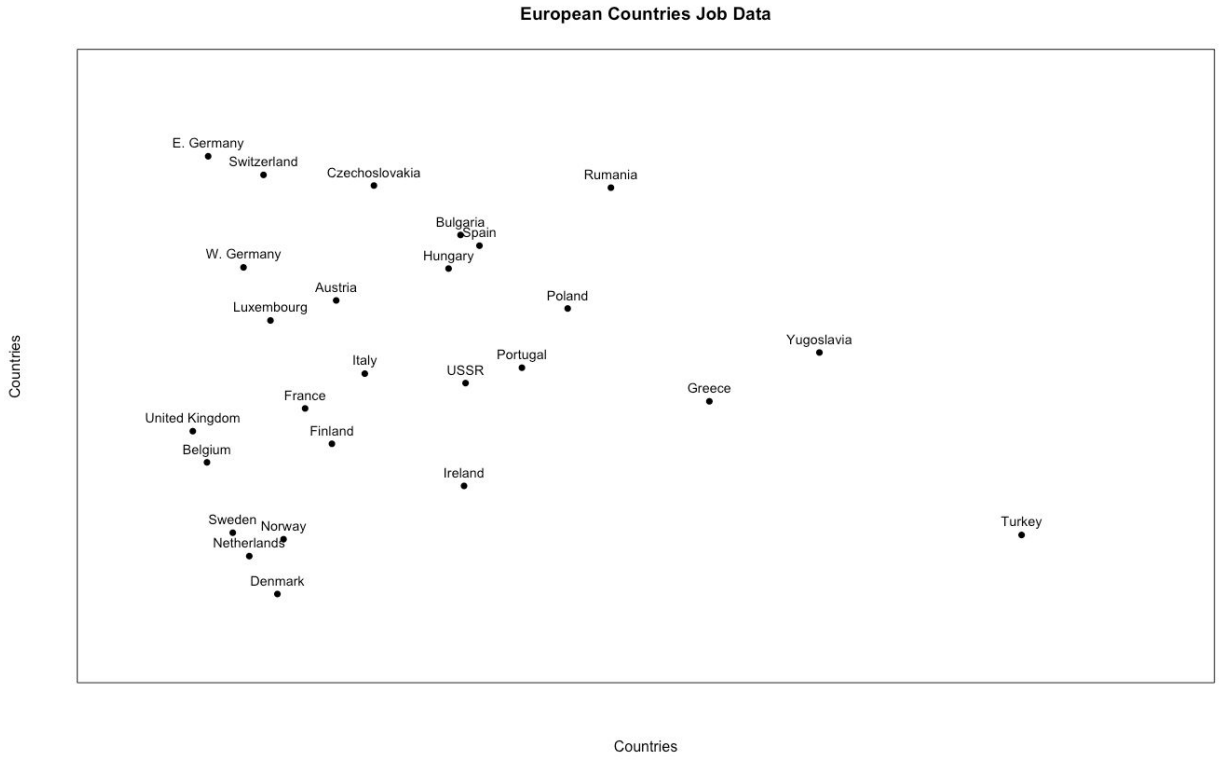
Homework 4

Problem 1

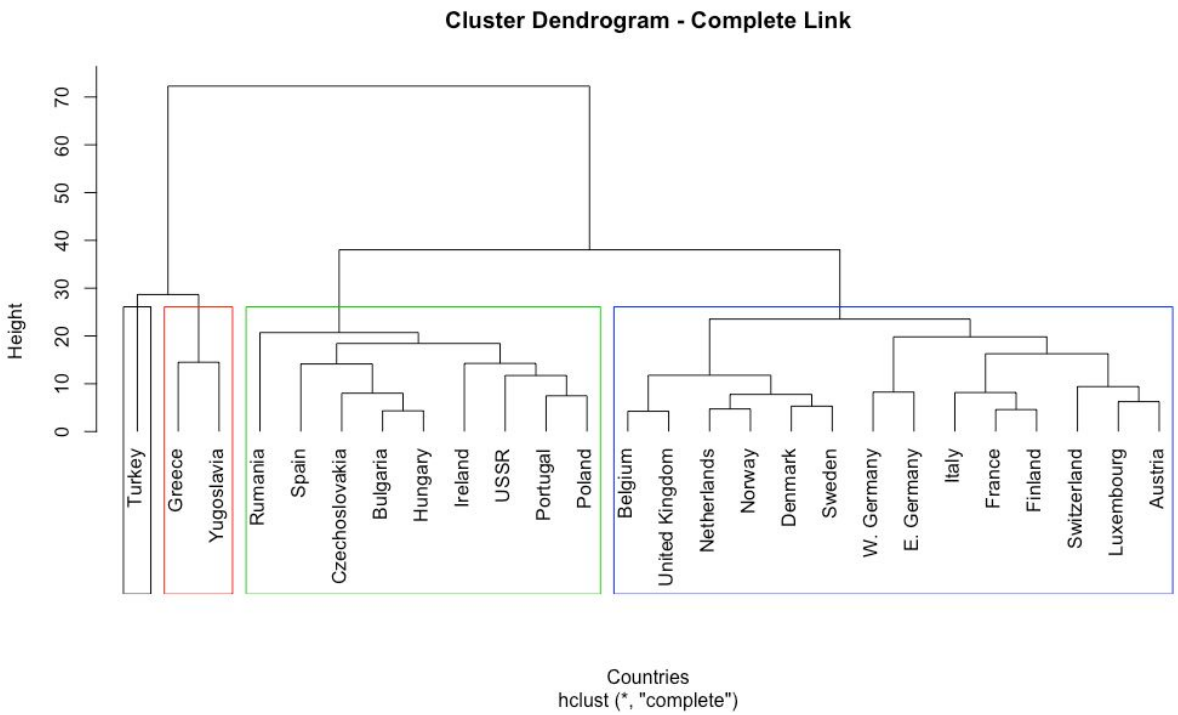
Source Code : HW4-1.R

Part 1

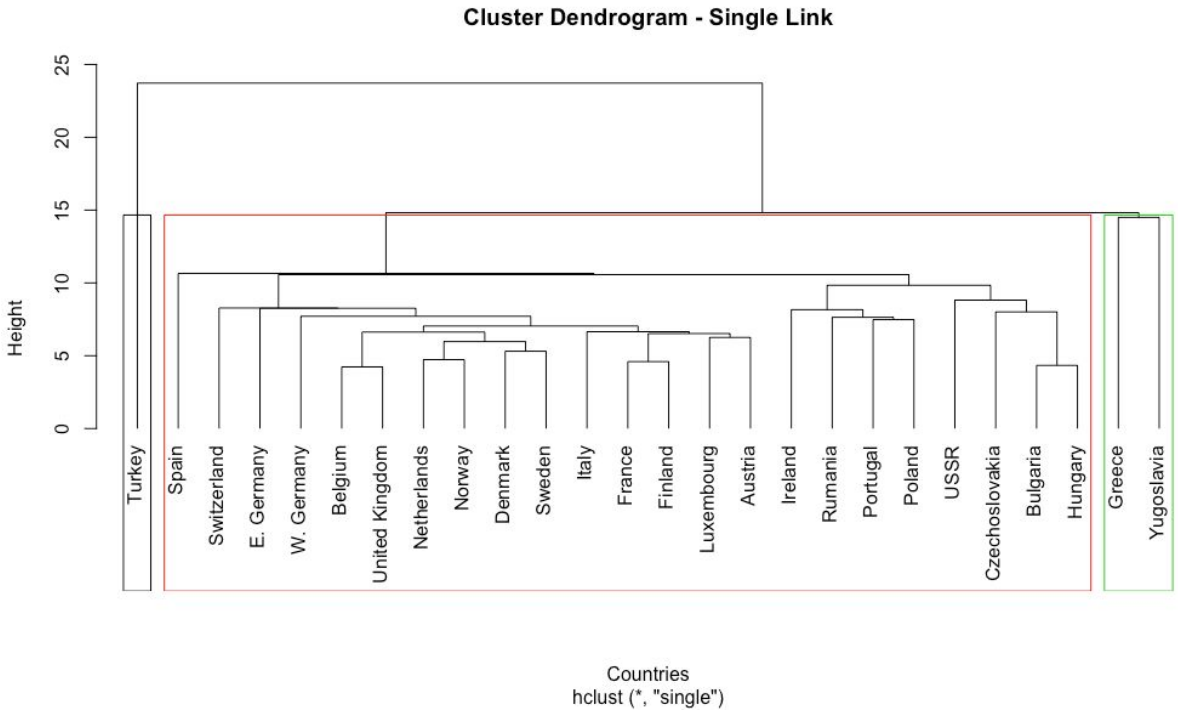
As part of preprocessing, I mapped the job data of countries in a two dimensional space as show in the chart below. Turkey is an outlier data point. Yugoslavia and Greece are closer but are quite far from Turkey on the right and rest of the countries on the left.



This section contains Dendrograms for Complete, Single and Group average link.

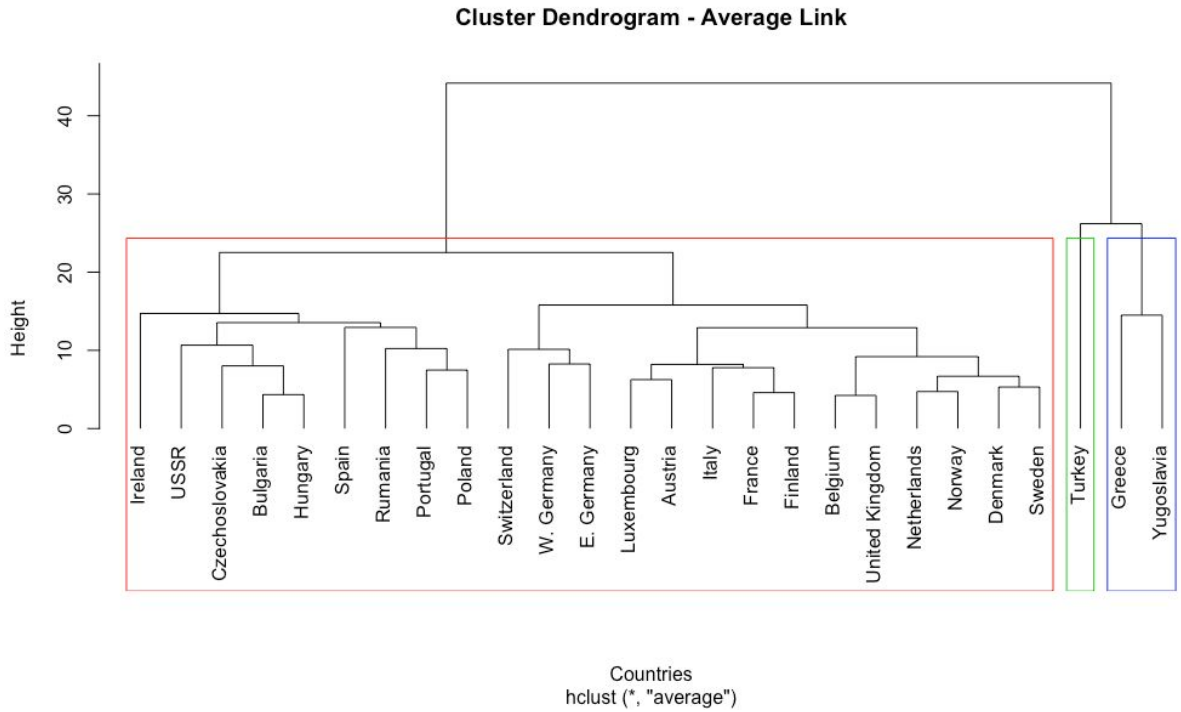


In the complete link we merge in each step the clusters whose merger has the smallest diameter. Since we are measuring smallest diameter, it can be sensitive to outliers.



Ashish Kumar (ashishk2@illinois.edu)

In the single link we merge in each step the clusters whose two closest members has smallest distance. Single link clustering has tendency to produce imbalance clusters which are not very compact. We see the same behavior in the chart above.



The average link clustering is a compromise between single and complete link clustering.

Key differences and Observations

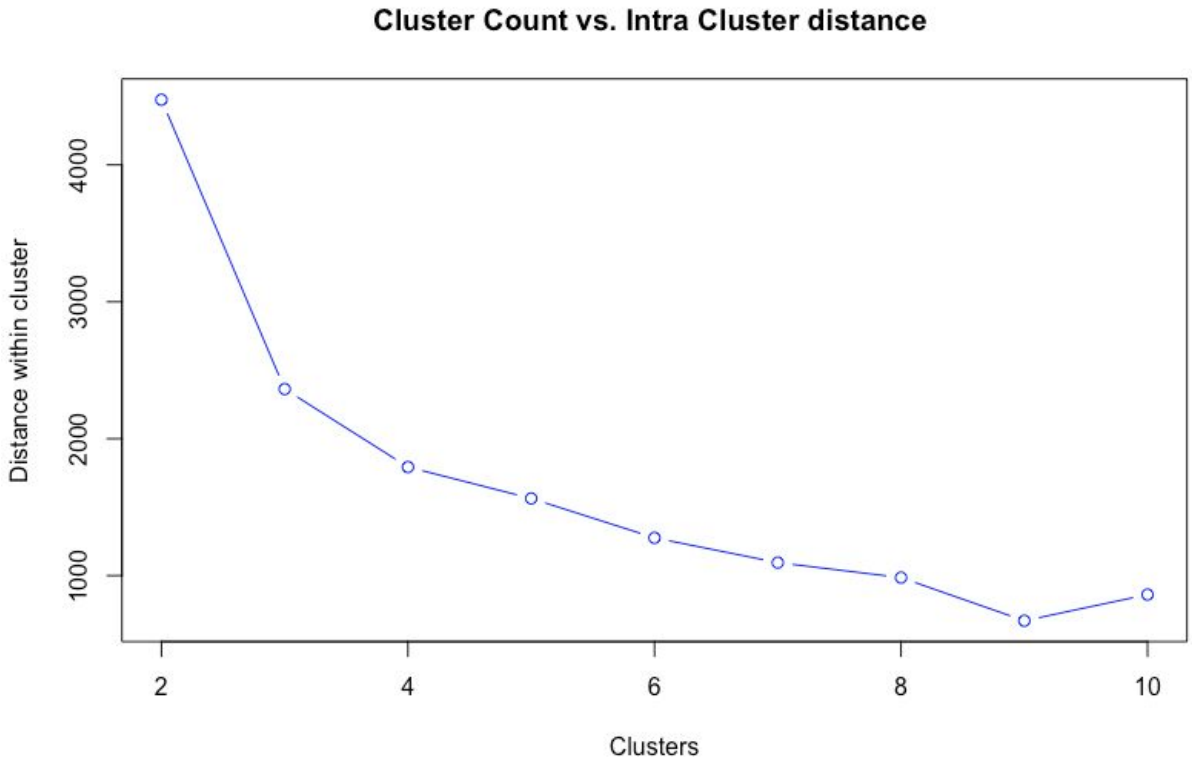
Turkey, Greece and Yugoslavia have 40% + people employed in the agriculture sector. Turkey had largest employment of 66% in the agriculture section. It explains the observations of Turkey being in its own cluster. Greece & Yugoslavia are in a separate cluster across all three type of hierarchical clustering.

Countries such as UK, Netherlands, Denmark, Italy, France has extremely low population working the agriculture sector but a high number of people are employed in the manufacturing sector. It explains why they are in same cluster.

Countries such as Spain, Rumania, Hungary had significant population working in both agriculture and manufacturing sector, which explains why these countries are clustered together.

Part 2

The **knee** of the line is around cluster count of 4. A value between 3 to 5 should be a good value for K-Means clustering for the Jobs dataset. In all three dendrogram charts, the data is divided in 3 or 4 clusters at medium height.



Problem 2

Source code HW4-2.R

My program run time varies from 10 to 15 minutes on a dual core, 8 GB macbook.

Part A Baseline Setup

BaseLine
SEGSIZE = 32
Cluster = 40*12 = 480
SLIDER=12
Error Rate : 31%
Accuracy :73%

Confusion Matrix

	Y_test													
predicted	1	2	3	4	5	6	7	8	9	10	11	12	13	14

1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	20	0	4	0	0	0	0	0	0	0	0	0	0	4
3	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	19	0	0	1	0	0	0	0	3	0	
6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	11	3	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	1	0	1	1	0	19	2	3	0	0	
11	1	0	0	2	0	0	0	0	0	0	13	0	0	1	
12	0	0	0	0	0	0	0	7	3	1	5	17	0	1	
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	0	0	0	1	0	0	0	0	0	0	0	0	0	14	

Part B

Error rates are listed within brackets ( ).

Segment Size 32

Cluster/Slider	8	12
480	78.44%(21.56%)	73.05%(26.95%)
640	77.24%(22.76%)	76.04%(23.96%)
720	74.25%(25.75%)	71.25%(28.75%)

Slider value of 8 is better than 16. I decided to use slider value of 8 in the next round of testing and reduce the segment size to 16.

Segment Size 16, Slider = 8

Cluster = 480	Cluster = 640	Cluster = 720
78.44%(21.56%)	77.24%(22.76%)	74.85%(25.15%)

Segment Size 8, Slider = 8

I get best training accuracy of 93%(7%) but test accuracy is not as great. The best test accuracy score was 74.85%(21.56%)

Best Score 78.44% (21.56%)

Confusion Matrix

	Y_test													
predicted	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	2	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	18	0	4	0	0	0	0	0	0	0	0	0	0
3	0	0	6	0	0	0	0	0	0	0	0	0	0	0
4	0	2	0	1	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	20	0	0	0	0	0	0	0	2	0
6	0	0	0	0	0	1	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	12	5	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	1	0	0	20	1	0	1	0
11	0	0	0	2	0	0	0	1	1	0	13	0	0	1
12	0	0	0	1	0	0	0	7	0	0	6	20	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	18

Citation

I consulted with Aruna Neervannan throughout the assignment but we did not work in a group.