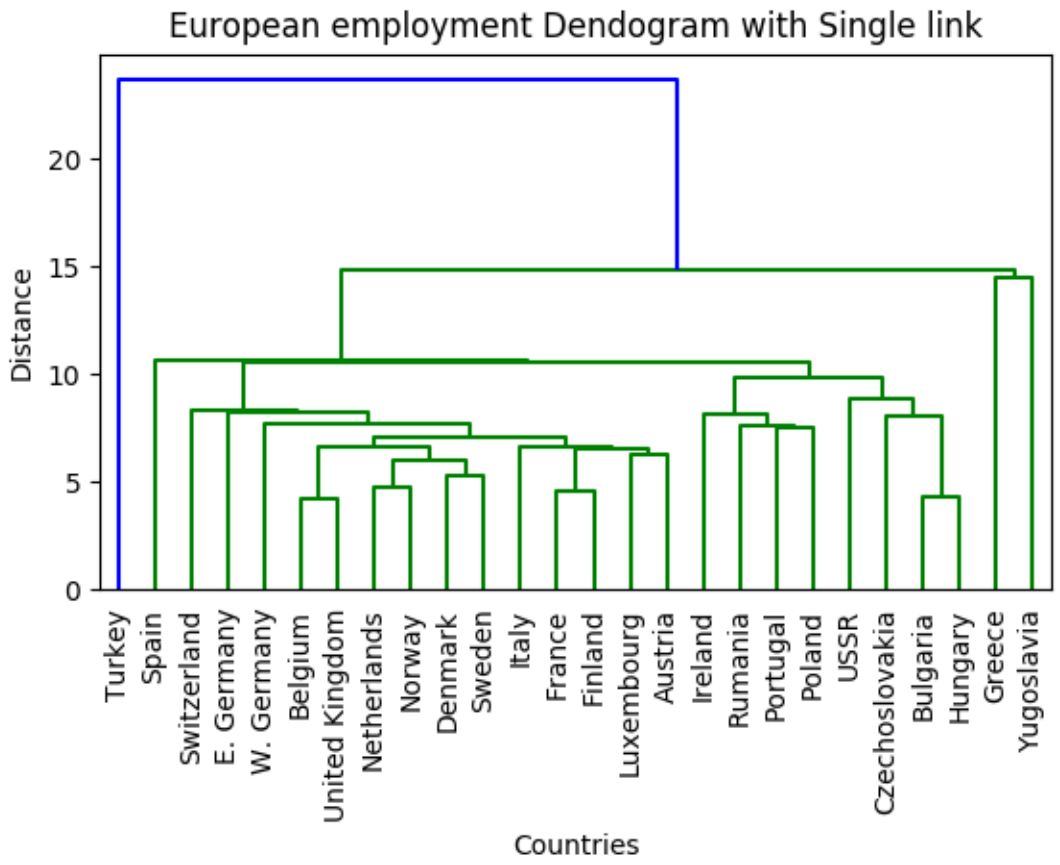
HOMEWORK4

Problem1, Part 1:

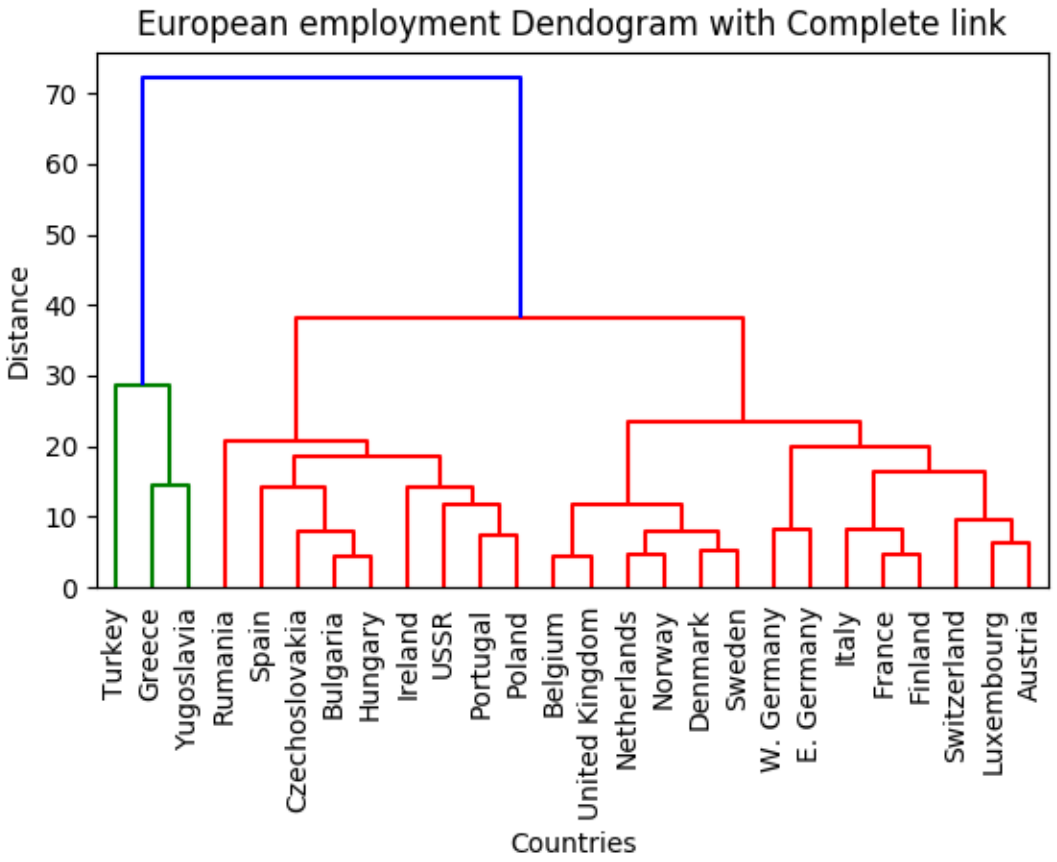
Using the European Jobs dataset, we were asked to do agglomerative clustering on the data using single link, complete link, group average.

Following are the dendograms obtained:

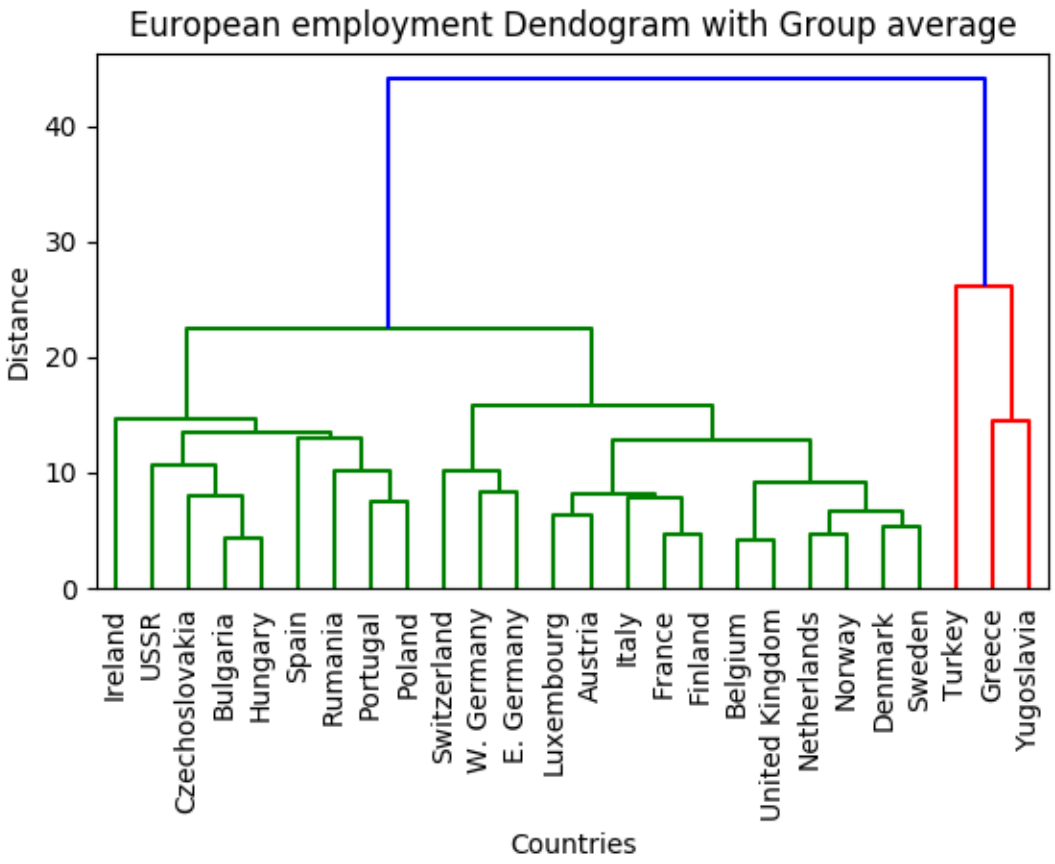
a.) Single Link: -



b.) Complete Link:



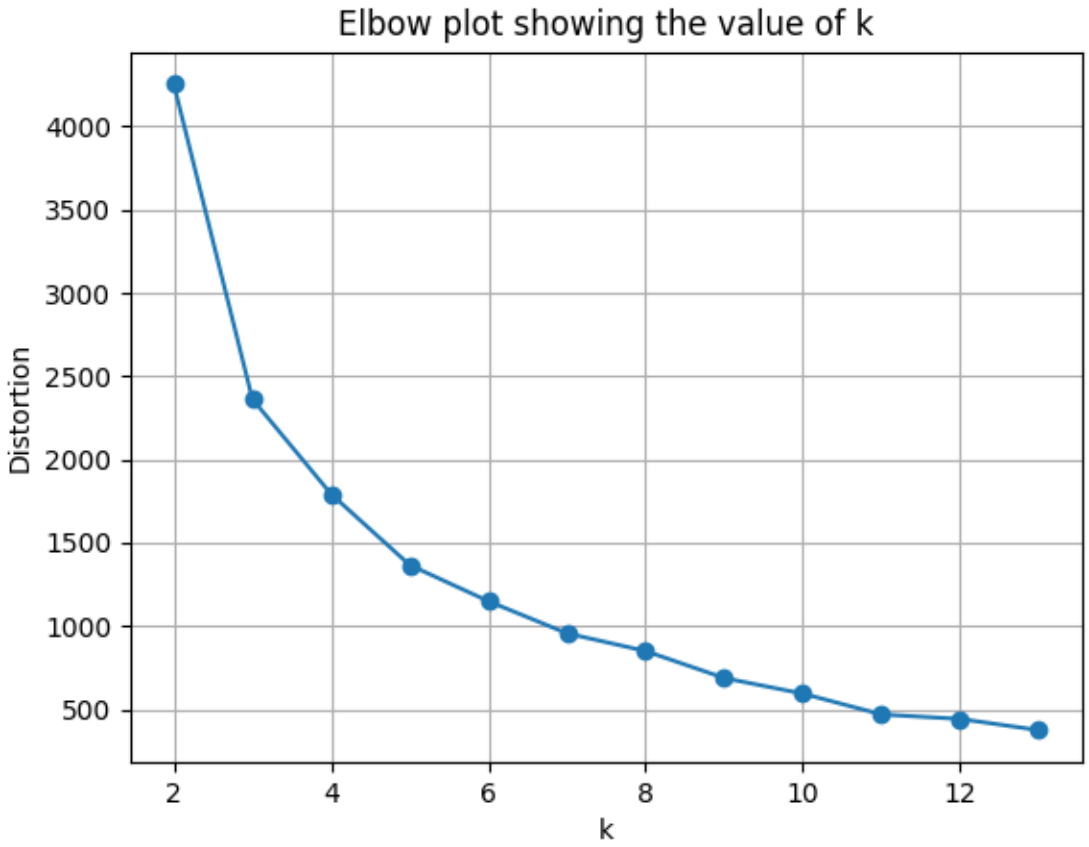
c.) Group Average: -



The above dendograns gives us a representation that shows how some of the countries are clustered together and lie close to each other, such nearness in the plots indicate some kind of common pattern that exists between such countries. To further explore a good no of cluster we shall be performing k-means in part 2.

Problem 1, Part2:

In order to check which would be an appropriate no of cluster for the above dataset we decided to do k-means clustering with different values of k and get the distortions for each value and use an elbow plot to better evaluate which would be a good choice for k. Following is the elbow plot:



Analyzing this plot shows that the gradient occurring at 4 and 5 are much more significant than going further beyond. Personally to me 4 seemed to be a better choice for k as the slope is huge and seeing the dendograms I am able to figure out **4 evident clusters**.

Problem 2, Part1:

We have been using daily life set data which measures the acceleration rate in x,y,z as a function of time across 14 different activities performed by multiple subjects. The aim was to build a classifier that takes a signal and classify it among one of the 14 categories. The classifier needs feature that has to be generated by vector quantizing the signals.

We started by first splitting signals within a category into train and test. Now we created segments for both train and test by taking 32 rows and decomposing them into a 96-unit long vector, this was done for every signal and for those signals where a segment could not be created because of less data points we discarded those data points. We performed the above procedure for all the categories and stacked the segments differently for train and test. Now we pushed the training matrix into K-means clustering with 480 clusters and thus got the 480 cluster centers generated. We now have a matrix of cluster centers with shape 480*96.

The next step was to box each of the segments into one of these clusters centers by finding which amongst the 480 cluster centers is close to a given segment. After this we created a histogram of frequency vector for both test and train which will serve as a training and testing feature for Random Forest classifier. Note that while creating the histogram vector we also labeled them with the category to which they belong which will help us in training and testing the classifier.

One this is all done we push these features to random forest classifier and then perform prediction on the held out testing set. Following were the outputs:

a.)Total error rate:

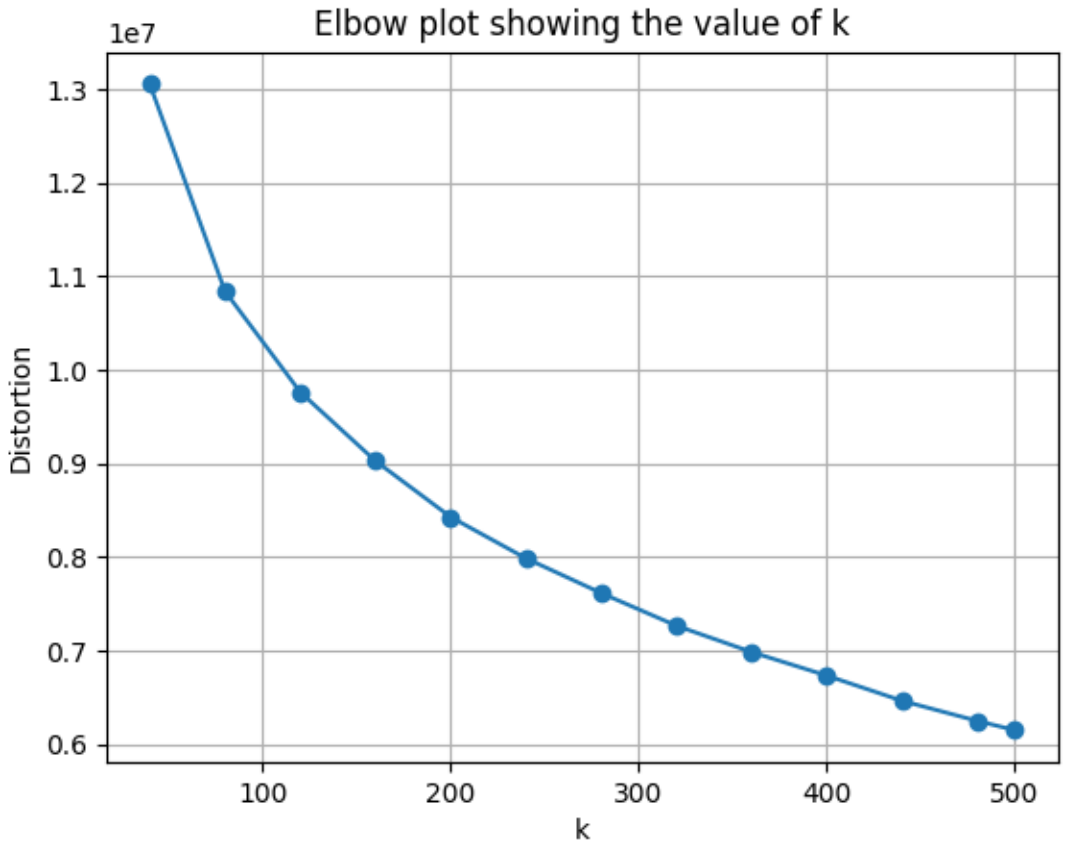
Accuracy achieved is **75.4491017964%**
Error rate for the classifier is **24.5508982036%**

b.)Confusion Matrix:

```
[[ 2  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0 18  0  0  1  0  0  0  0  0  0  0  0  1]
 [ 0  0  3  0  3  0  0  0  0  0  0  0  0  0]
 [ 0  3  0  5  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0 20  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  1  0  0  0  0]
 [ 0  0  0  0  0  0  3  0  0  0  0  0  0  0]
 [ 0  1  0  1  0  0  0 11  0  3  1  3  0  0]
 [ 0  0  0  0  1  0  0  0  0  3  0  1  0  0]
 [ 0  0  0  0  0  0  0  0  0 18  2  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  1 17  2  0  0]
 [ 0  2  0  0  0  0  0  1  0  0  4 13  0  0]
 [ 0  0  0  0  2  0  0  0  0  0  0  0  0  0]
 [ 0  3  0  0  0  0  0  1  0  0  0  0  0 16]]
```

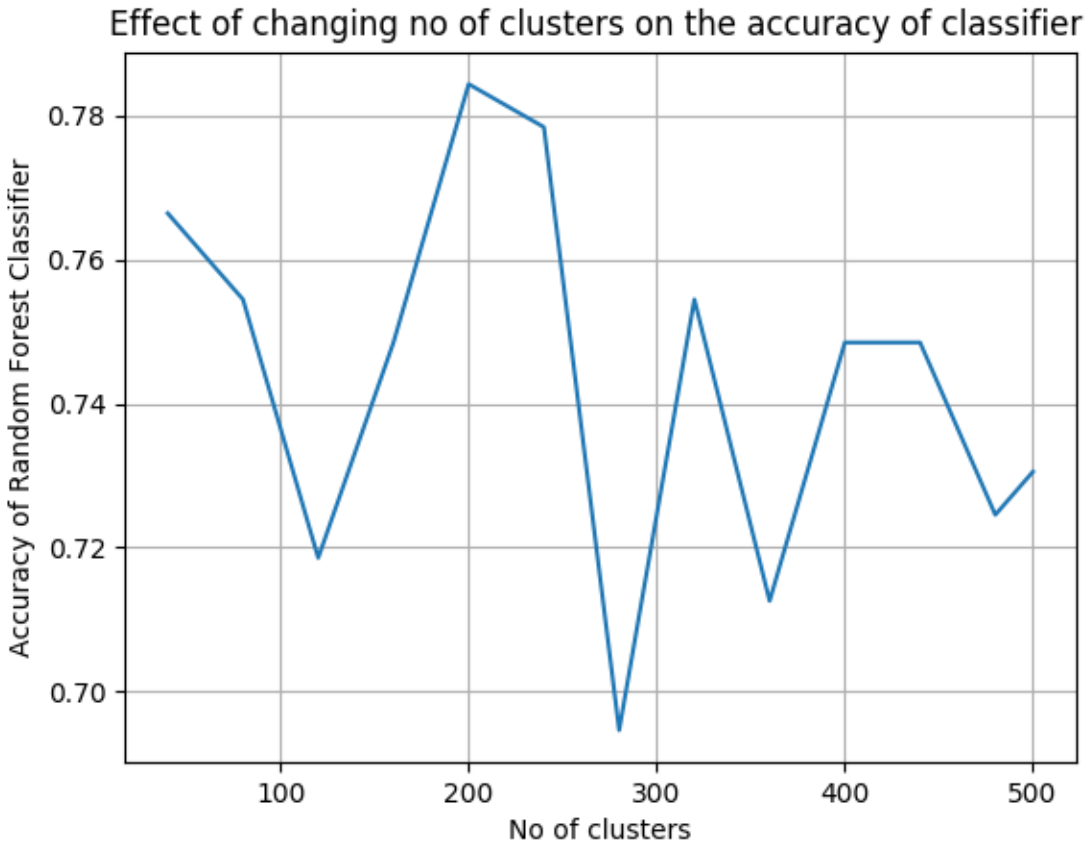
Problem2, Part2:

To further analyze how to improve the classification accuracy we decided to conduct some test. We started by plotting the elbow plot for the k-means clustering to get a sense how the distortion gets changed with the no of cluster. Below is the plot showing that:



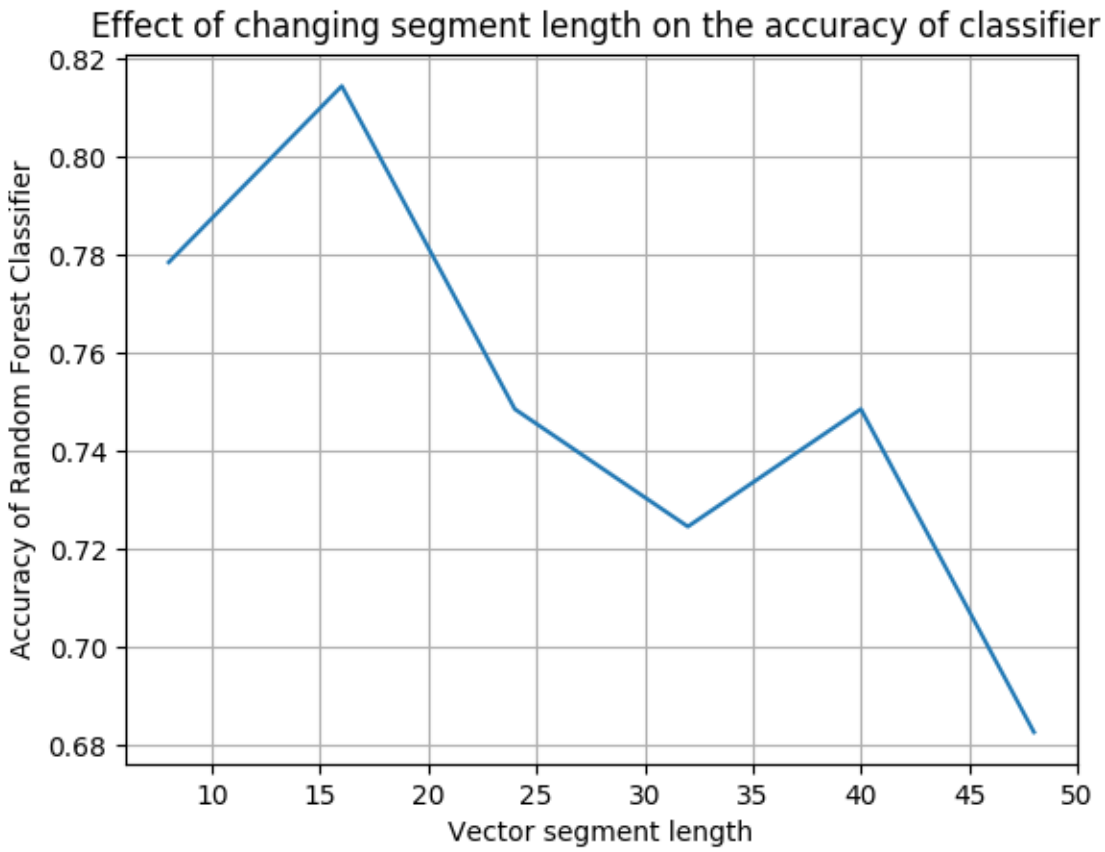
The following shows that 200 should be good no for k but we decide to further dig and check the behavior of classifier by changing the no of cluster and also by changing the segment length that we used for vector segmentation. Below are the results:

a.) Effect of no of clusters:



Here we see that for different values of k we see the accuracy varies between the 70-80% bracket and peaks up at 200 which was also an observation from the previous elbow plot.

b.) Effect of segment length:



Here we see that for segment length = 16 we get maximum accuracy and the variation for the other segment length is between 70-80% bracket as well.