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

Lancelot LECLERCQ

novembre 2021

Problématique

- Application au service de la santé publique
- Idées d'application en lien avec l'alimentation

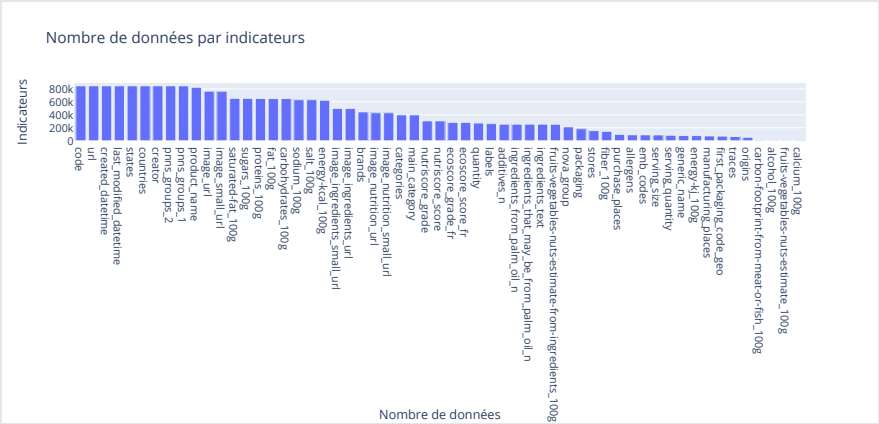
Jeu de données

- Jeu de données issu de Open Food Facts
- Base de donnée ouverte et participative
- Repertorie les ingrédients, les allergènes, composition nutritionnelle, les labels, l'origine, etc
- Grosse base de données (environ 4,5 Go)

Produits vendus en France

- Sélection des produits vendus en France
- 850 000 entrées sur les 2 milliards d'origine

- Suppression des colonnes dans lesquelles on a moins de 1% de données et des colonnes contenant des tags et des versions anglaises qui sont redondantes avec les colonnes du mêmes nom
- 57 colonnes restantes sur les 186 d'origine

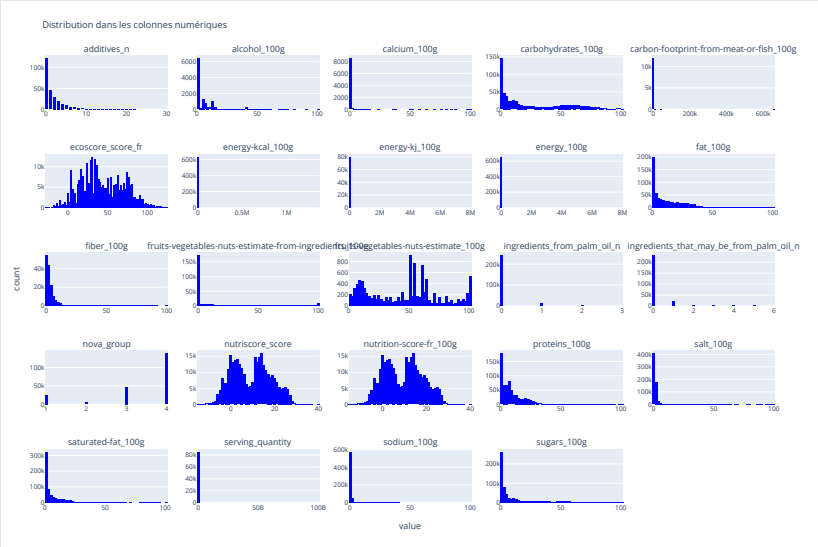


Nettoyage des outliers

- Valeurs extrèmes pour :
  - Quantité servie
  - Empreinte carbone
  - Énergie

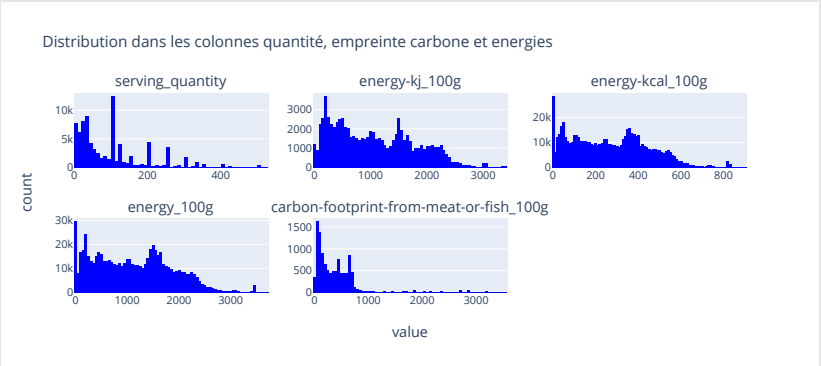
- Des valeurs très importantes là où le nombre de valeurs est très faible

- Nous écartons ces valeurs lorsqu'elles sont en dehors du quartile qui contient 99% des valeurs.



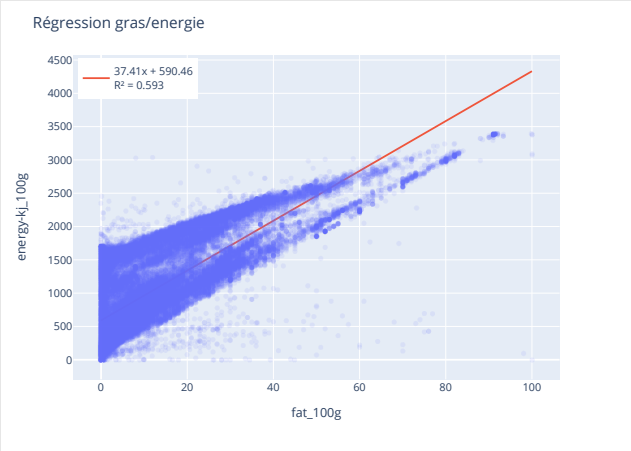
Nettoyage des outliers

- On obtient des distributions plus intéressantes
- Fonctionnement assez similaire des différents types d'énergies



Données redondantes

- Mêmes valeurs energy_100g et energy-kj_100g
- Variation pour ce qui est de l'energie en kcal pour 100g
- Nous n'allons garder que les colonnes avec des unités d'énergie : kcal et kj
- De même nous allons supprimer les colonnes finissant par _tags et _en qui contiennent des données redondantes des colonnes du même nom sans suffixes



Analyse univariée

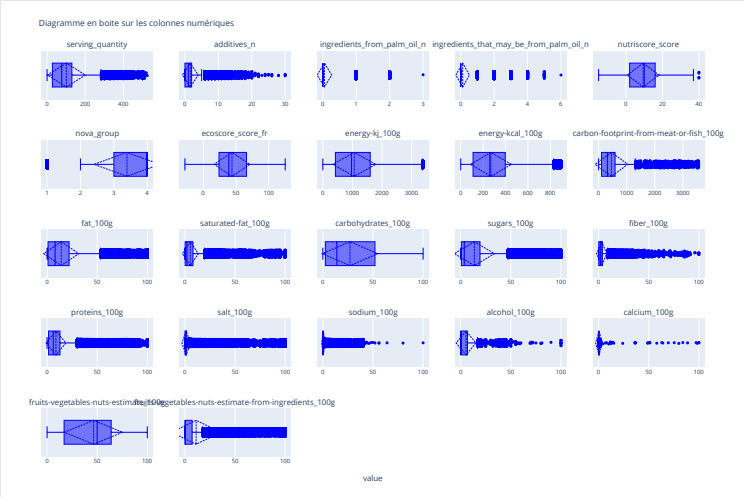
	serving_quantity	additives_n	ingredients_from_palm_oil_n	ingredients_that_may_be_from_palm_oil_n	nutriscore_score	nova_group	ecoscore_score_fr
count	84973,00	256284,00	256284,00	256284,00	308035,00	215382,00	283986,00
mean	101,26	1,60	0,05	0,12	9,44	3,39	43,80
std	99,35	2,37	0,22	0,40	8,76	0,98	27,29
min	0,00	0,00	0,00	0,00	-15,00	1,00	-28,00
25%	28,00	0,00	0,00	0,00	2,00	3,00	24,00
50%	77,00	1,00	0,00	0,00	10,00	4,00	39,00
75%	130,00	2,00	0,00	0,00	16,00	4,00	66,00
max	525,00	30,00	3,00	6,00	40,00	4,00	125,00

	energy-kj_100g	energy-kcal_100g	carbon-footprint-from-meat-or-fish_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g
count	79519,00	625208,00	11521,00	651724,00	655473,00	651661,00	655057,00
mean	1061,27	269,51	457,22	14,15	5,38	27,06	13,63
std	726,93	183,05	582,22	17,49	7,94	27,71	19,97
min	0,00	0,00	0,05	0,00	0,00	0,00	0,00
25%	425,00	110,00	103,60	1,00	0,20	2,50	0,60
50%	971,00	260,00	310,80	8,00	2,00	14,00	3,50
75%	1601,00	398,00	585,00	22,00	8,00	52,00	19,20
max	3388,00	895,80	3569,26	100,00	100,00	100,00	100,00

	fiber_100g	proteins_100g	salt_100g	sodium_100g	alcohol_100g	calcium_100g	fruits-vegetables-nuts-estimate_100g
count	145645,00	653554,00	634236,00	634249,00	11085,00	8699,00	10523,00
mean	3,03	8,86	1,24	0,50	5,29	0,35	46,34
std	5,08	9,93	4,27	1,73	9,91	3,36	29,02
min	0,00	0,00	0,00	0,00	0,00	0,00	0,00
25%	0,12	1,50	0,06	0,02	0,00	0,03	17,00
50%	1,70	6,20	0,52	0,21	0,00	0,12	50,00
75%	3,70	12,90	1,30	0,52	6,70	0,18	64,00
max	100,00	100,00	100,00	100,00	100,00	100,00	100,00

TAB. : Tableaux des statistiques sur chaques variables numériques

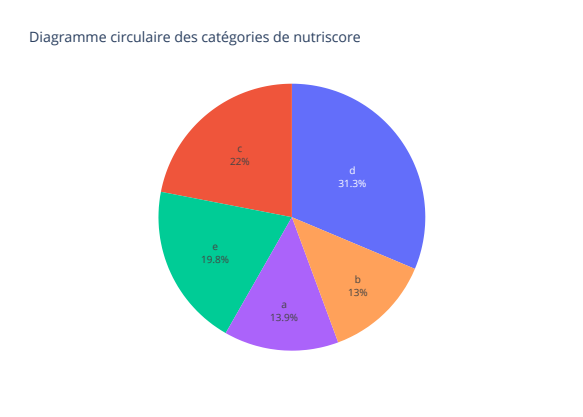
Analyse univariée



Analyse univariée

- Observation du nombre d'occurences de chaques catégories de nutriscore

- Diagramme circulaire permet de visualiser les proportions

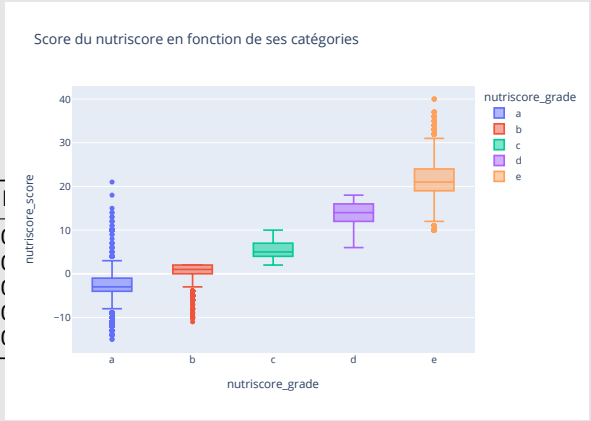


Analyse univariée

- Les catégories du nutriscore sont corrélées avec le score

- Vérification par ANOVA à une dimension

	coef	std err	t	
Intercept	-3.1898	0.013	-245.363	0
nutriscore_grade[T.b]	4.0284	0.019	215.532	0
nutriscore_grade[T.c]	8.7035	0.017	523.851	0
nutriscore_grade[T.d]	17.0136	0.016	1089.153	0
nutriscore_grade[T.e]	24.5820	0.017	1449.063	0



- Pouvoir scanner un produit et trouver des alternatives similaires de meilleure qualité
  - Meilleures pour la santé (nutriscore plus élevé)
    - Avec moins d'additifs (nombre d'additifs inférieurs)
  - Meilleures pour l'environement (ecoscore plus élevé)
    - Origine plus proche (France, Europe)
  - Voir les deux

- Pouvoir choisir selon un régime alimentaire particulier
  - Végétarien
  - Vegan
  - Halal
  - Kasher
  - Sans gluten
  - etc

Implémenter un système de reconnaissance d'image permettant d'extraire les données manquantes des photos des étiquettes