Assigment #4

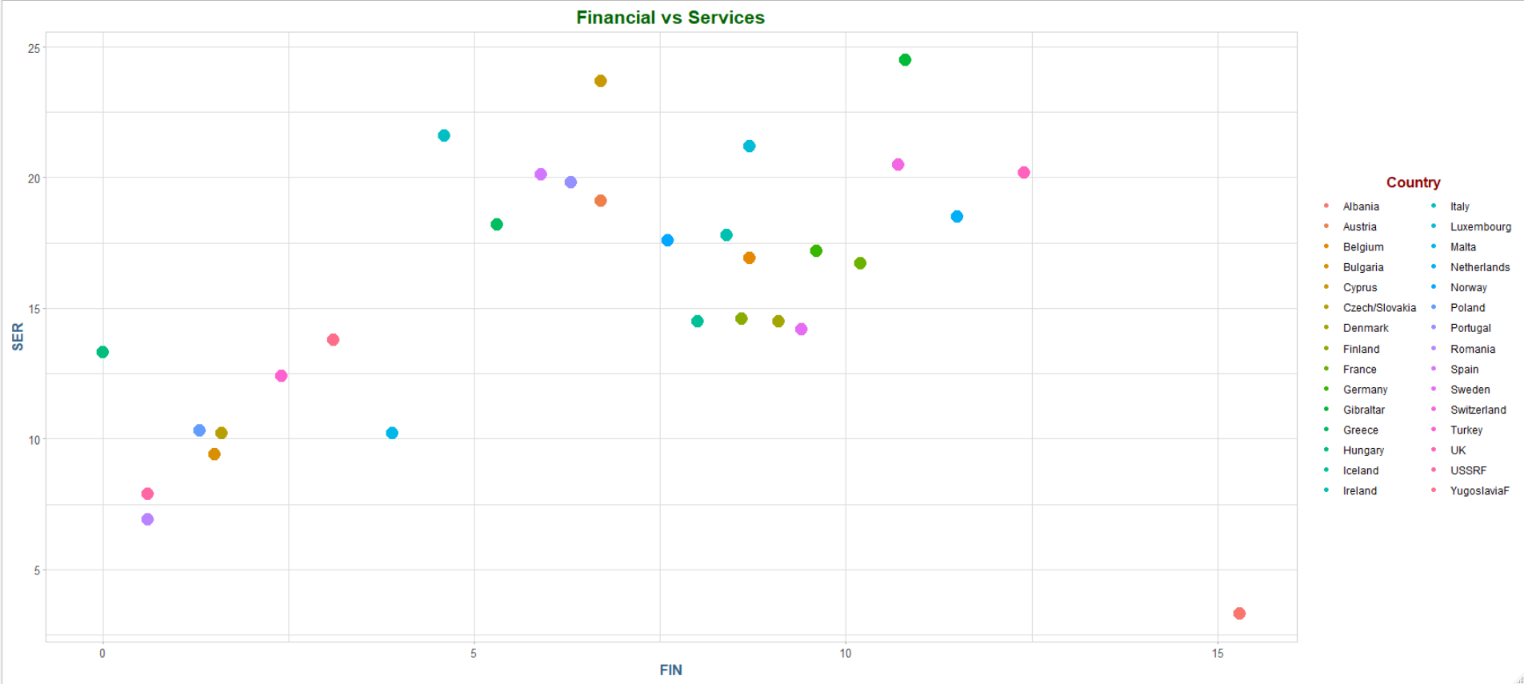
Brandon Moretz

### Tasks

1. Since we have a relatively small number of variables, we will begin our exploratory data analysis with a pairwise scatterplot. Obtain a pairwise scatterplot of the data. Note that when you have a small number of variables, the pairwise scatterplot is a useful statistical graphic. Another note about scatterplots – they are not very useful when we have too many data points.

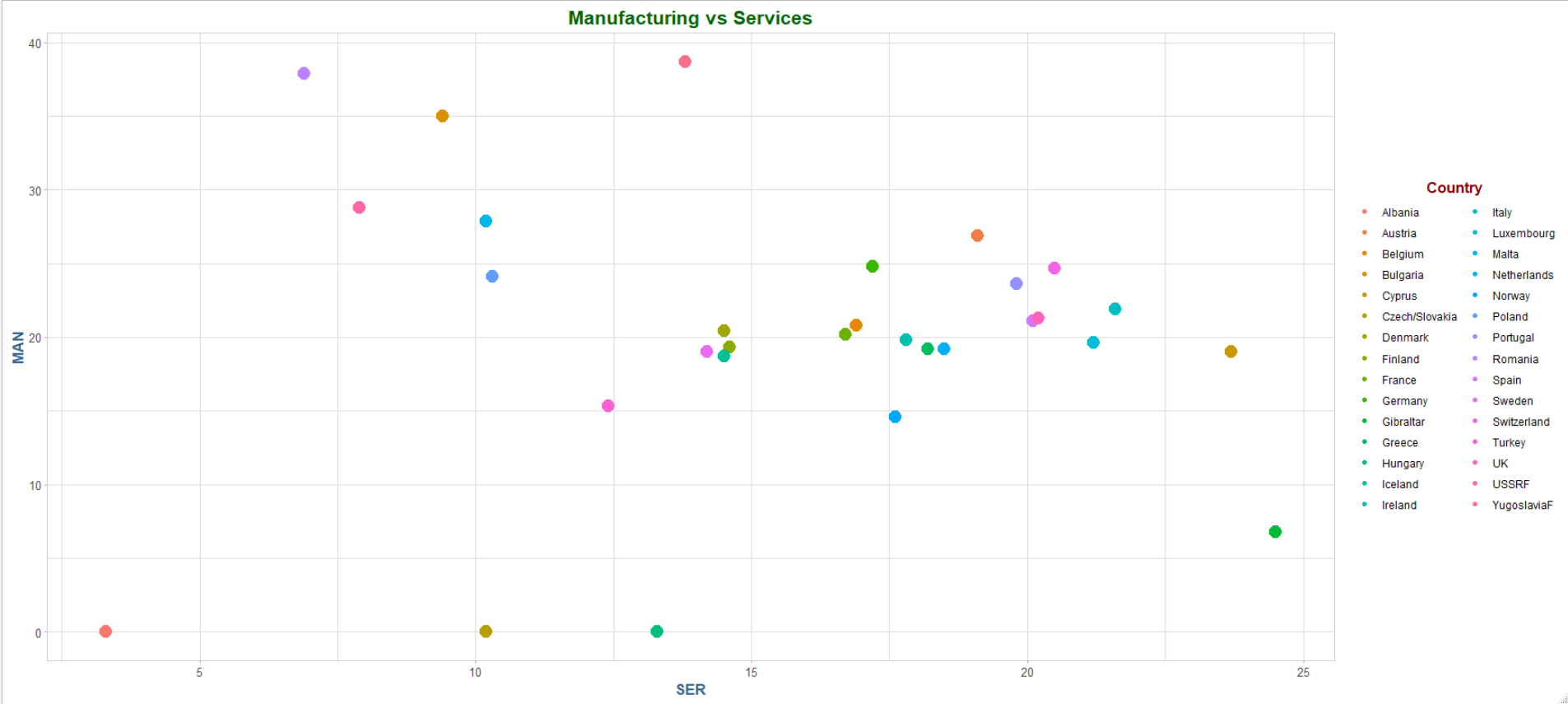
We are looking for groupings in the lower diagonal of the above chart. Some notables for further examination include, MAN/SER, SER/FIN, TC/FIN, SPS/SER.

1. Zooming in on some selections from above:
   1. FIN/SER



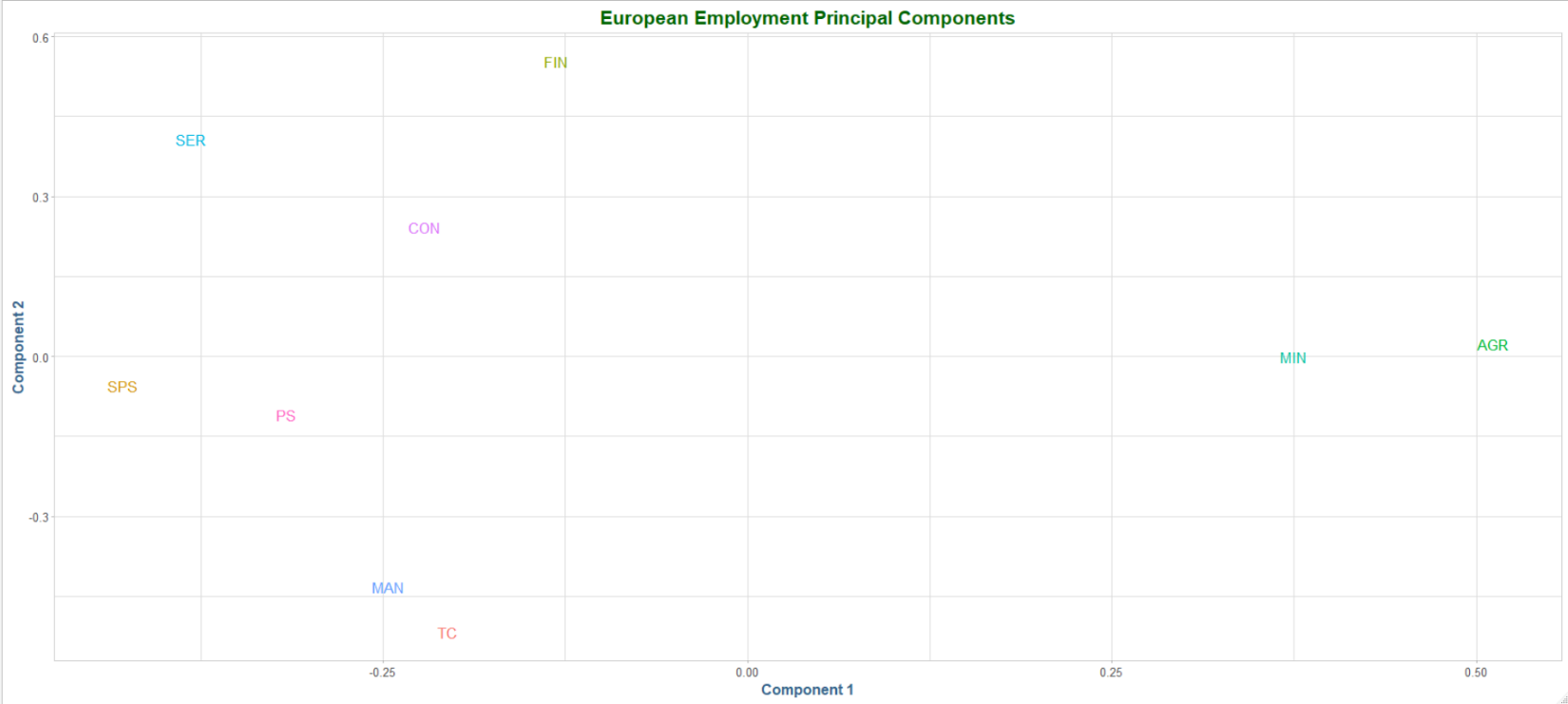
In the above chart I would think there are two or three distinct clusters/segments.

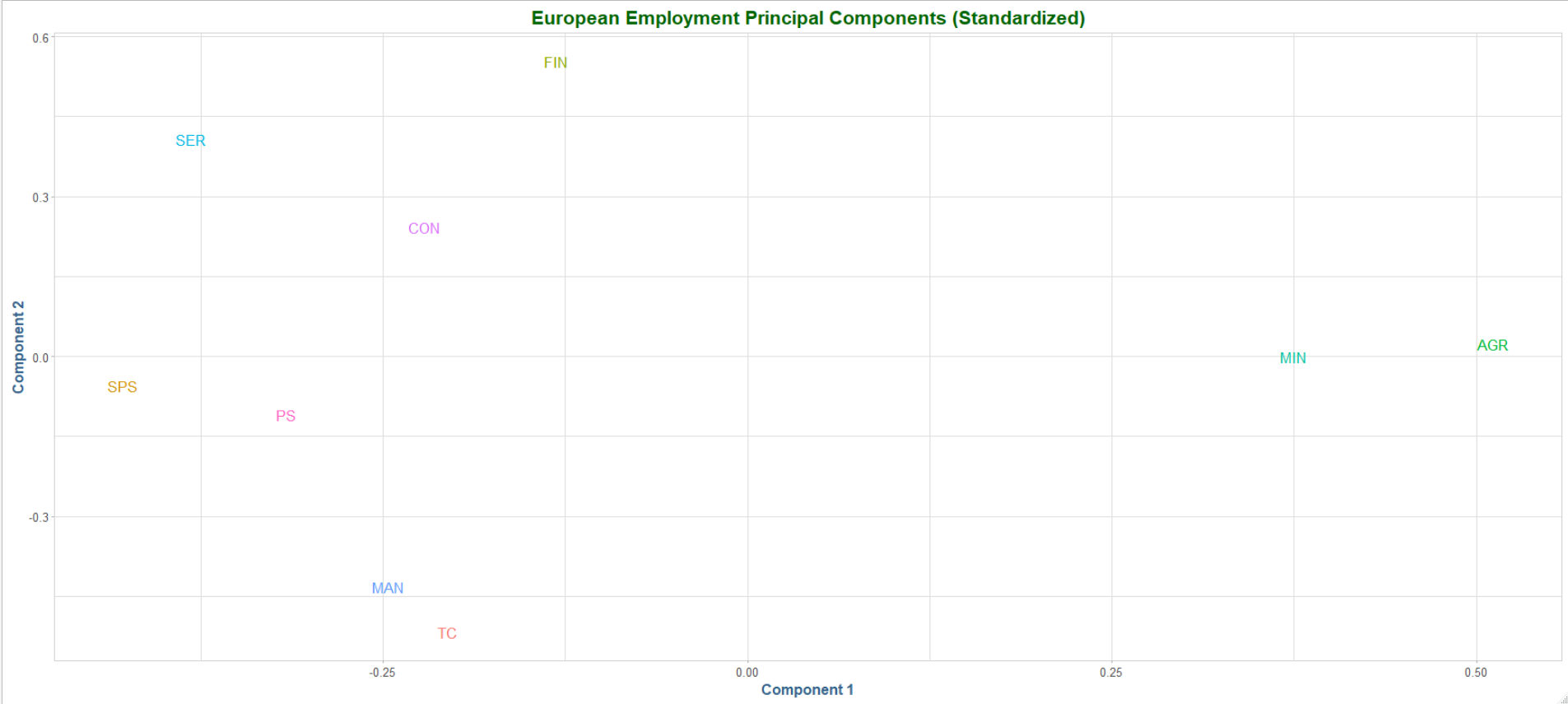
* 1. MAN/SER

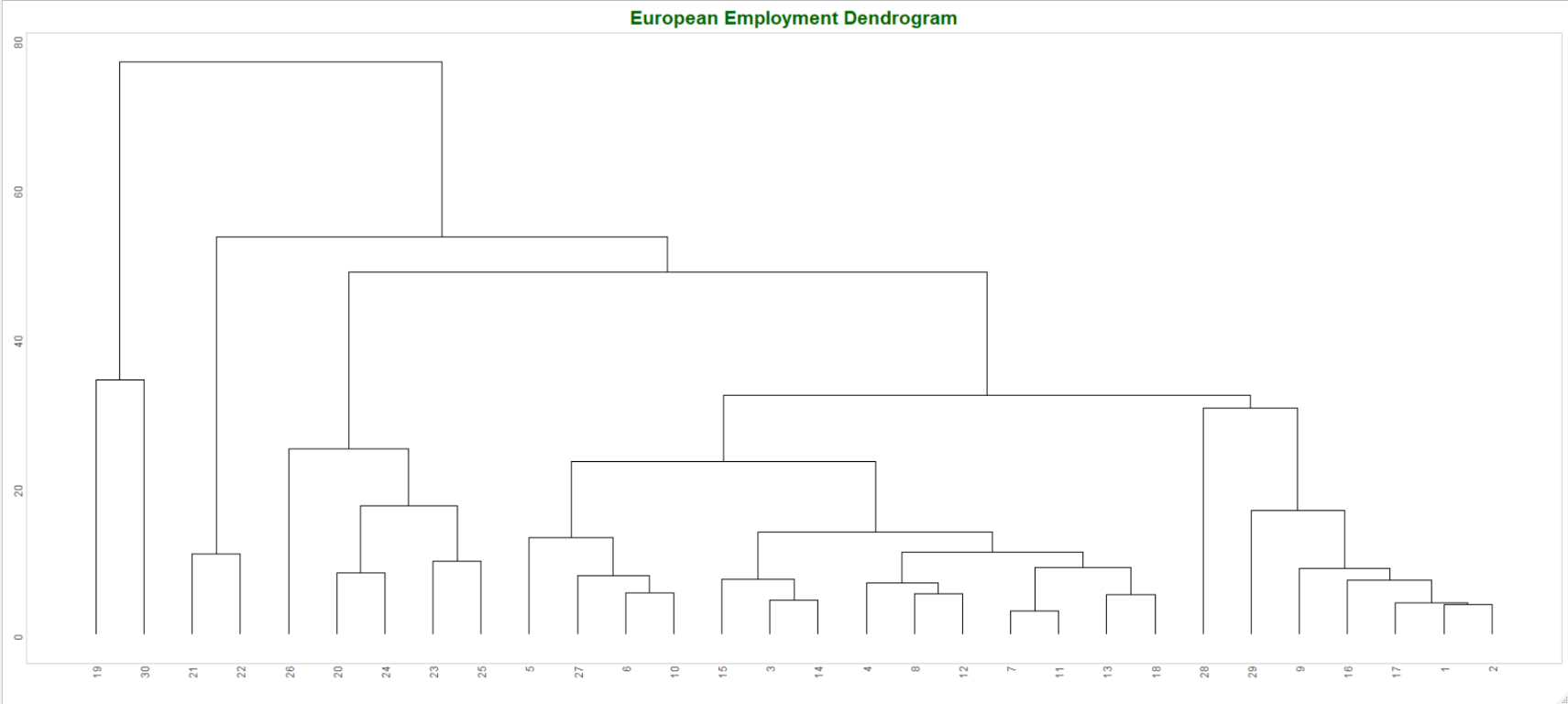


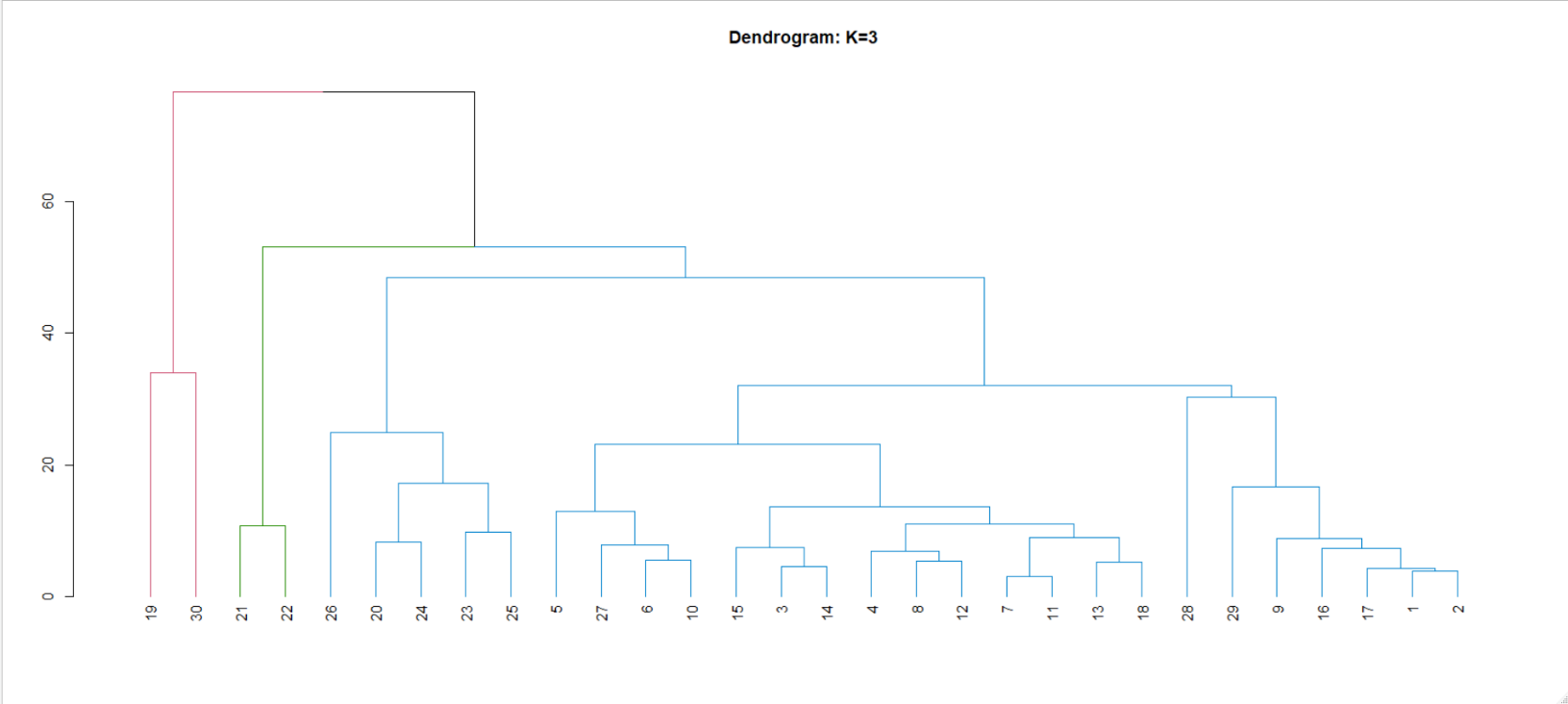
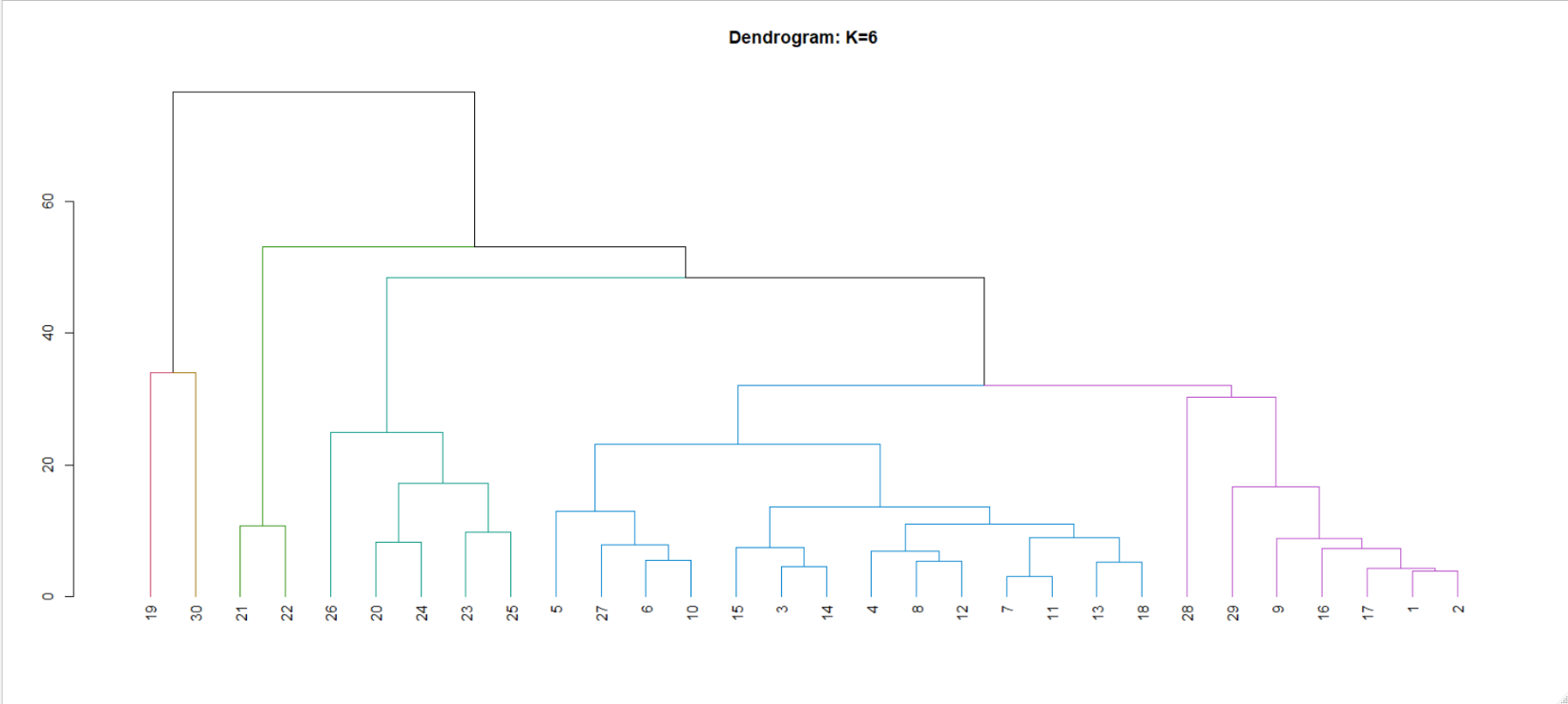
In the above plot I see four distinct clusters, which are much different than the previous plot.

* 1. I would think that that the first graph would be easier for an algorithm to cluster due to the clearer lines of separation.

1. We can use principal components analysis to reduce the dimension of the data. We can project the data down from 9D to 2D by performing PCA and using the first and second principal components. By doing so we are creating a new 2D view of the data, and a view of the data that contains information from more than two dimensions.
   1. The first two principal component loadings:
   2. There seems to be no noticeable difference between the standardized and non-standardized versions of the PCA.

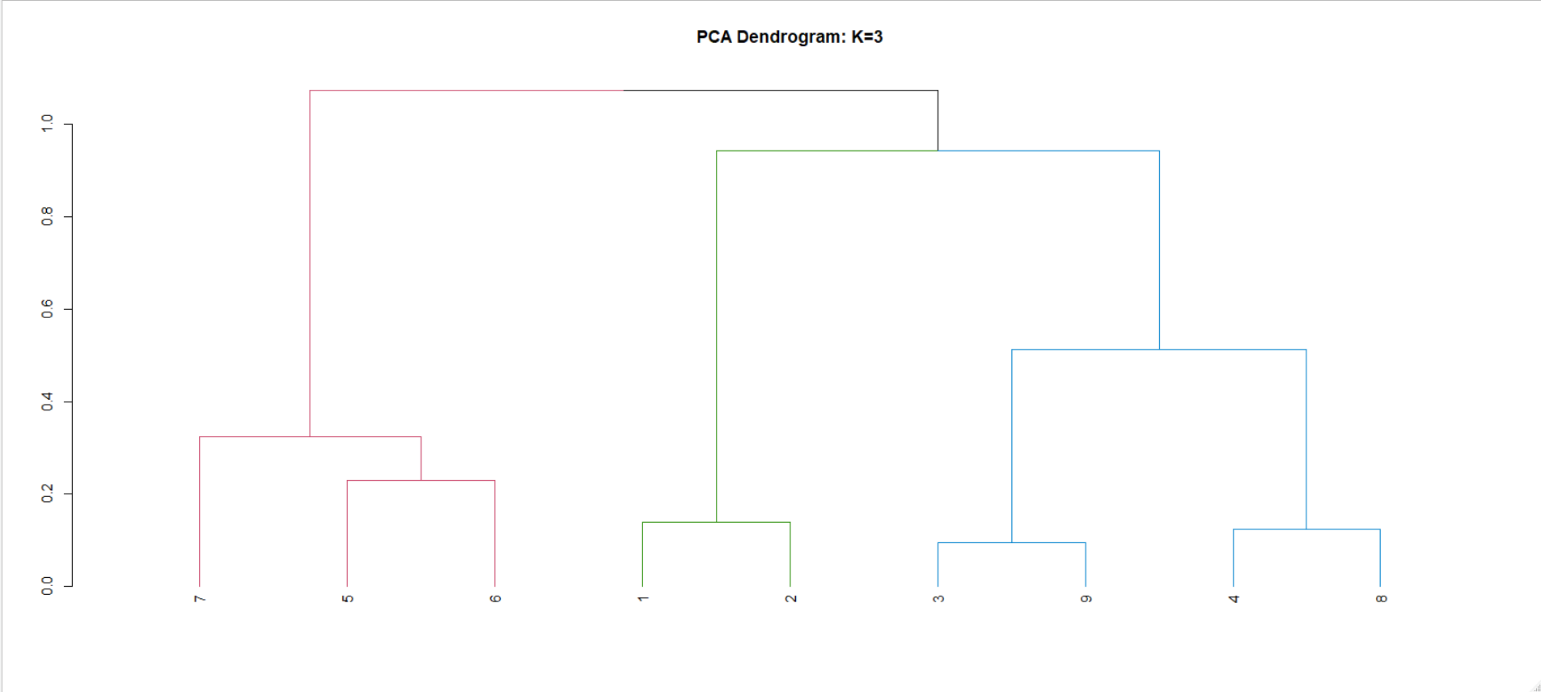
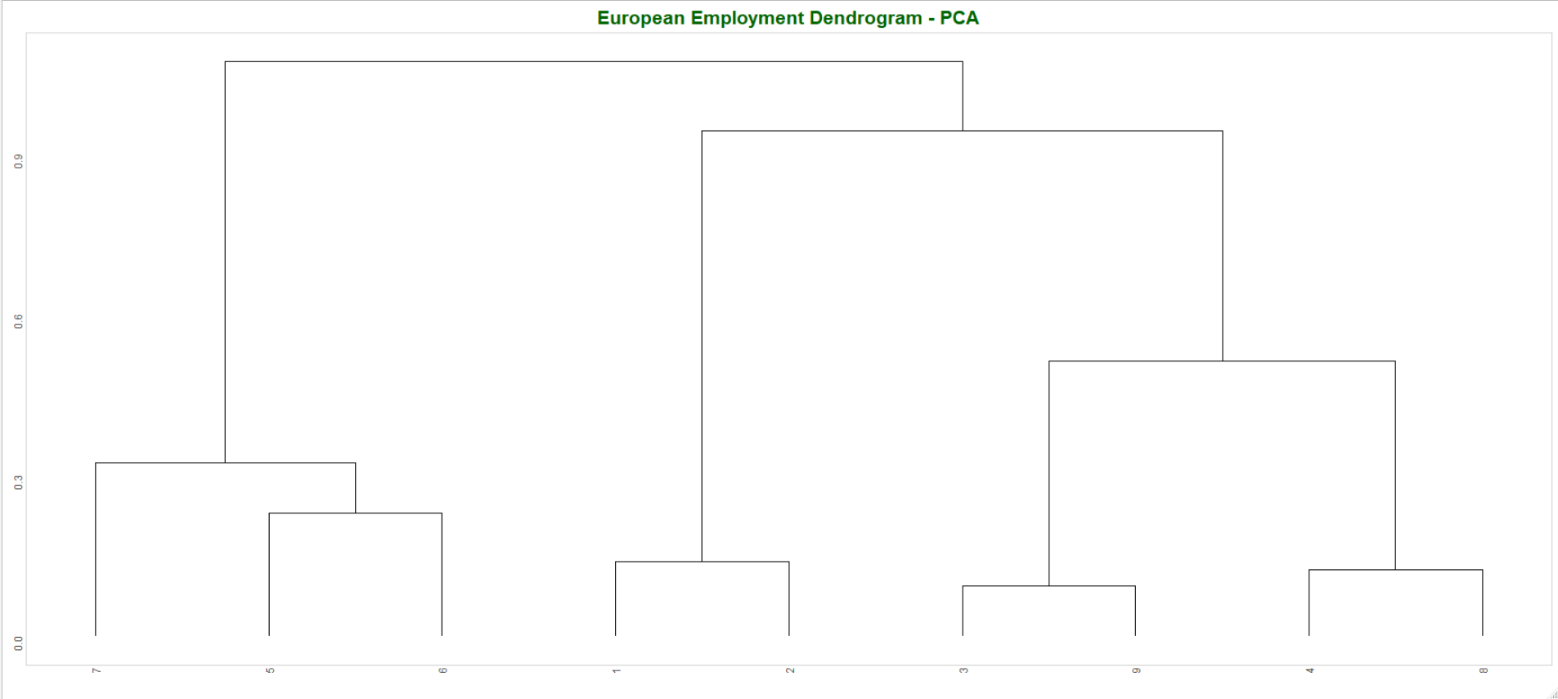
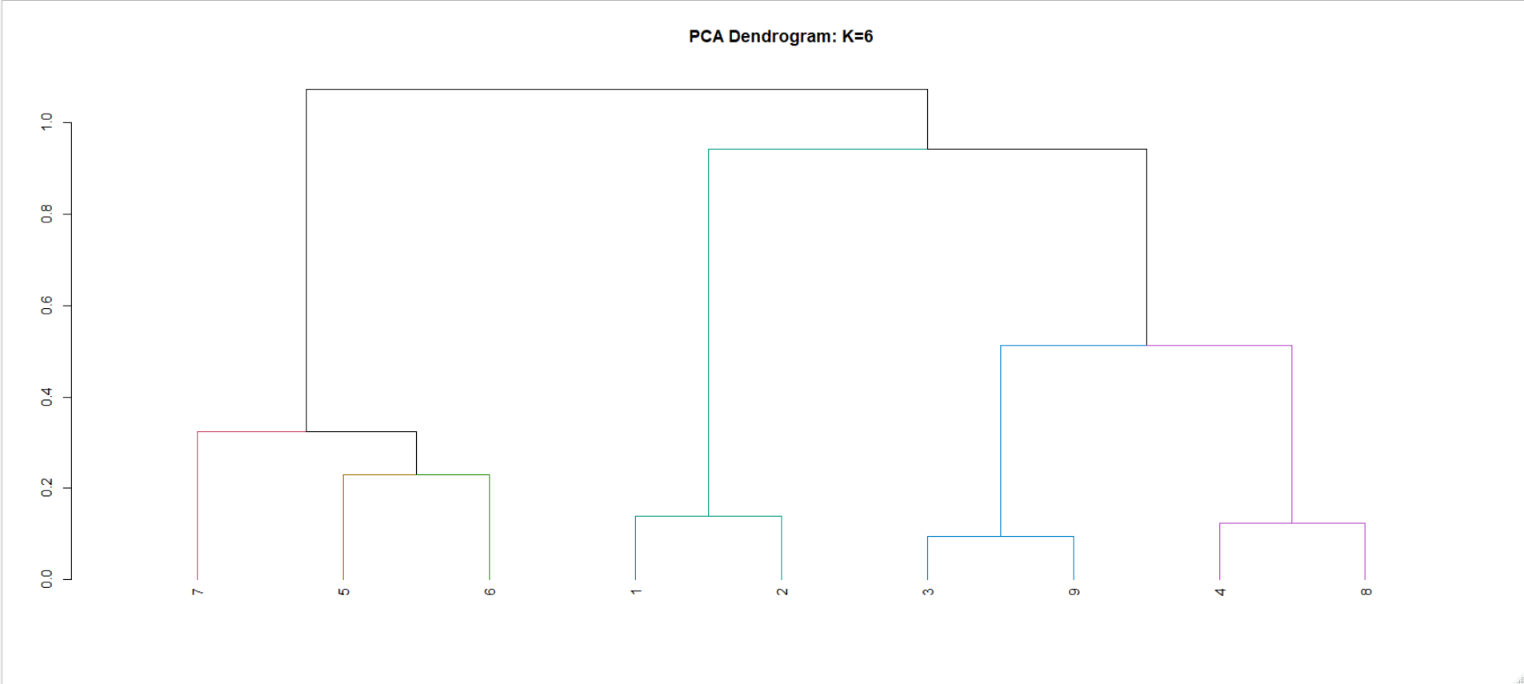


1. Hierarchical clustering algorithms fit a tree of clusters from k=2 to k=N, where N is the number of data points in the sample.  As you know, this tree of clusters can be visualized using a dendrogram. Since the cluster tree stores all possible cluster assignments, we must cut the tree using cutree() to force an assignment of the observations to a particular number of clusters.
   1. Obtain a dendrogram:

Dendrogram cuts:

Six clusters are a better fit here.

* 1. Perform the same analysis, but this time use the principal component space using the first and second principal components. Of these four ‘cluster models’ which one is the most accurate? Make a table to display their accuracy for easy comparison.



### Conclusion