No	Name	ID
1	يوسف عزت احمد عبدالمجيد	2018170476
2	احمد مصدق ابراهيم	2018170057
3	Peter Nyuol Majok	2017170530
4	احمد خالد صابر عبدالهادي	2018170512

Project 3: Clustering Heart Disease Patients Description

Problem Statement: Doctors frequently study former cases to learn how to best treat their patients. A patient who has a similar health history or symptoms to a previous patient could benefit from
undergoing the same treatment. This project investigates whether doctors might be able to
group together patients to target treatments using common *unsupervised learning* techniques.

Our Team Methodology:-

Data Exploration and Cleaning

- Clean the data by removing any unnecessary columns or rows.
- Fixing any errors or inconsistencies.
- Visual Representation of the Dataset.

Data Manipulation and Feature Engineering

- Visual Representation of Feature Relationships.

Modeling with different machine learning algorithms (more than one Clustering algorithm) to see which one performs best.

- Clustering:-
  - What's the Best Value of Clusters ?
  - What's the Impact of different features on the clusters ?

  Visualize Model Predictions.
- Hierarchical Clustering
  Visualize Model Predictions.

Demonstration in more Details:-

For this project, we conducted data exploration and cleaning on the heart disease dataset. The following steps were performed:

1. Checking Numerical Features for K-means:

We identified the numerical features in the dataset to determine their suitability for the clustering analysis.

2. Checking for Missing Values:
We checked for any missing values in the dataset and found one instance of missing data.

3. Handling Missing Values:
To address the missing value, we removed the corresponding row from the dataset.

4. Removing the ID Column:
As the ID column does not contribute to the clustering analysis, we removed it from the dataset.

5. Scaling the Data:
To ensure that all features have a similar scale and prevent any bias in the clustering process, we performed data scaling.

Before scaling:

After scaling:

```
> summary(scaled)
      age                sex                cp              trestbps            chol         
 Min.   :-2.8122   Min.   :-1.4620   Min.   :-2.2563   Min.   :-2.1481   Min.   :-2.34596  
 1st Qu.:-0.7226   1st Qu.:-1.4620   1st Qu.:-0.1741   1st Qu.:-0.6720   1st Qu.:-0.68183  
 Median : 0.1572   Median : 0.6818   Median :-0.1741   Median :-0.1042   Median :-0.08197  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000  
 3rd Qu.: 0.7071   3rd Qu.: 0.6818   3rd Qu.: 0.8670   3rd Qu.: 0.4635   3rd Qu.: 0.57594  
 Max.   : 2.4667   Max.   : 0.6818   Max.   : 0.8670   Max.   : 3.8699   Max.   : 6.12950  
      fbs             restecg            thalach            exang            oldpeak        
 Min.   :-0.4099   Min.   :-1.008271   Min.   :-3.3753   Min.   :-0.7077   Min.   :-0.9034  
 1st Qu.:-0.4099   1st Qu.:-1.008271   1st Qu.:-0.7409   1st Qu.:-0.7077   1st Qu.:-0.9034  
 Median :-0.4099   Median :-0.003295   Median : 0.1228   Median :-0.7077   Median :-0.2124  
 Mean   : 0.0000   Mean   : 0.000000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
 3rd Qu.:-0.4099   3rd Qu.: 1.001681   3rd Qu.: 0.7274   3rd Qu.: 1.4084   3rd Qu.: 0.4786  
 Max.   : 2.4315   Max.   : 1.001681   Max.   : 2.2820   Max.   : 1.4084   Max.   : 4.4519  
     slope        
 Min.   :-0.9871  
 1st Qu.:-0.9871  
 Median : 0.6403  
 Mean   : 0.0000  
 3rd Qu.: 0.6403  
 Max.   : 2.2677  
> 
```

These data exploration and cleaning steps resulted in a clean and prepared dataset for the subsequent clustering analysis.

The Data Transformation Visually:
Before:

	id	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
1	1	63	1	1	145	233	1	2	150	0	2.3	3
2	2	67	1	4	160	286	0	2	108	1	1.5	2
3	3	67	1	4	120	229	0	2	129	1	2.6	2
4	4	37	1	3	130	250	0	0	187	0	3.5	3
5	5	41	0	2	130	204	0	2	172	0	1.4	1
6	6	56	1	2	120	236	0	0	178	0	0.8	1
7	7	62	0	4	140	268	0	2	160	0	3.6	3
8	8	57	0	4	120	354	0	0	163	1	0.6	1
9	2	67	1	4	160	286	0	2	108	1	1.5	2
10	9	63	1	4	130	254	0	2	147	0	1.4	2

After:

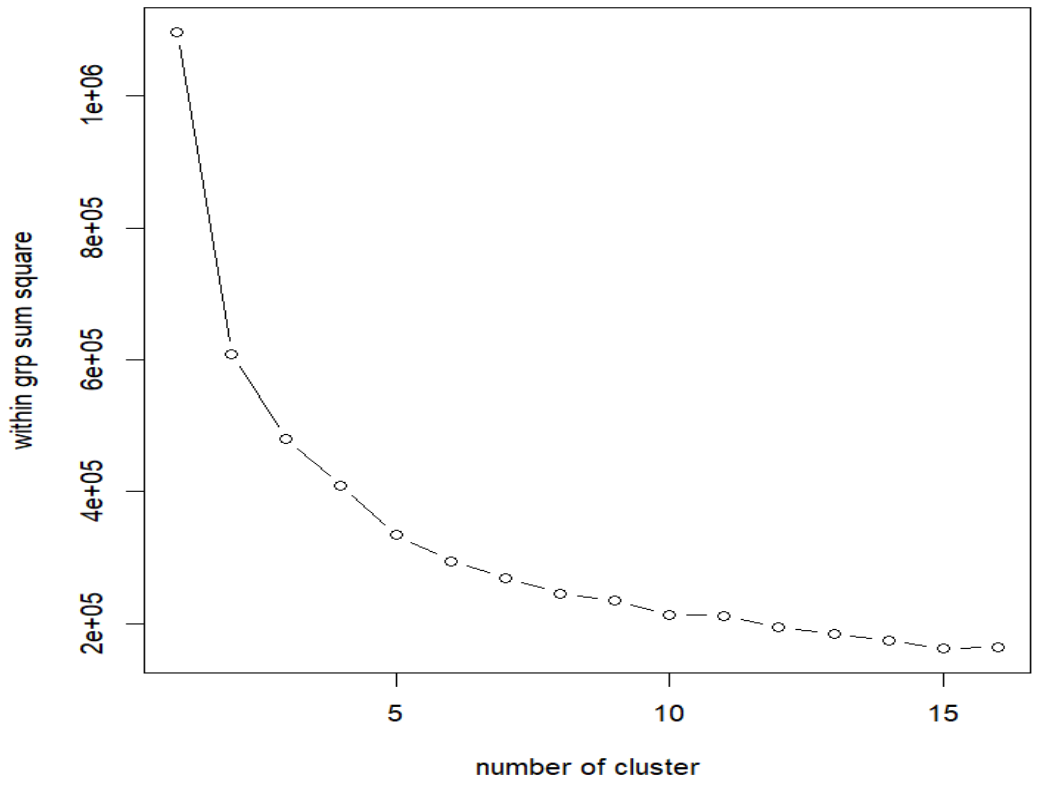
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
1	0.92704727	0.681775	-2.2562829	0.747352796	-0.275472961	2.4315365	1.001682	0.036389019	-0.7076833	1.08324651	2.2676694
2	1.36695301	0.681775	0.8670134	1.598944088	0.750095996	-0.4099142	1.001682	-1.777398283	1.4084285	0.39223436	0.6402831
3	1.36695301	0.681775	0.8670134	-0.671966026	-0.352874392	-0.4099142	1.001682	-0.870504632	1.4084285	1.34237607	0.6402831
4	-1.93234007	0.681775	-0.1740854	-0.104238497	0.053483120	-0.4099142	-1.008272	1.634249262	-0.7076833	2.11976474	2.2676694
5	-1.49243432	-1.461951	-1.2151841	-0.104238497	-0.836633334	-0.4099142	1.001682	0.986468083	-0.7076833	0.30585784	-0.9871032
6	0.15721222	0.681775	-1.2151841	-0.671966026	-0.217421888	-0.4099142	-1.008272	1.245580554	-0.7076833	-0.21240128	-0.9871032
7	0.81707083	-1.461951	0.8670134	0.463489031	0.401789558	-0.4099142	1.001682	0.468243139	-0.7076833	2.20614126	2.2676694
8	0.26718865	-1.461951	0.8670134	-0.671966026	2.065920319	-0.4099142	-1.008272	0.597799375	1.4084285	-0.38515432	-0.9871032
9	1.36695301	0.681775	0.8670134	1.598944088	0.750095996	-0.4099142	1.001682	-1.777398283	1.4084285	0.39223436	0.6402831
10	0.92704727	0.681775	0.8670134	-0.104238497	0.130884550	-0.4099142	1.001682	-0.093167217	-0.7076833	0.30585784	0.6402831

Clustering

(K-means)

For the clustering analysis, we employed the k-means algorithm to group the heart disease patients based on their features. Here's an overview of the process:

1. Determining the Number of Clusters (Elbow Method):
To determine the optimal number of clusters, we used the elbow method. We calculated the within-group sum of squares (WSS) for different values of k and plotted them on a line graph. This helped us identify the number of clusters that provided a significant reduction in WSS.
Here’s the graph:



From the graph, we found that the best choice of k was 5, where the WSS showed a significant decrease.

2. Running the K-means Algorithm:

After determining the optimal number of clusters, we ran the k-means algorithm. We initialized the algorithm with the chosen number of clusters and executed it on the preprocessed dataset.

3. Number of Patients in Each Cluster:

We examined the sizes of the clusters obtained from the k-means algorithm. This helped us understand the distribution of patients across different clusters.

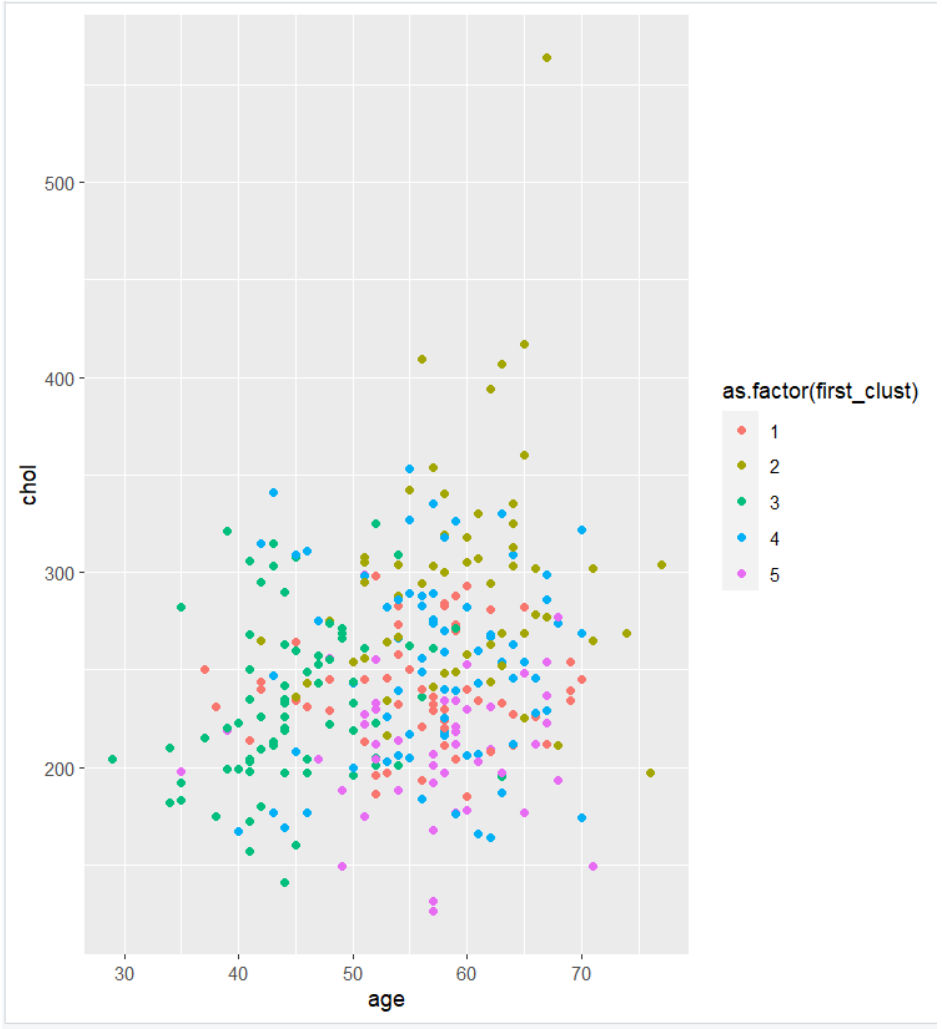
4.Checking Clustering Stability:

To ensure the robustness of the clustering results, we performed the clustering algorithm a second time by changing the seed. This allowed us to check the stability of the clusters and verify if similar groupings were obtained.

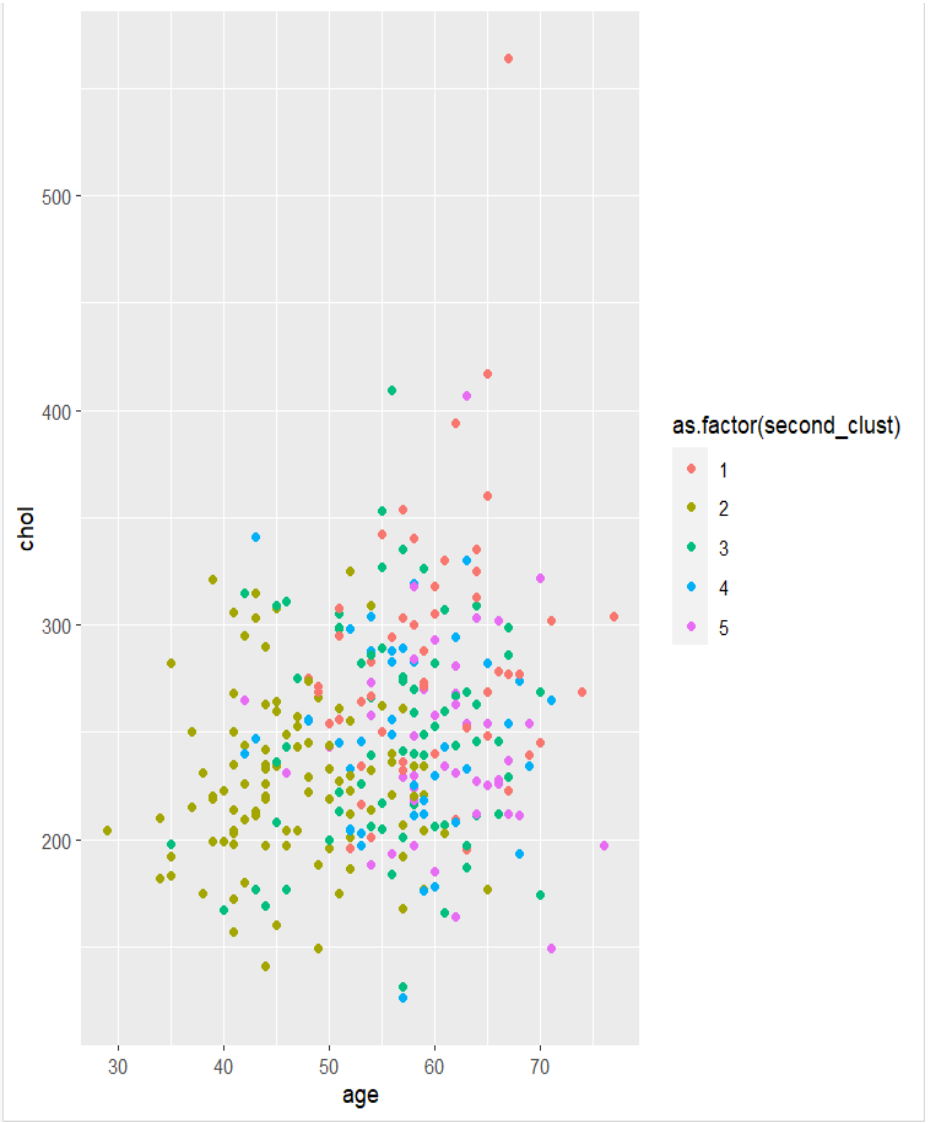
5. Visualizing the Clusters:

To visualize the clustering results, we created scatter plots of selected features. By mapping the cluster assignments to colors, we were able to visually analyze the distribution of patients in each cluster.

#plot_1

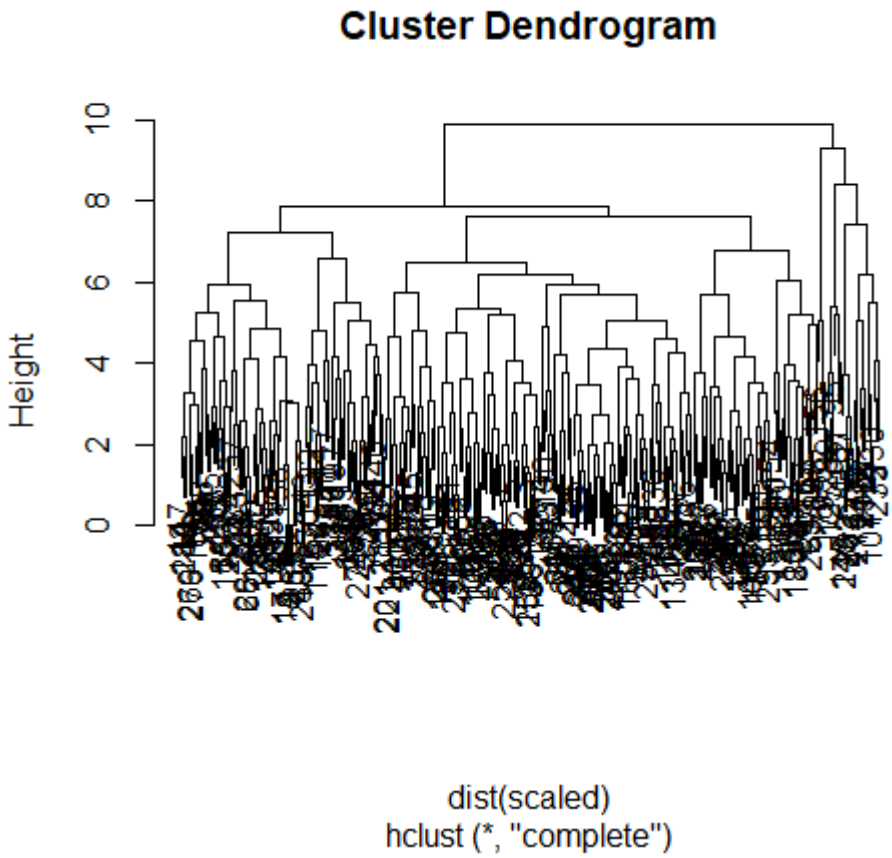

500
400
300
200
chol
30
40
50
60
70
age
as.factor(first_clust)
1
2
3
4
5

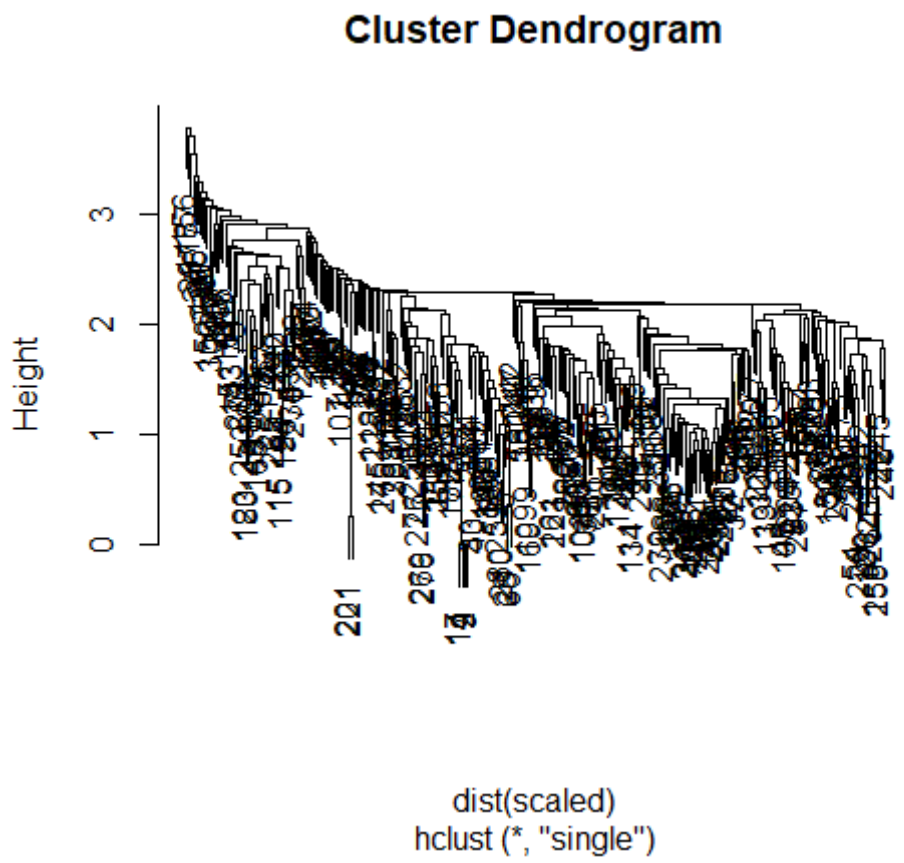
#plot_2



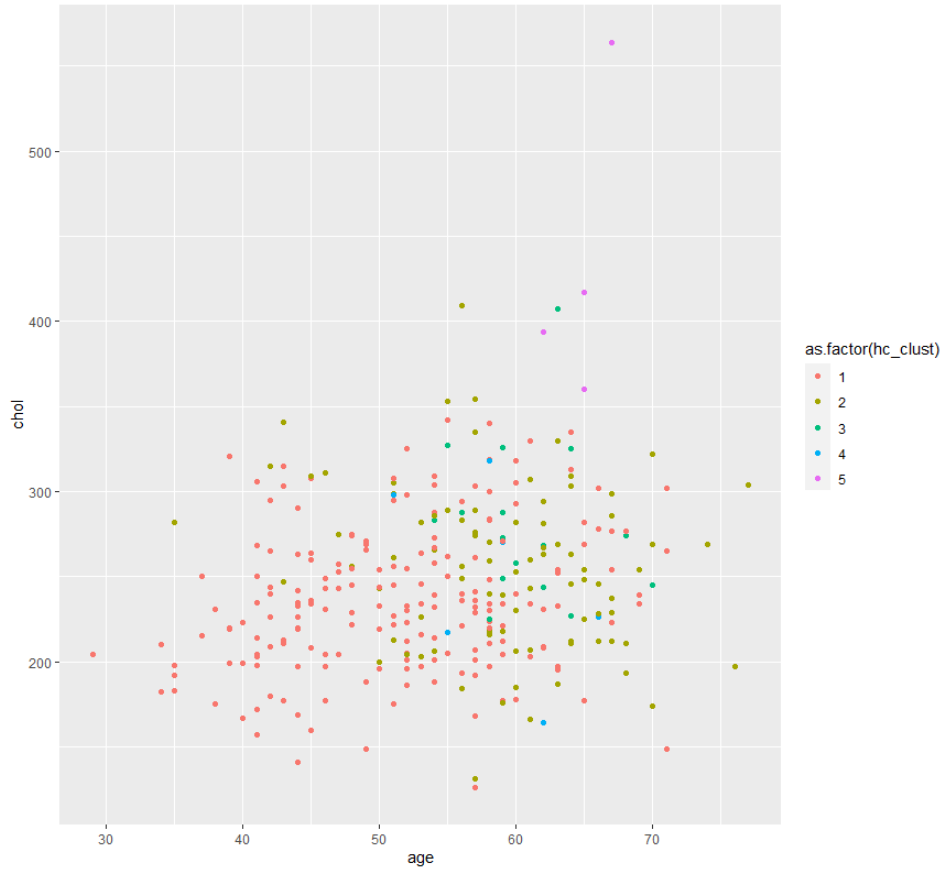
We found that the clustering results remained stable across different iterations, indicating the reliability of the clustering algorithm.

Hierarchical Clustering




Cluster Dendrogram
Height
0
1
2
3
dist(scaled)
hclust (*, "single")

Stability is remarked also here despite different Clustering Algorithms! Interesting!!



Different Perspective on Clustering using two different Factors