Tiffany Wong A20442087
CS422 - HW 8 Exercises 11/21/2021

1. Exercises
1.1 (2 points) **Tan, Chapter 7 #7, 16**

7. Suppose that for a data set

- *there are m points and K clusters,*
- *half the points and clusters are in "more dense" regions,*
- *half the points and clusters are in "less dense" regions, and*
- *the two regions are well-separated from each other.*

For the given data set, which of the following should occur in order to minimize the squared error when finding K clusters:

a. *Centroids should be equally distributed between more dense and less dense regions.*
b. *More centroids should be allocated to the less dense region.*
c. *More centroids should be allocated to the denser region.*

Note: *Do not get distracted by special cases or bring in factors other than density. However, if you feel the true answer is different from any given above, justify your response.*

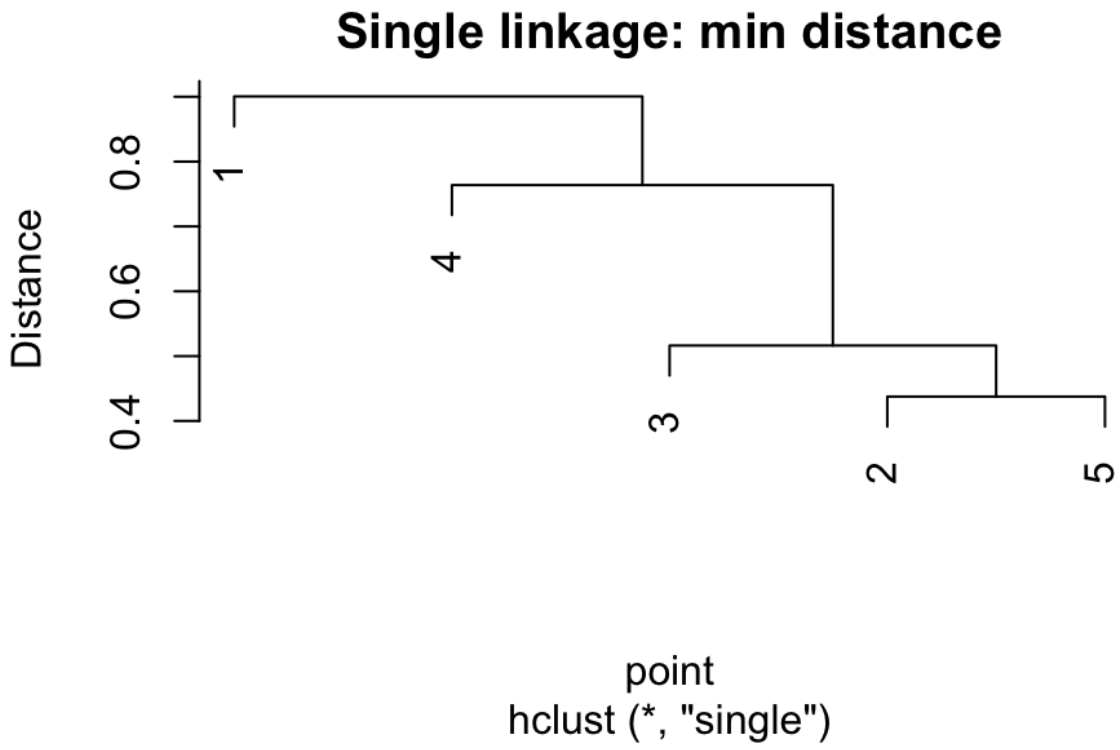
To minimize the squared error with finding K clusters, you have to focus on the dense regions. Option (c) would allow for a higher proportion of points to have lower squared errors in the denser regions, meaning the squared estimate of errors (SSE) would be minimized. The thinking behind this is that if you put the centroid in the denser region, then squares of a lot of the data points will be small. But if you put a lot of the data points in a less dense region, then all of the data points from the dense region will contribute high squared distances. Essentially, option (c) will minimize the distance, therefore minimizing the squared error when finding K clusters.

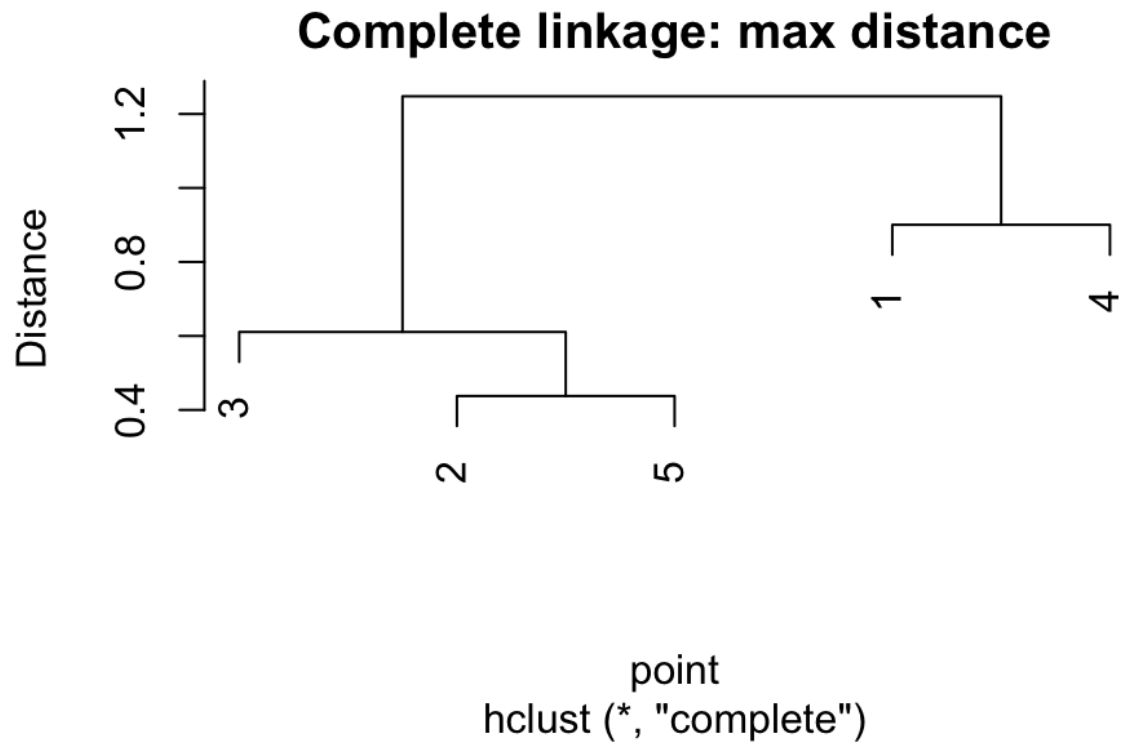
16. Use the similarity matrix in Table 7.13 to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	1.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

(For 16, note that Table 7.13 for Exercise 16 has a similarity matrix, not a distance matrix. Similarity and distance are related to each other by the formula *distance = 1.0 – similarity*.)

I used the hclust() function in R to draw the dendrogram.
I used single and complete linkage to contrast the minimum distance with the maximum distance of the data points.





My code:

```r
library(ggplot2)

A = matrix(c(0.00, 0.10, 0.41, 0.55, 0.35, 0.10, 1.00, 0.64, 0.47,
0.98, 0.41, 0.64, 1.00, 0.44, 0.85, 0.55, 0.47, 0.44, 1.00, 0.76,
0.35, 0.98, 0.85, 0.76, 1.00), nrow=5, ncol=5, byrow = TRUE)

D = 1 - A

D <- dist(D[], diag=TRUE)

hcsingle <- hclust(D,method='single')
plot(hcsingle, xlab="point", ylab="Distance", ylim=c(0,1),
main="Single linkage: min distance")


hccomplete <- hclust(D, method='complete')
plot(hccomplete, xlab="point", ylab="Distance", ylim=c(0,1),
main="Complete linkage: max distance")
```