Homework 5

CS 57300

(Taking one extension day)

Submitted by:-

Name : Rashmi Soni

PUID: 0029136200

(Note: The colors used in the graph denotes the following clusters:

0:"black",

1:"blue",

2:"darkgray",

3:"darkgreen",

4:"darkred",

5:"purple",

6:"orange",

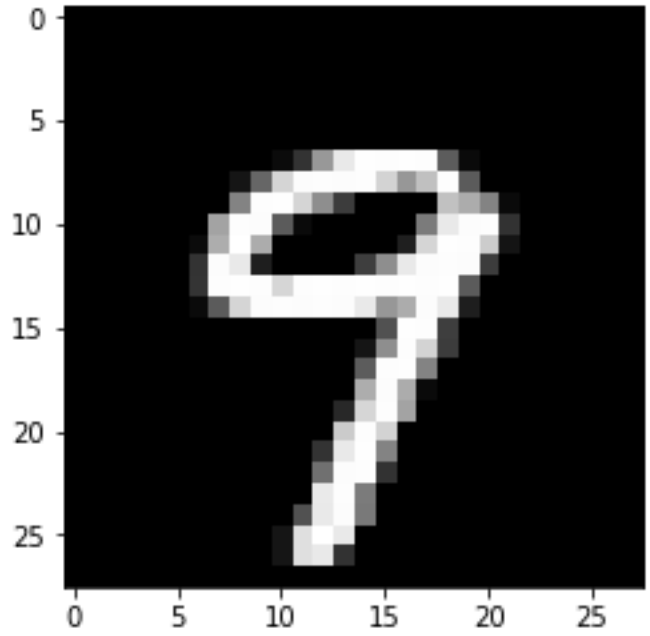
7:"yellow",

8:"chocolate",

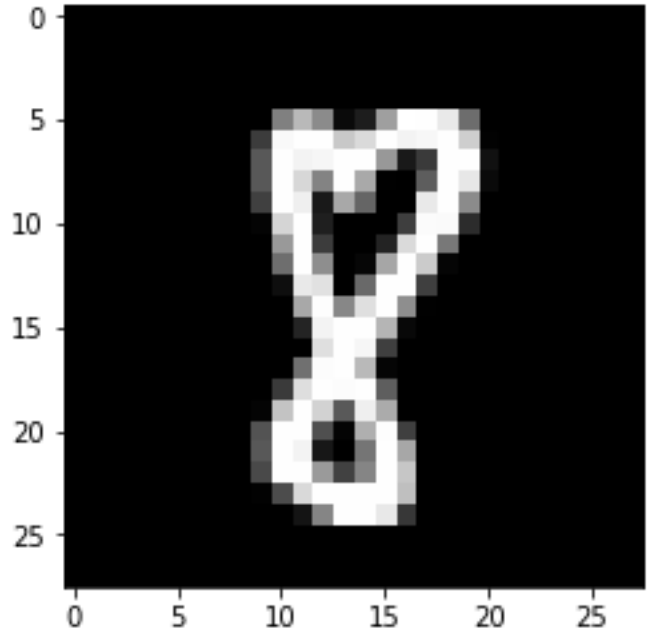
9:"deeppink")

A. Exploration

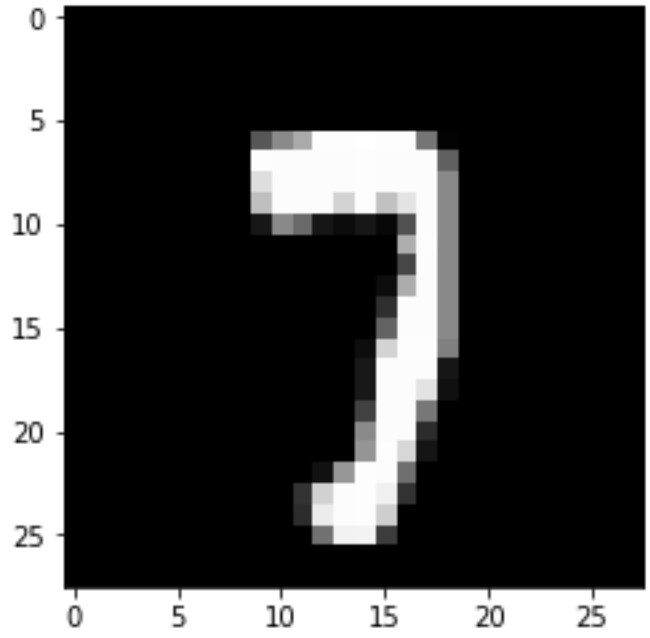
Digit 9:-



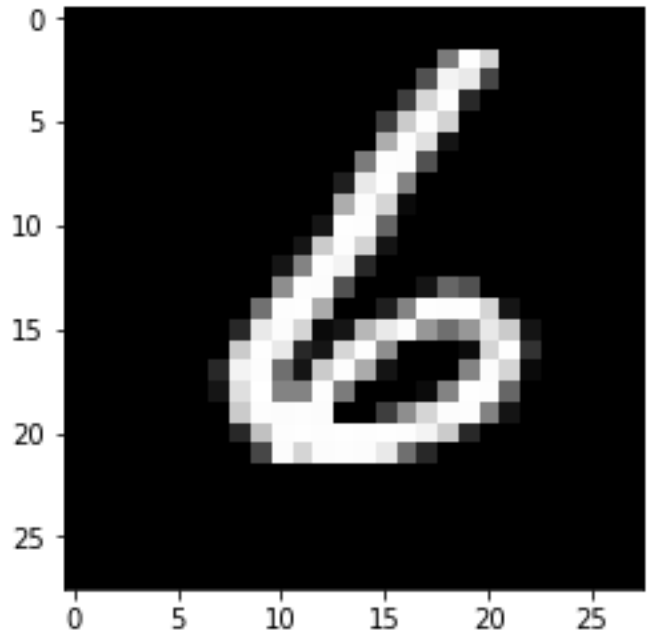
Digit 8:-



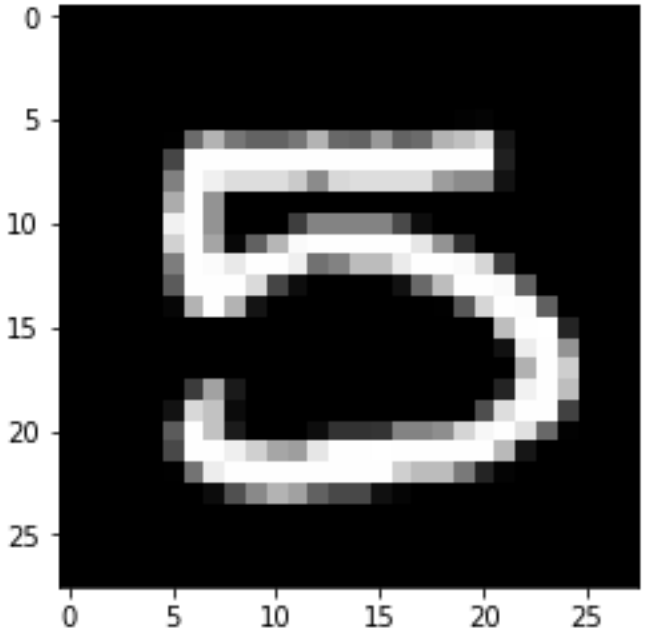
Digit 7:-



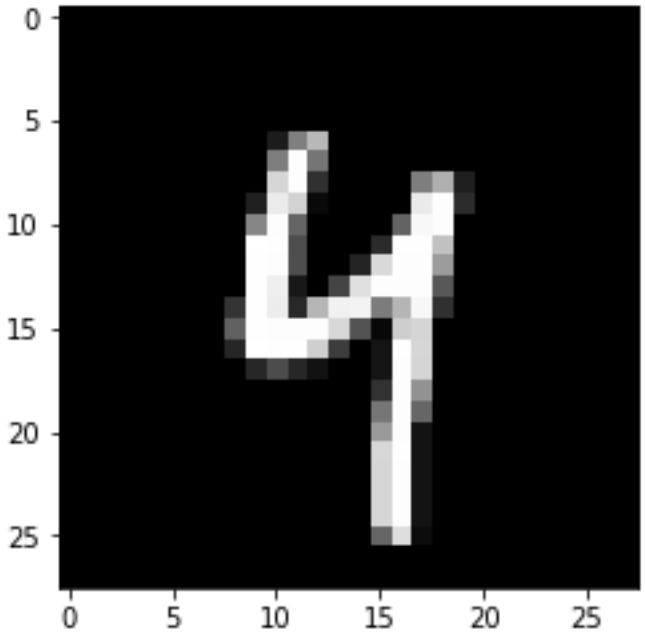
Digit 6:-



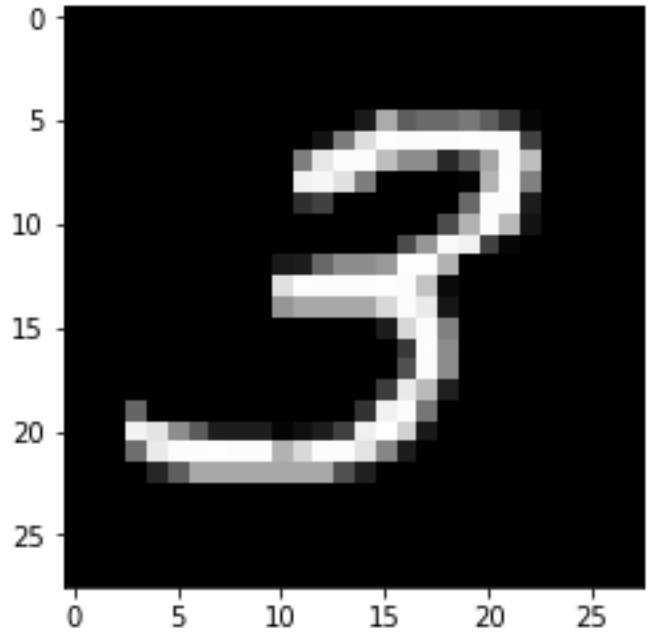
Digit 5:-



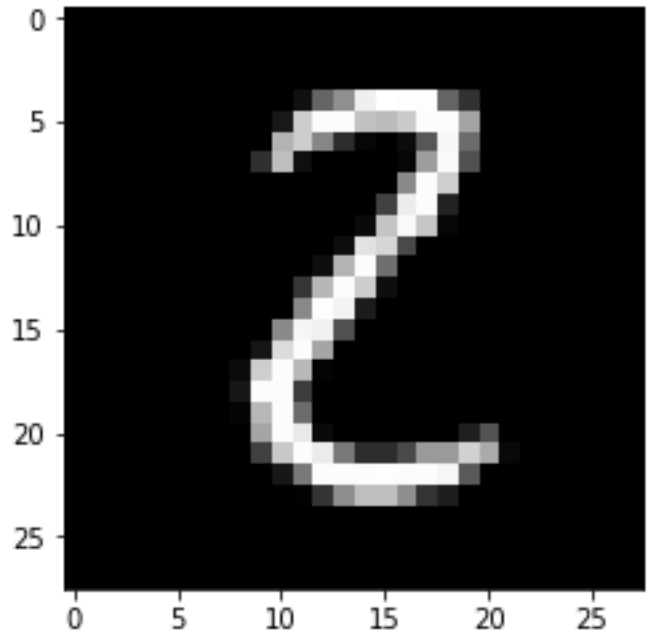
Digit 4:-



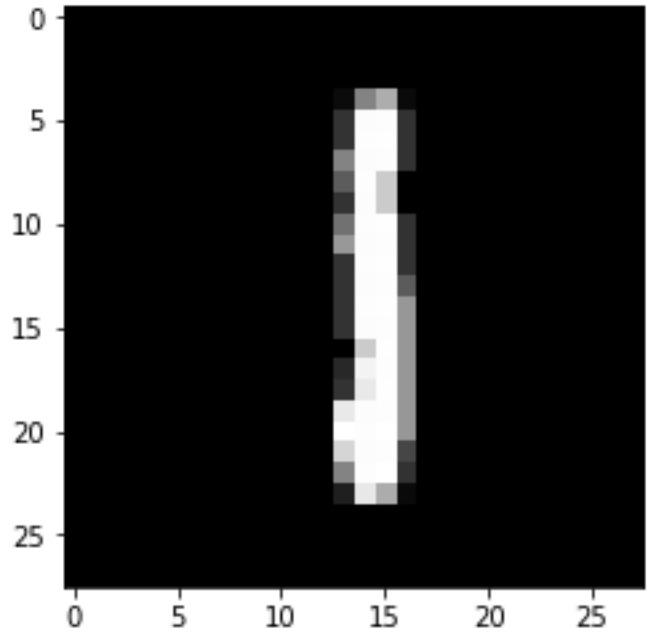
Digit 3:-



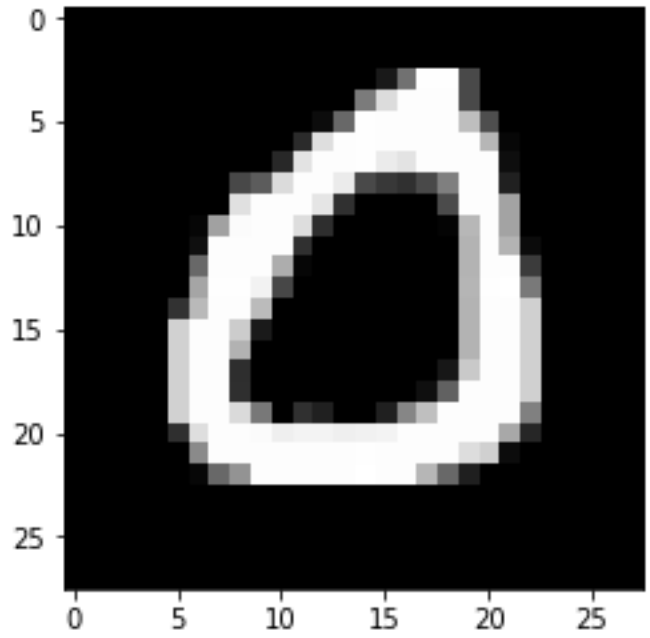
Digit 2:-



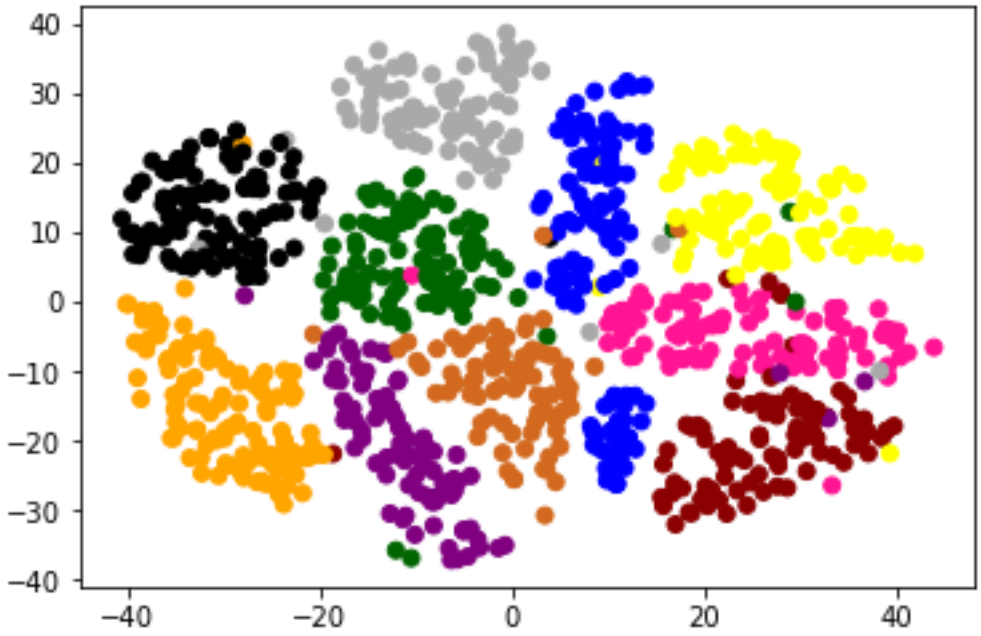
Digit 1:-



Digit 0:-



2. Exploration:-



B. Analysis of k-means :

1.)

Values of K= [2, 4, 8, 16, 32]

WC on full data= [8983224.0424403436, 4215072.7411717102, 1904472.4816357121, 896153.54493203235, 421654.26141275774]

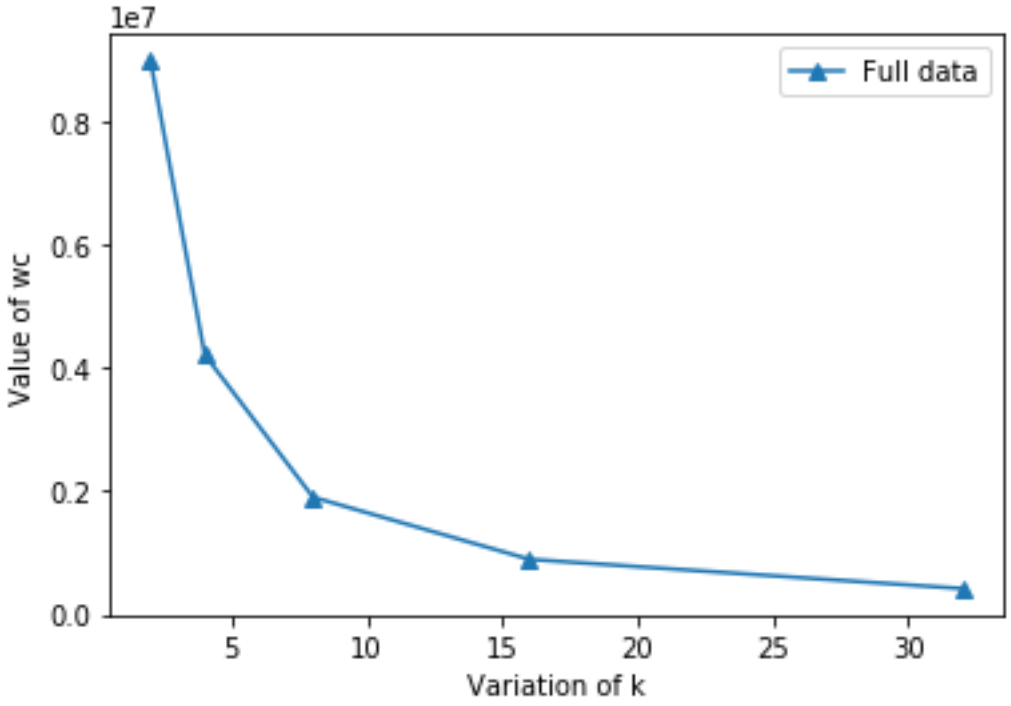
WC on 2467 data= [4211155.6885079658, 623865.31116823317, 416469.4187857986, 244363.99576883996, 87659.651405968369]

WC on 67 data= [340372.41942807229, 223070.1212773922, 92908.685225715511, 51322.16095098418, 25361.416668169433]

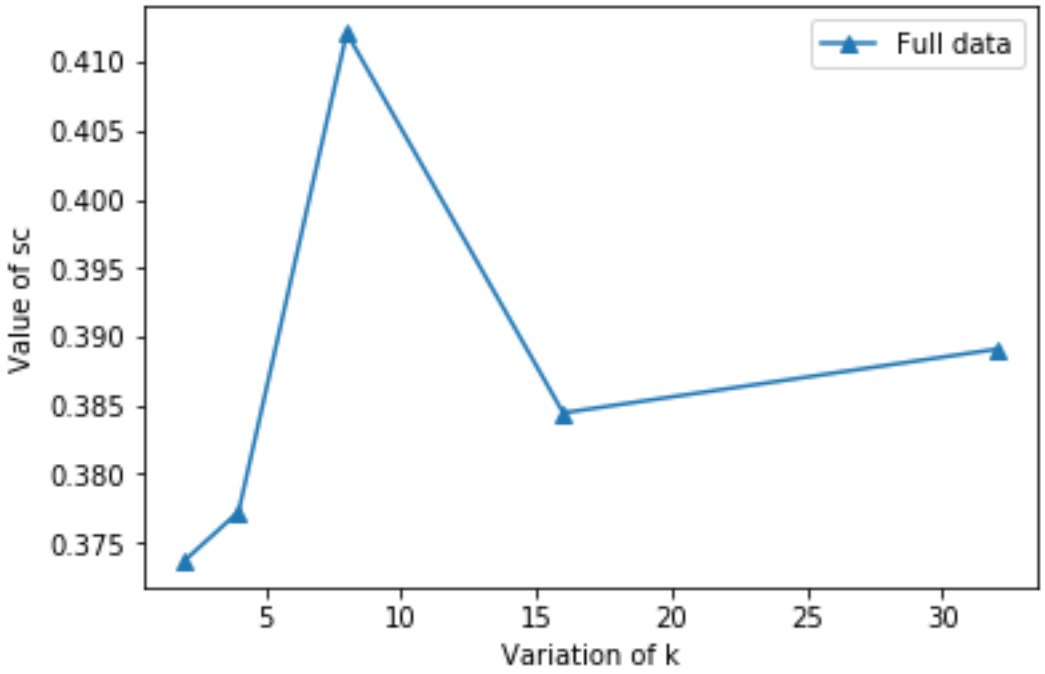
SC on full data= [0.3736915531487520 2, 0.3772461057811468 1, 0.4121101453253875 9, 0.3844542784391885, 0.3890843322470203 8]

SC on 2467 data= [0.493488075375761, 0.6953962989148753 1, 0.553548984729074 1, 0.4376415036360395 4, 0.361965866183722 9]

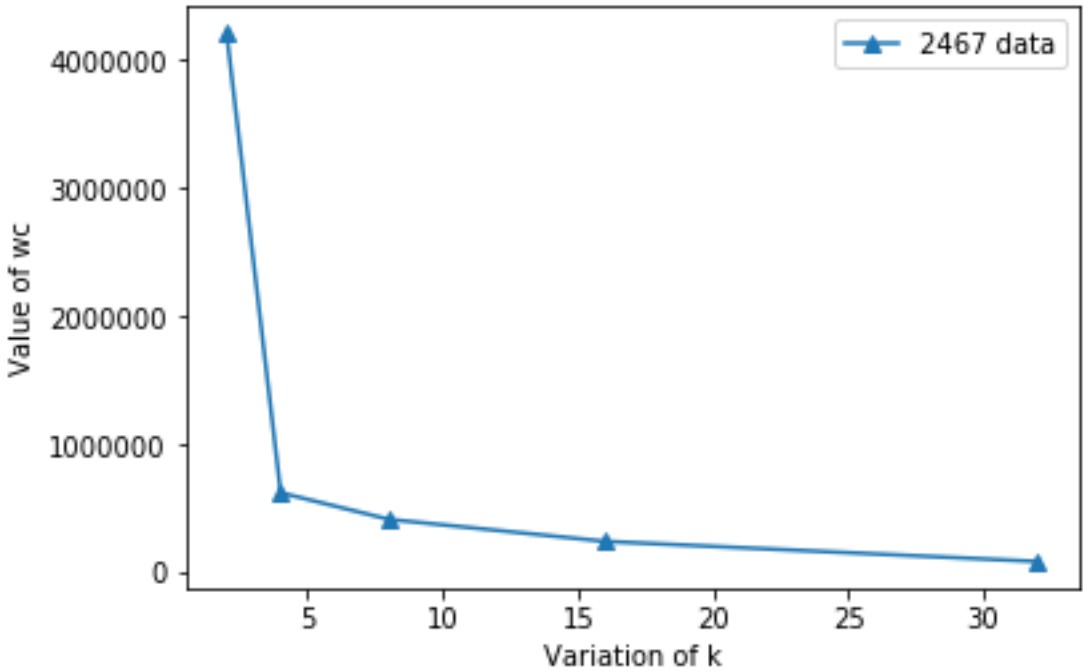
SC on 67 data= [0.8217453505137588 5, 0.6130007294390573 7, 0.3928796450574030 3, 0.3529927291750821 8, 0.358411021472612]



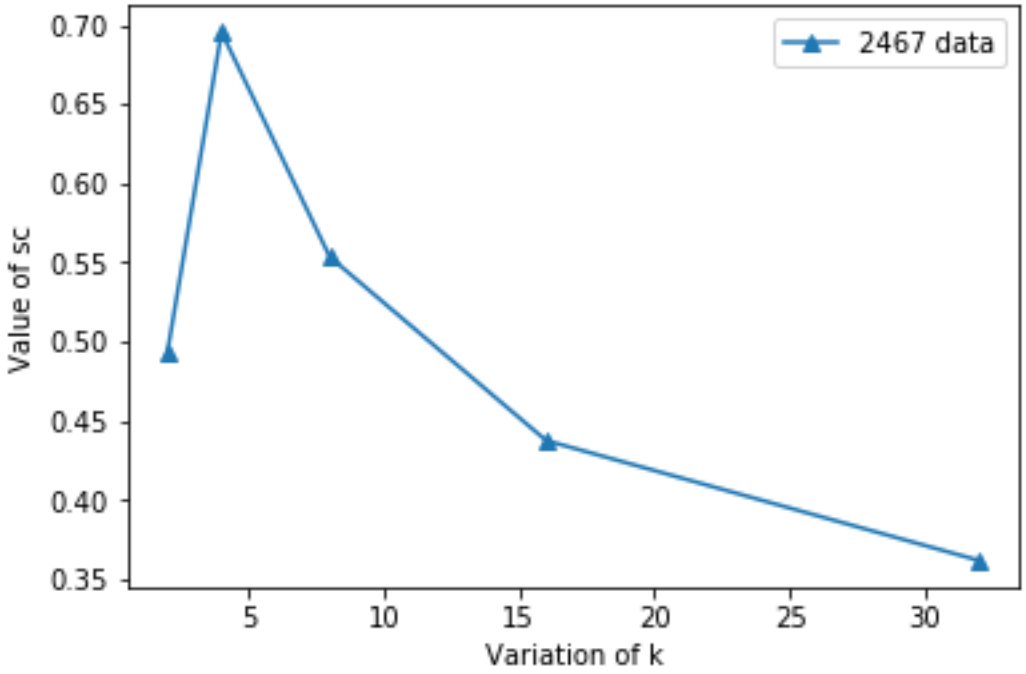
Graph 1


Full data
0.410
0.405
0.400
0.395
0.390
0.385
0.380
0.375
Value of sc
5
10
15
20
25
30
Variation of k

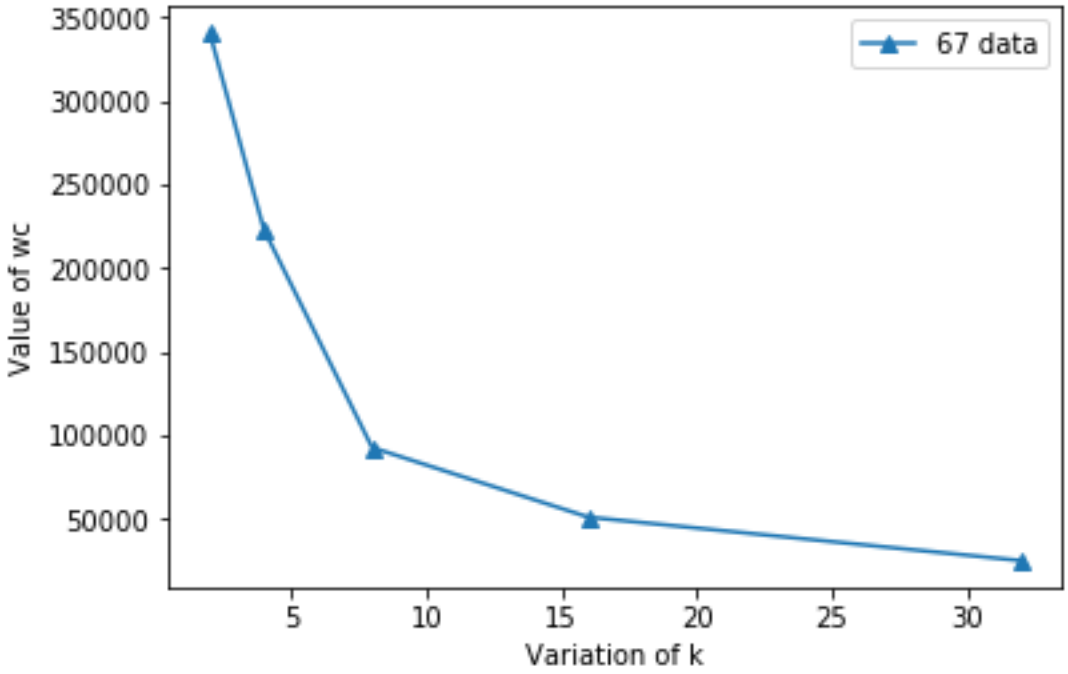
Graph 2


2467 data
4000000
3000000
2000000
1000000
0
Value of wc
5
10
15
20
25
30
Variation of k

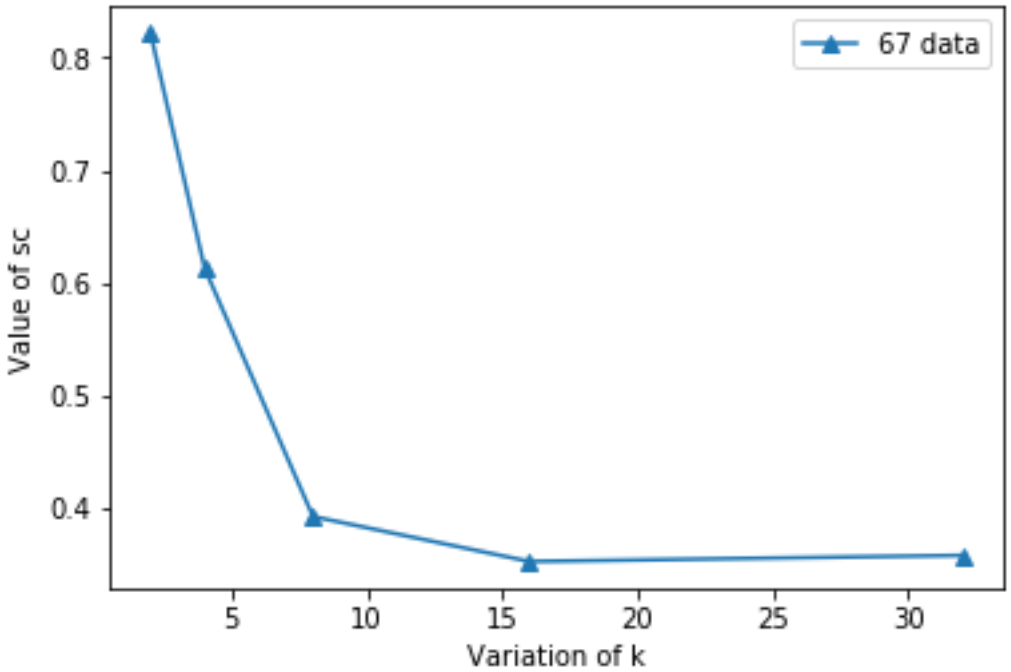
Graph 3



Graph 4


67 data
350000
300000
250000
200000
150000
100000
50000
Value of wc
5
10
15
20
25
30
Variation of k

Graph 5


67 data
0.8
0.7
0.6
0.5
0.4
Value of sc
5
10
15
20
25
30
Variation of k

Graph 6

2.) Best Values of K:-

We analyze the graphs showing the variation of k and the value of SC to pick our k values as they clear indication of peak values.

For full data set, best K:

We should choose the value of k as 8 because we can see in graph 2 (variation of k and value of sc) that there is a peak reached at the value of 8 and after that point there is a decrease in the value of silhouette coefficient. We can see from graph1 also that the value of WC is not decreased that much after k=8.So, we choose the best value of k as 8.

For 2467 dataset, best K:

We should choose the value of k as 4 because we can see in graph 4 (variation of k and value of sc) that the value of sc reaches the peak value at k=4 and after that the value of SC is decreased. We can see from graph3 also that the value of WC decreased tremendously at k=4.Thus, the best value of k is 4.

For 67 dataset, best K:

We shall choose the value of k as 2 since we can see in graph 6(variation of k and value of SC) that at k=2 only, the value of SC is maximum, and after that point it decreases, hence we choose k=2.

Comparision between two scores:

For full data-set:

WC_SSE: From graph1, we see that the value of wc reaches a saturation point at k=8, and after k=8, the value of wc is not decreased that much, so we choose k=8 from graph1.

SC: As discussed, from graph2, we see that the value of SC reaches a peak point at k=8, and after k=8, the value of SC is decreased, so we choose k=8 from graph2.

For 2467 data-set:

WC_SSE: From graph3, we see that the value of wc decreased tremendously at k=2 and reaches a saturation point at k=4, and after k=4, the rate of decrease of WC is not that much, so we choose k=4 from graph3.

SC: As discussed, from graph4, we see that the value of SC reaches a peak point at k=4, and after k=4, the value of SC is decreased, so we choose k=4 from graph4.

For 67 data-set:

WC_SSE: From graph5, we see that the value of wc reaches a saturation point at k=8, and after k=8, the value of wc is not changed that much, so we choose k=8 from graph5.

SC: But As discussed, from graph6, we see that the value of SC reaches a peak point at k=2, and after k=2, the value of SC is decreased, so we choose k=2 from graph6.

So, only considering graph 5, we might chose a different value of k, because WC graph does not provide sufficient and clear indication about the value of k to be picked.

3.) For 10 trials:

Values of K= [2, 4, 8, 16, 32]

Average WC on full data= [8983697.212393444, 4280988.0682722991, 1898912.1748344109, 866863.59961230529, 418572.91046116577]

Std WC on full data= [97.955486504909217, 26461.386649854732, 4380.14962225813, 3874.6218242672908, 4752.3161110296005]

Average WC on 2467 data= [4300944.3878043555, 732354.72340806259, 383419.42237532843, 196621.21419464101, 87386.475004545209]

Std WC on 2467 data= [43372.062301809739, 102922.09340724314, 14967.973813704391, 9416.7245148612947, 2064.4570517999045]

Average WC on 67 data= [340372.41942807229, 204507.77665352757, 117762.13344719952, 53027.784422013341, 26637.309100319148]

Std WC on 67 data= [0.0, 9482.8450724801714, 12628.000588666948, 1323.9523019476105, 369.8243908808426]

Average SC on full data= [0.3735915996551413 8, 0.3720074222431093 4, 0.4037375171070403 1, 0.4025061697440217 6, 0.3877152949920831 3]

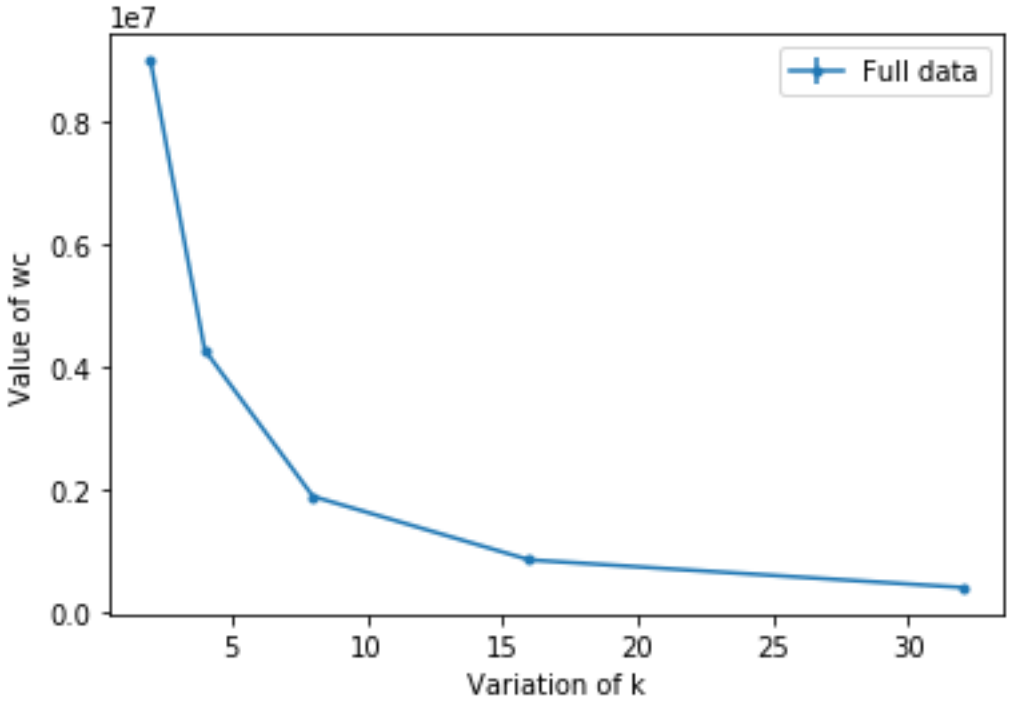
Std SC on full data= [2.0692339043001161e-05, 0.0021927164888313 62, 0.0021759032817198392, 0.0021402776479522235, 0.0016160609866678521]

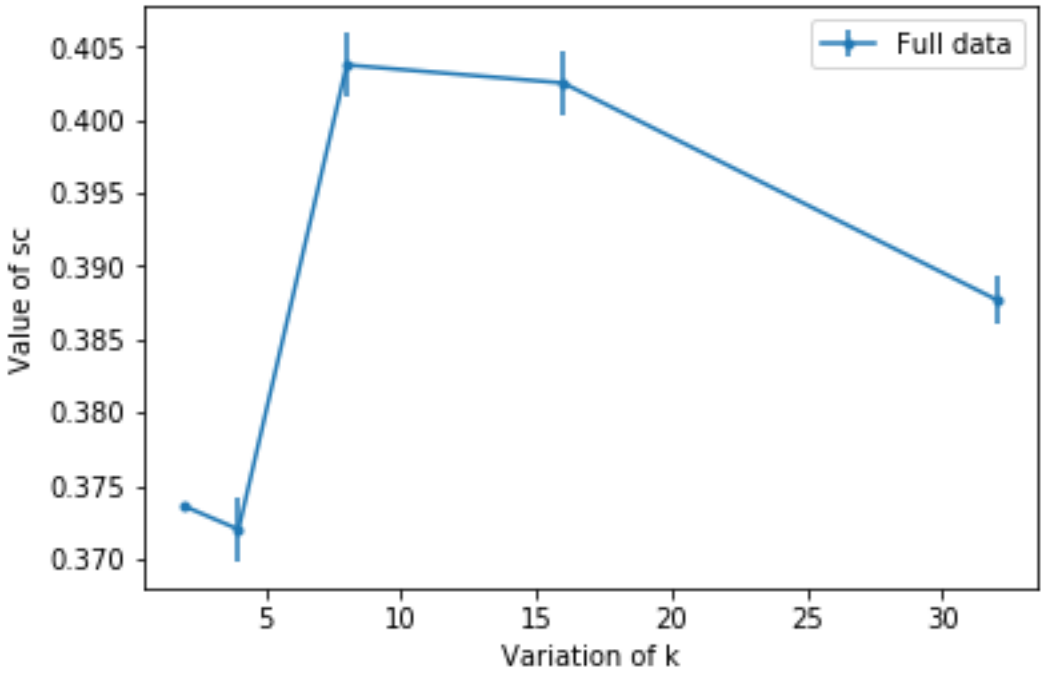
Average SC on 2467 data= [0.4976487723951156 2, 0.6780344984106492 2, 0.5045889874068907 3, 0.4107194612920698 2, 0.3716441783856603 8]

Std SC on 2467 data= [0.0020098076011405347, 0.0164708501624443 52, 0.0139711570842526 74, 0.0068696472417228 93, 0.0019602029429627828]

Average SC on 67 data= [0.8217453505137587 4, 0.5387329003661806, 0.4272753042974919 4, 0.3686034879985139 5, 0.3555173942761298 3]

Std SC on 67 data= [3.5108334685767011e-17, 0.0251021476755145 31, 0.0233631316617338222, 0.0038448372034328 29, 0.0019138358215287146]

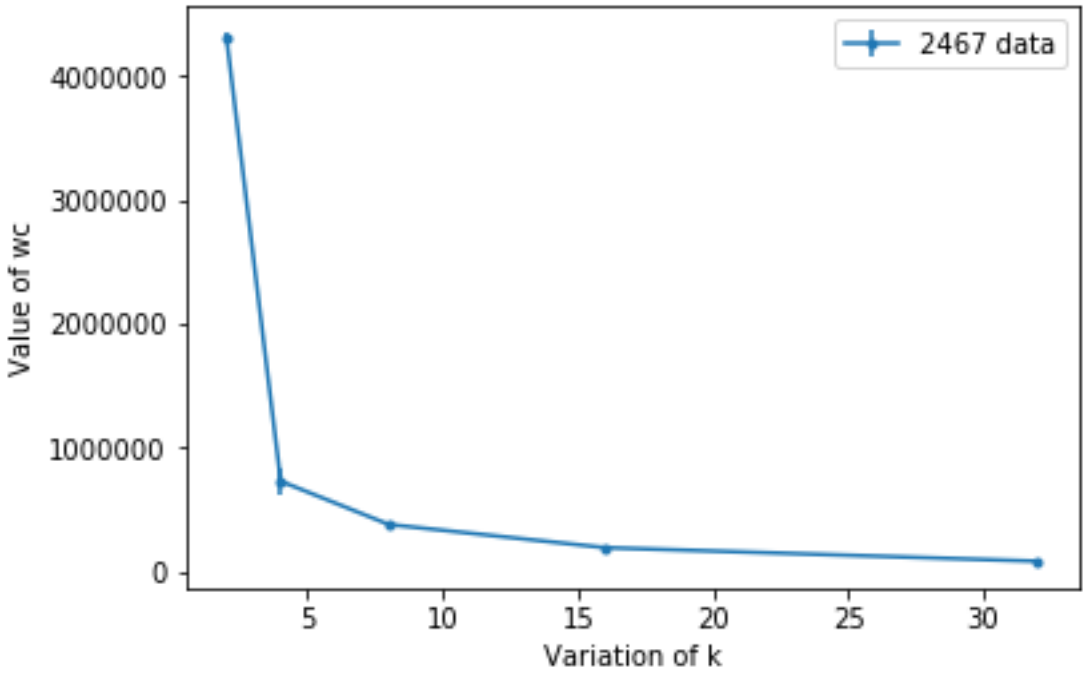


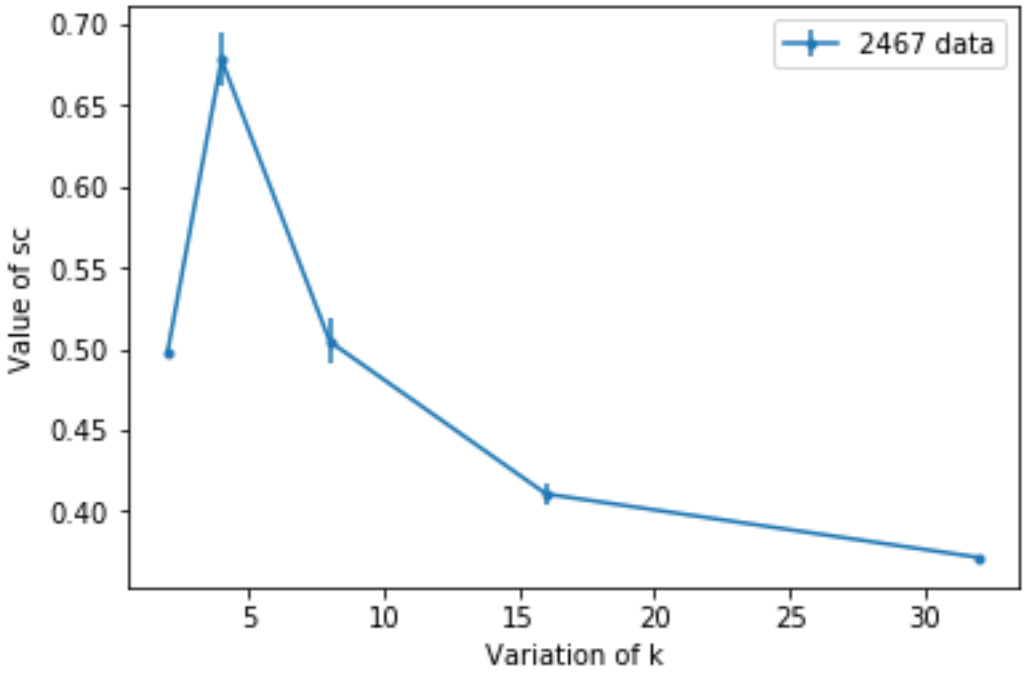


Analysis: As observed from the WC graph, we see that there are no significant error bars. And we can conclude that k-means is not sensitive to the initial random starting conditions, since the

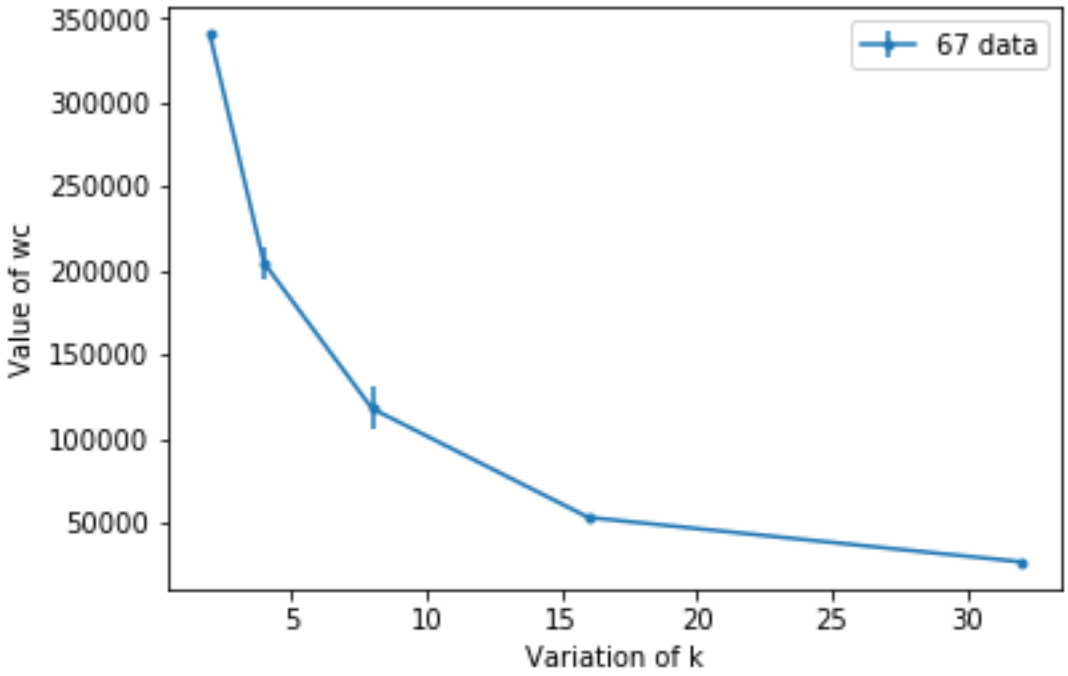
random starting clusters does not affect the within cluster distances.

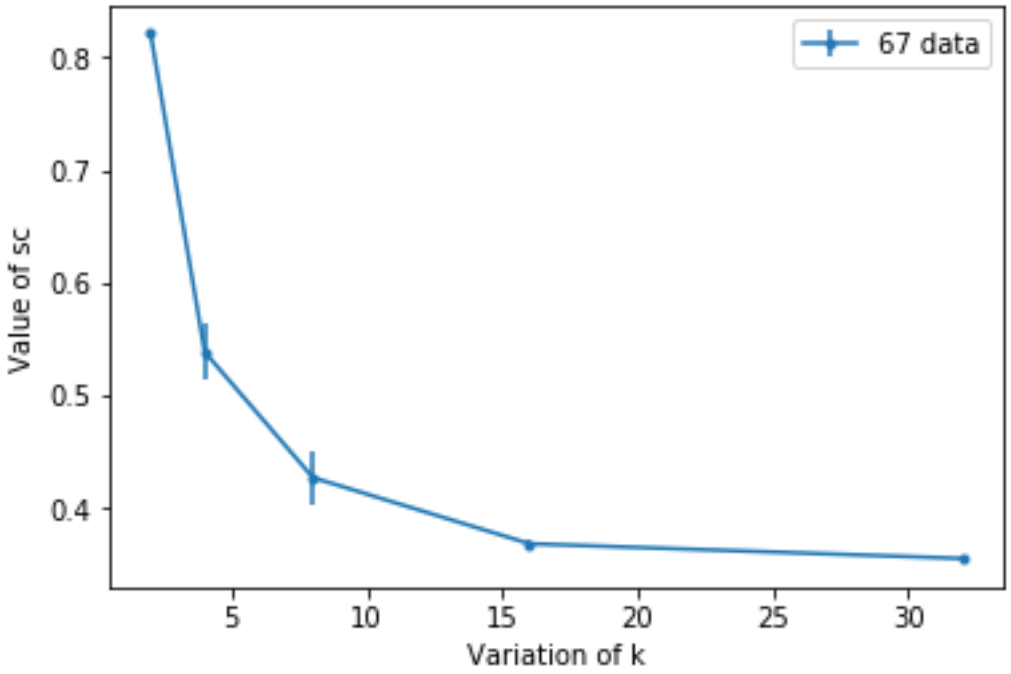
But in the SC graph, we can observe significant error bars and there is a lot of variance for k=4,8,16. So k-means is sensitive enough to the initial random starting conditions, since the inter-cluster distances varies with the initial selected centroid points.





Analysis: As we see from the graphs that there is significant error bars for k=4,8 and thus the intra as well as inter cluster distances varies according to the initial random seeds.

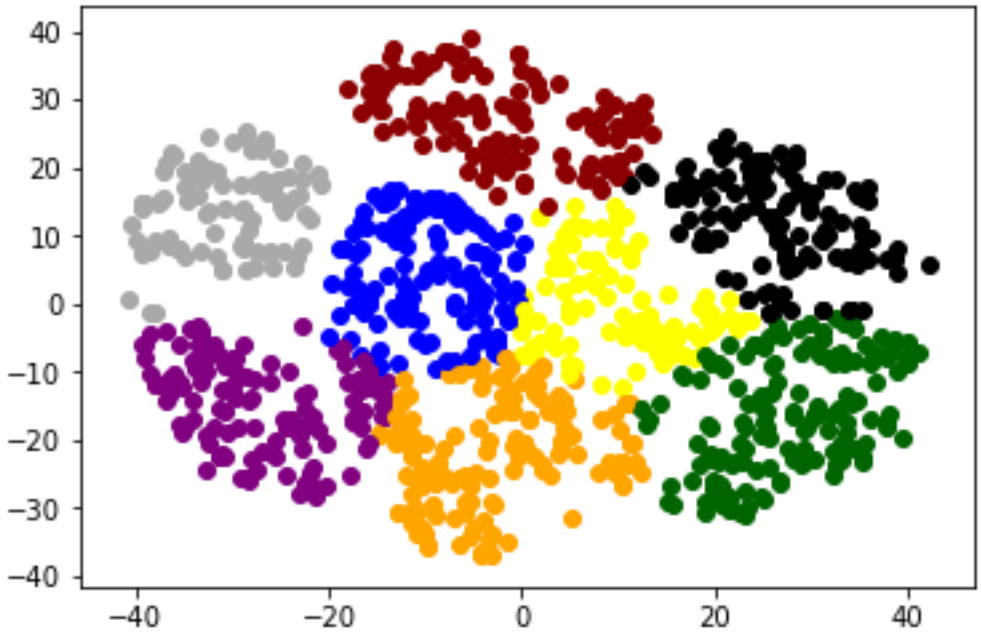




Analysis: As we see from the graphs that there is significant error bars for k=4,8 and thus the intra as well as inter cluster distances varies according to the initial random seeds.

4.) For full data set, we chose k=8 in B.2

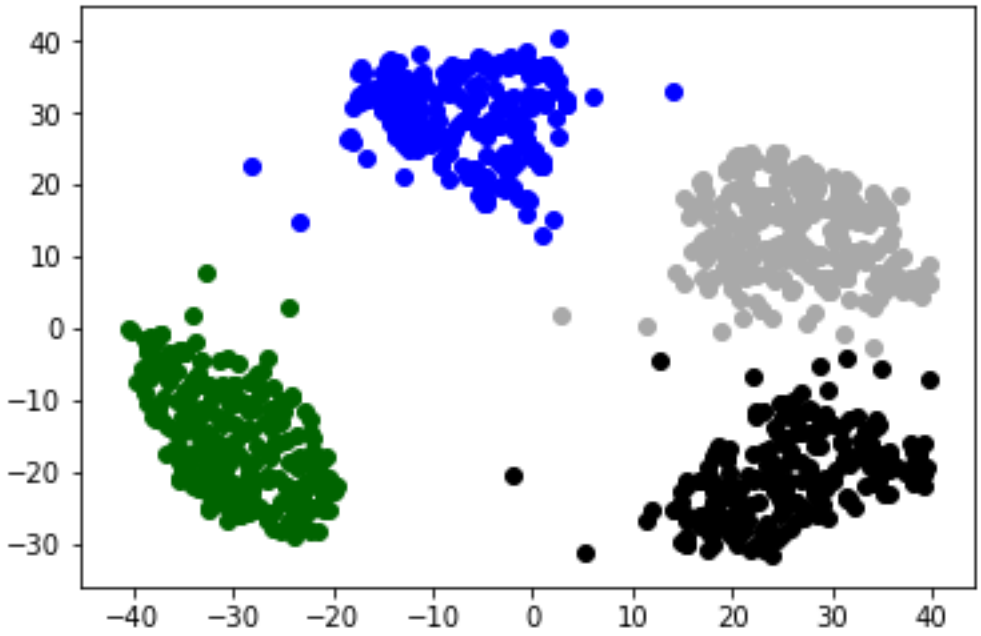
NMI: 0.342957062384



Analysis: As we know that the maximum value of NMI reaches 0.5 in our implementation, and in this case we see that the value of NMI is coming as 0.342957062384.This NMI value indicates that the purity of the clusters has not reached its maximum. And from the visualization graph, we can see that the eight clusters are not that well-separated and there can be chances that points which lie on the boundary of other clusters are clustered in the false group.

For 2467 data-set, we chose k=4 in B.2

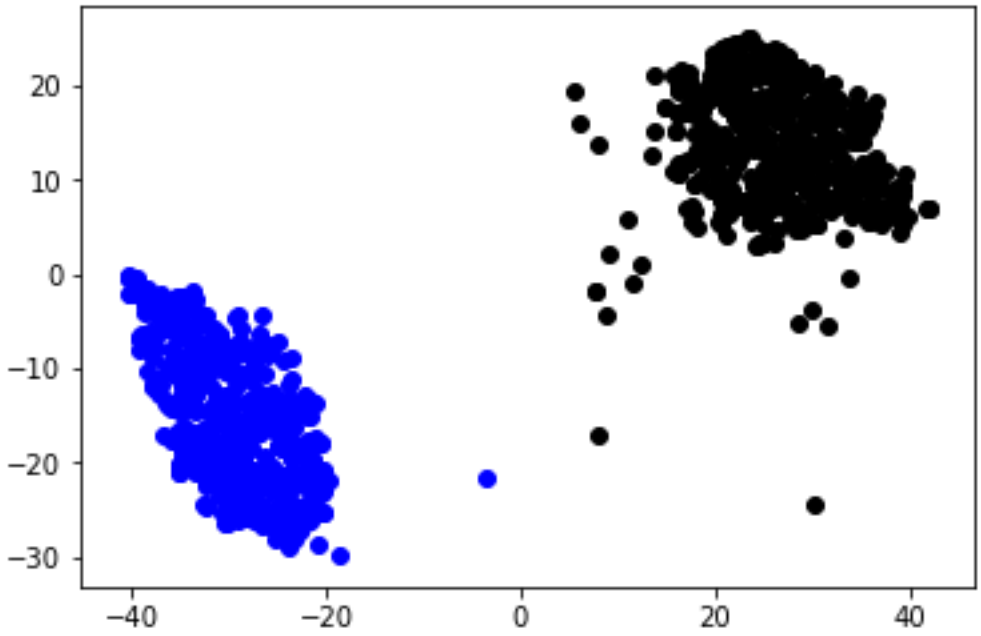
NMI: 0.45465341281



Analysis: As we know that the maximum value of NMI reaches 0.5 in our implementation, and in this case we see that the value of NMI is coming as 0.45465341281. This NMI value indicates that the purity of the clusters has not reached its maximum. And from the visualization graph, we can see that the four clusters are much well-separated except some points which are lying close to other clusters. So there can be slight chances in clustering points which lie on the boundary of other clusters into false group.

For 67 data-set, we chose k=2 in B.2

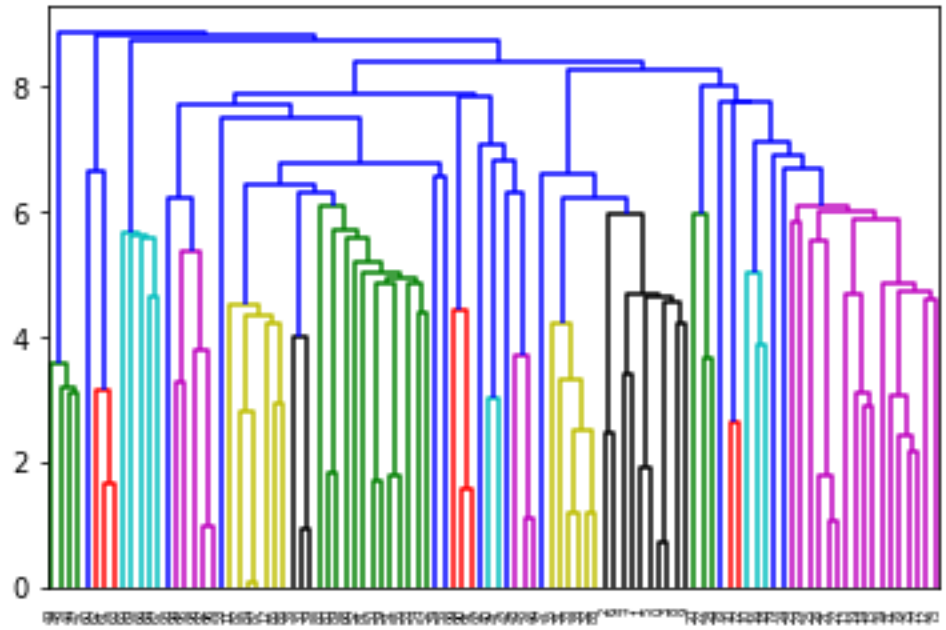
NMI: 0.490710990204



Analysis: As we know that the maximum value of NMI reaches 0.5 in our implementation, and in this case we see that the value of NMI is coming as 0.490710990204. This NMI value indicates that the purity of the clusters has almost reached its maximum. And from the visualization graph, we can see that the two clusters are very much well-separated .So there are very slight chances in clustering points in a particular cluster into a false group.

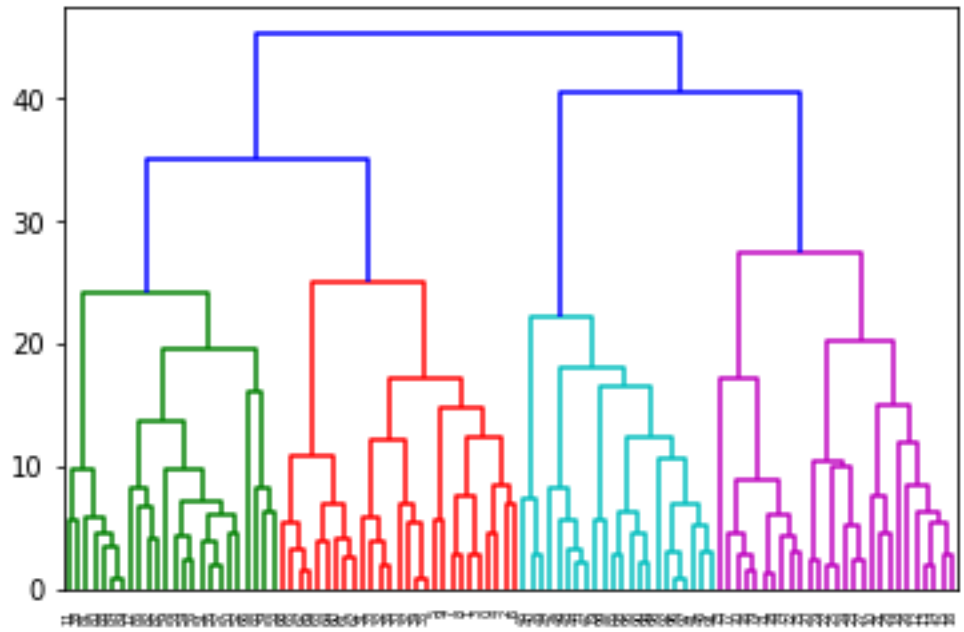
C. Comparision to hierarchical clustering

1)

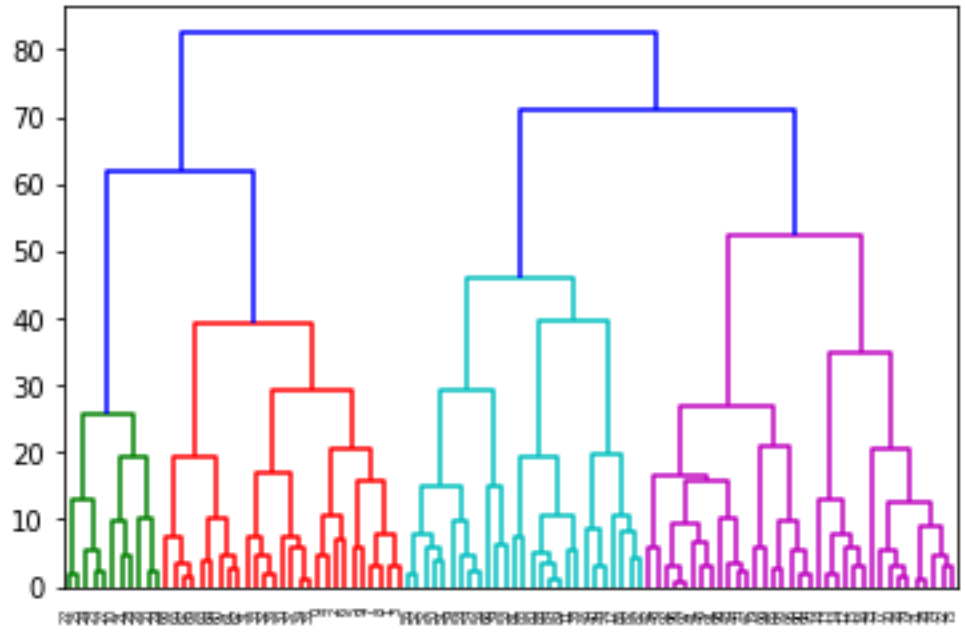


Graph Single Linkage

2)

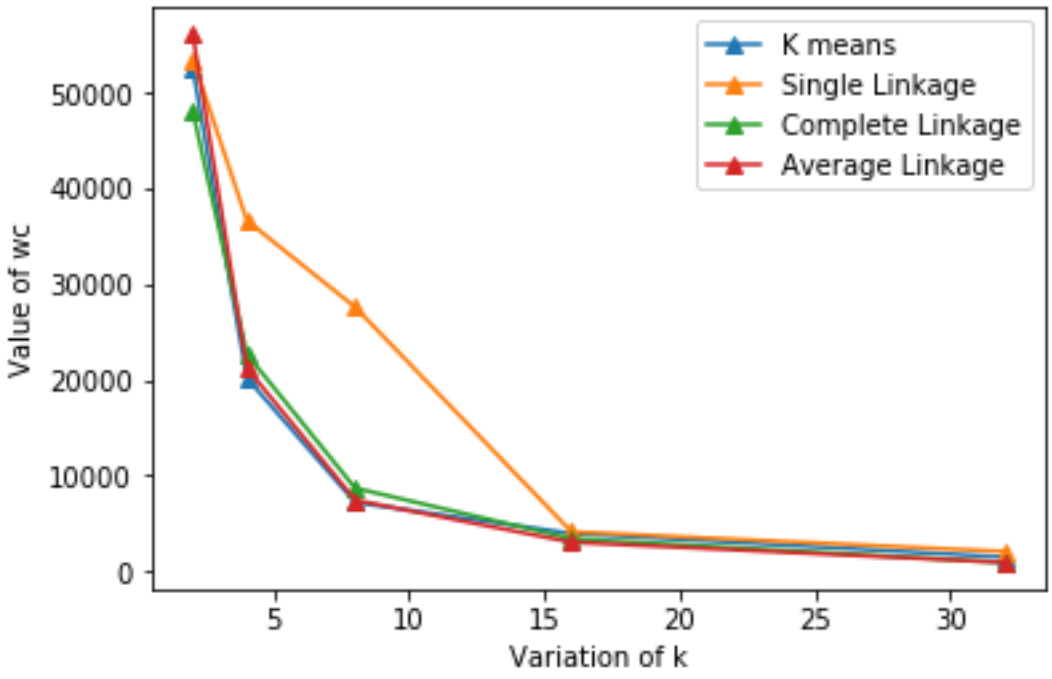


Graph Average Linkage

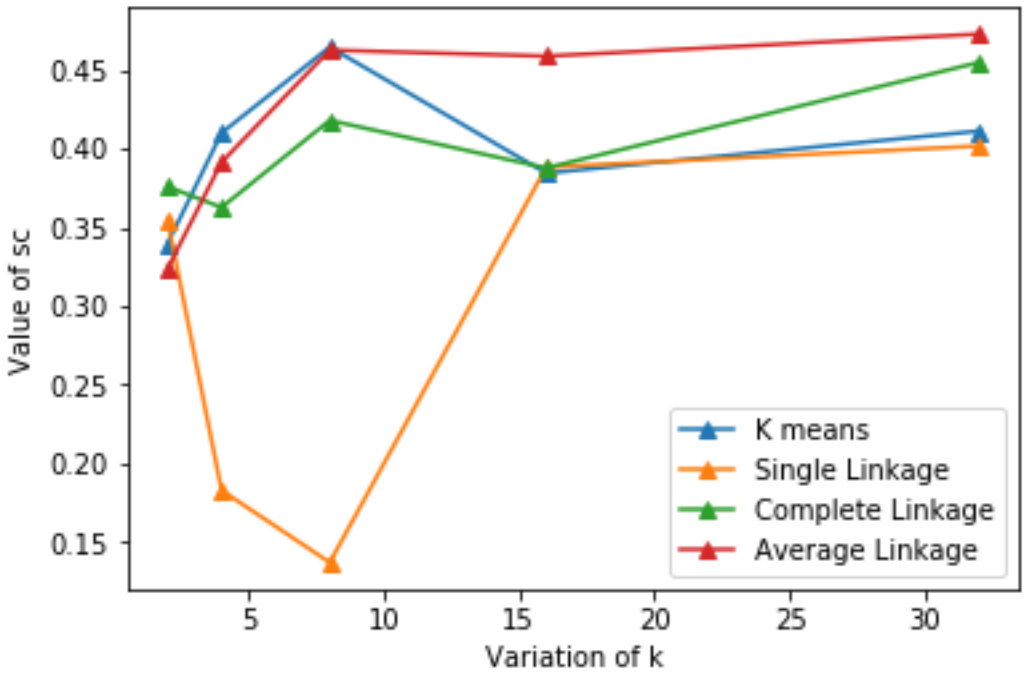


Graph Complete Linkage

3)



Graph 1



Graph 2

4) Value of K:

For Single Linkage:

We see that in point(3), graph 1, the value of k decreases tremendously as we increase k and reaches a saturation value at k=16. After k=16, the within cluster distance is not decreased that much.

And in graph 2, we see that at k=8, the value of SC is very low and after that it reaches the peak at k=16, which should be selected as the value of K.

So from analyzing both the graphs, we select k=16 for single linkage.

For average linkage:

We see that in point(3), graph 1, the value of k decreases tremendously as we increase k and reaches a saturation value at k=16. After k=16, the within cluster distance is not decreased that much.

And in graph 2, we see that at k=8, the value of SC reaches the peak and after that the value of SC starts to decrease.

So from analyzing both the graphs, we select k=8 for average linkage as in graph 1 also, the variation in WC value is not that much for values of k=8 and k=16.

For complete linkage:

We see that in point(3), graph 1, the value of k decreases tremendously as we increase k and reaches a saturation value at k=16. After k=16, the within cluster distance is not decreased that much.

And in graph 2, we see that at k=8, the value of SC reaches the peak and after that the value of SC starts to decrease, and again starts to increase.

So from analyzing both the graphs, we select k=8 for complete linkage as in graph 1 also, the variation in WC value is not that much for values of k=8 and k=16.

For k-means:

We see that in point(3), graph 1, the value of k decreases tremendously as we increase k and reaches a saturation value at k=16. After k=16, the within cluster distance is not decreased that much.

And in graph 2, we see that at k=8, the value of SC reaches the peak and after that the value of SC starts to decrease.

So from analyzing both the graphs, we select k=8 for k-means as in graph 1 also, the variation in WC value is not that much for values of k=8 and k=16.

In part B, we chose the value of k as 8 in k-means implementation and thus it does not defer with the value of k selected from this implementation.

5) Computation of NMI:

NMI Single: 0.368906724863

NMI Average: 0.39338546880

NMI Complete: 0.389921073456

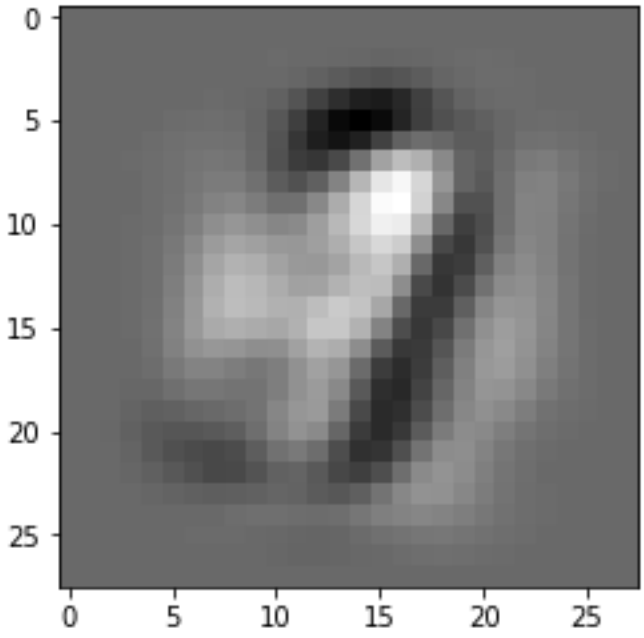
NMI k-means : 0.3766854603238

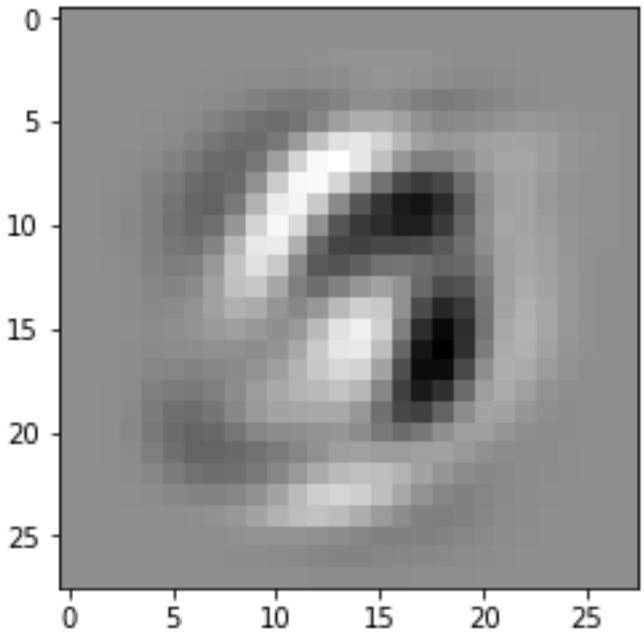
Now, we compare these results across all distance measures; and the dataset is small and we are randomly selecting 100 data points. For single linkage, we see that it underperforms across all distance measures and the values for average, complete and

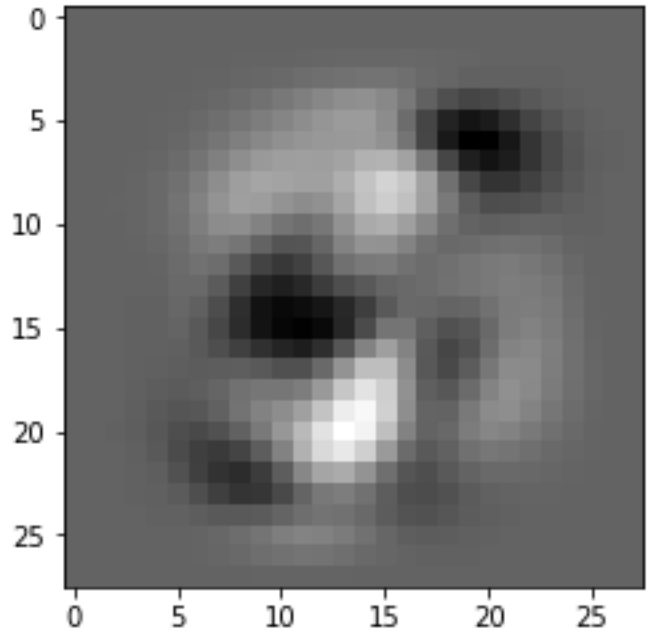
k-means are comparable. The variance between the values, in general, varies as the dataset is really small.

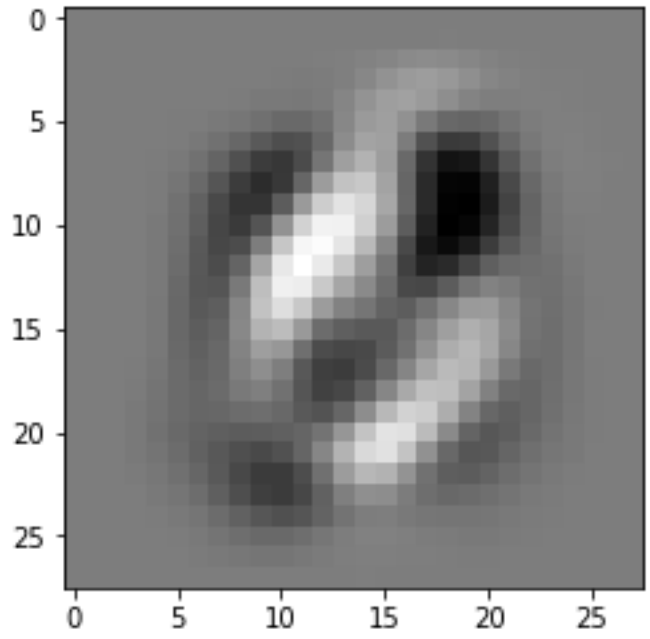
Bonus:

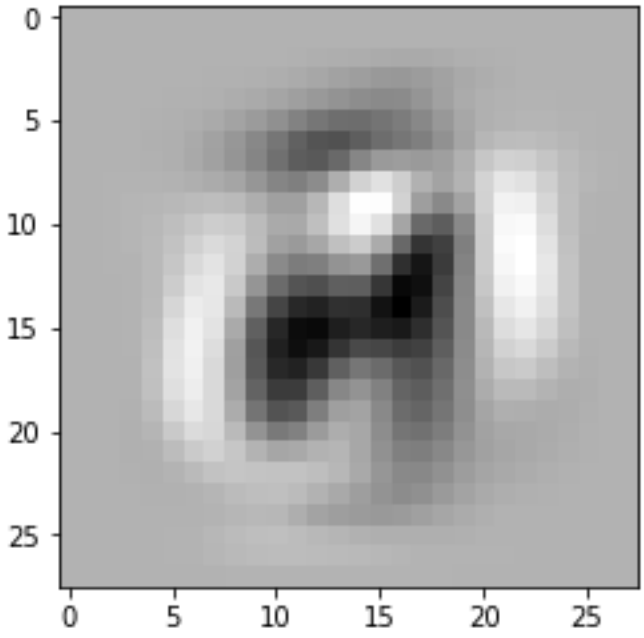
1) Implemented.
2)

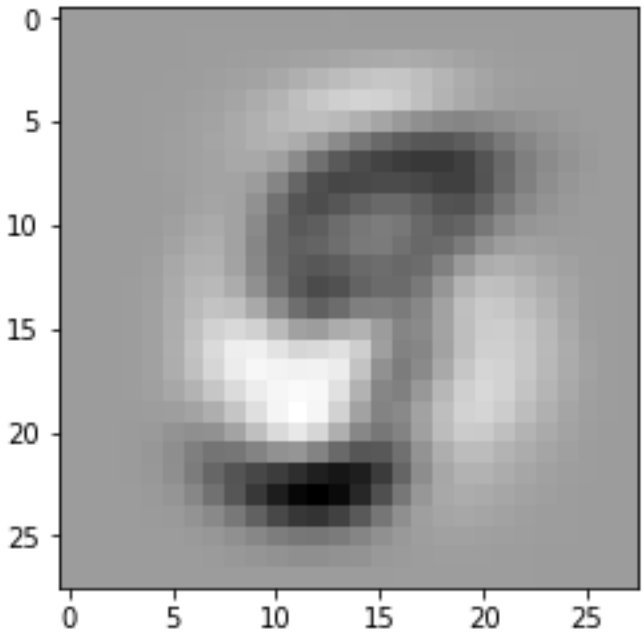

0
5
10
15
20
25
0
5
10
15
20
25

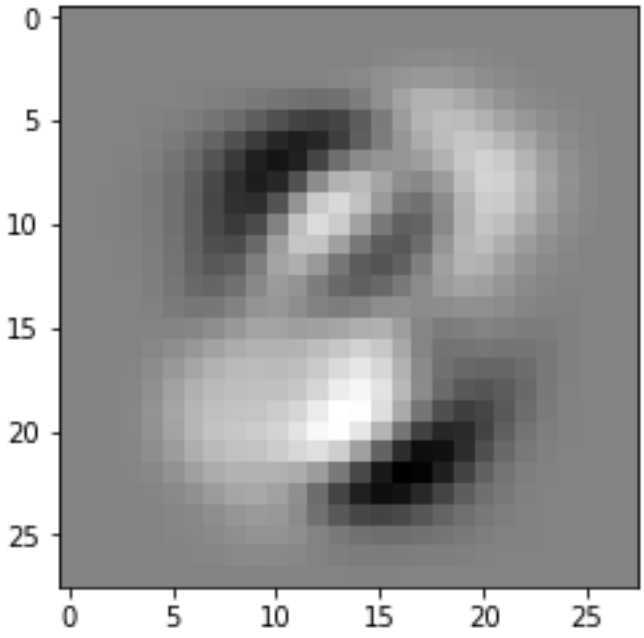

0
5
10
15
20
25
0
5
10
15
20
25

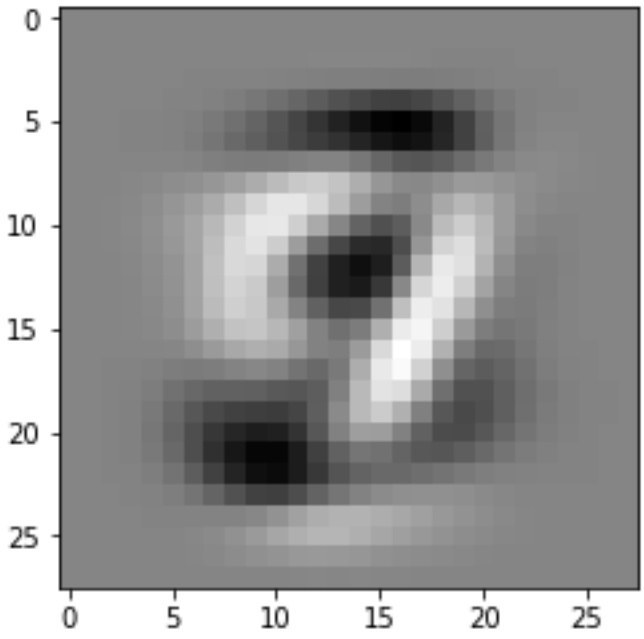

0
5
10
15
20
25
0
5
10
15
20
25

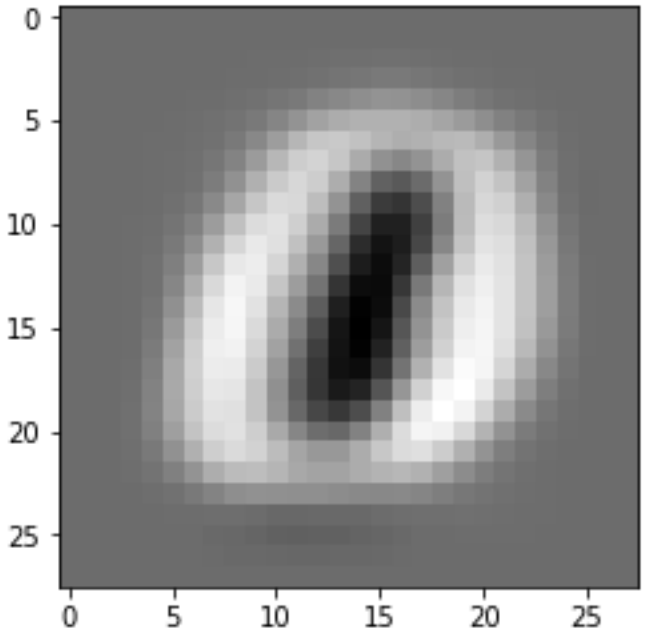

0
5
10
15
20
25
0
5
10
15
20
25

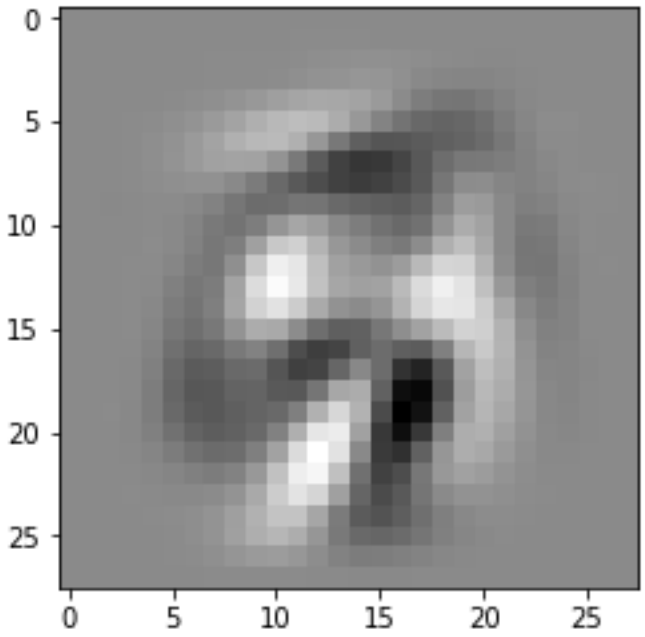

0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25

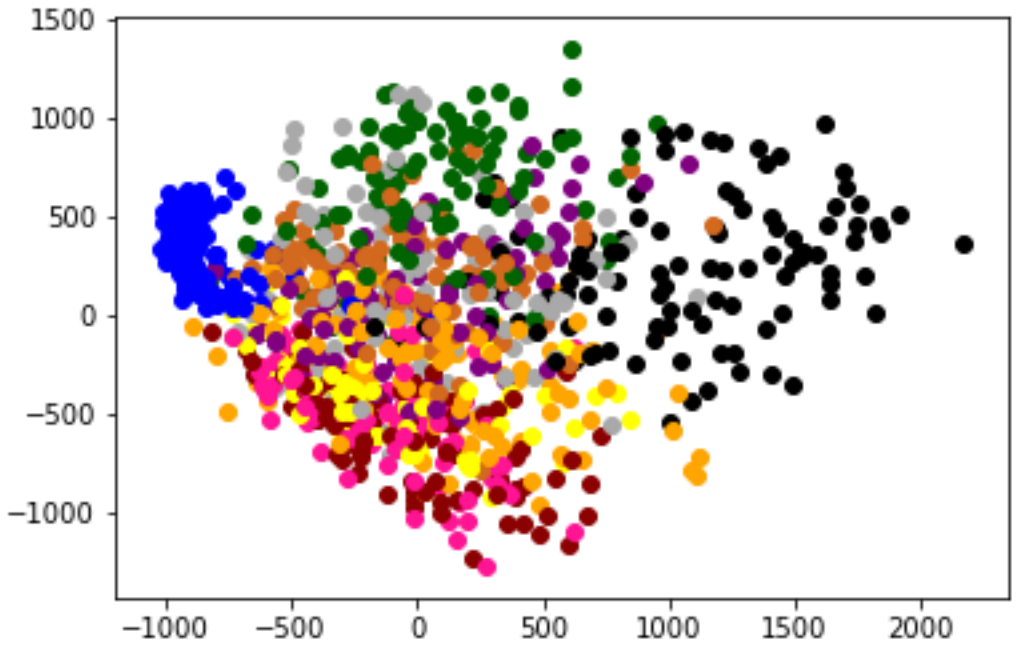

0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25





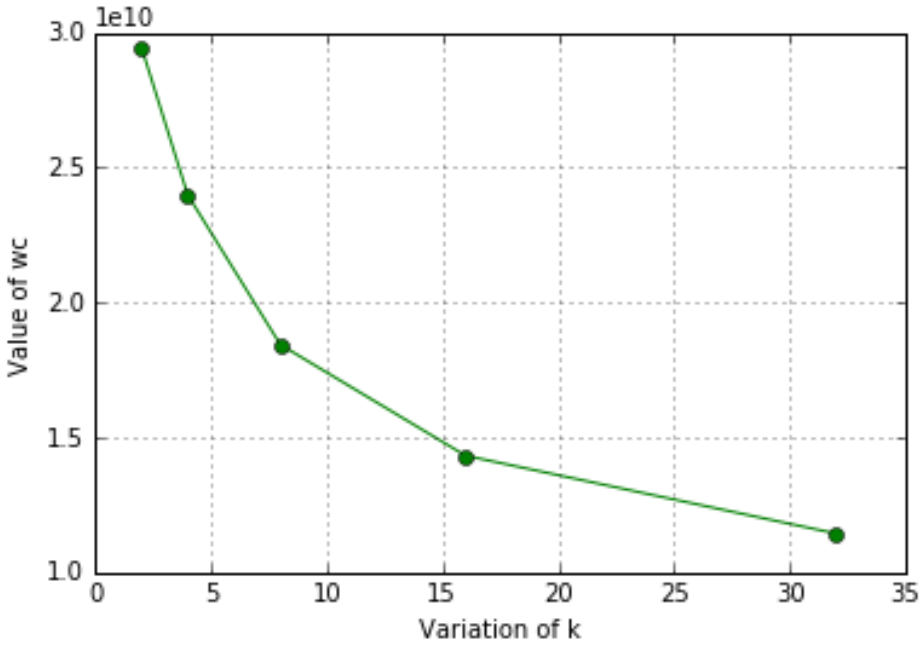
3) Visualization of randomly selected examples using the first two principle components:



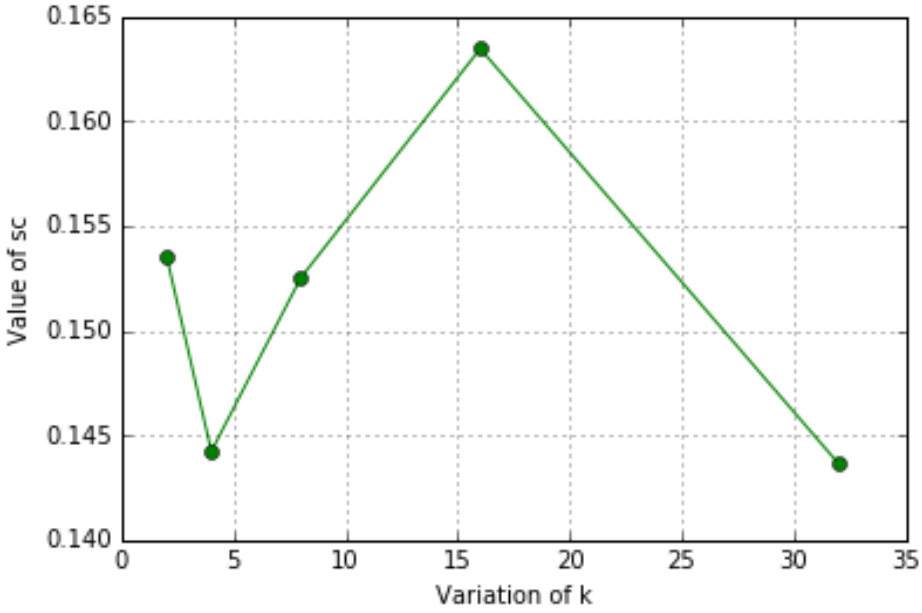
In this graph, we see that the clusters are not that well-separated, after the dimensionality reduction to 10 and then using the first two principal components.

Comparing to the clusters found with the tSNE embedding: We saw that the clusters were comparably separable and hence the dimensionality reduction method of tSNE is better.

1) For full data set:
   Experiment B.1 :



Graph- full data (value of WC)



Graph- full data (value of SC)

B.2

Choosing value of k:

We observe from the SC graph that there is a clear peak for k=16 and after k=16, the value of SC is decreasing. Since

we choose the value of k from the graph of SC, we select k=16.

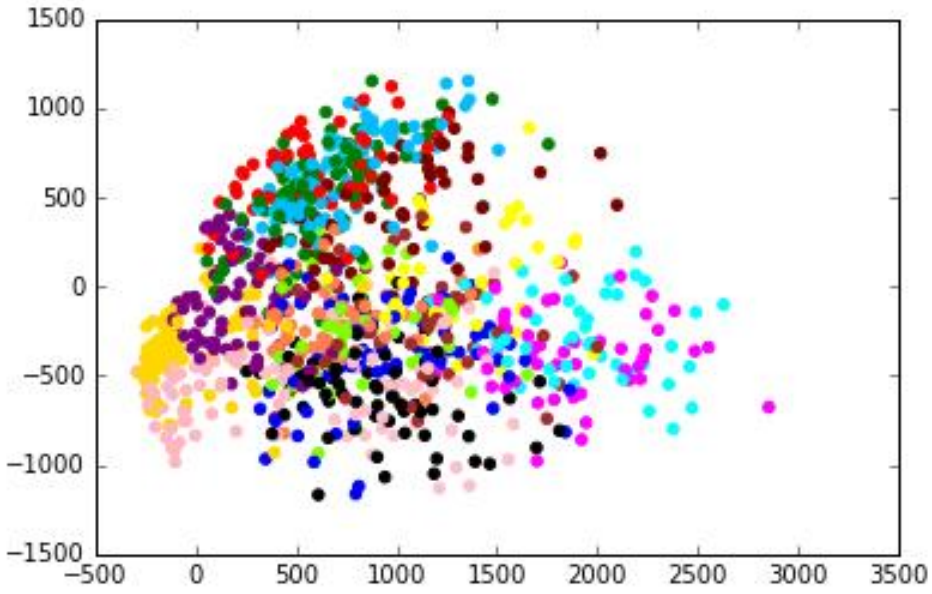
Comparing two scores:

From the graph of WC, we observe that the value of WC is tremendously decreased till k=8, and after k=8 there is not much significant rate of decrease in the value of WC. So we choose k=8 from this graph. While as discussed, from the SC graph, we can state that we choose k=16, as there is a clear peak value of SC at this value of k.

B.4:

We chose value of k as 16.

NMI: 0.2371422633445



Analysis: As we know that the maximum value of NMI reaches 0.5 in our implementation, and in this case we see that the value of NMI is coming as 0.2371422633445.This NMI value

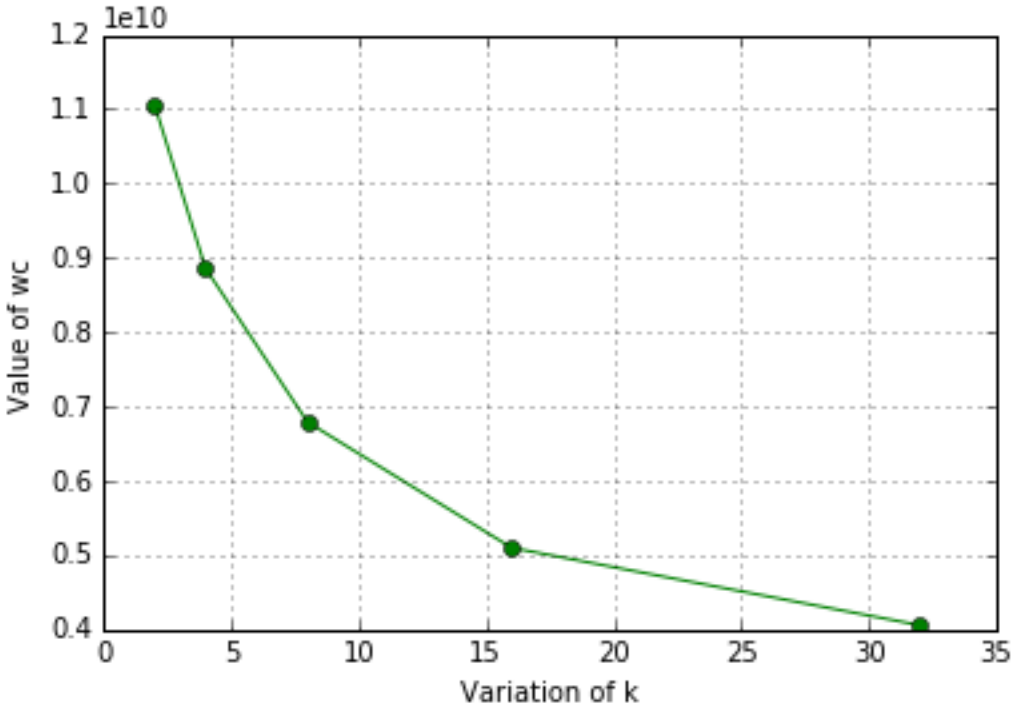
indicates that the purity of the clusters has not reached its maximum. And from the visualization graph, we can see that the sixteen clusters are not well-separated and there can be high chances that points which lie on the boundary of other clusters are clustered in the false group. And the purity of the clusters is very low.

Comparing to tSNE: We observed that in tSNE clustering, the clusters were comparably separable, and the purity of the clusters were comparable high than this graph. So tSNE clustering was better than clustering using pca.

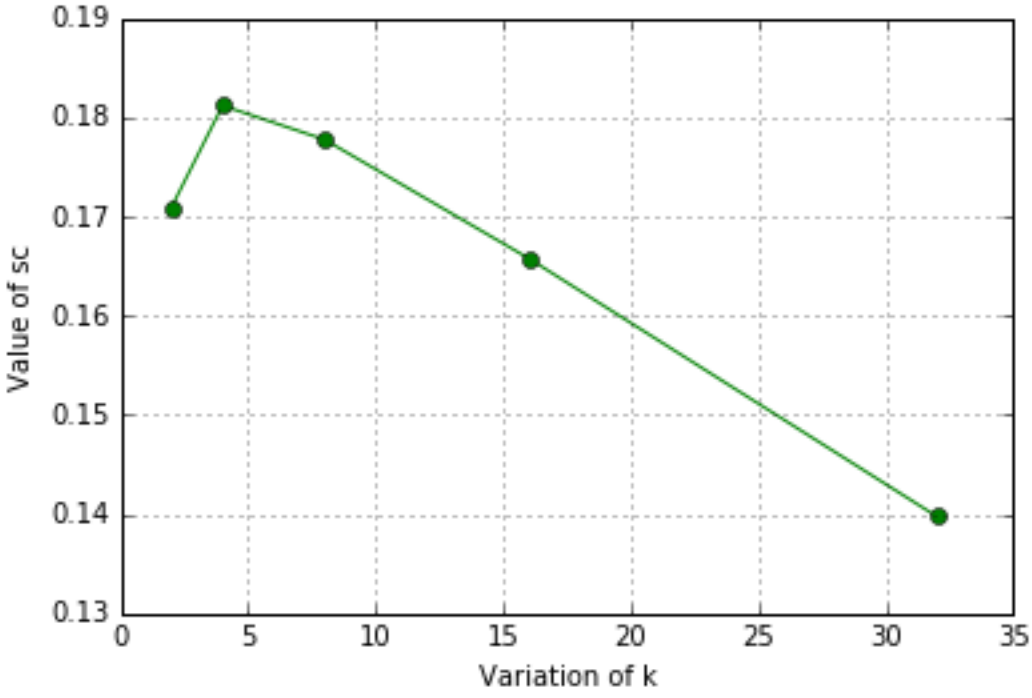
5) Repeating part 1 and 4 for 2467 dataset:

1) Implemented

4) Experiment B.1 :



Graph- 2467 dataset (Value of WC)



Graph- 2467 dataset (Value of SC)

B.2

We should choose the value of k as 4 because we can see in graph (variation of k and value of SC) that the value of SC reaches the peak value at k=4 and after that the value of SC is decreased. We can see from WC graph that the value of WC decreases significantly till k=16.Thus, the best value of k is 4 from the SC graph.

Comparing across two scores:

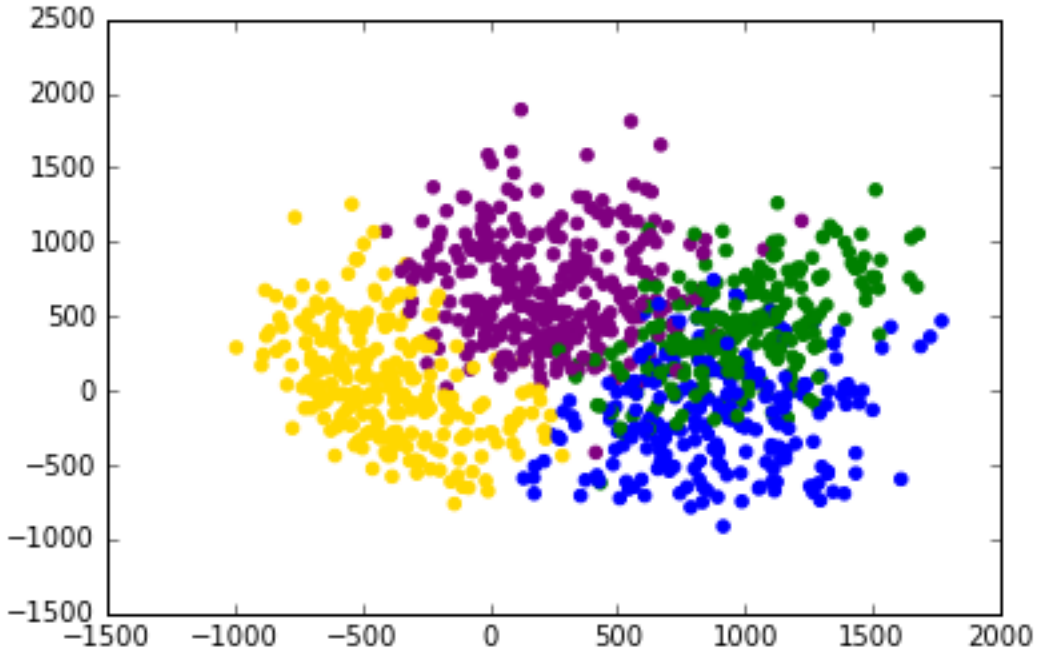
WC_SSE: From this graph, we see that the value of WC decreases till k=16 and reaches a saturation point at k=16, and after k=16, the rate of decrease of WC is not that much, so we choose k=16 from this graph.

SC: As discussed, from this graph, we see that the value of SC reaches a peak point at k=4, and after k=4, the value of SC is decreased, so we choose k=4 from SC graph.

B.4

We chose k=4 for 2467 dataset:

NMI : 0.3397325202115



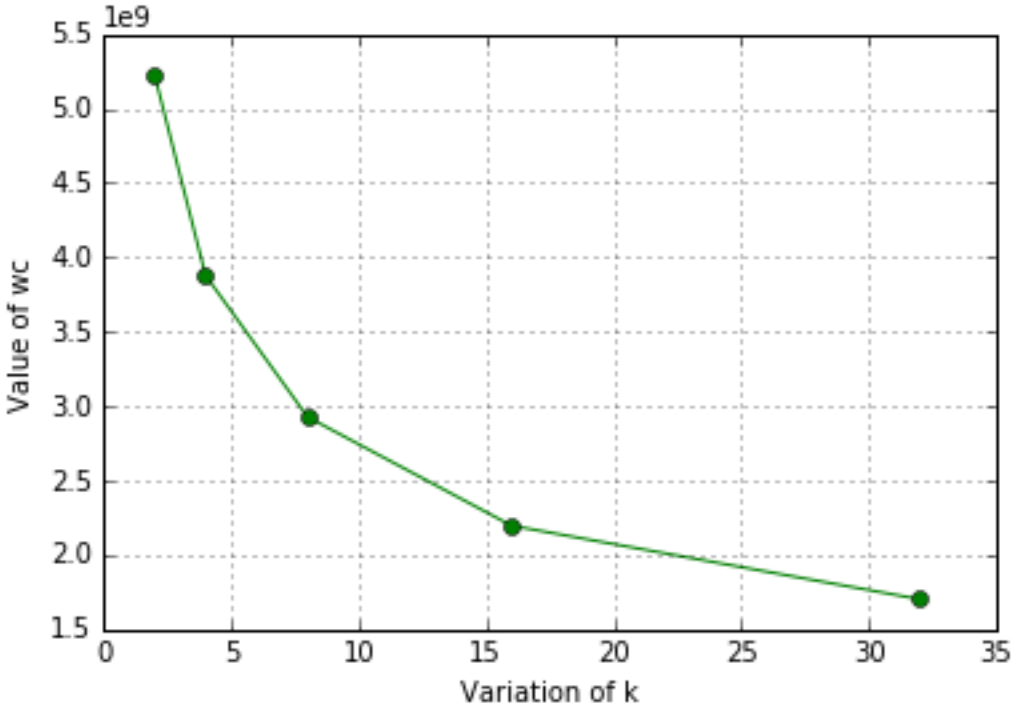
Analysis: As we know that the maximum value of NMI reaches 0.5 in our implementation, and in this case we see that the value of NMI is coming as 0. 3397325202115. This NMI value indicates that the purity of the clusters has not reached its maximum. And from the visualization graph, we can see that the four clusters are not well-separated and points in a cluster are lying close to other clusters. So there are chances in clustering points which lie on the boundary of other clusters into false group.

Comparing to tSNE: We observed that in tSNE clustering, the clusters were well-separated, and the purity of the clusters were comparable high than this graph. So tSNE clustering was better than clustering using pca.

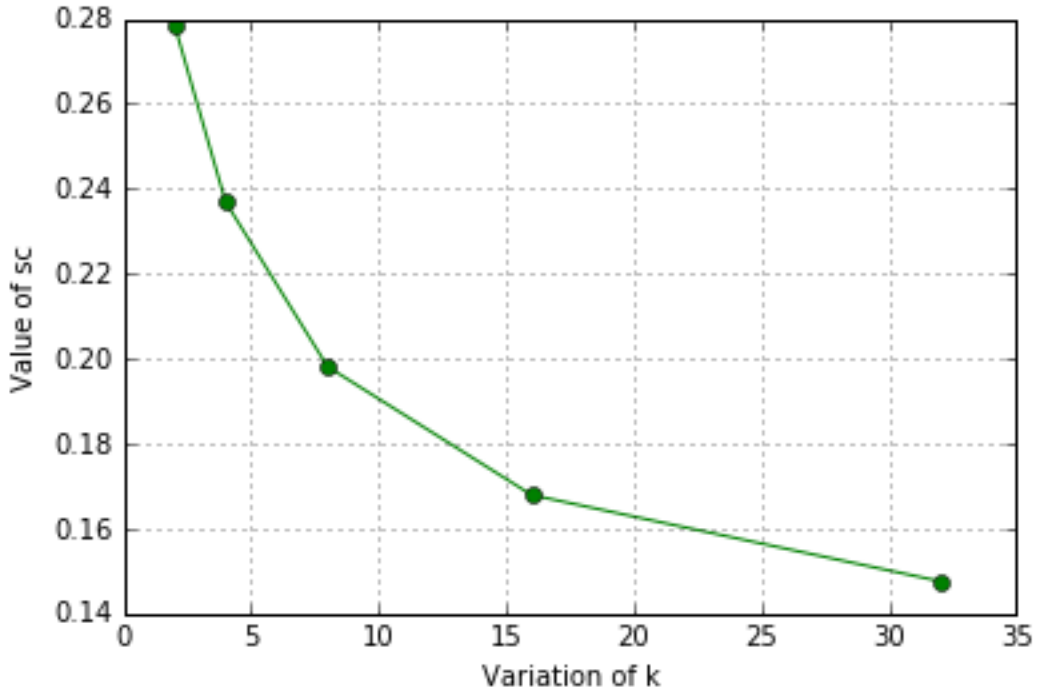
For 67 data-set:

1) Implemented

4) Experiment B.1:



Graph- 67 dataset (Value of WC)



Graph- 67 dataset (Value of WC)

B.2:

We shall choose the value of k as 2 since we can see in graph (variation of k and value of SC) that at k=2 only, the value of SC is maximum(peak value), and after that point it decreases, hence we choose k=2.

Comparision between two scores:

WC_SSE: From graph, we see that the value of wc reaches a saturation point at k=16, and after k=16, the value of wc is not changed that much, so we choose k=16 from WC graph.

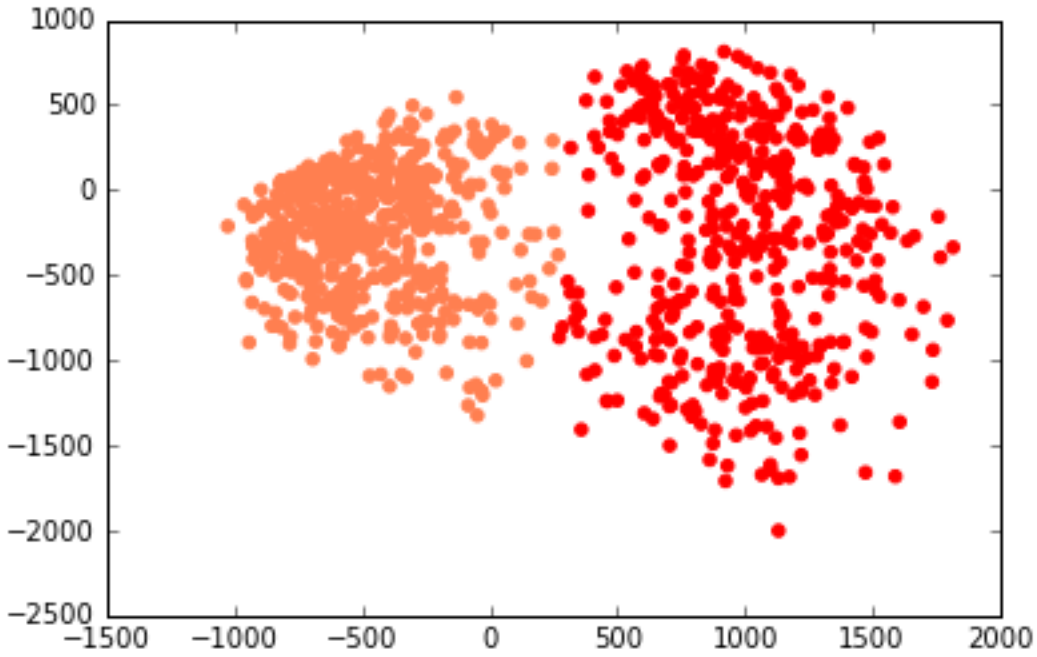
SC: But As discussed, from SC graph, we see that the value of SC reaches a peak point at k=2, and after k=2, the value of SC is decreased, so we choose k=2 from SC graph.

So, only considering WC graph, we might choose a different value of k, because WC graph does not provide sufficient and clear indication about the value of k to be picked.

B.4:

We chose value of k as 2

NMI : 0.4609830869805



Analysis: As we know that the maximum value of NMI reaches 0.5 in our implementation, and in this case we see that the value of NMI is coming as 0.4609830869805. This NMI value indicates that the purity of the clusters has not reached its maximum. And from the visualization graph, we can see that the two clusters are well-separated except some points .So there are slight chances in clustering points in a particular cluster into a false group.

Comparing to tSNE: We observed that in tSNE clustering, the clusters were very well-separated, and the purity of the clusters were comparable high than this graph. So tSNE clustering was better than clustering using pca.