



SOUTENANCE

ACCESSIBILITE DES DONNEES

Cédric Dietzi

Sommaire

- Appel à projet
- Nettoyage du jeu de données
- Prototype DataXPlor
- Démo
- Annexe: guide d'utilisation



DataXPlor

Données exploitables

Données nettoyées

Données accessibles

Navigation facile dans les données

Visualisation pertinente automatique

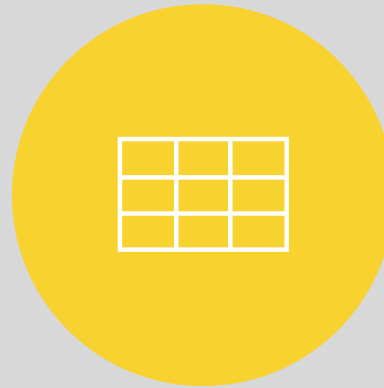
Données informatives

Calculs statistiques de base embarqués

Le jeu de données



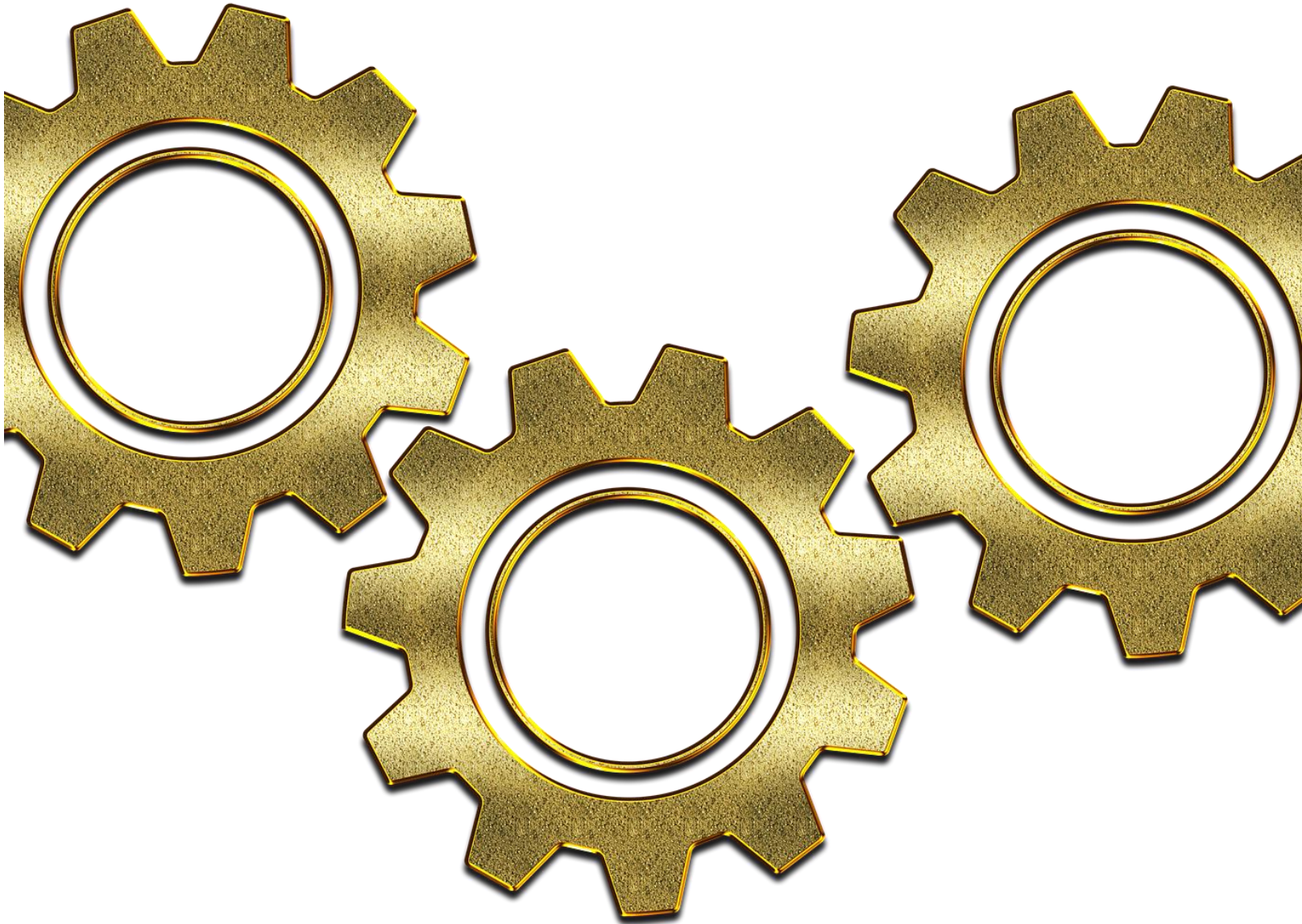
320 000 PRODUITS



162 VARIABLES



MARQUE, MAGASIN,
CATEGORIE,
COMPOSITION, ...



Nettoyage du jeu de données

Nettoyage des variables

Lorsque qu'une information est disponible sous plusieurs variables, on supprime les **variables non-normées**. Si disponible, on conserve la variable en français.

Ex: coutries, countries_tags, countries_fr => on garde coutries_fr

Suppression de **22 variables**

On rajoute **4 variables informatives** contenant la somme: 1) des ingrédients au total, 2) des acides gras, 3) des sucres, 4) des protéines

Ajout de **4 variables**

Suppression des **variables à valeur unique**: elles ne portent pas d'information

Suppression de **26 variables**

Suppression des **variables peu remplies** devant les autres: elles portent peu d'information et sont comparativement de mauvaise qualité. Seuil: 10%

Suppression de **65 variables**: 6 qualitatives et 59 quantitatives (les autres quantitatives sont remplies à plus de 50%)

Imputation des valeurs manquantes: Qualitatives par le mot clé 'missing' et Quantitatives par la médiane.

On est passé de 162 à 53 variables

Nettoyage des produits

Suppression des produits avec des **valeurs métiers aberrantes**. Cohérence des sommes d'ingrédients, des valeurs d'énergie, d'empreinte carbone.

Suppression de **90 000 produits** dont 86 000 pour lesquels la somme des compositions par 100g est ≤ 0 ou > 100

Suppression des **produits peu renseignés** avec moins de 23 éléments sur 53.

Suppression de **1 900 produits**

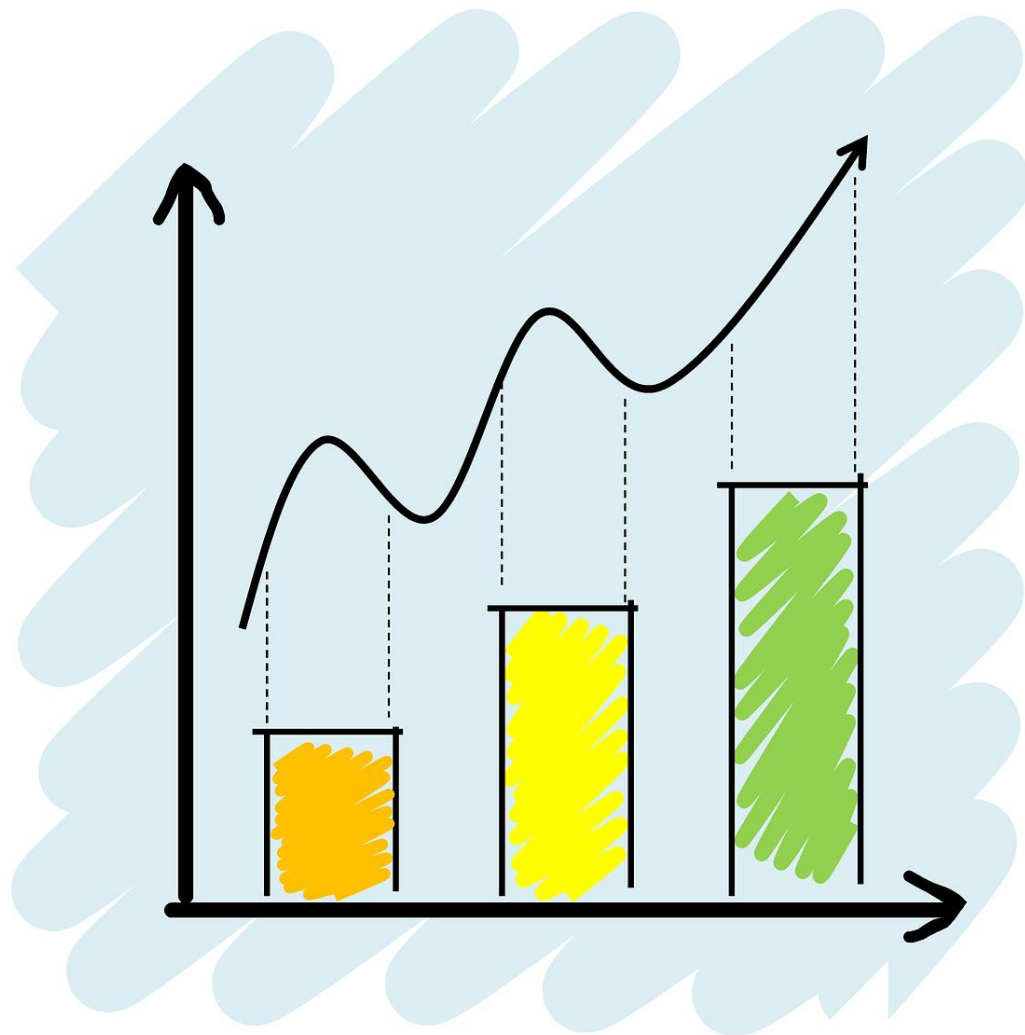
Traitement des **doublons**: 53 000 produits sont des doublons sur la partie quantitative. Parmi les variables qualitative, 'product_name' est la plus discriminante et fait tomber ces doublons de 53 000 à 13 000.

Suppression de **13 000 produits**

Valeurs statistiques aberrantes : si la catégorie pnns_groups_2 est remplie, suppression des outliers au-delà de 1,5 fois la valeur interquartile, 5 fois sinon.

Suppression de **36 000 produits**

On est passé de 320 000 à 179 000 produits



DataXPlor

Prototype
d'exploration
de données

Dashboard

DataXplor

PCA: Eboulis

PC1-PC2


PC3-PC4

PC5-PC6

PC7-PC8

DataXplor

Sélectionner pour filtrer:

 Vider la cellule !

stores

missing

Carrefour

Auchan

Leclerc

Cora

Intermarché

Lidl

Franprix

Super U

Casino

Aldi

Monoprix

Dia

Migros

Leader Price

Carrefour Market

LIDL

Netto

Picard

__Autres__

 Filtrer

 Réinitialiser

Taille du jeu de données

(25089, 53)

Sélectionner pour visualiser:

code

creator

product_name

generic_name

quantity

packaging_tags

brands_tags

categories_fr

manufacturing_places_tags

labels_fr

emb_codes_tags

purchase_places

stores

countries_fr

ingredients_text

allergens

serving_size

additives_n

additives

ingredients_from_palm_oil_n

ingredients_that_may_be_from_palm_oil_n

nutrition_grade_fr

pnns_groups_1

pnns_groups_2

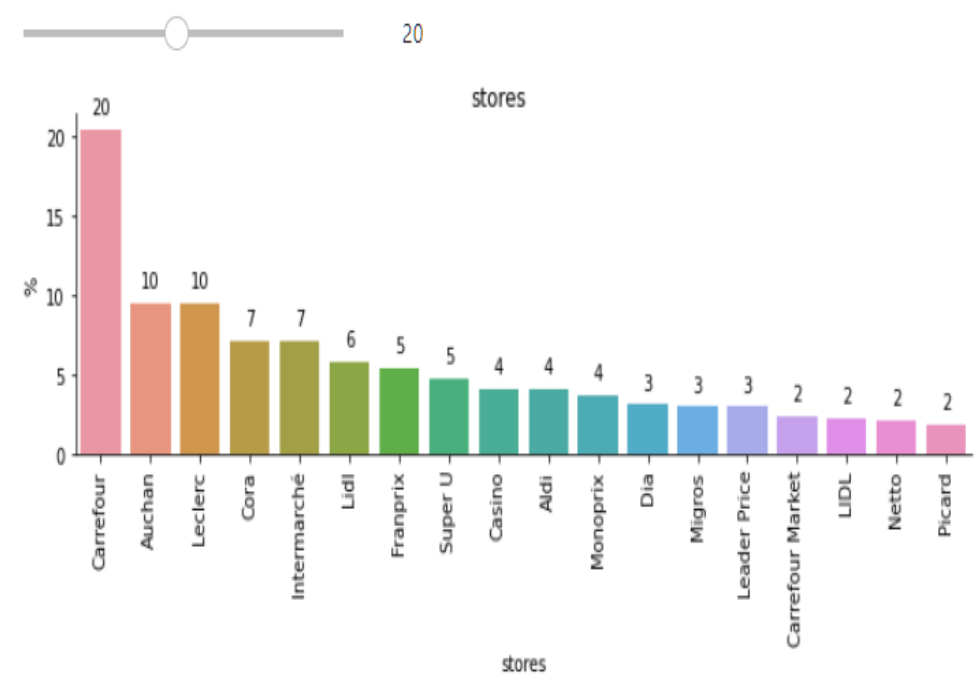
states_fr

main_category_fr


energy_100g

fat 100g

Nombre de bars:



1. Zone de filtration

Sélectionner pour filtrer:  Vider la cellule !

stores

missing

Carrefour

Auchan

Leclerc

Cora

Intermarché

Lidl

Franprix

Super U

Casino

Aldi

Monoprix

Dia

Migros

Leader Price

Carrefour Market

LIDL

Netto

Picard

__Autres__

Filtrer

Réinitialiser

Taille du jeu de données

(25089, 53)

2. Zone de sélection

Sélectionner pour visualiser:

code

creator

product_name

generic_name

quantity

packaging_tags

brands_tags

categories_fr

manufacturing_places_tags

labels_fr

emb_codes_tags

purchase_places

stores

countries_fr

ingredients_text

allergens

serving_size

additives_n

additives

ingredients_from_palm_oil_n

ingredients_that_may_be_from_palm_oil_n

nutrition_grade_fr

pnns_groups_1

pnns_groups_2

states_fr

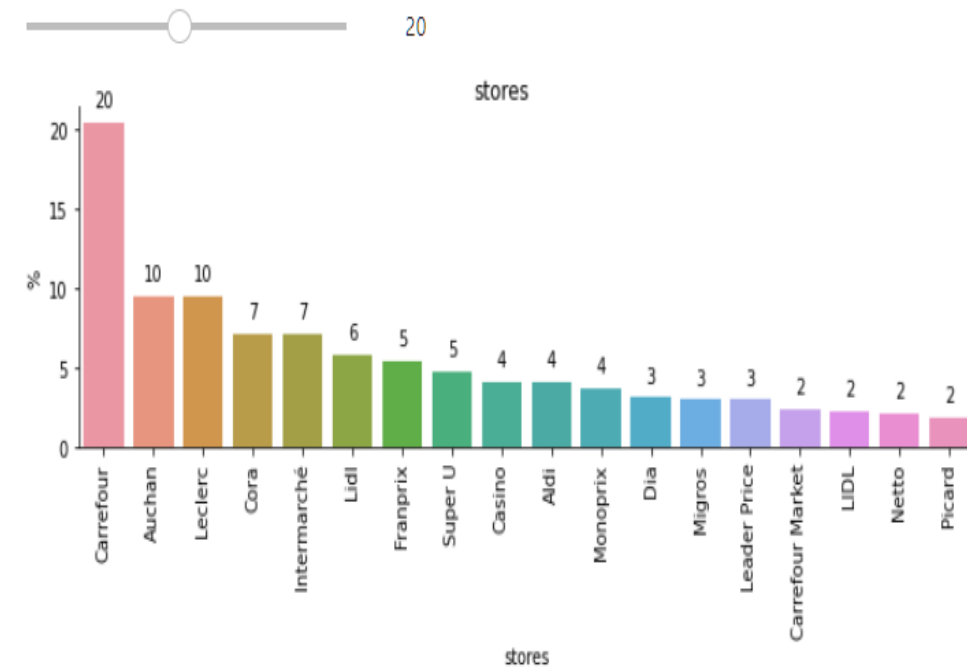
main_category_fr

energy_100g

fat_100g

3. Zone de visualisation et d'information

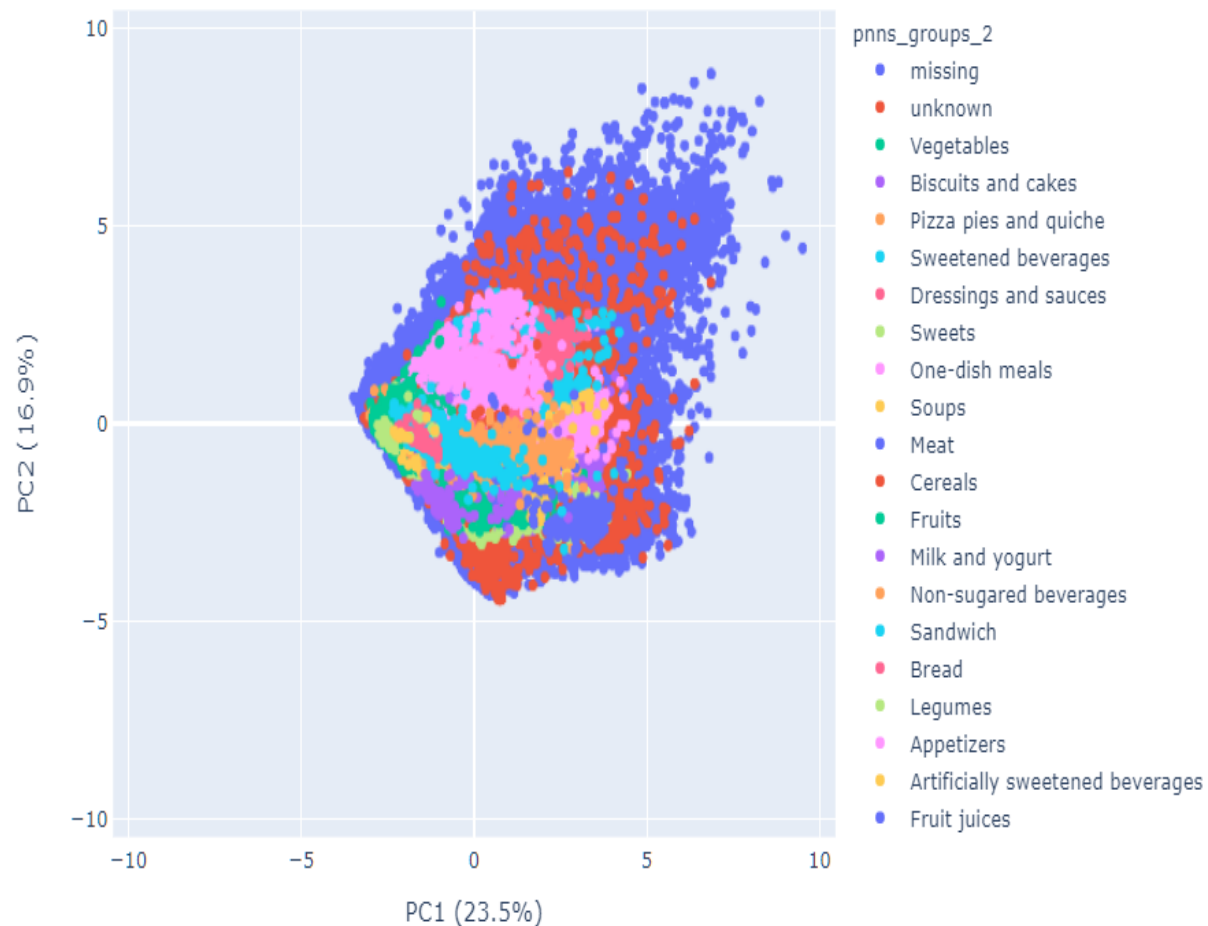
Nombre de bars:



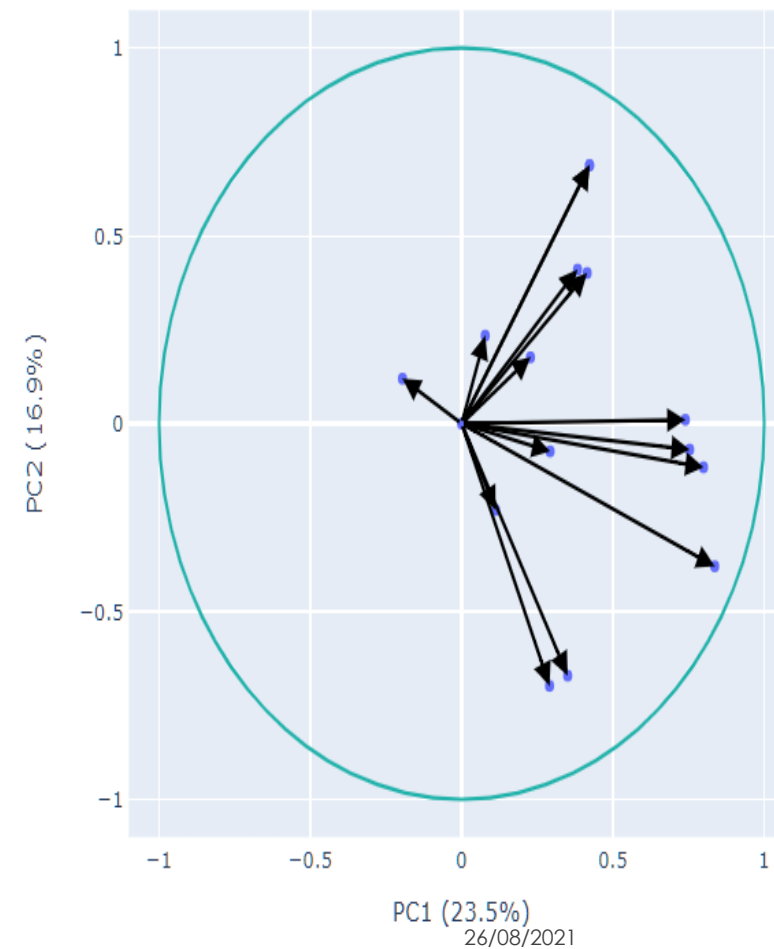
⚡ (Re)Lancer l'ACP



Plan factoriel PC1 / PC2

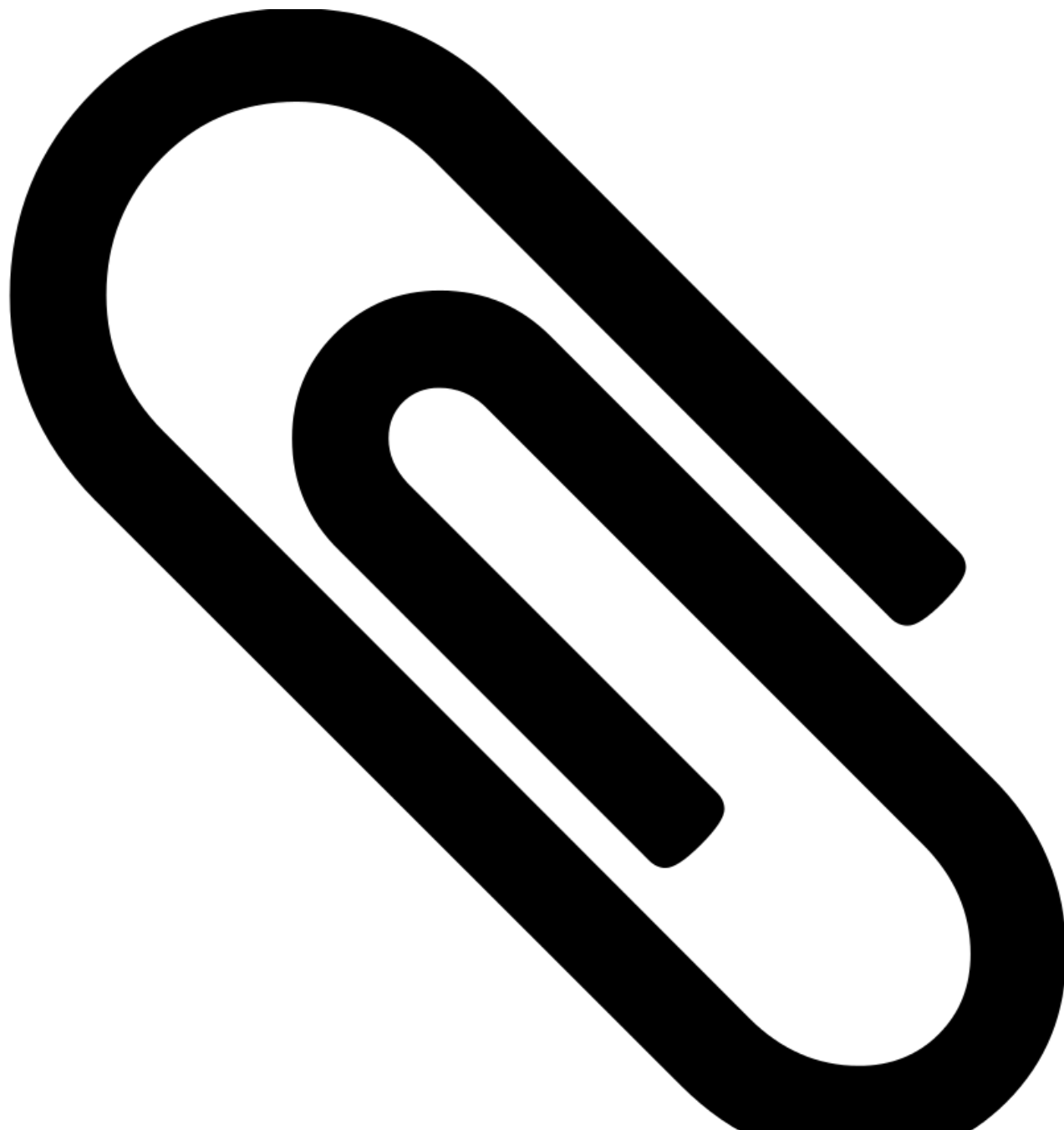


Cercle des corrélations - PC1 / PC2





Demo



Annexe Guide d'utilisation

Dashboard

DataXplor

PCA: Eboulis

PC1-PC2


PC3-PC4

PC5-PC6

PC7-PC8


DataXplor

Sélectionner pour filtrer:

 Vider la cellule !

- stores
- missing
- Carrefour
- Auchan
- Leclerc
- Cora
- Intermarché
- Lidl
- Franprix
- Super U
- Casino
- Aldi
- Monoprix
- Dia
- Migros
- Leader Price
- Carrefour Market
- LIDL
- Netto
- Picard
- __Autres__

 Filtrer

 Réinitialiser

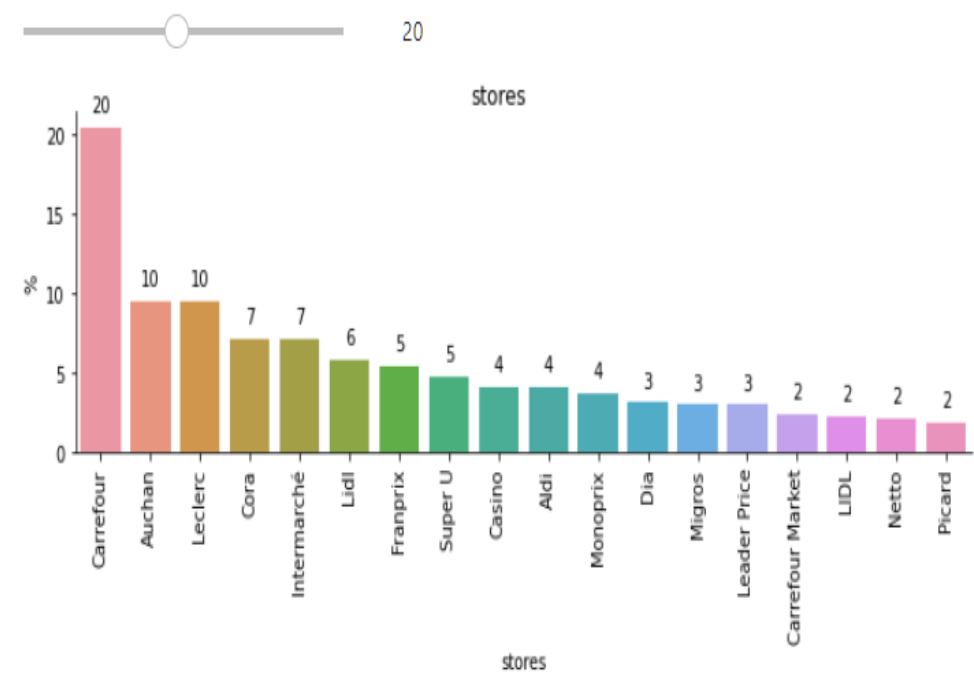
Taille du jeu de données

(25089, 53)


Sélectionner pour visualiser:

- code
- creator
- product_name
- generic_name
- quantity
- packaging_tags
- brands_tags
- categories_fr
- manufacturing_places_tags
- labels_fr
- emb_codes_tags
- purchase_places
- stores
- countries_fr
- ingredients_text
- allergens
- serving_size
- additives_n
- additives
- ingredients_from_palm_oil_n
- ingredients_that_may_be_from_palm_oil_n
- nutrition_grade_fr
- pnns_groups_1
- pnns_groups_2
- states_fr
- main_category_fr
- energy_100g
- fat 100g

Nombre de bars:



1. Zone de filtration

Sélectionner pour filtrer:  Vider la cellule !

stores

missing

Carrefour

Auchan

Leclerc

Cora

Intermarché

Lidl

Franprix

Super U

Casino

Aldi

Monoprix

Dia

Migros

Leader Price

Carrefour Market

LIDL

Netto

Picard

__Autres__

▼ Filtrer

🔄 Réinitialiser

Taille du jeu de données

(25089, 53)

Sélectionner la variable qualitative à filtrer ici.

Sélectionner les catégories sur lesquelles filtrer ici.


On peut filtrer autant de variables qu'on veut, revenir sur une variable, changer les catégories sélectionnées, etc.

Les 19 catégories les plus importantes sont affichées par ordre décroissant, les autres catégories sont rassemblées dans '__Autres__'

Déclencher la filtration ou réinitialiser le jeu de données ici.

Vérifier le résultat de la filtration ici. Le jeu de données est maintenant modifié.

1. Zone de filtration

Sélectionner pour filtrer:  Vider la cellule !

ingredients_text

salt
water
sugar
missing
citric acid
sel
niacin
sucre
riboflavin
eau
dextrose
corn syrup
folic acid)
spices
reduced iron
natural flavor
Water
thiamine mononitrate
sea salt
Autres

▼ Filtrer

🔄 Réinitialiser

Taille du jeu de données

(215981, 53)

DataXplor

Variables dont les valeurs sont des listes

Certaines variables ont pour valeur des listes de mots. Ces variables terminent par _tags, _fr ou _text.

Elles ont un traitement spécifique: chaque mot des listes est considéré comme une catégorie et les mots sont classés par le nombre d'individus dont la variable contient ce mot.

Exemple avec la variables 'ingredient_text' ici

salt est le mot qui revient le plus souvent dans les listes d'ingrédients des produits, suivi par water.

2. Zone de sélection

Sélection des données à visualiser ici

Sélectionner pour visualiser:

code
creator
product_name
generic_name
quantity
packaging_tags
brands_tags
categories_fr
manufacturing_places_tags
labels_fr
emb_codes_tags
purchase_places
stores
countries_fr
ingredients_text
allergens
serving_size
additives_n
additives
ingredients_from_palm_oil_n
ingredients_that_may_be_from_palm_oil_n
nutrition_grade_fr
pnns_groups_1
pnns_groups_2
states_fr
main_category_fr
energy_100g
fat_100g

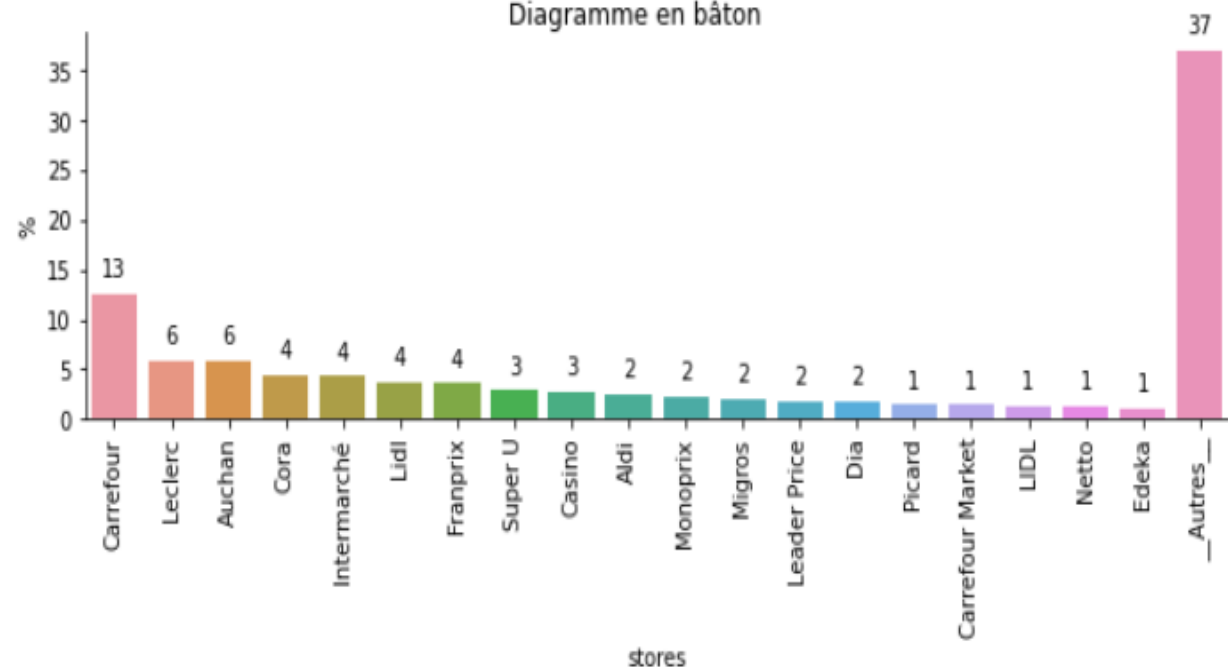
3. Zone de visualisation et d'information

1 variable qualitative => diagramme en bâtons

Nombre de bars:

20

Diagramme en bâton



2. Zone de sélection

Sélection des
données à visualiser
ici

Sélectionner pour visualiser:

ingredients_text
allergens
serving_size
additives_n
additives
ingredients_from_palm_oil_n
ingredients_that_may_be_from_palm_oil_n
nutrition_grade_fr
pnns_groups_1
pnns_groups_2
states_fr
main_category_fr
energy_100g
fat_100g
saturated-fat_100g
trans-fat_100g
cholesterol_100g
carbohydrates_100g
sugars_100g
fiber_100g
proteins_100g
salt_100g
sodium_100g
vitamin-a_100g
vitamin-c_100g
calcium_100g
iron_100g
nutrition-score-fr_100g

3. Zone de visualisation et d'information

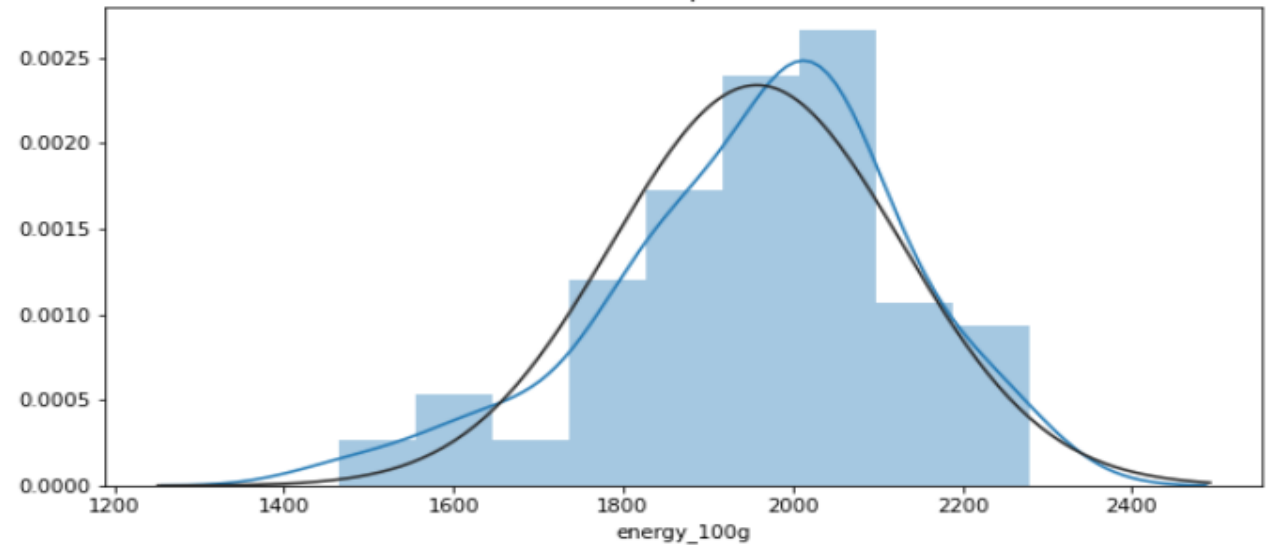
1 variable quantitative => densité de probabilité
avec tests d'ajustement à une loi normale

Plage des abscisses:

1189.3 – 2553.7

Test statistique d'ajustement
Hypothèse: distribution normale
Test de Shapiro: pvalue = 4.75%
Test de Kolmogorov: pvalue = 34.32%

Densité de probabilité



2. Zone de sélection

Sélection des données à visualiser ici

Sélectionner pour visualiser:

categories_fr
manufacturing_places_tags
labels_fr
emb_codes_tags
purchase_places
stores
countries_fr
ingredients_text
allergens
serving_size
additives_n
additives
ingredients_from_palm_oil_n
ingredients_that_may_be_from_palm_oil_n
nutrition_grade_fr
pnns_groups_1
pnns_groups_2
states_fr
main_category_fr
energy_100g
fat_100g
saturated-fat_100g
trans-fat_100g
cholesterol_100g
carbohydrates_100g
sugars_100g
fiber_100g
proteins_100g
salt_100g
sodium_100g
vitamin-a_100g
vitamin-c_100g
calcium_100g
iron_100g
nutrition-score-fr_100g

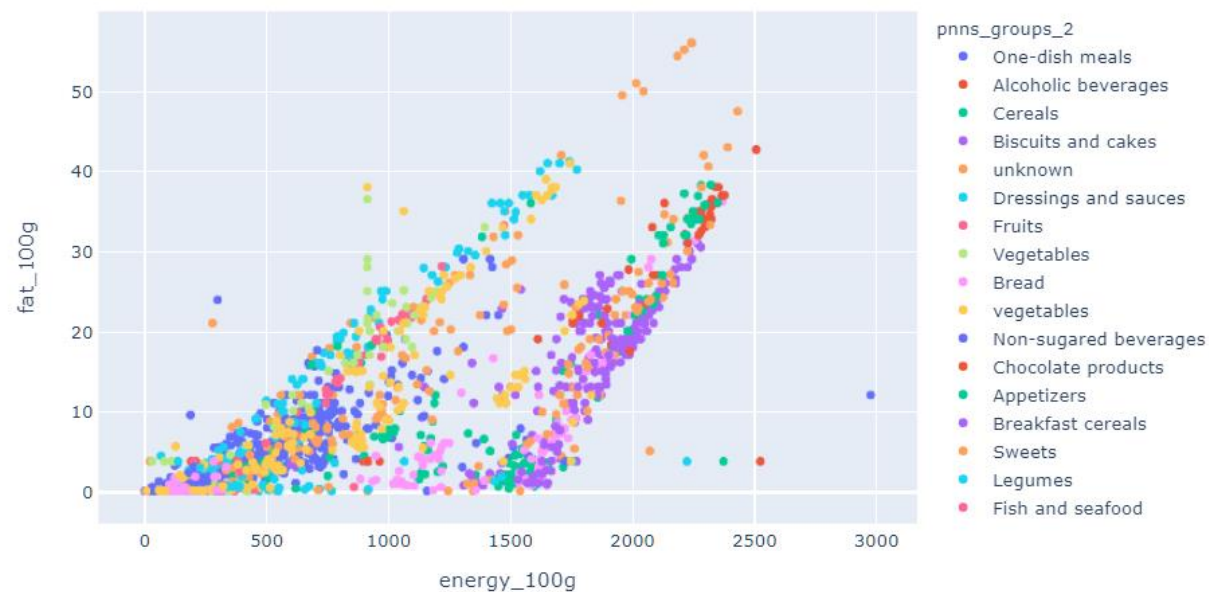
3. Zone de visualisation et d'information

2 variables quantitatives => nuage de point avec la corrélation entre les deux variables. Possibilité de colorer des catégories selon 3^{ème} variable.

pnns_groups_2

La corrélation entre les deux variables est de: 0.686

Nuage de points



2. Zone de sélection

Sélection des données à visualiser ici

Selectionner pour visualiser:

categories_fr
manufacturing_places_tags
labels_fr
emb_codes_tags
purchase_places
stores
countries_fr
ingredients_text
allergens
serving_size
additives_n
additives
ingredients_from_palm_oil_n
ingredients_that_may_be_from_palm_oil_n
nutrition_grade_fr
pnns_groups_1
pnns_groups_2
states_fr
main_category_fr
energy_100g
fat_100g
saturated-fat_100g
trans-fat_100g
cholesterol_100g
carbohydrates_100g
sugars_100g
fiber_100g
proteins_100g
salt_100g

3. Zone de visualisation et d'information

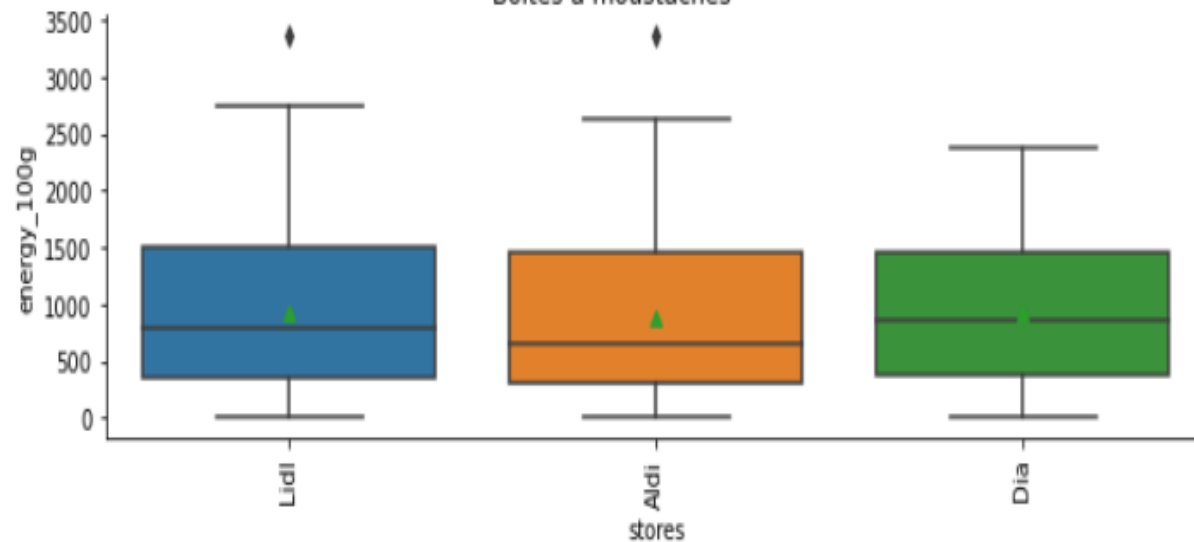
2 variables quantitative/ quantitative => boîtes à moustaches avec analyse de la variance.

Test statistique d'analyse de la variance

Hypothèse: 'stores' n'a pas d'influence sur 'energy_100g'

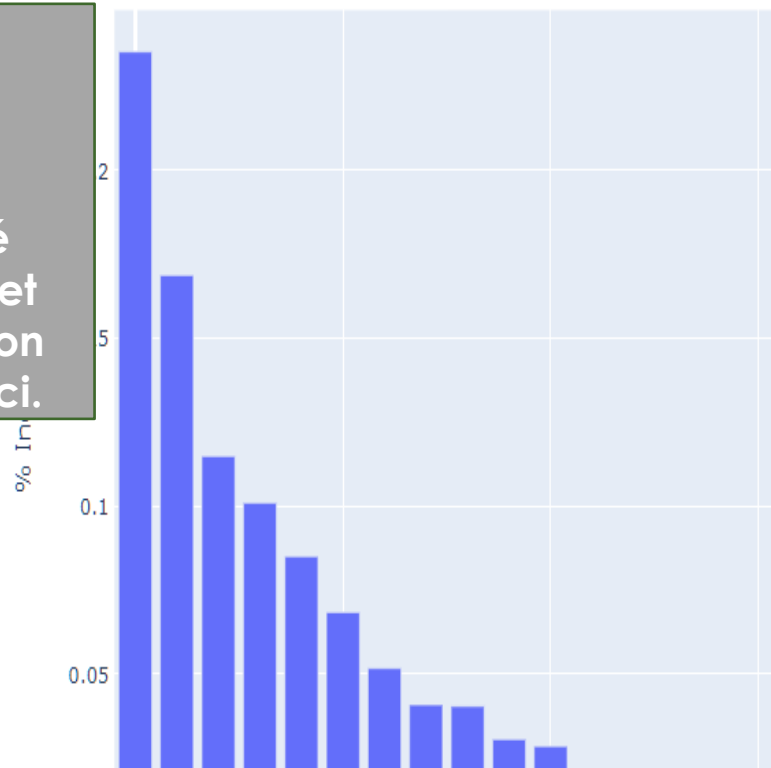
ANOVA: pvalue = 20.84%

Boîtes à moustaches

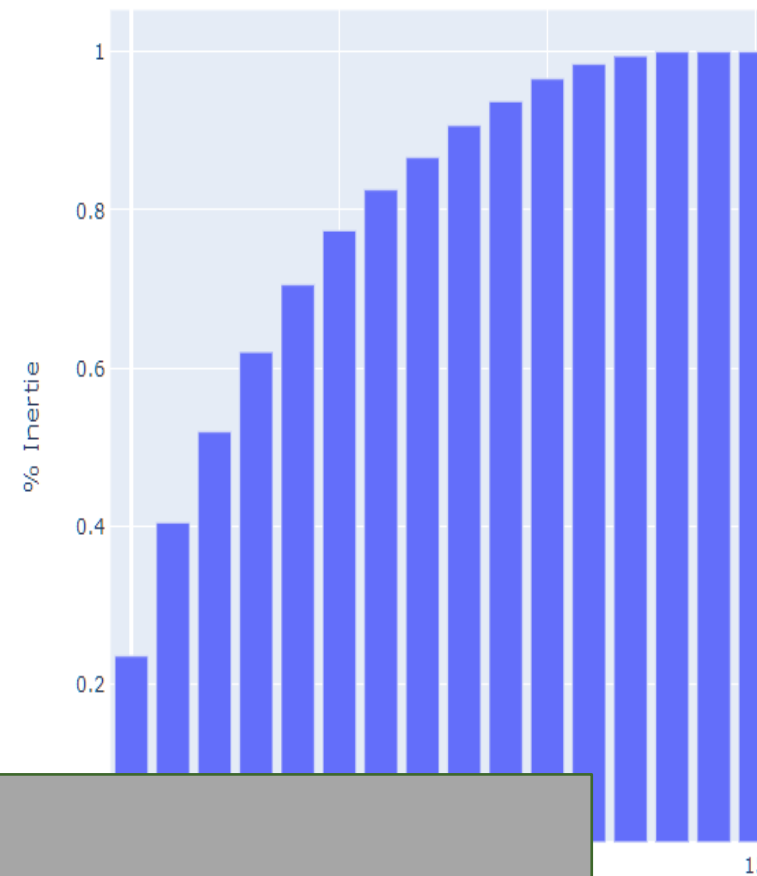


⏮ (Re)Lancer l'ACP

Eboulis de la contribution des facteurs à l'inertie



Eboulis cumulés



Relancer l'ACP sur le jeu de données sélectionné dans l'onglet d'exploration générale, ici.

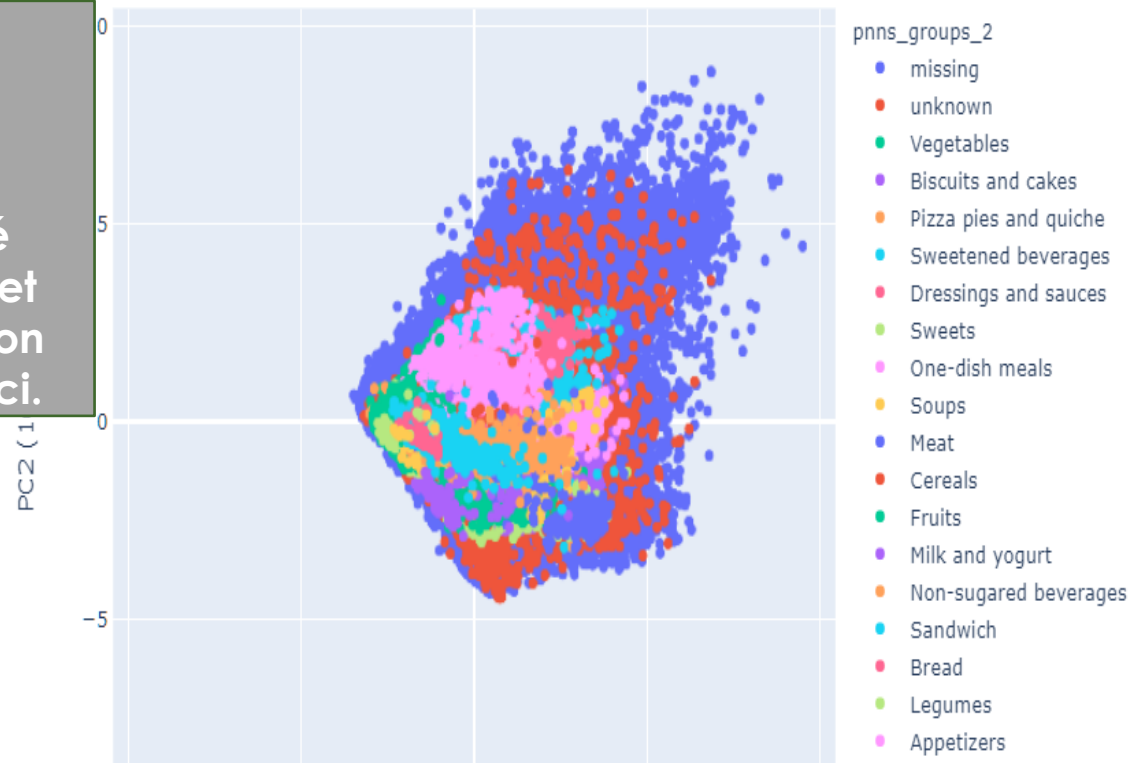
Les onglets donnent accès:

- Aux graphes des éboulis
- Aux graphes des plans factoriels et cercles de corrélations selon les 8 premiers facteurs issus de l'ACP.

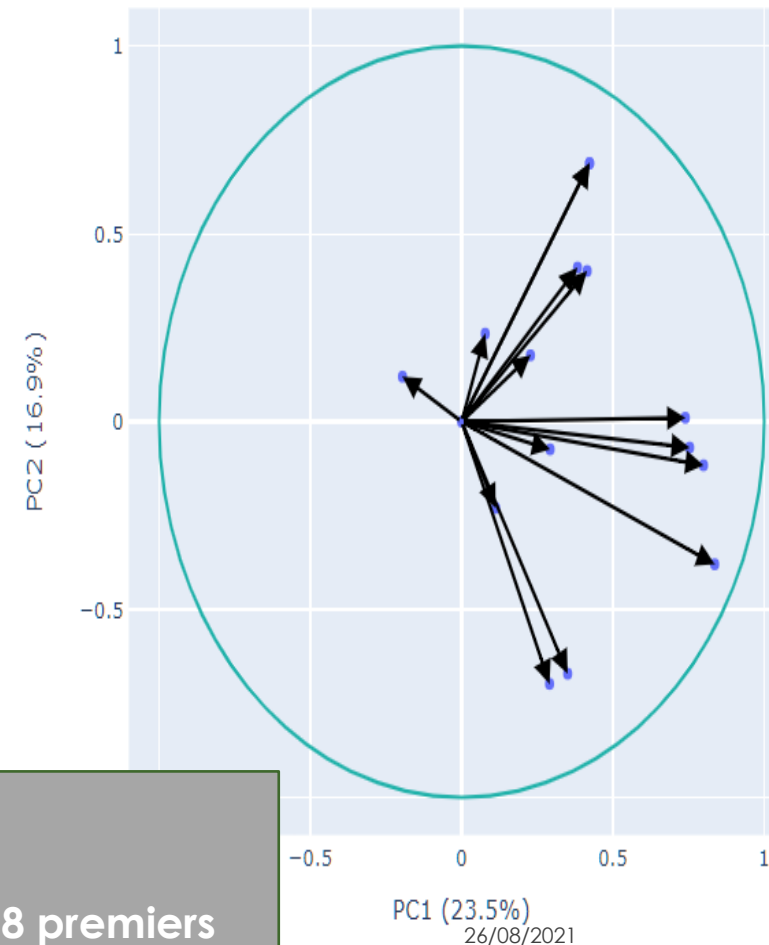
Dashboard

(Re)Lancer l'ACP

Plan factoriel PC1 / PC2



Cercle des corrélations - PC1 / PC2



Relancer l'ACP sur le jeu de données sélectionné dans l'onglet d'exploration générale, ici.

Les onglets donnent accès:

- Aux graphes des éboulis
- Aux graphes des plans factoriels et cercles de corrélations selon les 8 premiers facteurs issus de l'ACP.