Analysez des données nutritionnelles

Client : Lamarmite

Base de données : OpenFoodFacts

Version figée : lien de téléchargement

Objets de la mission

- Préparation d'une base de données pour la génération de recettes saines :
  - Préparer la base de données en supprimant les données inutiles
  - Créer des variables si nécessaire

- Réalisation d'une analyse exploratoire :
  - Comprendre les interactions entre les variables
  - Identifier les composantes permettant de différencier les produits sains de ceux à éviter

Problématique

Qu’est-ce qu’une alimentation saine ?

- D’après le Programme national nutrition santé (PNNS) :
  - Limiter les aliments gras, salés, sucrés et ultra-transformés
  - Privilégier l’eau comme boisson

  Par jour :
  - Manger au moins 5 fruits et légumes
  - 3 produits laitiers
  - 1 à 2 fois de la viande, volaille, poisson et œufs
- Un repère simple : le nutri-score, de A à E

Problématique

- Deux possibilités d'utilisation de la base de données :
  - Utiliser les produits comme des éléments de base. Exemple : les œufs, les huiles, la viande

  => On calcule la valeur énergétique d'une recette à partir de ses composants

  - Utiliser les ingrédients des produits transformés pour en extraire les associations. Exemple : on souhaite une recette à base de chocolat, on constate que les ingrédients souvent associés sont la farine et le sucre.

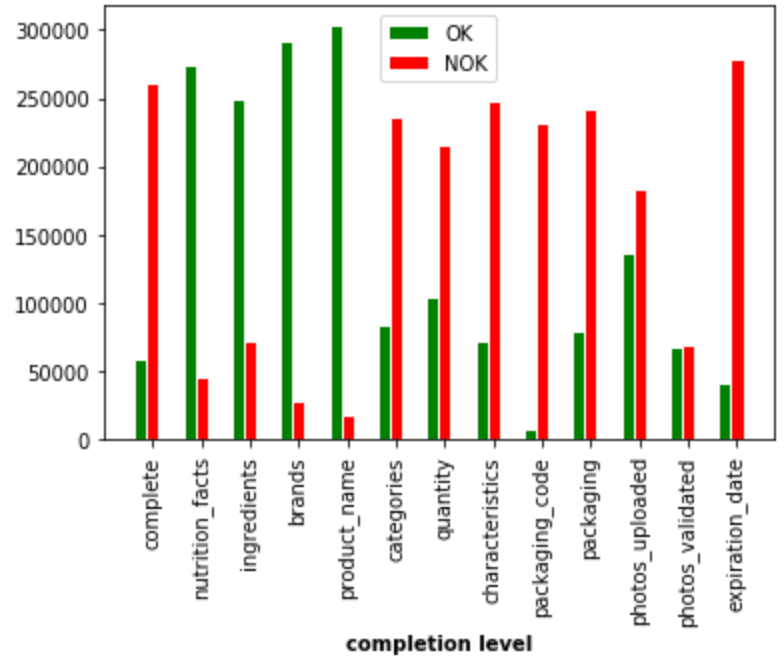
  => On propose une recette avec ces éléments

- S'assurer que les produits correspondent bien aux pays de l'utilisateur (disponibilité des produits et habitudes d'association)

1. Nettoyage des données

a. De nombreuses valeurs manquantes

- 162 colonnes pour plus de 320 000 produits
- 76% de données manquantes
- 34 colonnes remplies à plus de 50%
- 16 colonnes complètement vides
- Seulement 59000 produits qui sont considérés comme complètement remplis



Des données reprises des emballages des produits, par nature incomplets

1. Nettoyage des données

b. Des lignes et colonnes en double

- Des doublons sur les codes des produits
- Des colonnes qui contiennent les mêmes informations (anglais et français)
- Des colonnes qui ont la même signification. Exemple folates_100g et vitamine-b9_100g mais pas les mêmes données

=>Regroupement des valeurs et suppression des doublons

1. Nettoyage des données

c. Des valeurs aberrantes

- Des nutriments à plus de 100g pour 100g de produit ou à moins de 0
- Des valeurs d'énergie > 4000kJ par 100g (limite théorique : 3700)

=> Remplacement par des valeurs nulles

1. Nettoyage des données

d. Traitement des données

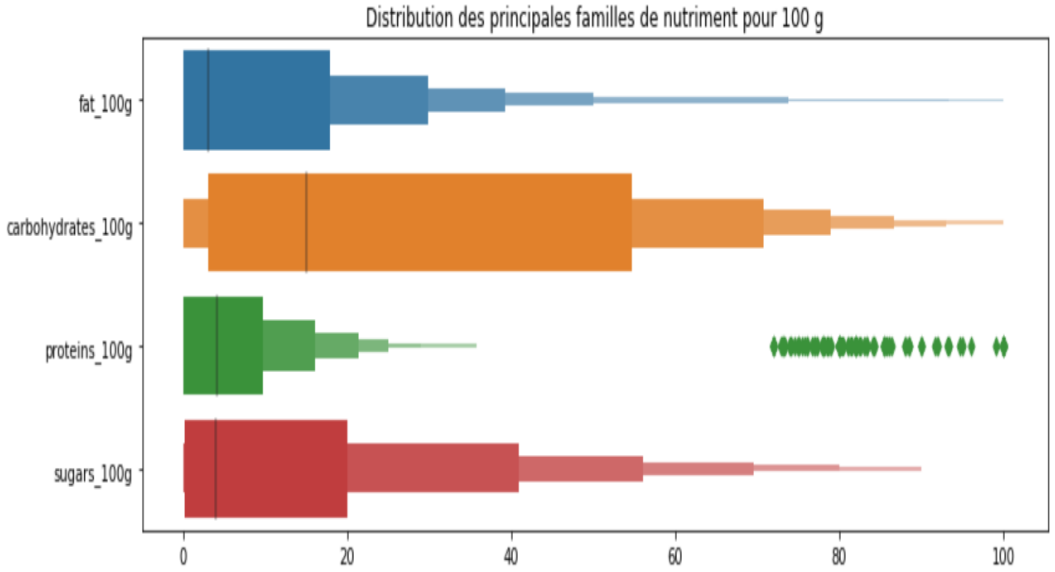
- Forte corrélation linéaire entre l'énergie et ses principales sources : les lipides, les glucides et les protéines

  ⇒Remplacement des valeurs manquantes
- Remplacement des valeurs nutritionnelles vides par des 0 lorsque la famille mère est nulle. Exemple : si glucides = 0 alors sucres = 0
- Manipulations du champ « ingrédients » pour la standardisation et la mise en forme de liste
- Nouvelles colonnes : nombre d'ingrédients, produit biologique ou non, décomposition de la colonne quantité

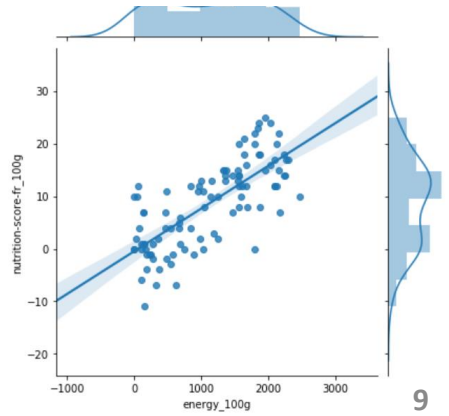
2. Analyse des données

a. Analyse des relations entre les principales variables

- Des nutriments pour 100g qui ne sont pas répartis de la même manière



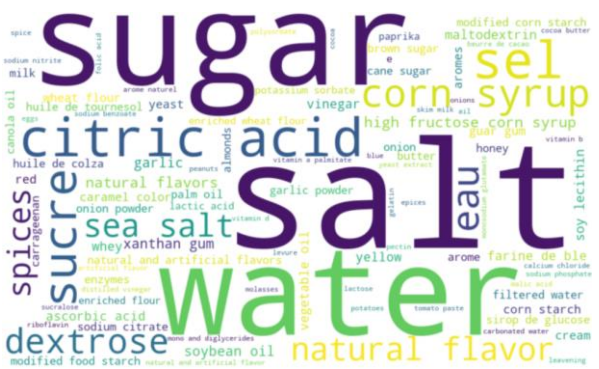
- Comme cela est prévu, le nutriscore est lié aux variables d'énergie, de graisses saturées, des sucres, sels, fibres et protéines (corrélation linéaire : $R^2 = 0,85$).

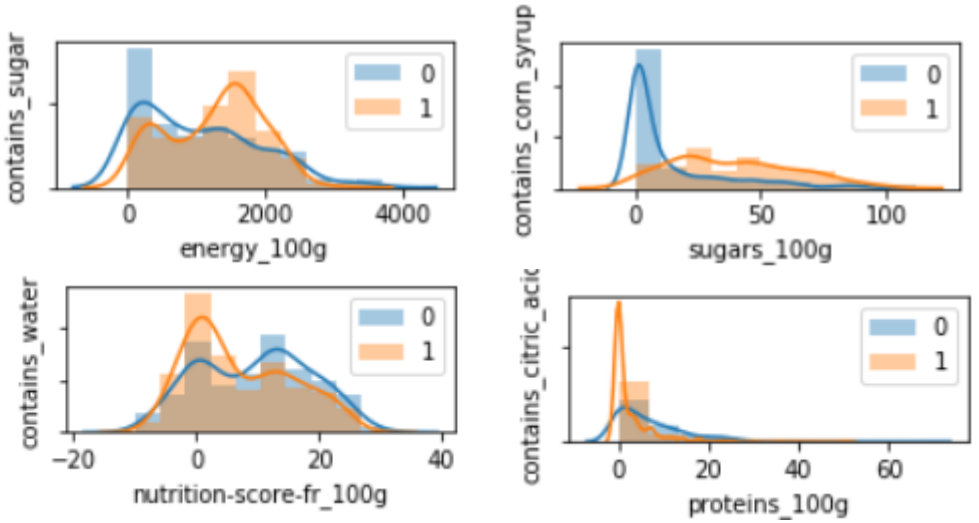


2. Analyse des données

b. Analyse de l'impact de la présence ou de l'absence d'ingrédients

- Certains ingrédients sont très utilisés
- Les ingrédients les plus présents ont des effets notables sur les variables :
  - Sucre sur l'énergie
  - Sirop de maïs
  - Eau sur le nutriscore
  - Acide citrique associé aux protéines





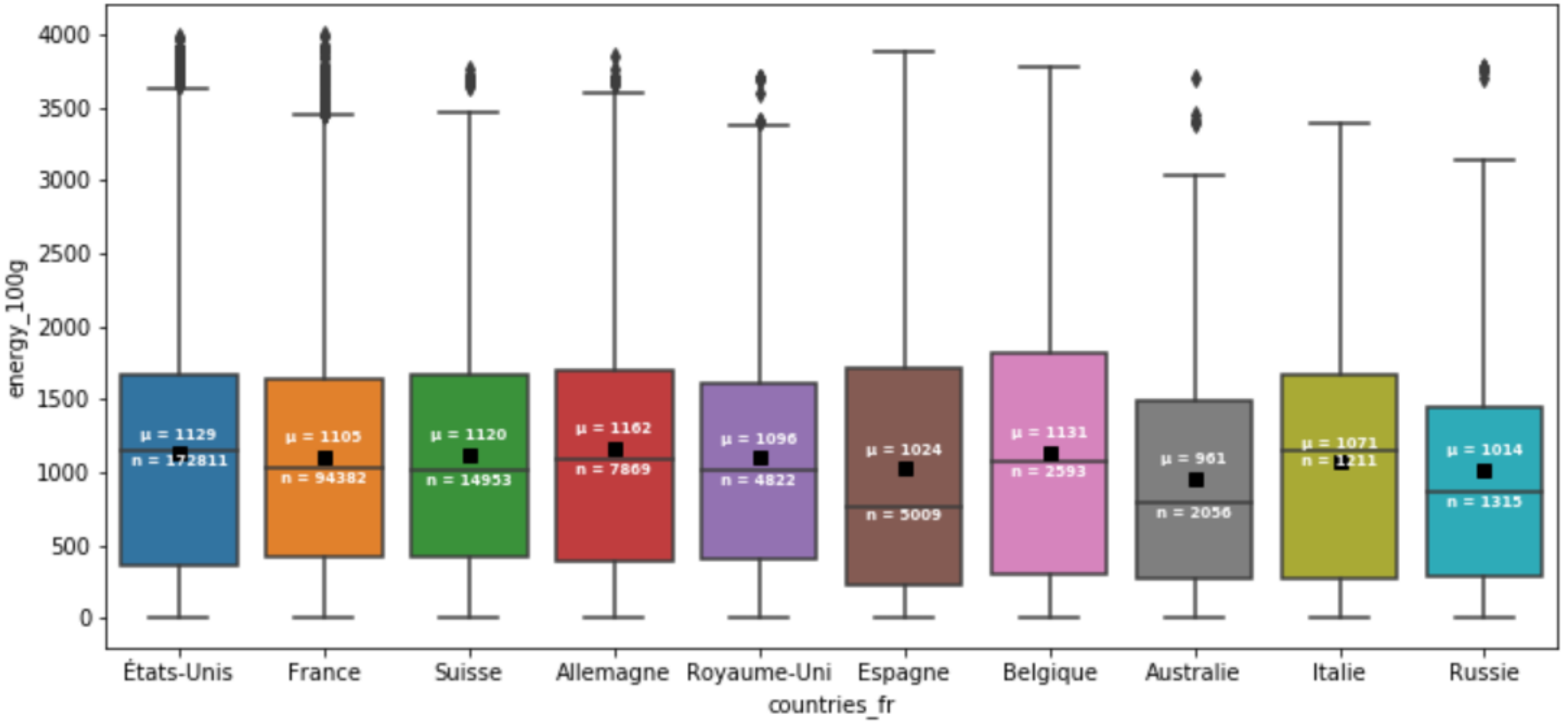
2. Analyse des données

c. Feature engineering

- Décomposition de la colonne "quantité" pour en extraire les valeurs numériques (en mL ou g)
- Calcul du nombre d’ingrédients
- Colonne binaire pour savoir si le produit est bio d’après les labels
- Colonne binaire pour savoir si le produit est une boisson et analyse des différences entre boissons et aliments sur les valeurs de nutrition

2. Analyse des données

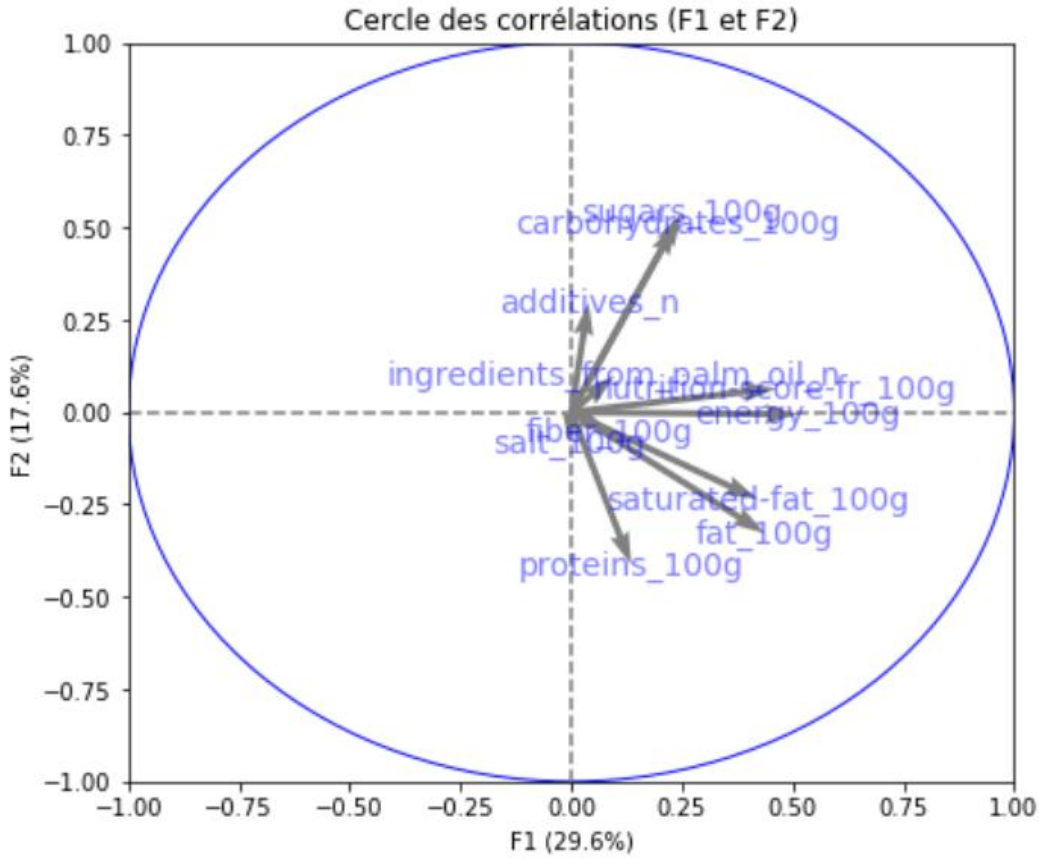
d. Analyse par pays



Des différences relativement peu marquées entre les pays sur le plan énergétique

3. Classification et réduction dimensionnelle

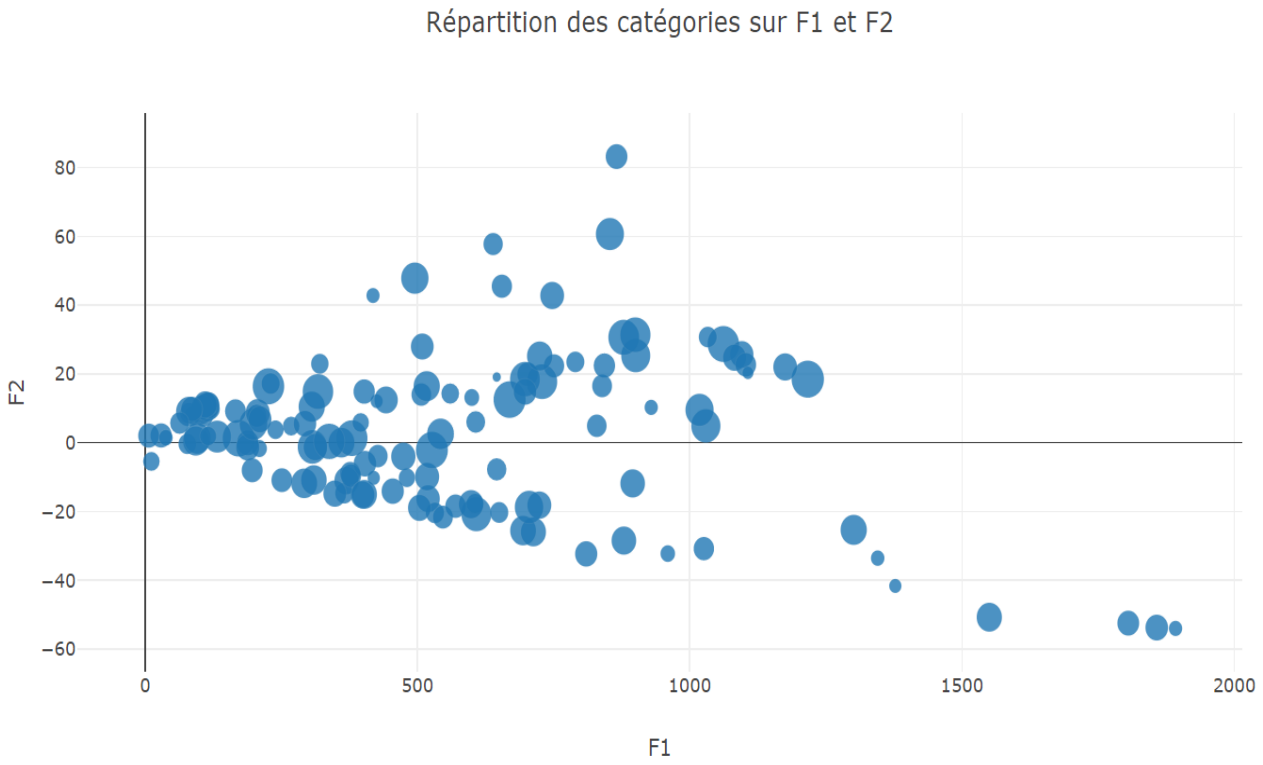
a. Analyse en composantes principales



- 2 premiers plans factoriels :
  - 1. Type d'énergie (transformée : sucres et additifs vs. naturelle : protéines et lipides)
  - 2. Quantité d'énergie
- Variables corrélées rassemblées

3. Classification et réduction dimensionnelle

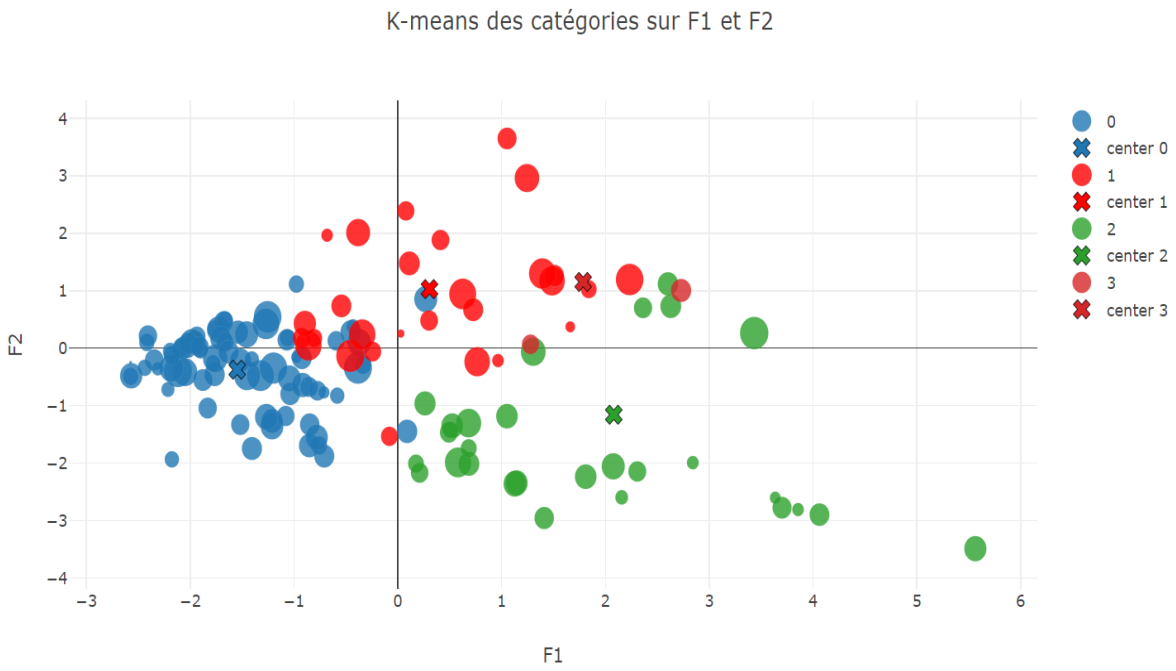
a. Analyse en composantes principales



- Les catégories agrégées par leurs valeurs moyennes sont similaires quand elles sont proches

3. Classification et réduction dimensionnelle

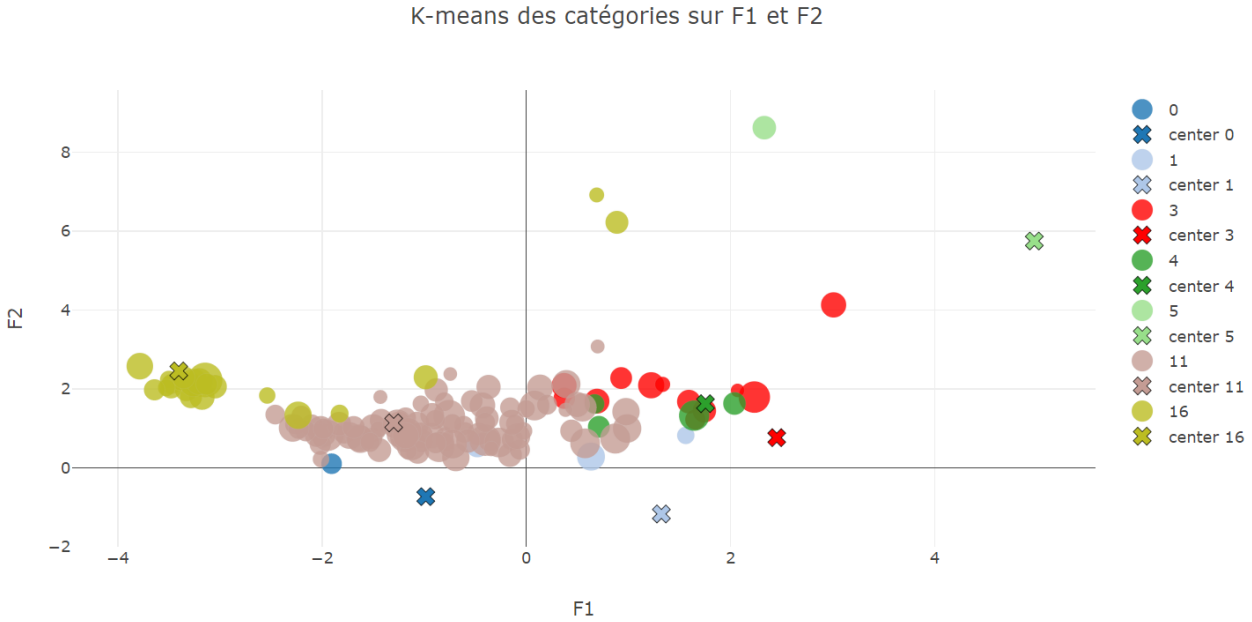
b. Application du K-Means sur les moyennes et les plans projetés



- 4 familles d'après la méthode du coude
- Des produits que le PCA avait rapprocher sont séparés :
  - Sablés / Biscuits / Gauffres
  - Charcuterie / Lardons

3. Classification et réduction dimensionnelle

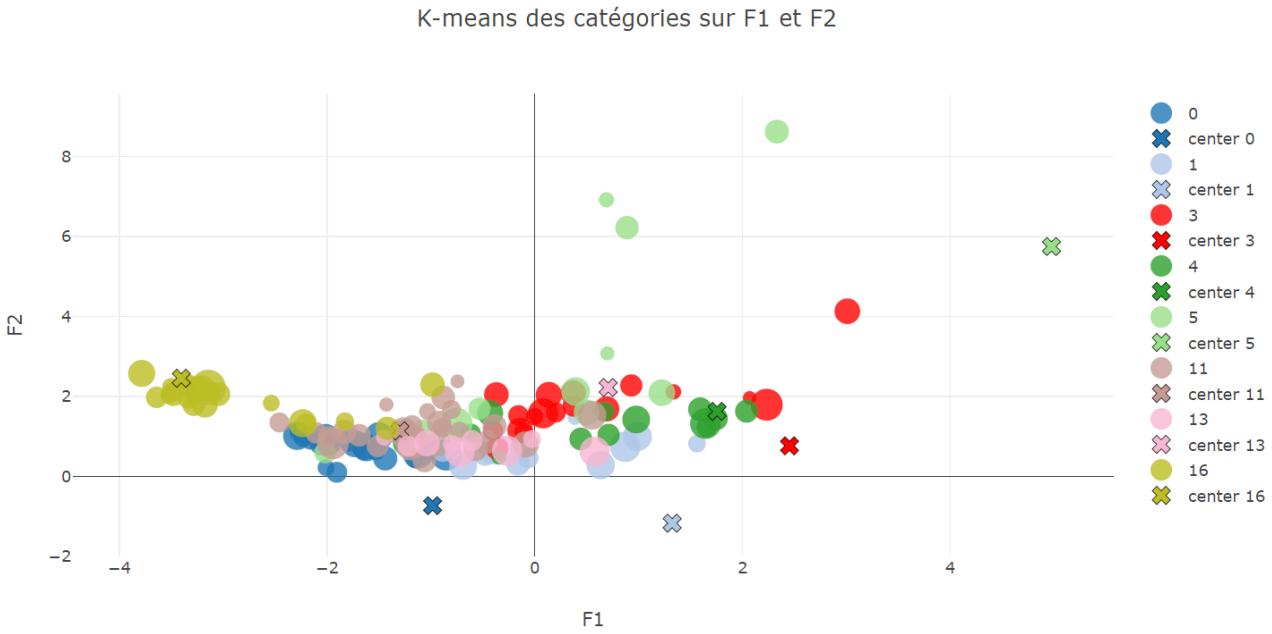
c. K-Means sur toutes les valeurs



- K-Means au niveau de chaque produit
- PCA après K-Means et agrégation par catégorie pour visualisation
- Beaucoup de clusters créés pour quelques valeurs extrêmes
- Un cluster central trop important

3. Classification et réduction dimensionnelle

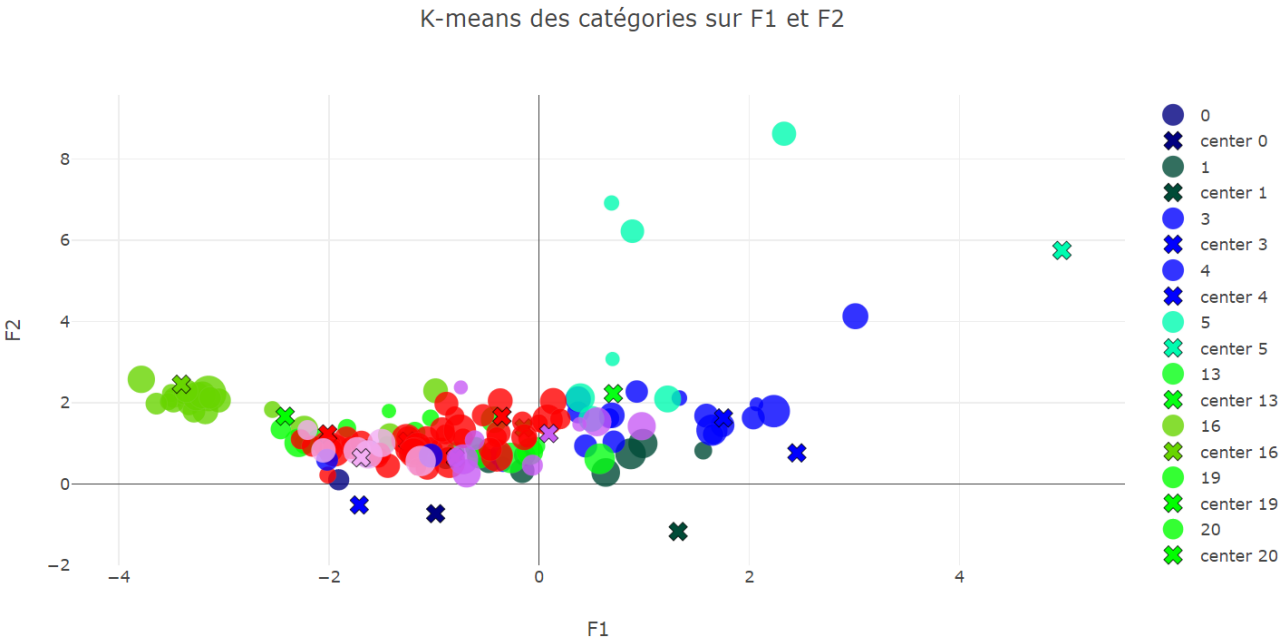
d. Décomposition du cluster central : 1.pondération



- On pondère l'attribution d'un cluster en fonction du nombre d'éléments dans chaque cluster
- Rassemblement assez cohérent de catégories
- Exception : rassemble les liquides malgré les divergences (eaux et crèmes)

3. Classification et réduction dimensionnelle

e. Décomposition du cluster central : 2.nouveau clustering



- 17 clusters affichés après pondération
- Bonne séparation des groupes même si les liquides restent ensemble
- Des produits qu'il est recommandé de peu consommer sont regroupés (clusters 1, 3 et 4)

Conclusion

a. Conclusion de l’analyse

- Les 2 premiers plans factoriels du PCA sur les principales valeurs montrent une zone où les produits sont à éviter : valeurs élevées sur F1/F2
- La classification permet de regrouper les groupes de produits à privilégier où à éviter
- On peut se servir de ces deux informations pour construire un modèle prédictif de la « qualité » d’un produit d’après ces 2 éléments.

Conclusion

b. Comment identifier les produits pertinents

- Des produits sains :
  - Regarder le nutri-score du produit s'il est présent, A et B sont de bons produits, les autres sont à éviter
  - Utiliser le PCA et/ou le K-Means pour identifier les produits à éviter

- Des produits disponibles :
  - S'assurer de la concordance entre le pays de l'utilisateur et le pays où le produit est disponible

Conclusion

c. Générer des recettes en sélectionnant les meilleurs produits

- Proposer une recette en choisissant les meilleurs ingrédients :

  => Pour un ingrédient de la recette, retourner ceux de la base qui correspondent et qui ont les meilleurs scores. Par exemple, quels sont les meilleurs chocolats sur le plan nutritionnel ?

- Proposer les associations les plus pertinentes :

  => Pour un ingrédient, on retourne les plus fréquemment associés dans les produits qui ont un bon score. Par exemple : quels sont les ingrédients les plus souvent associés au chocolat pour lesquels le produit final a une note de A ou B ?