Kelly Noëlla Sota 20/10/2024

Projet: Open Food Facts

Rapport de projet : nettoyage des données, détection des valeurs aberrantes et analyse à l'aide de l'ACP

1. **Outils et environnement**

Le projet a été développé à l'aide de Visual Studio Code (VSCode), un IDE puissant et flexible.

J'ai utilisé Data Wrangler pour un aperçu initial approfondi de l'ensemble de données, combiné avec des fonctions Pandas essentielles telles que `.info()`, `.describe()` et `.shape()` pour une compréhension globale de la structure de l'ensemble de données, y compris les types de données, la plage de valeurs et la forme globale.

2. **Prétraitement des données**

A. **Gestion des valeurs manquantes**

Les valeurs manquantes se trouvaient dans plusieurs colonnes. Certaines colonnes présentaient des lacunes importantes, ce qui nécessitait des approches différentes pour leur traitement.

Méthodes :

- Suppression : pour les colonnes contenant plus de 60 % de données manquantes, j'ai choisi de supprimer ces colonnes, réduisant ainsi l'ensemble de données à un nombre de fonctionnalités plus gérable.

- Imputation :

    - Pour la colonne `countries_fr`, j'ai utilisé KNN algorithm, qui estime les valeurs manquantes en fonction des voisins les plus proches.

- Pour les autres colonnes, j'ai utilisé l'imputation moyenne (mean) ou j'ai rempli avec « NA » le cas échéant, en veillant à ce que les données soient aussi complètes que possible pour l'analyse ultérieure.

B. Détection et traitement des valeurs aberrantes

- Visualisation : j'ai utilisé des boxplots pour identifier visuellement les valeurs aberrantes dans les colonnes numériques.

- Méthode Z-score : j'ai également utilisé la méthode statistique Z-score pour détecter les anomalies, en signalant les valeurs avec un Z-score supérieur à 3 comme des valeurs aberrantes.

- Correction : après avoir identifié les valeurs aberrantes, je les ai traitées en conséquence, en comparant l'ensemble de données avant et après le traitement des valeurs aberrantes à l'aide de visualisations telles que des boxplots.

3. Analyse univariée et multivariée

A. Analyse univariée

J'ai identifié des tendances clés, en utilisant des visualisations comme des boxplot, histogramme, scatterplot, telles que :

- Le top 5 des produits les plus fréquents et des meilleurs créateurs.

- Les 5 premiers pays qui produisent le plus de produits.

- Répartition des produits selon leur grade nutritionnel, le grade D étant le plus fréquent, représentant 69,5% des données.

B. Analyse multivariée

J'ai exploré les relations entre différentes variables, telles que :

- Sodium et Sel (avec la relation Sel = Sodium * 2,5).

- Énergie et matières grasses.

- La relation entre nutrition_grade_fr et nutrition-score-fr_100g.

4. Analyse en composantes principales (ACP)

J'ai divisé les données en colonnes numériques et catégorielles. Les colonnes catégorielles ont été encodées en utilisant one-hot encodage, mais la colonne « product_name » avait trop de valeurs distinctes, ce qui a donné lieu à un ensemble de données de grande dimension qui ne pouvait pas être traité efficacement. Par conséquent, j'ai limité l'ACP aux colonnes numériques.

Les colonnes numériques que j'ai standardisées pour l'ACP comprenaient :

`fat_100g`, `fat-satured_100g`, `trans-fat_100g`, `cholesterol_100g`, `carbohydrates_100g`, `sucres_100g`, `fiber_100g`, `proteins_100g`, `sel_100g`, `sodium_100g`, `vitamin-a_100g`, `vitamin-c_100g`, `calcium_100g`, `fer_100g`, `energie_100g`, `nutrition-score-fr_100g`.

Les valeurs manquantes ont été complétées avec la moyenne avant d'exécuter l'ACP.

Exécution de l'ACP : j'ai sélectionné 2 composantes principales pour l'analyse et visualisé le cercle de corrélation et les scatterplots pour mieux comprendre comment les variables contribuaient aux composantes principales.

5. Principales informations

- Produits les plus fréquents : Les produits de catégorie D étaient les plus courants.

- Matrice de corrélation : Le heatmap a mis en évidence une forte relation entre l'énergie et les carbohydrates, ainsi qu'entre les protéines et l'énergie.

- Sodium et sel : une relation linéaire entre le sodium et le sel était clairement visible dans le scatterplot.

- Influence énergétique : L'énergie semble influencer la nutrition grade.

- Composante principale 2 (PC2) : La plupart des variables ont montré une corrélation par rapport à PC2.