HW4 Lichen Liang

1.(a) Assume P1, P4, P7 are centroids of the three clusters

$\nearrow$ euclidean distance

P2: $d(P2,P1) = \sqrt{(2-2)^2 + (5-10)^2} = \sqrt{25}$

$d(P2,P4) = \sqrt{(2-5)^2 + (5-8)^2} = \sqrt{18}$

$d(P2,P7) = \sqrt{10}$ *

P3: $d(P3,P1) = \sqrt{72}$

$d(P3,P4) = \sqrt{25}$ *

$d(P3,P7) = \sqrt{53}$

P5: $d(P5,P_1) = \sqrt{89}$

$d(P5,P4) = \sqrt{40}$

$d(P5,P7) = \sqrt{36}$ *

P6: $d(P6,P1) = \sqrt{52}$

$d(P6,P4) = \sqrt{17}$ *

$d(P6,P7) = \sqrt{29}$

P8: $d(P8,P1) = \sqrt{5}$

$d(P8,P4) = \sqrt{2}$ *

$d(P8,P7) = \sqrt{58}$

P9: $d(P9,P1) = \sqrt{50}$

$d(P9,P4) = \sqrt{13}$ *

$d(P9,P7) = \sqrt{45}$

cluster 1: P1
cluster 2: P4, P3, P6, P8, P9
cluster 3: P7, P2, P5

b) $P2: d(P2,P1) = |2-2| + |5-10| = 5$
$d(P2,P4) = |2-5| + |5-8| = 6$
$d(P2,P7) = 4$ *

$P3: d(P3,P1) = 12$
$d(P3,P4) = 7$ *
$d(P3,P7) = 9$

$P5: d(P5,P_1) = 13$
$d(P5,P4) = 8$
$d(P5,P7) = 6$ *

$P6: d(P6,P1) = 10$
$d(P6,P4) = 5$ *
$d(P6,P7) = 7$

$P8: d(P8,P1) = 3$
$d(P8,P4) = 2$ *
$d(P8,P7) = 10$

$P9: d(P9,P1) = 10$
$d(P9,P4) = 5$ *
$d(P9,P7) = 9$

Cluster 1: P1
Cluster 2: P4, P3, P6, P8, P9
Cluster 3: P7, P5, P2

Clusters remained the same from part a)

c) Clustering depend on the data's distribution, sparsity, density, size, and shape. So one method cannot work for all kinds of data.
For example, the first row, if a centroid in kmeans is randomly chosen, then it measures the distance between a point to the centroids and the data is clustered wrong. In this case hierarchical clustering worked well.

2. a) k means involves picking k points as centers and each other point should belong to one cluster only. Then the distance between a center and the point should be minimal (error should be minimal). Therefore we want the sum of all the errors to be minimal

b) we can see that $x \in C_i$ meaning each point x is center of its Cluster itself. Therefore the distance $(X - C_i)$ is always 0 thus the SSE.