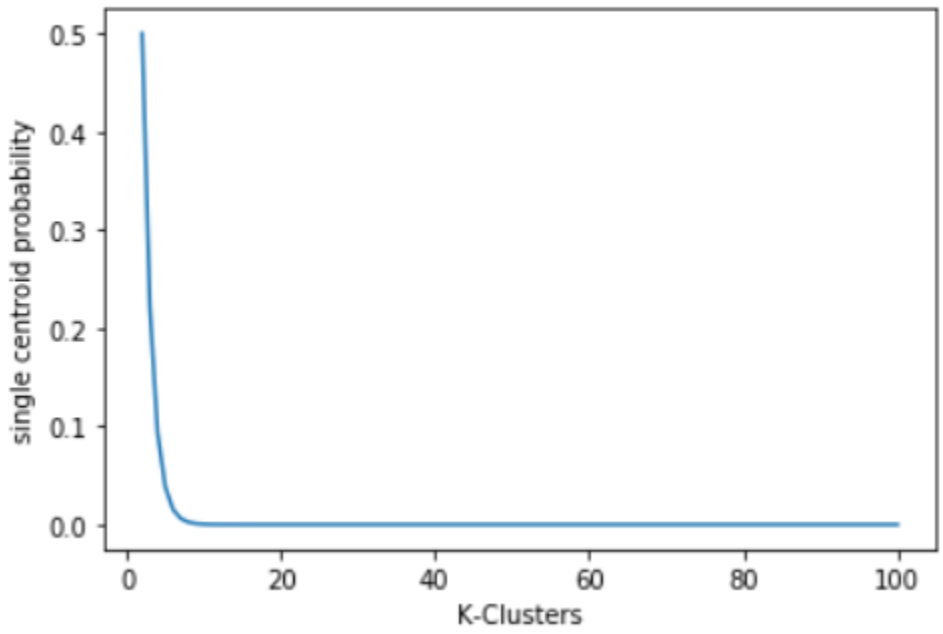
4)

A) For k= 2 to 100 and n=10

P=(2K!) / (K^K)

```
: [<matplotlib.lines.Line2D at 0x21c02a44a88>]
```



B)

For size 2K

P=(2K!) / (K^K)

When k=10

P= 0.00072576

When k=100

P= 1.8665243e-42

When k=1000

P= 0 or undefined

7)

Option 2 is the right answer more centroids should be present in less denser region to reduce the MSE. Reason being the sparsely populated area are distributed far from each other to reduce MSE we would require more centroid to reach all the points

11)

If the SSE is low for one variable in all the clusters, it means that the variable doesn't contribute much to cluster formation and is most likely a constant value.

If the SSE is low for just one cluster it is most helpful for defining that cluster

If the SSE is High for all clusters it is mostly likely not useful for any cluster formation and can be considered as noise

If the SSE is High for just one cluster means it does not add much to the cluster formation of which it belongs

The per variable SSE information is used to improve your clustering by eliminate the low SSE & high SSE for all the cluster as they don't add much to cluster formation as mentioned above in case1 & case3 as they are constants and do not add much value or noise making it hard to compute the clusters.

17)

{6, 12, 18, 24, 30, 42,48}

1)

1) {18, 45}

Cluster 1 :

(18-6)^2+(18-12)^2+(18-18)^2+(24-18)^2+(30-18)^2

Error=360

Cluster 2:

(45-42)^2+(48-45)^2

Error = 18

Total Error=378

2) {15, 40}

Cluster 1 :

(15-6)^2+(15-12)^2+(15-18)^2+(15-24)^2

Error=180

Cluster 2:

(40-30)^2+(40-42)^2+(40-48)^2

Error = 168

Total Error=348

2) Yes, the above centroids are stable as they are far from each other

3) the two clusters are

C1: {6, 12, 18, 24, 30}

C2: {42,48}

4) For single link the distance between the centroids are greater compared to that of kmeans. Hence single link provides more natural clustering

5) natural clustering here means the contiguous cluster formed by the single link.

6) The k-means work if the cluster are well separated i.e. its objective is reducing the MSE.If the K-means strategy is followed it forms unnatural clusters

22)

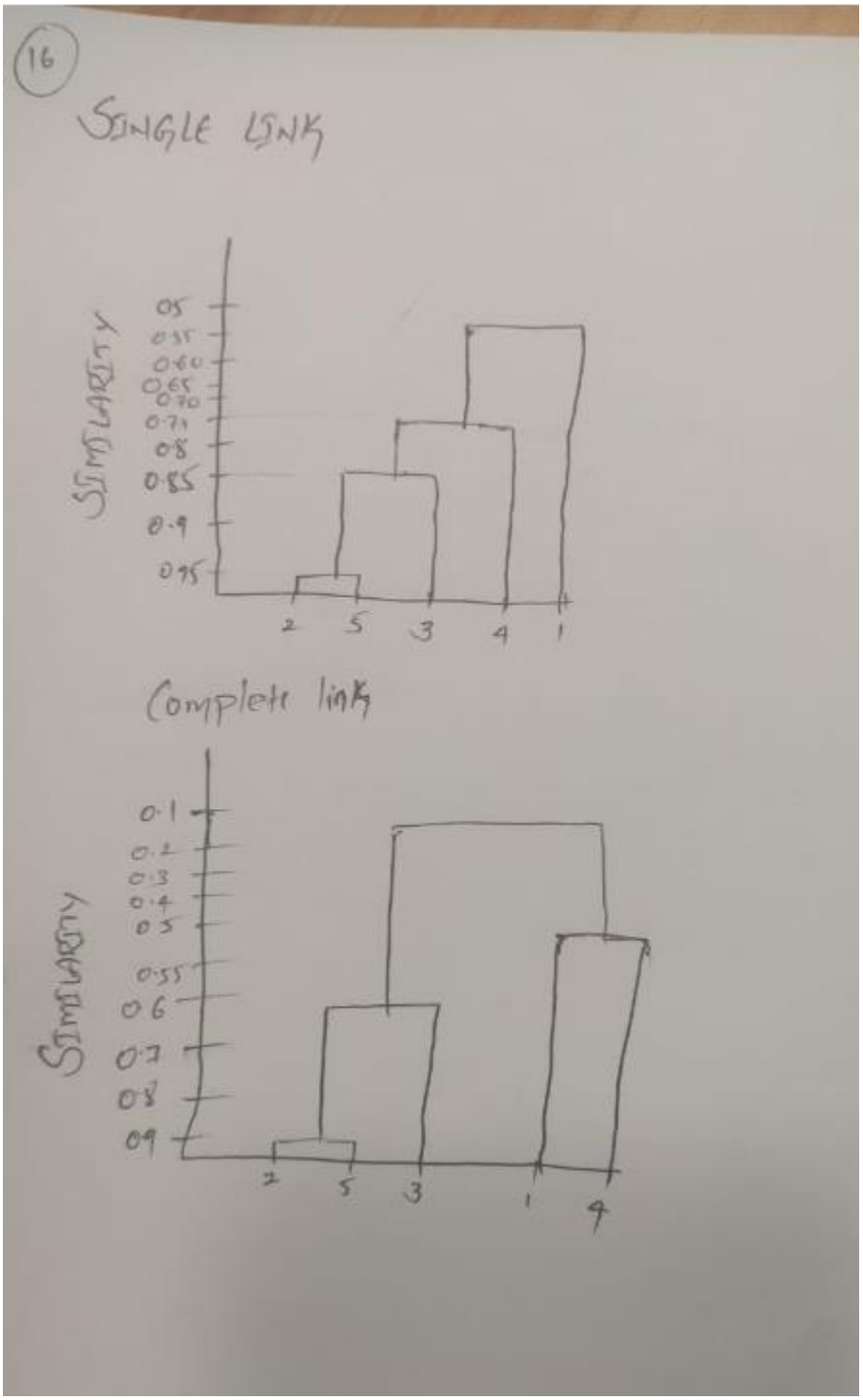
1) Uniform distanced points are at same distance from each other there by forming no cluster

But the points sampled from uniform distribution can for clusters of high or low density

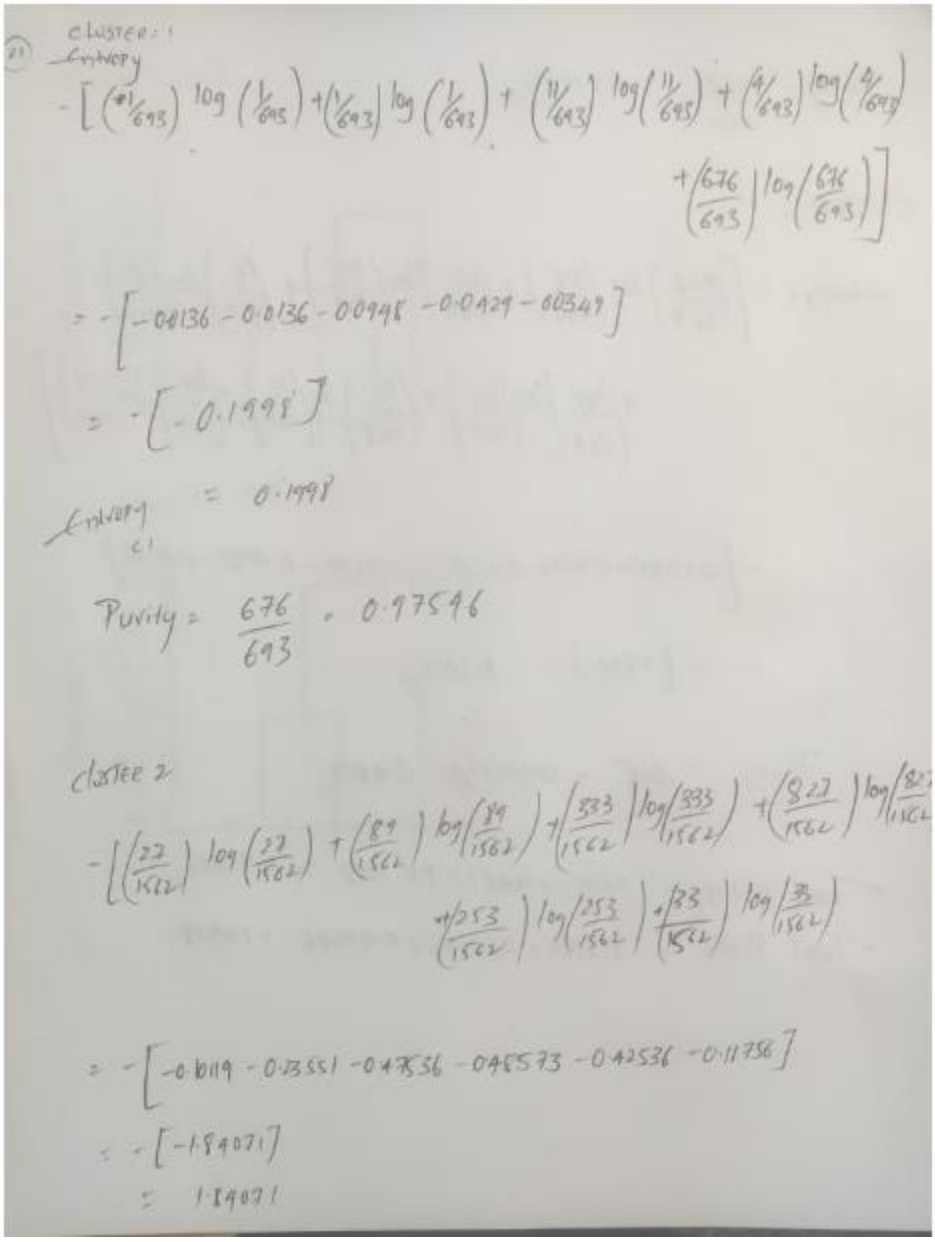
2) from the above reasoning the points sampled from uniform distribution has smaller SSE as it forms cluster.

3) For the uniform spaced data using DBSCAN everything will be considered as a single cluster and there is high chance it will be considered as noise. For the random data using DBSCAN it forms clusters with high/Low density.

16)


16
SINGLE LINK
SIMILARITY
05
0.55
0.60
0.65
0.70
0.75
0.8
0.85
0.9
0.95
2
5
3
4
1
Complete link
SIMILARITY
0.1
0.2
0.3
0.4
0.5
0.55
0.6
0.7
0.8
0.9
2
5
3
1
4

21)


cluster: 1
Entropy
= -[-0.0136 - 0.0136 - 0.0098 - 0.0429 - 0.0349]
= -[-0.1998]
= 0.1998
Entropy
c1
Purity = 676/693 = 0.97546
cluster 2
= -[-0.0119 - 0.13551 - 0.17536 - 0.08573 - 0.12536 - 0.11756]
= -[-1.84071]
= 1.84071

Purity = $\frac{827}{1562}$ = 0.52944 //

CLUSTER-3

Entropy = $-\left[\left(\frac{326}{949}\right)\log\left(\frac{326}{949}\right) + \left(\frac{465}{949}\right)\log\left(\frac{465}{949}\right) + \left(\frac{8}{949}\right)\log\left(\frac{8}{949}\right)\right.$

$\left. + \left(\frac{105}{949}\right)\log\left(\frac{105}{949}\right) + \left(\frac{16}{949}\right)\log\left(\frac{16}{949}\right) + \left(\frac{29}{949}\right)\log\left(\frac{29}{949}\right)\right]$

$= -\left[-0.5295 - 0.5092 - 0.0580 - 0.3514 - 0.0993 - 0.1532\right]$

$= +\left(1.6961\right) = 1.6961$ //

Purity = $\frac{465}{949}$ = 0.4879 //

Total Entropy: $\left(\frac{693}{3204}\right)(0.1998) + \left(\frac{1562}{3204}\right)(1.84071) + \left(\frac{949}{3204}\right)(1.8961)$

$= 0.043321 + 0.8774 + 0.5023$

$= 1.44291$

Total purity:

$\left(\frac{693}{3204}\right)\left(\frac{676}{693}\right) + \left(\frac{827}{1562}\right)\left(\frac{1562}{3204}\right) + \left(\frac{949}{3204}\right)\left(\frac{465}{949}\right)$

$= 0.2109 + 0.2581 + 0.1451$

$= 0.6141$