Task 4 Part a:

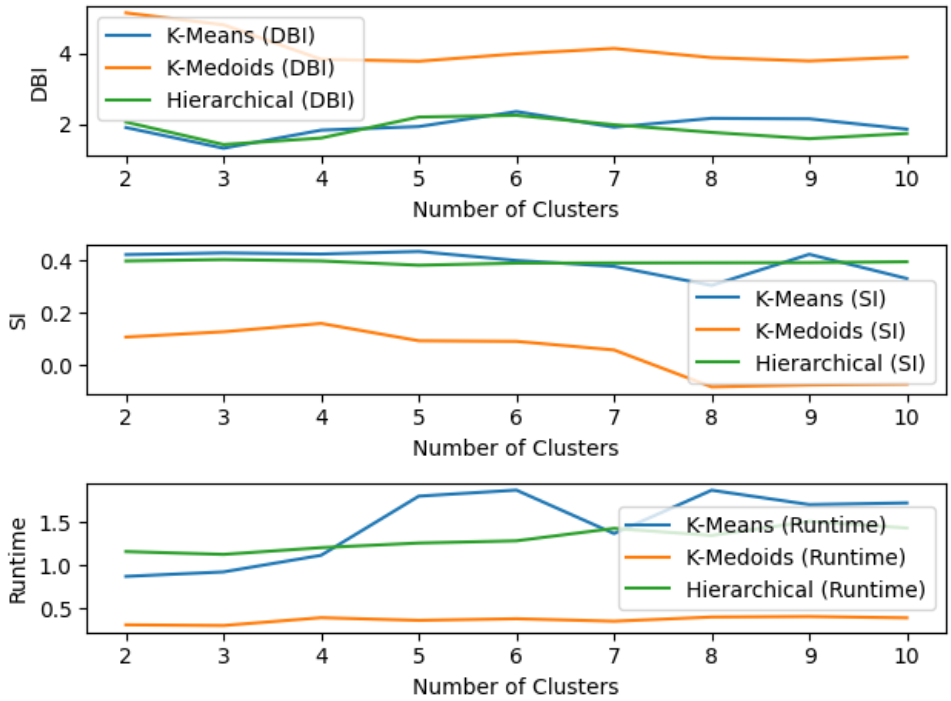


Figure 1: Unlabeled with K = 10

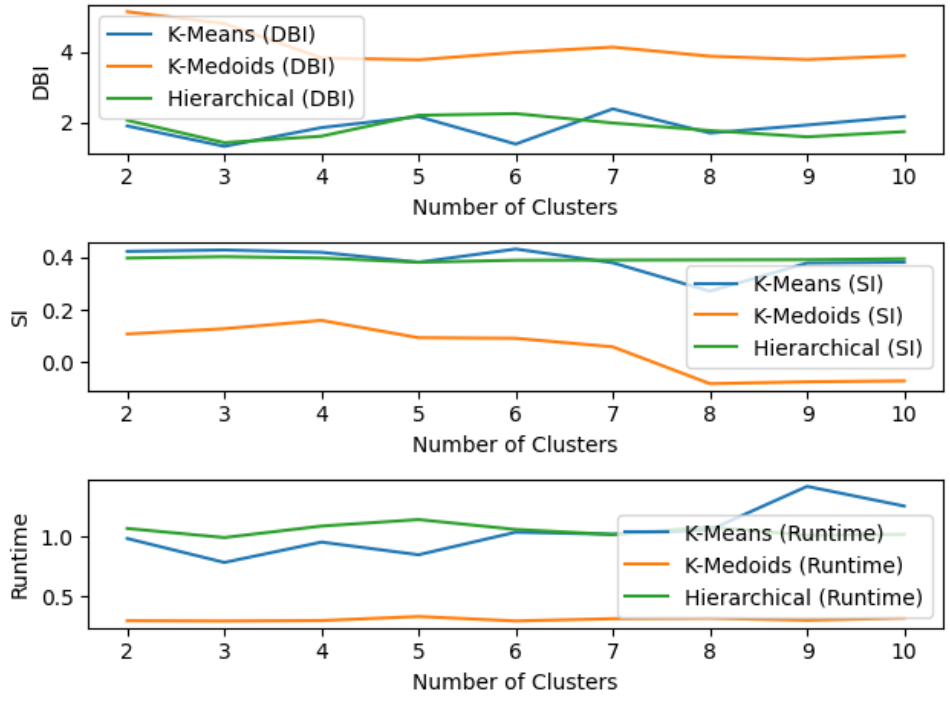


Figure 2: Labeled with K = 10

Part b:

We want to have the lowest DBI (Davies Bouldin Index) and the highest SI (Silhouette Index) for the best performance algorithm.

For K-means and Hierarchical clustering for unlabeled data, they have a very similar performance which have a similar low DBI and high SI with a similar running time, so it is hard to say when you are using unlabeled data with one is better. But it is easy to tell that K-medoid had the worst performance for all three algorithms, which it has high DBI with a low SI, only thing it is better that other algorithms are the running time is way faster than other 2 algorithms. But overall for unlabeled data I thing Hierarchical has a best performance

Part c:

| K # | K-mean (second) | K-medoids (second) | Hierarchical (second) |
|---|---|---|---|
| 2 | 0.75 | 0.24 | 0.89 |
| 3 | 0.69 | 0.25 | 0.89 |
| 4 | 0.74 | 0.25 | 0.89 |
| 5 | 0.76 | 0.25 | 0.93 |
| 6 | 0.85 | 0.25 | 0.92 |
| 7 | 0.89 | 0.25 | 0.89 |
| 8 | 1.00 | 0.25 | 0.88 |
| 9 | 1.12 | 0.26 | 0.99 |
| 10 | 1.19 | 0.27 | 0.94 |
| Total | 7.99 | 2.26 | 8.20 |

For Hierarchical K=6 had the best performance, which K-mean use 0.85s, K-medoids use 0.25s and Hierarchical use 0.92s.

K-mean is very similar hard to tell, for K-medoids is the k=2

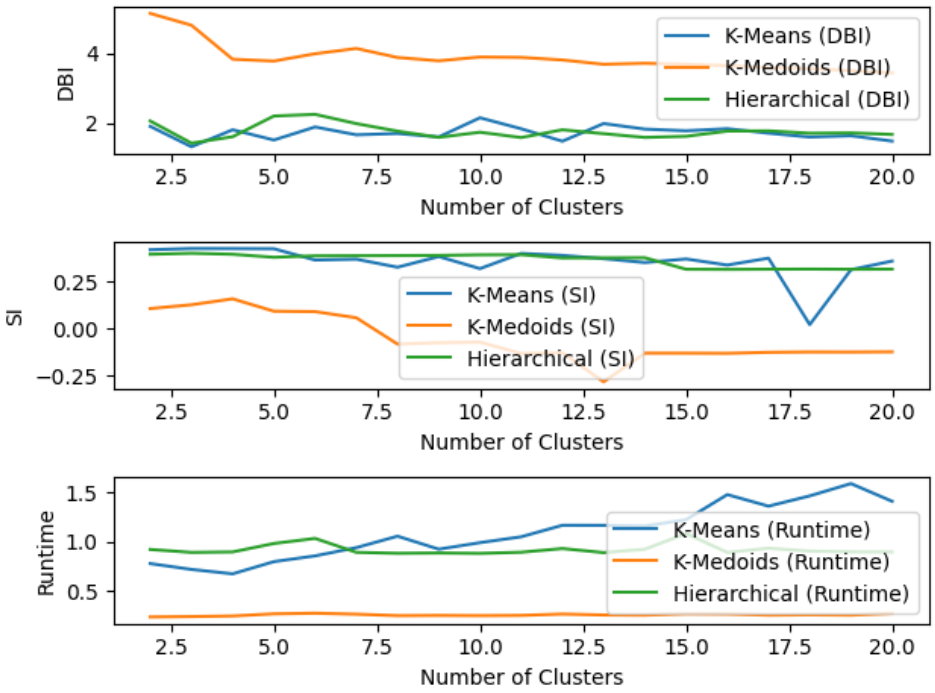


Figure 4: Unlabeled K= 20

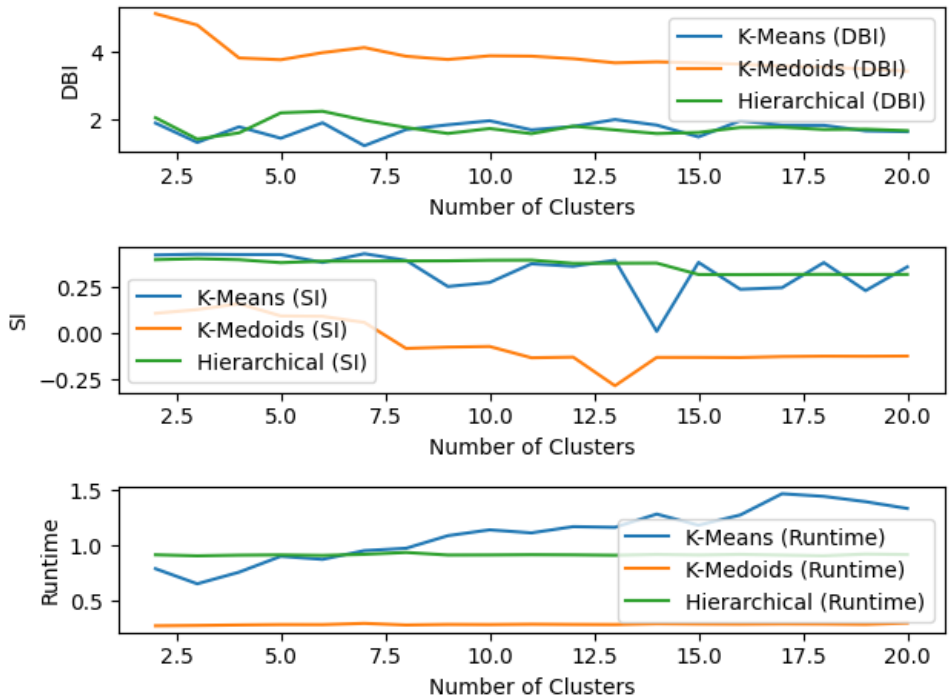

Figure 4: Labeled, K =20

Part d:

To answer part d and e, I am introducing figure 3 and figure 4.

From figure 3 and figure 4 we can see that when K increases K-mean's running time started increasing very fast (K > 7). And at the same time K-mean's SI score had rise and fall for both types of data. K-medoids are always the worst so we do not consider it at. Overall, based on running time and stable performance, I recommend Hierarchical. But K-mean has some local optimal solutions, it can be considered too for special requirements.

Part e:

Labeled were poorly. Did not have the output data but from the graph, looks like K-medoids and Hierarchical looking similar score for DBI and SI for both labeled and unlabeled, but K-mean labeled looks like it has the worst SI score. SI measures the similar an object is to its own cluster compared to other clusters, so worst SI means poorly the data is.