



Concevez une application au service de la santé publique

Projet 3 du cursus Data Scientist

Sommaire

1. Présentation du projet
2. Présentation de l'application
3. Opérations de nettoyage effectuées
4. Opérations de remplissage effectuées
5. Focus sur le traitement des doublons
6. Traitement des valeurs aberrantes
7. Focus sur les principales étapes du Machine Learning
8. Descriptions et Analyses univariées des différentes variables importantes
9. Analyse multivariée et les résultats statistiques associés
10. Analyse de la Variance (ANOVA)
11. Analyse en Composante Principale (ACP)
12. Conclusion

Présentation du projet



Objectif à atteindre

- Trouver des idées innovantes d'applications en lien avec l'alimentation
- Traiter le jeu de données mis à disposition afin de repérer des variables pertinentes
- Produire des visualisations afin de mieux comprendre les données



Moyens mis à disposition

- Le jeu de données **Open Food Fact**

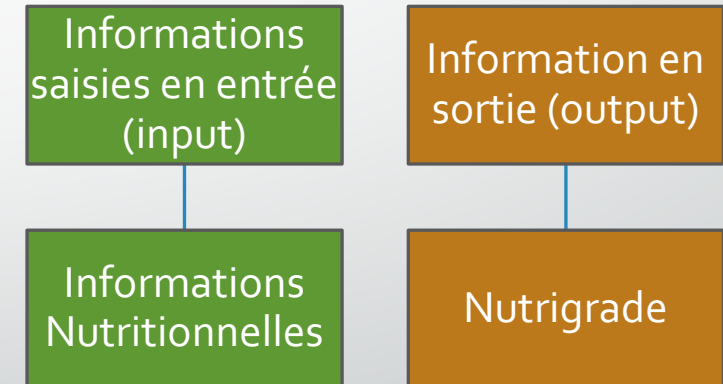
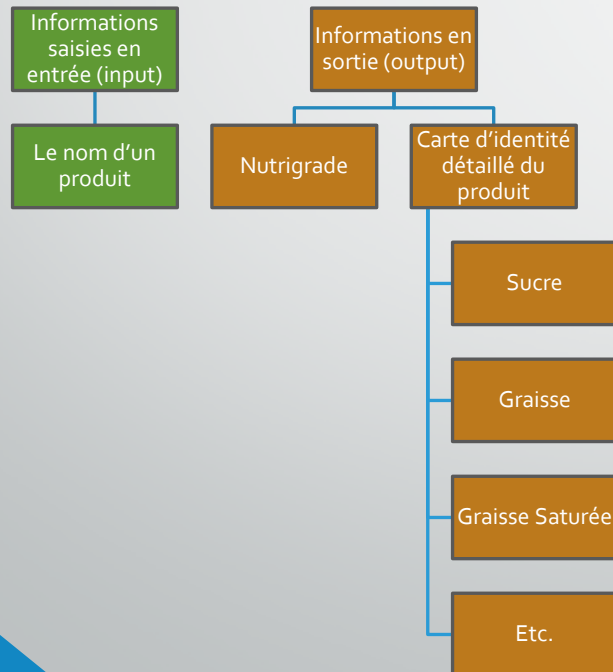
Présentation de l'idée d'application



Présentation de l'idée d'application



2 approches possibles



Opérations de nettoyage effectuées

Nettoyage par Pays

- Produits vendus en France uniquement

Traitement des colonnes contenant des **dates**

- Convertissage des colonnes au format **date**
- Parsage des dates au format **yyyymmdd**

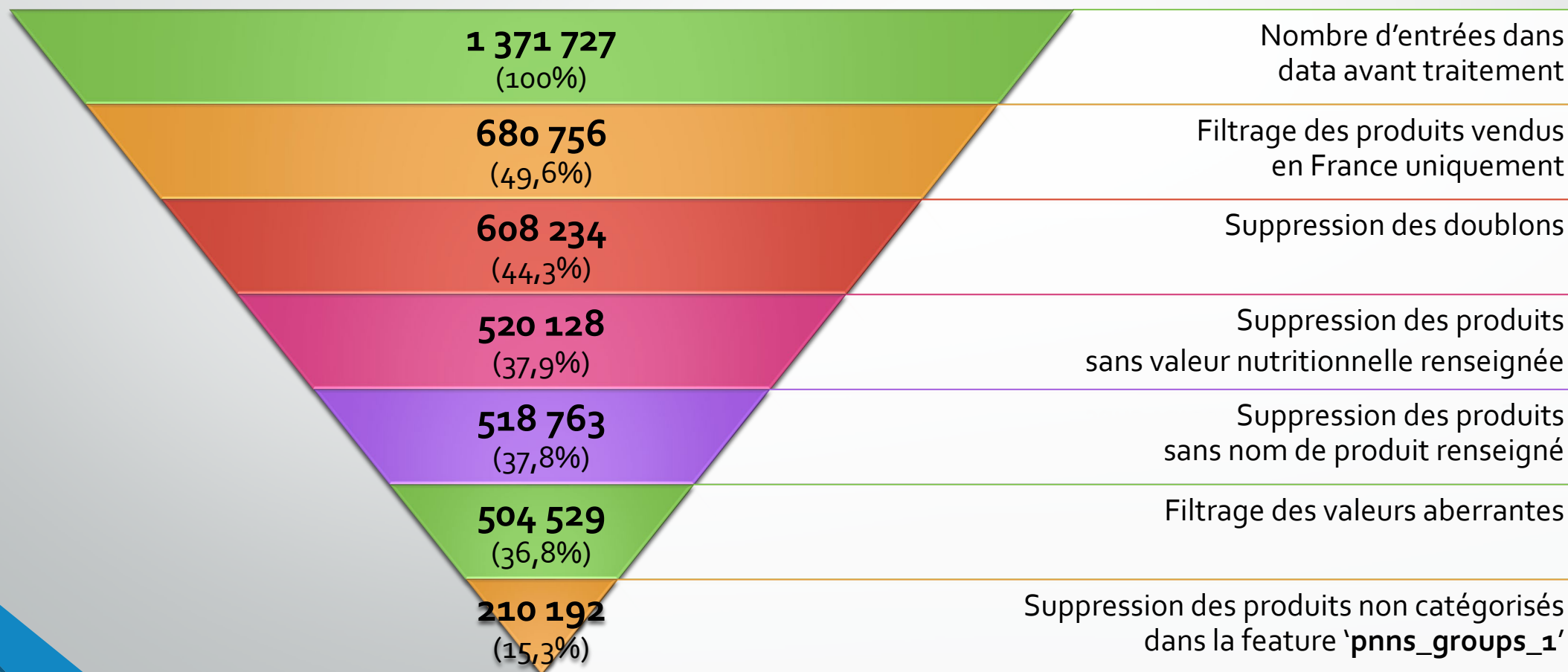
Nettoyage par features

- Suppression des colonnes inutiles
- Suppression des colonnes vides et peu renseignées

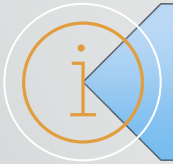
Nettoyage par produits

- Suppression des produits redondants
- Suppression des produits trop peu renseignés
- Traitement des données aberrantes
- Traitement des données manquantes

Opérations de nettoyage effectuées



Suppression des doublons



Pour l'ensemble des étapes suivantes, lors d'une suppression de doublon, je conserve systématiquement les premières occurrences

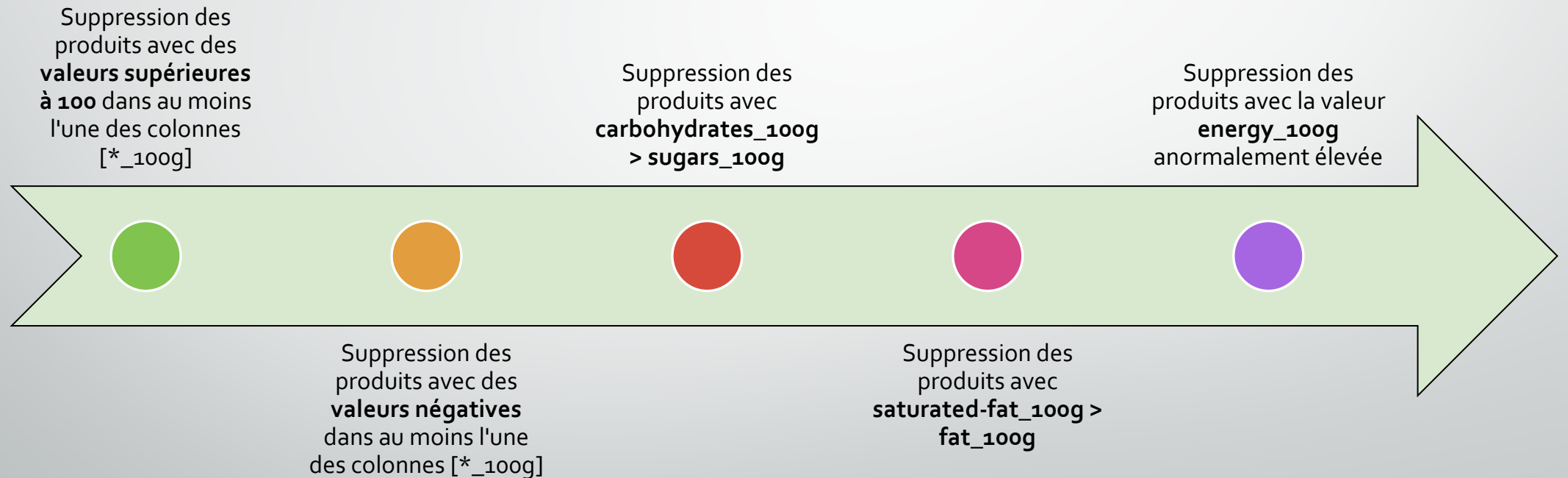
Recherche et suppression
des doublons en
comparant l'ensemble
des colonnes disponibles

Recherche et suppression
des doublons
en comparant toutes
les colonnes excepté :
code, **created_t**
et **last_modified_t**

Recherche des valeurs
en doublon en comparant
leurs noms (pas de
suppression)

Recherche et suppression
des valeurs en doublon en
comparant leurs noms
et leurs valeurs
nutritionnelles [***_100g**]

Filtrage des valeurs aberrantes



Opérations de remplissage effectuées

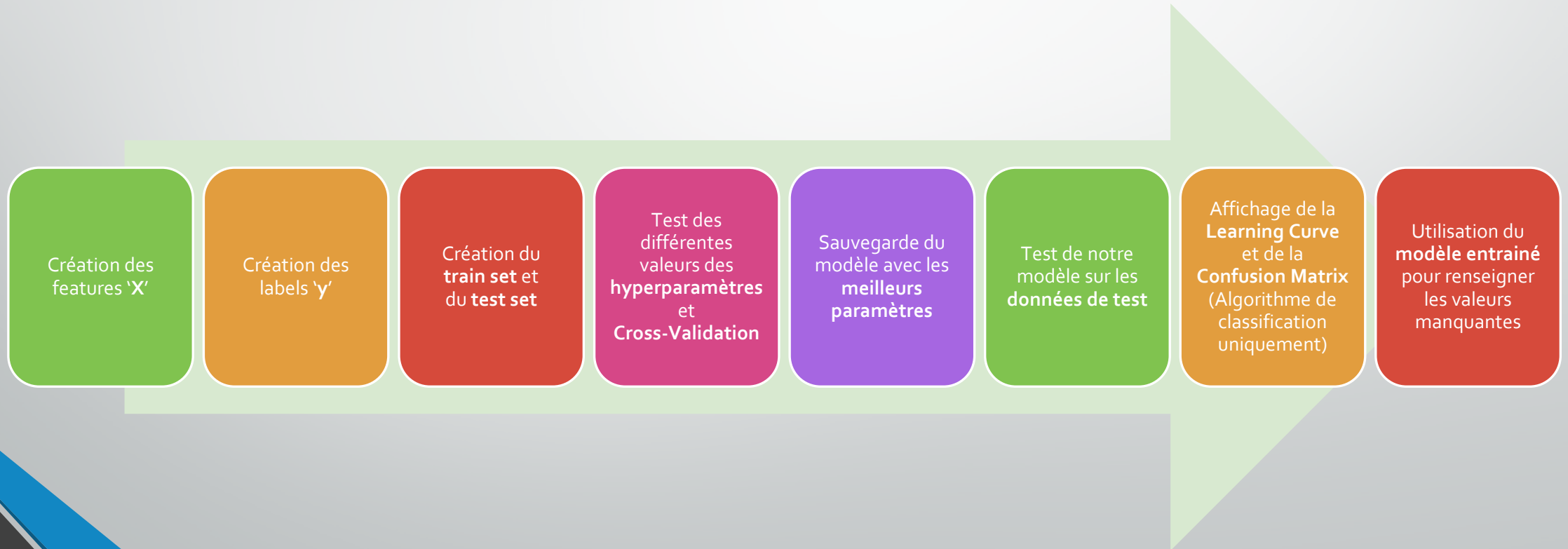
Remplissage par la valeur moyenne en fonction de la catégorie du produit

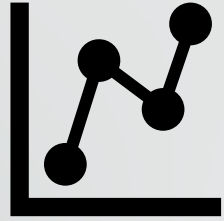
- Utilisée pour remplir les features informant sur la valeur nutritionnelle pour 100g des produits

Appel aux techniques de Machine Learning

- Algorithme de régression : **KNN Regressor**
 - ❖ pour renseigner la colonne '**Nutriscore**'
- Algorithme de classification : **KNN Classifieur**
 - ❖ pour renseigner la colonne '**Nutrigrade**'

Etapes principales de remplissage à l'aide du Machine Learning

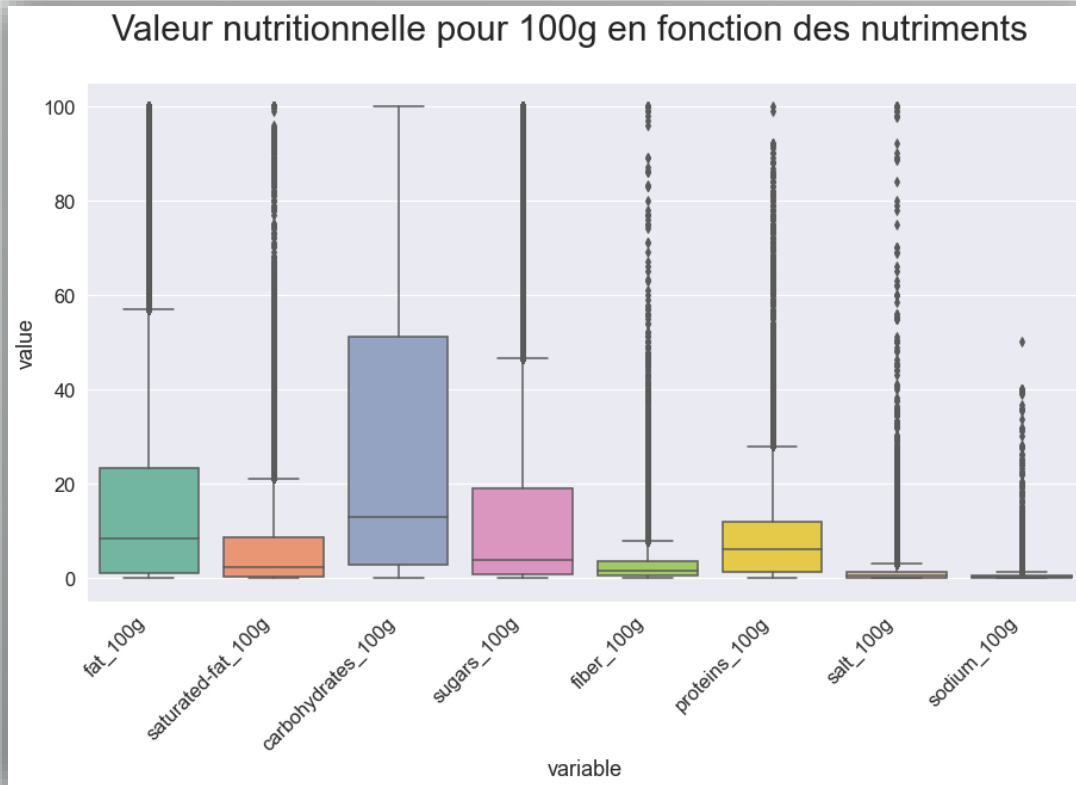




Description et Analyse univariée
des différentes variables importantes
avec les visualisations associées.



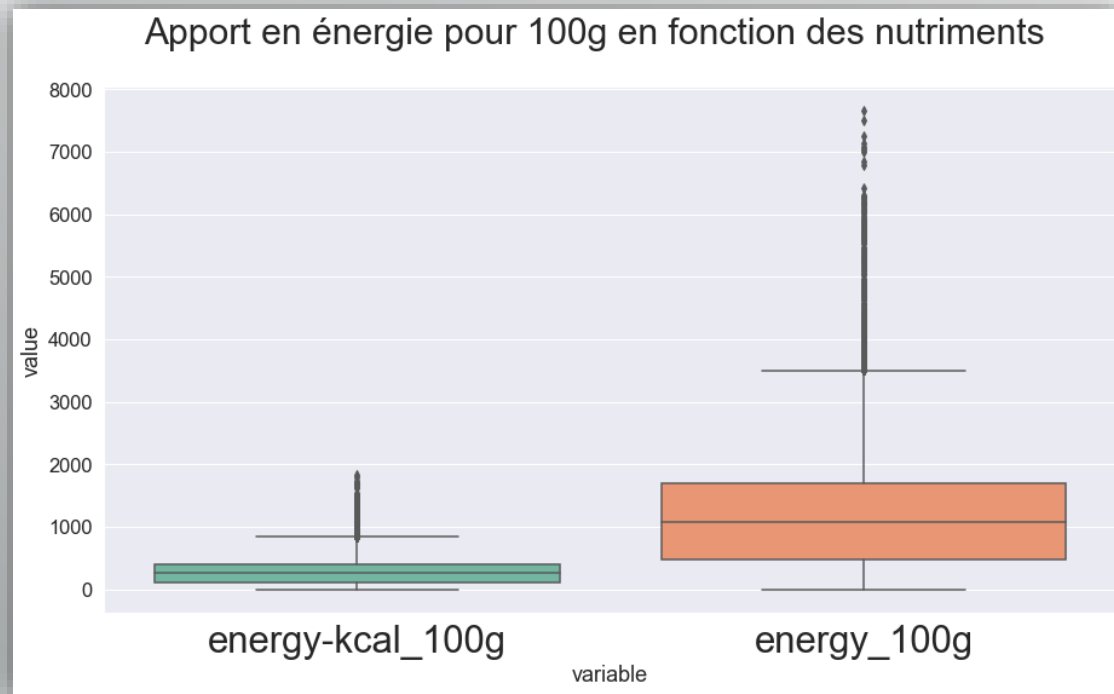
Analyse univariée des nutriments



De nombreux outliers proviennent d'erreurs de saisie de la part des utilisateurs d'Open Food Fact

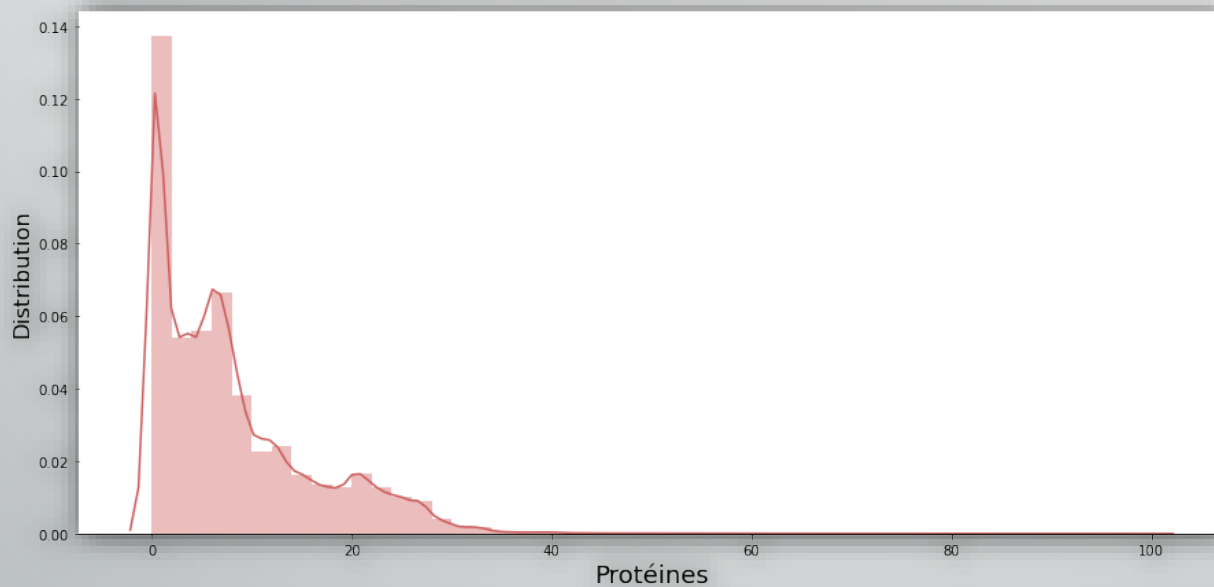
Les autres outliers sont légitimes et proviennent de divers aliments particulièrement riches pour un ou plusieurs nutriments

Analyse univariée des apports énergétiques



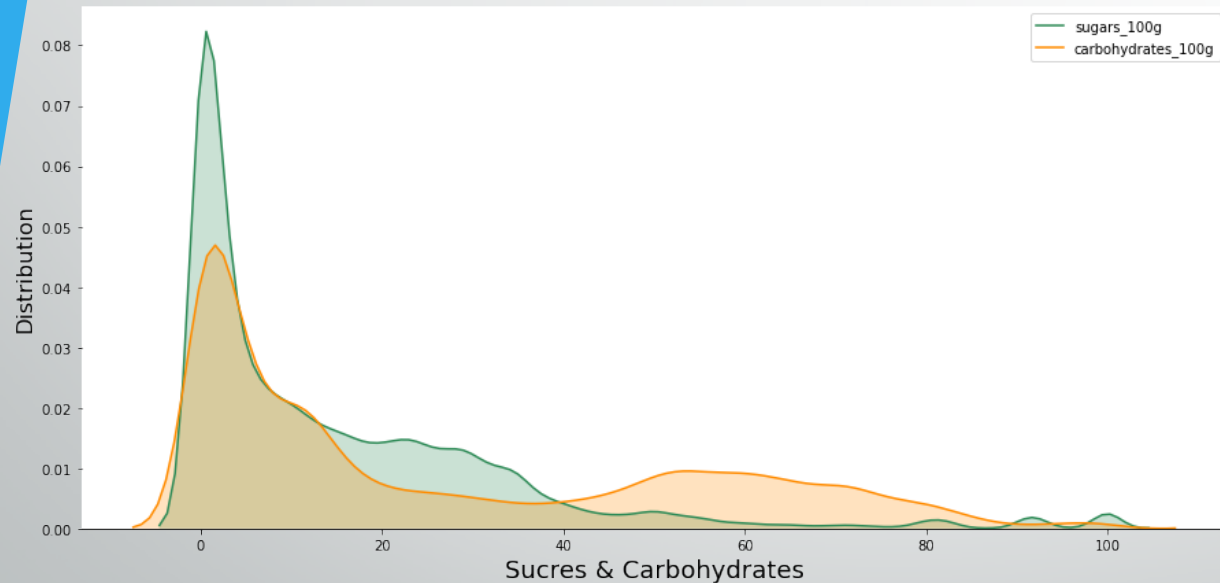
Des valeurs anormalement élevées dans l'apport énergétique de certains aliments doivent être contrôlés avec attention pour juger de leur pertinence.

Distribution de la valeur nutritionnelle des protéines sur 100g de produit




La distribution des protéines est multimodale non centrée

Distribution de la valeur nutritionnelle des Sucres et Carbohydrates sur 100g de produit



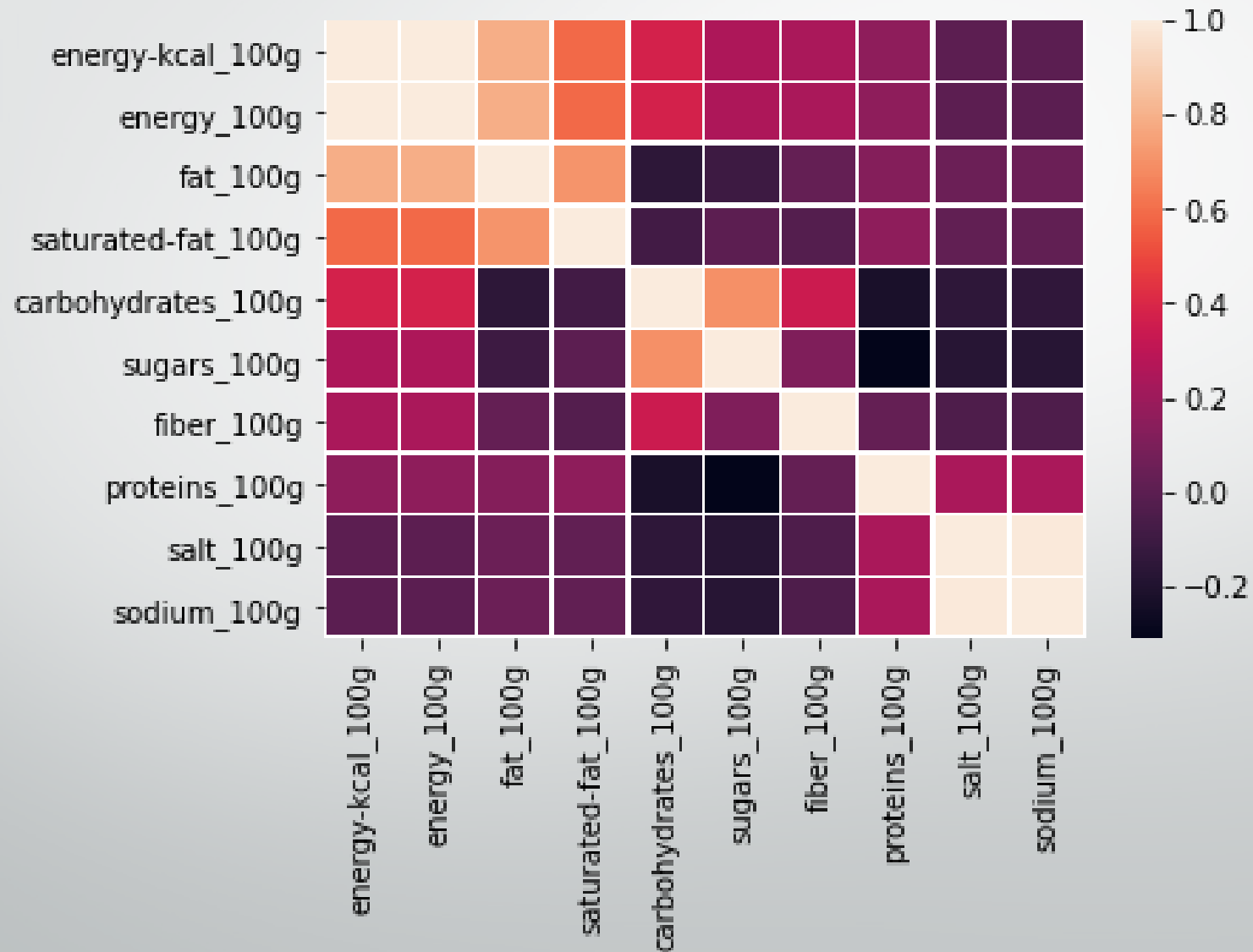
La distribution du sucre
est unimodale non centrée

La distribution du carbohydres
est bimodale non centrée



Analyses multivariées et les résultats statistiques
associés, en lien avec l'idée d'application

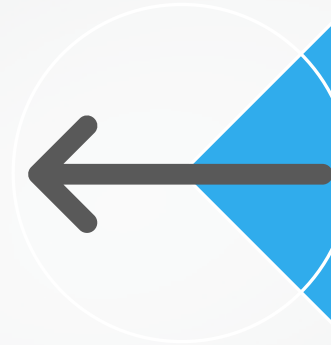
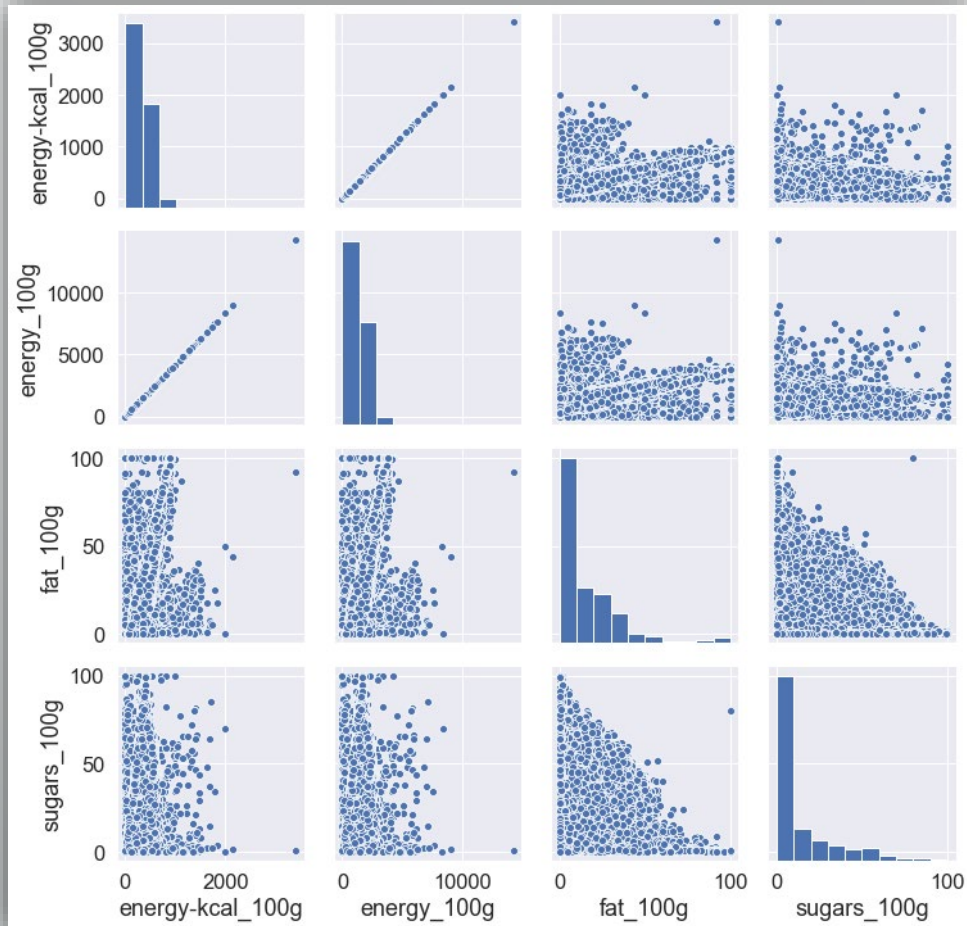
HeatMap de Corrélation



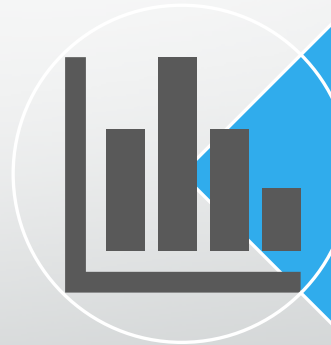


Graphique PairPlot

Graphique PairPlot



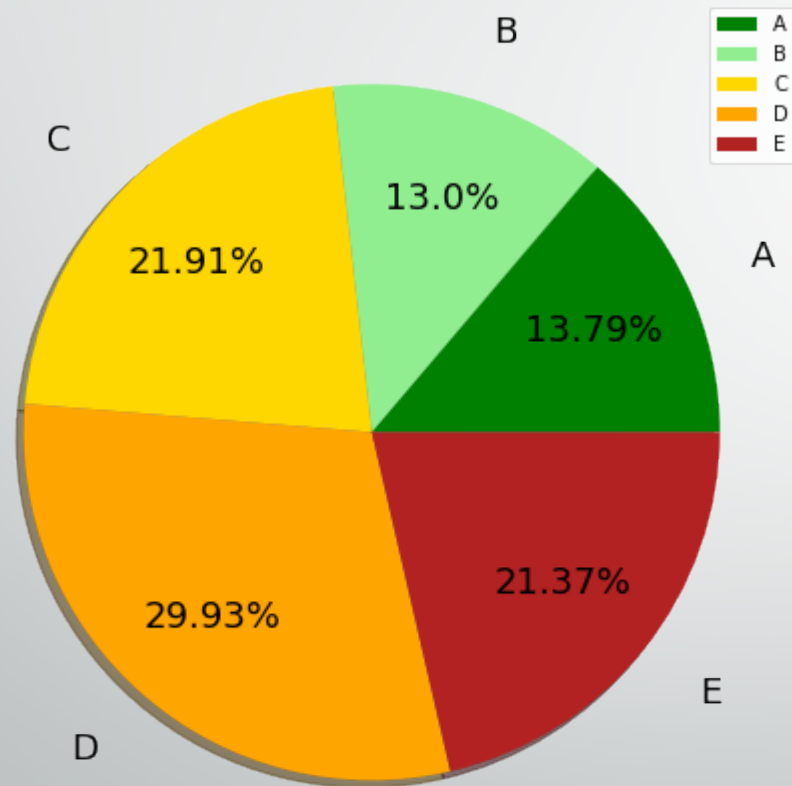
Ci-contre, une version réduite
du PairPlot complet



Le PairPlot met en évidence des
dépendances ou non entre
certains nutriments

- Par exemple le gras et le sucre semblent s'opposer sur leurs valeurs hautes.
- Un aliment très sucré est peu gras et un aliment très gras est peu sucré.
- La réciproque est fausse, des aliments peuvent être peu gras et peu sucré.

Répartition des produits en fonction du nutrigrade



La catégorie D est la mieux représentée avec près d'un tiers des produits



Les catégories C et E sont également majoritaires



Globalement le jeu de données recense majoritairement des produits peu recommandés nutritionnellement

ANOVA (ANalysis Of VAriance)

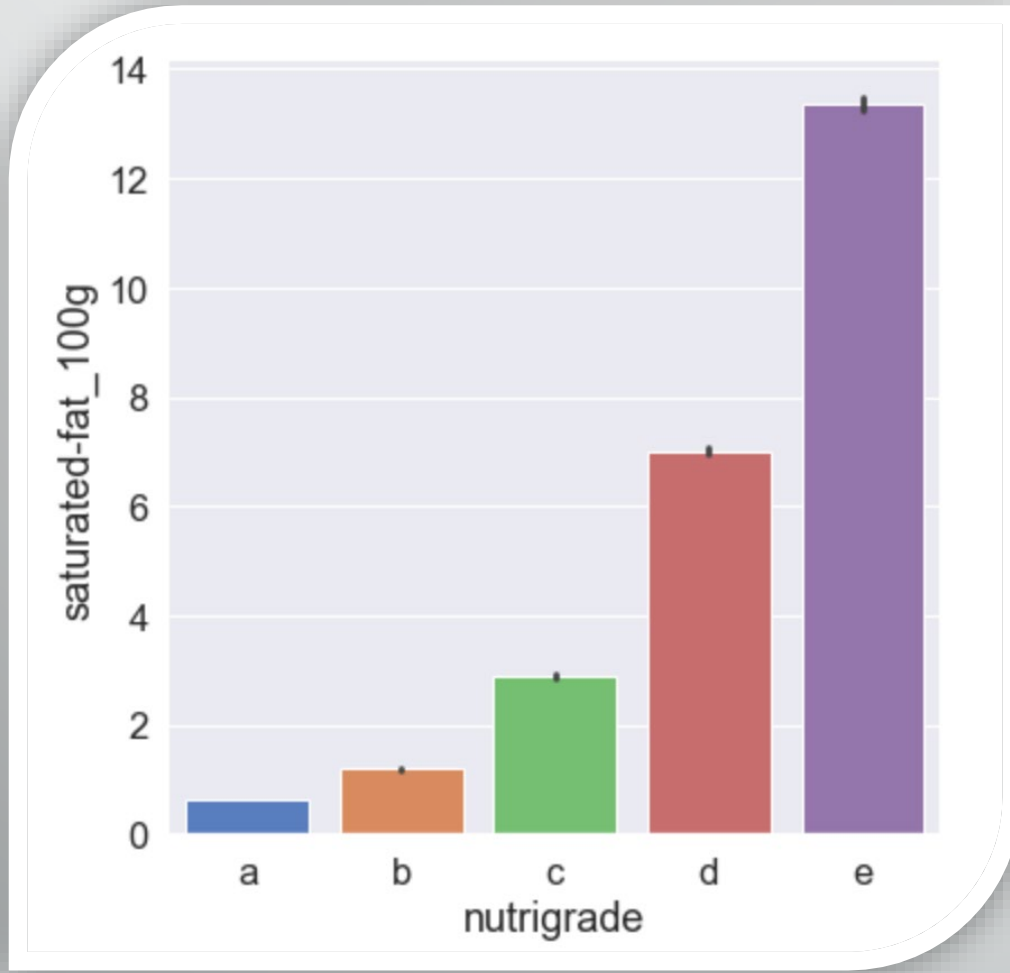
Analyse de la Variance



Peut-on prévoir la valeur
'**nutrigrade**' d'un produit à partir
d'une de ses valeurs nutritionnelles?

Si oui, quelle est la valeur
nutritionnelle qui explique le mieux
la valeur du '**nutrigrade**'?

ANOVA (ANalysis Of VAriance)



ANOVA (ANalysis Of VAriance)

Nutriments	Rapport de corrélation (eta squared)
saturated-fat_100g	0,277
energy_100g	0,238
energy-kcal_100g	0,238
fat_100g	0,179
sugars_100g	0,132
carbohydrates_100g	0,040
salt_100g	0,037
sodium_100g	0,037
fiber_100g	0,035
proteins_100g	0,026

Les valeurs
nutritionnelles
qui expliquent
le mieux la valeur
du 'nutrigrade'

- saturated-fat_100g
- energy_100g' et 'energy_100g
- fat_100g
- sugars_100g

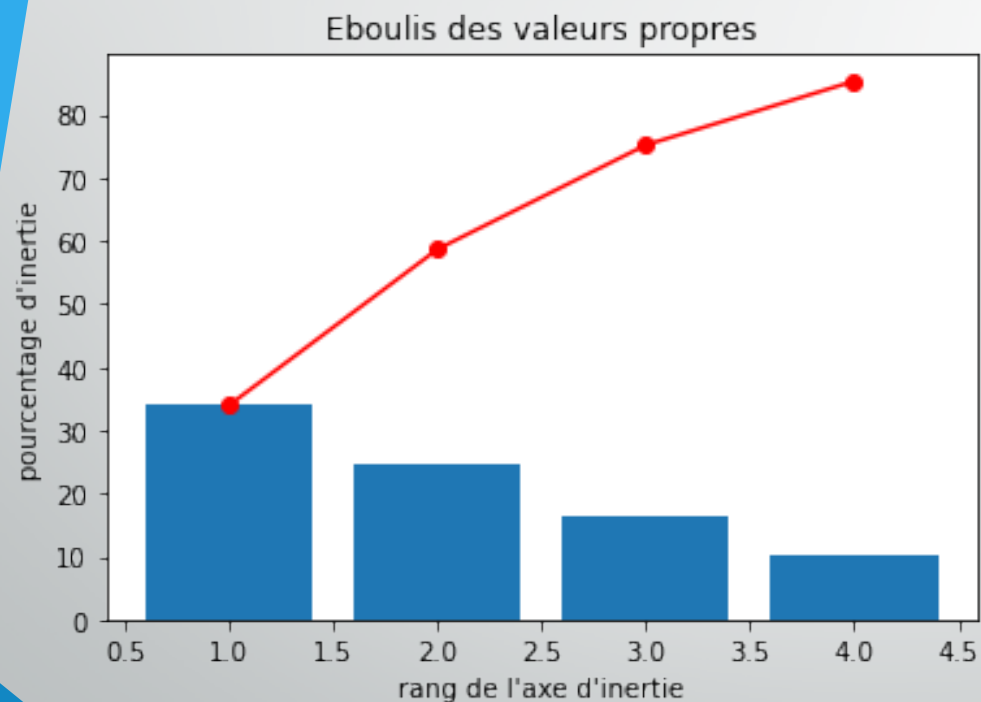
Analyse en Composante Principale



Existe t'il
des groupes de variables
très corrélées entre elles
qui peuvent être regroupées
en de nouvelles
variables synthétiques ?

Analyse en Composante Principale

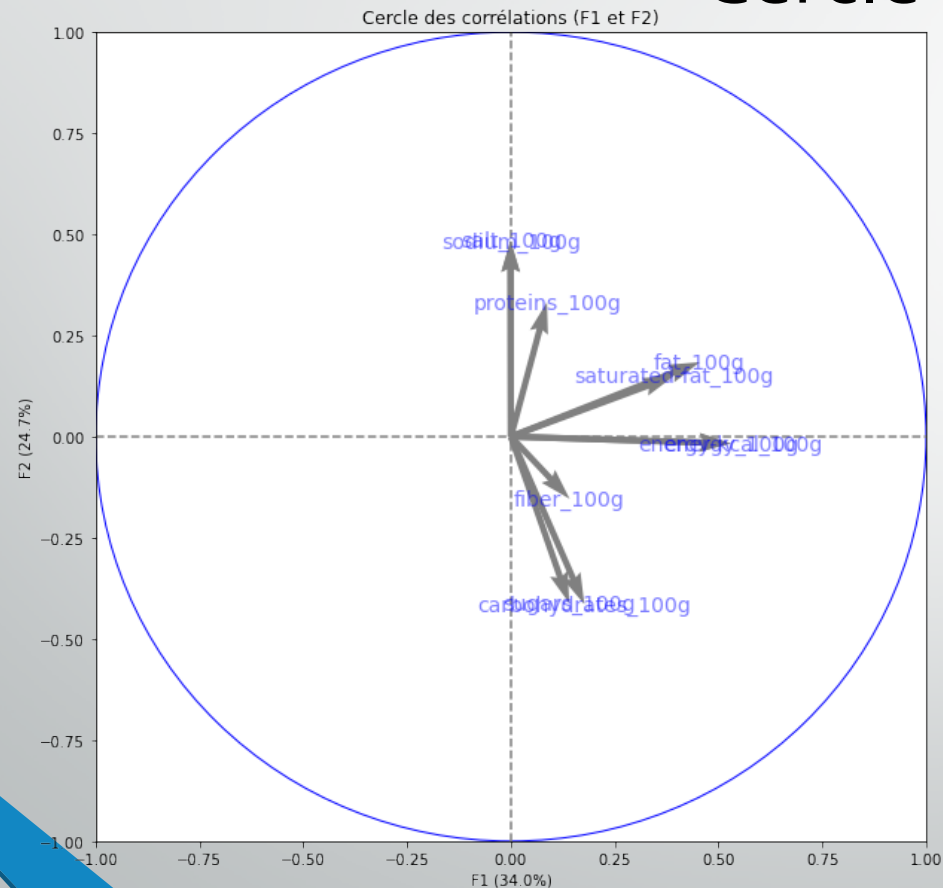
Eboulis des valeurs propres



Le graphique représentant l'éboulis des valeurs propres nous indique que 90% de l'inertie totale est associée aux 4 premiers axes d'inertie

Analyse en Composante Principale

Cercle des corrélations



Modèle plus simple

Modèle plus rapide

Modèle plus performant

Analyse en Composante Principale

Détail des 4 nouvelles variables F1 à F4

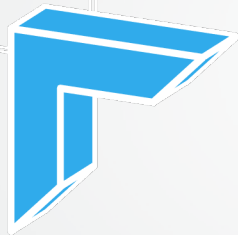
	energy-kcal_100g	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g
F1	0.52916	0.52916	0.45499	0.39302	0.17684	0.14036	0.14141	0.08681	-0.00080	-0.00096
F2	-0.01679	-0.01679	0.18408	0.14981	-0.41450	-0.41146	-0.15399	0.33190	0.48422	0.48338
F3	0.03930	0.03930	-0.24608	-0.23759	0.45487	0.35595	0.26589	-0.03898	0.48644	0.48730
F4	0.01792	0.01792	-0.11687	-0.18799	0.01951	-0.31456	0.71757	0.54216	-0.14446	-0.14596



Les variables F1 à F4 sont des combinaisons linéaires des autres variables [*_100g]

Synthèse des différentes conclusions sur la faisabilité du projet

- La 2^{ème} approche retenue :
 - Input: Informations Nutritionnelles
 - Output: Nutrigrade
- L'application est réalisable car :
 - La base de données est suffisamment nettoyée et optimisée
 - Elle n'a pas besoin d'avoir le produit recherché dans sa base de données pour connaître le nutrigrade
- Limitations connues :
 - La performance pour prédire les nutrigrades A et B est limitée
 - Pas de problème si les utilisateurs recherchent des produits dont le nutrigrade va de C à E
 - N'alimente pas la base de donnée d'Open Food Fact



Merci

