

Data Mining (DMG)

Assignment 4

Reshan Faraz (PhD19006)

Akhand Pratap Singh (MT20029)

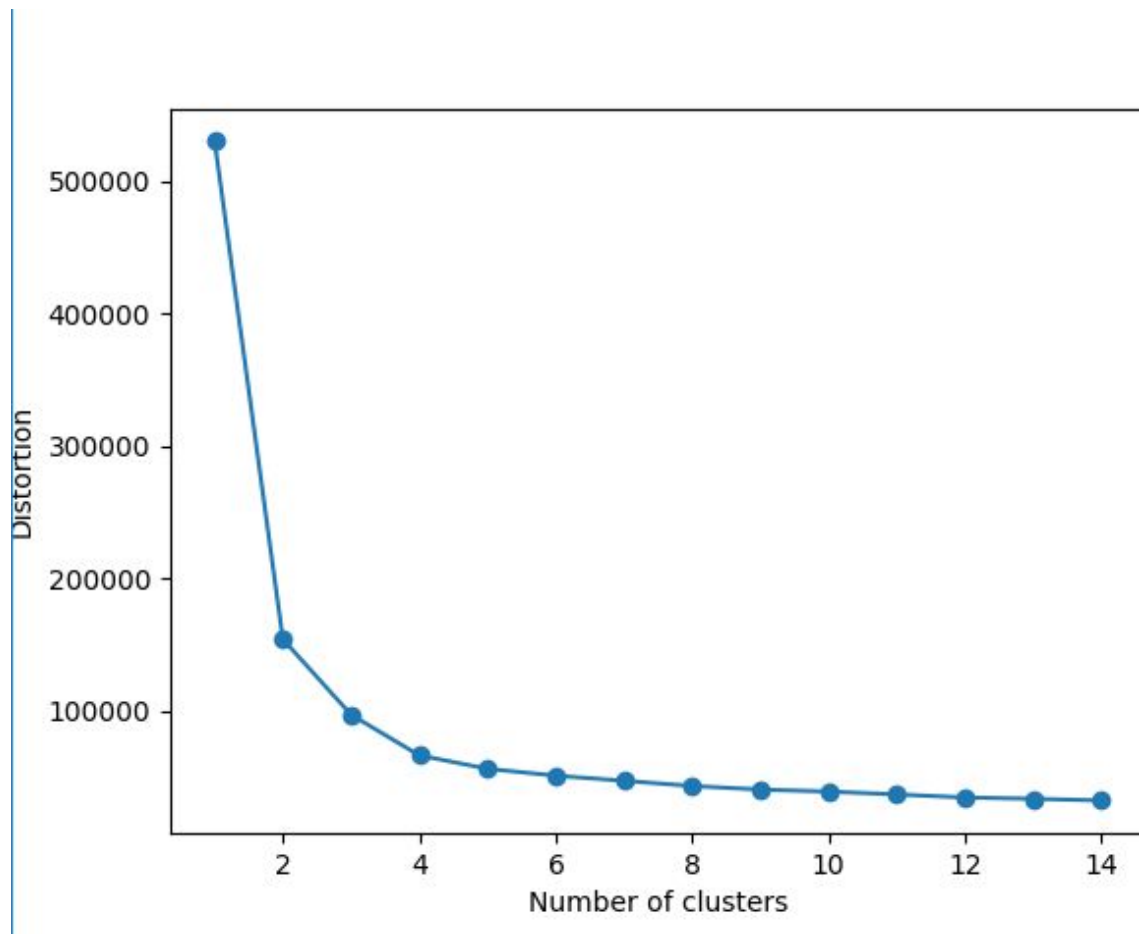
Date : 04 Dec 2020

Methodology and Approaches :-

- 1.) First I started with basic data preprocessing on the given dataset i.e.(clustering_data.csv) by loading it.During Data analysis I found that 'id' attributes is not needed for building my model.Therefore I dropped the 'id' attribute.
 - 2.) After that I preprocess the data using Label Encoder .And converted all the attributes into integers.
 - 3.) After that I performed basic Data Visualization on the dataset and found out the optimal cluster using the elbow method
 - 4.) I also scale data with mean and standard deviation.
 - 5.) After that I build Kmeans Clustering model with 4 clusters and plotted the graph for the same.Also We found out the centroid for each cluster.
 - 6.) Then I implemented second clustering method by build Kmeans Clustering model with 7 clusters and plotted the graph for the same.Also We found out the centroid for each cluster.we also match our centroid with true labels which is quite close with the true label.
 - 7.) After that I implemented third clustering method by build Agglomerative Clustering model .In this we form grouping between the clusters as there is no concept of centers and plotted the graph for the same and plotted the Dendrogram of it.
-

Visualize of data :-.

I tried different values of k to find the optimal k by using the elbow method.
K varies from 1 to 15.



Data Analysis steps :-

- 1.) First I started with basic data preprocessing on the given dataset i.e.(clustering_data.csv) by loading it.
- 2.) During Data analysis I found that the 'id' attribute is not needed for building my model. Therefore I dropped the 'id' attribute.
- 3.) After that I converted my data to integers using Label Encoder
- 4.) I also scale data with mean and standard deviation to get the clusters frequency close to the standard true label.
- 5.) After that I performed basic Data Visualization on the dataset and found out the optimal cluster using KMeans and its graph.
- 6.) For each clustering method I reduce the data to 2D using PCA for better visualization of Data.

Learning in doing this assignment :

Following are the list of things we learned from doing this assignment:

- 1.) First thing I learned: How to perform basic **Data Preprocessing** on any given dataset and how to scale the data
 - 2.) I also learned how to use Label Encoder which is useful when we have string as features
 - 3.) I learned How to perform basic **Data Visualization** on any given dataset.
 - 4.) I learned to build **KMeans clustering** Model with centroids.
 - 5.) I learned to build **KMeans++ clustering** Model with centroids.
 - 6.) I learned to build **Agglomerative clustering** Model and forming grouping between the clusters.
 - 7.) I build strong concepts and have good understanding of different clustering algorithm.
-

Report Centroid/representative object/prototype of each cluster.

Kmeans++ with 7 clusters (before reducing the data to 2D):

```
[ 0.25225027 -2.13483579 -0.47700953  0.16261746  0.2244684  0.14845087
-2.77144516 -1.69182583 -1.93432238  0.26812981] -> cluster 1
[ 1.89120682  1.0760088 -0.56425447  0.23956468  1.64881182  0.71945148
 0.54393327  1.63976196  1.2070449  0.40423389] -> cluster 2
[-0.34582322  0.81657511 -0.56541759 -0.82366428 -0.04675298  0.05088636
 0.88600478 -2.47299029 -1.99808643  0.2952747 ] -> cluster 3
[-0.74655412  0.93783726  0.6357562 -0.65998184 -1.95774319  1.408553
 1.00137773  0.85310725  1.00456502  0.33039341] -> cluster 4
[-0.63375753  0.15470177  1.30844013 -0.07938762 -0.17111139 -2.83979133
 1.39408065  0.75254259  1.0379963  0.35890126] -> cluster 5
[ 0.14835865 -2.55172906  0.09006649 -0.32359597  1.08977057  0.48926841
-2.78943461  1.59079389  1.61899845  0.32025125] -> cluster 6
[-0.48672043  0.21041061 -0.59668882  3.09983968 -0.31938217 -0.32523077
-0.04456828  0.08664872 -0.74257425 -3.34841875] -> cluster 7
```

Kmeans ++ with 7 clusters (after reducing the data to 2D) :

```
1.96234133 -2.81490817 -> cluster 1
-3.38012978 -0.2438359 -> cluster 2
1.95939914  3.51163197 -> cluster 3
-2.69270703  3.19547824 -> cluster 4
-2.44225038 -2.73172359 -> cluster 5
-0.03654496 -0.08932585 -> cluster 6
```



```
3.23127833 0.18018128 -> cluster 7
```

Kmeans

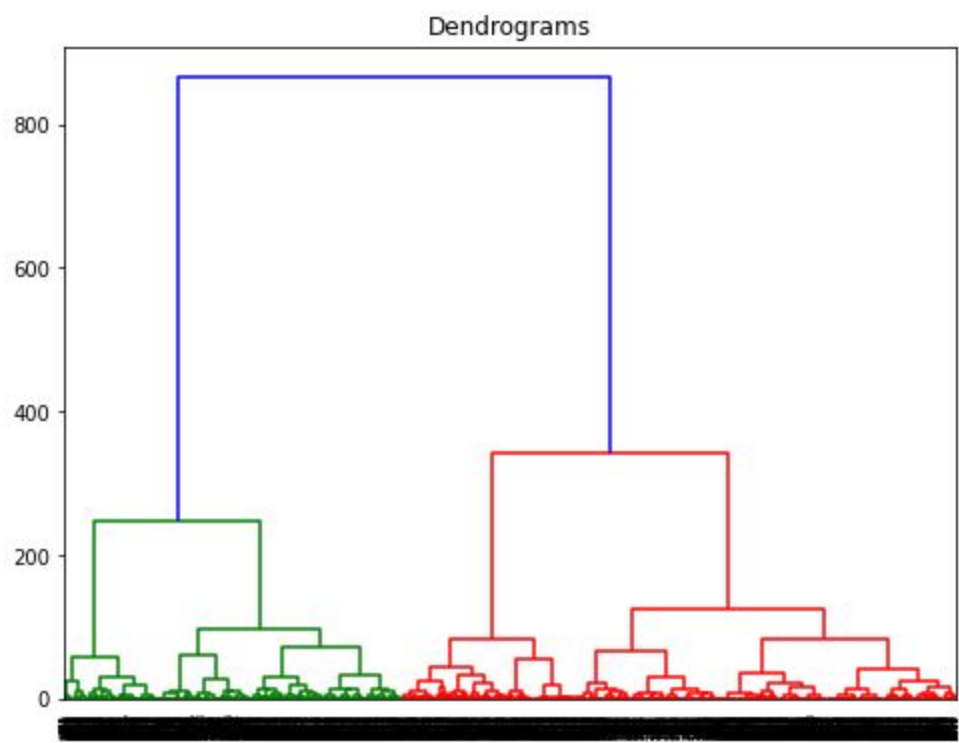
Kmeans with cluster 4 (before reducing the data to 2D):

```
[1.36638452 1.43047158 1.48004837 1.1668682 13.66868198 0.61064087
 0.62515115 2.15961306 1.40749698 1.45707376] -> cluster 1
[2.14311111 1.29422222 1.49422222 1.26133333 23.60088889 0.67022222
 0.56622222 3.10133333 2.008 1.39111111] -> cluster 2
[1.20474263 1.61133603 1.52573742 1.32851359 3.47484095 0.70040486
 0.75477154 2.80161943 1.89994216 1.41873916] -> cluster 3
[1.52164009 1.38496583 1.30979499 0.52164009 33.50341686 0.64692483
 0.61047836 2.85193622 1.97722096 1.41002278] -> cluster 4
```

Kmeans with cluster 4 (after reducing the data to 2D):

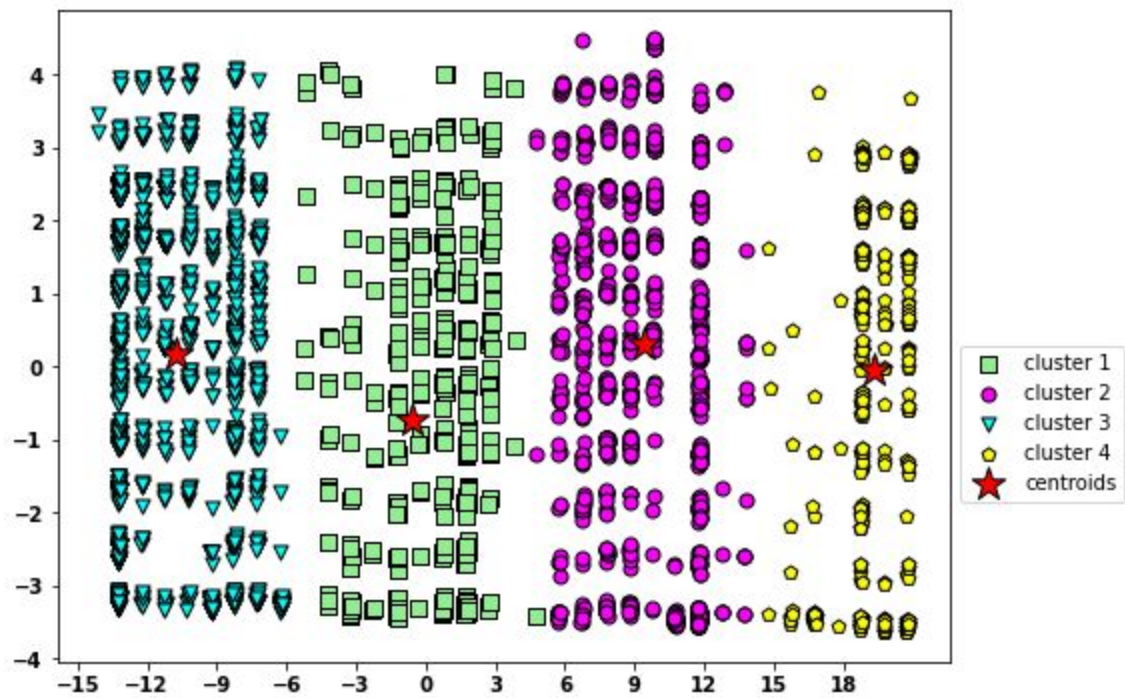
```
-0.55085953 -0.74063583 -> cluster 1
9.40536874 0.30668611 -> cluster 2
-10.74713911 0.17143439 -> cluster 3
19.28537754 -0.06350517 -> cluster 4
```

AgglomerativeClustering: In Agglomerative Clustering we form grouping and form a tree like structure :



Visualizing the clusters :I used PCA to reduce the data dimension to 2D for better visualization.

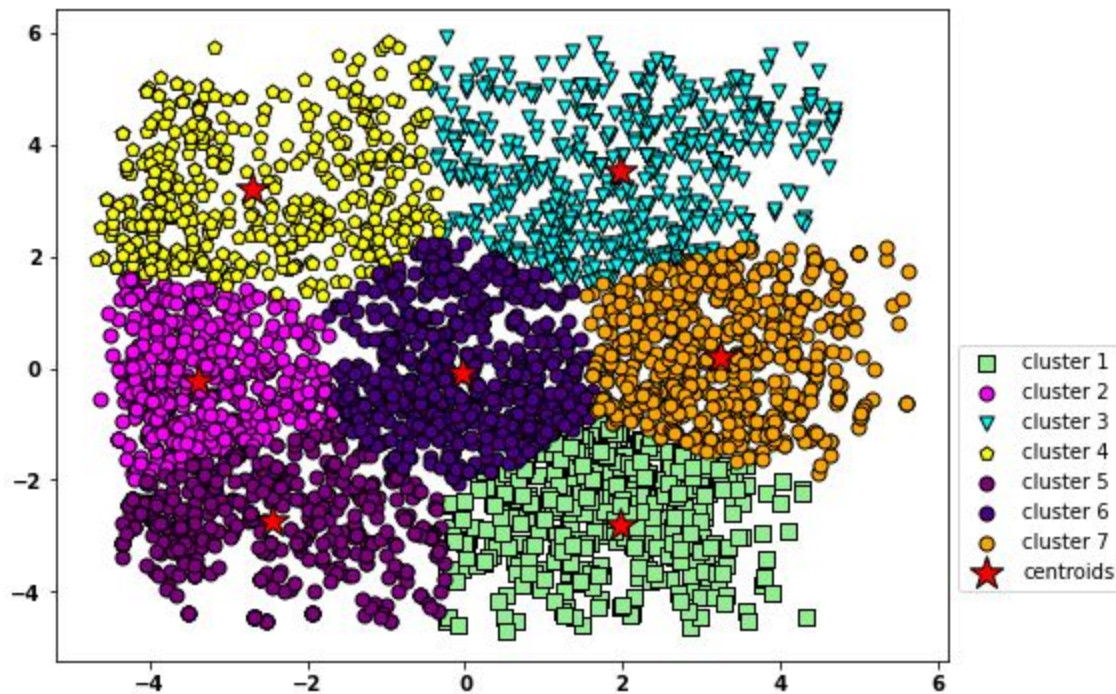
For Kmeans with k= 4



For Kmeans++

I used PCA to reduce the data dimension to 2D for better visualization.

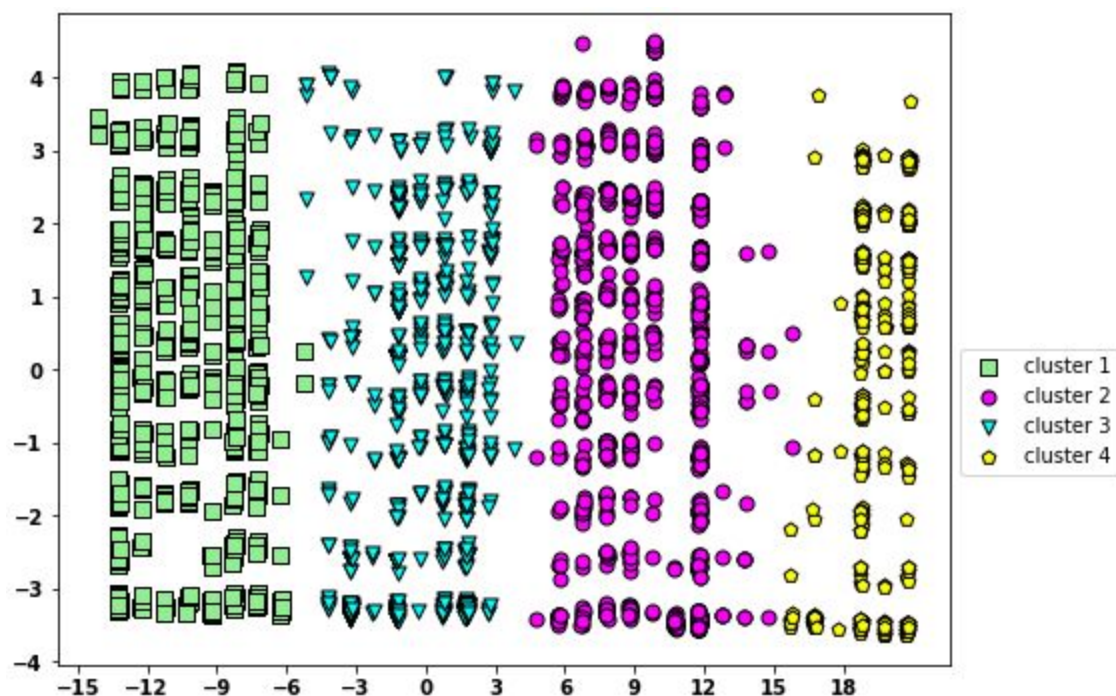
For Kmeans with $k=7$



For Agglomerative clustering

I used PCA to reduce the data dimension to 2D for better visualization.


$k=4$



Here in Agglomerative Clustering there is no centroid concept ,we form a group of similar clusters, better representation is with trees.(graph attached above)

Comparing my Cluster distribution with the true labels

Cluster	TRUE LABEL	Kmeans++
1	540	576
2	542	572
3	743	742
4	540	510
5	540	604
6	675	626
7	540	490



Contributions: Reshan Faraz(PhD19006) and Akhand Pratap Singh (MT20029) both contribute almost the same with coding and report preparation.