Préparez des données pour un organisme de santé publique

Sommaire

Rappel de la mission

- **Objectif principal:** Préparer les données de la base Open Food Facts pour créer un système de suggestion ou d'auto-complétion des champs, aidant ainsi les usagers à remplir plus efficacement la base de données.

- **Méthodologie:** Commencer par nettoyer et explorer les données pour déterminer la faisabilité de cette application.

Démarche méthodologique de nettoyage des données

- **Importation des bibliothèques et chargement des données:** Utilisation de pandas et numpy pour manipuler les données.

- **Sélection des colonnes pertinentes pour l'analyse:** Focus sur les colonnes importantes pour notre étude.

Identification des features pertinentes

- **Importation des bibliothèques**
- **Chargement des données**
- **Sélection des colonnes importantes**

	code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	generic_name	quantity	...	ph_100g	fruits-vegetables-nuts_100g	collagen-meat-protein-ratio_100g	cocoa_100g	chlorophyl_100g	foo
0	0000000003087	http://world-fr.openfoodfacts.org/produit/0000...	openfoodfacts-contributors	1474103866	2016-09-17T09:17:46Z	1474103893	2016-09-17T09:18:13Z	Farine de blé noir	NaN	1kg	...	NaN	NaN	NaN	NaN	NaN	
1	0000000004530	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Banana Chips Sweetened (Whole)	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
2	0000000004559	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Peanuts	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
3	0000000016087	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055731	2017-03-09T10:35:31Z	1489055731	2017-03-09T10:35:31Z	Organic Salted Nut Mix	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	
4	0000000016094	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055653	2017-03-09T10:34:13Z	1489055653	2017-03-09T10:34:13Z	Organic Polenta	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	

5 rows × 162 columns

Compréhension des colonnes pour l'harmonisation

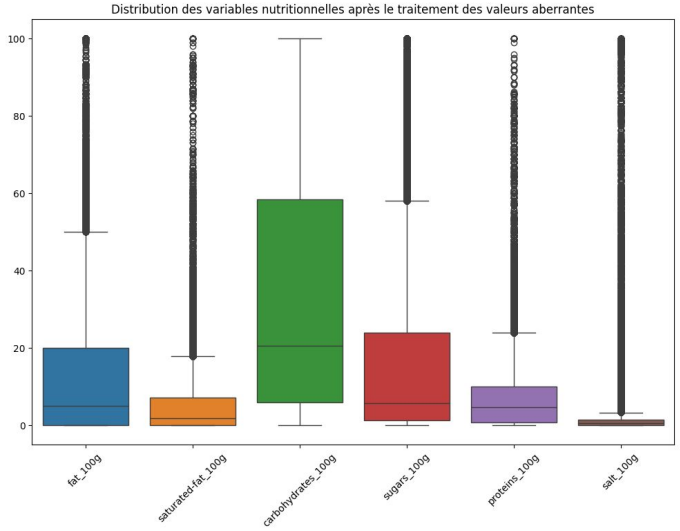
- **Analyse des valeurs uniques des colonnes:** Analyse des valeurs uniques de 'pnns_groups_1' et 'pnns_groups_2'.
- **Suppression des colonnes redondantes:** Suppression de 'pnns_groups_2' en raison de la redondance et de la non-nécessité.

Identification des valeurs manquantes et élimination des doublons

- **Calcul du pourcentage de valeurs manquantes:** Identification des colonnes avec des valeurs manquantes.
- **Techniques de traitement des valeurs manquantes:** Utilisation de l'imputation, de la suppression et d'autres méthodes.
- **Élimination des doublons:** Suppression des doublons en utilisant la colonne 'code'.

Traitement des valeurs aberrantes

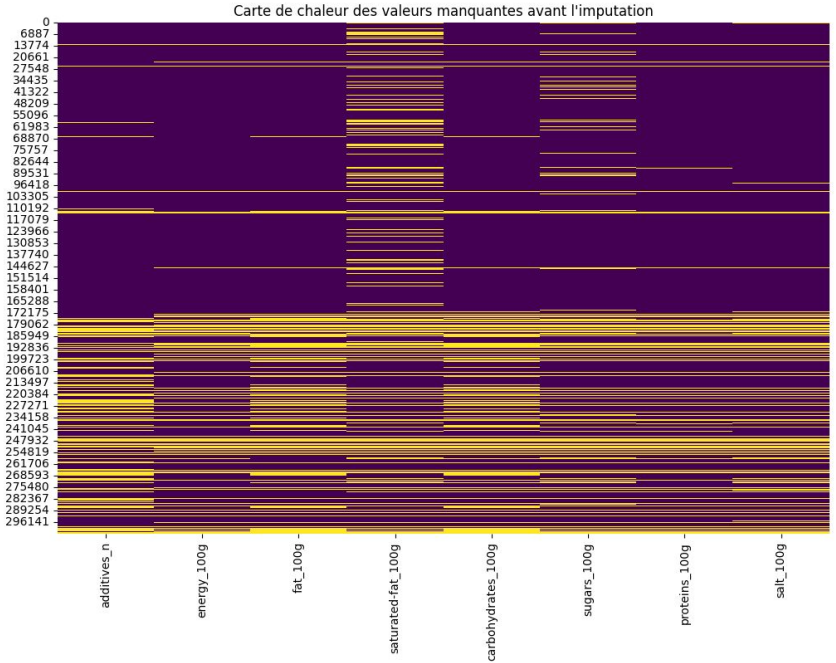
- **Visualisation des valeurs aberrantes avec des boxplots:** Identification des valeurs au-delà des quartiles.
- **Remplacement par la médiane des valeurs non aberrantes:** Utilisation de la médiane pour chaque colonne.
- **Imputation itérative pour les valeurs manquantes:** Utilisation de LinearRegression pour estimer les valeurs manquantes.



Traitement des valeurs manquantes avec SimpleImputer

- **Utilisation de SimpleImputer pour 'nutrition-score-fr_100g':** Stratégie de médiane pour imputer les valeurs manquantes.

- **Complétion du dataset:** Utilisation de cette colonne dans l'analyse sans introduire de biais significatif.

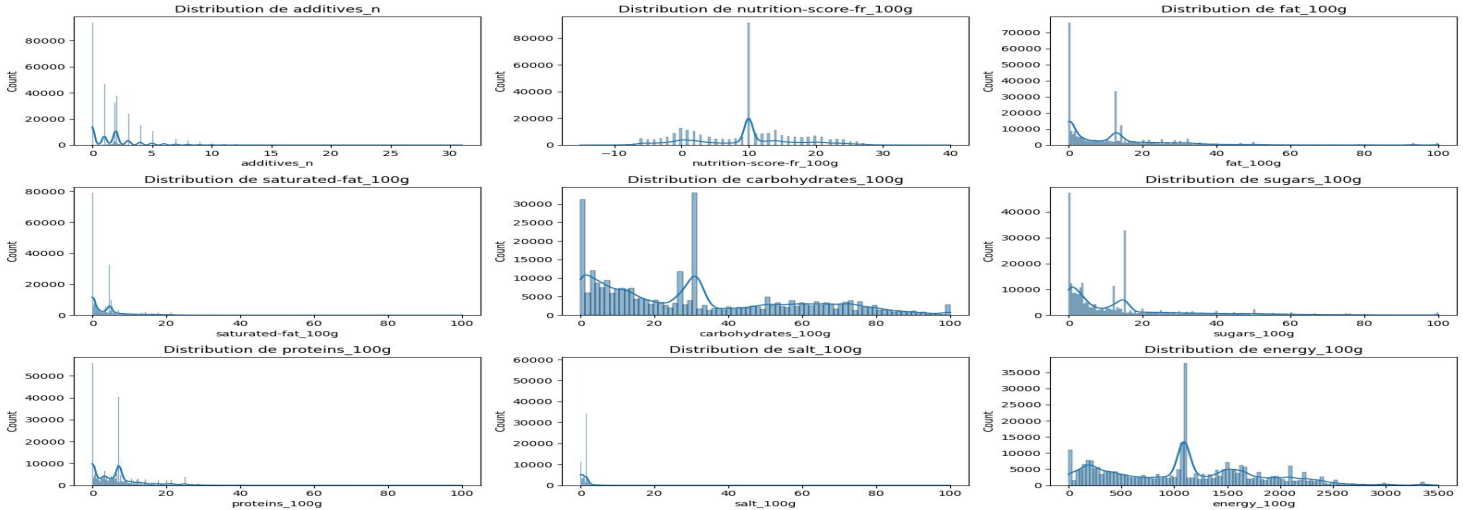


Analyse univariée

- **Statistiques descriptives pour chaque feature numérique:** Analyse des distributions de 'fat_100g' et 'energy_100g'.

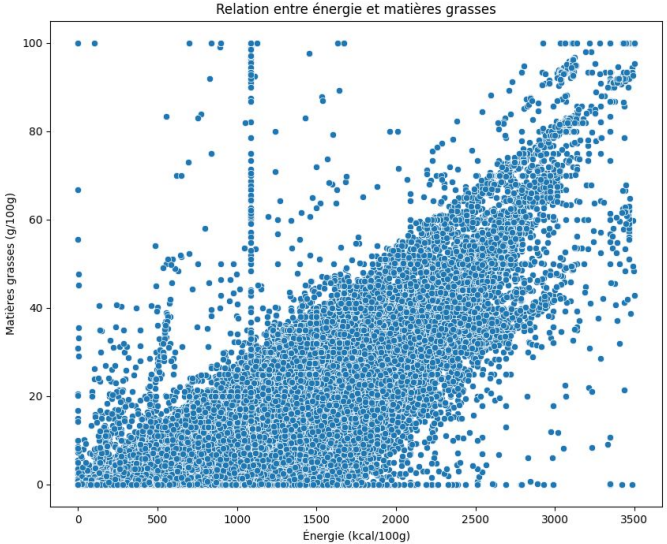
- **Observation des asymétries et des queues longues:** 'fat_100g' montre une asymétrie avec une longue queue vers la droite.

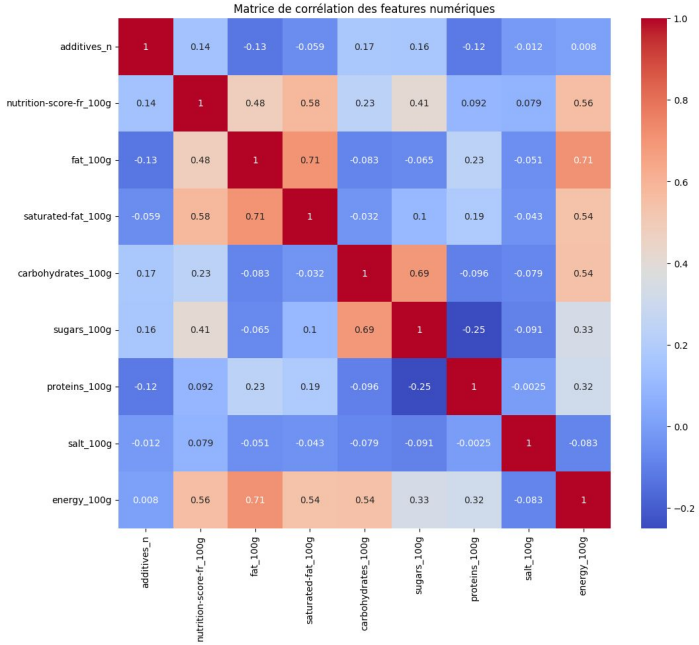
- **Centres et symétries:** 'energy_100g' est plus symétrique et centrée autour de la moyenne.



Analyse bivariée

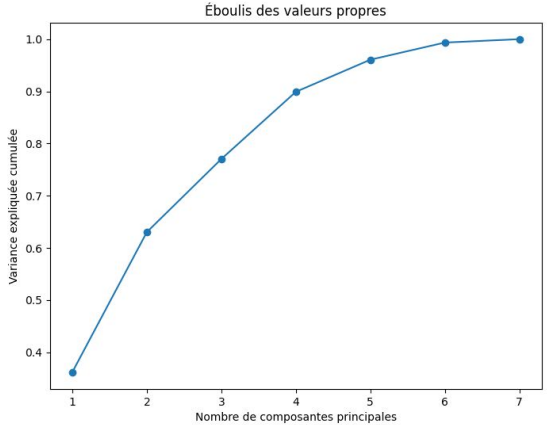
- **Création de scatterplots pour examiner les relations:** Examen des relations entre différentes paires de features.
- **Tendance positive entre 'energy_100g' et 'fat_100g':** Plus d'énergie implique généralement plus de matières grasses.
- **Utilisation de heatmap de corrélation:** Visualisation des relations entre toutes les paires de features numériques.





Analyse multivariée

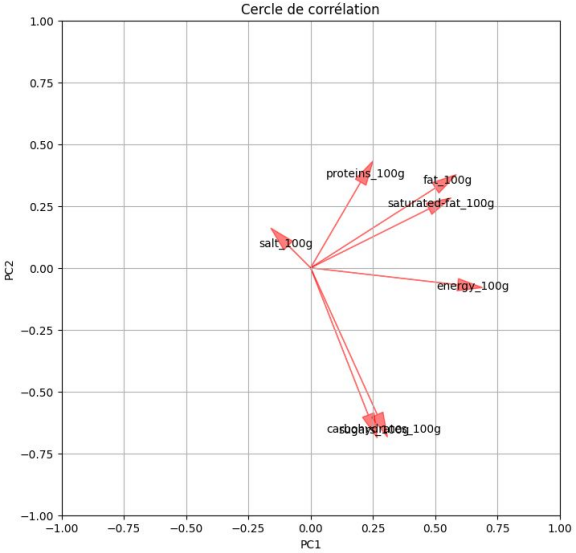
- **Analyse en Composantes Principales (ACP):** Réduction de la dimensionnalité et visualisation de la structure globale.
- **Éboulis des valeurs propres:** Les deux premières composantes principales expliquent la majorité de la variance.
- **Cercle de corrélation:** Des features comme 'fat_100g' et 'saturated-fat_100g' sont fortement corrélées avec la première composante.
- **ANOVA:** Vérificati onde la significativité des différences entre les groupes de Nutri-Score.



Analyse Multi-varié:

- **Analyse en Composantes Principales (ACP):** Réduction de la dimensionnalité des données et visualisation de la structure globale. Les deux premières composantes principales expliquent la majorité de la variance.

- **Cercle de corrélation de l'ACP:** Les features comme 'fat_100g' et 'saturated-fat_100g' sont fortement corrélées avec la première composante principale.



Démarche méthodologique d’exploration des données

- **ANOVA:** Vérification de la significativité des différences entre les groupes de Nutri-Score. Les résultats montrent des différences significatives pour toutes les features analysées.

- **Nutri-Score et classification des produits:** Le Nutri-Score est une mesure pertinente pour la classification des produits.

Faits pertinents pour l’application

- **Corrélation positive forte:** Les analyses de corrélation ont révélé des relations positives fortes entre certaines features, comme 'fat_100g' et 'saturated-fat_100g'.

- **Implication pratique:** Un produit riche en matières grasses est également susceptible d'avoir une teneur élevée en acides gras saturés.

Faits pertinents pour l’application

- **Amélioration de la complétude du dataset:** L'utilisation de l'imputation itérative et de SimpleImputer a considérablement amélioré la complétude de notre dataset.

- **Réduction des lacunes:** Comparaison des heatmaps avant et après imputation montrant une réduction significative des valeurs manquantes.

Faits pertinents pour l’application

- **Cohérence et complétude des données:** Les tests statistiques et les visualisations ont montré que les données nettoyées et imputées sont suffisamment complètes et cohérentes pour être utilisées dans un système de suggestion.

- **Boxplots selon le Nutri-Score:** Les produits avec un meilleur Nutri-Score (A et B) ont tendance à avoir des teneurs plus faibles en matières grasses comparés aux produits avec des scores D et E.

Conclusion et faisabilité du projet

- **Nettoyage et complétion des données:** Nos analyses ont montré que les données de la base Open Food Facts peuvent être nettoyées et complétées efficacement pour permettre l'auto-complétion des champs.

- **Cohérence et complétude:** Les données nettoyées sont cohérentes et suffisamment complètes pour être utilisées dans le système de suggestion proposé par Santé publique France.

- **Conformité avec le RGPD**: Ce projet utilise un dataset public ne contenant pas de données personnelles, mais nous avons respecté les grands principes du RGPD.

- **Principes du RGPD**: Les principes respectés incluent la licéité, la transparence, la limitation de la finalité et des données, l'exactitude et la limitation de la conservation.

Questions et discussion