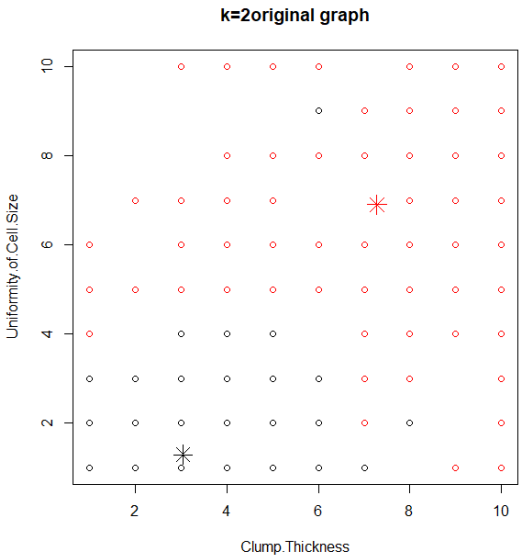
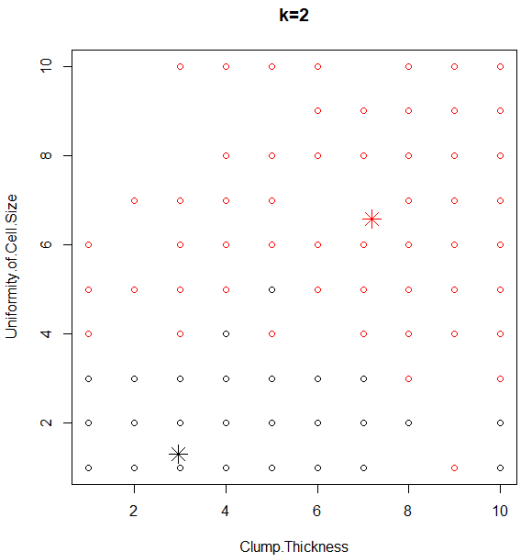
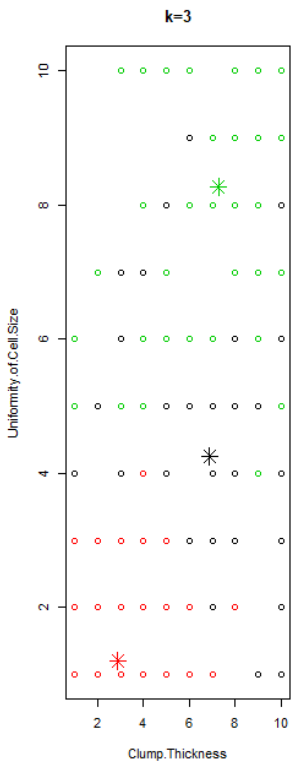
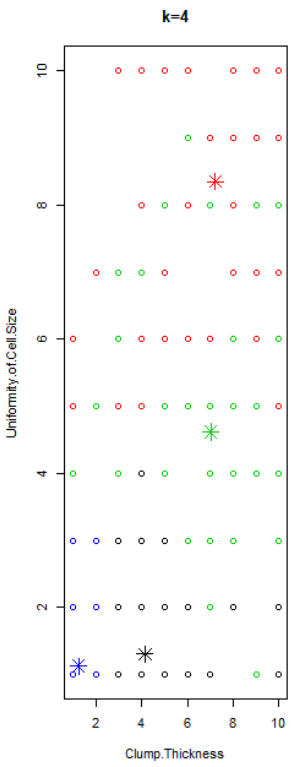
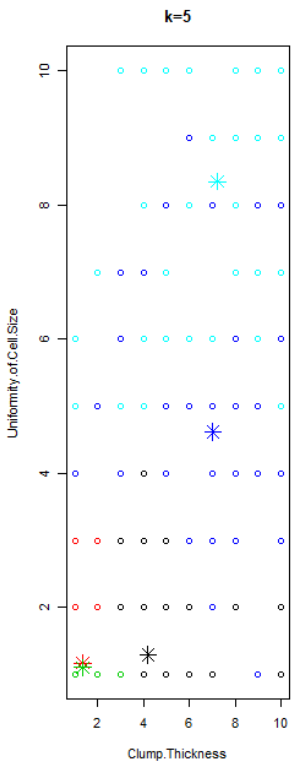
After the initial cleanup of the first task, the k means clustering method have been used to cluster the data in 2 clusters based on all the variables except 'class'. As we can see below, the original class-based classification and the results from the kmeans clustering cluster are not much different visually on the clump thickness and uniformity of cell size dimensions. Only a few points are not assigned in the places it supposed to be.
And then, the parameter k have been changed into 3, 4, 5 to compare the SSE(Sum of Square Error) to compare the effectiveness of each model in the case of different k.







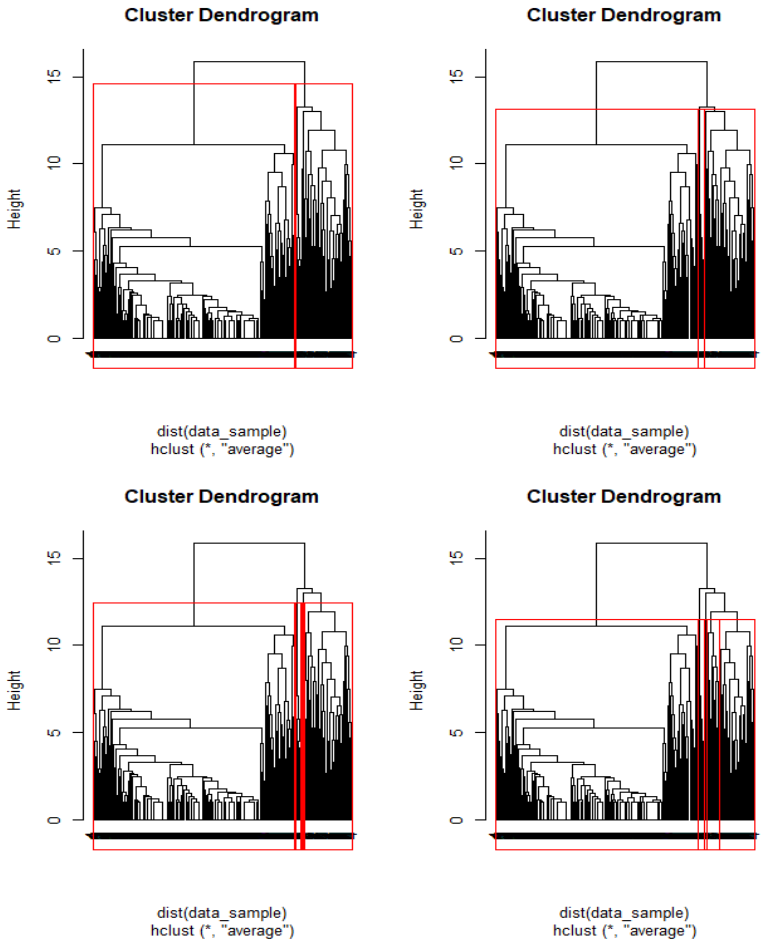




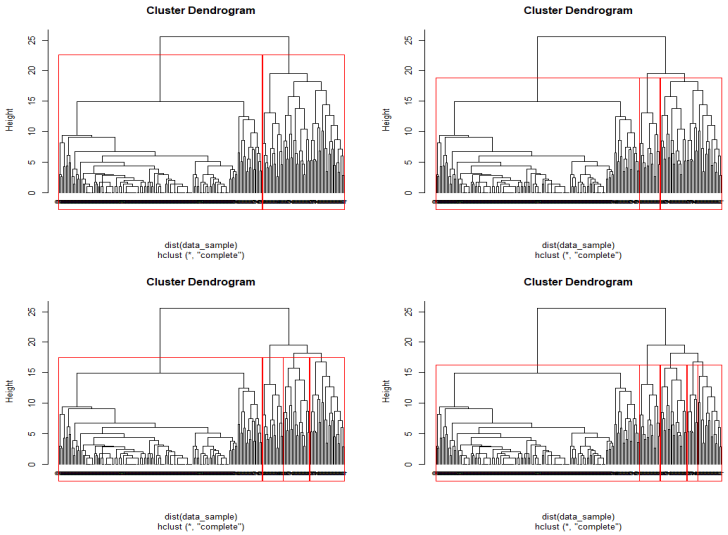
	sse
1	NA
2	18829.18
3	15747.95
4	14871.51
5	13640.28

After building the models using different K, the SSE have been calculated respectively. The SSE is getting lower when k is getting bigger. The SSE drops dramatically when k changes from 2 to 3, which means k = 3 would be the most appropriate model when using k means clustering algorithm. But actually, there are only 2 classes for now, we can do some further scientific study to find out whether there is a new type of class or whether a subclass can be divided from a super class based on the k means result. After that the hierarchical clustering was applied using different agglomeration methods and try to set the different n clusters to compare performances.

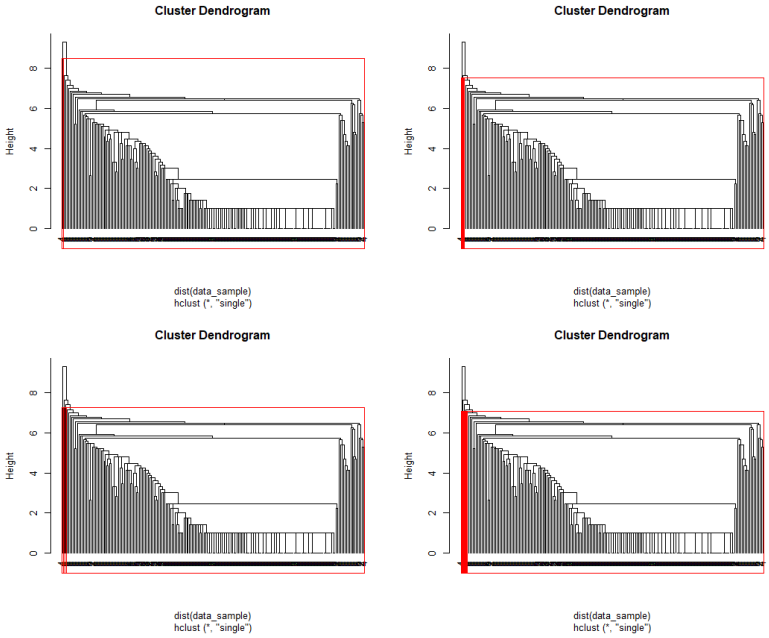
Based on the plots when n cluster= 2 and 3, we can clear see that the second cluster can be divided into 2 subtypes, but not necessarily have to. If to divide the disease types into 3 clusters, the number of the third one might relatively less. We can see the disease can basically be divided into 2 types , but can also have a subtype under the second type if the theoretical study has proven that it is necessary. The outcomes of using "single" method have great differences with using other 2 methods. And the outcomes of "complete" and "average" are basically the same. In my opinion, using average agglomeration method is the best, because its cluster structure looks more reasonable. Branches have been clustered from higher level to lower level.



(Using average method)



(Using complete method)



(Using single method)

To do the classification, before build the model, we need to know which variables are relevant. Assume all the variables are obey to normal distribution, t-test was used to determine whether the variable should put into the model. For each variable, there is a statistical difference between the different classes. So, all the variables should put into the model. So, all the variables should put into the model. a tree model to predict the class was built and the accuracy, precision, and recall was calculated. As shown below, the tree model to predict the class2 performs better than class4. And totally, the model is good and the accuracy is 91.7%.

```
> accuracy
[1] 0.9170732
> eva_result
  precision    recall        f1
2  0.964539 0.9189189 0.9411765
4  0.812500 0.9122807 0.8595041
```

And then, use knn method to do the classification. The k was set to 1-5 to evaluate the performances when using different n. The results show that there is no big difference among all the models and the accuracy is only 68.78%

```
> conclusion=cbind.data.frame(accuracy,precision,recall,f1)
> conclusion
   accuracy precision    recall        f1
1 0.6878049 0.9791667 0.9527027 0.9657534
2 0.6829268 0.9722222 0.9459459 0.9657534
3 0.6878049 0.9791667 0.9527027 0.9657534
4 0.6878049 0.9791667 0.9527027 0.9657534
5 0.6926829 0.9861111 0.9594595 0.9657534
```