PREPARER DES DONNEES POUR UN ORGANISME DE SANTE PUBLIQUE

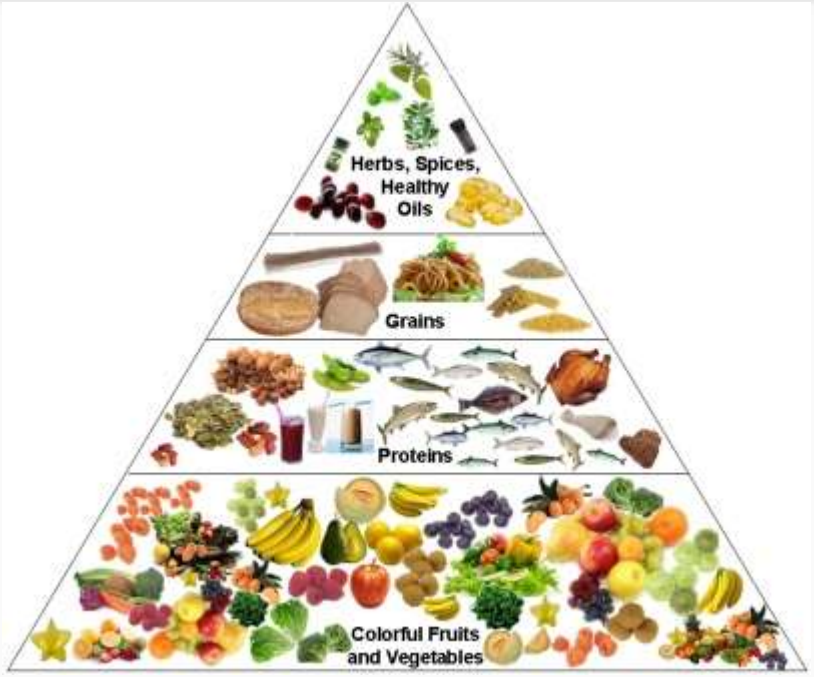
Présentation “Projet 3” chez “OPENCLASSROOM”

Jaoid KRAIRI

(Juin 2021)

SOMMAIRE

PRESENTATION DE L'APPEL A PROJET :
Rendre les données de santé plus accessible



DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES : 1/ Importer les librairies Python et le jeu de données

Importer les librairies

Entrée [1]:

```
1  import pandas as pd
2  import numpy as np
3  import seaborn as sb
4  from scipy import stats
5  from scipy.stats import uniform
6  from scipy.stats import norm
7  from scipy.stats import beta
8  from sklearn import decomposition
9  from sklearn import preprocessing
10 from functions import *
11 import matplotlib.pyplot as plt
12 from sklearn.preprocessing import StandardScaler
13 import statsmodels.api as sm
```

Charger le fichier dans un dataframe

Entrée [2]:

```
1  missing_values = ["n/a", "na", "--", "0"]
```

Entrée [3]:

```
1  df = pd.read_table("fr.openfoodfacts.org.products.csv", na_values= missing_values)
```

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
2/ Affichage partiel du contenu des données

Nombre de lignes et de colonnes

```
Entrée [4]:  1 df.shape
Out[4]:  (320772, 162)
```

Afficher les 5 première lignes de ma dataframe

```
Entrée [5]:  1 df.head(5)
```

Out[5]:

	code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	generic_name	c
0	3087	http://world-fr.openfoodfacts.org/produit/0000...	openfoodfacts-contributors	1474103866	2016-09-17T09:17:46Z	1474103893	2016-09-17T09:18:13Z	Farine de blé noir	NaN	
1	4530	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Banana Chips Sweetened (Whole)	NaN	
2	4559	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Peanuts	NaN	
3	16087	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055731	2017-03-09T10:35:31Z	1489055731	2017-03-09T10:35:31Z	Organic Salted Nut Mix	NaN	
4	16094	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055653	2017-03-09T10:34:13Z	1489055653	2017-03-09T10:34:13Z	Organic Polenta	NaN	

Afficher les 5 dernières lignes de ma dataframe

```
Entrée [6]:  1 df.tail(5)
```

Out[6]:

	code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	g
320767	9948282780603	http://world-fr.openfoodfacts.org/produit/9948...	openfoodfacts-contributors	1490631299	2017-03-27T16:14:59Z	1491244498	2017-04-03T18:34:58Z	Tomato & ricotta	
320768	99567453	http://world-fr.openfoodfacts.org/produit/9956...	usda-ndb-import	1489059076	2017-03-09T11:31:16Z	1491244499	2017-04-03T18:34:59Z	Mint Melange Tea A Blend Of Peppermint, Lemon ...	
320769	9970229501521	http://world-fr.openfoodfacts.org/produit/9970...	tomato	1422099377	2015-01-24T11:36:17Z	1491244499	2017-04-03T18:34:59Z	乐吧泡菜味香片	
320770	9980282863788	http://world-fr.openfoodfacts.org/produit/9980...	openfoodfacts-contributors	1492340089	2017-04-16T10:54:49Z	1492340089	2017-04-16T10:54:49Z	Tomates aux Vermicelles	
320771	999990026839	http://world-fr.openfoodfacts.org/produit/9999...	usda-ndb-import	1489072709	2017-03-09T15:18:29Z	1491244499	2017-04-03T18:34:59Z	Sugar Free Drink Mix, Peach Tea	

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES : 3/ Déterminer le nombre de variables qualitatives et quantitatives

Verifier les différents types de nos variable

Entrée [7]:
```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 320772 entries, 0 to 320771
Columns: 162 entries, code to water-hardness_100g
dtypes: float64(106), object(56)
memory usage: 396.5+ MB
```

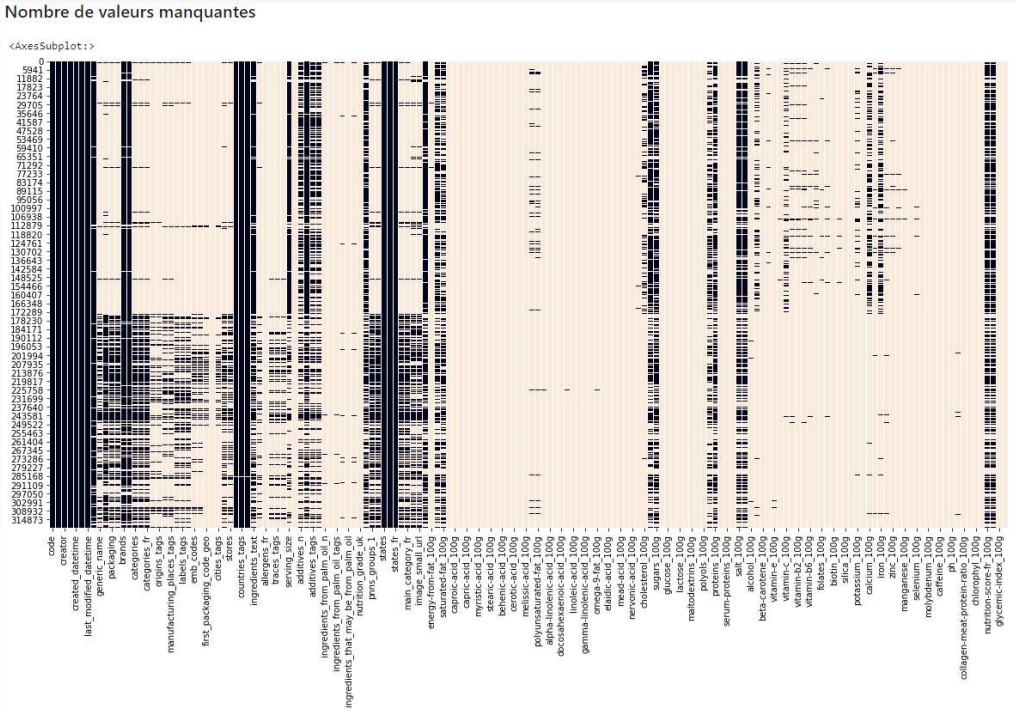
Entrée [8]:
```
1 df.dtypes.value_counts()
```

Out[8]:
```
float64    106
object      56
dtype: int64
```

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
4/ Vérifier si le jeu de données comporte des valeurs manquantes

```
Entrée [11]:  1  df.isnull().sum().sum()

Out[11]:  41113087
```



DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
5/ Supprimer et déterminer le nouveaux nombres de variables

Entrée [15]:
```
for column in column_with_nan:
    if df[column].isnull().sum()*100.0/df.shape[0] > 50:
        df.drop(column,1, inplace=True)
```

Entrée [16]:
```
df.shape
```

Out[16]:
```
(320772, 30)
```

Entrée [17]:
```
df.dtypes.value_counts()
```

Out[17]:
```
object     20
float64    10
dtype: int64
```

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
6/ Couper notre data frame en 2 parties ((Catégorique : Qualitative) et (Numérique : Quantitative))

cat_data: Categorique et num_data: Numerique

Entrée [18]:

```
cat_data=[]
num_data=[]

for i,c in enumerate(df.dtypes):
    if c==object:
        cat_data.append(df.iloc[:,i])
    else:
        num_data.append(df.iloc[:,i])
cat_data=pd.DataFrame(cat_data).transpose()
num_data=pd.DataFrame(num_data).transpose()
```

Cette condition me permet de traiter de manière séparer les variables qualitative dite 'categorique' et les variables quantitative dite 'numerique'

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
7/ Remplacer les valeurs manquantes Catégorique

Remplacer les valeurs manquantes Categorique

Pour les variables qualitative on va remplacer les valeurs manquantes par les valeurs qui se répétent le plus souvent.

```
code                      0
url                       0
creator                   0
created_t                 0
created_datetime          0
last_modified_t           0
last_modified_datetime    0
product_name              0
brands                    0
brands_tags               0
countries                 0
countries_tags            0
countries_fr              0
ingredients_text          0
serving_size              0
additives                 0
nutrition_grade_fr        0
states                    0
states_tags               0
states_fr                 0
dtype: int64
```

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
8/ Remplacer les valeurs manquantes Numérique

Remplacer les valeurs manquantes Numerique

```
energy_100g                0
fat_100g                   0
saturated-fat_100g         0
carbohydrates_100g         0
sugars_100g                0
proteins_100g              0
salt_100g                  0
sodium_100g                0
nutrition-score-fr_100g    0
nutrition-score-uk_100g    0
dtype: int64
```

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
9/ Résumer statistique des variables Catégorique de manière rapide

Résumer statistique des variables Categorique de manière rapide

	code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	brands	brands_tags	countries	countries_tags	countries_fr	ingredients_text	serving_size	additives	nutrition_grade_fr	states	states_tags
count	320772	320772	320772	320772	320772	320772	320772	320772	320772	320772	320772	320772	320772	320772	320772	320772	320772	320772	320772
unique	320638	320749	3535	189635	189568	180639	180495	221347	58783	50252	1434	725	722	205520	25422	196069	5	1021	1021
top	722810	http://world-fr.openfoodfacts.org/produit/0011...	usda-ndb-import	1489077120	2017-03-09T16:32:00Z	1439141742	2015-08-09T17:35:42Z	Ice Cream	Carrefour	carrefour	US	en:united-states	États-Unis	Carbonated water, natural flavor.	240 ml (8 fl oz)	[extra-virgin-olive-oil -> en:extra-virgin-o...	d	en:to-be-completed, en:nutrition-facts-complet...	en:to-be-completed,en:nutrition-facts-complete...
freq	25	24	169870	24	29	33	33	18172	31391	31575	170208	173278	173278	72032	114946	72264	162325	168951	168951

states_fr
320772
1021
A compléter,Informations nutritionnelles compl...
168951

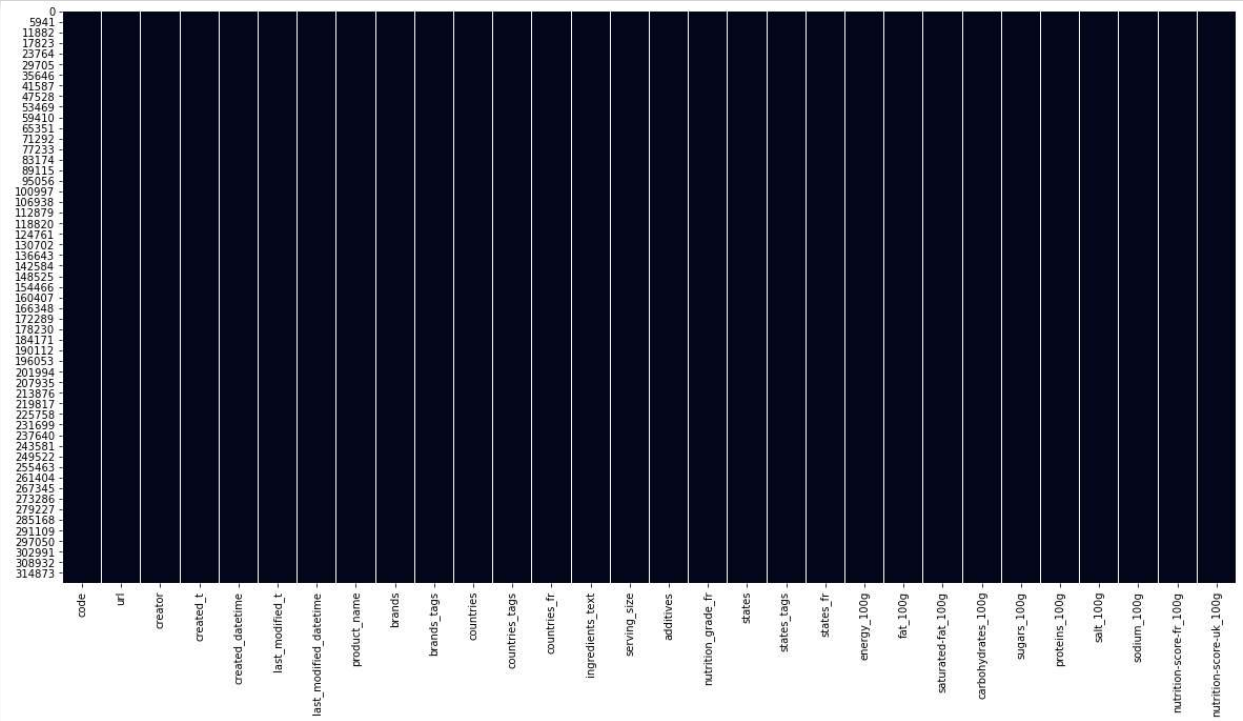
DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
10/ Résumer statistique des variables numérique de manière rapide

Résumer statistique des variables numerique de manière rapide

	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	proteins_100g	salt_100g	sodium_100g	nutrition-score-fr_100g	nutrition-score-uk_100g
count	3.207720e+05	320772.000000	320772.000000	320772.000000	320772.000000	320772.000000	3.207720e+05	3.207720e+05	320772.000000	320772.000000
mean	1.175999e+03	14.933142	5.915072	32.425402	15.423975	8.114114	1.856043e+00	7.307277e-01	10.172593	9.774017
std	5.813586e+03	14.064055	6.317097	24.748338	19.171269	6.921555	1.144799e+02	4.507081e+01	7.305940	7.381454
min	2.000000e-02	0.000100	0.000100	0.001000	-17.860000	-800.000000	5.000000e-08	1.968504e-08	-15.000000	-15.000000
25%	5.980000e+02	9.820000	4.460000	13.900000	5.000000	5.000000	4.000000e-01	1.574803e-01	6.000000	6.000000
50%	1.153000e+03	11.920000	4.500000	26.200000	9.100000	6.670000	7.747000e-01	3.050000e-01	11.000000	10.000000
75%	1.569000e+03	14.290000	4.500000	49.900000	14.840000	8.240000	1.115060e+00	4.390000e-01	13.000000	13.000000
max	3.251373e+06	714.290000	550.000000	2916.670000	3520.000000	430.000000	6.431280e+04	2.532000e+04	40.000000	40.000000

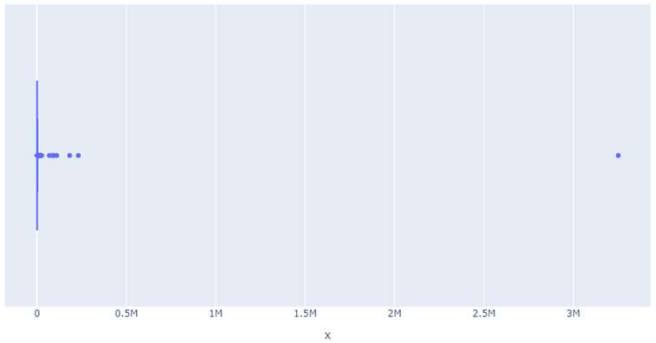
DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :

11/ Réunir en une seul Data frame et vérifier la présence ou non de valeurs manquante

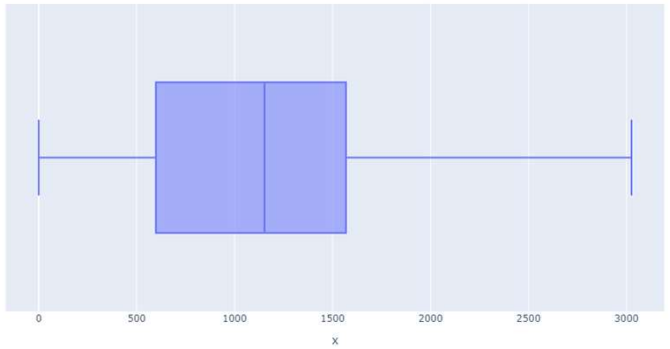


DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
12/ Remplacer les valeurs Outliers Numérique

AVANT



APRES



DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
13/ Supprimer ligne dupliquée

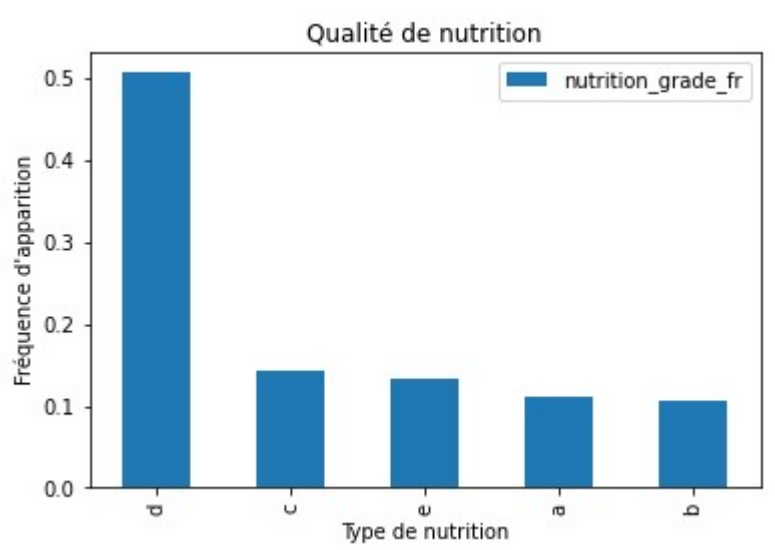
AVANT

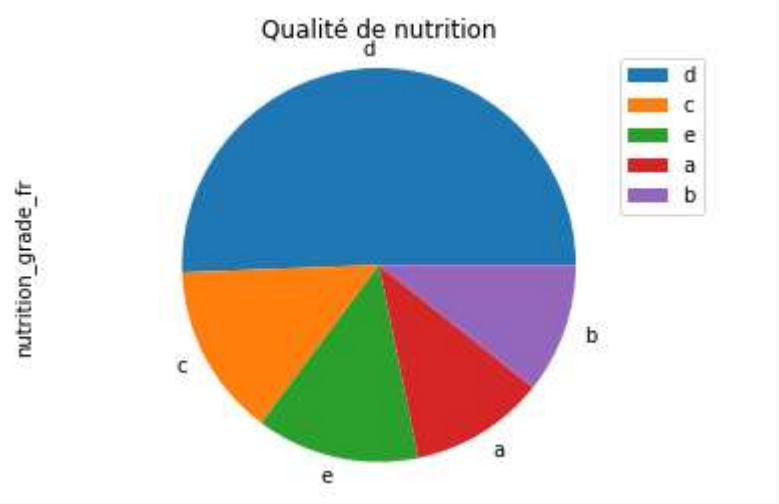
```
Entrée [66]:  1  df.shape
Out[66]:  (320772, 30)
```

APRES

```
Entrée [68]:  1  df.shape
Out[68]:  (320771, 30)
```

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
14/ Supprimer ligne dupliquée





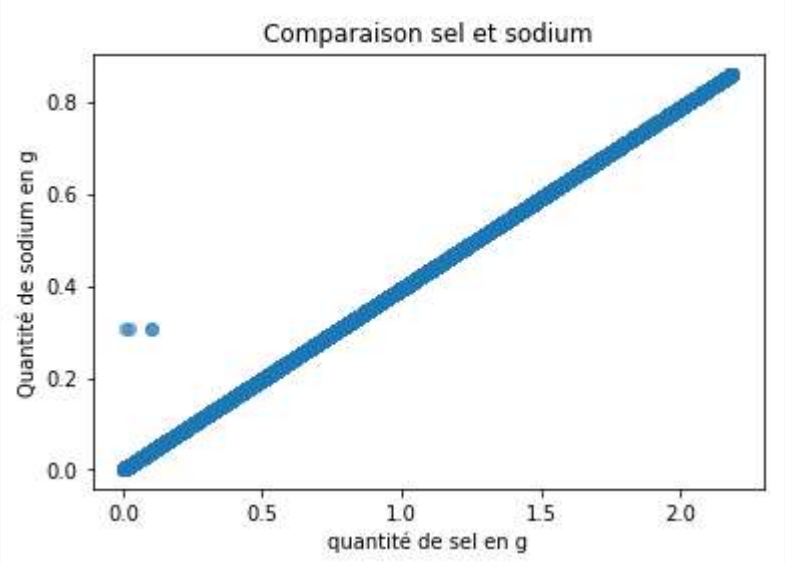
DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
15/ Analyse statistique par qualité nutritionnel

nutrition_grade_fr	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	proteins_100g	salt_100g	sodium_100g	nutrition-score-fr_100g	nutrition-score-uk_100g
a	719.664484	15.065217	4.528177	29.470175	5.102043	7.355846	0.470875	0.185384	-2.904052	-2.903898
b	547.141751	13.484720	4.531970	19.675720	6.537832	5.607392	0.605583	0.238419	4.990261	4.645101
c	1018.341438	13.340310	4.538973	32.069334	10.722789	6.823762	0.968172	0.381171	6.367561	6.307139
d	1241.531184	13.751148	4.520024	33.601931	12.686343	7.131537	0.924207	0.363868	12.185838	11.628040
e	1832.609116	18.715844	4.556386	40.882469	19.777586	7.531806	1.084118	0.426818	12.957553	12.354543

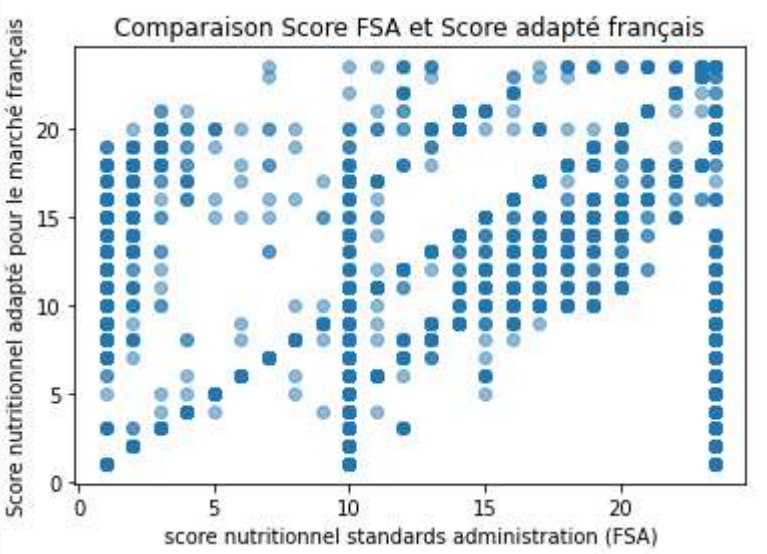
Quantité de sucre en gramme

Score nutritionnel

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
16/ Analyser la corrélation entre 2 variables quantitative 'salt_100g' et 'sodium_100g'



DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
17/ Analyser la corrélation entre 2 variables quantitative 'nutrition-score-uk_100g' et 'nutrition-score-fr_100g'



DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
18/ Loi normale variable quantitative
1/ Présentation de la méthode utilisée

Qu'est-ce qu'une distribution normale ?

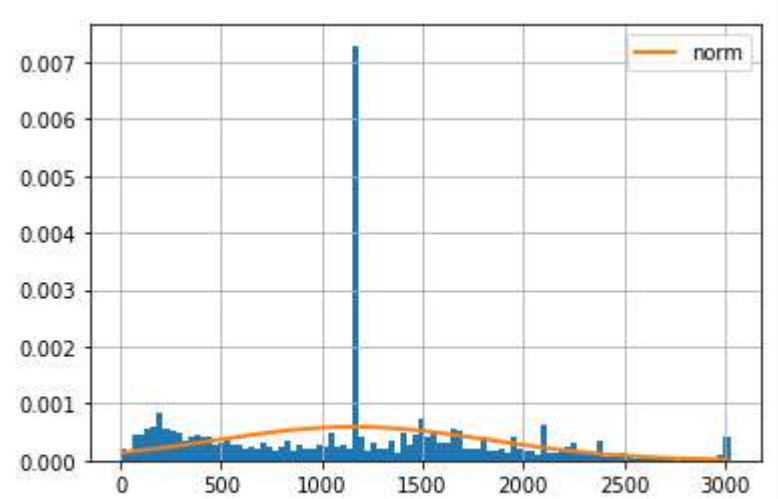
Pourquoi la loi normale est-elle intéressante ?

Fitter qu'est-ce que c'est?

Test de Kolmogorov-Smirnov qu'est-ce que c'est?

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
18/ Loi normale variable qualitative
2/ Mise en application

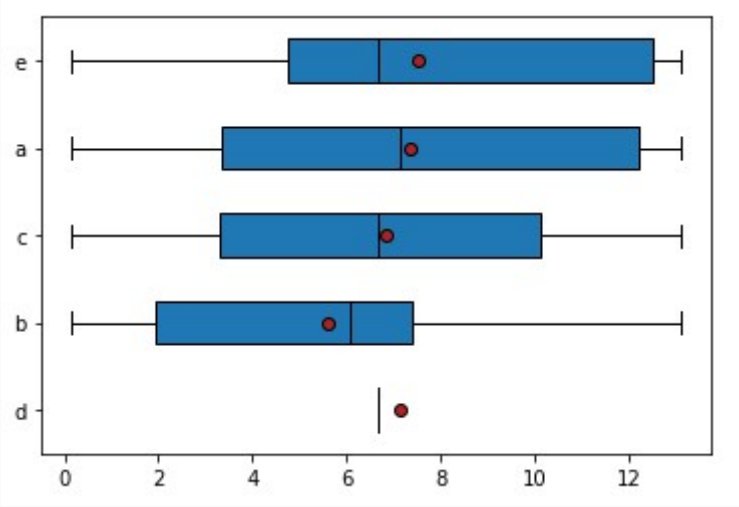
Loi normale variable qualitative 'energy_100g'



P < 0,1 → (0.10925805318432746, 0.0009999999999998899) ← Renvoi 'table'

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :

18/ Méthode d'analyse explicative (ANOVA) exemple variable quantitative 'proteins_100g' et qualitative 'nutrition_grade_fr'



DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
19/ Méthode d'analyse descriptive (ACP)
1/Préparation ACP

Importer les librairies

Entrée [1]:
```
import pandas as pd
import numpy as np
import seaborn as sb
from sklearn import decomposition
from sklearn import preprocessing
from functions import *
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
```

Entrée [99]:
```
n_comp = 9
```

Entrée [100]:
```
df_pca = df[['energy_100g', 'fat_100g','saturated-fat_100g','carbohydrates_100g', 'sugars_100g', 'proteins_100g',
             'salt_100g', 'sodium_100g','nutrition-score-fr_100g', 'nutrition-score-uk_100g']]
```

Entrée [101]:
```
X = df_pca.values
X
```

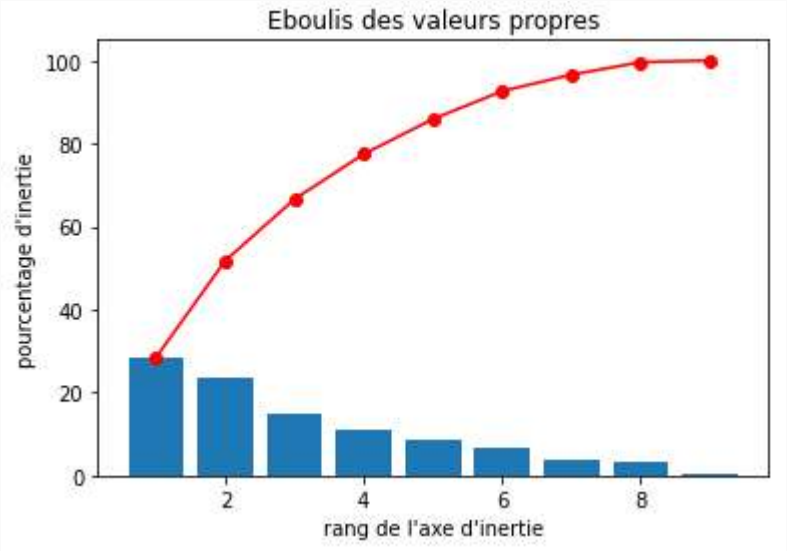
Entrée [102]:
```
names = df['code']
features = df_pca.columns
```

Entrée [103]:
```
std_scale = preprocessing.StandardScaler().fit(X)
X_scaled = std_scale.transform(X)
```

Entrée [104]:
```
pca = decomposition.PCA(n_components=n_comp)
pca.fit(X_scaled)
```

Entrée [105]:
```
display_scree_plot(pca)
```

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
19/ Méthode d'analyse descriptive (ACP)
2/Affichage de l'inertie

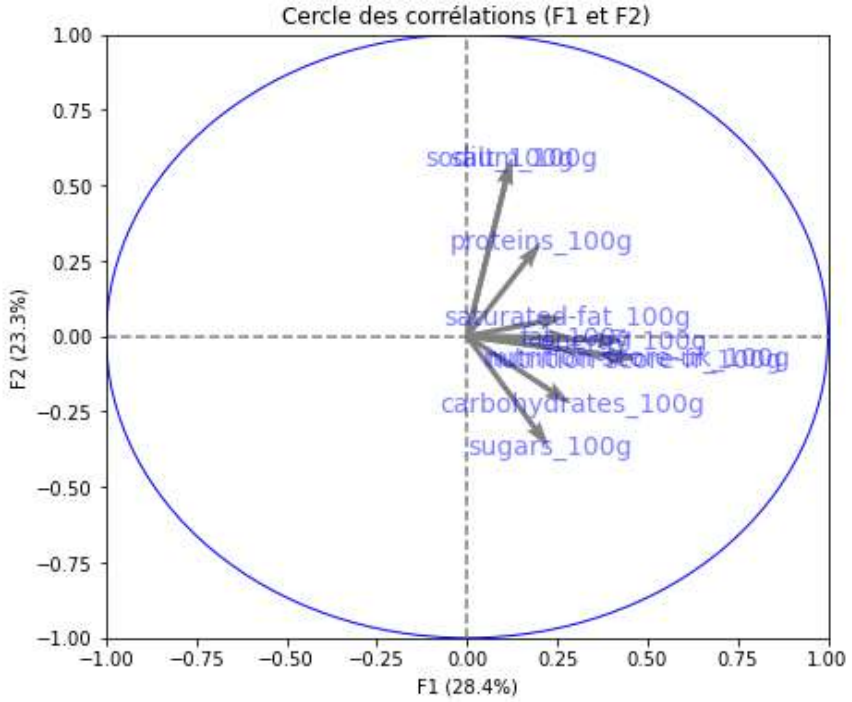


DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
19/ Méthode d'analyse descriptive (ACP)
3/Calcule des nouvelles composantes

Entrée [112]:

```
pcs = pca.components_
display_circles(pcs, n_comp, pca, [(0,1)], labels = np.array(features))
```

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
19/ Méthode d'analyse descriptive (ACP)
4/Cercle des corrélations

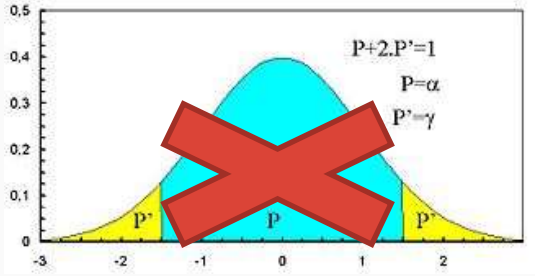


DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
20/ Classement Pays et Marque en fonction du score nutritionnel

countries	
ALLEMAGNE	11.000000
Albania	11.000000
Albania,Italia	14.000000
Albania,Italia, en:denmark	11.000000
Albania,Italy	2.666667
Alemanha,Portugal,Espanha	7.000000
Alemania, España	4.500000
Alemania,España	8.464286
Algeria	11.000000
Algeria, en:france	11.000000

brands	
365 дней	11.0
Act II	11.0
Annie's	11.0
Boucherie	11.0
Carrefour	11.0
Casino	11.0
Casino	11.0
Casino délices	11.0
Core Meal, Core Method	-2.0
Debic	11.0

DEMARCHE METHODOLOGIQUE DE NETTOYAGE ET D'EXPLOITATION DE DONNEES :
21/ Conclusion générale



La combinaison de 2 pays
La combinaison de 2 marques

LE PROTOTYPE REALISE :

PAGE WEB

CONCLUSION

- ✓ Le nettoyage des données est complet
- ✓ Le nettoyage des données est pertinent
- ✓ Le nettoyage des données est présentable
- ✓ L'analyse statistique multivariée est complète
- ✓ L'analyse statistique multivariée est pertinente
- ✓ La communication des résultats à l'aide de représentations graphiques est complète
- ✓ Résultat pertinente et présentable

REMERCIEMENT

- Remercier mon Mentor
- Remercier l’équipe pédagogique OPENCLASSROOM
- Merci de m’avoir écouter

REPONDRE AUX QUESTIONS