Assignment 1

Name: M Abdullah Awan

Roll No: 21L-5663

Section: BDS-4C

Course: Fundamentals Of Big Data

Q1

Pre Processing:

In pre processing stage I have first checked if there are any null values in the data or not. I then droped the ID column which is nominal and does not help us in giving any information for clustering. Then I did dicritization on age column into 3 bins of 3 different categories i.e young, middle and old and formeda new column named agebins. Then I did normalization using min max scalar on age and salary attribute to bring data in range of 0 to 1 and to convert big dominating values into small ones. Then I checked the correlation of data which showed me that age and salary are moderately positively correlated but I didn't droped one of the attribute to see their formation of clusters. Then I also encoded the values of some attributes like ordinal into 0 1 2 3. Following are the results of above mentioned techniques.

#Checking if there are any null values in data frame

data.isna().sum()

```
id          0
age         0
sex         0
region      0
salary      0
married     0
children    0
car         0
dtype: int64
```

#Checking the correlation

	age	salary	children
age	1.000000	0.752726	0.023572
salary	0.752726	1.000000	0.036761
children	0.023572	0.036761	1.000000

```
#Discretize the 'age' attribute into three bins
```

	age	agebins
0	48	MiddleAged
1	40	MiddleAged
2	51	Old
3	23	Young
4	57	Old
...	...	...
595	61	Old
596	30	Young
597	31	MiddleAged
598	29	Young
599	38	MiddleAged

600 rows × 2 columns

Q2

Selecting the Subsets:

The subsets I selected for clustering are:

Age and Salary

Salary and Children

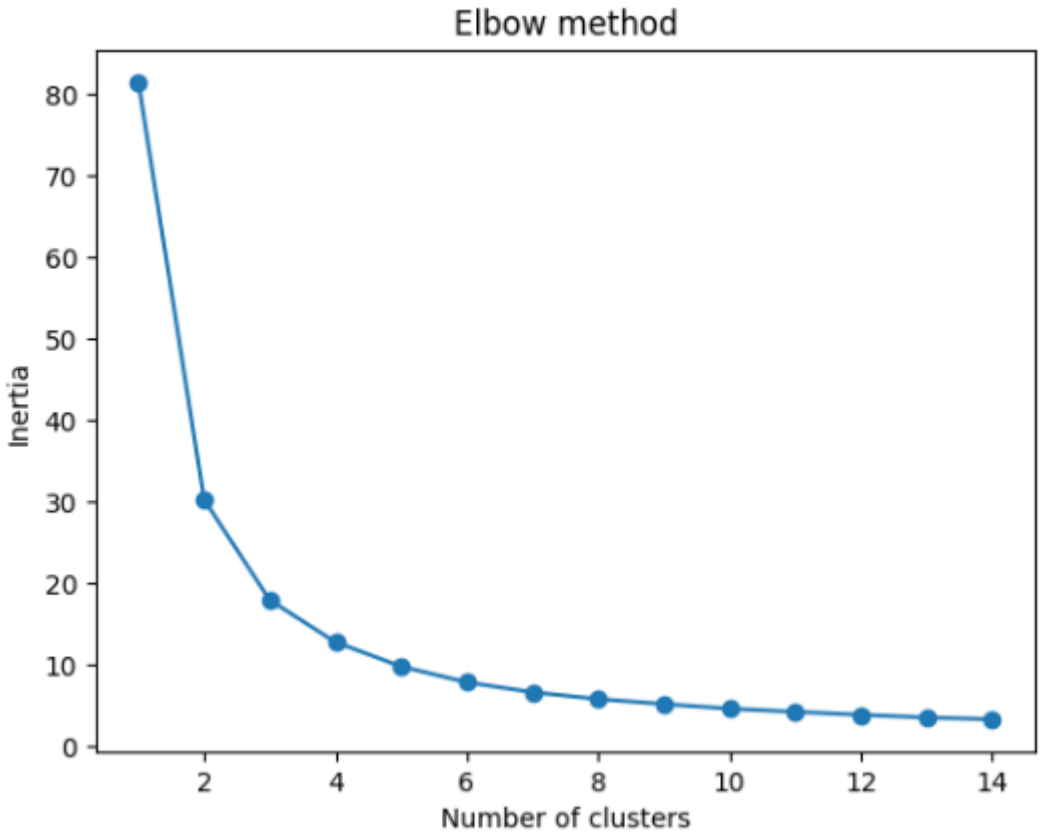
Region and Salary

Q3

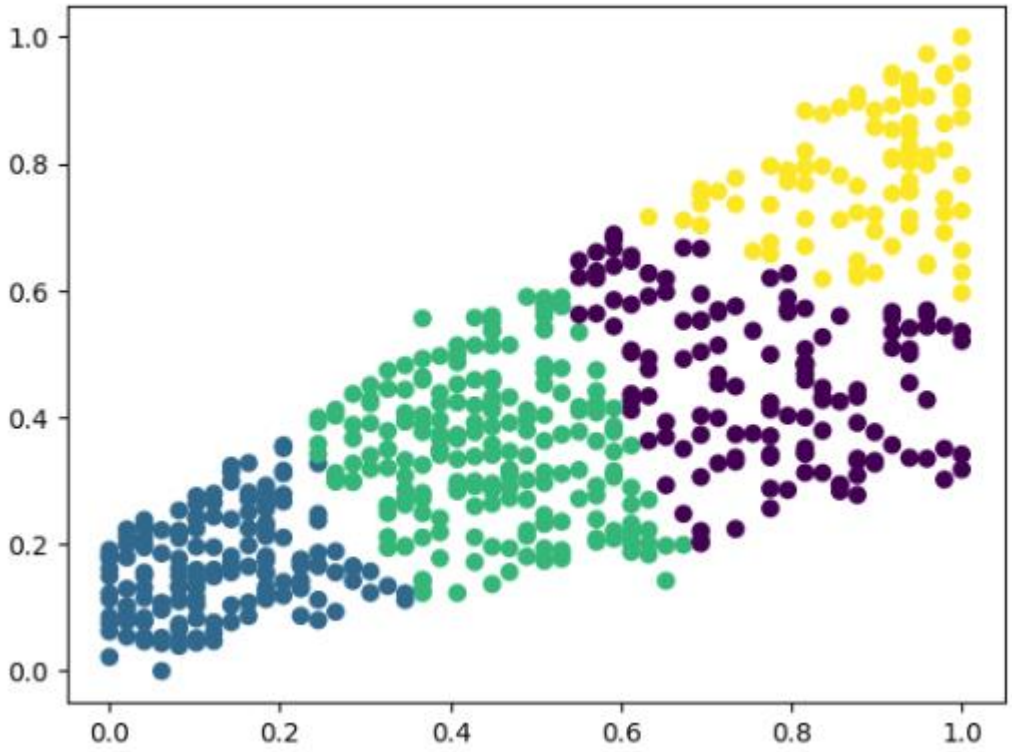
K Means Clustering:

Following are the results of k means clustering for different subsets.

#For Subset1



K means Clustering

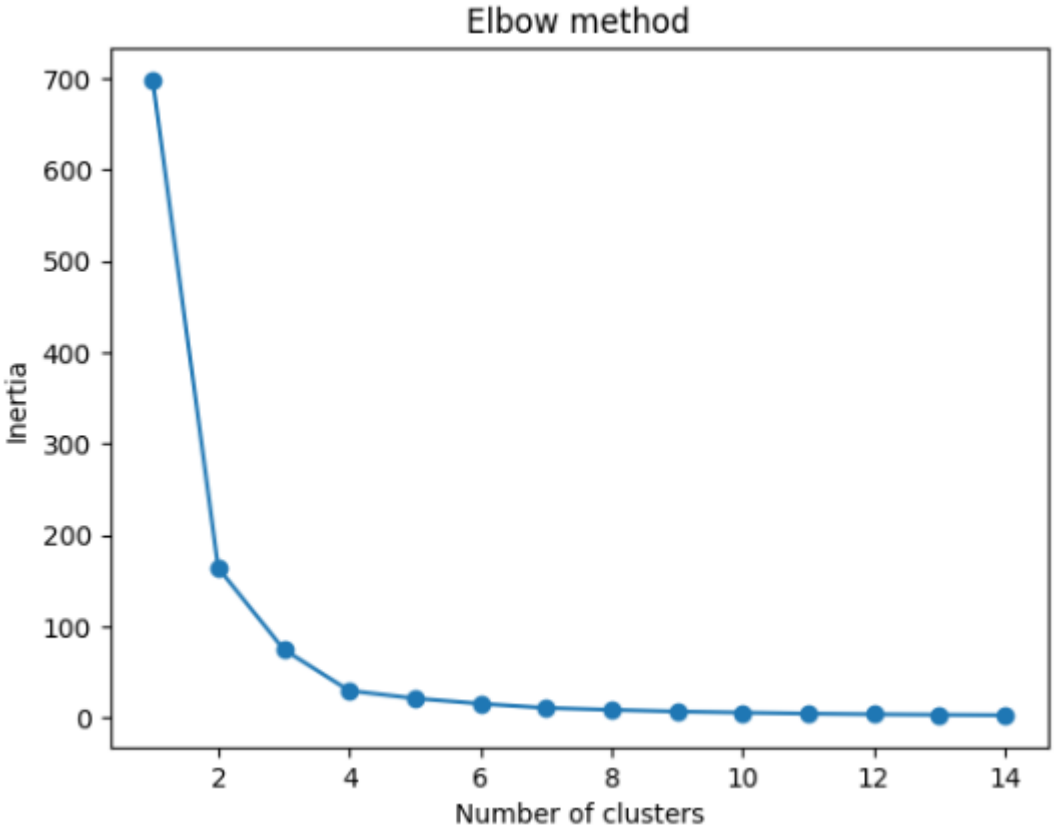


Analysis On different values of K

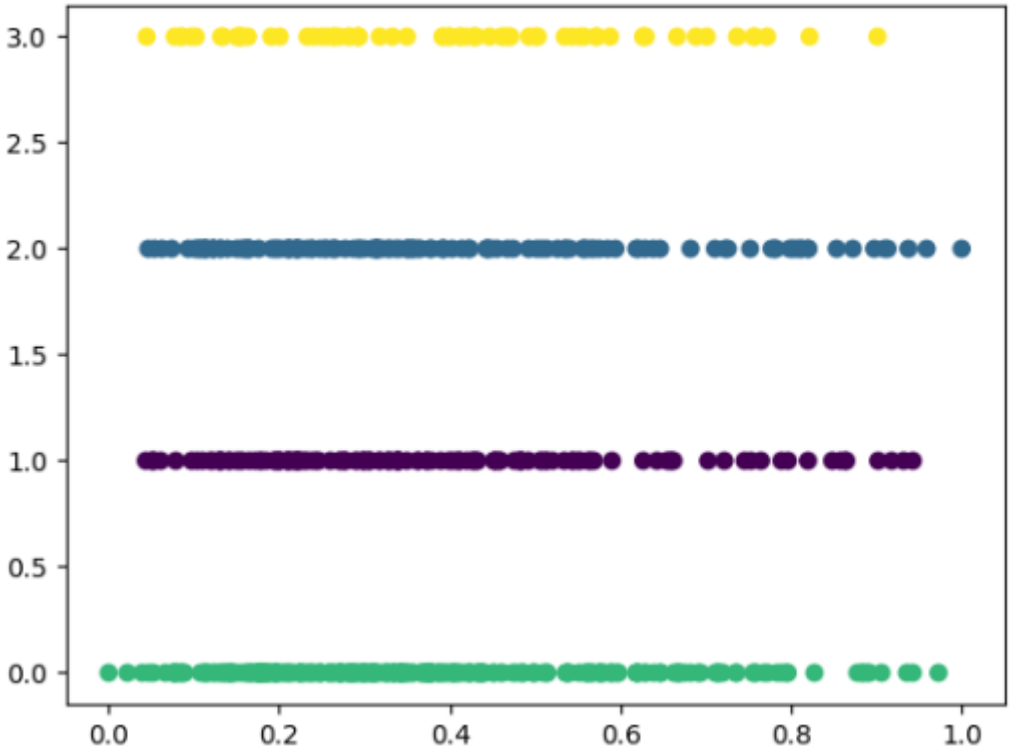
```
K =  2
Sum of square error =  30.23200360480138
Silhouette score =  0.5073406400014452
Iterations to convergence =  9
Time =  0.05508589744567871
---------------------------------------------------------------
K =  3
Sum of square error =  17.920831948779256
Silhouette score =  0.4549948702686407
Iterations to convergence =  5
Time =  0.06589651107788086
---------------------------------------------------------------
K =  4
Sum of square error =  12.774236009780495
Silhouette score =  0.4374267556576339
Iterations to convergence =  6
Time =  0.07052779197692871
---------------------------------------------------------------
K =  5
Sum of square error =  9.778215078523898
Silhouette score =  0.442348269610333
Iterations to convergence =  7
Time =  0.09198451042175293
---------------------------------------------------------------
K =  6
Sum of square error =  7.838935105813766
Silhouette score =  0.4403832133335851
Iterations to convergence =  16
Time =  0.07237720489501953
---------------------------------------------------------------
K =  7
Sum of square error =  6.619268628868602
Silhouette score =  0.4058593999265336
Iterations to convergence =  13
Time =  0.0787954330444336
---------------------------------------------------------------
```

As a result, I can say that as we increase the K the sse decreases but our elbow method shows that clustering should be performed for k = 4 as the inertia started to drop after that also the iterations of convergence are less relatively for k = 4

#For Subset2



K means Clustering

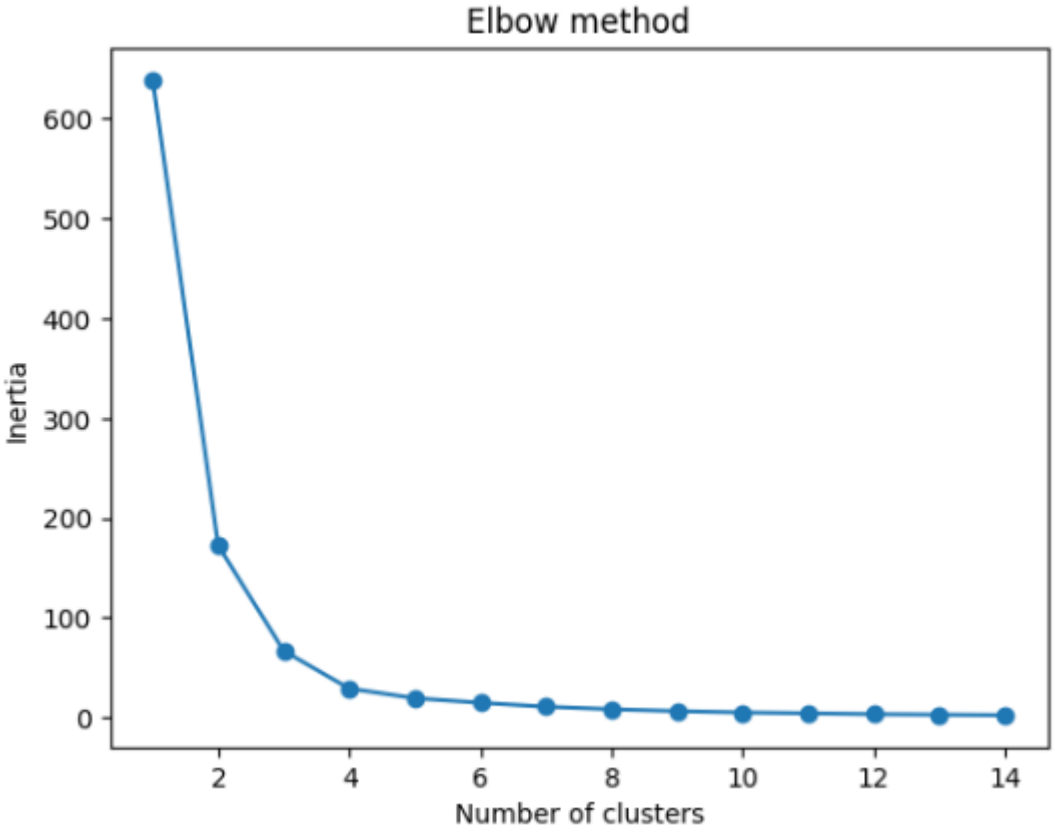


Analysis for different values of k

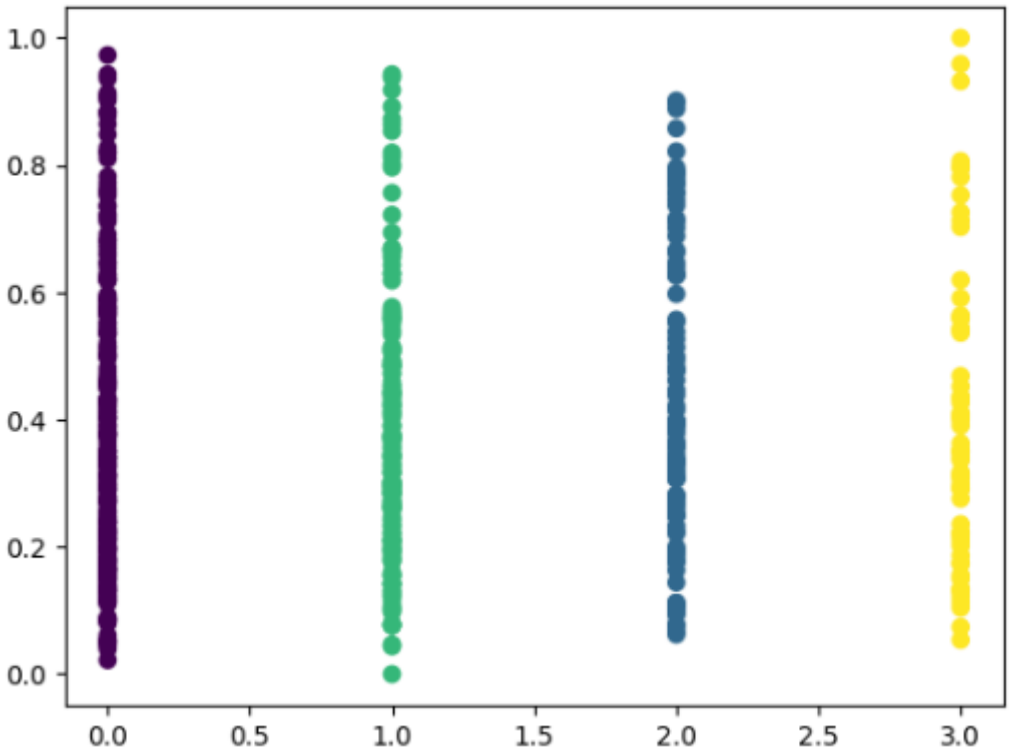
```
K =  2
Sum of square error =  163.77811306344714
Silhouette score =  0.6745770432613153
Iterations to convergence =  2
Time =  0.04313802719116211
----------------------------------------------------------------
K =  3
Sum of square error =  74.56802503260347
Silhouette score =  0.6899505922084933
Iterations to convergence =  2
Time =  0.044747114181518555
----------------------------------------------------------------
K =  4
Sum of square error =  29.455869385500627
Silhouette score =  0.7631819273447408
Iterations to convergence =  2
Time =  0.05311918258666992
----------------------------------------------------------------
K =  5
Sum of square error =  21.040178041428327
Silhouette score =  0.7002982609325543
Iterations to convergence =  3
Time =  0.07086014747619629
----------------------------------------------------------------
K =  6
Sum of square error =  15.246236974153174
Silhouette score =  0.6768057981612081
Iterations to convergence =  5
Time =  0.08310079574584961
----------------------------------------------------------------
K =  7
Sum of square error =  10.664547661060531
Silhouette score =  0.641718425465857
Iterations to convergence =  6
Time =  0.09171414375305176
----------------------------------------------------------------
```

According to elbow method we have to do clustering for k = 4 which also have the least iterations to convergence and relatively low sse

#For Subset 3



K means Clustering



Analysis for different values of k

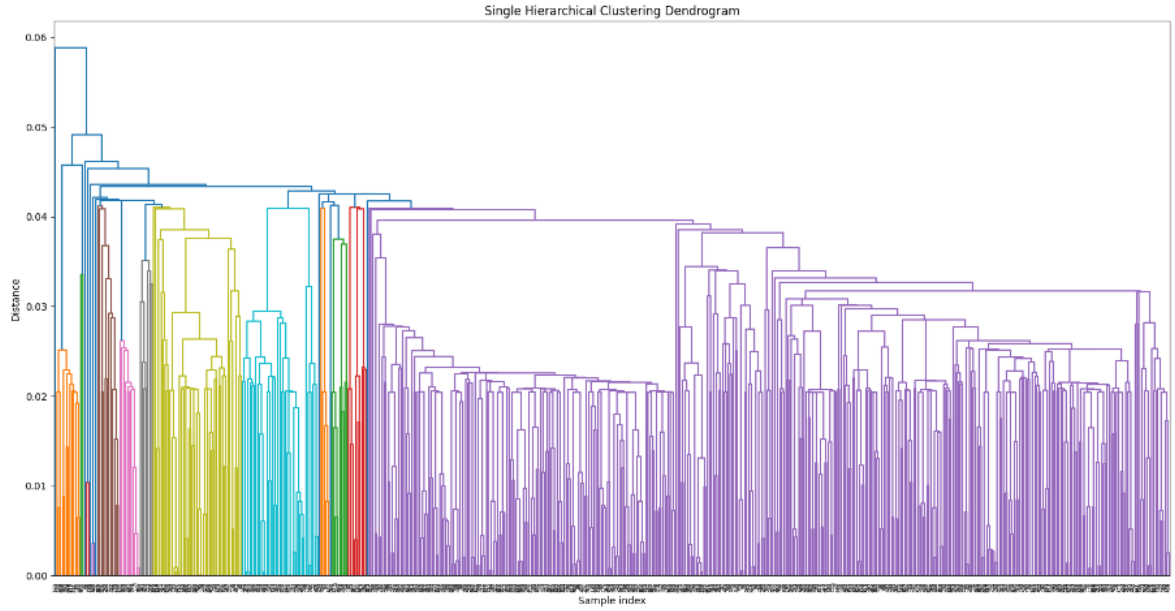
```
K =  2
Sum of square error =  172.2236712317982
Silhouette score =  0.6642639667338099
Iterations to convergence =  2
Time =  0.06429243087768555
---------------------------------------------------------------
K =  3
Sum of square error =  66.93623730793264
Silhouette score =  0.7048406024562126
Iterations to convergence =  2
Time =  0.05653190612792969
---------------------------------------------------------------
K =  4
Sum of square error =  29.244378820574482
Silhouette score =  0.7638263981227069
Iterations to convergence =  2
Time =  0.07020282745361328
---------------------------------------------------------------
K =  5
Sum of square error =  19.613178323520316
Silhouette score =  0.7021014616777228
Iterations to convergence =  7
Time =  0.07245731353759766
---------------------------------------------------------------
K =  6
Sum of square error =  15.044338578885322
Silhouette score =  0.6422448389108046
Iterations to convergence =  6
Time =  0.0777580738067627
---------------------------------------------------------------
K =  7
Sum of square error =  11.094383570634879
Silhouette score =  0.6284642730298938
Iterations to convergence =  5
Time =  0.0666961669921875
---------------------------------------------------------------
```

According to elbow method we have to do clustering for k = 4 which also have the least iterations to convergence and relatively low sse

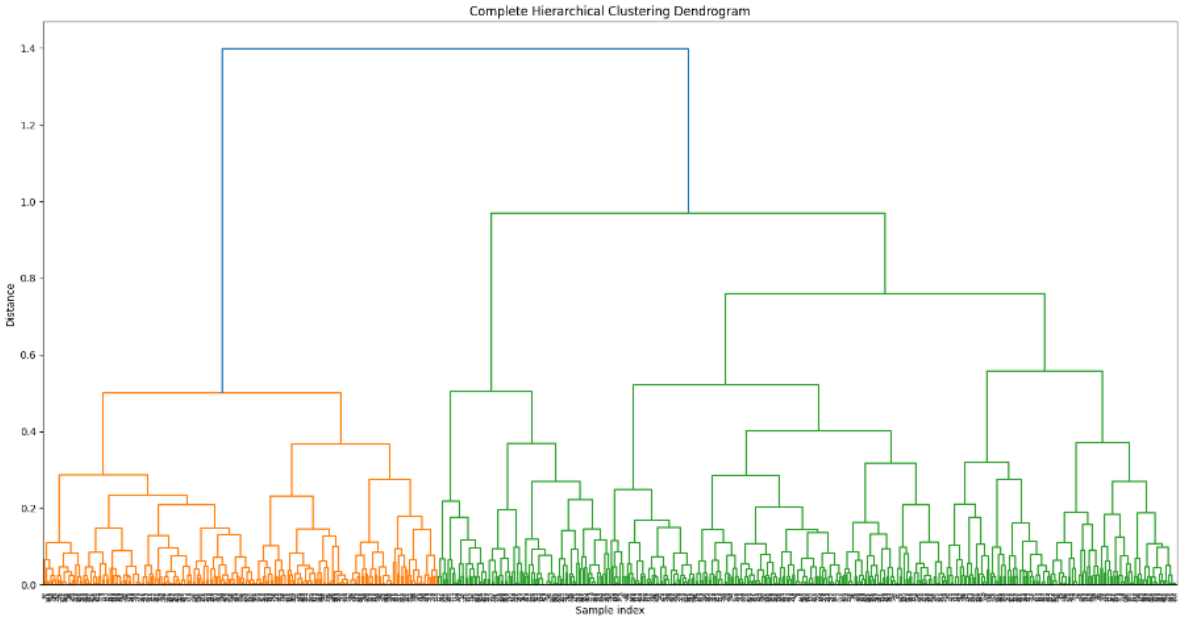
Q4

Hierarchical Clustering (single link, complete link, average link) and DBSCAN:

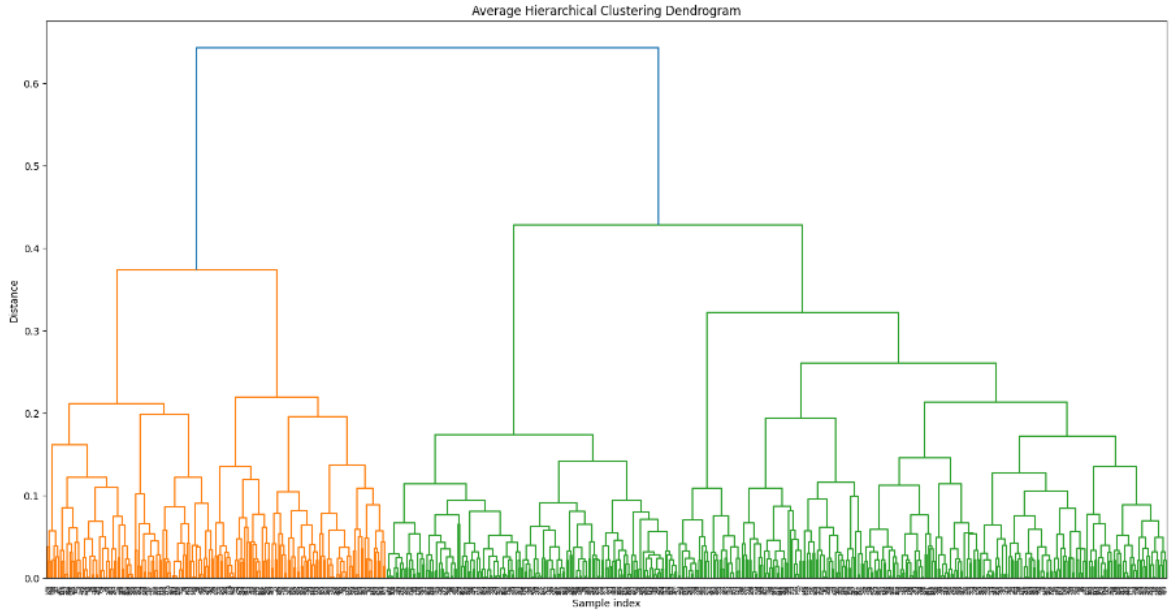
#For subset1



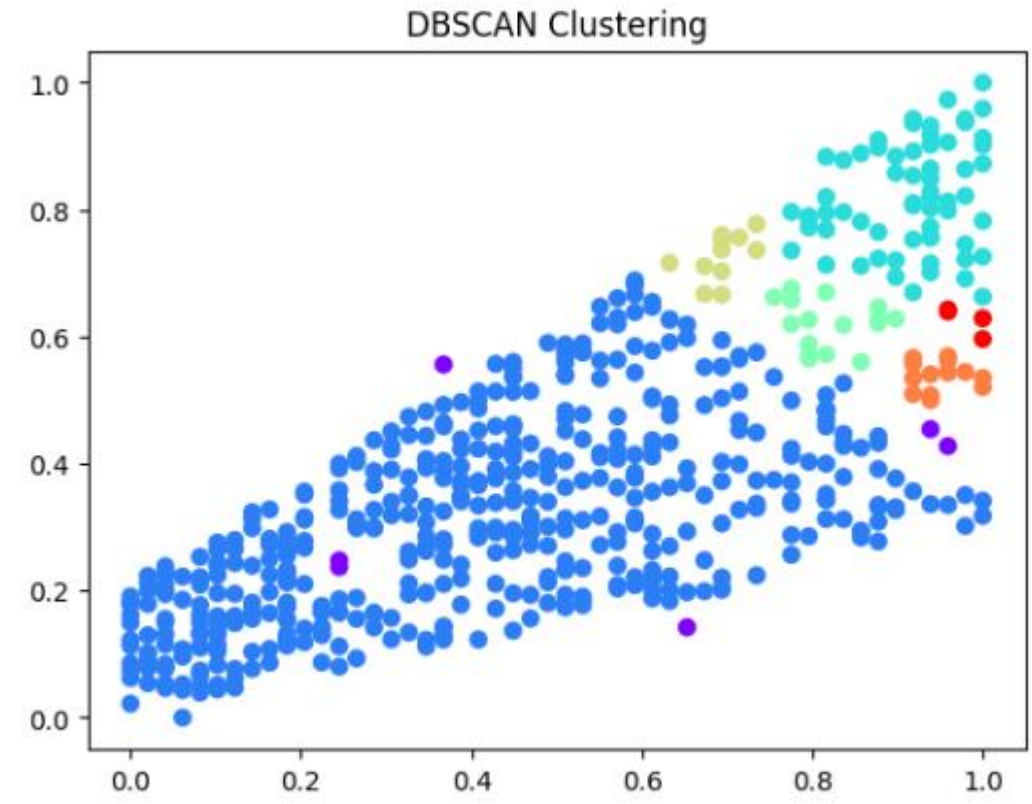

```
Time =  8.4323091506958
```




```
Time =  8.947044134140015
```



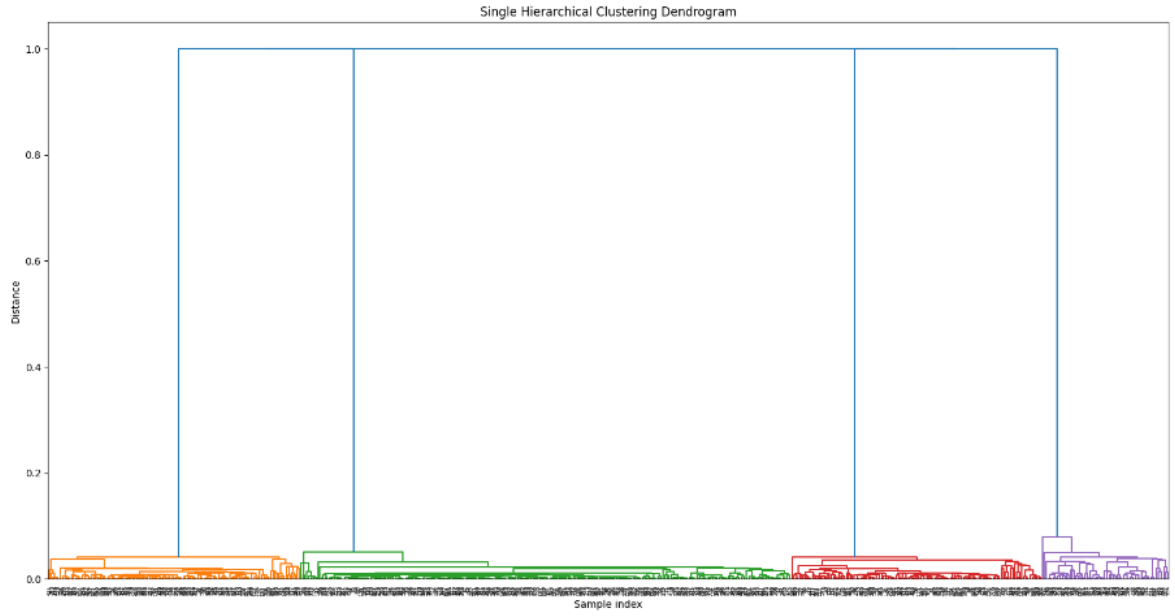

```
Time =  9.540706157684326
```



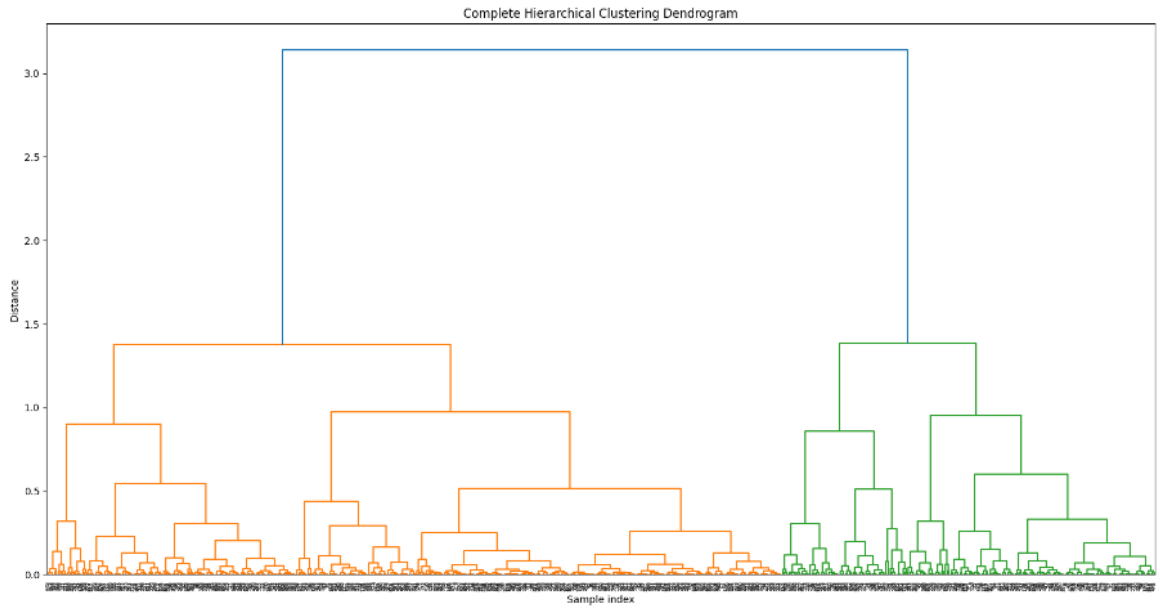

```
Time =  0.25728511810302734
```

In dbscan we can see that clustering is not that good and it is changed when we change the parameters because the points are merged with each other a lot. Min points for dbscan are derived by formula 2*dimensions and eps is derived from built in function with library nearest neighbor

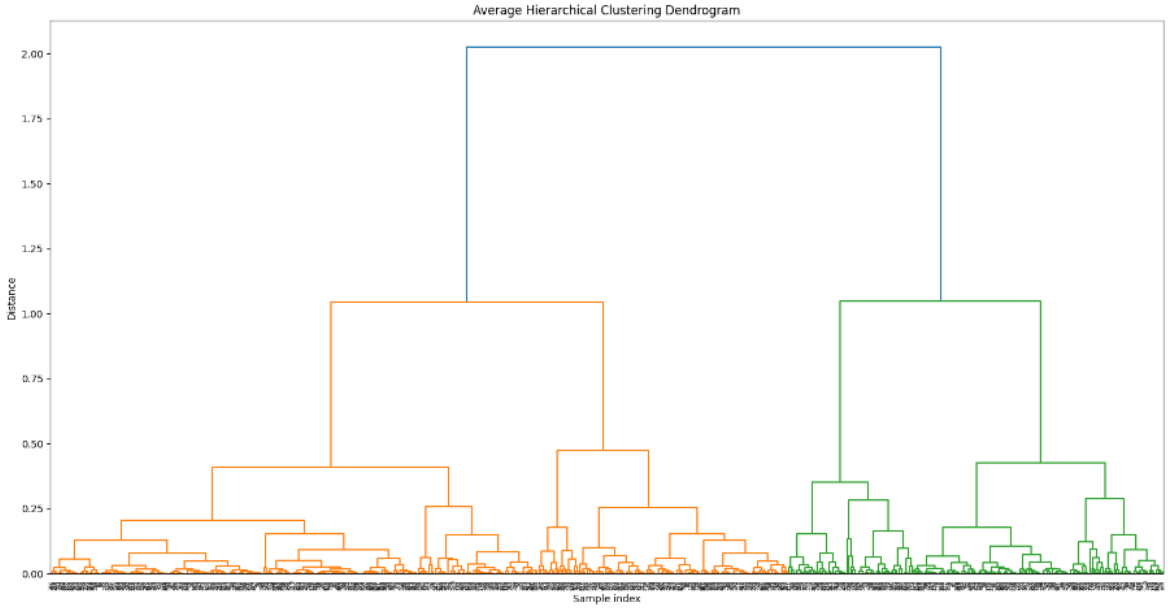
#For Subset 2



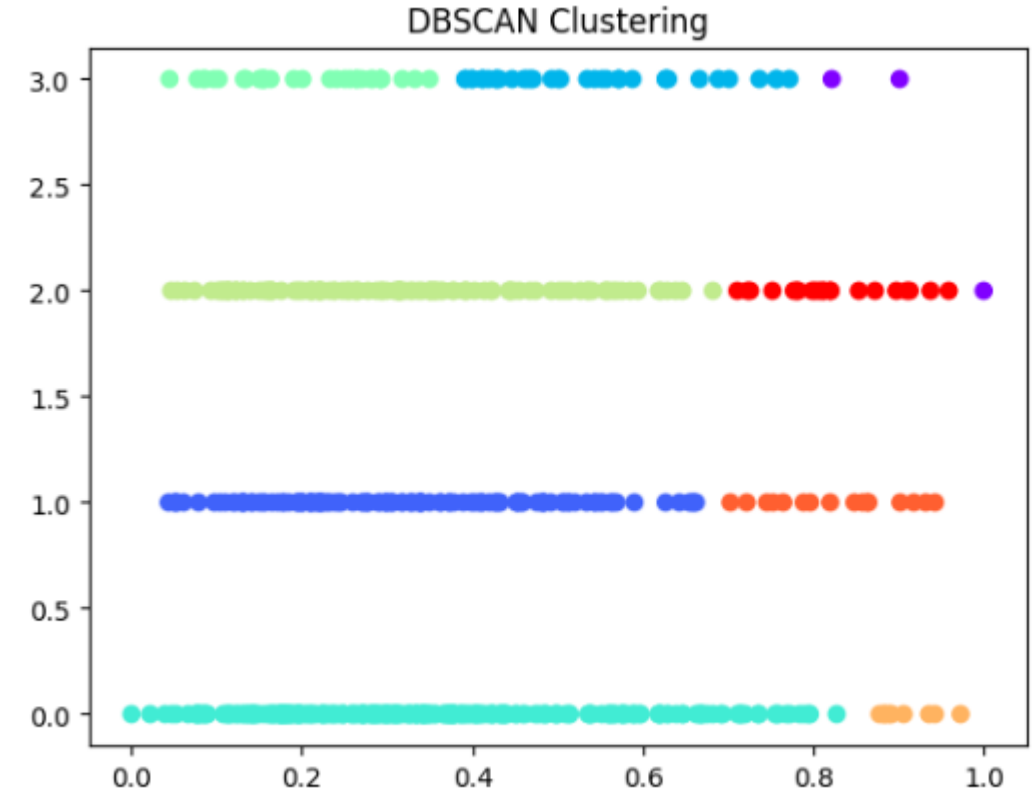

```
Time =  8.476863384246826
```




```
Time =  8.596288442611694
```



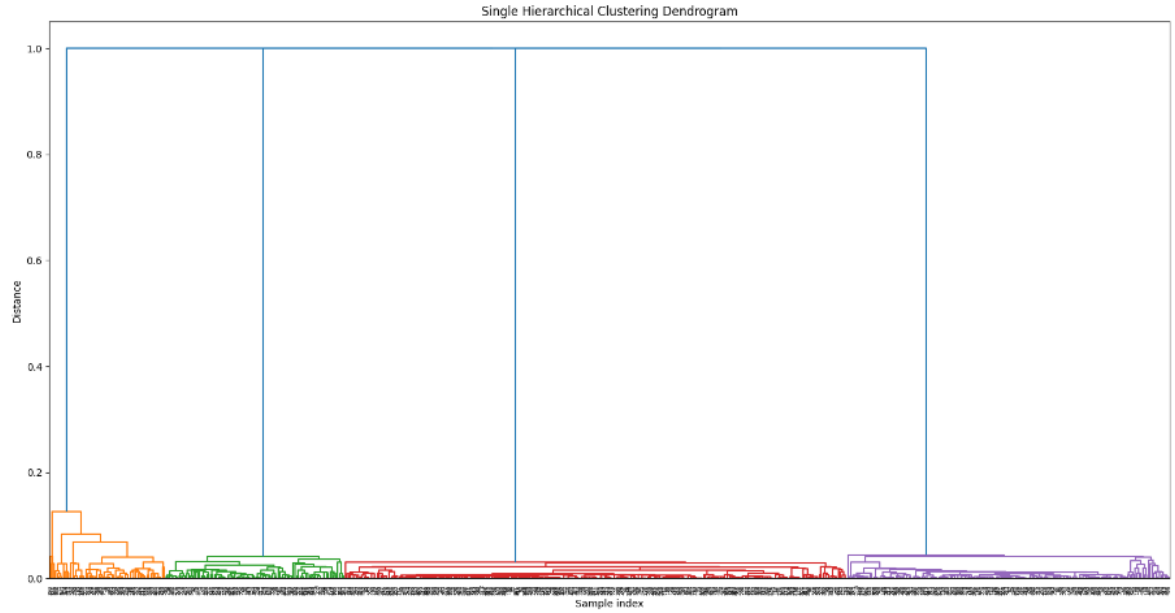

```
Time =  8.208831310272217
```



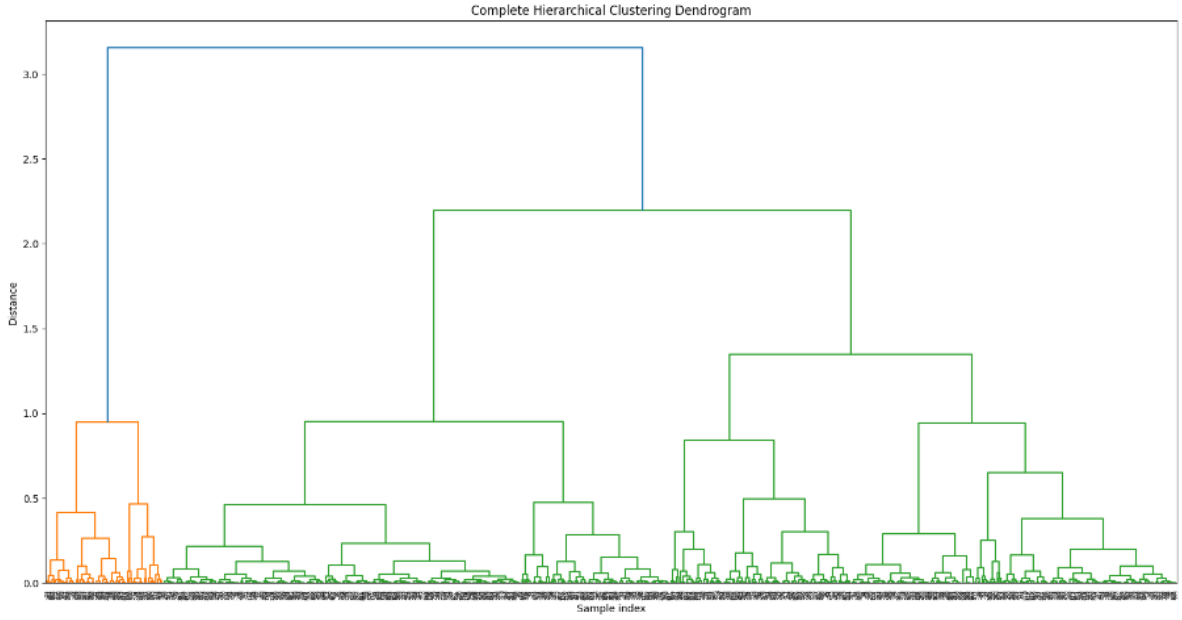

```
Time =  0.1522226333618164
```

We can see that dbscan clustering performed for eps 0.04 does not give good result but increasing eps gives good cluster also the other hierarchical technique takes much time

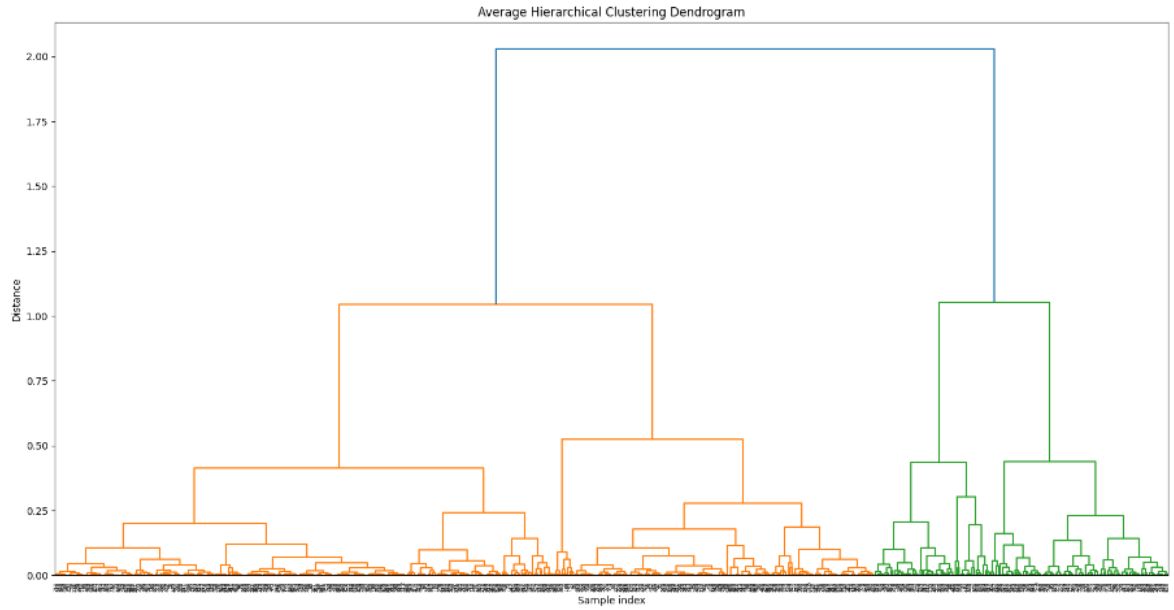
#For Subset 3



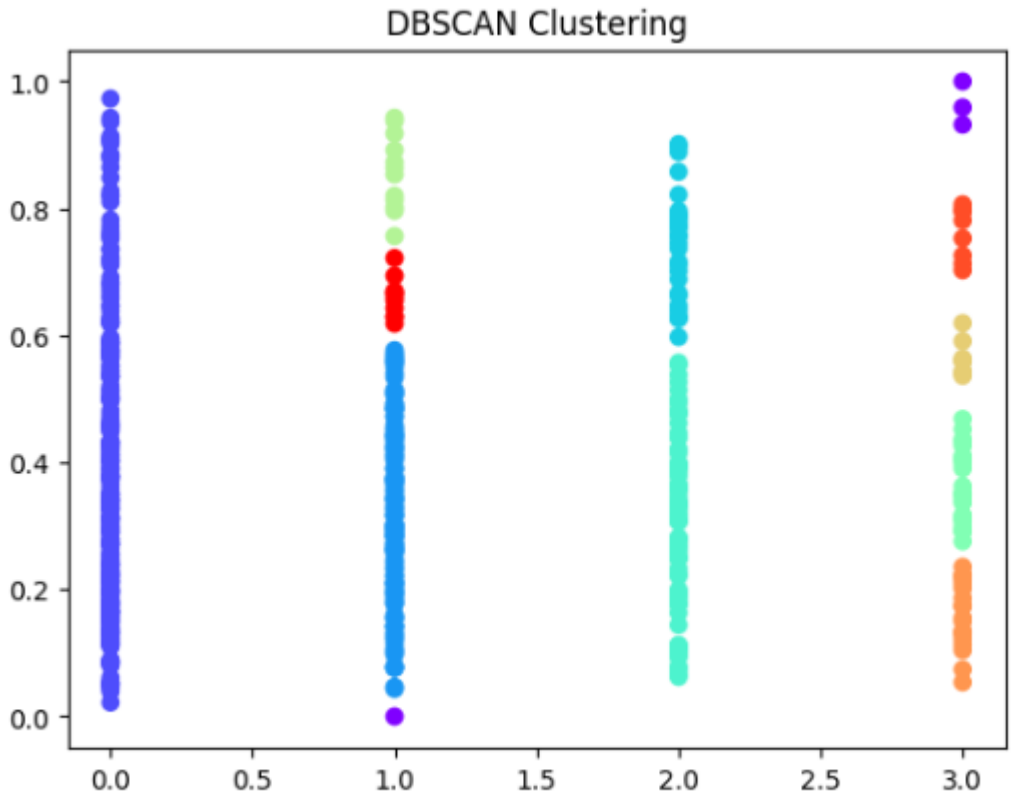

```
Time =  8.190628051757812
```




```
Time =  8.691598892211914
```




```
Time =  9.480721712112427
```




```
Time =  0.15239882469177246
```

We can see that dbscan clustering performed for eps 0.04 does not give good result but increasing eps gives good cluster also the other hierarchical technique takes much time

Conclusion:

As a conclusion I can say that the more suitable mean of doing clustering for this data is k means as it takes less time and gives good clusters for specific value of K obtained from elbow method. K means algorithm takes least time to perform clustering as compare to other hierarchical clustering techniques which is shown above. The single complete and average hierarchical clustering techniques takes much more time and are not that good for clustering. In case of dbscan the clusters formed are not that good with the parameters generated on the base of data but if I change the parameters the clusters formation changes also it takes less time than the other hierarchical techniques but more than k means. So, I can say that the best method for doing clustering on this data is K means algorithm which is efficient and makes good clusters which can be easily visualized from above graphs.