clustering_by_silhouette

created by Etzion Harari

etzionhar@gmail.com (mailto:etzionhar@gmail.com)

Introduction

import & define essential modules + variables

In [1]:

```
from clustering_by_silhouette import silhouette_clustering
from scluster import SCluster
import pandas as pd
from create_3d_gif import pd_to_gif
from sklearn.decomposition import PCA

MPL_Colors=['blue', 'green','red', 'gold', 'purple', 'lime', 'tomato', 'navy', 'teal', 'mar
            'olive', 'orange', 'sienna', 'indigo','yellow', 'darkgreen', 'darkblue', 'choco
            'black']

def pca(df,dim):
    return PCA(n_components=dim).fit(df.T).components_
```

import data

In [2]:

```
df1 = pd.read_csv(r'Example\data_1.csv')
df2 = pd.read_csv(r'Example\data_2.csv')
fields1 = ['a','b','c','d','e','f','g','h','i','j','k','l','m','n']
fields2 = ['f' + str(i) for i in range(1, 15)] # fields names
```

Analysis

run silhouette_clustering for hdbscan & kmeans

In [3]:

```
m_label = SCluster(typ='meanshift', org=10, lim=12).fit(df1[fields1]).labels_ # MeanShift
h_label = SCluster(typ='hdbscan').fit(df2[fields2]).labels_                    # HDBSCAN
k_label = SCluster(org=3 ,lim=20).fit(df2[fields2]).labels_                    # Kmeans (def
```

```
cluster kind: meanshift, input value = 10, silhouette = 0.38
cluster kind: meanshift, input value = 11, silhouette = 0.28
cluster kind: meanshift, input value = 12, silhouette = 0.34
cluster kind: hdbscan, input value = 2, silhouette = 0.08
cluster kind: hdbscan, input value = 3, silhouette = -0.22
cluster kind: hdbscan, input value = 4, silhouette = 0.18
cluster kind: hdbscan, input value = 5, silhouette = -0.11
cluster kind: hdbscan, input value = 6, silhouette = 0.03
cluster kind: hdbscan, input value = 7, silhouette = 0.05
cluster kind: hdbscan, input value = 8, silhouette = 0.05
cluster kind: hdbscan, input value = 9, silhouette = 0.06
cluster kind: hdbscan, input value = 10, silhouette = 0.31
cluster kind: hdbscan, input value = 11, silhouette = 0.31
cluster kind: hdbscan, input value = 12, silhouette = 0.31
cluster kind: hdbscan, input value = 13, silhouette = 0.31
cluster kind: hdbscan, input value = 14, silhouette = 0.3
cluster kind: hdbscan, input value = 15, silhouette = 0.3
cluster kind: hdbscan, input value = 16, silhouette = 0.31
cluster kind: hdbscan, input value = 17, silhouette = 0.31
cluster kind: hdbscan, input value = 18, silhouette = 0.3
cluster kind: hdbscan, input value = 19, silhouette = 0.47
cluster kind: hdbscan, input value = 20, silhouette = 0.47
cluster kind: kmeans, input value = 3, silhouette = 0.37
cluster kind: kmeans, input value = 4, silhouette = 0.45
cluster kind: kmeans, input value = 5, silhouette = 0.46
cluster kind: kmeans, input value = 6, silhouette = 0.46
cluster kind: kmeans, input value = 7, silhouette = 0.36
cluster kind: kmeans, input value = 8, silhouette = 0.36
cluster kind: kmeans, input value = 9, silhouette = 0.36
cluster kind: kmeans, input value = 10, silhouette = 0.36
cluster kind: kmeans, input value = 11, silhouette = 0.32
cluster kind: kmeans, input value = 12, silhouette = 0.31
cluster kind: kmeans, input value = 13, silhouette = 0.31
cluster kind: kmeans, input value = 14, silhouette = 0.29
cluster kind: kmeans, input value = 15, silhouette = 0.27
cluster kind: kmeans, input value = 16, silhouette = 0.27
cluster kind: kmeans, input value = 17, silhouette = 0.27
cluster kind: kmeans, input value = 18, silhouette = 0.26
cluster kind: kmeans, input value = 19, silhouette = 0.27
cluster kind: kmeans, input value = 20, silhouette = 0.26
```

In [4]:

```
print(f'meanshift number of clusters: {len(set(m_label))}\nkmeans    number of clusters: {l
```

```
meanshift number of clusters: 16
kmeans    number of clusters: 5
hdbscan   number of clusters: 4
```

Plot Results

Arrange Data

In [5]:

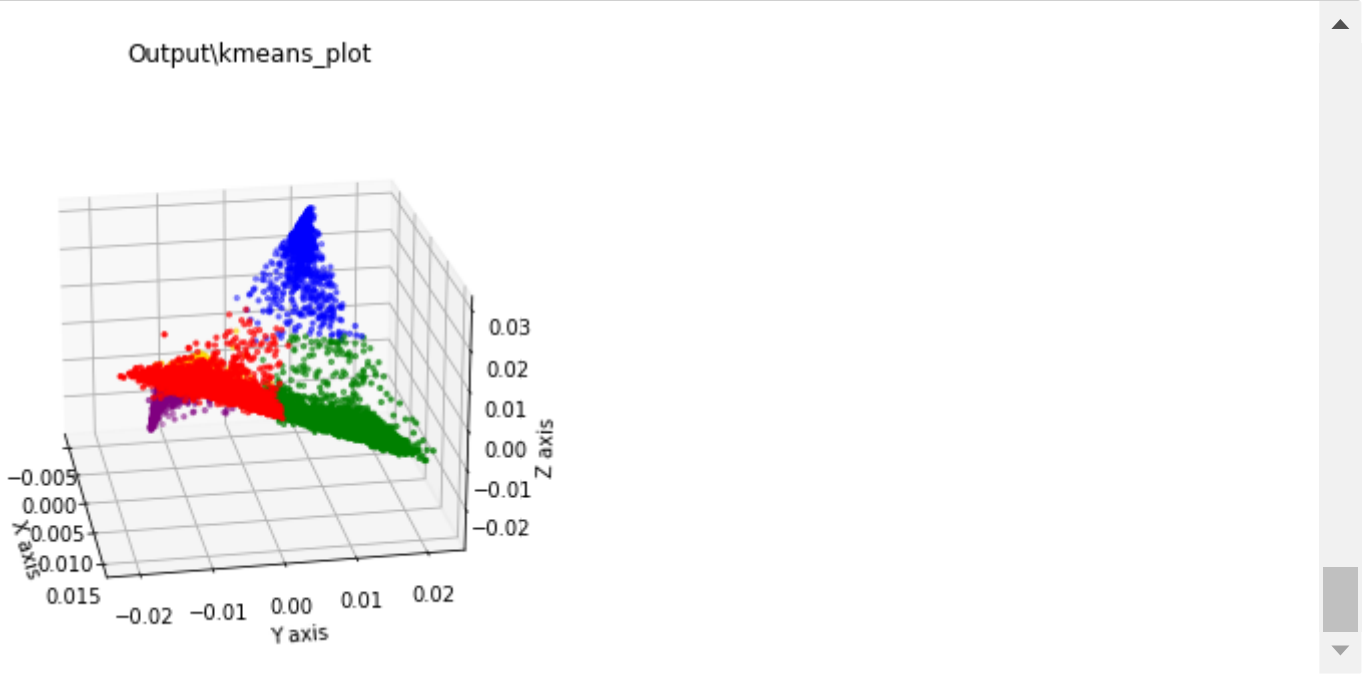
```
df1['x'], df1['y'], df1['z'] = pca(df1[fields1],3)
df2['x'], df2['y'], df2['z'] = pca(df2[fields2],3)

df1['m_color'] = [MPL_Colors[i] for i in m_label]
df2['h_color'] = [MPL_Colors[i] for i in h_label]
df2['k_color'] = [MPL_Colors[i] for i in k_label]
```

Plot Data

In [6]:

```
pd_to_gif(df1 ,['x','y','z'] ,'Output\meanshift_plot',clrs='m_color')
pd_to_gif(df2 ,['x','y','z'] ,'Output\hdbscan_plot',clrs='h_color')
pd_to_gif(df2 ,['x','y','z'] ,'Output\kmeans_plot' ,clrs='k_color')
```



GIF version for this graphs available in here (https://github.com/EtzionData/Clustering_by_Silhouette)

. . . . .