Data

You are being provided with the required information in dataset 'Round1_2-of-3_Dataset.csv'.

Background

The assignment focuses on solving the problem of computing the Davies Bouldin Index, a similarity measure which indicates the quality of clustering that has been performed. You are being provided with required concept below. You can use the dataset 'Round1_2-of-3_Dataset.csv' to exploit your solution.

Problem

You need to implement from scratch in Python.

The following code snippet guides on implementing the clustering algorithm in Python scikit---learn.

```
import pandas as pd
from sklearn.cluster import k_means
df=pd.read_csv("Round2_Dataset_G2.csv")
df=df.dropna()
X1=df.copy()
del X1['Customer']
del X1['Effective To Date']
X4=pd.get_dummies(X1)
n=10
clf=k_means(X4,n_clusters=n)
centroids=clf[0]
labels=clf[1]
```

Here the centroids of the n clusters have been stored in the list 'centroids'.
The labels of each record have been stored in the list 'labels'.

The davies bouldin index has to be calculated for any value of n as follows:

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i, \text{ where}$$

$$R_i = \max_{j=1...nc, i\neq j}\left(R_{ij}\right), \; i = 1...n_c$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

$$d_{ij} = d(v_i, v_j), \; s_i = \frac{1}{\|c_i\|} \sum_{x \in c_i} d(x, v_i)$$

Where:

- d(x,y) is the Euclidean distance between x and y.
- c_i is the cluster i.
- v_i is the centroid of cluster c_i
- $\|c_i\|$ refers to the norm of c_i

Submission

You are required to submit code and processed dataset.

Evaluation

You shall be scored on the following:

1. Approach
2. Understanding
3. Implementation