MN_bfill_Clustering

May 3, 2024

[1]:
```
!pip install scikit-learn-extra
```

```
Requirement already satisfied: scikit-learn-extra in
c:\users\tky12\anaconda3\lib\site-packages (0.3.0)
Requirement already satisfied: numpy>=1.13.3 in
c:\users\tky12\anaconda3\lib\site-packages (from scikit-learn-extra) (1.26.4)
Requirement already satisfied: scipy>=0.19.1 in
c:\users\tky12\anaconda3\lib\site-packages (from scikit-learn-extra) (1.11.4)
Requirement already satisfied: scikit-learn>=0.23.0 in
c:\users\tky12\anaconda3\lib\site-packages (from scikit-learn-extra) (1.2.2)
Requirement already satisfied: joblib>=1.1.1 in
c:\users\tky12\anaconda3\lib\site-packages (from scikit-learn>=0.23.0->scikit-
learn-extra) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in
c:\users\tky12\anaconda3\lib\site-packages (from scikit-learn>=0.23.0->scikit-
learn-extra) (2.2.0)
```

[2]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from mpl_toolkits.mplot3d import Axes3D
from scipy.stats import multivariate_normal
from scipy.stats import norm
import plotly.express as px
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import silhouette_score
import pickle
```

[3]:
```
# Surpress warnings
def warn(*args, **kwargs):
    pass
import warnings
warnings.warn = warn
```

[4]:
```
pd.DataFrame.iteritems = pd.DataFrame.items
```

[5]:
```
# Retrieve dataset and read first 5 rows
macroNutrient_bfill = pd.read_csv(r"../Dataset/Dataset_for_EDA/
 ↪macroNutrient_bfill.csv", encoding= 'unicode_escape')
macroNutrient_bfill
```

[5]:

	No.	Description	Category	\
0	15155	ABALONE,MIXED SPECIES,RAW	ABALONE	
1	15156	ABALONE,MXD SP,CKD,FRIED	ABALONE	
2	9427	ABIYUCH,RAW	ABIYUCH	
3	9002	ACEROLA JUICE,RAW	ACEROLA JUICE	
4	9001	ACEROLA,(WEST INDIAN CHERRY),RAW	ACEROLA	
…	…	…	…	
8785	1119	YOGURT,VANILLA,LOFAT,11 GRAMS PROT PER 8 OZ	YOGURT	
8786	1220	YOGURT,VANILLA,LOFAT,11 GRAMS PROT PER 8 OZ,FO…	YOGURT	
8787	1295	YOGURT,VANILLA,NON-FAT	YOGURT	
8788	16004	YOKAN,PREP FROM ADZUKI BNS & SUGAR	YOKAN	
8789	3217	ZWIEBACK	ZWIEBACK	

	Energy (Kcal)	Carbohydrate(g)	Protein(g)	Total Lipid(g)	\
0	105	6.01	17.10	0.76	
1	189	11.05	19.63	6.78	
2	69	17.60	1.50	0.10	
3	23	4.80	0.40	0.30	
4	32	7.69	0.40	0.30	
…	…	…	…	…	
8785	85	13.80	4.93	1.25	
8786	85	13.80	4.93	1.25	
8787	78	17.04	2.94	0.00	
8788	260	60.72	3.29	0.12	
8789	426	74.20	10.10	9.70	

	Monounsaturated Fatty Acids(g)	Polyunsaturated Fatty Acids(g)	\
0	0.107	0.104	
1	2.741	1.676	
2	0.082	0.090	
3	0.082	0.090	
4	0.082	0.090	
…	…	…	
8785	0.343	0.036	
8786	0.343	0.036	
8787	0.000	0.000	
8788	0.011	0.026	
8789	4.244	2.073	

	Saturated Fatty Acids(g)
0	0.149
1	1.646

```
2                       0.014
3                       0.068
4                       0.068
…                       …
8785                    0.806
8786                    0.806
8787                    0.000
8788                    0.043
8789                    2.525

[8790 rows x 10 columns]
```

[6]:
```
macroNutrient_bfill.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8790 entries, 0 to 8789
Data columns (total 10 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   No.                             8790 non-null   int64
 1   Description                     8790 non-null   object
 2   Category                        8790 non-null   object
 3   Energy (Kcal)                   8790 non-null   int64
 4   Carbohydrate(g)                 8790 non-null   float64
 5   Protein(g)                      8790 non-null   float64
 6   Total Lipid(g)                  8790 non-null   float64
 7   Monounsaturated Fatty Acids(g)  8790 non-null   float64
 8   Polyunsaturated Fatty Acids(g)  8790 non-null   float64
 9   Saturated Fatty Acids(g)        8790 non-null   float64
dtypes: float64(6), int64(2), object(2)
memory usage: 686.8+ KB
```

[7]:
```
macroNutrient_bfill.describe()
```

[7]:
```
                No.  Energy (Kcal)  Carbohydrate(g)   Protein(g)  \
count   8790.000000    8790.000000      8790.000000  8790.000000
mean   15663.495222     226.317634        22.127710    11.342849
std     9251.413586     169.877539        27.270822    10.530474
min     1001.000000       0.000000         0.000000     0.000000
25%     9086.250000      91.000000         0.050000     2.380000
50%    14427.500000     191.000000         9.340000     8.000000
75%    20142.750000     337.000000        34.910000    19.880000
max    93600.000000     902.000000       100.000000    88.320000

       Total Lipid(g)  Monounsaturated Fatty Acids(g)  \
count     8790.000000                     8790.000000
mean        10.553725                        3.960572
std         15.814842                        6.925061
```

```
min          0.000000                      0.000000
25%          0.950000                      0.200000
50%          5.140000                      1.790000
75%         13.720000                      4.938500
max        100.000000                     83.689000

       Polyunsaturated Fatty Acids(g)  Saturated Fatty Acids(g)
count                     8790.000000               8790.000000
mean                         2.252043                  3.508612
std                          5.233486                  6.463679
min                          0.000000                  0.000000
25%                          0.202250                  0.200000
50%                          0.671000                  1.506000
75%                          2.049000                  4.240000
max                         74.623000                 95.600000
```

[8]:
```
macroNutrient_bfill.head()
```

[8]:
```
     No.                       Description       Category  Energy (Kcal)  \
0  15155         ABALONE,MIXED SPECIES,RAW        ABALONE            105
1  15156          ABALONE,MXD SP,CKD,FRIED        ABALONE            189
2   9427                       ABIYUCH,RAW        ABIYUCH             69
3   9002                ACEROLA JUICE,RAW  ACEROLA JUICE             23
4   9001  ACEROLA,(WEST INDIAN CHERRY),RAW        ACEROLA             32

   Carbohydrate(g)  Protein(g)  Total Lipid(g)  \
0             6.01       17.10            0.76
1            11.05       19.63            6.78
2            17.60        1.50            0.10
3             4.80        0.40            0.30
4             7.69        0.40            0.30

   Monounsaturated Fatty Acids(g)  Polyunsaturated Fatty Acids(g)  \
0                           0.107                           0.104
1                           2.741                           1.676
2                           0.082                           0.090
3                           0.082                           0.090
4                           0.082                           0.090

   Saturated Fatty Acids(g)
0                     0.149
1                     1.646
2                     0.014
3                     0.068
4                     0.068
```

[9]:
```
float_columns = [x for x in macroNutrient_bfill.columns if x not in ['No.',␣
↪'Description', 'Category']]
```

[10]:
```
skew_columns = (macroNutrient_bfill[float_columns]
                .skew()
                .sort_values(ascending=False))

skew_columns = skew_columns.loc[skew_columns > 0.75]
print("{} of the 7 columns are skewed with the vast majority being heavily␣
↪skewed".format(len(skew_columns)))
skew_columns
```

```
7 of the 7 columns are skewed with the vast majority being heavily skewed
```

[10]:
```
Saturated Fatty Acids(g)          6.582713
Polyunsaturated Fatty Acids(g)    6.281052
Monounsaturated Fatty Acids(g)    4.626117
Total Lipid(g)                    3.309724
Protein(g)                        1.166368
Energy (Kcal)                     1.148610
Carbohydrate(g)                   1.127598
dtype: float64
```

[11]:
```
# Perform log transform on skewed columns
for col in skew_columns.index.tolist():
    macroNutrient_bfill[col] = np.log1p(macroNutrient_bfill[col])
```

[12]:
```
macroNutrient_bfill[float_columns]
```

[12]:

	Energy (Kcal)	Carbohydrate(g)	Protein(g)	Total Lipid(g)	\
0	4.663439	1.947338	2.895912	0.565314	
1	5.247024	2.489065	3.026746	2.051556	
2	4.248495	2.923162	0.916291	0.095310	
3	3.178054	1.757858	0.336472	0.262364	
4	3.496508	2.162173	0.336472	0.262364	
…	…	…	…	…	
8785	4.454347	2.694627	1.780024	0.810930	
8786	4.454347	2.694627	1.780024	0.810930	
8787	4.369448	2.892592	1.371181	0.000000	
8788	5.564520	4.122608	1.456287	0.113329	
8789	6.056784	4.320151	2.406945	2.370244	

	Monounsaturated Fatty Acids(g)	Polyunsaturated Fatty Acids(g)	\
0	0.101654	0.098940	
1	1.319353	0.984323	
2	0.078811	0.086178	
3	0.078811	0.086178	
4	0.078811	0.086178	

```
…                                …                                …
8785                          0.294906                         0.035367
8786                          0.294906                         0.035367
8787                          0.000000                         0.000000
8788                          0.010940                         0.025668
8789                          1.657085                         1.122654

      Saturated Fatty Acids(g)
0                     0.138892
1                     0.973049
2                     0.013903
3                     0.065788
4                     0.065788
…                        …
8785                  0.591114
8786                  0.591114
8787                  0.000000
8788                  0.042101
8789                  1.259880

[8790 rows x 7 columns]
```

[13]:
```
macroNutrient_bfill.describe()
```

[13]:
```
                No.  Energy (Kcal)  Carbohydrate(g)   Protein(g)  \
count   8790.000000    8790.000000      8790.000000  8790.000000
mean   15663.495222       5.065342         2.137177     2.047259
std     9251.413586       0.997539         1.615109     1.068937
min     1001.000000       0.000000         0.000000     0.000000
25%     9086.250000       4.521789         0.048790     1.217876
50%    14427.500000       5.257495         2.336020     2.197225
75%    20142.750000       5.823046         3.581016     3.038792
max    93600.000000       6.805723         4.615121     4.492225

       Total Lipid(g)  Monounsaturated Fatty Acids(g)  \
count     8790.000000                     8790.000000
mean         1.770891                        1.102790
std          1.183944                        0.930723
min          0.000000                        0.000000
25%          0.667829                        0.182322
50%          1.814823                        1.026042
75%          2.689207                        1.781456
max          4.615121                        4.438986

       Polyunsaturated Fatty Acids(g)  Saturated Fatty Acids(g)
count                     8790.000000               8790.000000
mean                         0.759121                  1.039531
```

```
std                          0.758034                  0.890128
min                          0.000000                  0.000000
25%                          0.184195                  0.182322
50%                          0.513422                  0.918688
75%                          1.114814                  1.656321
max                          4.325760                  4.570579
```

[14]:
```
scaler = MinMaxScaler()
macroNutrient_bfill[float_columns] = scaler.
 ↪fit_transform(macroNutrient_bfill[float_columns])

macroNutrient_bfill.describe()
```

[14]:
```
                No.  Energy (Kcal)  Carbohydrate(g)   Protein(g)  \
count   8790.000000    8790.000000      8790.000000  8790.000000
mean   15663.495222       0.744277         0.463082     0.455734
std     9251.413586       0.146574         0.349960     0.237953
min     1001.000000       0.000000         0.000000     0.000000
25%     9086.250000       0.664410         0.010572     0.271107
50%    14427.500000       0.772511         0.506167     0.489117
75%    20142.750000       0.855610         0.775931     0.676456
max    93600.000000       1.000000         1.000000     1.000000

       Total Lipid(g)  Monounsaturated Fatty Acids(g)  \
count     8790.000000                     8790.000000
mean         0.383715                        0.248433
std          0.256536                        0.209670
min          0.000000                        0.000000
25%          0.144705                        0.041073
50%          0.393234                        0.231143
75%          0.582695                        0.401321
max          1.000000                        1.000000

       Polyunsaturated Fatty Acids(g)  Saturated Fatty Acids(g)
count                     8790.000000               8790.000000
mean                         0.175489                  0.227440
std                          0.175237                  0.194752
min                          0.000000                  0.000000
25%                          0.042581                  0.039890
50%                          0.118689                  0.201000
75%                          0.257715                  0.362388
max                          1.000000                  1.000000
```

[15]:
```
X = macroNutrient_bfill[float_columns]
X
```

[15]:

	Energy (Kcal)	Carbohydrate(g)	Protein(g)	Total Lipid(g)	\
0	0.685223	0.421947	0.644650	0.122492	
1	0.770972	0.539328	0.673774	0.444529	
2	0.624253	0.633388	0.203973	0.020652	
3	0.466968	0.380891	0.074901	0.056849	
4	0.513760	0.468498	0.074901	0.056849	
…	…	…	…	…	
8785	0.654500	0.583869	0.396246	0.175712	
8786	0.654500	0.583869	0.396246	0.175712	
8787	0.642026	0.626764	0.305234	0.000000	
8788	0.817624	0.893283	0.324179	0.024556	
8789	0.889955	0.936086	0.535802	0.513582	

	Monounsaturated Fatty Acids(g)	Polyunsaturated Fatty Acids(g)	\
0	0.022900	0.022872	
1	0.297219	0.227549	
2	0.017754	0.019922	
3	0.017754	0.019922	
4	0.017754	0.019922	
…	…	…	
8785	0.066435	0.008176	
8786	0.066435	0.008176	
8787	0.000000	0.000000	
8788	0.002465	0.005934	
8789	0.373303	0.259528	

	Saturated Fatty Acids(g)
0	0.030388
1	0.212894
2	0.003042
3	0.014394
4	0.014394
…	…
8785	0.129330
8786	0.129330
8787	0.000000
8788	0.009211
8789	0.275650

[8790 rows x 7 columns]

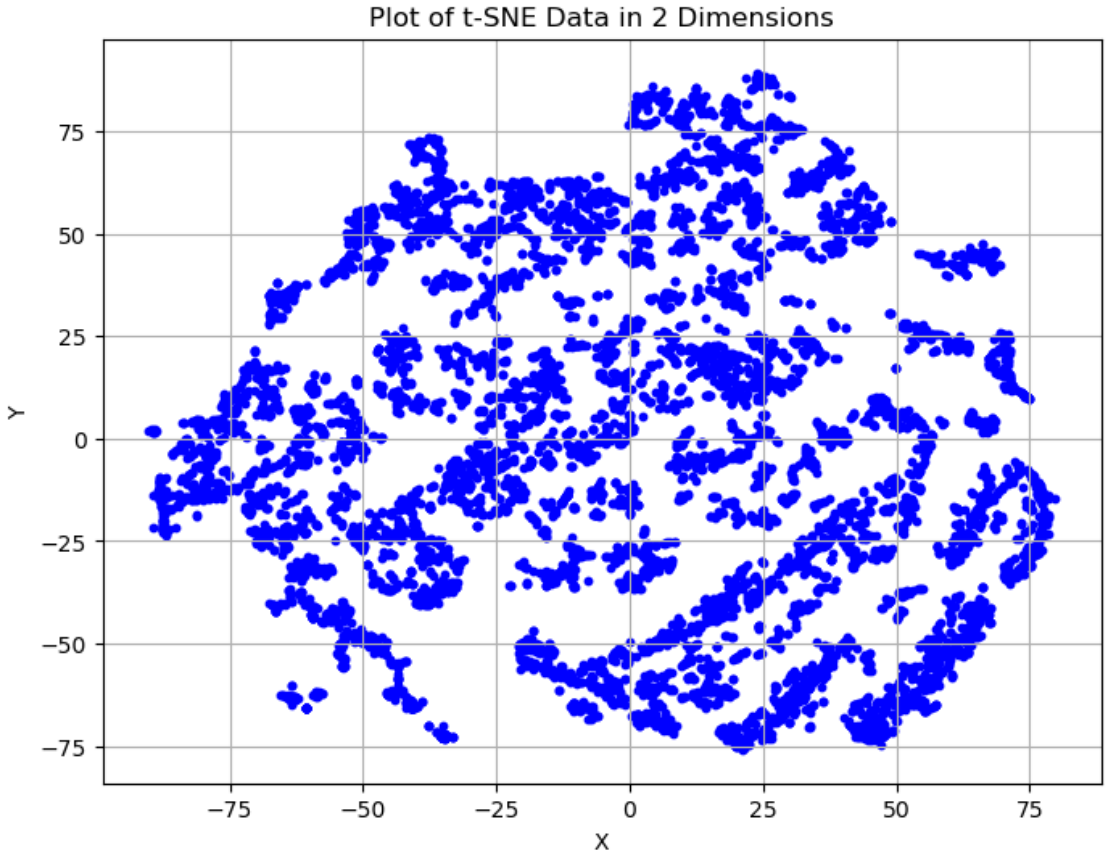
0.1 Dimensionality Reduction

[16]:
```
from clustering_function import dimensionality_reduction,scatter_plot_clustering
```

[17]:
```
# Reduced to 2 dimensions using tsne
```

```
X_tsne_reduced, tsne_x_data, tsne_y_data, reduction_method =␣
↪dimensionality_reduction(X, 't-SNE', n_components=2)
```

[18]:
```
scatter_plot_clustering(tsne_x_data, tsne_y_data, reduction_method)
```



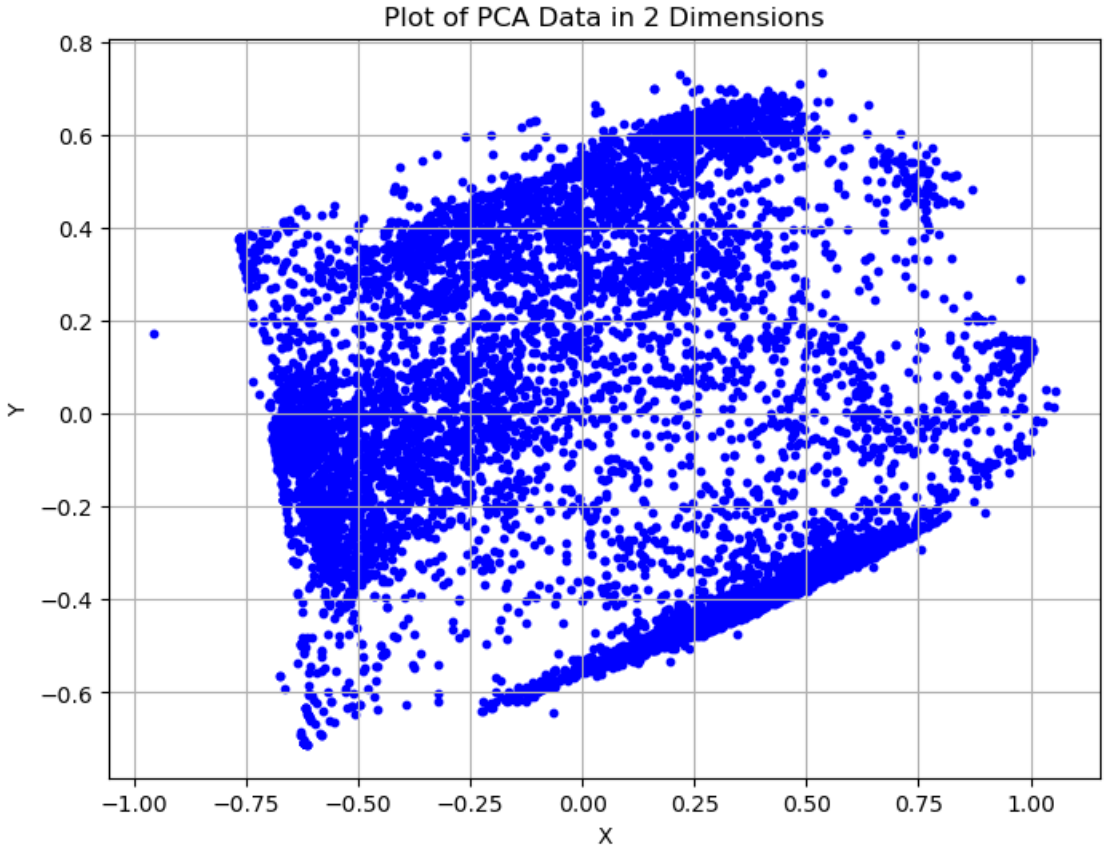
[19]:
```
X_tsne_reduced
```

[19]:
```
array([[-11.827961  , -32.04308   ],
       [ 13.443113  ,  -1.1108218 ],
       [-62.286602  ,  -0.43208718],
       …,
       [-57.835987  ,   2.8110402 ],
       [-52.79809   ,  41.900913  ],
       [  1.3407313 ,  45.89533   ]], dtype=float32)
```

[20]:
```
pd.DataFrame(X_tsne_reduced).to_csv("../Dataset/Dataset_for_reduced_data/
↪bfill_reduced.csv")
```

[21]:
```
# Reduced to 2 dimensions using pca
X_pca_reduced, pca_x_data, pca_y_data, reduction_method =␣
↪dimensionality_reduction(X, 'PCA', n_components=2)
```

[22]:
```
scatter_plot_clustering(pca_x_data, pca_y_data, reduction_method)
```



[23]:
```
X_pca_reduced
```

[23]:
```
array([[-0.29310583, -0.22923045],
       [ 0.11862504,  0.07327452],
       [-0.60098264, -0.02049183],
       …,
       [-0.58501763, -0.04923963],
       [-0.60972057,  0.23254428],
       [ 0.06793785,  0.51259567]])
```

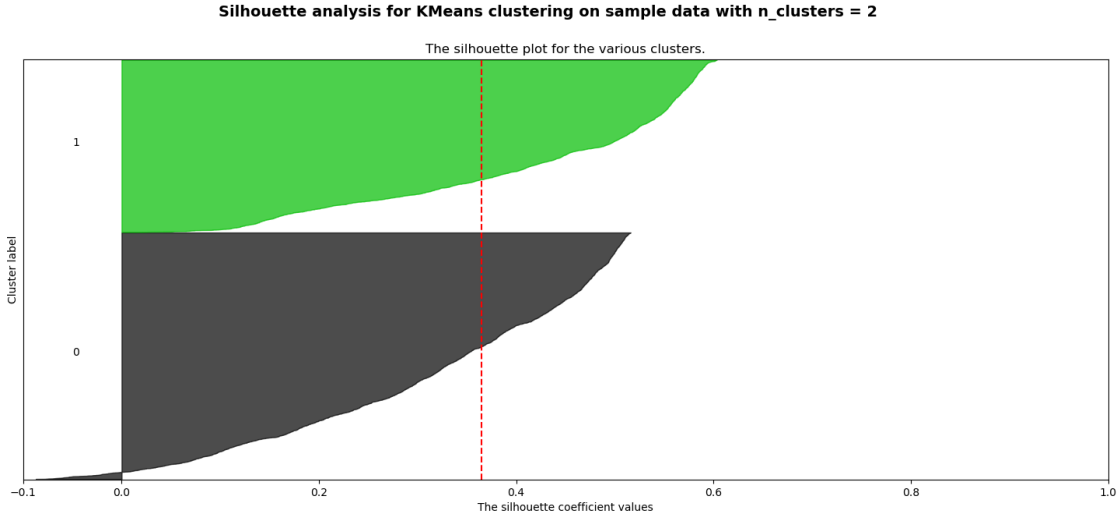
0.2 K Means: Optimal number of clusters

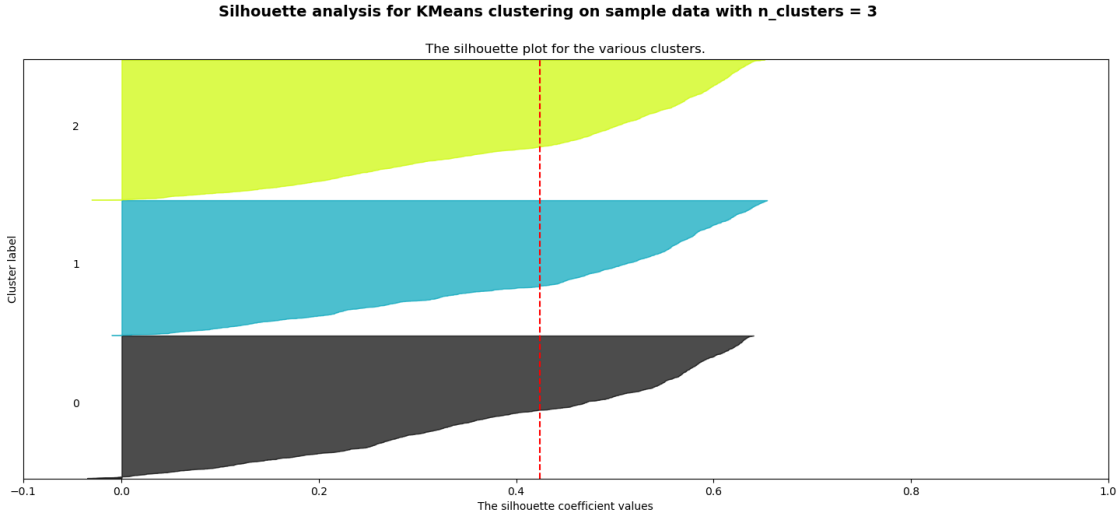
Silhoutte Analysis

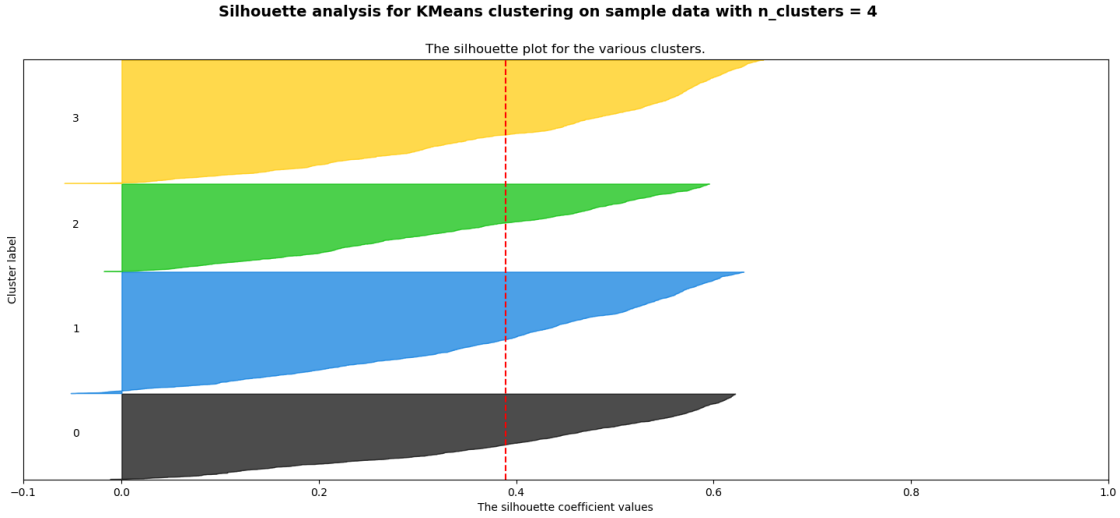
[24]:
```
from clustering_function import plot_silhouette_analysis
```

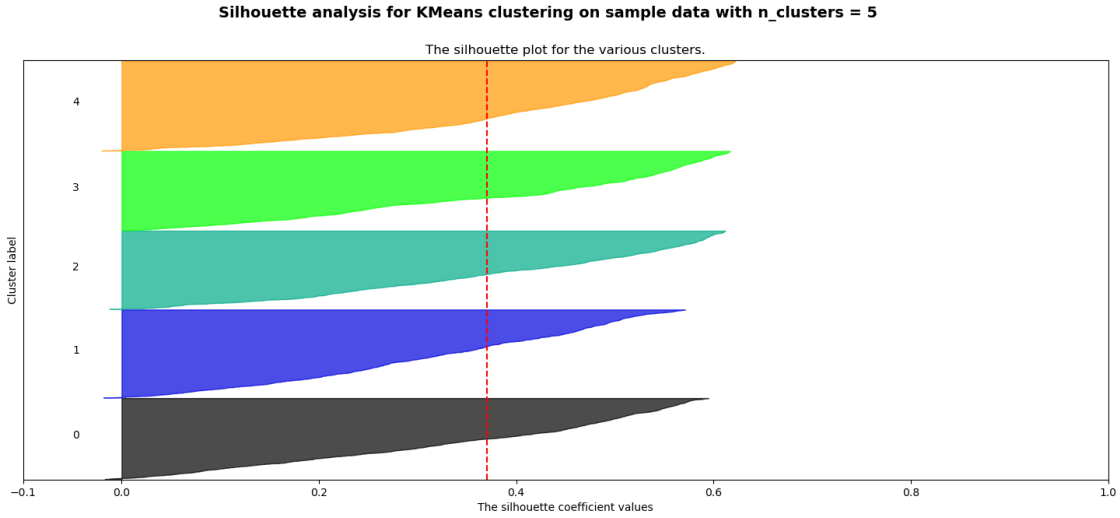
[25]:
```
tsne_list_num_clusters = list(range(2,8))
plot_silhouette_analysis(X_tsne_reduced, tsne_list_num_clusters)
```

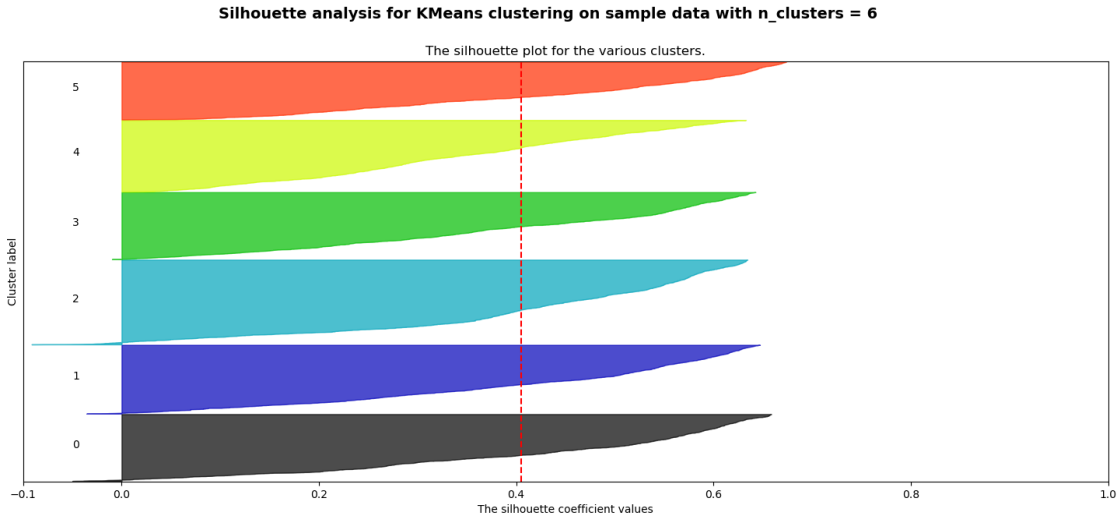
```
For n_clusters = 2 The average silhouette_score is : 0.36472055
For n_clusters = 3 The average silhouette_score is : 0.42388904
For n_clusters = 4 The average silhouette_score is : 0.3892632
For n_clusters = 5 The average silhouette_score is : 0.37000164
For n_clusters = 6 The average silhouette_score is : 0.40517595
For n_clusters = 7 The average silhouette_score is : 0.39805335
```

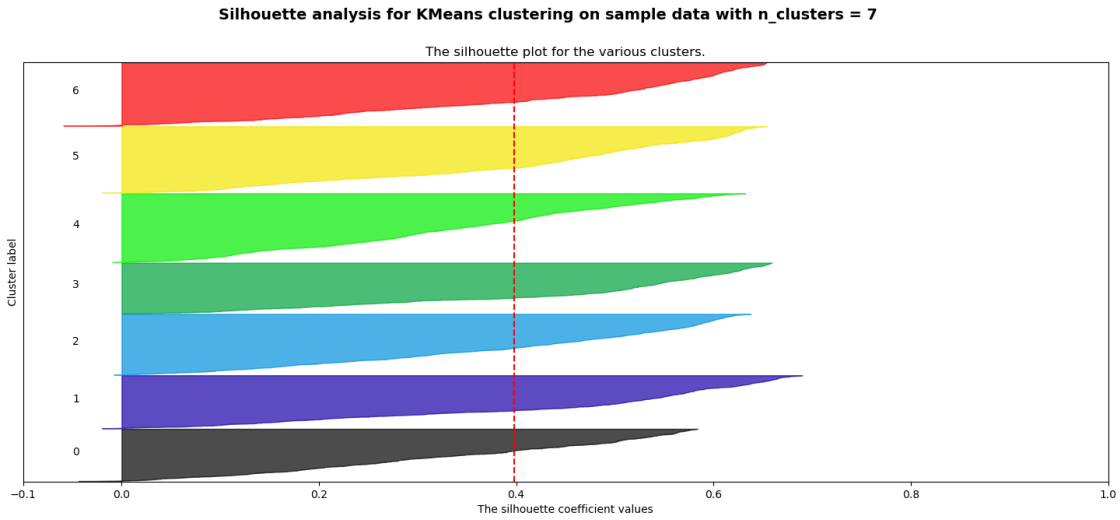





Silhouette analysis for KMeans clustering on sample data with n_clusters = 4
The silhouette plot for the various clusters.
Cluster label
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


Silhouette analysis for KMeans clustering on sample data with n_clusters = 5
The silhouette plot for the various clusters.
Cluster label
4
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values

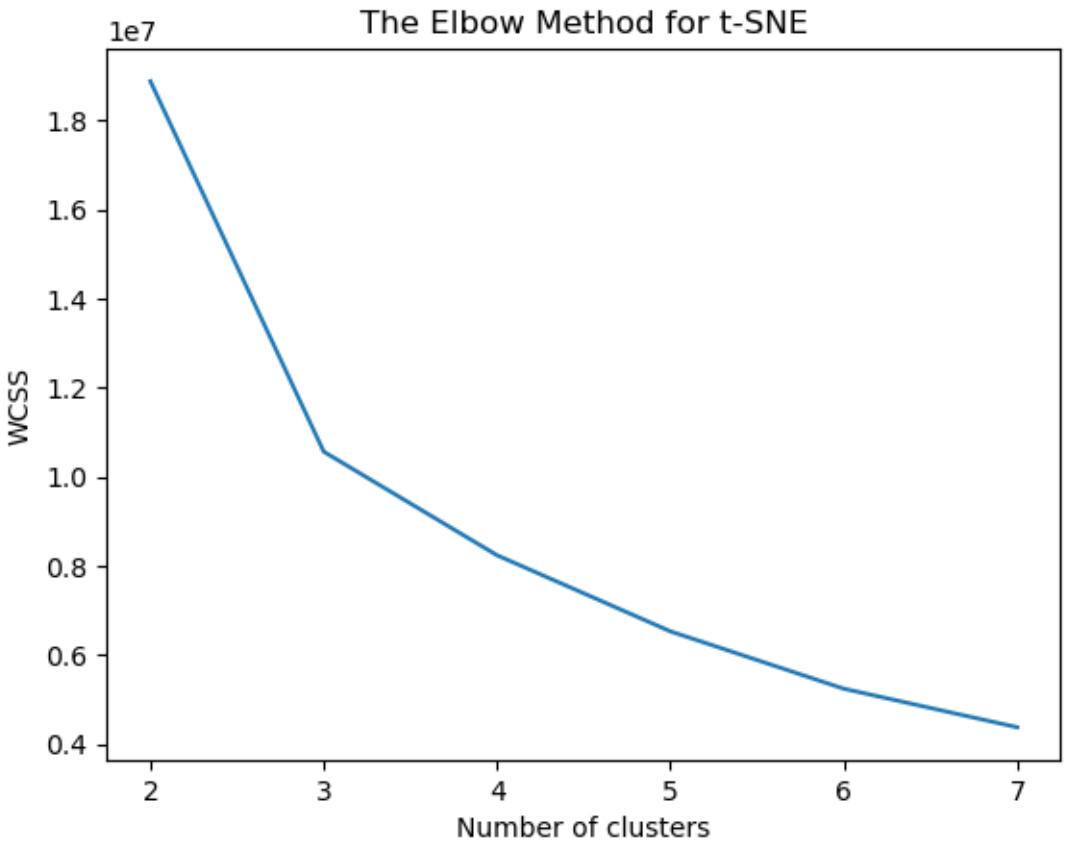




Elbow Method

[26]:
```
from clustering_function import elbow_method_analysis
```

[27]:
```
elbow_method_analysis(X_tsne_reduced)
```



0.3 Clustering

[28]:
```
# Trackers throughout each model
scores = {} # to track the silhoutte score of the tuned model
food_groups = {} # to track the counts of each group
```

[29]:
```
X_with_labels = X.copy()
X_with_labels
```

[29]:

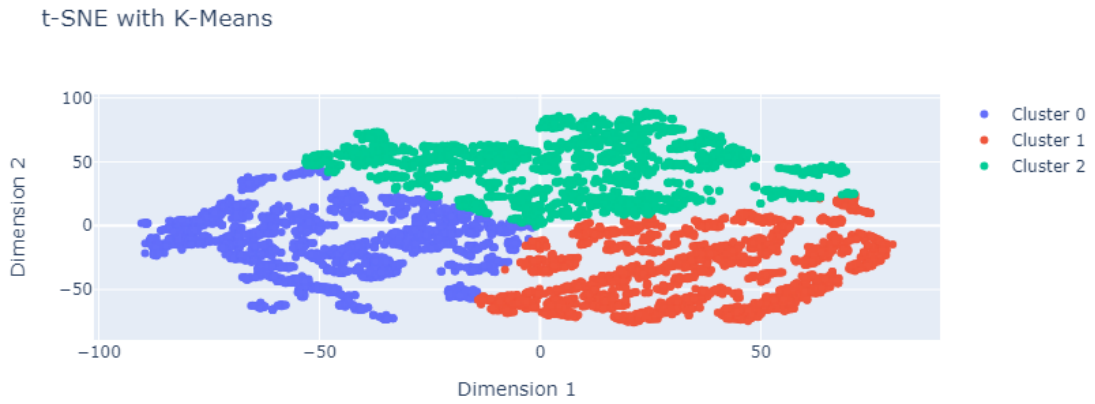
	Energy (Kcal)	Carbohydrate(g)	Protein(g)	Total Lipid(g)	\
0	0.685223	0.421947	0.644650	0.122492	
1	0.770972	0.539328	0.673774	0.444529	
2	0.624253	0.633388	0.203973	0.020652	
3	0.466968	0.380891	0.074901	0.056849	
4	0.513760	0.468498	0.074901	0.056849	
…	…	…	…	…	
8785	0.654500	0.583869	0.396246	0.175712	
8786	0.654500	0.583869	0.396246	0.175712	
8787	0.642026	0.626764	0.305234	0.000000	
8788	0.817624	0.893283	0.324179	0.024556	

```
8789       0.889955          0.936086    0.535802         0.513582

      Monounsaturated Fatty Acids(g)  Polyunsaturated Fatty Acids(g)  \
0                           0.022900                        0.022872
1                           0.297219                        0.227549
2                           0.017754                        0.019922
3                           0.017754                        0.019922
4                           0.017754                        0.019922
…                                …                               …
8785                        0.066435                        0.008176
8786                        0.066435                        0.008176
8787                        0.000000                        0.000000
8788                        0.002465                        0.005934
8789                        0.373303                        0.259528

      Saturated Fatty Acids(g)
0                     0.030388
1                     0.212894
2                     0.003042
3                     0.014394
4                     0.014394
…                          …
8785                  0.129330
8786                  0.129330
8787                  0.000000
8788                  0.009211
8789                  0.275650

[8790 rows x 7 columns]
```

0.4 K Means Clustering

[30]:
```
from clustering_function import perform_kmeans_clustering, plot_clusters
```

[31]:
```
tsne_clust_labels, kmean_model = perform_kmeans_clustering(X_tsne_reduced, 3)
plot_clusters(X_tsne_reduced, tsne_clust_labels, 'K-Means')
```



[32]:
```
# export model
with open('../Model_fitted/Kmeans_model_bfill_pkl', 'wb') as files:
    pickle.dump(kmean_model, files)
```

[33]:
```
X_with_labels['kmeans_tsne'] = tsne_clust_labels
scores['kmeans_tsne'] = (silhouette_score(X_tsne_reduced,␣
 ↪X_with_labels['kmeans_tsne']))

X_with_labels
```

[33]:

	Energy (Kcal)	Carbohydrate(g)	Protein(g)	Total Lipid(g)	\
0	0.685223	0.421947	0.644650	0.122492	
1	0.770972	0.539328	0.673774	0.444529	
2	0.624253	0.633388	0.203973	0.020652	
3	0.466968	0.380891	0.074901	0.056849	
4	0.513760	0.468498	0.074901	0.056849	
…	…	…	…	…	
8785	0.654500	0.583869	0.396246	0.175712	
8786	0.654500	0.583869	0.396246	0.175712	
8787	0.642026	0.626764	0.305234	0.000000	
8788	0.817624	0.893283	0.324179	0.024556	
8789	0.889955	0.936086	0.535802	0.513582	

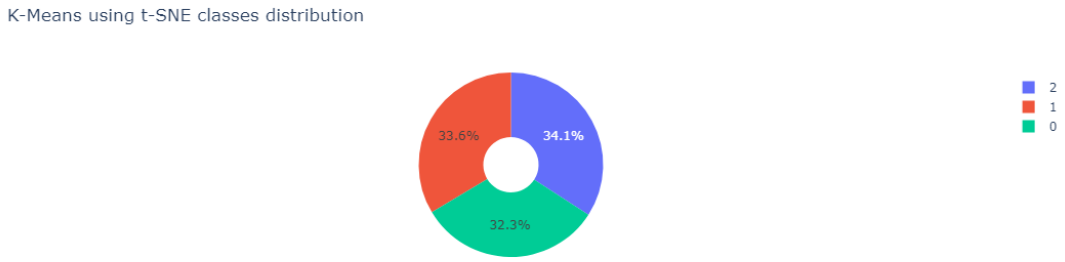
	Monounsaturated Fatty Acids(g)	Polyunsaturated Fatty Acids(g)	\
0	0.022900	0.022872	
1	0.297219	0.227549	
2	0.017754	0.019922	
3	0.017754	0.019922	
4	0.017754	0.019922	
…	…	…	

```
8785                          0.066435                          0.008176
8786                          0.066435                          0.008176
8787                          0.000000                          0.000000
8788                          0.002465                          0.005934
8789                          0.373303                          0.259528

      Saturated Fatty Acids(g)  kmeans_tsne
0                     0.030388            0
1                     0.212894            1
2                     0.003042            0
3                     0.014394            0
4                     0.014394            0
…                          …            …
8785                  0.129330            0
8786                  0.129330            0
8787                  0.000000            0
8788                  0.009211            0
8789                  0.275650            2

[8790 rows x 8 columns]
```

[34]:
```
from clustering_function import get_food_groups, plot_cluster_distribution
```

[35]:
```
food_groups, value_counts = get_food_groups(X_with_labels['kmeans_tsne'],␣
↪'KMeans_tsne',food_groups)
```

[36]:
```
plot_cluster_distribution(value_counts, 'K-Means')
```



[37]:
```
kmeans_tsne = X_with_labels.groupby('kmeans_tsne')
```

[38]:
```
n = kmeans_tsne['kmeans_tsne'].count().count() # number of cluster
for i in range(0,n):
    display(kmeans_tsne.get_group(i).describe())
```

```
       Energy (Kcal)  Carbohydrate(g)   Protein(g)  Total Lipid(g)  \
count    2838.000000      2838.000000  2838.000000     2838.000000
mean        0.590238         0.499393     0.231250        0.100672
std         0.136855         0.207507     0.172197        0.099689
min         0.000000         0.000000     0.000000        0.000000
25%         0.530571         0.391107     0.104626        0.020652
50%         0.606421         0.510831     0.215257        0.065026
75%         0.665998         0.628618     0.319460        0.155541
max         0.881090         1.000000     1.000000        0.533865

       Monounsaturated Fatty Acids(g)  Polyunsaturated Fatty Acids(g)  \
count                     2838.000000                     2838.000000
mean                         0.042634                        0.044501
std                          0.068626                        0.064199
min                          0.000000                        0.000000
25%                          0.001571                        0.006609
50%                          0.008619                        0.021928
75%                          0.060702                        0.055072
max                          0.464472                        0.562505

       Saturated Fatty Acids(g)  kmeans_tsne
count               2838.000000       2838.0
mean                   0.040263          0.0
std                    0.061411          0.0
min                    0.000000          0.0
25%                    0.003042          0.0
50%                    0.012335          0.0
75%                    0.045604          0.0
max                    0.350064          0.0

       Energy (Kcal)  Carbohydrate(g)   Protein(g)  Total Lipid(g)  \
count    2954.000000      2954.000000  2954.000000     2954.000000
mean        0.780218         0.076973     0.662125        0.524394
std         0.076482         0.150940     0.156890        0.179101
min         0.203695         0.000000     0.000000        0.000000
25%         0.734514         0.000000     0.657116        0.401543
50%         0.775541         0.000000     0.697886        0.507837
75%         0.815354         0.066226     0.740156        0.632921
max         1.000000         0.757620     1.000000        1.000000

       Monounsaturated Fatty Acids(g)  Polyunsaturated Fatty Acids(g)  \
count                     2954.000000                     2954.000000
mean                         0.372198                        0.184486
std                          0.168995                        0.144902
min                          0.000000                        0.000000
25%                          0.250068                        0.084296
50%                          0.354566                        0.137161
75%                          0.473042                        0.245454
```

```
max                          1.000000                          0.943431

       Saturated Fatty Acids(g)  kmeans_tsne
count               2954.000000       2954.0
mean                   0.338964          1.0
std                    0.165382          0.0
min                    0.000000          1.0
25%                    0.213472          1.0
50%                    0.311995          1.0
75%                    0.441126          1.0
max                    1.000000          1.0

       Energy (Kcal)  Carbohydrate(g)   Protein(g)  Total Lipid(g)  \
count    2998.000000      2998.000000  2998.000000     2998.000000
mean        0.854682         0.809149     0.464876        0.513038
std         0.059789         0.167215     0.154716        0.199707
min         0.619994         0.000000     0.000000        0.000000
25%         0.823148         0.719626     0.387440        0.365407
50%         0.866902         0.871162     0.478997        0.531641
75%         0.897908         0.928343     0.559377        0.661329
max         0.999674         0.997844     0.964519        1.000000

       Monounsaturated Fatty Acids(g)  Polyunsaturated Fatty Acids(g)  \
count                     2998.000000                     2998.000000
mean                         0.321300                        0.290619
std                          0.189231                        0.189414
min                          0.000000                        0.000000
25%                          0.178973                        0.144298
50%                          0.318302                        0.250632
75%                          0.441564                        0.407909
max                          0.923391                        1.000000

       Saturated Fatty Acids(g)  kmeans_tsne
count               2998.000000       2998.0
mean                   0.294739          2.0
std                    0.175334          0.0
min                    0.000000          2.0
25%                    0.140432          2.0
50%                    0.296644          2.0
75%                    0.419405          2.0
max                    0.903884          2.0
```

0.4.1 Insights

- Cluster 0 shows very low values in total lipids among all clusters, with low values in protein. This cluster is probably related to foods that are generally healthier than those in the other two clusters.
- Cluster 1 indicates very low values for carbohydrates and moderate to high values in protein

and lipid, suggesting a high protein and low carbohydrate food group for this cluster.

- Cluster 2 exhibits highest values in energy and carbohydrates and slightly moderate to low values in protein.

0.5 Agglomerative Clustering

[39]:
```
from clustering_function import tuning_agglomerative, perform_agg_clustering
```

[40]:
```
agg_param_grid = {
      'n_clusters': [2, 3, 4, 5, 6],
      'linkage': ['ward'],
      'affinity' : ['euclidean']
  }
```

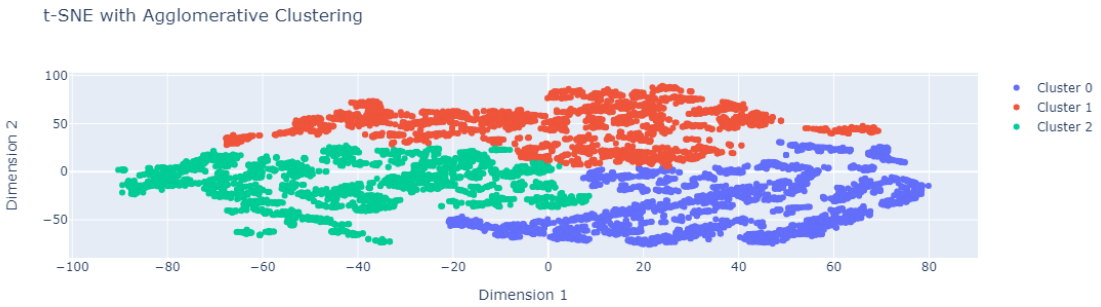
[41]:
```
tuning_agglomerative(X_tsne_reduced, agg_param_grid)
```

```
Best silhouette score: 0.40360528
Best parameters: {'affinity': 'euclidean', 'linkage': 'ward', 'n_clusters': 3}
```

[42]:
```
Agg_tsne_clust_labels,agg_model = perform_agg_clustering(X_tsne_reduced, 3)
```

[43]:
```
# export model
with open('../Model_fitted/Agg_model_bfill_pkl', 'wb') as files:
    pickle.dump(agg_model, files)
```

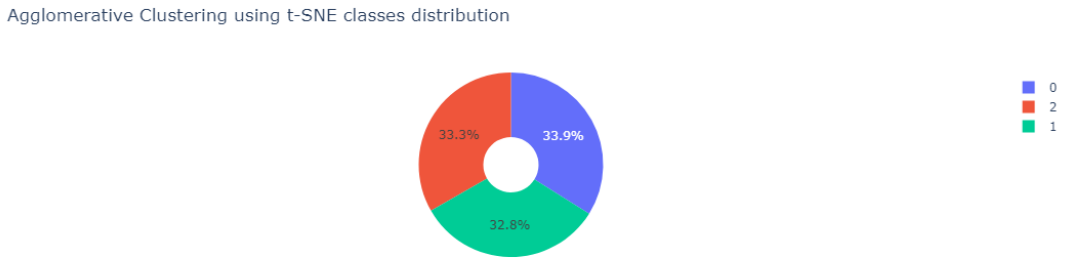
[44]:
```
plot_clusters(X_tsne_reduced, Agg_tsne_clust_labels, "Agglomerative Clustering")
```



[45]:
```
X_with_labels['Agg_tsne'] = Agg_tsne_clust_labels
scores['Agg_tsne'] = (silhouette_score(X_tsne_reduced,␣
 ↪X_with_labels['Agg_tsne']))
```

[46]:
```
food_groups, value_counts = get_food_groups(X_with_labels['Agg_tsne'],␣
 ↪'Agglomerative_tsne',food_groups)
```

[47]:
```
plot_cluster_distribution(value_counts, 'Agglomerative Clustering')
```



[48]:
```
Aggtsne = X_with_labels.groupby('Agg_tsne')
```

[49]:
```
n = Aggtsne['Agg_tsne'].count().count() # number of cluster
for i in range(0,n):
    display(Aggtsne.get_group(i).describe())
```

	Energy (Kcal)	Carbohydrate(g)	Protein(g)	Total Lipid(g)	\
count	2978.000000	2978.000000	2978.000000	2978.000000	
mean	0.787592	0.070352	0.649686	0.536957	
std	0.074686	0.151909	0.179226	0.189934	
min	0.530571	0.000000	0.000000	0.020652	
25%	0.738185	0.000000	0.653097	0.411580	
50%	0.778509	0.000000	0.697595	0.522022	
75%	0.822469	0.028391	0.739775	0.647590	
max	1.000000	0.917097	1.000000	1.000000	

	Monounsaturated Fatty Acids(g)	Polyunsaturated Fatty Acids(g)	\
count	2978.000000	2978.000000	
mean	0.378068	0.199322	
std	0.180452	0.173111	
min	0.000000	0.000000	
25%	0.252981	0.082402	
50%	0.364386	0.139512	
75%	0.486983	0.259509	
max	1.000000	1.000000	

	Saturated Fatty Acids(g)	kmeans_tsne	Agg_tsne
count	2978.000000	2978.000000	2978.0
mean	0.340571	1.005709	0.0
std	0.171659	0.255870	0.0
min	0.003042	0.000000	0.0
25%	0.216075	1.000000	0.0

```
50%                  0.318520     1.000000       0.0
75%                  0.452404     1.000000       0.0
max                  1.000000     2.000000       0.0

       Energy (Kcal)  Carbohydrate(g)   Protein(g)  Total Lipid(g)  \
count    2885.000000      2885.000000  2885.000000     2885.000000
mean        0.859816         0.844905     0.467051        0.486045
std         0.048214         0.120934     0.159041        0.212871
min         0.656199         0.189696     0.000000        0.000000
25%         0.827949         0.770369     0.391311        0.339886
50%         0.867705         0.890634     0.481570        0.523478
75%         0.895691         0.934641     0.562600        0.648896
max         0.965905         1.000000     0.964519        0.939034

       Monounsaturated Fatty Acids(g)  Polyunsaturated Fatty Acids(g)  \
count                     2885.000000                     2885.000000
mean                         0.300224                        0.269855
std                          0.190089                        0.179451
min                          0.000000                        0.000000
25%                          0.144595                        0.131079
50%                          0.306596                        0.241218
75%                          0.431838                        0.384699
max                          0.877617                        0.895755

       Saturated Fatty Acids(g)  kmeans_tsne  Agg_tsne
count               2885.000000  2885.000000    2885.0
mean                   0.280631     1.902253       1.0
std                    0.181432     0.429671       0.0
min                    0.000000     0.000000       1.0
25%                    0.117891     2.000000       1.0
50%                    0.288018     2.000000       1.0
75%                    0.412873     2.000000       1.0
max                    0.903884     2.000000       1.0

       Energy (Kcal)  Carbohydrate(g)   Protein(g)  Total Lipid(g)  \
count    2927.000000      2927.000000  2927.000000     2927.000000
mean        0.586326         0.486310     0.247247        0.126941
std         0.128596         0.169471     0.174845        0.127010
min         0.000000         0.000000     0.000000        0.000000
25%         0.530571         0.387875     0.118121        0.024556
50%         0.606421         0.503213     0.230785        0.074449
75%         0.672186         0.610688     0.328292        0.210347
max         0.938210         1.000000     1.000000        0.894264

       Monounsaturated Fatty Acids(g)  Polyunsaturated Fatty Acids(g)  \
count                     2927.000000                     2927.000000
mean                         0.065491                        0.058227
std                          0.100382                        0.083044
min                          0.000000                        0.000000
```

```
25%                          0.002242                          0.008176
50%                          0.012275                          0.027434
75%                          0.101401                          0.075460
max                          0.959965                          0.887772
```

```
       Saturated Fatty Acids(g)  kmeans_tsne  Agg_tsne
count               2927.000000  2927.000000    2927.0
mean                   0.059909     0.159549       2.0
std                    0.087235     0.486485       0.0
min                    0.000000     0.000000       2.0
25%                    0.003903     0.000000       2.0
50%                    0.016027     0.000000       2.0
75%                    0.087542     0.000000       2.0
max                    0.808184     2.000000       2.0
```

0.5.1 Insights

- Cluster 0 contains the lowest values for carbohydrate among all three clusters.
- Cluster 1 is characterized by the high values for carbohydrate, energy and protein.
- Cluster 2 contains moderate values of energy, carbohydrate, and is quite low in protein and fats.

0.6 GMM

[50]:
```
from clustering_function import gmm_bic_score, perform_gmm_clustering
from sklearn.mixture import GaussianMixture
from sklearn.model_selection import GridSearchCV
```

[51]:
```
gmm_param_grid = {
    "n_components": range(1, 6),
    "covariance_type": ["spherical", "tied", "diag", "full"],
}
grid_search = GridSearchCV(
    GaussianMixture(), param_grid=gmm_param_grid, scoring=gmm_bic_score
)
```

[52]:
```
grid_search.fit(X_tsne_reduced)
Gmm_tsne_results = grid_search.cv_results_
```

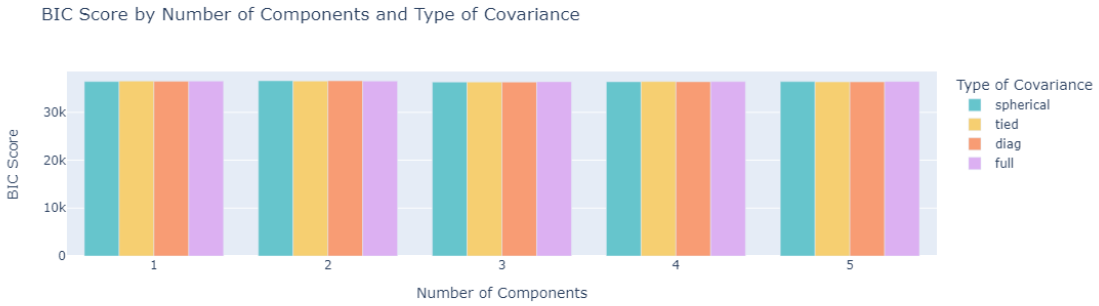
[53]:
```
Gmm_tsne_df = pd.DataFrame(Gmm_tsne_results)[
    ["param_n_components", "param_covariance_type", "mean_test_score"]
]
Gmm_tsne_df["mean_test_score"] = -Gmm_tsne_df["mean_test_score"]
Gmm_tsne_df = Gmm_tsne_df.rename(
    columns={
        "param_n_components": "Number of Components",
        "param_covariance_type": "Type of Covariance",
        "mean_test_score": "BIC Score",
```

```
    }
)
Gmm_tsne_df.sort_values(by="BIC Score").head()
```

[53]:

	Number of Components	Type of Covariance	BIC Score
12	3	diag	36316.739564
2	3	spherical	36323.584894
7	3	tied	36337.504887
9	5	tied	36361.144114
14	5	diag	36361.202482

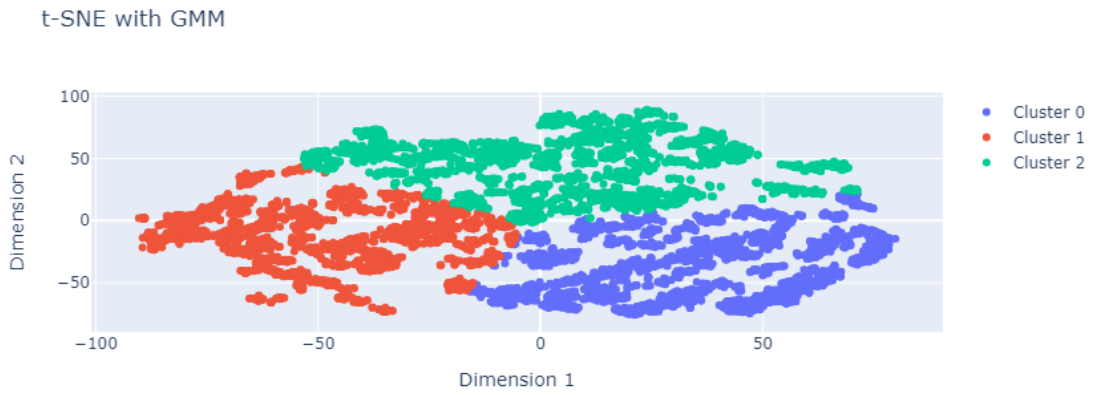
[54]:
```
fig = px.bar(Gmm_tsne_df, x="Number of Components", y="BIC Score",
                 color="Type of Covariance", barmode="group",
                 title="BIC Score by Number of Components and Type of␣
 ↪Covariance",
                 color_discrete_sequence=px.colors.qualitative.Pastel)
fig.show()
```



[55]:
```
Gmm_tsne_labels,gmm_model = perform_gmm_clustering(X_tsne_reduced, 3,␣
 ↪'spherical')
```

[56]:
```
# export model
with open('../Model_fitted/Gmm_model_bfill_pkl', 'wb') as files:
    pickle.dump(gmm_model, files)
```

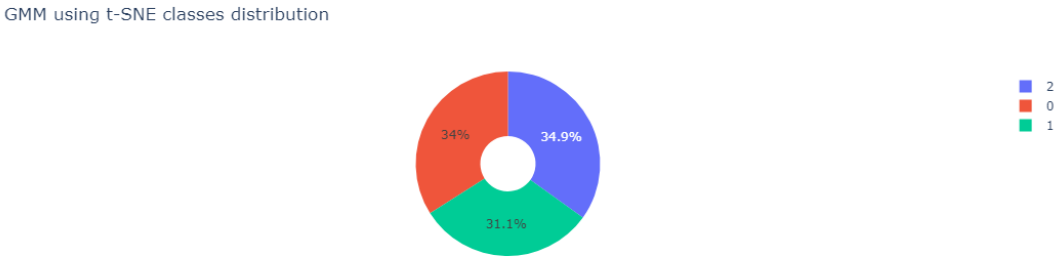
[57]:
```
plot_clusters(X_tsne_reduced, Gmm_tsne_labels, "GMM")
```



[58]:
```
X_with_labels['Gmm_tsne'] = Gmm_tsne_labels
scores['Gmm_tsne'] = (silhouette_score(X_tsne_reduced,␣
 ↪X_with_labels['Gmm_tsne']))
```

[59]:
```
food_groups, value_counts = get_food_groups(X_with_labels['Gmm_tsne'],␣
 ↪'GMM_tsne',food_groups)
```

[60]:
```
plot_cluster_distribution(value_counts, 'GMM')
```



[61]:
```
GmmTsne = X_with_labels.groupby('Gmm_tsne')
```

[62]:
```
n = GmmTsne['Gmm_tsne'].count().count() # number of cluster
for i in range(0,n):
    display(GmmTsne.get_group(i).describe())
```

	Energy (Kcal)	Carbohydrate(g)	Protein(g)	Total Lipid(g)	\
count	2988.000000	2988.000000	2988.000000	2988.000000	
mean	0.777702	0.079502	0.661774	0.517353	

```
std         0.078108         0.153540     0.158991         0.182426
min         0.203695         0.000000     0.000000         0.000000
25%         0.733270         0.000000     0.656970         0.395604
50%         0.774034         0.000000     0.697837         0.504484
75%         0.814782         0.072906     0.740096         0.628677
max         1.000000         0.757620     1.000000         1.000000

       Monounsaturated Fatty Acids(g)  Polyunsaturated Fatty Acids(g)  \
count                     2988.000000                     2988.000000
mean                         0.367588                        0.181972
std                          0.169846                        0.139884
min                          0.000000                        0.000000
25%                          0.246721                        0.083653
50%                          0.351866                        0.136522
75%                          0.470545                        0.244292
max                          1.000000                        0.900156

       Saturated Fatty Acids(g)  kmeans_tsne     Agg_tsne  Gmm_tsne
count               2988.000000  2988.000000  2988.000000    2988.0
mean                   0.333914     0.984940     0.130522       0.0
std                    0.167231     0.121813     0.492697       0.0
min                    0.000000     0.000000     0.000000       0.0
25%                    0.209204     1.000000     0.000000       0.0
50%                    0.309270     1.000000     0.000000       0.0
75%                    0.436556     1.000000     0.000000       0.0
max                    1.000000     1.000000     2.000000       0.0

       Energy (Kcal)  Carbohydrate(g)   Protein(g)  Total Lipid(g)  \
count    2735.000000      2735.000000  2735.000000     2735.000000
mean        0.586491         0.500337     0.224780        0.096117
std         0.136506         0.204034     0.164320        0.095514
min         0.000000         0.000000     0.000000        0.000000
25%         0.526545         0.391463     0.101826        0.020652
50%         0.601603         0.511548     0.212703        0.061792
75%         0.661180         0.628737     0.314366        0.146916
max         0.881090         1.000000     0.993116        0.533865

       Monounsaturated Fatty Acids(g)  Polyunsaturated Fatty Acids(g)  \
count                     2735.000000                     2735.000000
mean                         0.039133                        0.040522
std                          0.064905                        0.056598
min                          0.000000                        0.000000
25%                          0.001571                        0.006159
50%                          0.007967                        0.020769
75%                          0.055083                        0.050658
max                          0.464472                        0.429054

       Saturated Fatty Acids(g)  kmeans_tsne     Agg_tsne  Gmm_tsne
```

count	2735.000000	2735.0	2735.000000	2735.0
mean	0.038344	0.0	1.903473	1.0
std	0.060078	0.0	0.372086	0.0
min	0.000000	0.0	0.000000	1.0
25%	0.003042	0.0	2.000000	1.0
50%	0.011299	0.0	2.000000	1.0
75%	0.040981	0.0	2.000000	1.0
max	0.350064	0.0	2.000000	1.0

	Energy (Kcal)	Carbohydrate(g)	Protein(g)	Total Lipid(g)	\
count	3067.000000	3067.000000	3067.000000	3067.000000	
mean	0.852419	0.803558	0.460954	0.509986	
std	0.063480	0.174356	0.156926	0.202382	
min	0.619994	0.000000	0.000000	0.000000	
25%	0.820963	0.711312	0.383301	0.361357	
50%	0.866499	0.867832	0.476393	0.527693	
75%	0.897009	0.928164	0.558110	0.660148	
max	0.999674	0.997844	0.964519	1.000000	

	Monounsaturated Fatty Acids(g)	Polyunsaturated Fatty Acids(g)	\
count	3067.000000	3067.000000	
mean	0.318991	0.289529	
std	0.190436	0.191854	
min	0.000000	0.000000	
25%	0.173487	0.141463	
50%	0.313479	0.248512	
75%	0.439620	0.406529	
max	0.923391	1.000000	

	Saturated Fatty Acids(g)	kmeans_tsne	Agg_tsne	Gmm_tsne
count	3067.000000	3067.000000	3067.000000	3067.0
mean	0.292334	1.958591	1.024780	2.0
std	0.176360	0.278462	0.317009	0.0
min	0.000000	0.000000	0.000000	2.0
25%	0.139277	2.000000	1.000000	2.0
50%	0.293062	2.000000	1.000000	2.0
75%	0.418970	2.000000	1.000000	2.0
max	0.903884	2.000000	2.000000	2.0

0.6.1 Insights

- Cluster 0 is a low-carb, high-protein food group with minimal carbohydrate values and slightly higher than moderate calorie and protein values.
- Cluster 1 represents a lower fat food group with moderate values for energy and carbohydrates, low values for protein, and total lipids.
- Cluster 2 is the highest-scoring of the three clusters in terms of energy and carbohydrates, indicating a high-carbohydrate food group.

0.7 K-Medoids

[63]:
```
from clustering_function import tuning_kmedoids, perform_kmd_clustering
```

[64]:
```
kmd_param_grid = {
    'n_clusters': [3, 4],
    'method': ['alternate', 'pam'],
    'init' : ['random', 'heuristic', 'k-medoids++', 'build']
}
```

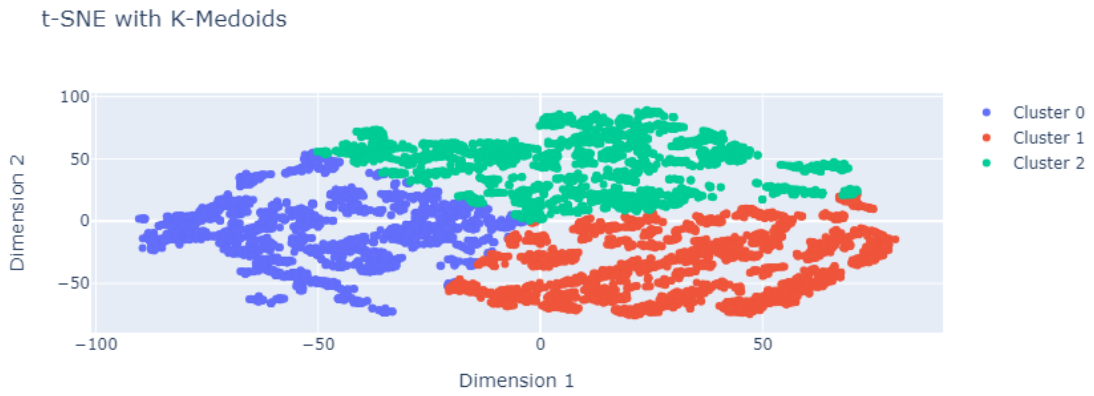
[65]:
```
tuning_kmedoids(X_tsne_reduced, kmd_param_grid)
```

```
Best silhouette score: 0.42373863
Best parameters: {'init': 'build', 'method': 'alternate', 'n_clusters': 3}
```

[66]:
```
Kmd_tsne_labels,kmd_model = perform_kmd_clustering(X_tsne_reduced, 3,␣
 ↪'heuristic', 'alternate')
```

[67]:
```
# export model
with open('../Model_fitted/Kmd_model_bfill_pkl', 'wb') as files:
    pickle.dump(kmd_model, files)
```

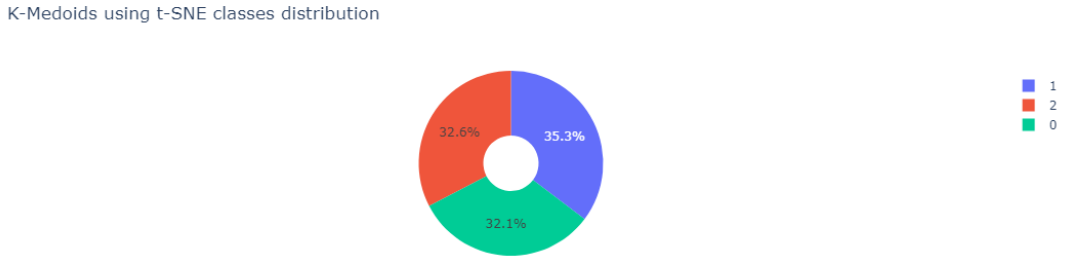
[68]:
```
plot_clusters(X_tsne_reduced, Kmd_tsne_labels, "K-Medoids")
```



[69]:
```
X_with_labels['Kmd_tsne'] = Kmd_tsne_labels
scores['Kmd_tsne'] = (silhouette_score(X_tsne_reduced,␣
 ↪X_with_labels['Kmd_tsne']))
```

[70]:
```
food_groups, value_counts = get_food_groups(X_with_labels['Kmd_tsne'],␣
 ↪'KMedoids_tsne',food_groups)
```

[71]:
```
plot_cluster_distribution(value_counts, 'K-Medoids')
```



[72]:
```
Kmdtsne = X_with_labels.groupby('Kmd_tsne')
```

[73]:
```
n = Kmdtsne['Kmd_tsne'].count().count() # number of cluster
for i in range(0,n):
    display(Kmdtsne.get_group(i).describe())
```

	Energy (Kcal)	Carbohydrate(g)	Protein(g)	Total Lipid(g)	\
count	2820.000000	2820.000000	2820.000000	2820.000000	
mean	0.597638	0.532936	0.225523	0.101138	
std	0.144407	0.206038	0.152106	0.100246	
min	0.000000	0.000000	0.000000	0.000000	
25%	0.530571	0.411499	0.104626	0.020652	
50%	0.606421	0.534785	0.220279	0.065026	
75%	0.673693	0.651590	0.323139	0.154481	
max	0.881090	1.000000	0.716366	0.533865	

	Monounsaturated Fatty Acids(g)	Polyunsaturated Fatty Acids(g)	\
count	2820.000000	2820.000000	
mean	0.038945	0.040564	
std	0.057583	0.053589	
min	0.000000	0.000000	
25%	0.001571	0.006609	
50%	0.008402	0.021823	
75%	0.058584	0.052231	
max	0.351014	0.364110	

	Saturated Fatty Acids(g)	kmeans_tsne	Agg_tsne	Gmm_tsne	\
count	2820.000000	2820.000000	2820.000000	2820.000000	
mean	0.041236	0.110638	1.901418	1.074113	
std	0.062809	0.457285	0.313238	0.262002	
min	0.000000	0.000000	0.000000	1.000000	
25%	0.003042	0.000000	2.000000	1.000000	
50%	0.012542	0.000000	2.000000	1.000000	
75%	0.046711	0.000000	2.000000	1.000000	

```
max                  0.350064     2.000000     2.000000     2.000000
```

```
       Kmd_tsne
count    2820.0
mean        0.0
std         0.0
min         0.0
25%         0.0
50%         0.0
75%         0.0
max         0.0
```

```
       Energy (Kcal)  Carbohydrate(g)   Protein(g)  Total Lipid(g)  \
count    3101.000000      3101.000000  3101.000000     3101.000000
mean        0.771667         0.087196     0.660734        0.500109
std         0.079178         0.161903     0.156963        0.189585
min         0.203695         0.000000     0.000000        0.000000
25%         0.728185         0.000000     0.654631        0.378260
50%         0.770972         0.000000     0.697401        0.496087
75%         0.811883         0.100481     0.739213        0.620550
max         1.000000         0.757620     1.000000        1.000000
```

```
       Monounsaturated Fatty Acids(g)  Polyunsaturated Fatty Acids(g)  \
count                      3101.000000                      3101.000000
mean                          0.354960                         0.175927
std                           0.172767                         0.130497
min                           0.000000                         0.000000
25%                           0.235699                         0.081062
50%                           0.343981                         0.133298
75%                           0.463553                         0.241218
max                           1.000000                         0.887772
```

```
       Saturated Fatty Acids(g)  kmeans_tsne     Agg_tsne     Gmm_tsne  \
count               3101.000000  3101.000000  3101.000000  3101.000000
mean                   0.320417     0.944856     0.171558     0.043857
std                    0.172492     0.229706     0.559012     0.214051
min                    0.000000     0.000000     0.000000     0.000000
25%                    0.196145     1.000000     0.000000     0.000000
50%                    0.299612     1.000000     0.000000     0.000000
75%                    0.426652     1.000000     0.000000     0.000000
max                    1.000000     2.000000     2.000000     2.000000
```

```
       Kmd_tsne
count    3101.0
mean        1.0
std         0.0
min         1.0
25%         1.0
```

```
50%         1.0
75%         1.0
max         1.0
```

```
       Energy (Kcal)  Carbohydrate(g)   Protein(g)  Total Lipid(g)  \
count    2869.000000      2869.000000  2869.000000     2869.000000
mean        0.858806         0.800702     0.460435        0.535660
std         0.059058         0.183475     0.163015        0.191553
min         0.638258         0.000000     0.000000        0.000000
25%         0.825301         0.715658     0.379489        0.392173
50%         0.869298         0.867516     0.478738        0.547273
75%         0.901530         0.926664     0.561710        0.672699
max         1.000000         0.997844     0.964519        1.000000

       Monounsaturated Fatty Acids(g)  Polyunsaturated Fatty Acids(g)  \
count                     2869.000000                     2869.000000
mean                         0.339202                        0.307636
std                          0.188510                        0.193664
min                          0.000000                        0.000000
25%                          0.206419                        0.160237
50%                          0.332539                        0.268889
75%                          0.455408                        0.421787
max                          0.923391                        1.000000

       Saturated Fatty Acids(g)  kmeans_tsne     Agg_tsne     Gmm_tsne  \
count               2869.000000  2869.000000  2869.000000  2869.000000
mean                   0.309967     1.989543     0.991635     1.988149
std                    0.173036     0.108377     0.289157     0.153523
min                    0.000000     0.000000     0.000000     0.000000
25%                    0.160970     2.000000     1.000000     2.000000
50%                    0.309988     2.000000     1.000000     2.000000
75%                    0.430355     2.000000     1.000000     2.000000
max                    0.903884     2.000000     2.000000     2.000000

       Kmd_tsne
count    2869.0
mean        2.0
std         0.0
min         2.0
25%         2.0
50%         2.0
75%         2.0
max         2.0
```

0.7.1 Insights

- Cluster 0 contains the lowest values in terms of the fatty acids compared to the other two clusters.

- Cluster 1 shows a moderate value in energy and high value in protein, but slightly low value in carbohydrates on average.
- Cluster 2 contains the highest values in energy and carbohydrates compared to the other two clusters.

[74]:
```
scores
```

[74]:
```
{'kmeans_tsne': 0.42386162,
 'Agg_tsne': 0.40360528,
 'Gmm_tsne': 0.42308033,
 'Kmd_tsne': 0.4204784}
```

The K-Means model was shown to be the most effective one when the dataset's missing values were filled in using the backward filling approach.