APSTA-GE 2011 Project #1

Name: Yuan Ding
NetID: yd1400

Part 1: Features Exploration and Standardization

Code:

Use hist() function to draw histograms

Use describe() function in library(psych)

For the data preparation part, first let us draw histograms to explore the distribution of each feature in our dataset:

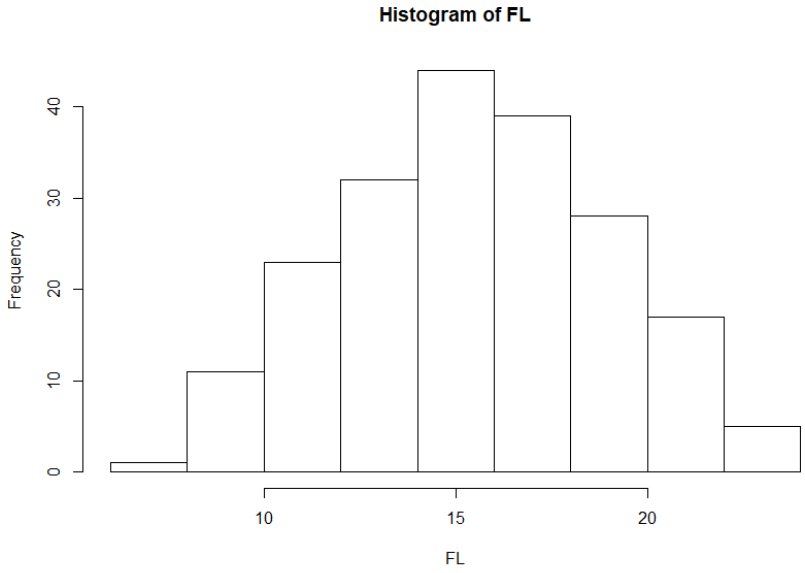


Fig. 1-a Histogram of FL

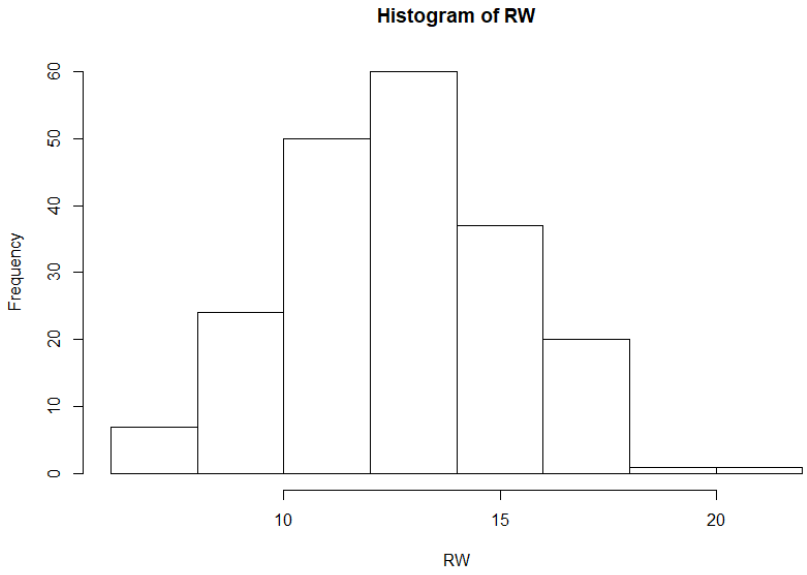


Fig. 1-b Histogram of RW

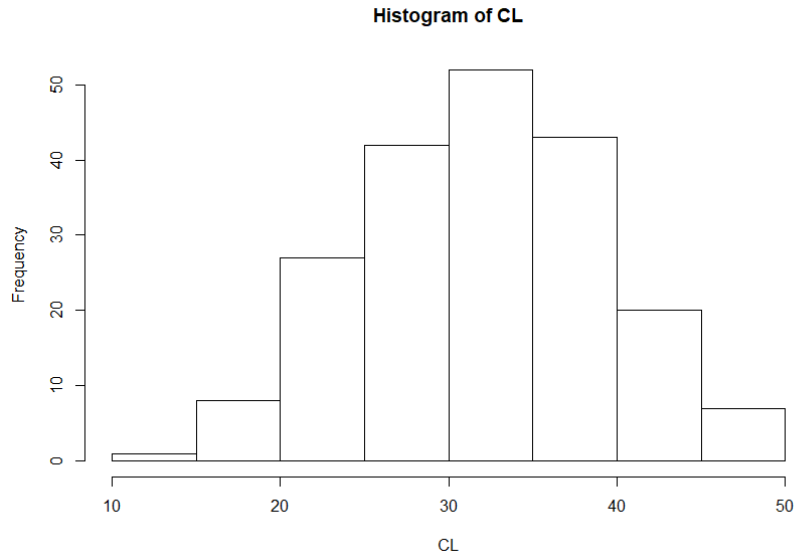


Fig. 1-c Histogram of CL

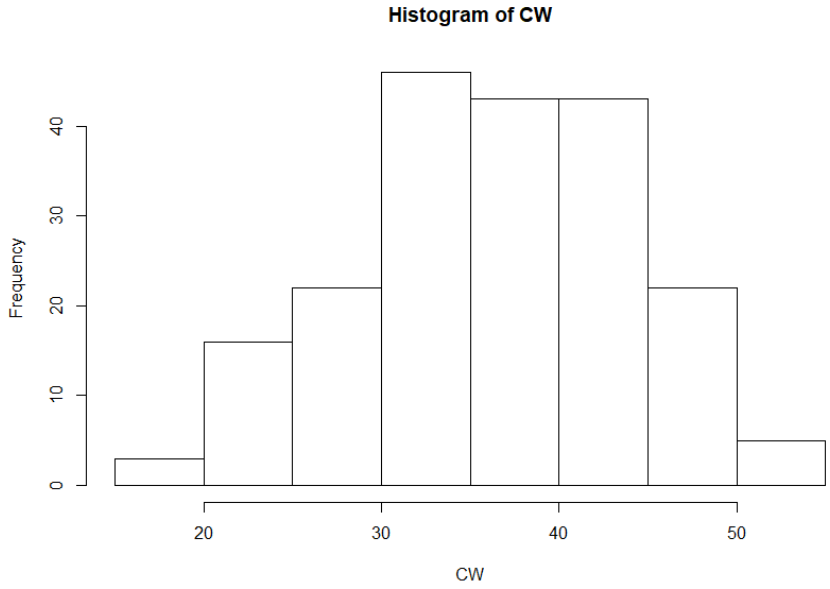


Fig. 1-d Histogram of CW

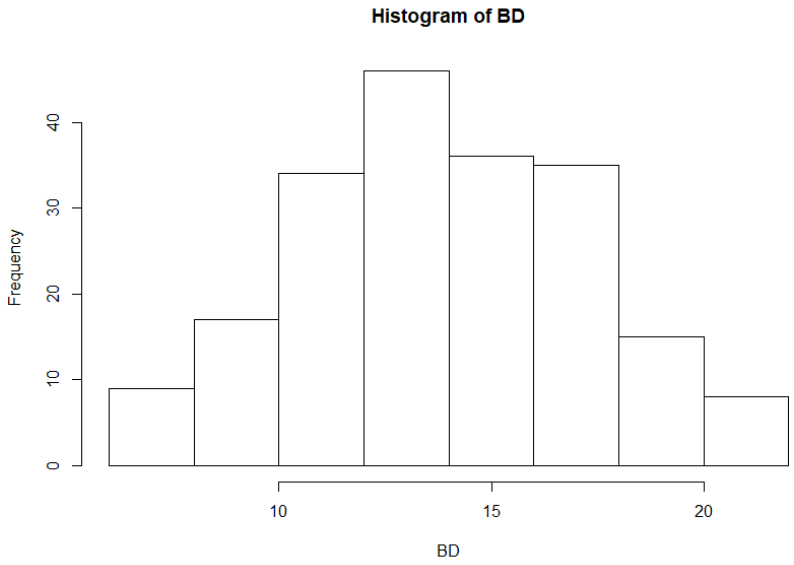


Fig. 1-e Histogram of BD

From the above histograms we can see that the distribution of all the five features are approximately normal distributions. The range and standard deviation of data are not quite large so we do not have to take logs.

Second, we need to discover some statistics about the distribution of our five features. Here we use describe() function:

Table 1 The Statistics on Five Crab Features (unit: mm)

Features	Mean	SD	min	max	range
FL	15.58	3.5	7.2	23.1	15.9
RW	12.74	2.57	6.5	20.2	13.7
CL	32.11	7.12	14.7	47.6	32.9
CW	36.41	7.87	17.1	54.6	37.5
BD	14.03	3.42	6.1	21.6	15.5

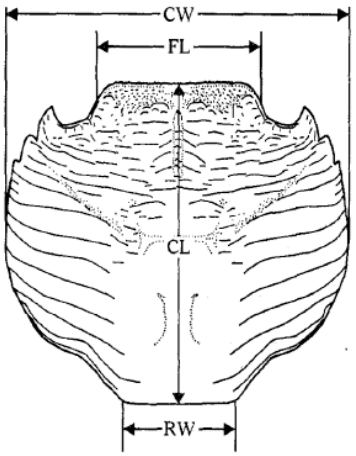


Fig. 1. Dorsal view of carapace of *Leptograpsus*, showing measurements taken. *FL*, width of frontal region just anterior to frontal tubercles. *RW*, width of posterior region. *CL*, length along midline. *CW*, maximum width. The body depth was also measured; in females but not in males the abdomen was first displaced.

Fig.2 The Measurements for Crabs (cited from the slides)

We see from the above table that although all the data are in millimeters, the range of CL and CW (32.9, 37.5) are more than twice the range of FL(15.9), RW(13.7) and BD(15.5), and their standard deviations are also larger. From the Fig.2 and the common sense we can know that for a crab, CL and CW are usually larger than FL and RW, and those features with higher variance (standard deviation) are not more important than features with lower variance (standard deviation). In this crab case, the importance of features is independent of the variance of the features, and we would better standardize the original data so that the data for five features can have the same scale.

Part 2: Principal Components Analysis (PCA)

Code:

```
> princomp(crabs.stdz[,4:8],cor = T)
Call:
princomp(x = crabs.stdz[, 4:8], cor = T)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
2.18834065 0.38946785 0.21594669 0.10552420 0.04137243

 5  variables and  200 observations.
```

```
> loadings(princomp(x = crabs.stdz[, 4:8], cor = T))

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
FL  0.452  0.138  0.531  0.697
RW  0.428 -0.898
CL  0.453  0.268 -0.310        -0.792
CW  0.451  0.181 -0.653         0.575
BD  0.451  0.264  0.443 -0.707  0.176

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
SS loadings       1.0    1.0    1.0    1.0    1.0
Proportion Var    0.2    0.2    0.2    0.2    0.2
Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

First we use the function princomp to generate the principle components based on the newly standardized data. Then we draw bivariate plots based on the new five components. From the bivariate plots using original features (see Fig.3-a), we can see that the clusters are “stringy”, but they also closely connected and not fairly well separated. From the bivariate plots using principal components (see Fig.3-b), things become better, but the clusters are still not fairly separated and connected to each other. So based on the result of bivariate plots, we would choose to use the **centroid linkage** method hierarchical clustering. We will verify the choice again using the dendrograms of both the methods.

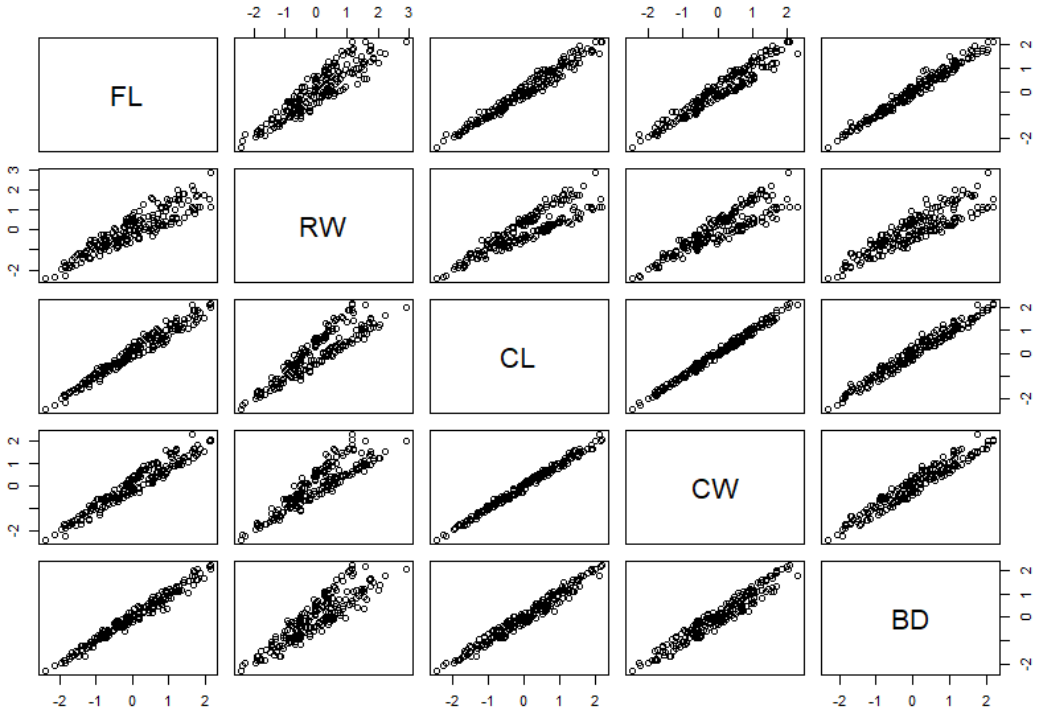


Fig.3-a The Matrix of Bivariate Scatterplots (original variables)

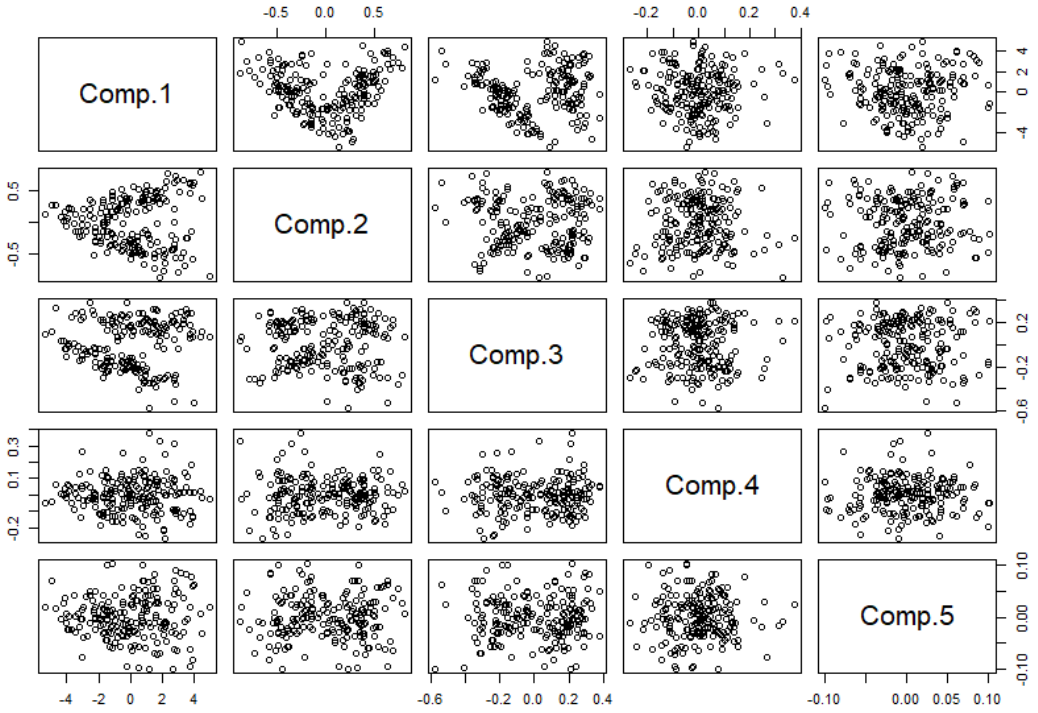


Fig.3-b The Matrix of Bivariate Scatterplots (PCA)

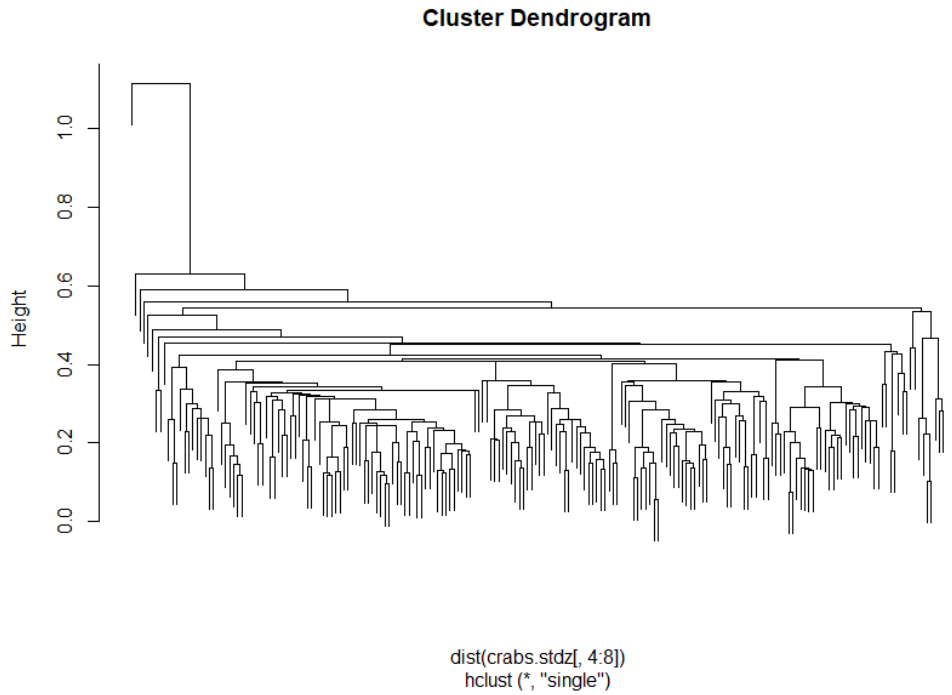


Fig.4-a Dendrogram of Single Linkage

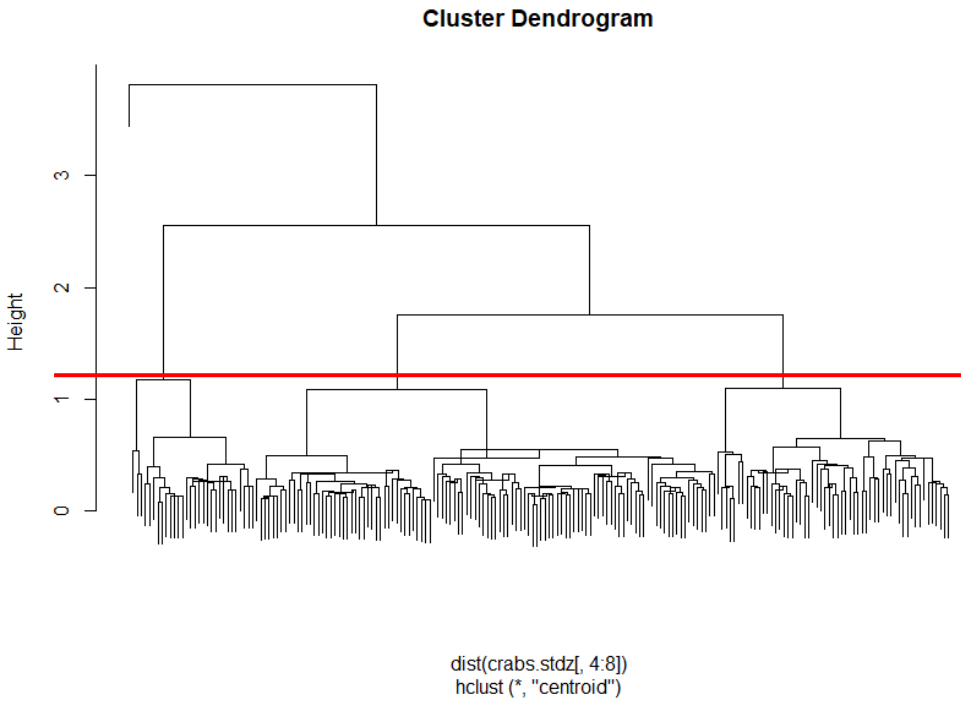


Fig.4-b Dendrogram of Centroid Linkage

From the above dendrograms of both **single linkage** and **centroid linkage**, we can see that the method of **centroid linkage** can give us more fairly separated clusters. The dendrogram shows that the clusters of **single linkage** are not fairly separated, and can have too many small clusters which only have one or several points. So

we confirm again that in the crab case, the method **centroid linkage** is a better choice for hierarchical clustering.

Part 3: Hierarchical Clustering and K-means Clustering

1. Hierarchical Clustering

(1) Choose the number of clusters

By examining the dendrogram, the number of 4 would be a perfect choice. We draw a horizontal line above Height 1 (see Fig.4-b) and we can see that in this horizontal level, there are four group which are pretty well separated. Because the height is only about 1, the distance between points within each group won't be that large and the homogeneity within each group is still pretty good. From the dendrogram, we can see that number 3 or 7 is also a good cut point of getting fairly separated clusters. We will evaluate the alternative choices using the following plots.

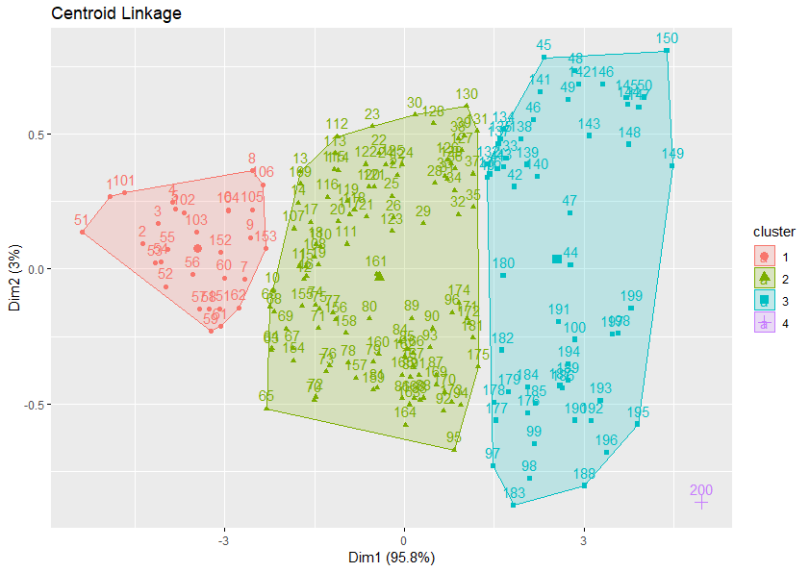


Fig.5-a Assigned Cluster Labels (centroid linkage, k=4)

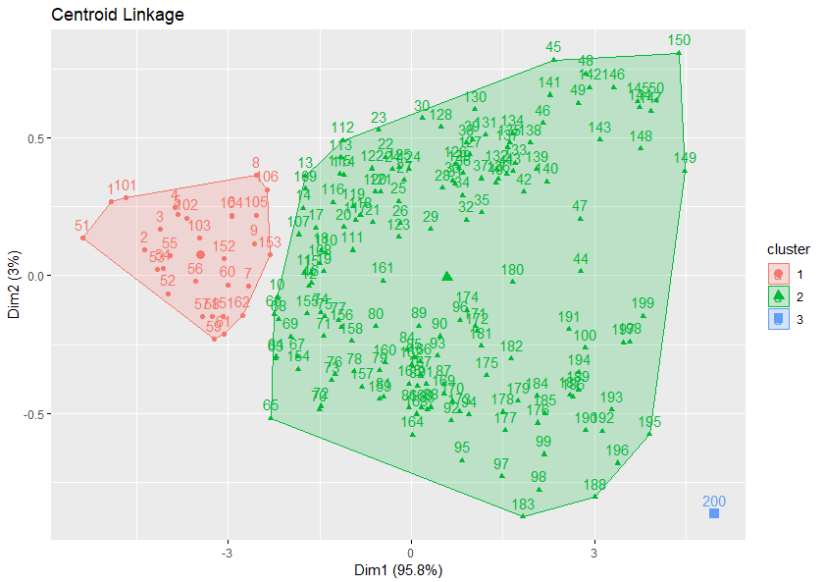


Fig.5-b Assigned Cluster Labels (centroid linkage, k=3)

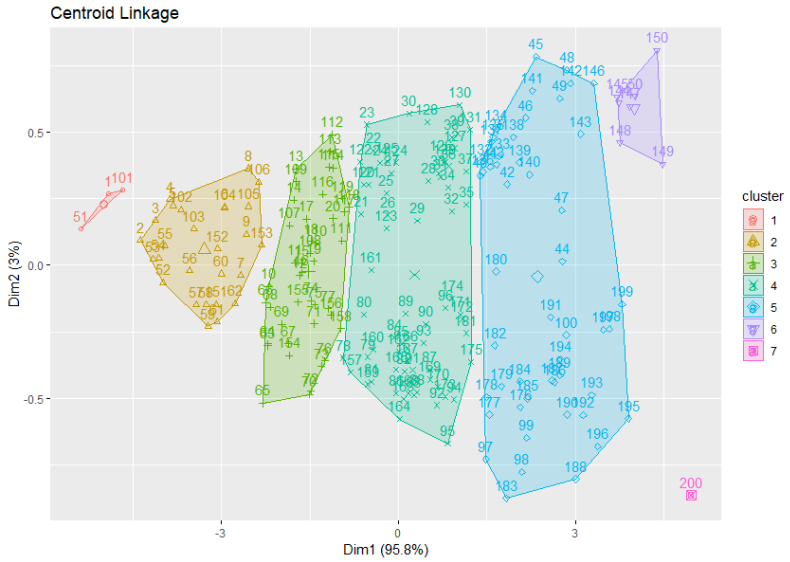


Fig.5-c Assigned Cluster Labels (centroid linkage, k=7)

From the above plots we can find out that number 4 is still a better solution. When k=3 (see Fig.5-b), the green cluster is too big and the points within this cluster are not closely connected. When k=7(see Fig.5-c), the size of each cluster is good, but the number of clusters is a little bit more and there are three clusters only containing one or just several points. When k=4, we have three main clusters (not too many, not too few) and they definitely have different values along the Dimension 1, so the number 4 would be a good choice of cut point.

(2) Determine the optimal number of clusters based on C(g)

Code:

```
> NbClust(crabs.stdz[,4:8],method='centroid', index='ch')
$`All.index`
       2        3        4        5        6        7        8        9       10
  5.2112  79.0970 234.9100 183.8227 164.9569 277.0511 255.3027 283.1213 265.1371
      11       12       13       14       15
262.5162 238.9288 292.7189 270.1934 251.4228

$Best.nc
Number_clusters     Value_Index
        13.0000        292.7189
```

By default, the maximal number of clusters is 15. So if we run the above code, we would get the result that the best choice for the number of clusters is 13. However, we only have 200 observations and 13 clusters are too much. Based on what we have discussed in (1), it would be better if the number of clusters is no more than 7. So we change the parameters of NbClust function and run it again.

```
> NbClust(crabs.stdz[,4:8],method='centroid',max.nc=6,index='ch')
$`All.index`
       2        3        4        5        6
  5.2112  79.0970 234.9100 183.8227 164.9569

$Best.nc
Number_clusters     Value_Index
           4.00          234.91
```

When we limit the number of clusters, the best choice for cut point would be **4**, which is consistent with our previous observation of the dendrogram.

2. K-means Clustering

Code:

```
> set.seed(2011)
> NbClust(crabs.stdz[,4:8],method='kmeans',distance="euclidean",index='ch')
$`All.index`
       2        3        4        5        6        7        8        9       10
361.0123 395.6452 451.7746 462.2850 413.8731 393.7106 391.2940 380.0450 398.1781
      11       12       13       14       15
395.5734 394.6369 372.9139 360.1467 363.8643

$Best.nc
Number_clusters     Value_Index
          5.000          462.285
```

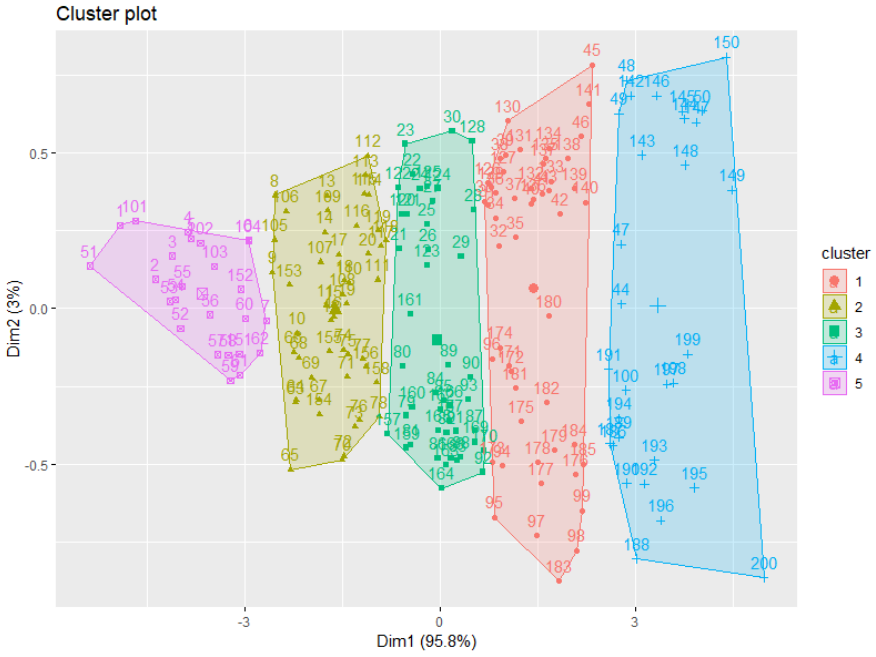


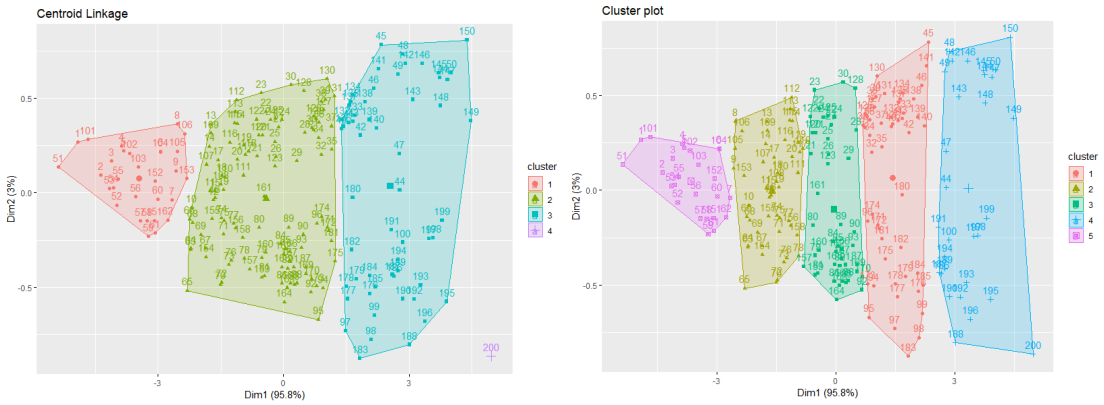
Fig.6 Assigned Cluster Labels (kmeans, g=5)

For kmeans, we get the number **5** from the NbClust function. We can see from the above figure (Fig.6) that five clusters are fairly separated and have different ranges of values along Dimension 1, which is perfect.

3. Compare results from two methods

We will compare the optimal result of centroid linkage(k=4) and kmeans (g=5).

```
> xtabs(~lbls.centroid4+km.5$cluster)
              km.5$cluster
lbls.centroid4  1  2  3  4  5
             1  0  5  0  0 25
             2 23 44 45  0  0
             3 28  0  0 29  0
             4  0  0  0  1  0
```



(a) (b)

Fig.7 Assigned Cluster Labels (a) centroid linkage, k=4 (b) kmeans, g=5

The maximal agreement between two methods was that clusters are mainly formed according to their values along the Dimension 1. However, due to the different number of clusters, the size of each cluster is very different in two methods. The entry '23', '44' and '45' in the (2, 1) (2, 3) and (2, 3) position, and the entry '28' and '29' in the (3, 1) and (3, 4) position suggests that two main clusters in **centroid linkage** split to several small parts and form new clusters instead. The red cluster in **centroid linkage** and the purple cluster in **kmeans** are the most similar: they have similar size and position, and the main difference is that the purple cluster has less points. In **centroid linkage**, Cluster 4 only has one point and is quite small, but in **kmeans**, the Cluster 4 is much larger and has 30 points. In general, the clusters in **kmeans** are more evenly separated and have similar size.

Part 4: Demographics Analysis after Clustering

In this part, we are supposed to evaluate the distribution of the known demographics (sex, species) for the k-means cluster solution.

Code:

1) The distribution of sex:

```
> count(crabs.stdz, vars="sex")
     sex freq
1 Female  100
2   Male  100
```

```
> count(crabs.stdz, c("sex", "km.5$cluster"))
      sex km.5.cluster freq
1  Female           1   21
2  Female           2   21
3  Female           3   28
4  Female           4   16
5  Female           5   14
6    Male           1   30
7    Male           2   28
8    Male           3   17
9    Male           4   14
10   Male           5   11
```

2) The distribution of species:

```
> count(crabs.stdz, "species")
  species freq
1    Blue  100
2  Orange  100
```

```
> count(crabs.stdz, c("species", "km.5$cluster"))
   species km.5.cluster freq
1     Blue           1   21
2     Blue           2   29
3     Blue           3   25
4     Blue           4    6
5     Blue           5   19
6   Orange           1   30
7   Orange           2   20
8   Orange           3   20
9   Orange           4   24
10  Orange           5    6
```

Table 2-a The Frequency Distribution of Sex within Five Clusters

Clusters	Sex		Total
	Female	Male	
1	21	30	51
2	21	28	49
3	28	17	45
4	16	14	30
5	14	11	25

Table 2-b The Frequency Distribution of Species within Five Clusters

Clusters	Species		Total
	Blue	Orange	
1	21	30	51
2	29	20	49
3	25	20	45
4	6	24	30
5	19	6	25

From the above two tables we can see that the clusters are divide in a manner consistent with demographic differences. Because in some clusters, the two kind of sexes or species are not evenly distributed. For example, in the frequency distribution of sex (see Table 2-a), females are more likely to be classified as Cluster 3 and males are more likely to be classified as Cluster 1 and 2. In the frequency distribution of species (see Table 2-b), blue crabs are more likely to

be classified as Cluster 2 and 5, and orange crabs are more likely to be classified as Cluster 1 and 4.