1) Briefly describe the hierarchical and partitioning clustering approaches, including their advantages and disadvantages.

Clustering methods try to produce results that have high intra-class similarity with low inter-class similarity. Also, the a quality clustered model with show some hidden patterns.

There are many different methods that clustering can be done to achieve these goals. I will give a brief outline of a few of these methods.

Hierarchical:

With this approach you take N samples into C clusters. It starts at the head where everything belongs to one group to the base where everything is its own group. With each level from the head more samples are divided into smaller clusters until the groups are just one sample by themselves. Each of these clusters says that they have something in common more so than the samples that are not part of the cluster. A great benefit with this approach is that more information can be derived from the data since there is information at each level. However, with this approach it is very slow to the point where depending on the data and the time frame to calculate it the approach could be cost prohibited .

Partitioning(K-Means):

The idea with K-means is that it will take N objects and put them into K groups. It does this by finding groups that are close to each other over all dimensions. The algorithm will minimize the overall distance of all objects from their k-means center group. With this method it is difficult to know how many k groups the data should produce and need to be specified by the user. Also, the data needs to be scaled to each other so each dimension is being represented fairly. A great benefit with this approach is that the algorithm is fast and scales well.

2) Give an example how clustering can be used....

A Japanese candlestick is a shape given off of four data points on a given timeframe in a financial data set(open, high, low, close). Traders use these shapes to give insight on where the buyers and sellers are in the market. Using clustering techniques one can possibly develop support for using this trading tool. Before, processing this data the price data needs to be independent of the previous price before it. Thus, for the following examples the first step is to offset the open, high, low and close by the previous close.

a) As a pre-processing step

One method to test the idea of trading candlesticks would be to use a clustering technique as a pre-process step to feed into the association rules. Once the data is offset by the previous close take each candlestick data apply k-means to make them into groups. Take these groups and feed it into association rules program with the rules on the left side being the candlestick group and which are present. On the right side of the association rules would be the next candlestick group.

b)As a post processing step

Ask traders to group different candlesticks and patterns. The traders may disagree with each other about when and where patterns are at times. With the given results feed the data into a MDS statistical program. Then feed the eigendecomposition data in a k-mean program to group the candlesticks and patterns across different dimensions into groups.

c) As an investigatory tool in its own right

Using a clustering technique to make many groups with the candlesticks. Then you can look at the distribution of the high, low and close of the next candlestick. With these distribution look for any time the data is not normally distributed or say if the mean of the close is not 0 and has a bias towards a direction.

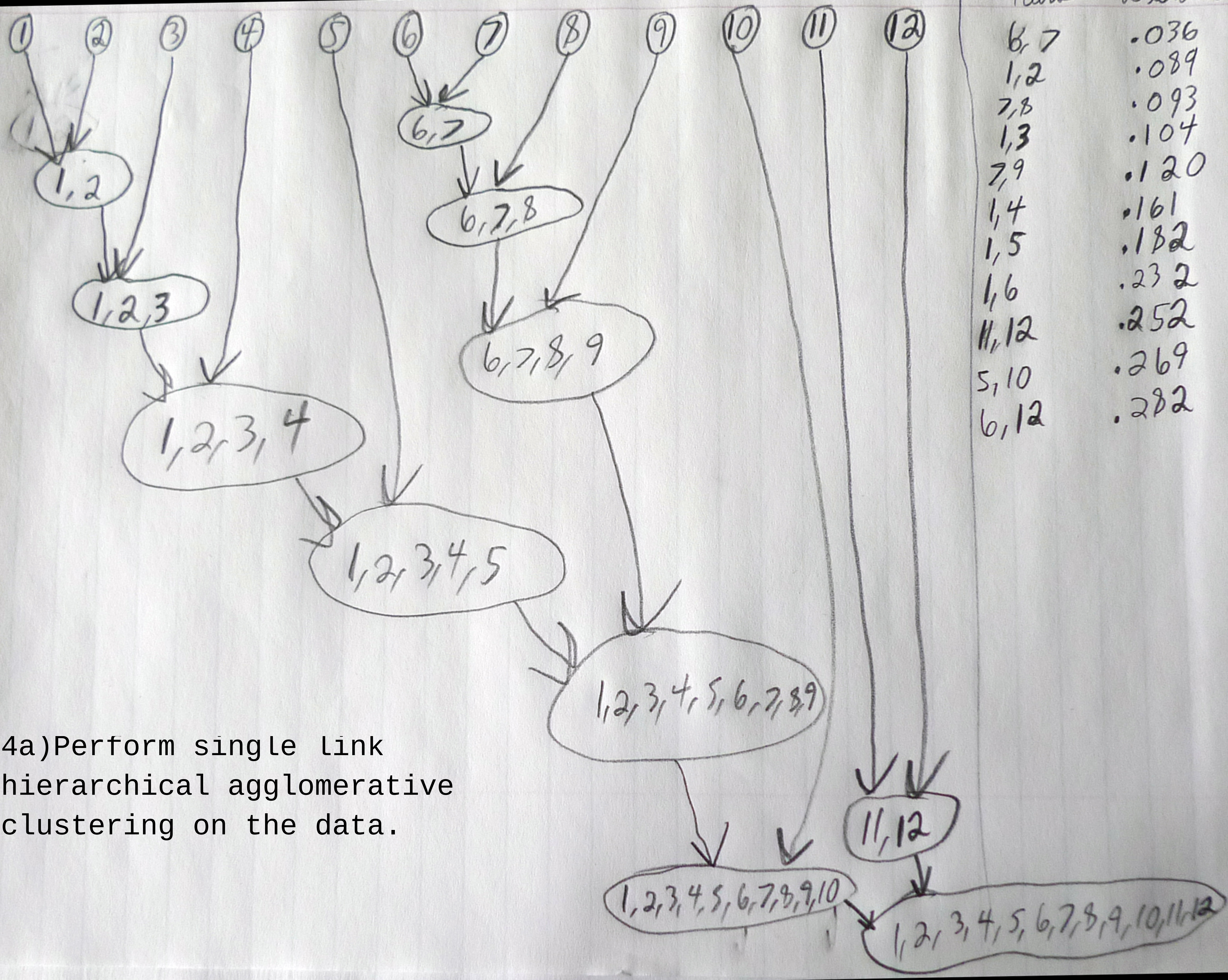
3) Give three distant ways we might evaluate quality of a clustering.

1) High quality clusters will have high intra-class similarities.

2)High quality clusters will have low inter-class similarities.

3)High quality clusters will show some or all hidden patterns in the data.

4a)Perform single link hierarchical agglomerative clustering on the data.



6,7	.036
1,2	.089
7,8	.093
1,3	.104
7,9	.120
1,4	.161
1,5	.182
1,6	.232
11,12	.252
5,10	.269
6,12	.282

4bi) How are primates grouped?

They are grouped by the shortest distance they have in the matrix. Since the hierarchical cluster is of a single link it is form by the two pairs having the shortest distance to each other. Then what group they both currently belong to get grouped together. Here are the shortest paired links at each level.

pairs	distances
6,7	0.036
1,2	0.089
7,8	0.093
1,3	0.104
7,9	0.12
1,4	0.161
1,5	0.182
1,6	0.232
11,12	0.252
5,10	0.269
6,12	0.282

4bii) Into how many clusters would you break the dendogram? Why?

I would stop the grouping right before the group 5 and 6 where added together with each other creating the massive group 1,2,3,4,5,6,7,8,9. Thus, the groups I would have clustered are [(1,2,3,4,5),(6,7,8,9),(10),(11),(12)]. The reasons I would have stopped here is that there are a few different clusters without having most of groups belong into one cluster.

4biii)Does the structure of the dendogram tell you anything?

Sure each level of the dendogram tells you a small message. The strongest messages are of the clusters that form in the beginning steps. Thus, from this structure I would say you can see that pairs (6,7), (1,2) , (7,8), and (1,3) form the clusters (6,7,8) and (1,2,3). These two clusters have a high intra-class similarities and low inter-class similarities between eachother.