STAT430_HW8

Enguang Fan

4/6/2020

1

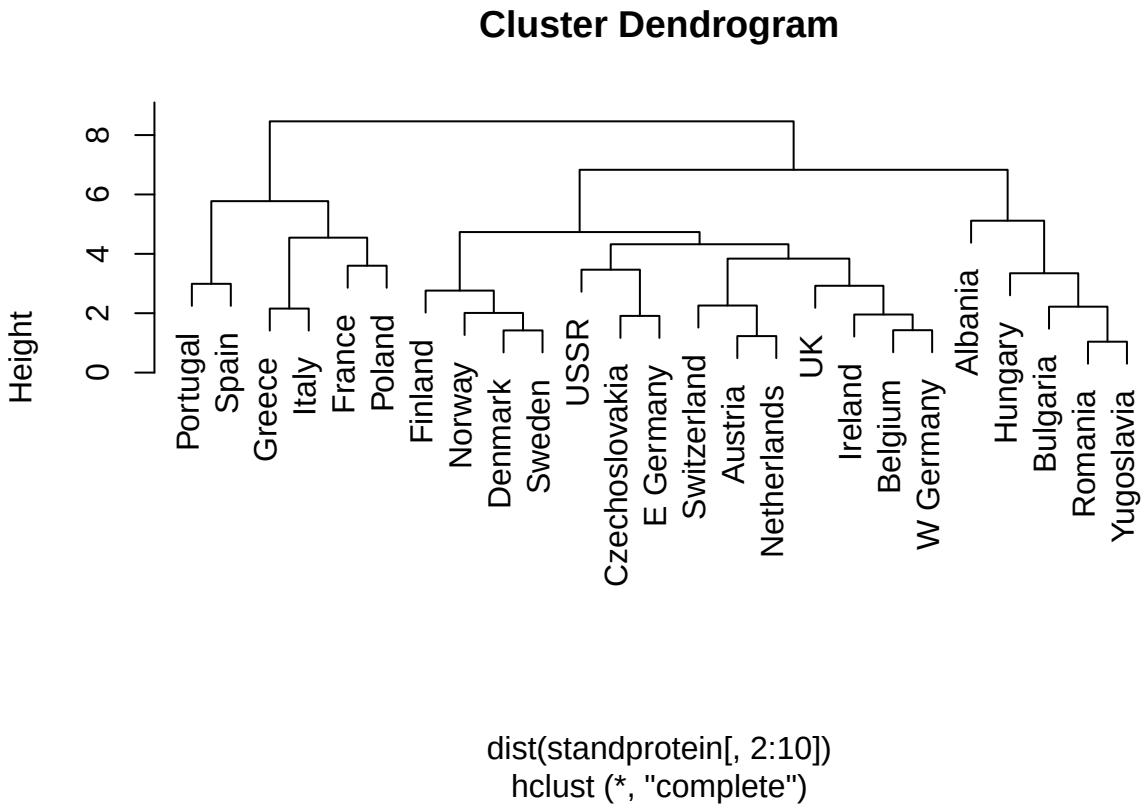
```
url='http://www.biz.uiowa.edu/faculty/jledolter/DataMining/protein.csv'
protein=read.csv(url)
head(protein)
```

```
##          Country RedMeat WhiteMeat Eggs Milk Fish Cereals Starch Nuts Fr.Veg
## 1        Albania    10.1       1.4  0.5  8.9  0.2    42.3    0.6  5.5    1.7
## 2        Austria     8.9      14.0  4.3 19.9  2.1    28.0    3.6  1.3    4.3
## 3        Belgium    13.5       9.3  4.1 17.5  4.5    26.6    5.7  2.1    4.0
## 4       Bulgaria     7.8       6.0  1.6  8.3  1.2    56.7    1.1  3.7    4.2
## 5 Czechoslovakia     9.7      11.4  2.8 12.5  2.0    34.3    5.0  1.1    4.0
## 6        Denmark    10.6      10.8  3.7 25.0  9.9    21.9    4.8  0.7    2.4
```

```
standprotein = protein
#let us standardized the data
for (i in 2:9) {
  standprotein[,i] = (protein[,i] - mean(protein[,i]))/sd(protein[,i])
}
head(standprotein)
```

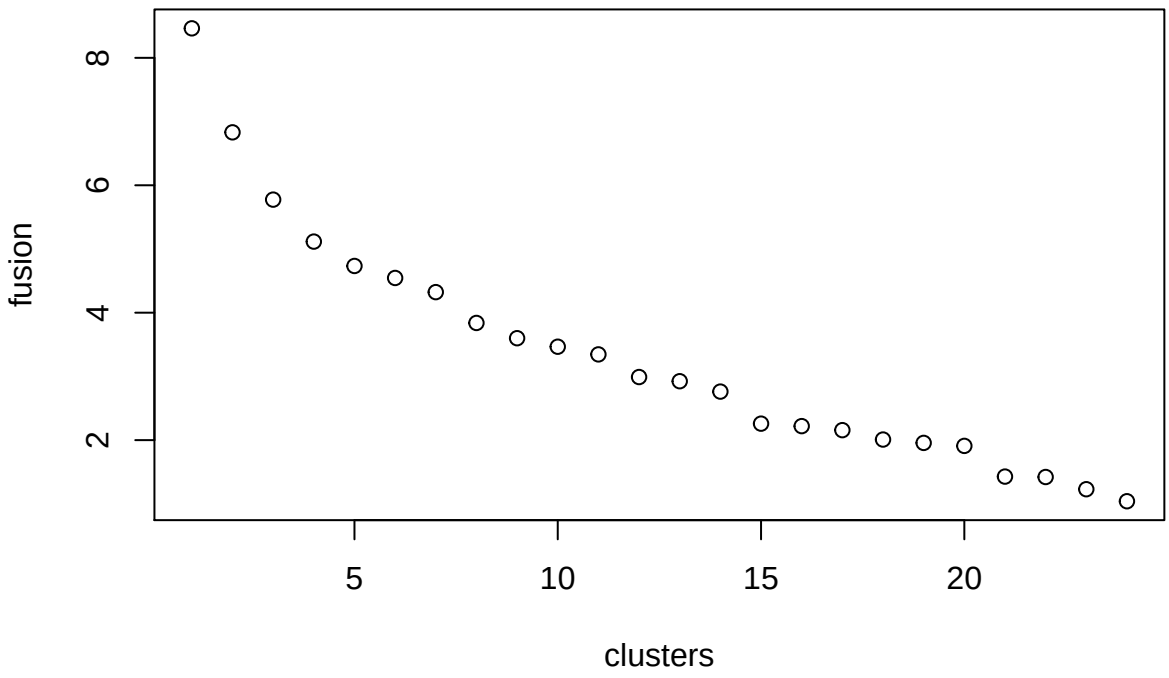
```
##          Country     RedMeat  WhiteMeat       Eggs        Milk        Fish
## 1        Albania  0.08126490 -1.7584889 -2.1796385 -1.15573814 -1.20028213
## 2        Austria -0.27725673  1.6523731  1.2204544  0.39237676 -0.64187467
## 3        Belgium  1.09707621  0.3800675  1.0415022  0.05460623  0.06348211
## 4       Bulgaria -0.60590157 -0.5132535 -1.1954011 -1.24018077 -0.90638347
## 5 Czechoslovakia -0.03824231  0.9485445 -0.1216875 -0.64908235 -0.67126454
## 6        Denmark  0.23064892  0.7861225  0.6835976  1.11013912  1.65053488
##      Cereals     Starch       Nuts Fr.Veg
## 1  0.9159176 -2.2495772  1.2227536    1.7
## 2 -0.3870690 -0.4136872 -0.8923886    4.3
## 3 -0.5146342  0.8714358 -0.4895043    4.0
## 4  2.2280161 -1.9435955  0.3162641    4.2
## 5  0.1869740  0.4430614 -0.9931096    4.0
## 6 -0.9428885  0.3206688 -1.1945517    2.4
```

```
countries=protein[,1]
clust=hclust(dist(standprotein[,2:10]),method="complete")
plot(clust,labels=countries)
```



##2

```
## look at fusion coefficients
fusion=clust$height
n=nrow(protein)
clusters=(n-1):1
plot(clusters,fusion)
```



I will choose cluster numbers = 12 Justification: Because of the standardized of the data, the curve is flattened compared to the one we see in lecture, so In this plot, there is no obvious elbow, I will choose cluster numbers = 12

3

```
twelveclust=cutree(clust,k=12)
## Check countries in cluster 3, for example
twelveclust
```

```
##  [1]  1  2  3  4  5  6  5  6  7  8  9  3  8  2  6 10 11  4 11  6  2  3 12  3  4
```

```
countries[twelveclust==1]
```

```
## [1] Albania
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
twelveclust[10]
```

```
## [1] 8
```

```r
#record how many countries each clusters
size = rep(0,12)

result = data.frame(matrix(0, nrow = 12, ncol = 9))
#df delete the country
df = standprotein[,2:10]
```

```r
for (i in 1:25) {
  for (c in 1:12) {
    if (twelveclust[i] == c) {
      result[c,] = df[i,] + result[c,]
      size[c] = size[c] +1
    }
  }
}
```

```r
#calculate the mean value
for (i in 1:12) {
  result[i,] = result[i,] / size[i]
}
names(standprotein)
```

```
##  [1] "Country"   "RedMeat"   "WhiteMeat" "Eggs"      "Milk"      "Fish"
##  [7] "Cereals"   "Starch"    "Nuts"      "Fr.Veg"
```

```r
finalresult = setNames(result, c( "RedMeat",   "WhiteMeat", "Eggs"   ,   "Milk"     , "Fish"    ,  "Cer
finalresult
```

```
##         RedMeat  WhiteMeat        Eggs        Milk       Fish    Cereals
## 1   0.081264904 -1.7584889 -2.17963852 -1.15573814 -1.2002821  0.9159176
## 2   0.200772117  1.2643650  0.65377225  0.73952979 -0.5830949 -0.6300502
## 3   1.261398625  0.4003702  1.30993054  0.50144849 -0.2010267 -0.8084895
## 4  -1.103848283 -0.5764176 -1.34452798 -1.05253048 -0.9847564  1.9880721
## 5  -0.232441528  0.9756148  0.28095505 -0.74759876 -0.1716368 -0.2549480
## 6   0.006572897 -0.2290150  0.19147892  1.34587480  1.1582546 -0.8722721
## 7   2.441532345  0.5424895  0.32569311  0.33608167  0.4161605 -0.3779572
## 8  -0.068119111 -1.0411250 -0.07694947 -0.20575854  0.1075669  0.6380079
## 9  -1.352821642  1.2192478 -0.03221141 -1.04314796 -1.1708923  0.7154581
## 10 -0.874792793  0.6237005 -0.21116367  0.30793413 -0.3773659  0.3509863
## 11 -0.949484801 -1.1764767 -0.74802044 -1.45832423  1.8562639 -0.3779572
## 12 -0.157749520 -0.8922382 -0.74802044 -0.07205771 -0.3773659  1.0343709
##        Starch       Nuts Fr.Veg
## 1  -2.2495772  1.2227536  1.700
## 2  -0.4544848 -0.6237991  4.300
## 3   0.7184449 -0.4643241  3.500
## 4  -1.1480432  0.9205905  3.400
## 5   0.9020339 -1.0686504  3.800
## 6   0.1676780 -0.9553392  2.125
## 7   0.3206688 -0.3384228  6.500
## 8  -1.3010340  1.4997366  6.600
## 9  -0.1689019  1.1723931  4.200
## 10  0.9938284 -0.5398649  6.600
```

```
## 11  0.9326321  1.1220326  7.550
## 12  1.2998101  0.1651825  2.900
```

```
for (i in 1:12) {
  print(countries[twelveclust==i])
}
```

```
## [1] Albania
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Austria     Netherlands Switzerland
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Belgium   Ireland   UK        W Germany
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Bulgaria   Romania    Yugoslavia
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Czechoslovakia E Germany
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Denmark Finland Norway  Sweden
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] France
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Greece Italy
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Hungary
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Poland
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Portugal Spain
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] USSR
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

Cluster 1: Albania very low whitemeat, eggs, milk, fish, starch. This cluster's country has relatively low protein and meat in the diet, possibly due to their awful economic situation

Cluster 2: decent amount of white meat eggs and milk but low fish, diet has many protein input but low seafood possibly due to its geolocation

Cluster 3: a lot of red meat, decent number of whitemeat eggs and milk, low in fish and cereals, the life quality is pretty good

Cluster 4: low in red and white meat, also eggs, milk and fish. Whereas abundant cereals, this clusters has poor Animal husbandry but strong crop farming

Cluster 5: low in redmeat, decent amout in whitemeat, eggs starch, life quality and economic situation should be good so people is stopping eating redmeat for their health

Cluster 6:high in milk and fish, possibly near the ocean so the seafood is abundant, also has good Dairy industry

Cluster 7: decent amount of red and white meat, eggs fish and milk, should be a developed country and diet has a lot protein

Cluster 8: low in meat, egg and milk but high in fish, possibly near coast but has poor Animal husbandry and crop industry.

Cluster 9: :high in milk and fish, possibly near the ocean so the seafood is abundant, also has good Dairy industry

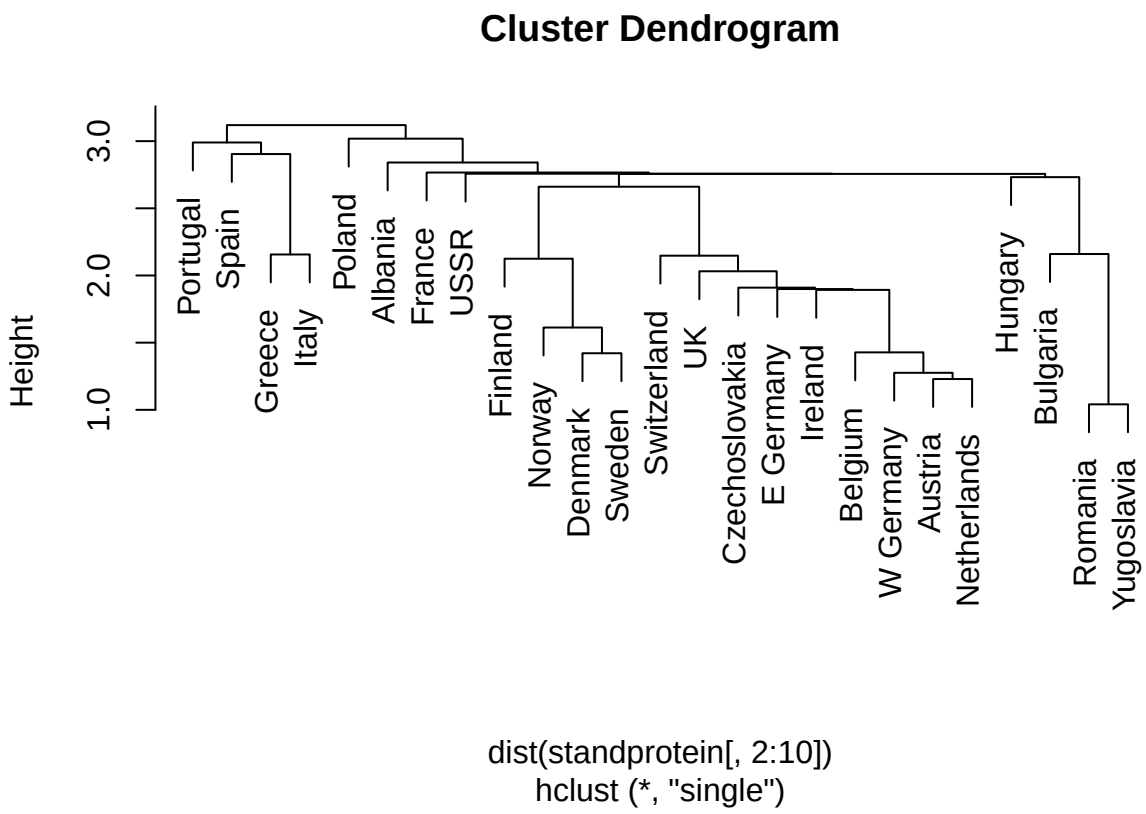
Cluster 10: low in meat, egg and milk but high in fish, possibly near coast so has some fish but has poor Animal husbandry and crop industry.

Cluster 11: low in red meat, egg and milk but high in fish, possibly near coast but has poor Animal husbandry and crop industry.

Cluster 12: low in red meat, egg and milk but high in fish, possibly near coast but has poor Animal husbandry and crop industry.

##4

```
#countries=protein[,1]
singleclust=hclust(dist(standprotein[,2:10]),method="single")
plot(singleclust,labels=countries)
```




```
sing=cutree(singleclust,k=12)
## Check countries in cluster 3, for example
for (i in 1:12) {
  print(countries[sing==i])
}
```

```
## [1] Albania
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Austria        Belgium        Czechoslovakia E Germany      Ireland
## [6] Netherlands    Switzerland    UK             W Germany
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
## [1] Bulgaria
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Denmark Finland Norway  Sweden
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] France
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Greece Italy
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Hungary
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Poland
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Portugal
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Romania    Yugoslavia
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] Spain
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
## [1] USSR
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

Comparison: The differenc is not drastic, since there are 12 clusters which is quite large if considered the size is 25, so the number of countries in each cluster will tend to be quite small, and therefore, the distance between clusters by using single and complete linkage will not be so different.

5

```
set.seed(5)
cityclusters=kmeans(standprotein[,2:10],centers=12)



countries[cityclusters$cluster==1]
```

```
## [1] Austria     Netherlands Switzerland W Germany
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
countries[cityclusters$cluster==2]
```

```
## [1] Albania    Bulgaria   Romania    Yugoslavia
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
countries[cityclusters$cluster==3]
```

```
## [1] Finland
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
countries[cityclusters$cluster==4]
```

```
## [1] Czechoslovakia E Germany
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
countries[cityclusters$cluster==5]
```

```
## [1] France
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
countries[cityclusters$cluster==6]
```

```
## [1] USSR
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
countries[cityclusters$cluster==7]
```

```
## [1] Portugal Spain
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
countries[cityclusters$cluster==8]
```

```
## [1] Belgium Ireland UK
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
countries[cityclusters$cluster==9]
```

```
## [1] Denmark Norway  Sweden
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
countries[cityclusters$cluster==10]
```

```
## [1] Greece Italy
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
countries[cityclusters$cluster==11]
```

```
## [1] Hungary
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
countries[cityclusters$cluster==12]
```

```
## [1] Poland
## 25 Levels: Albania Austria Belgium Bulgaria Czechoslovakia ... Yugoslavia
```

```
#countries[cityclusters$cluster==13]
```

Similarities: The K-means and complete linkage analysis both generate clusters in which the number is quite average, the number of countries range from 1 to 4, and the cluster is actually quite similar, like France and USSR are longly cluster in both methods. And Greece with Italy cluster also didn't change.

Differences: They generate quite different clusters, e.g. Albania is a solely cluster by complete linkage analysis but in k-means it's in cluster with Bulgaria Romania Yugoslavia. In complete linkage analysis, Finland is with Swedena and Danmark yet in K-means it's lonly cluster.