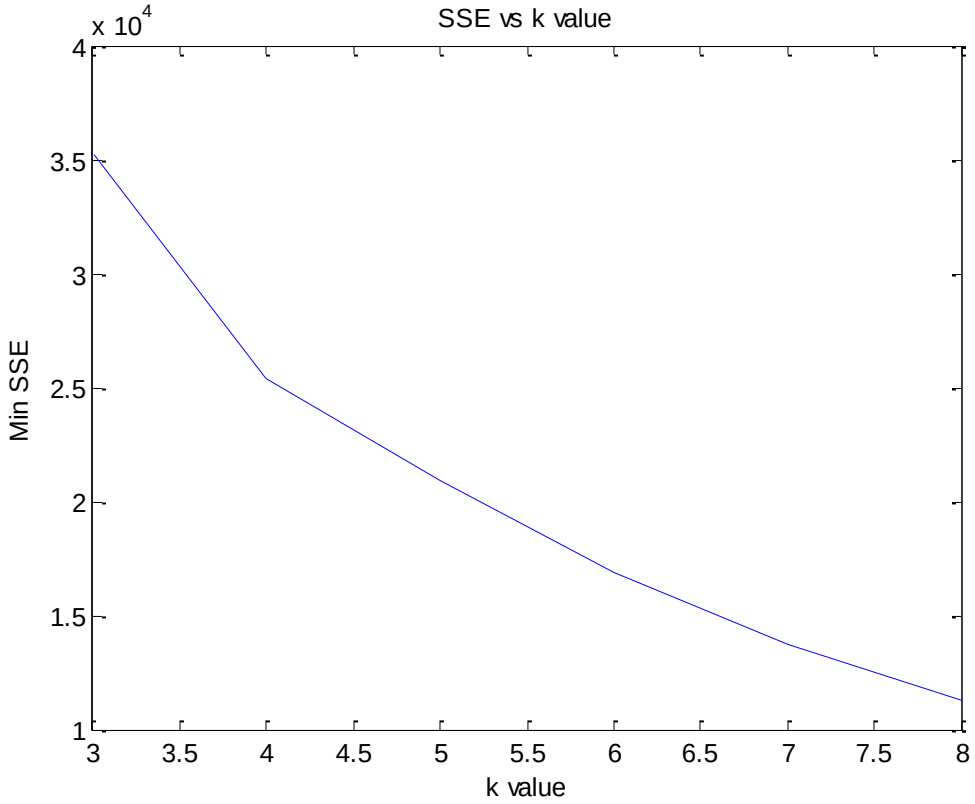
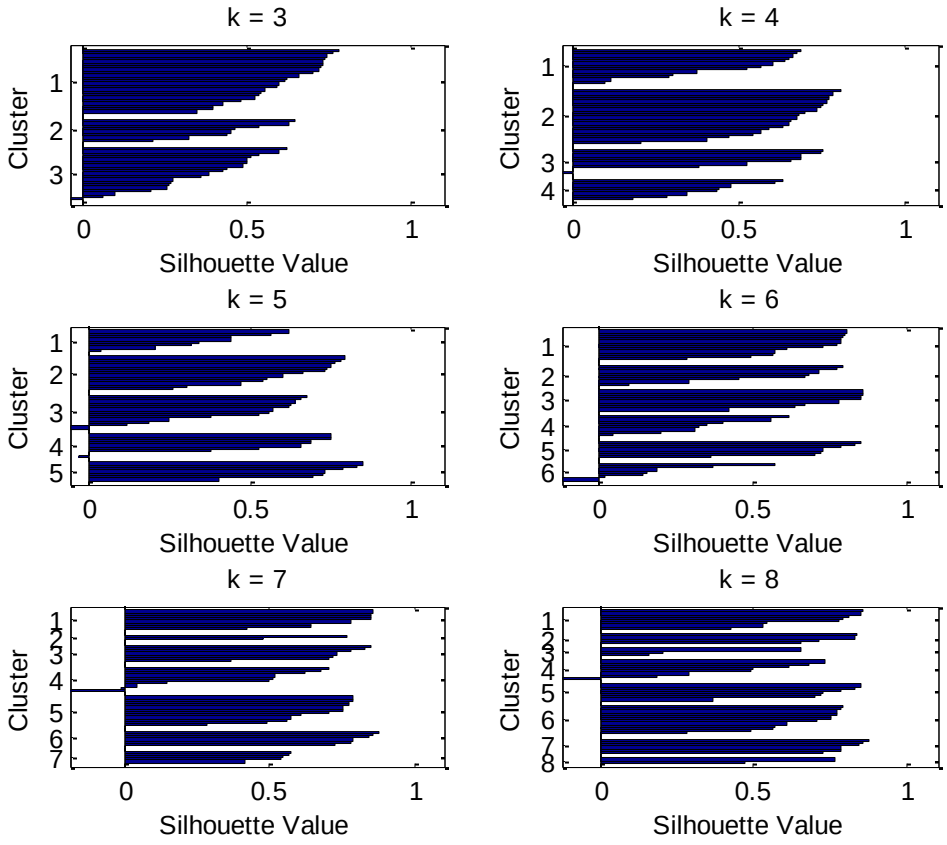
Intelligent Data Analysis
Homework #3
Due Date: Nov 14th, 2016, 7PM

Consider the data file attached with this homework. It contains scores for fifty students in four different subjects (Physics, Maths, English, and Music). Perform the following tasks with this data set.

1. Perform k-means clustering with this dataset for values of k to be 3, 4, 5, 6, 7, and 8. For each case of k run the clustering algorithm with three different initial cluster centers and select the one with the lowest SSE value. Plot the SSE against the values of k. Report the following in the submitted work: (Use Matlab kmeans function or any other similar toolbox)
    a. A plot of the SSE values against the values of k.
    Running k-means with k = 3
    SSE = 35247.21
    Running k-means with k = 4
    SSE = 25360.99
    Running k-means with k = 5
    SSE = 20927.03
    Running k-means with k = 6
    SSE = 16887.00
    Running k-means with k = 7
    SSE = 13762.85
    Running k-means with k = 8
    SSE = 11287.77



b. A plot of the silhouette coefficients for the data points in each clustering. (Each value of k results in one clustering)



c. What is the best number of clusters for this dataset? Justify your choice for the best number of clusters.

I selected 4 as the best number of clusters. Looking at the plot of the SSE values, you can see that there is a knee at k = 4. This indicates the there was a large advantage from k = 3 to k = 4 and of an advantage when moving from k = 4 to k = 3. Also, the silhouette values for k = 4 only has one node that is negative, but are fairly small. Silhouette values for k = 5, 6 are also pretty good, but there's no knee in the SSE values, which makes k = 4 the best choice.

d. For your choice of the best number of clusters report the centroids of all the clusters (Call this as Clustering-1).

Centroids for selected clustering:

Centroid 1	64.4615	54.7692	82.4615	70.9231
Centroid 2	92.2500	91.8000	70.1500	70.1500
Centroid 3	49.8889	37.2222	44.7778	64.3333
Centroid 4	56.1250	90.5000	56.2500	80.8750

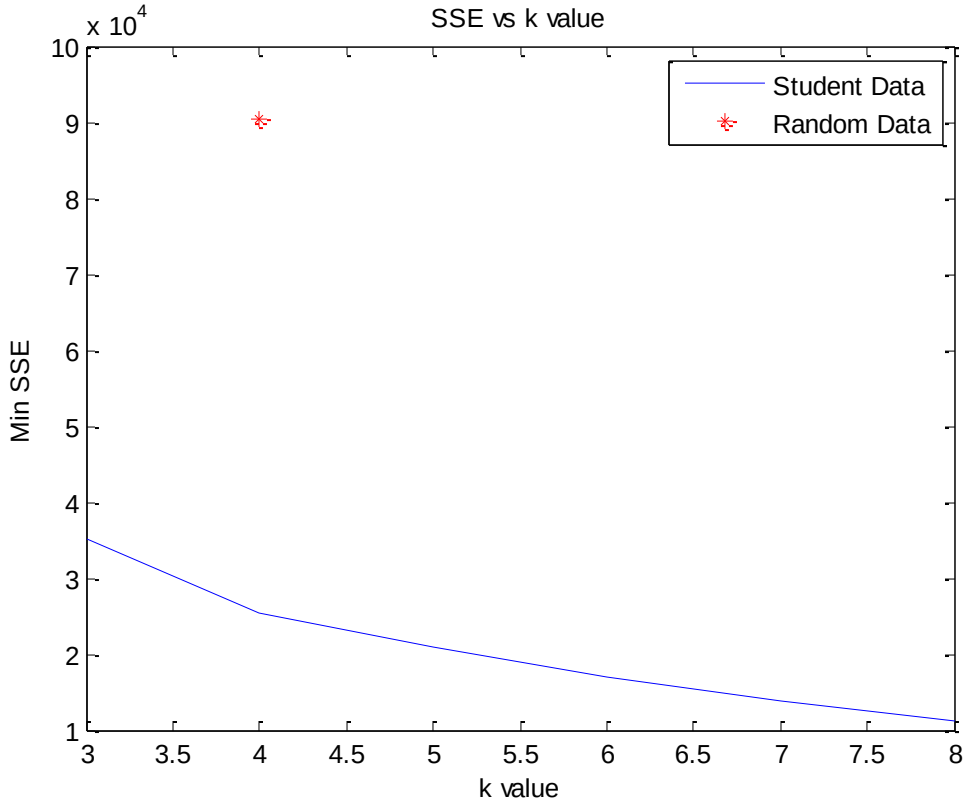
e. Generate 50 random 4-dimensional random data points such that each attribute can take values between 0 and 100. With this dataset form the same number of clusters as selected by you in (c) above. Report the centroids and populations of the clusters. Compare the SSE for this dataset with the SSE for the provided dataset. Comment on the differences between the two values.

Code used to generate the random data is available in the "Code section".

SSE value for random data: 90294.22
Centroids for random data:

Centroid 1	30.4211	28.3684	38.4211	52.6842
Centroid 2	84.1000	71.1000	35.8000	38.7000
Centroid 3	24.1818	85.1818	62.3636	47.9091
Centroid 4	60.0000	49.8000	77.3000	82.5000

Populations for clusters:
Cluster 1: 19
Cluster 2: 10
Cluster 3: 11
Cluster 4: 10

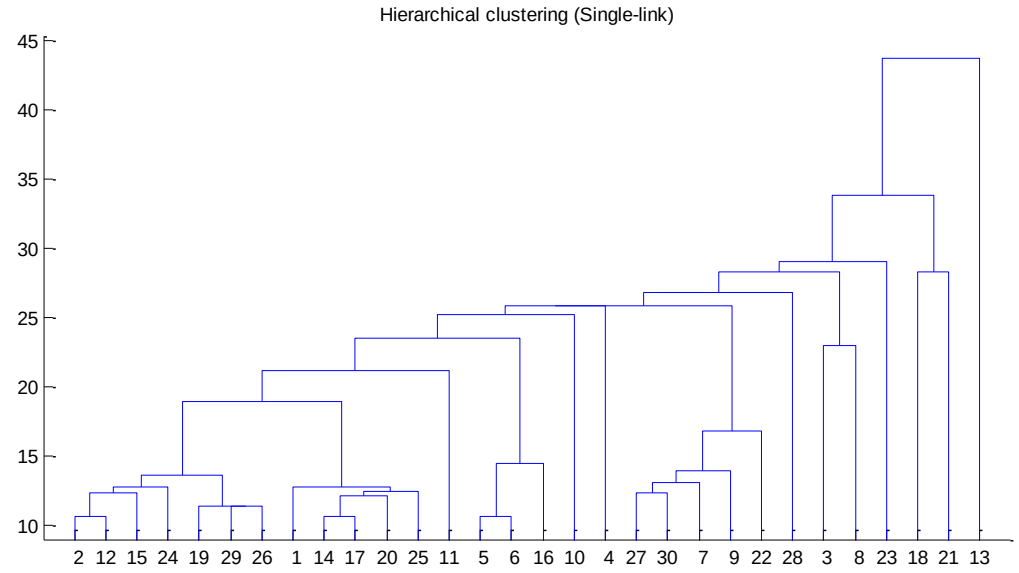


Looking at the SSE values, the random data has a value of **90,294,** the student data has a value of **25,360**. The random data is well above the SSE value for the student data. It is more than 3 times higher.

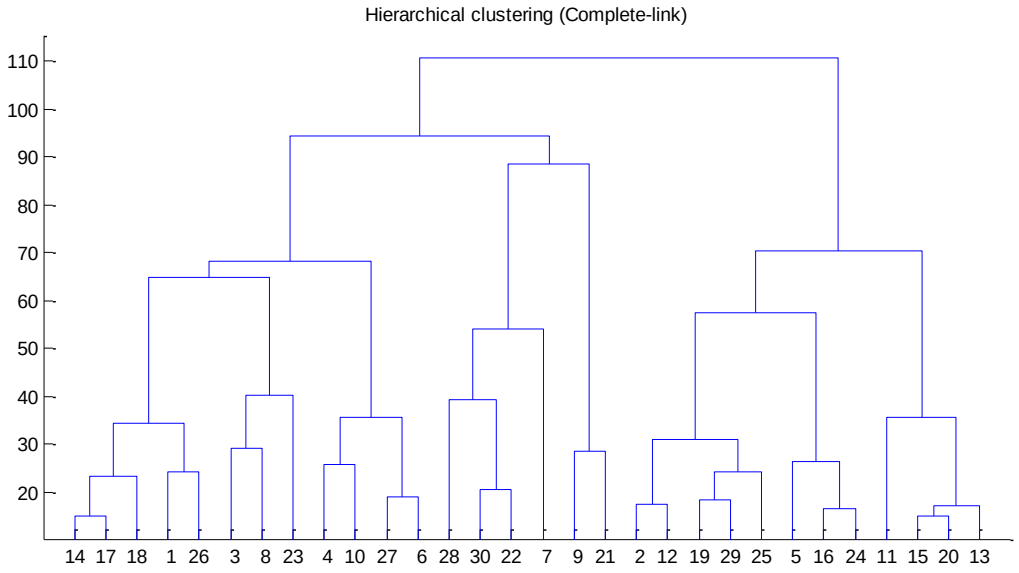
This indicates that the clustering on the student data is better than clustering on random data, indicating there is a pattern and the **clustering has significance**.

2. Perform hierarchical clustering for the students' scores dataset. Generate and show dendrograms for the cases (i) Single-Linkage clustering (Clustering-2), and (ii) Complete-Linkage clustering (Clustering-3). Use Euclidean distance for computing distance between data points. Report the following in the submitted work: (Use Matlab functions pdist and linkage, or any other similar toolbox.)
    a. Dendrograms for the two clusterings (Clustering-2 and Clustering-3)

Clustering-2



Clustering-3



b. Cluster compositions for each case when we need only four clusters. Write the data points included in each cluster and compute their centroids.

Code for calculating centroids is in the “Code” section.

-- Clustering-2 --
Centroids:
Centroid 1: 29 39 99 67

Centroid 2: 73.7174 73.1087 68.0435 71.5217

Centroid 3: 35.5000 86.5000 35.5000 62.5000

Centroid 4: 90 32 28 69

Cluster #1 - 1 points

Point 1:	29	39	99	67

Cluster #2 - 46 points

Point 1:	78	44	82	66
Point 2:	90	93	70	69
Point 3:	58	56	92	85
Point 4:	37	34	76	52
Point 5:	58	96	58	94
Point 6:	62	97	49	90
Point 7:	88	88	74	72
Point 8:	57	34	88	90
Point 9:	58	57	94	87
Point 10:	44	57	68	57
Point 11:	90	92	47	35
Point 12:	97	82	71	74
Point 13:	94	83	68	76
Point 14:	82	61	77	59
Point 15:	88	94	56	54
Point 16:	69	92	64	88
Point 17:	74	55	79	62
Point 18:	62	89	70	87
Point 19:	94	89	88	73
Point 20:	89	92	51	68
Point 21:	85	98	52	64
Point 22:	75	57	75	69
Point 23:	82	67	80	62
Point 24:	57	62	92	83
Point 25:	98	90	74	80
Point 26:	79	65	78	64
Point 27:	42	47	44	59
Point 28:	58	43	57	89
Point 29:	97	98	92	85

Point 30: 43 37 39 64
Point 31: 98 98 92 88
Point 32: 89 88 78 82
Point 33: 44 39 42 49
Point 34: 51 38 54 49
Point 35: 68 93 69 84
Point 36: 47 32 36 69
Point 37: 89 92 52 59
Point 38: 93 94 55 66
Point 39: 98 97 62 66
Point 40: 77 78 80 62
Point 41: 37 33 27 79
Point 42: 59 84 69 79
Point 43: 98 98 74 90
Point 44: 65 58 68 71
Point 45: 99 99 90 74
Point 46: 94 93 77 66
Cluster #3 - 2 points
Point 1: 29 94 32 53
Point 2: 42 79 39 72
Cluster #4 - 1 points
Point 1: 90 32 28 69

-- Clustering-3 --
Centroids:
Centroid 1: 35.5000 86.5000 35.5000 62.5000

Centroid 2: 55.0000 35.4000 37.4000 74.0000

Centroid 3: 60.5000 52.6667 76.0000 66.2778

Centroid 4: 85.8400 92.3600 68.0800 74.5200

Cluster #1 - 2 points
Point 1: 29 94 32 53
Point 2: 42 79 39 72
Cluster #2 - 5 points
Point 1: 58 43 57 89
Point 2: 43 37 39 64
Point 3: 47 32 36 69
Point 4: 90 32 28 69
Point 5: 37 33 27 79
Cluster #3 - 18 points
Point 1: 78 44 82 66
Point 2: 58 56 92 85
Point 3: 37 34 76 52
Point 4: 57 34 88 90

Point 5:	58	57	94	87
Point 6:	44	57	68	57
Point 7:	82	61	77	59
Point 8:	74	55	79	62
Point 9:	75	57	75	69
Point 10:	82	67	80	62
Point 11:	57	62	92	83
Point 12:	79	65	78	64
Point 13:	42	47	44	59
Point 14:	44	39	42	49
Point 15:	51	38	54	49
Point 16:	77	78	80	62
Point 17:	29	39	99	67
Point 18:	65	58	68	71
Cluster #4 - 25 points				
Point 1:	90	93	70	69
Point 2:	58	96	58	94
Point 3:	62	97	49	90
Point 4:	88	88	74	72
Point 5:	90	92	47	35
Point 6:	97	82	71	74
Point 7:	94	83	68	76
Point 8:	88	94	56	54
Point 9:	69	92	64	88
Point 10:	62	89	70	87
Point 11:	94	89	88	73
Point 12:	89	92	51	68
Point 13:	85	98	52	64
Point 14:	98	90	74	80
Point 15:	97	98	92	85
Point 16:	98	98	92	88
Point 17:	89	88	78	82
Point 18:	68	93	69	84
Point 19:	89	92	52	59
Point 20:	93	94	55	66
Point 21:	98	97	62	66
Point 22:	59	84	69	79
Point 23:	98	98	74	90
Point 24:	99	99	90	74
Point 25:	94	93	77	66

c. Comment on any differences in the cluster centers and cluster compositions for the two different clusterings as performed in (b) above.

Clustering-2 has two clusters that are have only one node (clusters 1 and 4), and one cluster with 2 nodes (cluster 3). This leaves most of the data points in cluster 2. It seems like Clustering-2 clustered a few noise data points, and put most of the data into one large cluster. This doesn't seem to give us much information. The largest cluster has a centroid value around 70 for every dimension.

Clustering-3 has more evenly spaced clusters. The sizes are 2, 5, 18 and 25. The largest cluster (cluster 4) in this case has fairly high test scores. Only one of the average test values for the cluster 4 centroid is lower than any of the other cluster centroids (dimension 3 has a value of 68, while cluster 3 centroid has a value of 76 in dimension 3).

d. Compute Rand Index for the comparison of Clustering-2 and Clustering-3 and show the counts a, b, c, and d as determined for computing the Rand index. Explain the meaning of each count and why such counts have been obtained for this dataset and these clusterings in this comparison.

Comparing Clustering-2 to Clustering-3
a = 443
b = 168
c = 593
d = 21
Rand Index: 0.4988

The value of 'a' means that there are 443 pairs that are in the same cluster in clustering-2 and they are in the same cluster in clustering-3. So, 443 pairs of points were clustered into the same cluster in both clusterings.

The value of 'b' means that there 168 pairs that are in different clusters in clustering-2 and they are in different clusters in clustering-3.

The value of 'c' means that there are 593 pairs that clustering-2 put into the same cluster, while clustering-3 put into different clusters.

The value of 'd' means that there are 21 pairs that clustering-2 put into different clusters, while clustering 3 put into the same clusters.

The 'a' and 'c' values are so high because clustering-2 put most values into the one cluster (cluster 2). That leaves few pairs left for clustering-2 to put into different pairs, which is required for values 'b' and 'c'.

A rand index value of 0.4988 means that the clusterings agree on about half the pair.

3. Compute Rand Index for the comparison of Clustering-1 and Clustering-2 and show the counts a, b, c, and d as determined for computing the Rand index. Explain the meaning of each count and why such counts have been obtained for this dataset and these clusterings in this comparison.

   Comparing Clustering-1 to Clustering-2
   a = 300
   b = 157
   c = 32
   d = 736
   Rand Index: 0.3731

   The value of 'a' means that there are 300 pairs that are in the same cluster in clustering-2 and they are in the same cluster in clustering-3. So, 443 pairs of points were clustered into the same cluster in both clusterings.

   The value of 'b' means that there 157 pairs that are in different clusters in clustering-2 and they are in different clusters in clustering-3.

   The value of 'c' means that there are 32 pairs that clustering-2 put into the same cluster, while clustering-3 put into different clusters.

   The value of 'd' means that there are 736 pairs that clustering-2 put into different clusters, while clustering 3 put into the same clusters.

   The 'd' value is very high in this case, which means that there the clustering algorithms do not agree. This is also reflecting in the low rand index value of 0.3731. This means that the kmeans clustering algorithm does not produce clusters very similar to single-link clustering algorithm.

Code

```
Part 1% K-Means clustering
clear all; clc; close all;

% We will be using the Squared Euclidean method for determining distance
% between two points.

% Load data. Only read first 50 entries.
% Exclude first column, it is student number.
Data = xlsread('StudentData2.xlsx','B2:E51');

startK = 3;
endK = 8;
kmeansIter = 3;

% Init minimum SSE for each value of k
minSSE = ones(endK,1) * 10^10;
clustering = cell(8,2);

% Run k-means for values 3 through 8
for k = startK : endK
    fprintf('Running k-means with k = %i\n',k);
    % For each k-means, run three times and choose the clustering with the
    % smallest SSE value.
    for c = 1:kmeansIter
        temp = randperm(50);
        seeds = temp(1:k);

        % K-means by default uses Squared Euclidean distance.
        [clusterID, centroids, sumD] = kmeans(Data, k, 'Start', Data(seeds,:));

        % Find SSE
        SSE = sum(sumD);
        if SSE < minSSE(k)
           minSSE(k) = SSE;
           clustering{k} = {centroids clusterID};
        end
    end
    fprintf('SSE = %0.2f\n', minSSE(k));
    subplot(3,2, k - startK + 1);
    silhouette(Data, clustering{k}{2});
end

figure
plot(3:8, minSSE(3:8))
title('SSE vs k value')
xlabel('k value')
ylabel('Min SSE')

% Select the best clustering
choice = input('Please enter the best clustering (3-8): ');
while choice < 3 || choice > 8
```

```
    fprintf('Invalid entry\n');
    choice = input('Please enter the best clustering (3-8): ');
end
labels1 = clustering{choice}{2};
centroids1 = clustering{choice}{1};
for i = 1:choice
   clusters1{i} =  Data(labels1 == i,:);
end

% Print the centroids
fprintf('Centroids for selected clustering:\n');
for i = 1:choice
    fprintf('Centroid %i',i)
    disp(centroids1(i,:));
end
save('Clustering-1','clusters1','labels1','centroids1');

% Compare to random data
randomData = randi([0 100],[50 4]);
temp = randperm(50);
seeds = temp(1:choice);
[clusterID, centroids, sumD] = kmeans(randomData, choice, 'Start', randomData(seeds,:));

% Find SSE
SSE = sum(sumD);
fprintf('SSE value for random data: %0.2f\n',SSE);
fprintf('Centroids for random data:\n')
for i = 1:choice
    fprintf('Centroid %i',i)
    disp(centroids(i,:));
end
fprintf('Populations for clusters:\n')
for i = 1:choice
   fprintf('Cluster %i: %i\n',i,sum(clusterID == i));
end
hold on; plot(choice, SSE, 'r*');
legend('Student Data','Random Data');
```

Part 2

```
clear all; close all; clc;

Data = xlsread('StudentData2.xlsx','B2:E51');

numClusters = 4;

% Perform clustering
dist = pdist(Data);
clustering2 = linkage(dist, 'single');
clustering3 = linkage(dist, 'complete');

% Display dendrograms
```

```
figure
dendrogram(clustering2);
title('Hierarchical clustering (Single-link)')

figure
dendrogram(clustering3);
title('Hierarchical clustering (Complete-link)')

% Create 4 clusters
labels2 = cluster(clustering2, 'maxclust', numClusters);
labels3 = cluster(clustering3, 'maxclust', numClusters);

% Calculate centroids
clusters2 = cell(numClusters,1);
centroids2 = zeros(numClusters,4);
clusters3 = cell(numClusters,1);
centroids3 = zeros(numClusters,4);
for i = 1:numClusters
    clusters2{i} = Data(labels2 == i,:);
    if sum(labels2 == i) > 1
        centroids2(i,:) = sum(clusters2{i}) / sum(labels2 == i);
    else
        centroids2(i,:) = clusters2{i};
    end

    clusters3{i} = Data(labels3 == i,:);
    if sum(labels3 == i) > 1
        centroids3(i,:) = sum(clusters3{i}) / sum(labels3 == i);
    else
        centroids3(i,:) = clusters3{i};
    end
end

% Print the clusters
fprintf('-- Clustering-2 --\n')
fprintf('Centroids:\n')
for i = 1:numClusters
    fprintf('Centroid %i: ',i)
    disp(centroids2(i,:))
end
for i = 1:numClusters
    curCluster = clusters2{i};
    [numRows numCols] = size(curCluster);
    fprintf('Cluster #%i - %i points\n',i,numRows);
    for row = 1:numRows
        point = curCluster(row,:);
        fprintf('Point %i:\t',row);
        for dim = 1:length(point)
            fprintf('%i\t',point(dim));
        end
        fprintf('\n');
    end
end
fprintf('\n\n');
```

```
fprintf('-- Clustering-3 --\n')
fprintf('Centroids:\n')
for i = 1:numClusters
    fprintf('Centroid %i: ',i)
    disp(centroids3(i,:))
end
for i = 1:numClusters
    curCluster = clusters3{i};
    [numRows numCols] = size(curCluster);
    fprintf('Cluster #%i - %i points\n',i,numRows);
    for row = 1:numRows
        point = curCluster(row,:);
        fprintf('Point %i:\t',row);
        for dim = 1:length(point)
            fprintf('%i\t',point(dim));
        end
        fprintf('\n');
    end
end

% Calculate rand index
a = 0;
b = 0;
c = 0;
d = 0;
for i = 1:length(labels2)
    for j = i+1:length(labels2)
        % Are elements i,j in same set in clustering 1?
        p2 = labels2(i) == labels2(j);

        % What set is pair i,j in clustering3?
        p3 = labels3(i) == labels3(j);

        if p2 && p3
            % They in the same set in both clusterings
            a = a + 1;
        elseif ~p2 && ~p3
            % They are in different sets in both clusterings
            b = b + 1;
        elseif p2 && ~p3
            % They are in the same set clustering2, diff sets clustering3
            c = c + 1;
        else
            % They are in diff set clustering2, same sets clustering3
            d = d + 1;
        end
    end
end
randIndex = (a + b)/(a + b + c + d);

fprintf('\n')
fprintf('Comparing Clustering-2 to Clustering-3\n')
fprintf('a = %i\n',a);
fprintf('b = %i\n',b);
```

```
fprintf('c = %i\n',c);
fprintf('d = %i\n',d);
fprintf('Rand Index: %0.4f\n',randIndex);

save('Clustering-2', 'clustering2','clusters2','labels2','centroids2');
save('Clustering-3', 'clustering3','clusters3','labels3','centroids3');
```

Part 3

```
% Compare Clustering-1 and Clustering-2
clear; clc;

load Clustering-1
load Clustering-2

% Calculate rand index
a = 0;
b = 0;
c = 0;
d = 0;
for i = 1:length(labels1)
    for j = i+1:length(labels1)
        % Are elements i,j in same set in clustering 1?
        p1 = labels1(i) == labels1(j);

        % What set is pair i,j in clustering3?
        p2 = labels2(i) == labels2(j);

        if p1 && p2
            % They in the same set in both clusterings
            a = a + 1;
        elseif ~p1 && ~p2
            % They are in different sets in both clusterings
            b = b + 1;
        elseif p1 && ~p2
            % They are in the same set clustering2, diff sets clustering3
            c = c + 1;
        else
            % They are in diff set clustering2, same sets clustering3
            d = d + 1;
        end
    end
end
randIndex = (a + b)/(a + b + c + d);

fprintf('\n')
fprintf('Comparing Clustering-1 to Clustering-2\n')
fprintf('a = %i\n',a);
fprintf('b = %i\n',b);
fprintf('c = %i\n',c);
```

```
fprintf('d = %i\n',d);
fprintf('Rand Index: %0.4f\n',randIndex);
```