Assignment5

Student: Lina Mi

1. Load `iris.csv` to R studio:

```
>iris<-read.csv("M:/Data Mining/week5/iris.csv")
```

2. Display `iris` data frame:

```
>iris
    X5.1 X3.5 X1.4 X0.2     Iris.setosa
1    4.9  3.0  1.4  0.2     Iris-setosa
2    4.7  3.2  1.3  0.2     Iris-setosa
3    4.6  3.1  1.5  0.2     Iris-setosa
4    5.0  3.6  1.4  0.2     Iris-setosa
```

3. Rename the columns of `iris` data frame:

```
>colnames(iris)<-c("sepal_length","sepal_width","petal_length","petal_width","class")
```

4. Display `iris` data frame with renamed columns:

```
>str(iris)
'data.frame':    149 obs. of  5 variables:
 $ sepal_length: num  4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 5.4 ...
 $ sepal_width : num  3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 3.7 ...
 $ petal_length: num  1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 1.5 ...
 $ petal_width : num  0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 0.2 ...
 $ class       : Factor w/ 3 levels "Iris-setosa",..: 1 1 1 1 1 1 1 1 1 1 ..
```

5. Select features of `sepal_length` and `sepal_width` for clustering in k means algorithm:

```
>kmeans_sepal<-data.frame(iris$sepal_length,iris$sepal_width)
```

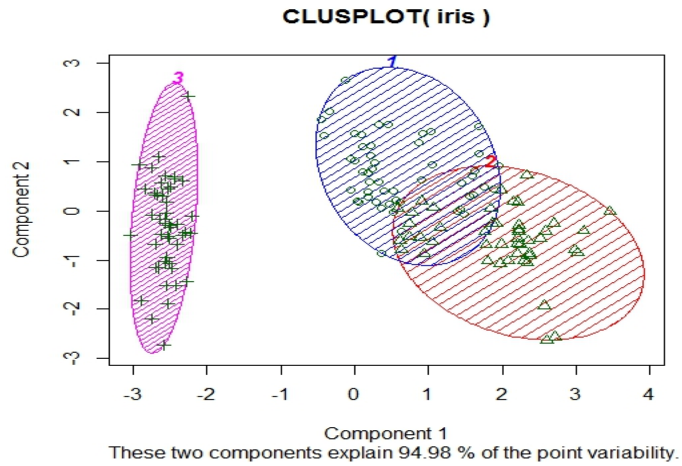
6. for `kmeans_sepal` variable, run k means algorithm with number of clusters equal 3:

```
>set.seed(1000)
> sepal<-kmeans(kmeans_sepal, 3)
```

7. plot the outcome of the clustering model using `clusplot:`

```
>clusplot(iris, sepal$cluster, color=TRUE,shade=TRUE, labels=5,lines=0)
```

The output plot is shown as following:



The clustering of iris data can also be seen with following R command

```
1) sepal$cluster
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3
  [45] 3 3 3 3 3 2 2 2 1 2 1 2 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 2 2 2
 1 1 1 1 1 1 1 1 2 1 1
  [89] 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2 2 2 1 1 2 2 2 2 1 2 1
 2 1 2 2 1 1 2 2 2 2 2
[133] 1 1 2 2 2 1 2 2 2 1 2 2 2 1 2 2 1

2) >sepal$size
[1] 53 47 49
```

From the clustering result using sepal_length and sepal_width features and cluster number of 3, it can be seen the clustering is fair good, the clustering group size is close to the class size indicated by the class column of iris data frame (which is 49, 50,50) though there are miss-clustered points in group 1 and group2

8. Select features of petal_length and petal_width for clustering in k means algorithm

```
>kmeans_petal<-data.frame(iris$petal_length,iris$petal_width)
```

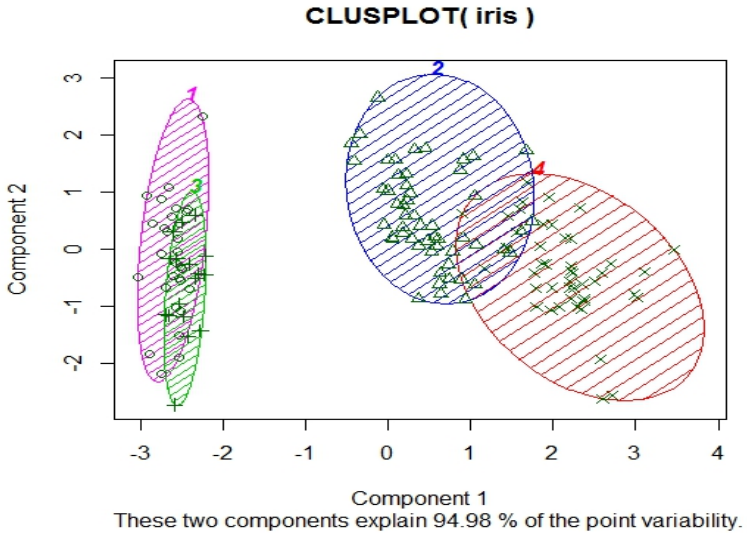
9. for kmeans_petal variable, run k means algorithm with number of clusters equal 4:

```
>set.seed(1000)
> petal_4=kmeans(kmeans_petal,4)
```

10. plot the outcome of clustering model using clusplot:

```
 > clusplot(iris, petal_4$cluster, color=TRUE, shade = TRUE, labels=5, lines=0)
```

The output plot is shown as following:



From the figure above, it can be seen that the group 1 and group3 overlapped greatly, that means the dissimilarity between group 1 and 3 decrease greatly with cluster number of 4 in k means algorithm, the cluster result using `sepal_length` and `sepal_width` and cluster number of 3 is more accurate since in that clustering, the similarity within each cluster is high and dissimilarities among different clusters are relatively high.

11. I would not consider training other models with a different cluster number, since the `class` column of iris data frame indicates that there are three classes in the data set. So using the cluster number of 3 in k means algorithm will get most accurate clustering result. `Class` value of each row of `iris` data frame can be used to examine the accuracy of clustering result of k-means algorithm. In other case where the classification information is not available, I will use different cluster number in the k means algorithm to get more accurate group results.