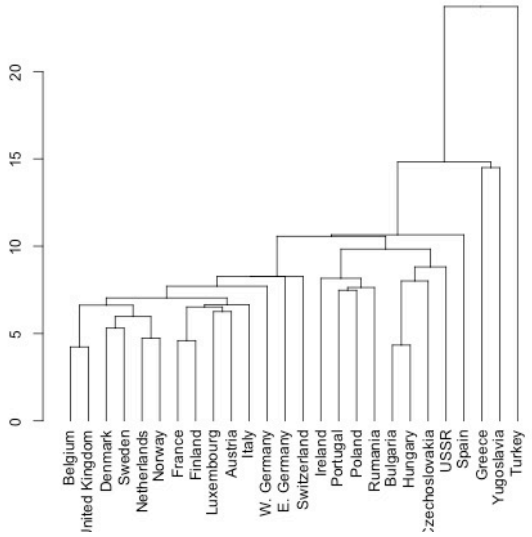
Homework 4 report

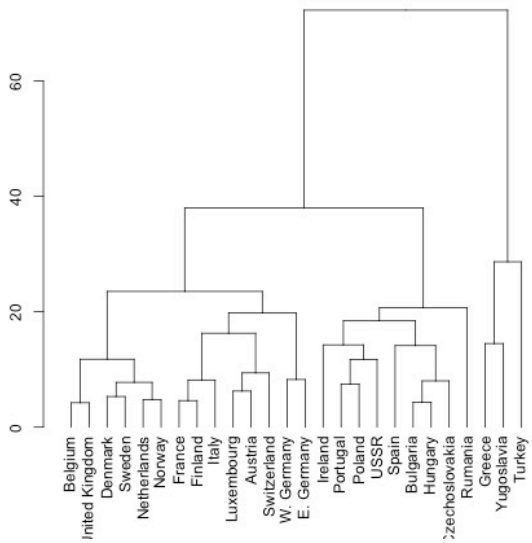
Problem 1

1. dendrogram of this data for each of single link, complete link, and group average clustering. In all types of clustering, Countries are clustered in 3 main groups, most of the capitalist Western Europe are clustered together, and countries of the communist East Bloc are tend to be grouped together. Yugoslavia, Turkey and Greece are together probably because of their location on map. One interesting difference I noticed in those three different cluster methods is that the single link method tends to produce extended clusters where as complete link and group average produced rounded clusters.

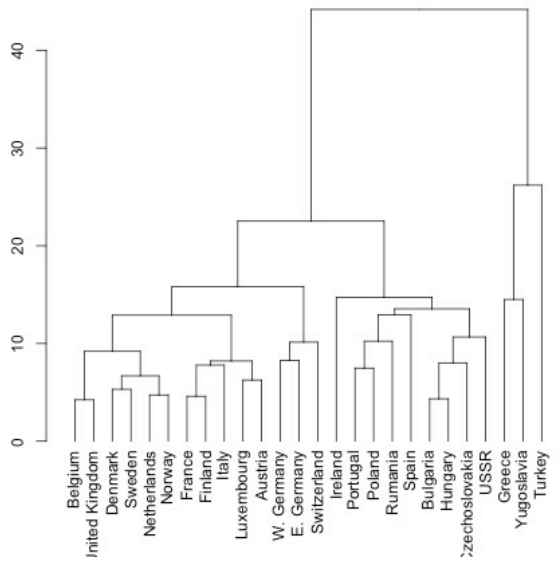
a. single link



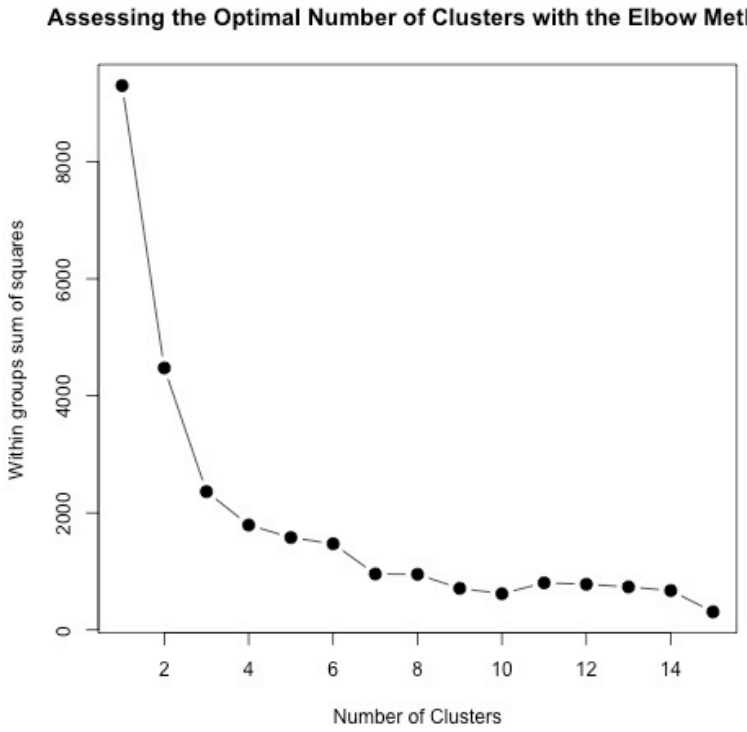
b. complete link



c. group average



2. Using k-means, cluster this dataset. What is a good choice of k for this data and why?
I used Elbow method to identify optimal K for this data. the Elbow method examines the within-cluster dissimilarity as a function of the number of clusters. **As the plot shows, the "knee" appears to be at 3 or 4, which is a good choice of k** and that is consistent with the observation in dendrogram.



When K=3, the countries are grouped as:
[Belgium, Denmark, France, W. Germany, Italy, Luxembourg, Netherlands, United Kingdom, Austria, Finland, Norway, Sweden, Switzerland, E. Germany],
[Ireland, Portugal, Spain, Bulgaria Czechoslovakia, Hungary, Poland, Rumania, USSR],
[Greece, Turkey, Yugoslavia]
When K=4, the countries are grouped as:
[Belgium, Denmark, France, Netherlands, United Kingdom, Finland, Norway, Sweden],
[W. Germany, Italy, Luxembourg, Austria, Switzerland, Czechoslovakia, E. Germany],
[Ireland, Portugal, Spain, Bulgaria, Hungary, Poland, Rumania, USSR],
[Greece, Turkey, Yugoslavia],

Problem 2

1. Build a classifier that classifies sequences into one of the 14 activities provided.
In my first attempt, I used 80% of data files as training data, I used chunk size of 32 to segment the data, in vector quantization, I used k means clustering with a k=480. Then I trained my data with a random forest algorithm(ntree=200).
The total error rate in test data is 0.3063584.
Confusion matrix showing below:

```
pred            Brush_teeth Climb_stairs Comb_hair Descend_stairs Drink_glass Eat_meat
  Brush_teeth             2            0         0              0           0        0
  Climb_stairs            0           18         0              3           0        0
  Comb_hair               0            0         5              0           0        0
  Descend_stairs          0            1         0              5           0        0
  Drink_glass             0            0         2              0          19        1
  Eat_meat                0            0         0              0           0        0
  Eat_soup                0            0         0              0           0        0
  Getup_bed               0            0         0              0           0        0
  Liedown_bed             0            0         0              0           0        0
  Pour_water              1            0         0              0           1        0
  Sitdown_chair           0            2         0              1           0        0
  Standup_chair           0            0         0              0           0        0
  Use_telephone           0            0         0              0           0        0
  Walk                    0            0         0              0           0        0

pred            Eat_soup Getup_bed Liedown_bed Pour_water Sitdown_chair Standup_chair
  Brush_teeth          0         0           0          0             0             0
  Climb_stairs         0         0           1          0             0             0
  Comb_hair            0         0           0          0             0             0
  Descend_stairs       0         0           0          0             0             0
  Drink_glass          0         1           0          0             0             0
  Eat_meat             0         0           0          0             0             0
  Eat_soup             0         0           0          0             0             0
  Getup_bed            0        12           1          0             0             0
  Liedown_bed          0         0           0          0             0             1
  Pour_water           1         1           0         19             0             0
  Sitdown_chair        0         2           2          1            14             9
  Standup_chair        0         5           2          0             6            11
  Use_telephone        0         0           0          0             0             0
  Walk                 0         0           0          0             0             0

pred            Use_telephone Walk
  Brush_teeth               0    0
  Climb_stairs              0    5
  Comb_hair                 0    0
  Descend_stairs            0    0
  Drink_glass               2    0
  Eat_meat                  0    0
  Eat_soup                  0    0
  Getup_bed                 0    0
  Liedown_bed               0    0
  Pour_water                0    0
  Sitdown_chair             0    1
  Standup_chair             0    0
  Use_telephone             1    0
  Walk                      0   14
```

2. Improving classifier

I tested a few combinations of number of cluster centers in k-means and size of the fixed length samples, error rates showing below. The results indicates that this vector quantization method produces relatively reliable and reproducible predictive power, **the best I can get through my experiment is when chunk size=16, k=400, or when chunk size=32, k=100, with an error rate of 0.2138728.**

chunk size	centroids	error rate
16	20	0.2369942
16	50	0.2369942
16	100	0.2312139

16	200	0.2254335
16	**400**	**0.2138728**
16	500	0.2312139
32	20	0.2485549
32	50	0.2369942
32	**100**	**0.2138728**
32	200	0.2254335
32	400	0.2485549
32	500	0.2601156
64	20	0.3121387
64	50	0.2601156
64	100	0.2601156
64	200	0.3121387
64	400	0.2947977
64	500	0.3294798

Reference

1. https://onlinecourses.science.psu.edu/stat857/node/136
example code of agglomerative clustering in problem1
2. https://rpubs.com/FelipeRego/K-Means-Clustering
example code of "Assessing the Optimal Number of Clusters with the Elbow Method"
2. https://piazza.com/class/jchzguhsowz6n9?cid=746
piazza discussion about read files in problem 2
3. https://stackoverflow.com/questions/7060272/split-up-a-dataframe-by-number-of-rows
example code for splitting dataframe by chunk size
4. https://stat.ethz.ch/pipermail/r-help/2005-March/068063.html
flatten a matrix