Projet 2

Parcours Machine Learning

Santé publique France

Concevoir une idée d'application au service de la santé publique

OPEN FOOD FACTS

Sommaire

Présentation de l’Application SweetMeat

- Envie d’une petite friandise alors que vous suivez un régime stricte? L’appli SweetMeat vous propose un top 5 de vos snacks préférées les plus **sains** parmi celles disponibles sur le marché.
- Régalez vos papilles sans culpabiliser!
- SweetMeat s’engage à respecter l’environnement en promouvant des produits **éco-responsables**.

EARTH FRIENDLY

Fonctionnement de l’application

- Scan du produit
- Check sur la base de données pour vérifier que le produit est répertorié

- Produit sain et éco responsable proposé par SweatMeat

Nettoyage des données

Étapes du nettoyage

- Chargement d'un chunk du jeu de données originel contenant 50.000 entrées ainsi que 182 colonnes.
- Sélection des colonnes estimées pertinentes pour notre idée d'application (ex. le nutriscore, les variables liées au calcul du nutriscore, pnns_groups_1 ainsi que pnns_groups_2 pour nous renseigner sur le type de produit alimentaire, cf. le notebook pour plus de précision). Notre jeu de données compte désormais 1875319 entrées et 21 colonnes.
- Filtrage sur les produits de type snacks (183760 produits)
- Suppression des doublons du jeu de données (ie. produits avec le même code barre, qui sont au nombre de 122).
- Fusion des catégories redondantes des colonnes pnns_groups_1 et pnns_groups_2 (ex. "pastries" et "Pastries").
- Filtrage des colonnes qui ont plus de la moitié de valeurs manquantes pour gagner en terme de qualité de donnée (3 colonnes filtrées).

Étapes du nettoyage (suite)

- Détection et filtrage des valeurs aberrantes (par ex: les valeurs qui ne sont pas comprises entre 0 et 100 de variables concernant la teneur en un nutriment donnée, ex. : glucides). Par cette opération, notre jeu de données est passé de 183648 produits à 82672)
- Imputation des valeurs manquantes à l'aide de KNNImputer. On s'est assuré que l'imputation n'impactait pas nos données en comparant les moyenne, médiane et le 3e quartile du jeu de données avant et après imputation. Les différences sont marginales après imputation.
- Notre dataset originel contenait 1875319 entrées et 186 colonnes. Après le nettoyage, il n'en compte plus que 82672 et 19 colonnes.

Label Eco-friendly

- Création d'une variable de type booléen indiquant si le produit est éco-friendly ou pas.
- On part du principe que seuls les produits provenant des pays proches (dans ce cas la France) sont éco-friendly, pas les autres.

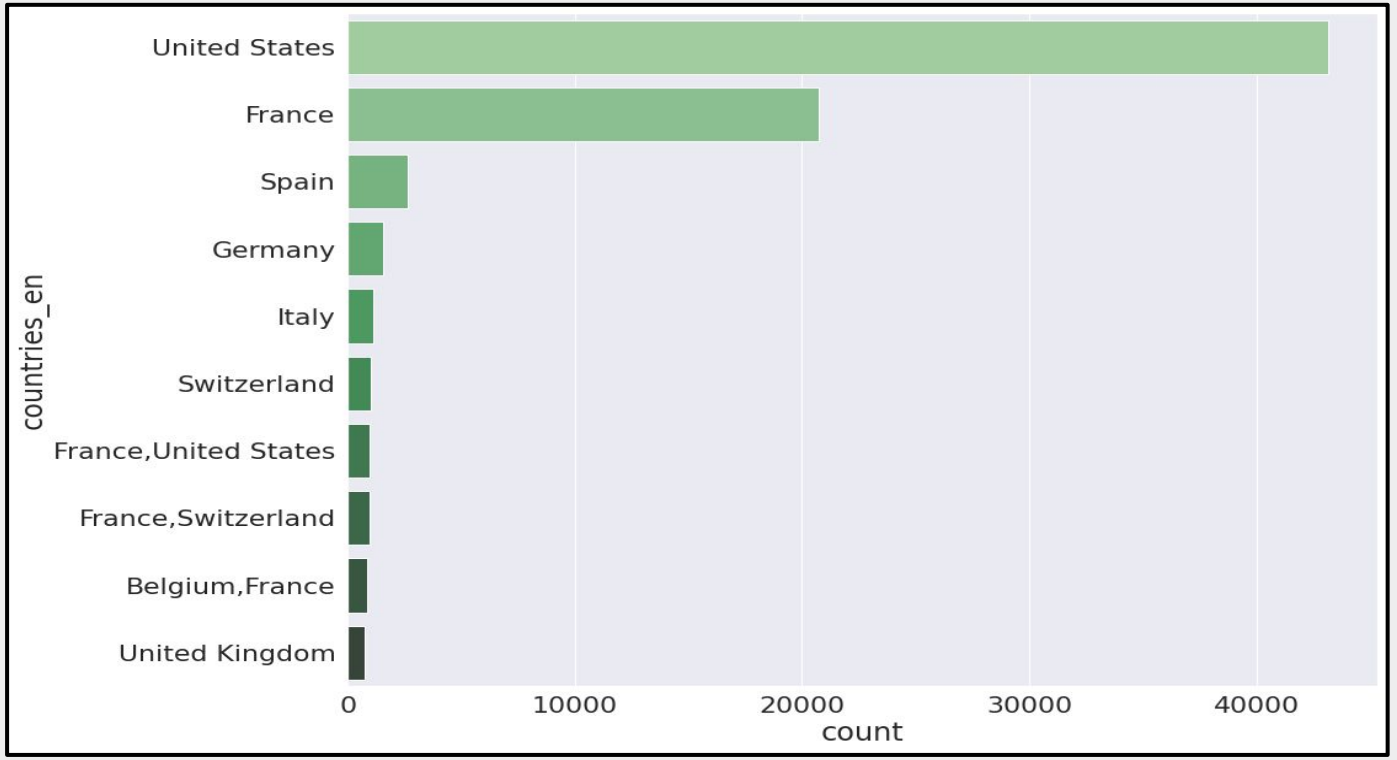
Pays	Eco-friendly
France	**x**
Etats-Unis	**-**
Italie	**x**
Japon	**-**

Analyse exploratoire

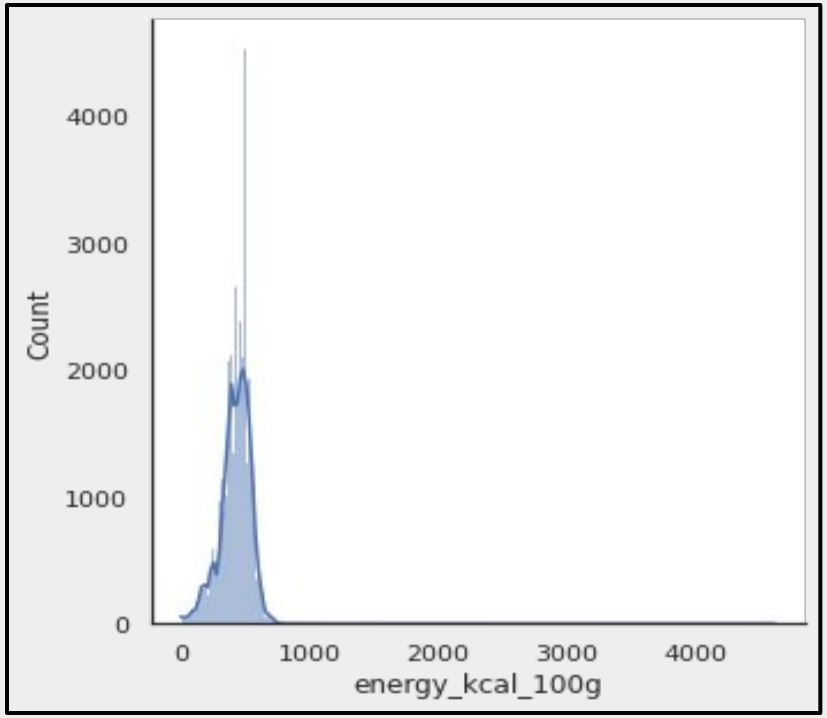
Analyse univariée

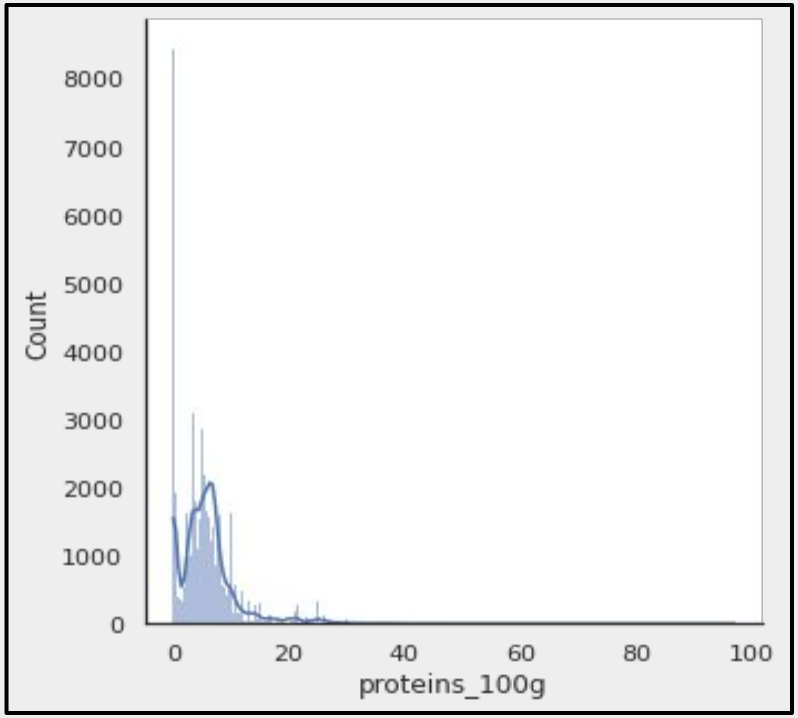
➔ L’analyse univariée nous permettra de décrire le comportement d’une variable, nous renseignera sur la nature de sa distribution ainsi que sur les variables utiles à l’application.

Top 10 des pays les plus représentés

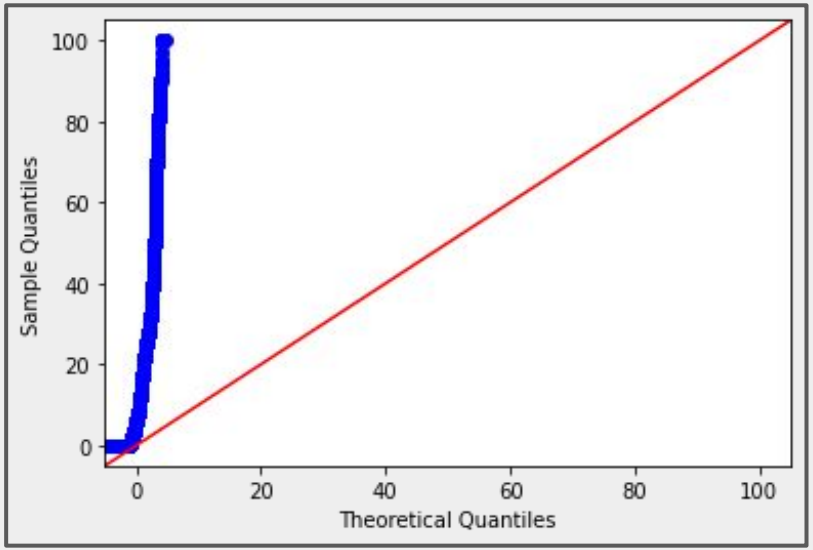


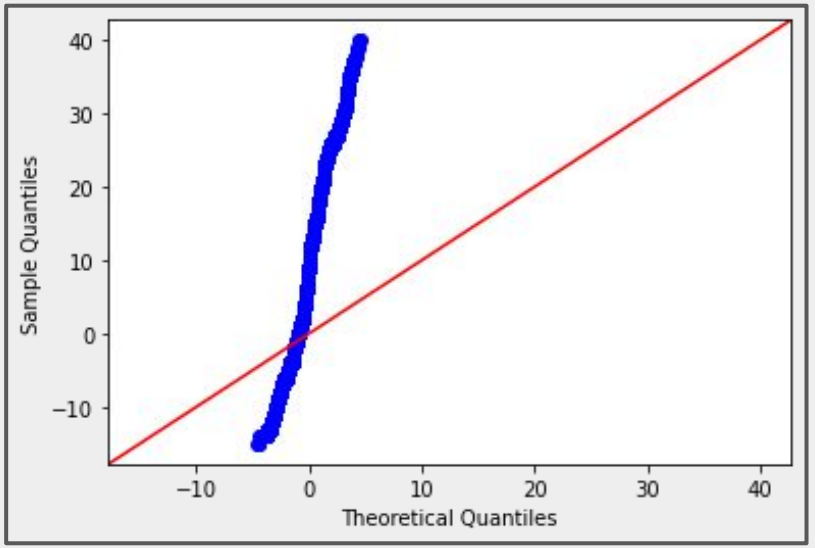
Distribution des variables quantitatives



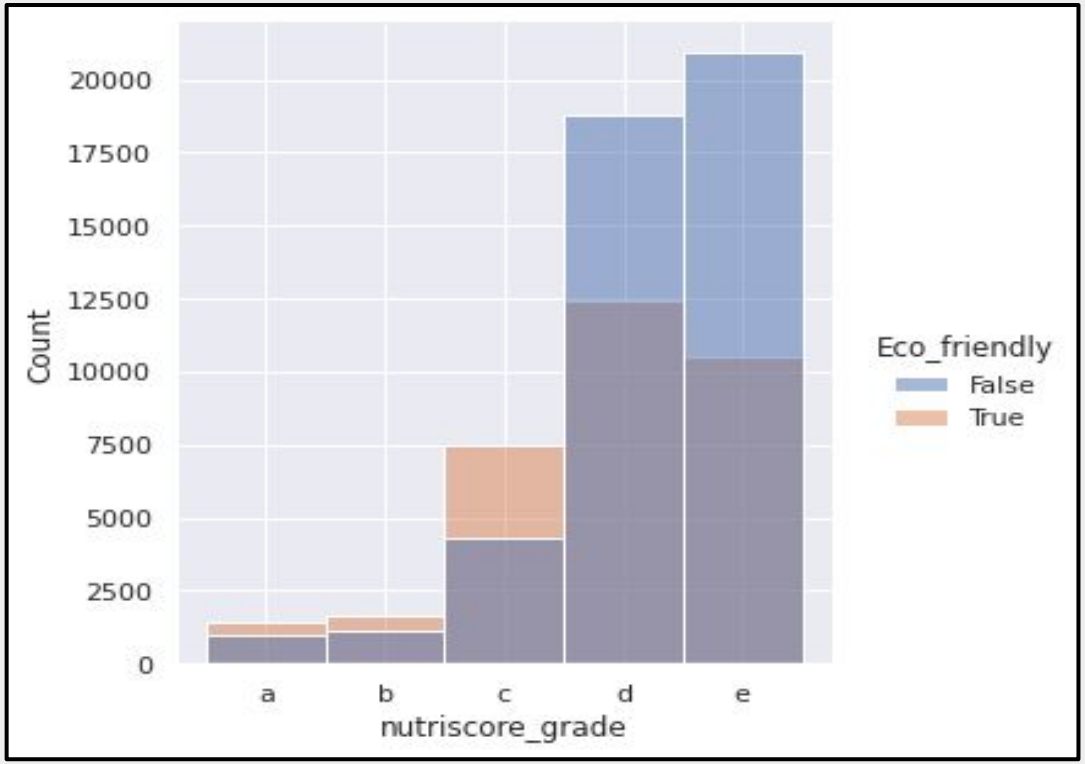


Test de normalité des variables quantitatives





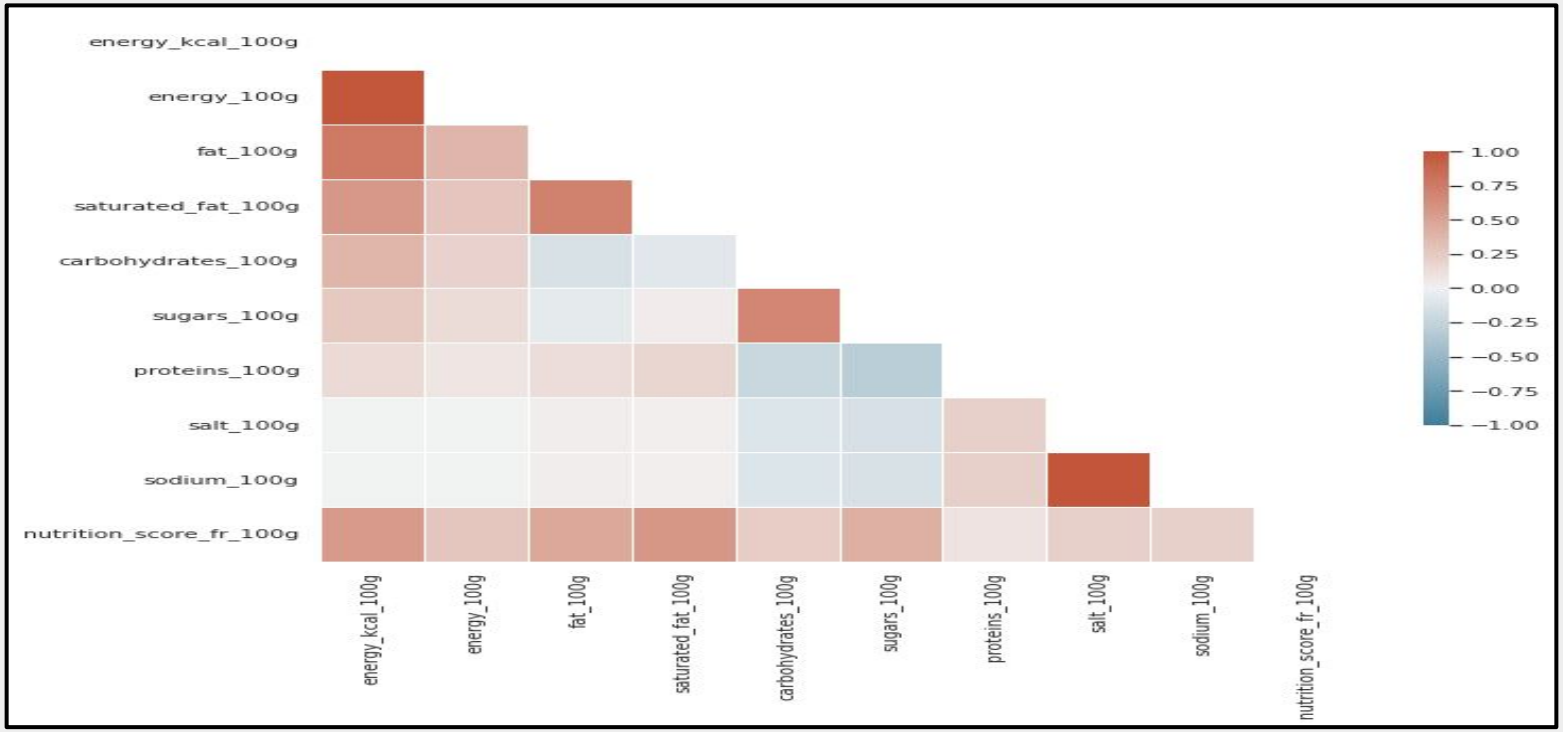
Distribution du grade nutriscore selon l'éco-responsabilité



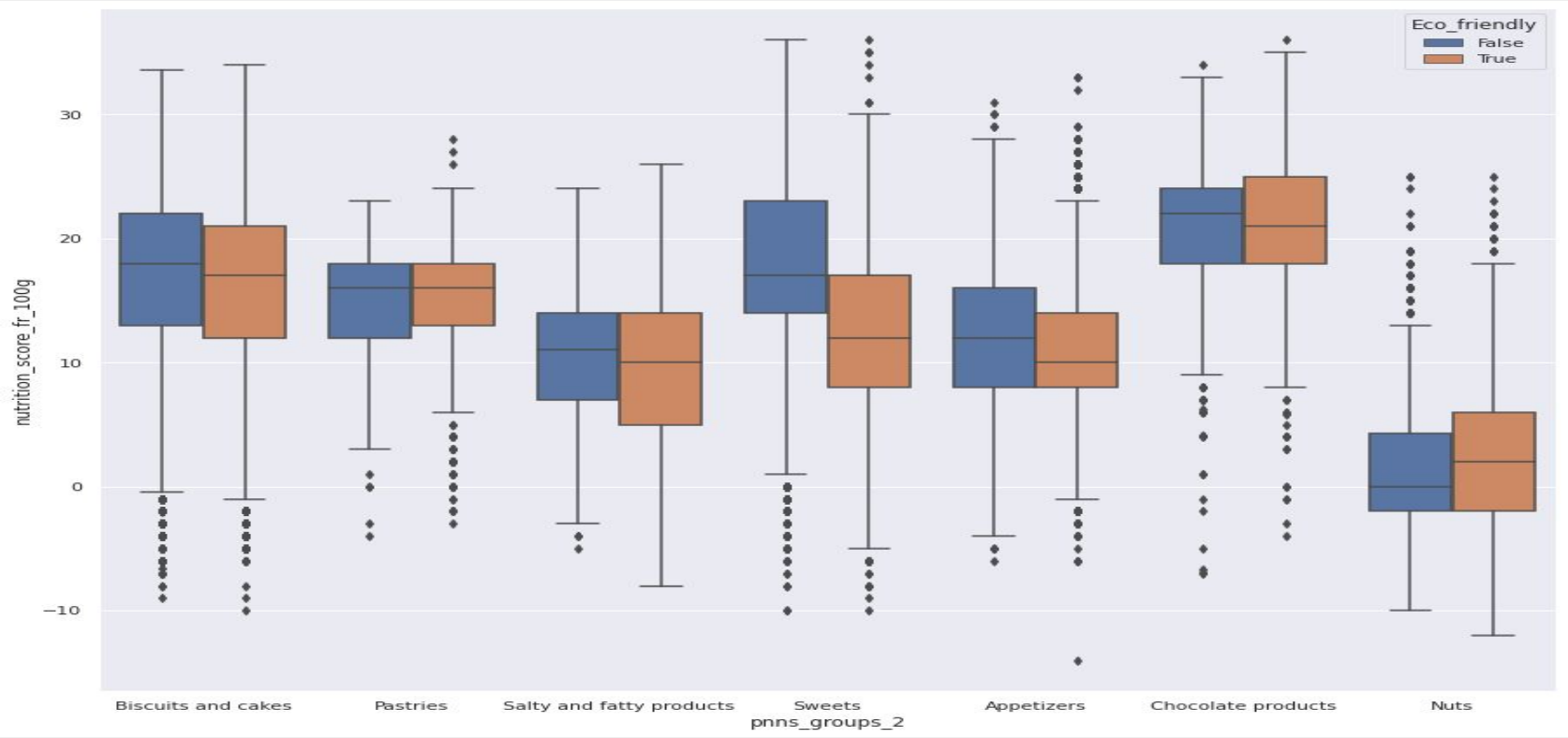
Analyse bivariée

➔ Grâce à l’analyse bivariée on pourra étudier les relations entre deux variables distinctes, quantitatives ou qualitatives.

Matrice de corrélation des variables quantitatives



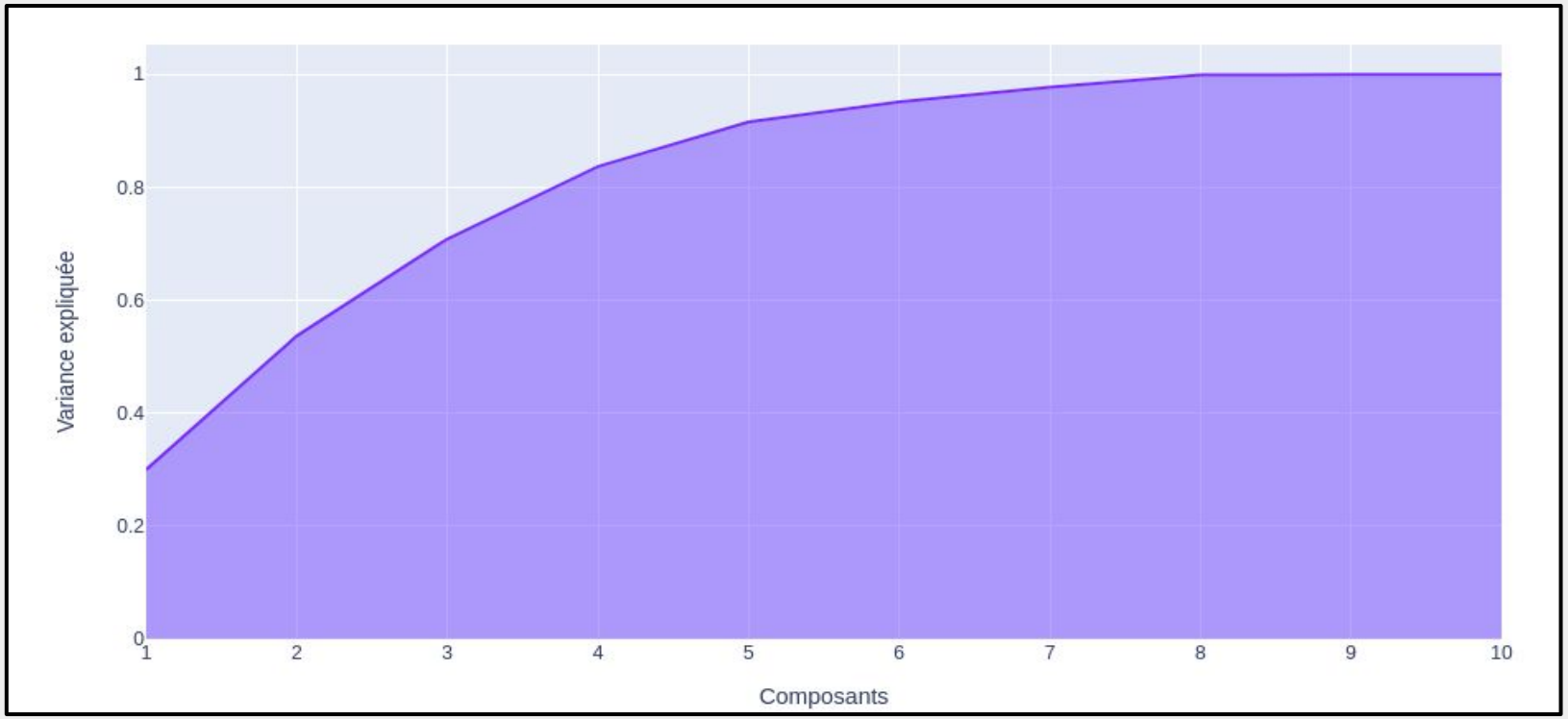
Distribution du nutriscore par produit et par éco-responsabilité



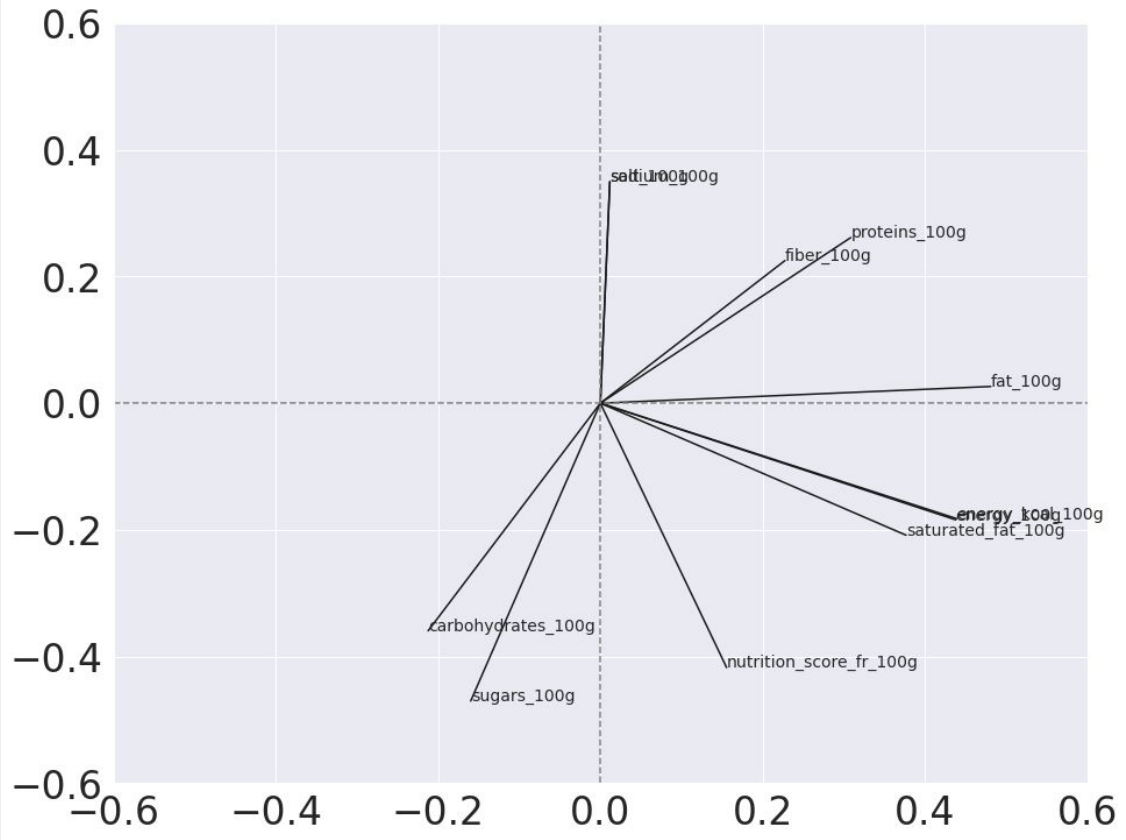
Analyse multivariée/PCA

➔ La PCA nous servira à transformer les variables corrélées en de nouvelles variables décorrélées les unes des autres. Cela permet de réduire le nombre de variables et de rendre ainsi l'information moins redondante.

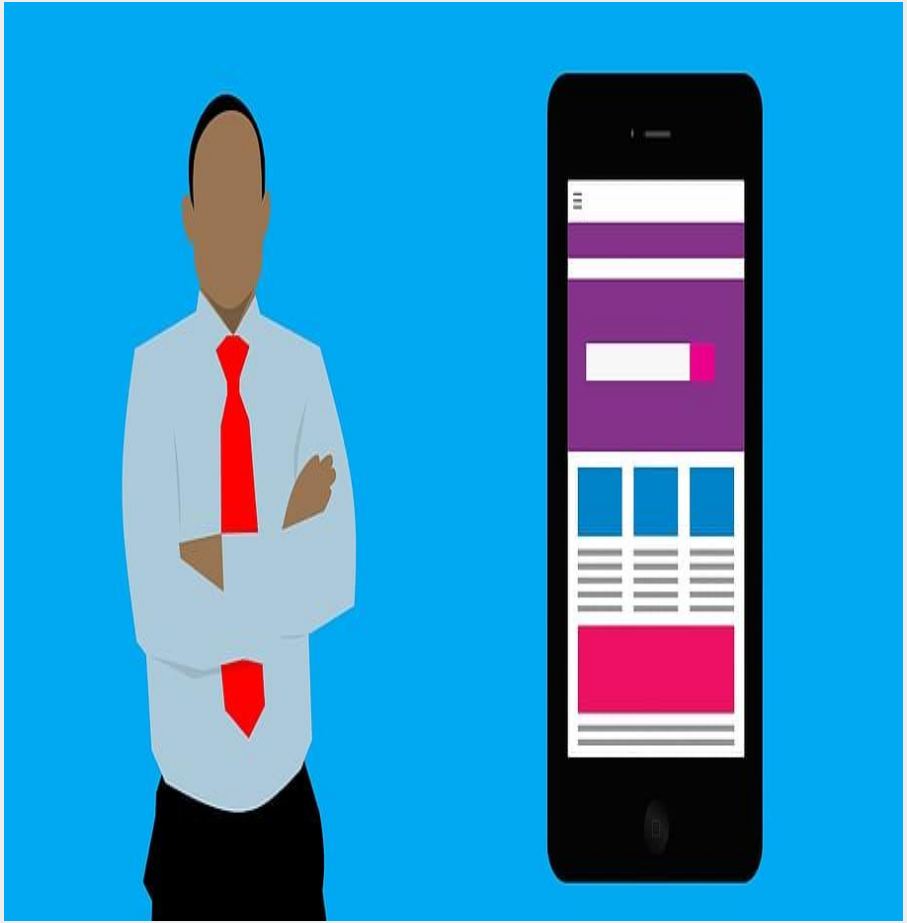
Variance expliquée et cumulée de la PCA



Contribution de chaque variable à PC1 et PC2



Conclusion
et axes d’améliorations



Conclusion

- Les analyses effectuées nous révèlent que les produits éco-friendly ont en moyenne des nutri-scores plus favorables que ceux non éco-friendly.
- Il est donc doublement recommandé de privilégier les produits éco-friendly, non seulement cela préserve l'environnement d'un coût écologique supplémentaire mais aussi cela offre la possibilité de consommer des produits plus sains.
- Vu le nombre de produits dont on dispose (plus de 80.000), le projet est tout à fait réalisable.
- Algorithme prêt au déploiement.

Axes d’amélioration

- Possibilité de faire participer l’utilisateur en lui permettant d’ajouter manuellement des produits à la base de données afin qu’elle soit plus conséquente.
- Ajout d’une catégorie “moyen de transport”, la prendre en compte pour calculer un score en tenant compte de la distance parcourue depuis le lieu de fabrication jusqu’au lieu de vente. Proposer les produits les plus éco responsables/sains.
- Créer des grades eco responsabilité (eco friendly --, eco friendly -, eco friendly, eco friendly +, eco friendly ++).


NOW $2.09
NOW $2.09
NOW $1.89
Corn Kernels
$2.29
89¢
Creamed Corn
89¢
NOW $1.19
HOME BRANDS LOWEST PRICES
Tomato Soup
89¢
Big Red Condensed Tomato Soup
$4.99
Thai Ready Meals
$2.99
AUSTRALIAN Sliced Beetroot
85¢
CONDENSED SOUP
Cherries
spaghetti
1222
BAKED BEANS
7975