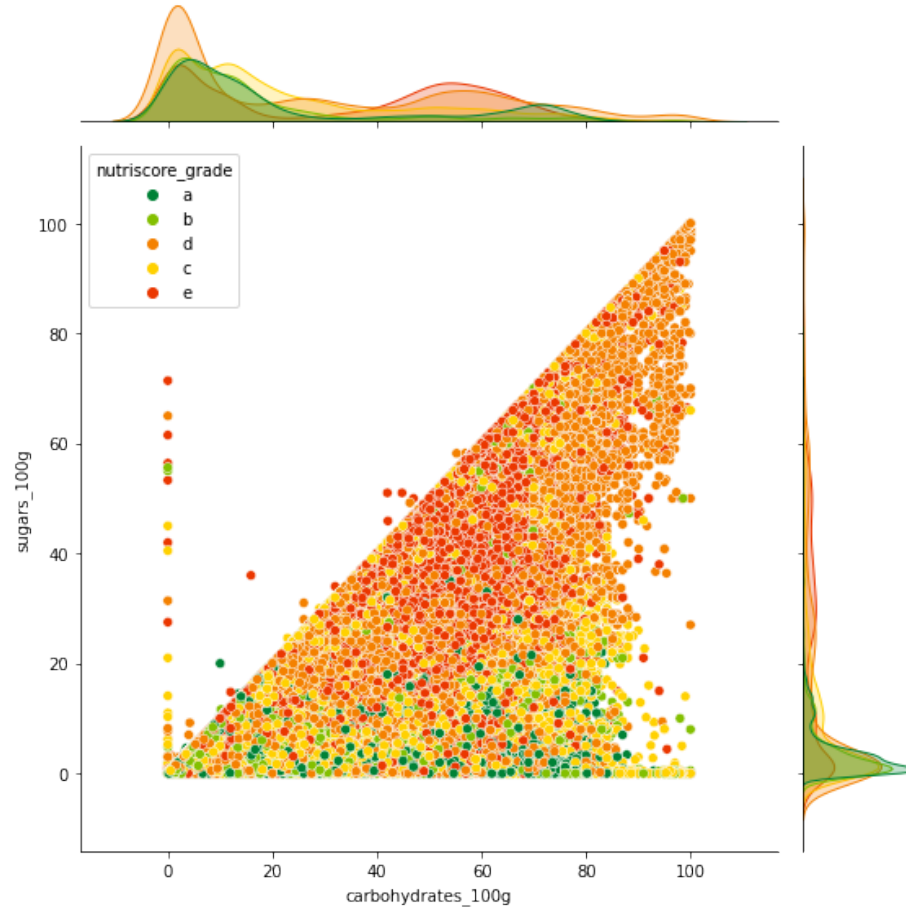
- Graphique boîte à moustache pour identifier la corrélation entre le sucre et le Nutri-score :



- Remarque : Si le sucre est élevé, l'aliment est moins bon en terme de qualité nutritionnelle (Nutri-score E).

Analyse bivariée

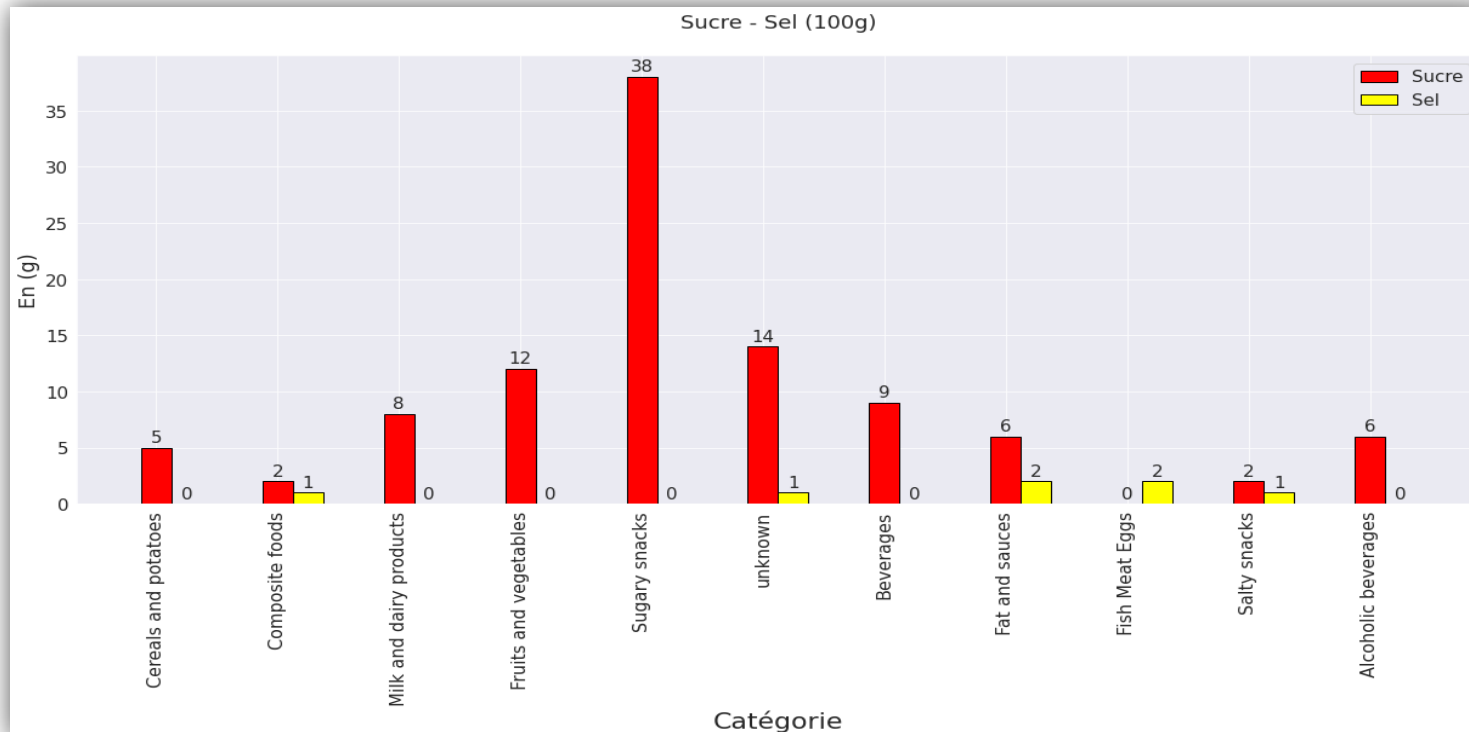
- Graphique pour identifier la relation entre le sucre et les glucide:



- Remarque : Si le sucre est élevé, la quantité des glucides augmente aussi.

Analyse multivariée

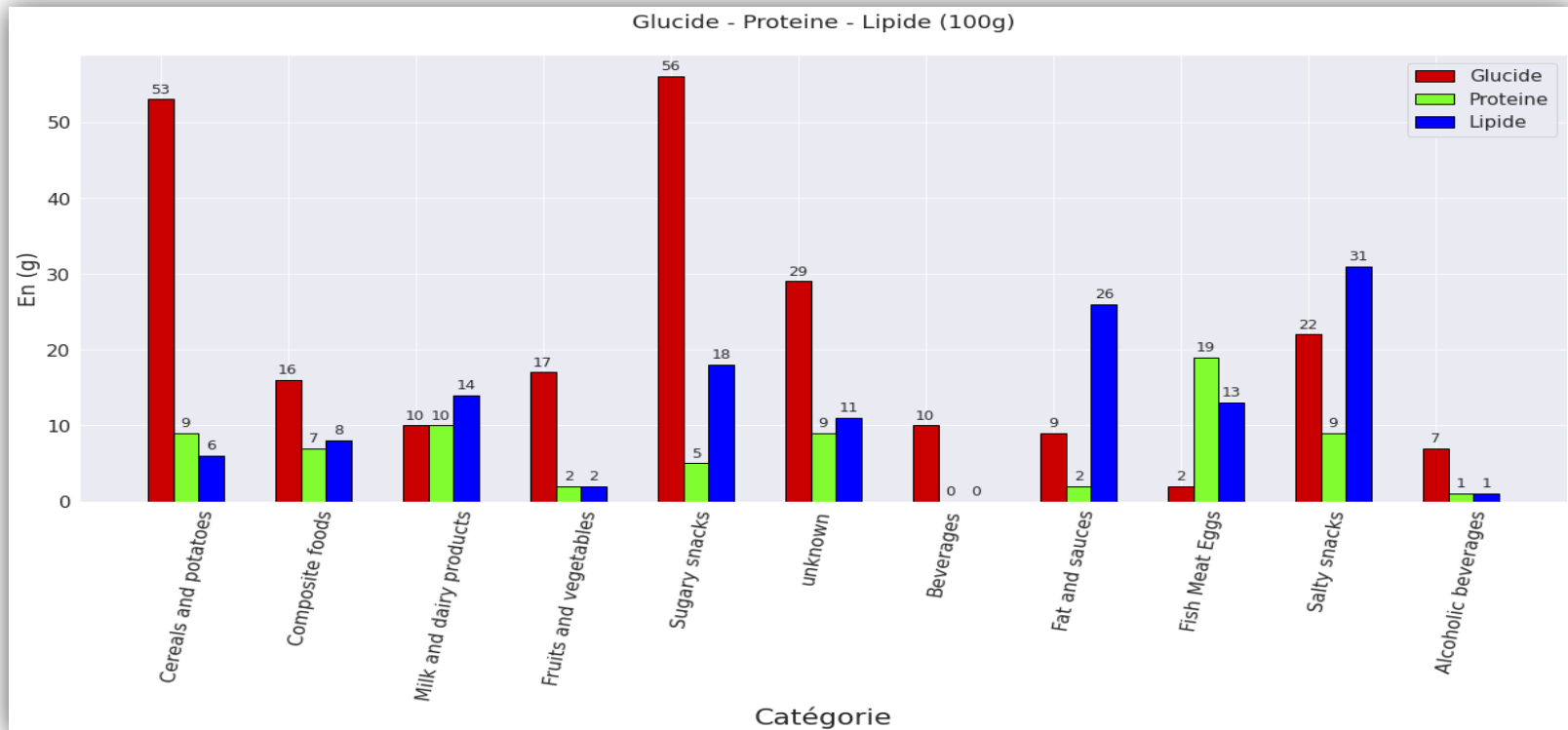
- Graphique pour déterminer la relation entre le sucre et le sel pour chaque catégorie d'aliments :



- Remarque : Les aliments qui contiennent une quantité élevée du sucre est dans la catégorie « collation sucrée » avec 38 grammes en moyenne.

Analyse multivariée

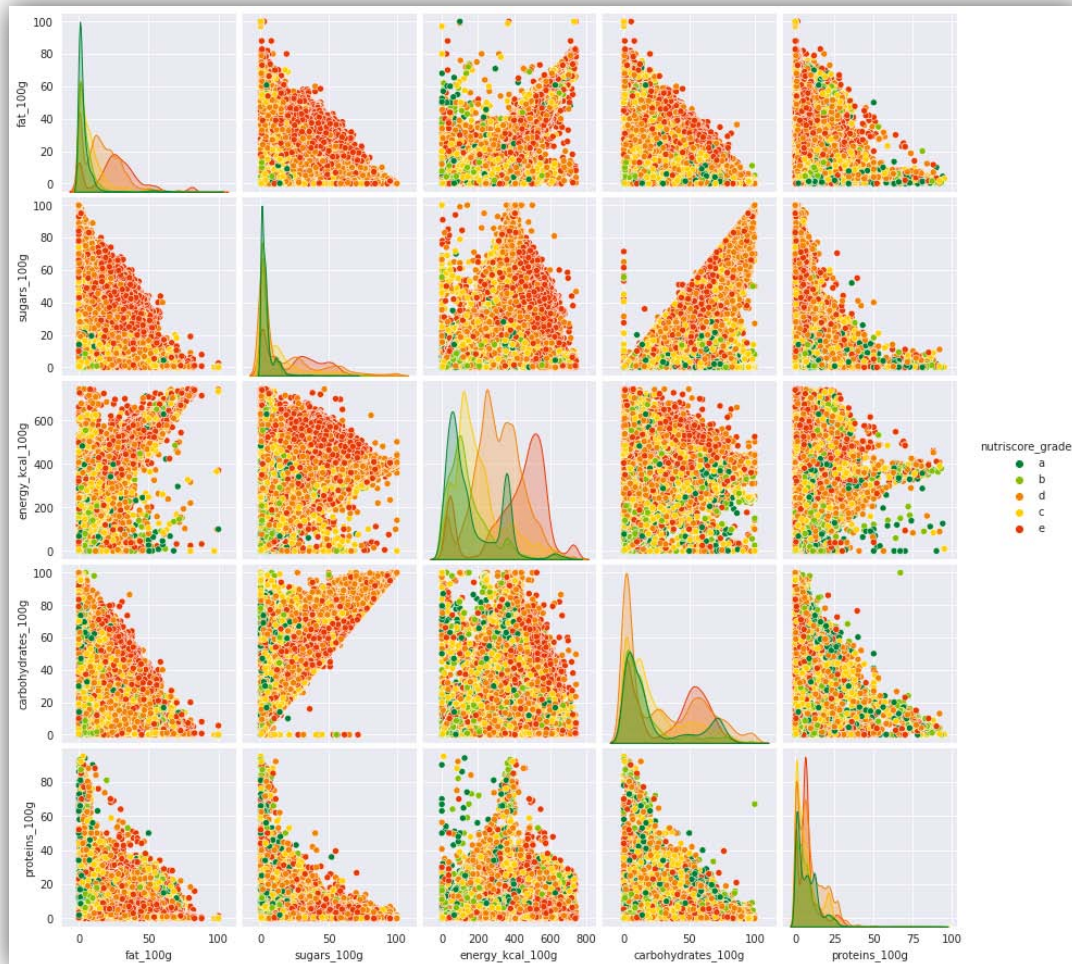
- Graphique pour déterminer la relation entre les glucides, les lipides et les protéines pour chaque catégorie d'aliments :



- Remarque : Les aliments riches en glucide appartiennent à la catégorie «collation sucré» et «céréales et pommes de terre », ceux riches en lipide appartiennent à «graisses et sauces » et «collations salées ».

Analyse multivariée

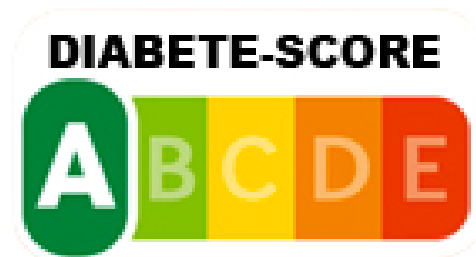
- Graphique pour déterminer la relation entre plusieurs variables : ('fat_100g', 'sugars_100g', 'energy_kcal_100g', 'carbohydrates_100g', 'proteins_100g')



Application

Application - Présentation

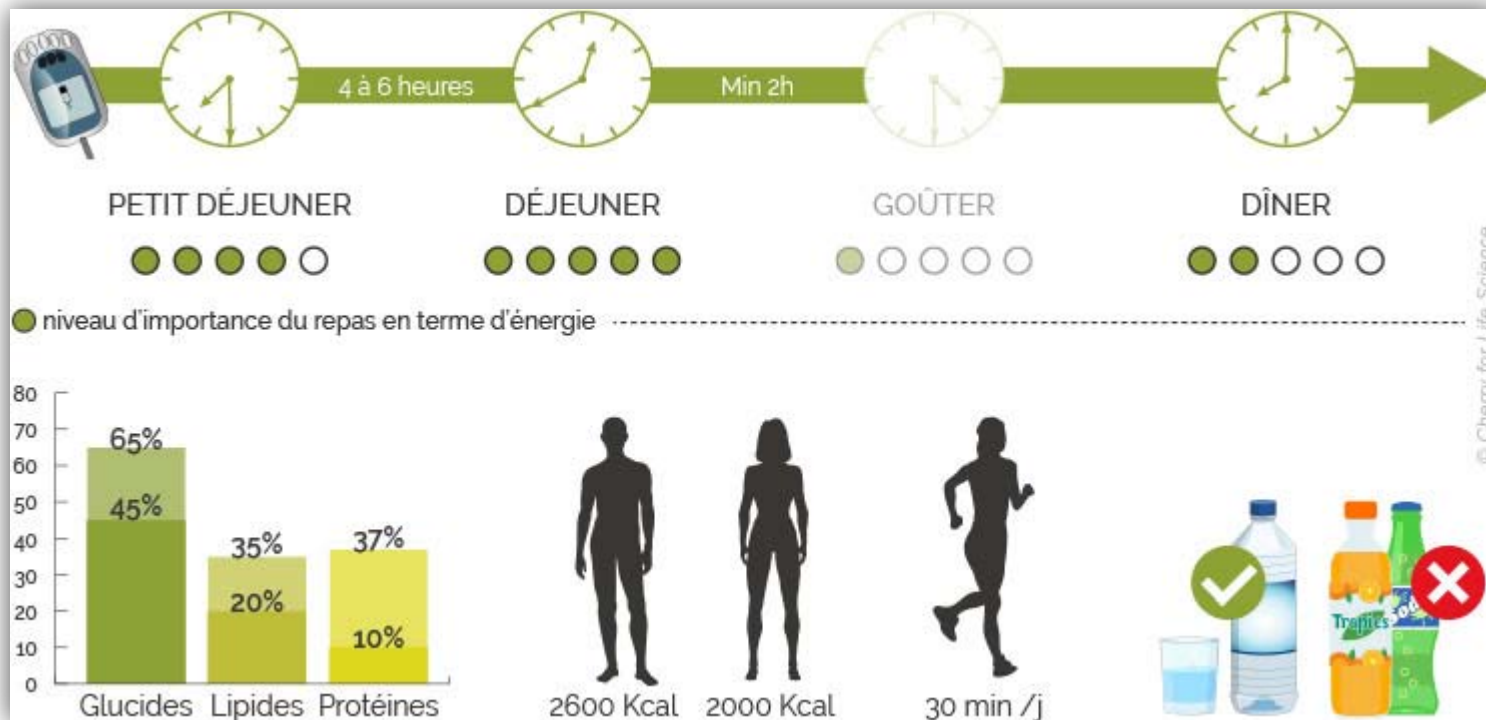
- Le diabète est un problème de santé publique qui atteint 5 millions de personnes en France, ce qui représente 8% de la population.
- Le consommateur diabétique a besoin d'être éclairé sur les aliments les plus adaptés à son régime alimentaire avec une bonne qualité nutritionnelle, un faible degré de transformation et un faible impact sur l'environnement.
- L'application « **DIABETE-SCORE** » va aider les consommateurs diabétiques à trouver un équilibre alimentaire via un système de notation qui prend en compte la quantité nutritionnelle nécessaire :



Application - Présentation

- Quel est le type de repas à privilégier lorsqu'on est diabétique ?

Pour un sujet diabétique, l'alimentation joue un rôle majeur pour limiter les risques de complications métaboliques. Avant de procéder à un quelconque régime, il est important d'adopter un mode de vie sain et ce, autant pour le sujet diabétique que pour sa famille :



Application - Calcul du score

- Pour cibler les aliments pour les diabétiques, il faut créer un repère nutritionnels qui permet de repérer la quantité de lipides, sucre, glucides et protéines contenue dans ces aliments ([source](#) : Ce repère a été inspirer du feu tricolore (*rouge, orange, vert*) mis en place par les autorités sanitaires de Grande Bretagne (Food Standards Agency - FSA).
- La formule de calcul définissant ce repère est décrite ci-dessous :

Pour 100 g	Faible	Modérée	Elevée
Glucides	jusqu'à 5g	de 5g à 20g	plus de 20g
Lipides	jusqu'à 3g	de 3g à 10g	plus de 10g
Sucre	jusqu'à 5g	de 5g à 13g	plus de 13g
Protéines	jusqu'à 5g	de 5g à 20g	plus de 20g
Nitru-score	Nutri-score D et E	Nutri-score B et C	Nutri-score A
Nova	Nova 4	Nova 2 et 3	Nova 1
Eco-grade	Eco-score D et E	Eco-score B et C	Eco-score A

Application - Calcul du score

- Ci-dessous un exemple d'aliments avec le repère nutritionnel tel affiché dans le dataframe :

product_name	fat_100g	carbohydrates_100g	proteins_100g	sugars_100g	quantite_sucre	quantite_glucide	quantite_lipide	quantite_proteine	nutrigrade	nova	ecograde
BAguette bressan	2.20	25.20	9.50	0.60	Faible	Elevee	Faible	Moderee	Elevee	Faible	Moderee
Blanquette de Volaille et son Riz	2.20	15.30	6.80	0.50	Faible	Moderee	Faible	Moderee	Moderee	Faible	Moderee
Compote de Pomme	0.50	93.00	1.50	66.00	Elevee	Elevee	Faible	Faible	Moderee	Faible	Moderee
Bonbons acidulés Raisin Fraise	0.00	93.30	0.00	93.30	Elevee	Elevee	Faible	Faible	Faible	Faible	Faible
Pâte d'Amandes	10.29	77.02	4.20	69.97	Elevee	Elevee	Elevee	Faible	Faible	Faible	Faible
...
Notre camembert bio	24.00	1.50	17.00	1.50	Faible	Faible	Elevee	Moderee	Faible	Moderee	Moderee
Thé vert Earl grey	0.20	0.50	0.50	0.50	Faible	Faible	Faible	Faible	Moderee	Faible	Elevee
Nectar de mangue	0.50	12.10	0.50	12.10	Moderee	Moderee	Faible	Faible	Faible	Moderee	Moderee
Jus multifruits	0.00	10.00	0.50	10.00	Moderee	Moderee	Faible	Faible	Faible	Moderee	Faible
Compote à Boire Pomme Poire	0.20	14.00	2.00	14.00	Elevee	Moderee	Faible	Faible	Elevee	Moderee	Moderee

Application - Calcul du score

- La prochaine étape est de mettre en place un système de notation qui prend en compte les valeurs nutritionnelles (sucre, lipide, glucide, protéine) qui intéressent les diabétiques et le nutri-score, l'eco-score et la classification nova.
- Ci-dessous le système de notation mis en place pour chaque produit:

	Note pour faible	Note pour modérée	Note pour élevée	Coefficient
Glucides	5	10	1	2
Lipides	10	5	1	1
Sucre	10	1	1	3
Protéines	5	10	5	3
Nitru-score	1	5	10	3
Nova	1	5	10	2
Eco-grade	1	5	10	1

Application - Calcul du score

- Pour rendre le « DIABETE-SCORE » visible et facile à comprendre j'ai créé une échelle graphique qui le divise en 5 classes (A, B, C, D, E) :















- Ce système de classification dépend de la note du produit :

DIABETE-SCORE	A	B	C	D	E
Note	Note ≥ 16	de 12 à 16	de 8 à 12	de 4 à 8	Note < 4

Application - Calcul du score


- Ci-dessous un exemple d'affichage avec le logo:

image_small_url	code	product_name	pnns_groups_2	Diabete_score	Diabete_grade	sugars_100g
	0000000274722	Blanquette de Volaille et son Riz	One-dish meals	16.00	DIABETE-SCORE 	0.50
	0000069006562	Galette de Pommes de Terre	One-dish meals	9.86	DIABETE-SCORE 	0.80
	0000069013508	Chicken Chips	One-dish meals	14.00	DIABETE-SCORE 	0.50
	0000069028045	Lasagnes aux saumon et aux épinards	One-dish meals	14.00	DIABETE-SCORE 	1.50
	0000069162176	Quenelles de brochet sauce Nantua	One-dish meals	10.14	DIABETE-SCORE 	0.60
	00010948	Taboulé	One-dish meals	6.00	DIABETE-SCORE 	5.00

Conclusions

Conclusions – Exemple 1

- Exemple avec un produit très sucré et avec un bon Nutri-score (code : 8594017144472) :







Mon application DIABETE-SCORE

Entrer le code barre

8594017144472









Produit sélectionné

1 : Produit(s)

image_small_url	code	product_name	pnns_groups_2	nutriscore_grade	ecoscore_grade_fr	nova_group	sugars_100g
	8594017144472	Emco Musli No Sugar Added 375g	Breakfast cereals				21.00

Les Produits proposés pour les diabétiques

1832 : Produit(s)

image_small_url	code	product_name	pnns_groups_2	Diabete_score	Diabete_grade	sugars_100g
	8436015595576	Flocons de riz	Breakfast cereals	19.29		0.50
	3266191035703	Flocon précuits de sarrasin	Breakfast cereals	18.14		1.50
	3760211821388	Corn flakes bio	Breakfast cereals	18.14		1.10
	3596710447299	Flocons 5 Céréales	Breakfast cereals	18.14		1.10

Produit choisi avec une quantité de sucre élevée 21g

La liste des produits proposés Avec une faible quantité de sucre

Conclusions – Exemple 2

DIABETE-SCORE





Mon application DIABETE-SCORE

Entrer le code barre

8435177050855









Produit sélectionné

1 : Produit(s)


image_small_url	code	product_name	pnns_groups_2	nutriscore_grade	ecoscore_grade_fr	nova_group	sugars_100g
	8435177050855	Abricots moelleux dénoyautés	Dried fruits				32.00

Les Produits proposés pour les diabétiques

1183 : Produit(s)

image_small_url	code	product_name	pnns_groups_2	Diabete_score	Diabete_grade	sugars_100g
	2609928037723	Cereaux de noix	Dried fruits	18.71		3.00
	3760190622204	Baies séchées de goji	Dried fruits	18.57		0.10
	3760159011605	Noix d'Amazonie	Dried fruits	18.00		2.10
	20905733	Cerneaux de Noix de Californie	Dried fruits	17.43		3.00

Conclusions – Exemple 3







Mon application DIABETE-SCORE

Entrer le code barre

3770012968250









Produit sélectionné

1 : Produit(s)

image_small_url	code	product_name	pnns_groups_2	nutriscore_grade	ecoscore_grade_fr	nova_group	sugars_100g
	3770012968250	La pâte à tartiner Choc !	Sweets				30.00

Les Produits proposés pour les diabétiques

9348 : Produit(s)

image_small_url	code	product_name	pnns_groups_2	Diabete_score	Diabete_grade	sugars_100g
	3273220081099	Nature	Sweets	18.57		0.00
	3273221091783	So soja & amande	Sweets	17.86		0.00
	4001686210512	Stevi Drop	Sweets	17.57		2.50
	3396410035396	Sojadelice nature	Sweets	17.14		1.20

Conclusions – Exemple 4

DIABETE-SCORE





Mon application DIABETE-SCORE

Entrer le code barre

3263859528812









Produit sélectionné

1 : Produit(s)


image_small_url	code	product_name	pnns_groups_2	nutriscore_grade	ecoscore_grade_fr	nova_group	sugars_100g
	3263859528812	Pilons de Poulet Rôtis Nature	One-dish meals				0.50

Les Produits proposés pour les diabétiques

12107 : Produit(s)

image_small_url	code	product_name	pnns_groups_2	Diabete_score	Diabete_grade	sugars_100g
	3596710389964	Lentilles	One-dish meals	20.00		0.50
	3021690029826	Lentilles cuisinées bio fondue d'oignons	One-dish meals	20.00		1.10
	3277510001705	Lentilles	One-dish meals	20.00		0.90
	3431590001899	Lentilles du Sud-Ouest cuisinée au tofu fumé	One-dish meals	20.00		0.50

Conclusions – Exemple 5







Mon application DIABETE-SCORE

Entrer le code barre

3560071083472









Produit sélectionné

1 : Produit(s)

image_small_url	code	product_name	pnns_groups_2	nutriscore_grade	ecoscore_grade_fr	nova_group	sugars_100g
	3560071083472	Petit Beurre multi céréales	Biscuits and cakes				23.00

Les Produits proposés pour les diabétiques

12184 : Produit(s)

image_small_url	code	product_name	pnns_groups_2	Diabete_score	Diabete_grade	sugars_100g
	9421019290380	Primo crackers	Biscuits and cakes	18.14		0.10
	3700413400721	16 Crêpes Peu Sucrées Surgelées	Biscuits and cakes	16.71		1.70
	3067850001318	Donuts de poulet	Biscuits and cakes	16.71		1.10
	3456300010684	Lunette chocolat noir	Biscuits and cakes	16.14		1.00