PARCOURS DATASCIENTIST

PROJET 2

ANALYSEZ DES DONNÉES NUTRITIONNELLES

mars 23, 2018

CONTENTS

Table of Contents

1 Introduction

Cette étude a pour objectif d'explorer une base de données alimentaire pour constituer un générateur de recettes saines.

Une recette est comprise ici comme une combinaison d'aliments de la base alimentaire.

2 Formulation du problème

Pour reformuler le problème posé en introduction, nous faisons les hypothèses suivantes :

- une recette saine est composée d'aliments sains.
- un aliment sain est composé de nutriments qui contribuent positivement à cette propriété.

Nous reformulons le problème dans les termes suivants : comment classer les aliments pour quantifier leur sanité ?

La qualité d'un aliment de la base est mesurée par une variable quantitative, le score.

Étant donné le rôle clé de l'apport énergétique des nutriments dans l'alimentation, nous avons étudié la corrélation entre le l'apport énergétique de l'ensemble des nutriments à un aliment et le score de l'aliment.

Pour pouvoir étendre le score à une plus large population d'aliments, un score obtenu par pondération des nutriments et de la densité énergétique a été établi. La pertinence de cette variable a été évaluée en comparant le score obtenu avec le score natif de la base des aliments.

3 Variables consistantes pour l'étude

Pour répondre au problème formulé, les variables consistantes pour l'étude qui ont été retenues sont présentées dans le tableau ci-dessous :

Variables qualitatives	Nom des variables qualitatives	Variables quantitatives	Nom des variables quantitatives
Dénomination des aliments	product_name	Apports en nutriments	fat saturated-fat cholesterol sugars sodium carbonhydrates fiber proteins vitamin-a vitamin-c calcium iron
Catégorie de nutriments	FIBRES ACIDES GRAS GLUCIDES LIPIDES PROTEINES SELS MINERAUX VITAMINES	Densité énergétique	energy
		Score natif	nutrition-score-fr nutrition-score-uk
		Score recalculé par pondération	score_1 score_2

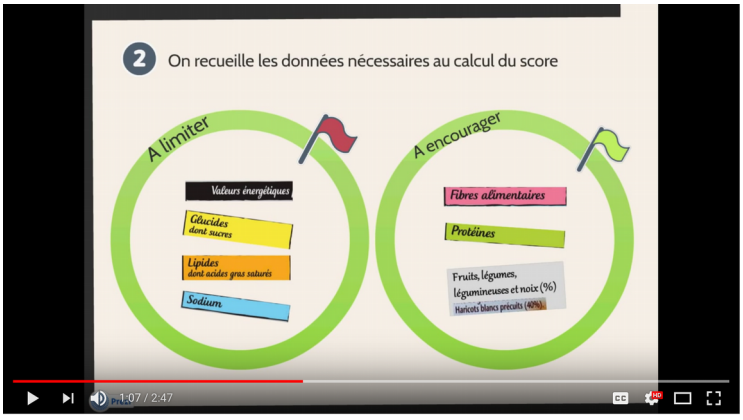
Les noms des colonnes se terminant avec le motif '_100g' ont été renommées en supprimant ce motif.

Les nutriments ont été décomposés en deux classes :

- les nutriments qui contribuent positivement à un aliment sain, en vert dans le tableau ci-dessus.
- les nutriments qui contribuent négativement à un aliment sain, en rouge dans le tableau ci-dessus.

Cette décomposition s'est faite sur la base de la littérature sur le sujet.

Cette approche intuitive est compatible avec le modèle d'openfact food : *https://www.youtube.com/watch?v=GAwTyEEHnOs*



4 Acquisition des données

Le fichier constituant la basse alimentaire est riche de plus de 300 000 références. La mémoire RAM nécessaire au traitement du modèle total est de plus de 3GB.

La complexité de certaines opérations de nettoyage étant de l'ordre de N^2, le parti a été pris de travailler sur une fraction aléatoire de la base pour la mise au point du modèle.

Une fois fait, la totalité de la base a été chargée pour l'analyse.

5 Nettoyage de la base alimentaire

Les opérations de nettoyage de la base ont deux conséquences directes :

- la diminution de la dimension du modèle par la suppression de certaines variables.
- La diminution de la complexité du modèle de données par la réduction de la population d'aliments dans la base.

Toutes les informations de la base ne sont pas exploitables.

Certaines donnés quantitatives sont absentes en valeurs ou sont aberrantes, et des données qualitatives ne sont pas exprimées.

5.1 Choix des variables consistantes

Les variables consistantes retenues sont :

- les nutriments
- la dénomination des aliments
- le score des aliments

La liste des nutriments retenus est issue du fichier datafields.txt qui décrit la base alimentaire. La liste est utilisée pour filtrer les variables retenues.

Cette liste est constituée a partir de la lecture du fichier data/nutrition_fact.txt
Ce dernier est issue d'opérations d'édition sur le fichier original datafields.txt.

5.2 Fusion des colonnes complémentaires

La variable permettant d'identifier un aliment, product_name, a été complétée par les variables generic_name et categories exprimées, lorsque l'identification (product_name) ne l'était pas.

Cette opération permet d'identifier le plus grand nombre d'aliments.

5.3 Purge des colonnes dont les min et max sont égaux.

Ces colonnes ne permettent pas d'exploiter des données statistiques.

5.4 Purge des colonnes dont les valeurs extrêmes des boxplot sont égales

Les valeurs extrêmes sont calculées de la façon suivante :

- zmin = min(max, Q3+1.5*(Q3-Q1))
- zmax = max(min, Q1-1.5*(Q3-Q1))

Les valeurs en dehors de ces mesures sont des outliers.

5.5 Purge des valeurs aberrantes

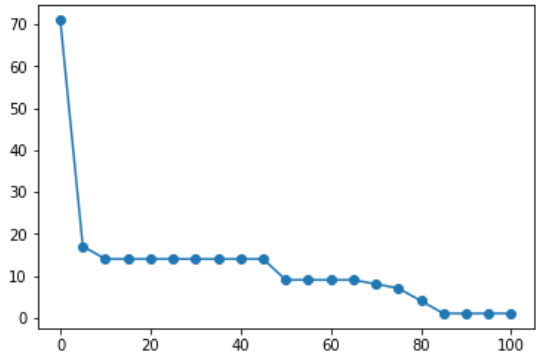
Les valeurs aberrantes détectées sont des observations négatives. La liste exhaustive de ces dernières est donnée ci-dessous :

Food : Grade A Fancy Chopped Spinach / Nutriment : sugars / Index= 8582 / Value= -1.2 --> 0.0
Food : Select, Spicy Red Bell Pepper Pasta Sauce / Nutriment : sugars / Index= 18209 / Value= -0.8 --> 0.0
Food : Xtra Butter Microwave Pop Corn, Butter / Nutriment : trans-fat / Index= 23576 / Value= -3.03 --> 0.0
Food : Traditional Tender Cracklins Chicharrones / Nutriment : fiber / Index= 23784 / Value= -6.7 --> 0.0
Food : Venus, Cuttlefish Balls / Nutriment : selenium / Index= 28324 / Value= -2e-06 --> 0.0
Food : Whole Cashews / Nutriment : proteins / Index= 33781 / Value= -3.57 --> 0.0
Food : Cheez Waffies / Nutriment : vitamin-c / Index= 41538 / Value= -0.0021 --> 0.0
Food : Cocoa Dyno-Bites, Sweetened Rice Cereal With Real Cocoa / Nutriment : copper / Index= 54954 / Value= -6.896552 --> 0.0
Food : Flavor Aid, Soft Drink Mix, Lemon / Nutriment : vitamin-a / Index= 80440 / Value= -0.0003396 --> 0.0
Food : Gelato Truffle On A Stick / Nutriment : trans-fat / Index= 107990 / Value= -0.7 --> 0.0
Food : Gourmet Blends, Seasoning, Garlic Pepper / Nutriment : proteins / Index= 115310 / Value= -500.0 --> 0.0
Food : Hummous, Black Truffle / Nutriment : sugars / Index= 117739 / Value= -3.57 --> 0.0
Food : Rugen Fisch, Salmon Fillets Skinless And Boneless / Nutriment : trans-fat / Index= 120692 / Value= -1.0 --> 0.0
Food : Crackers / Nutriment : sugars / Index= 146284 / Value= -6.67 --> 0.0
Food : Italianavera, Tomato Sauce With Gaeta Olives & Sicilian Capers, Olives & Sicilian Capers / Nutriment : sugars / Index= 150858 / Value= -6.25 --> 0.0
Food : Crispy Wheat Crackers / Nutriment : trans-fat / Index= 153498 / Value= -3.57 --> 0.0
Food : Organic Pumpkin Seeds / Nutriment : sugars / Index= 164030 / Value= -17.86 --> 0.0
Food : Lightly Dried Cilantro / Nutriment : proteins / Index= 169119 / Value= -800.0 --> 0.0
Food : France / Nutriment : biotin / Index= 189152 / Value= -1.0 --> 0.0
Food : France / Nutriment : pantothenic-acid / Index= 189152 / Value= -1.0 --> 0.0
Food : France / Nutriment : biotin / Index= 189160 / Value= -2.0 --> 0.0
Food : France / Nutriment : pantothenic-acid / Index= 189160 / Value= -2.0 --> 0.0
Food : France / Nutriment : biotin / Index= 189162 / Value= -1.0 --> 0.0
Food : France / Nutriment : pantothenic-acid / Index= 189162 / Value= -1.0 --> 0.0
Food : France / Nutriment : biotin / Index= 189260 / Value= -1.0 --> 0.0
Food : France / Nutriment : pantothenic-acid / Index= 189260 / Value= -1.0 --> 0.0
Food : France / Nutriment : biotin / Index= 189262 / Value= -1.0 --> 0.0
Food : France / Nutriment : pantothenic-acid / Index= 189262 / Value= -1.0 --> 0.0
Food : France / Nutriment : biotin / Index= 189269 / Value= -1.0 --> 0.0
Food : France / Nutriment : pantothenic-acid / Index= 189269 / Value= -1.0 --> 0.0
Food : France / Nutriment : biotin / Index= 189272 / Value= -1.0 --> 0.0
Food : France / Nutriment : pantothenic-acid / Index= 189272 / Value= -1.0 --> 0.0
Food : France / Nutriment : biotin / Index= 189417 / Value= -1.0 --> 0.0
Food : France / Nutriment : pantothenic-acid / Index= 189417 / Value= -1.0 --> 0.0
Food : Caprice des dieux / Nutriment : sugars / Index= 195761 / Value= -0.1 --> 0.0
Food : Lamthong, Fruit Cocktail In Syrup / Nutriment : iron / Index= 316892 / Value= -0.00026 --> 0.0

Ces valeurs ont été imputées à 0. La référence « Caprice des dieux » ou encore « Crakers » ont des valeurs d'apport en nutriment incohérentes.

5.6 Calcul du seuil de valeurs à Nan

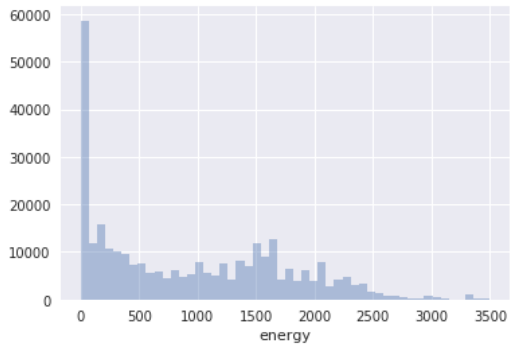
Ce seuil est un compromis entre le nombre de variables pour l'analyse du modèle et le nombre d'observations dont les valeurs nan ne sont pas calculables.



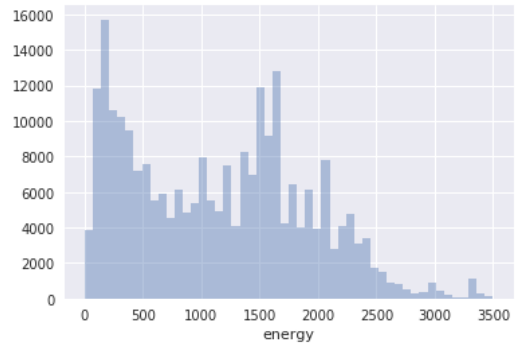
La valeur seuil de valeurs à nan admise est de 45 % autorisant à traiter 14 variables.

Ces valeurs sont imputées à la valeur 0.

5.7 Suppression des aliments à apport énergétique à 0



La distribution empirique de la densité énergétique met en évidence une sur-représentation du mode 0 pour la fréquence. Ces aliments représentent plus de 18 % du modèle.



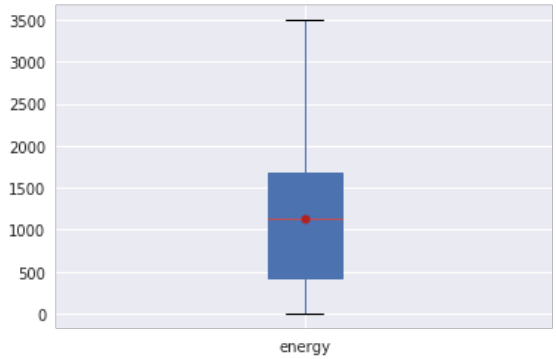
Deux modes apparaissent autour des valeurs 250 et 1500.

5.8 Synthèse de l’impact de la simplification du modèle de données

Opérations	Motivation	Dimension		Complexité	
		Avant	Après	Avant	Après
Fusion des colonnes	Identifier un maximum d’aliments.	162	160	320772	320772
Purge des lignes dont le nom de produit n'est pas renseigné	Identifier les aliments.	160	160	320772	304052
Suppression des variables non consistantes pour l'analyse	Obtenir un modèle de données consistant pour l’analyse.	160	86	304052	304052
Purge des colonnes dont toutes les valeurs sont nan	Obtenir un modèle de données calculable au regard des statistiques.	86	77	304052	304052
Purge des lignes dont toutes les valeurs sont NaN	Obtenir un modèle de données avec une population représentative.	77	77	304052	304052
Traitement des valeurs erratiques	Obtenir un modèle avec moins de biais.	77	77	304052	304052
Purge des colonnes dont min = max	Obtenir un modèle avec des variables représentatives pour le calcul statistique.	77	72	304052	304052
Purge des colonnes dans les quantiles extrêmes sont égaux	Obtenir un modèle avec des variables représentatives pour le calcul statistique.	72	71	304052	304052
Purge des colonnes dans les quantiles 1 et 3 sont égaux	Obtenir un modèle dont 50 % des observations sont contenues dans la boxplot.	71	69	304052	304052
Calcul du seuil de valeurs a NaN	Obtenir un modèle avec des variables représentatives pour le calcul statistique.	69	14	304052	304052
Purge des outliers de la variable energy	Faciliter l’analyse du modèle dans un premier temps.	14	14	304052	304052
Conservation des aliments dont la densité énergétique est > 0	Obtenir un modèle avec le moins de biais possible.	14	14	304052	247505

6 Analyse du modèle

6.1 Analyse uni-variée de la densité d'énergie

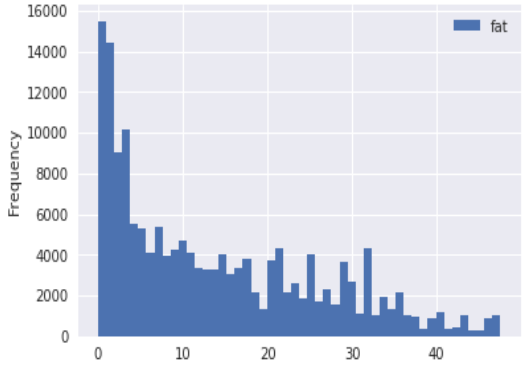


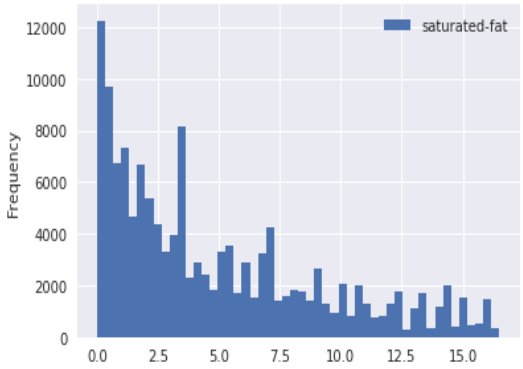
Le modèle retenu pour cette analyse est de supprimer les outliers de la densité énergétique.

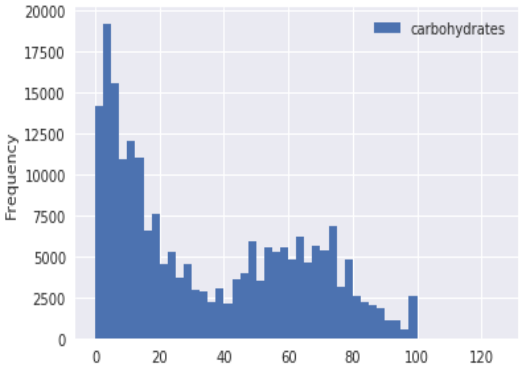
Ce choix a été motivé par le fait que la prise en compte des outliers, au vu de la qualité des observations de la base, compliquait considérablement l'analyse du modèle statistique. Ces outliers représentent 0,6 % de la base alimentaire.
Il serait nécessaire de s'assurer que ces aliments ne rentrent pas dans le commun d'une recette (sel, sucre…).

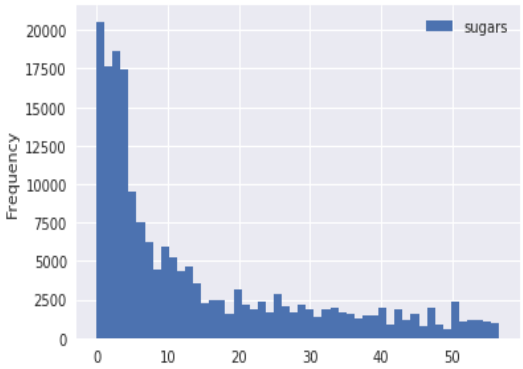
La représentation de montre la représentation de deux modes autour des valeurs 0 et 1500. La valeur 0 est aberrante.
L'analyse va être conduite autour des valeurs 0 < energy <=3500

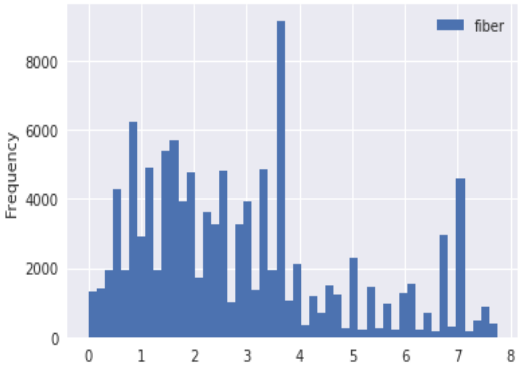
6.2 Distribution des nutriments

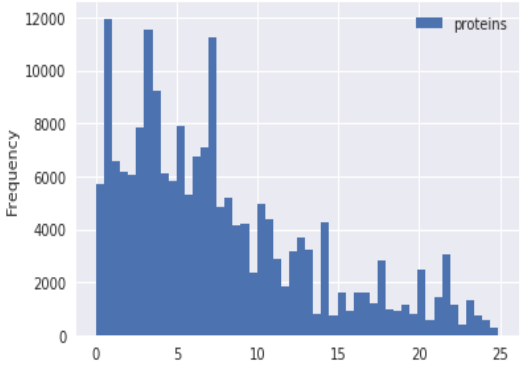


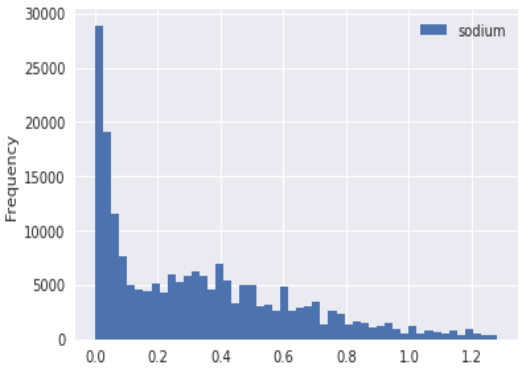


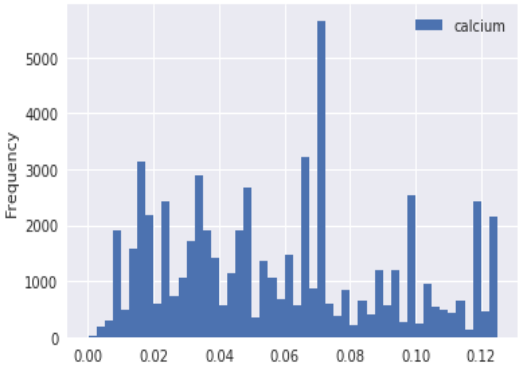


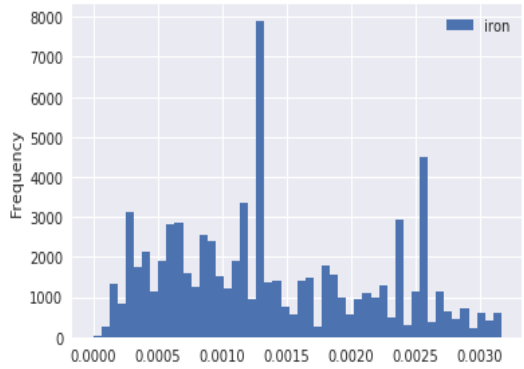








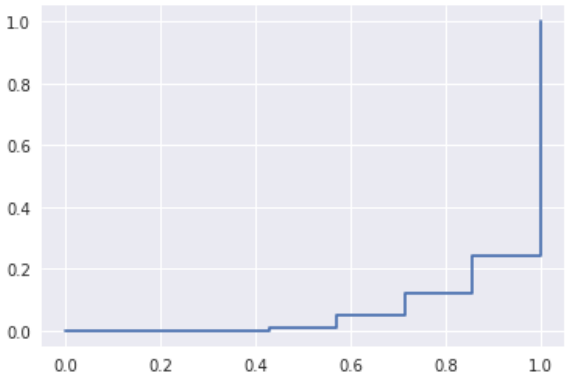




Certains nutriments ont une distribution multi-modale.
Par ailleurs, ils sont de fréquences inégales et leur apport relatif varie d'un facteur 1000 pour certains d'entre-eux.

Il sera tenu compte de ces éléments pour « scorer » les aliments.

6.3 Concentration des apports alimentaires

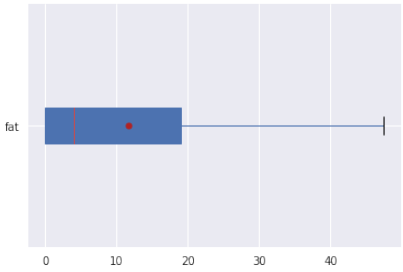


La courbe de Lorenz montre que les apports nutritionnels sont le fruit de quelques nutriments. 15% des nutriments concentrent 80% des apports.

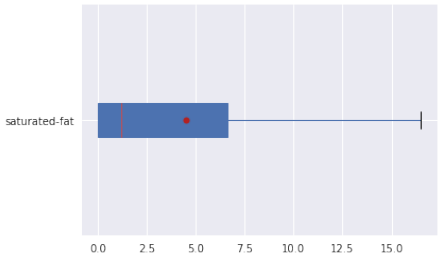
```
Coefficient de GINI= 0.89
```

6.4 Mesures de dispersion des nutriments

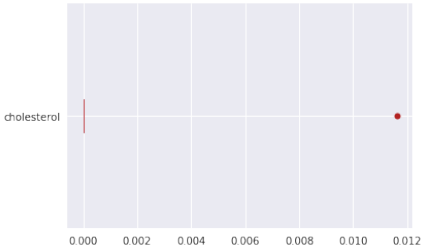
Moyenne: 11.685557912365406
Mediane: 4.0
Variance: 256.6102513869593
Ecart:16.019059004415936



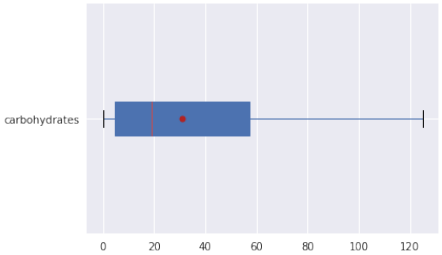
Moyenne: 4.520573131169067
Mediane: 1.18
Variance: 52.33467914940568
Ecart:7.234271155369122



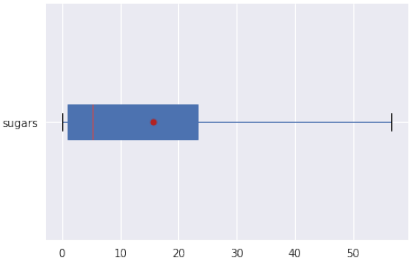
Moyenne: 0.011623153249429304
Mediane: 0.0
Variance: 0.0747282429874119
Ecart:0.27336467033508904



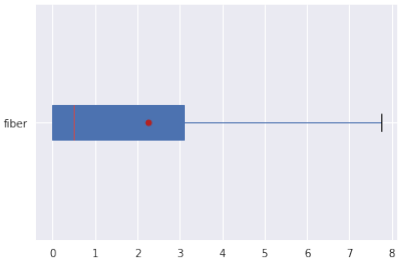
Moyenne: 30.91126301731278
Mediane: 18.9
Variance: 851.7991340524762
Ecart:29.18559805884533



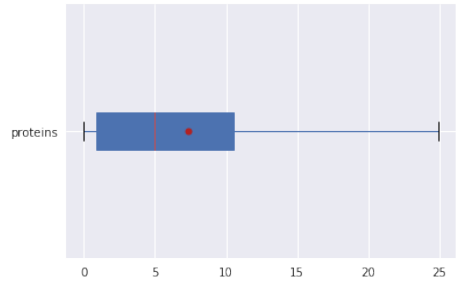
Moyenne: 15.54640064273449
Mediane: 5.21
Variance: 440.4314124144791
Ecart:20.98645783390992



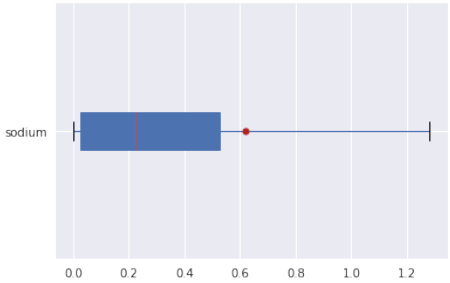
Moyenne: 2.264286934647785
Mediane: 0.5
Variance: 18.01369327954946
Ecart:4.244254148793337



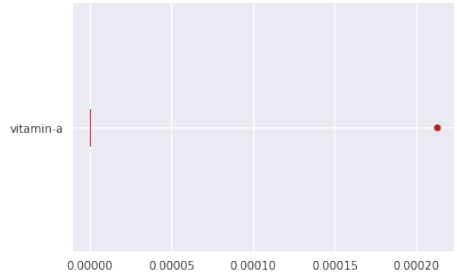
Moyenne: 7.322984347386921
Mediane: 5.0
Variance: 66.9055535942303
Ecart:8.17958150483448



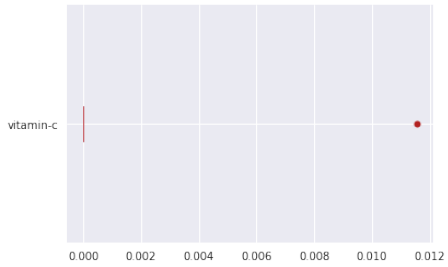
Moyenne: 0.6183285140418633
Mediane: 0.224
Variance: 2624.964154115974
Ecart:51.23440400859538



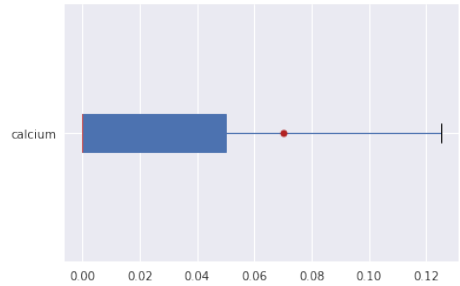
Moyenne: 0.00021272007166723905
Mediane: 0.0
Variance: 0.002984217706100184
Ecart:0.05462799379530777



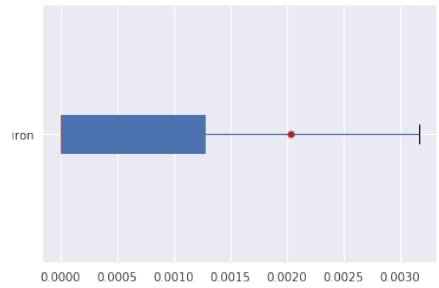
Moyenne: 0.011543384256479668
Mediane: 0.0
Variance: 2.786213397167533
Ecart:1.6691954340841977



Moyenne: 0.0699938222944991
Mediane: 0.0
Variance: 6.272702877524978
Ecart:2.504536459611834



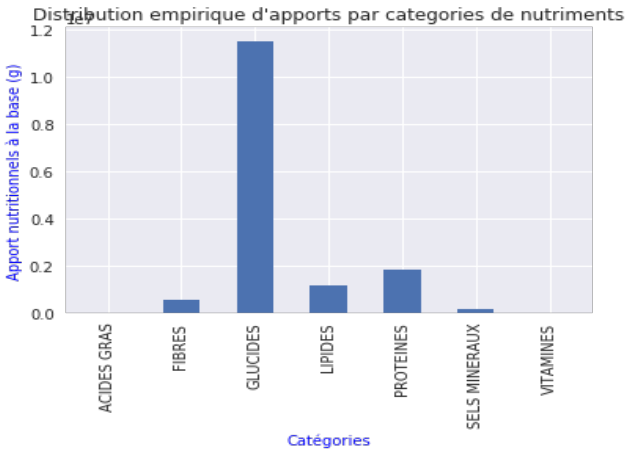
Moyenne: 0.00203051182932062
Mediane: 0.0
Variance: 0.025911525265955034
Ecart:0.1609705726707681



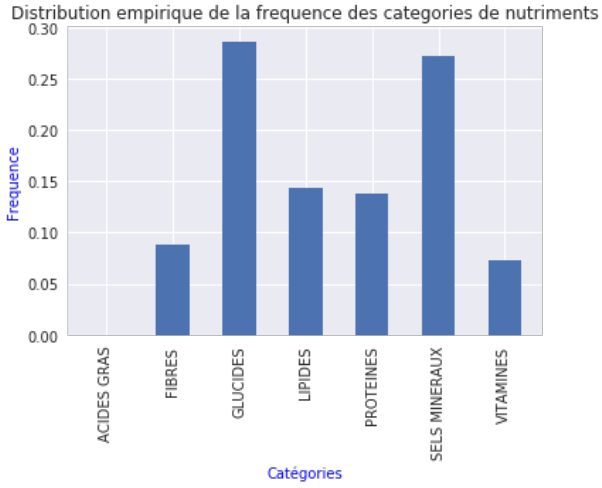
Le cholestérol et les vitamines ont des dispersions atypiques.

6.5 Distribution par catégories de nutriments

Les nutriments étudiés ont été classés en catégories.



Sur cet histogramme, l'apport en nutriment des vitamines n'est pas exprimé.



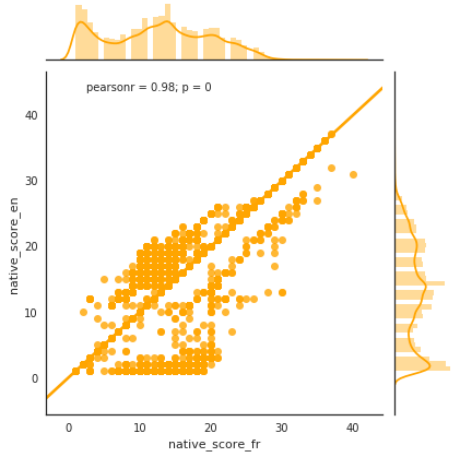
Le modèle des données a fait disparaître les acides gras.

La fréquence des vitamines, à l'état de traces en termes d'apports, sont représentées dans le modèle de données.

Les glucides et sels minéraux sont plus fréquents que les autres classes de nutriment. C'est sans doute une caractéristique de la production industrielle qui, pour l'expérience client, sur-représente cette classe d'aliments.

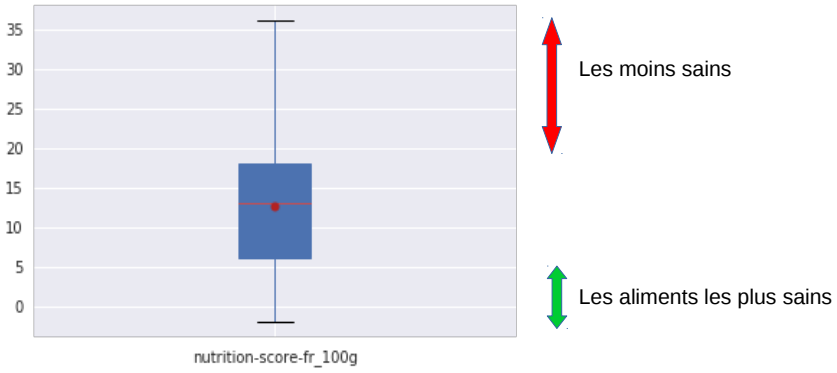
7 Analyse du scoring natif

7.1 Comparaison entre le scoring fr et uk.



Ces deux variables sont linéairement corrélées.
D'après le site d'open fact food : «Plus un produit est de bonne qualité, plus son score est bas»
https://www.youtube.com/watch?v=GAwTyEEHnOs

7.2 Critères de sanité et recommandations



Les 10 références alimentaires ayant le moins bon score :

- 18.0
- Dark Chocolate Coconut Chews
- Dark Chocolate Sea Salt & Turbinado Almonds
- Lion Peanut x2
- Milk Chocolate Peanut Butter Malt Balls
- Yogurt Pretzels
- Chili Mango
- Milk Chocolate Pretzels
- Butter Croissants
- Biscuit

Les 10 références alimentaires ayant le meilleur score :

- 6.0
- Pains chocolat x4 au beurre AOP
- Nice Tea, Raspberry
- Instant Oatmeal, Creamy Vanilla Bean & Honey
- Merl Veggi fresh vegetarischer Geflügelsalat
- Wok d'Autruche pré-grillé mariné
- Le Bon Paris, Avec Couenne (4 Tranches)
- L'emmental Francais
- Party Size Simply Naked Pita Chips
- poires Guyot
- Filets de sardines au naturel (2 parts)

8 Scoring par pondération

Une table de scoring est calculée.

Pour chacun des nutriments, les quantiles 1/4, 2/4 et 3/4 sont calculés.

Un poids est associé à chaque intervalle inter-quantile dont le signe dépend de la liste (+ ou -).

Méthode 1 : calcul de score_1 :

Valeur nutriment	score_1	
	Poids contribution >0	Poids contribution < 0
Val =0	0	0
quantile(1/4) < Val ≤ quantile(1/2)	1	-1
quantile(1/2) < Val ≤ max_extrem	2	-2

Méthode 2 : calcul de score_2 :

Valeur nutriment	score_2	
	Poids contribution >0	Poids contribution < 0
Val =0	0	0
0 < Val ≤ quantile(1/4)	1	-1
quantile(1/4) < Val ≤ quantile(1/2)	2	-2
quantile(1/2) < Val ≤ quantile(3/4)	3	-3
quantile(3/4) < Val ≤ max_extrem	4	-4

Le scoring d'un aliment va être la somme cumulée de ces poids.

Le calcul de ces quantiles est réalisé sur la base de la corrélation supposée entre l'énergie et l'apport en nutriments.

Une fois la pondération établie, un score est calculé pour chacun des aliments.

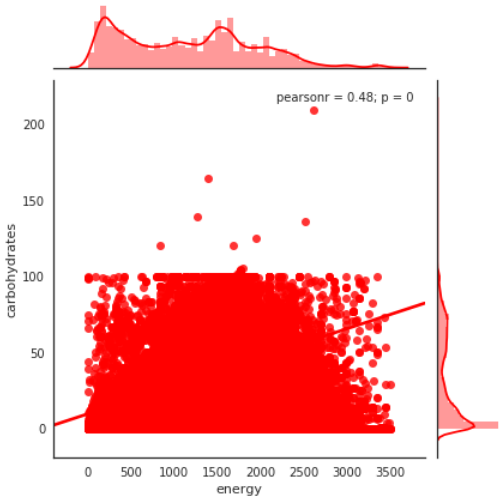
8.1 Corrélation energy / nutriments

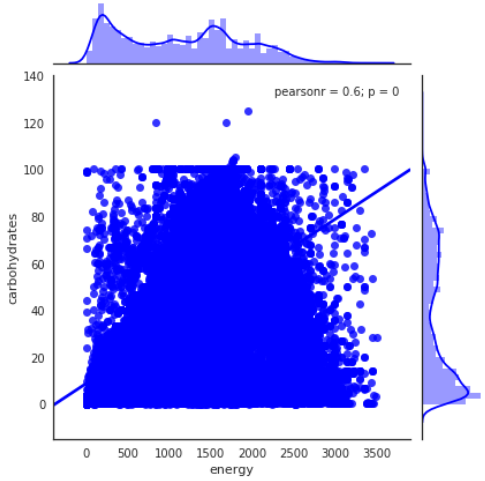
La littérature sur le sujet indique que les hydrates de carbone sont les nutriments qui contribuent le plus à notre apport énergétique.

Utilisation de la corrélation carbonydrates et energy.

Une corrélation est recherchée entre l'apport énergétique de ce nutriment et la variable energy.

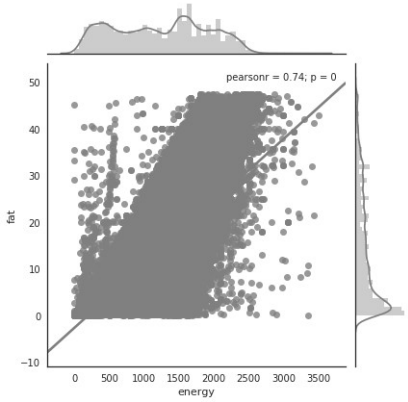
On prend le partie de contraindre le modèle afin d'optimiser cette corrélation linéaire.

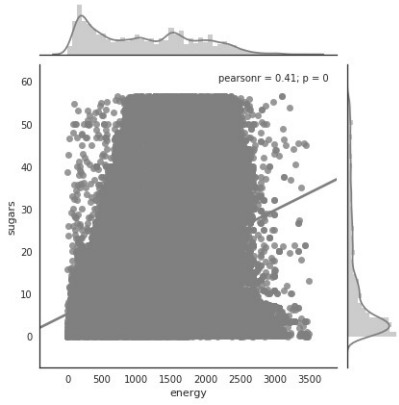




En traitant les outliers de l'hydrate de carbone, la corrélation est améliorée de 25 %. Les outliers représentent 12 % du modèle.

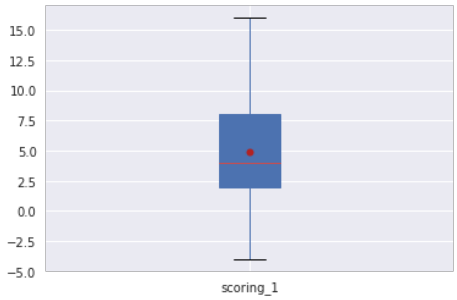
L'analyse bi-variée de la densité énergétique et des nutriments qui contribuent négativement à la sanité des aliments :

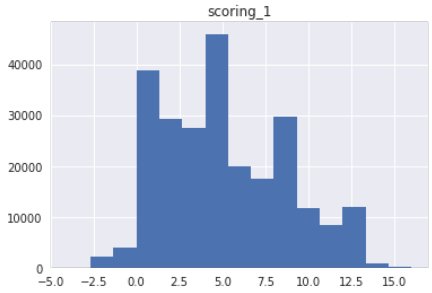




Le modèle de données prends en compte la corrélation entre les graisses la densité énergétique.
La relation avec le sucre est moins évidente.

8.2 Scoring méthode 1 (simplifiée)





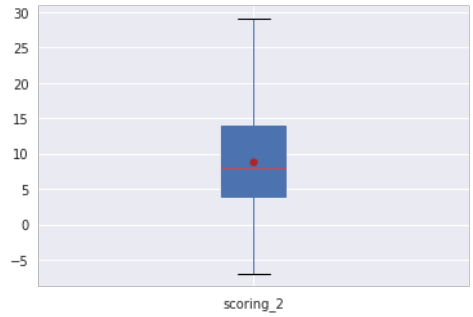
Scoring_1 : échantillon de 10 références alimentaires ayant le meilleur score :

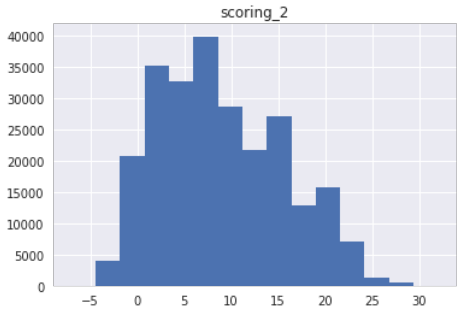
- 8.0
- Peanuts
- Organic Salted Nut Mix
- Organic Long Grain White Rice
- Organic Adzuki Beans
- Organic Penne Pasta
- Zen Party Mix
- Organic Golden Flax Seeds
- Organic Hazelnuts
- Organic Oat Groats
- Energy Power Mix

Scoring_1 : échantillon de 10 références alimentaires ayant le moins bon score:

- 2.0
- Sweeteners, Demerara Turbinado Sugar
- Marks % Spencer 2 Blueberry Muffins
- Sweeteners, Organic Fair Trade Sugar
- Vanilla Extract
- M&S Extrenely Chocolatey Milk, Dark & White Chocolate Biscuits
- diet lemonade by Sainsbury's
- Organic Unrefined Mascobado Sugar
- Pecan Halves
- Veggie Colin the Caterpillar
- Belgische Pralinen

8.3 Scoring méthode 2





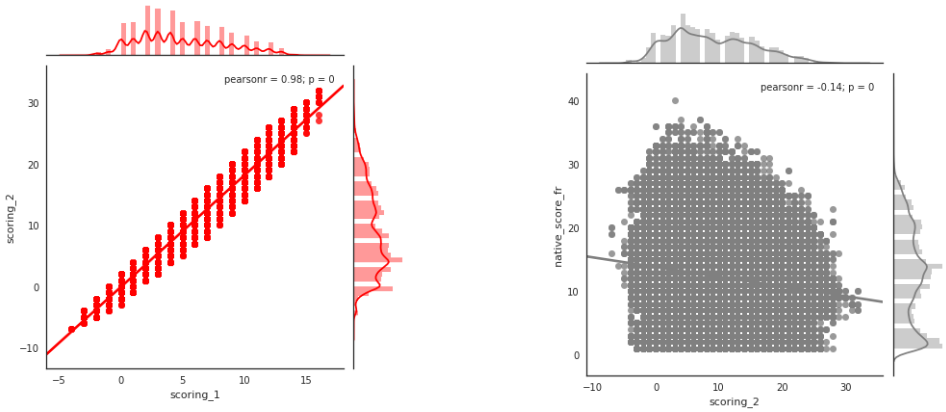
Scoring_2 : échantillon de 10 références alimentaires ayant le meilleur score :

- 14.0
- Peanuts
- Organic Salted Nut Mix
- Organic Long Grain White Rice
- Organic Muesli
- Organic Adzuki Beans
- Organic Penne Pasta
- Zen Party Mix
- Organic Golden Flax Seeds
- Cinnamon Nut Granola
- Organic Hazelnuts

Scoring_2 : échantillon de 10 références alimentaires ayant le moins bon score :

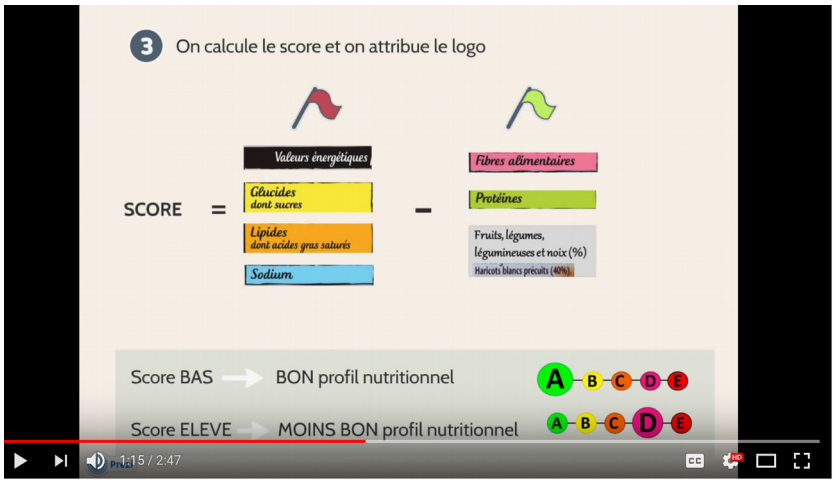
- 4.0
- Sweeteners, Demerara Turbinado Sugar
- Marks % Spencer 2 Blueberry Muffins
- Sweeteners, Organic Fair Trade Sugar
- Vanilla Extract
- M&S Extrenely Chocolatey Milk, Dark & White Chocolate Biscuits
- diet lemonade by Sainsbury's
- Organic Unrefined Mascobado Sugar
- Pecan Halves
- Veggie Colin the Caterpillar
- Mini Confettis

8.4 Corrélations entre scorings



Les deux méthodes de calcul du scoring sont corrélées.

Les deux méthodes de scoring, native et pondérées ne sont faiblement corrélées.
La méthode approximative de scoring par pondération va dans le sens du scoring natif qui stipule que plus la valeur du score natif est faible, meilleur est la sanité de l'aliment.



9 Recommandations pour le modèle

Le nettoyage du modèle a eu pour conséquences :

- de purger plus de 70 % des variables.
- De réduire de 50 % la population étudiée.

Le modèle issue des transformations successives a sûrement perdu de l'information nécessaire à la pertinence de son intégrité.

9.1 Prise en compte des outliers

Une prise en compte des outliers, notamment pour les aliments dont la densité énergétique est supérieure à 3500, permettrait d'affiner le modèle en prenant en compte des aliments qui échappent à l'analyse.

9.2 Renseignement de la base alimentaire

Un renseignement plus exhaustif de la base d'aliments permettrait d'affiner le modèle et prenant en compte des nutriments qui échappent à l'analyse. Pour ce faire, les nutriments d'aliments équivalents à la base pourraient être recherchés et imputés aux observations nono exploitables.

9.3 Extension du modèle

Certaines caractéristiques de la base alimentaire, comme les allergènes, méritent d'être prises en compte pour avoir un concept plus exhaustif de la sanité d'une référence alimentaire.

Le schéma de cette étude supporte l'extension de nouvelles variables.