NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Assignment #4: Cluster Analysis
MSDS 411

Data:

The data for this assignment is the European employment data set. This data is posted in Canvas.

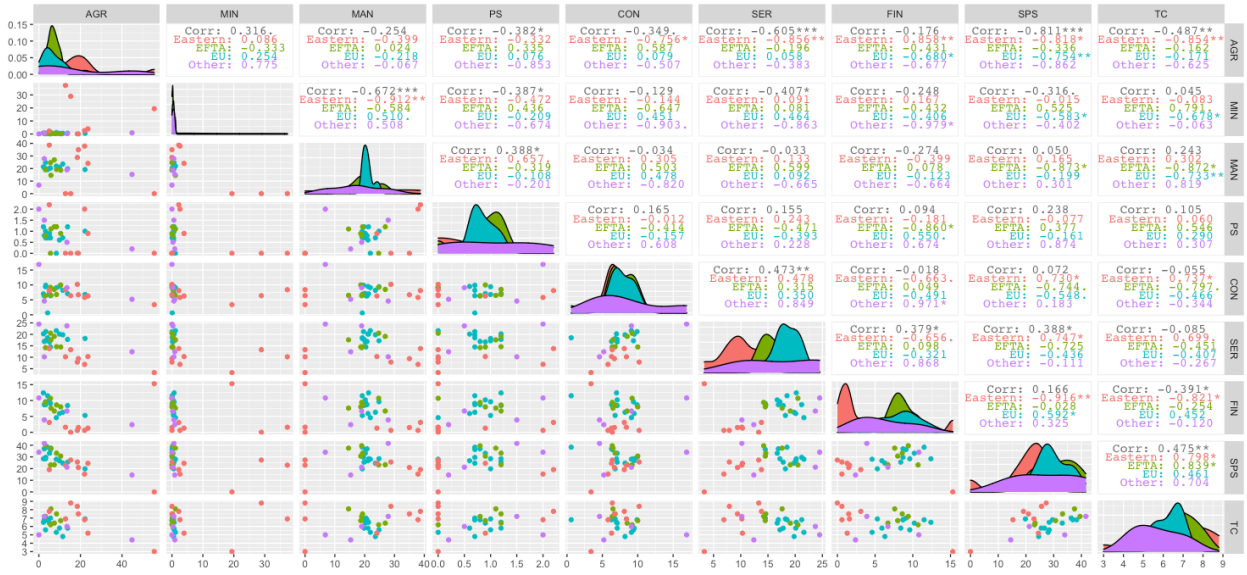
Data Description:

Employment in various industry segments reported as a percent for thirty European nations. Note that EU stands for European Union, EFTA stands for European Free Trade Association, and Eastern stand for Eastern European nations or the former Eastern Block.

For convenience here are the definitions of the abbreviated industries.

AGR: agriculture
MIN: mining
MAN: manufacturing
PS: power and water supply
CON: construction
SER: services
FIN: finance
SPS: social and personal services
TC: transport and communications

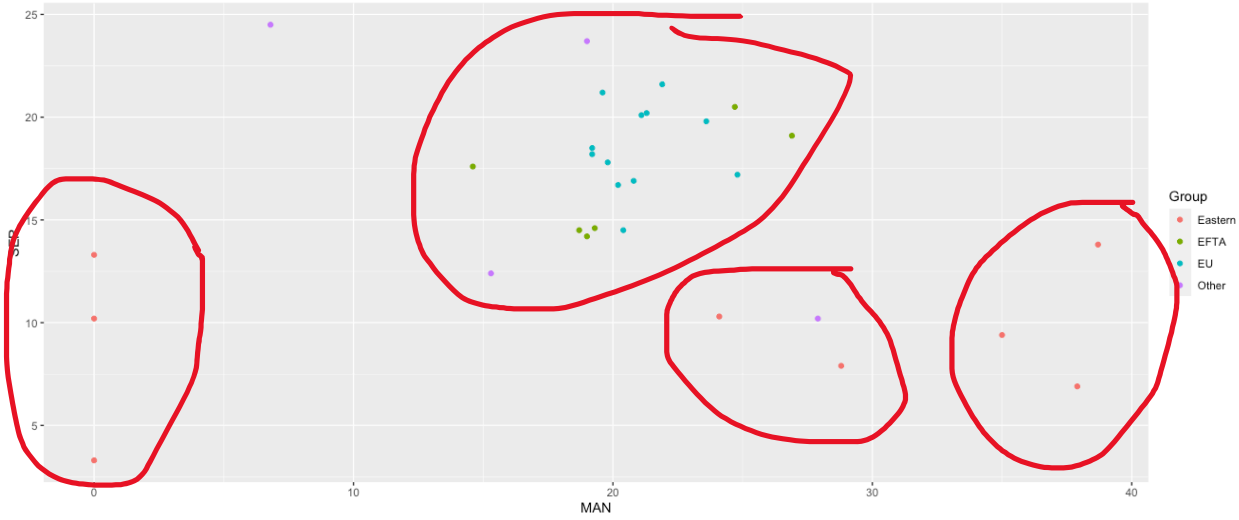
(1) Initial Exploratory Data Analysis: Obtain a pairwise scatterplot of the data.

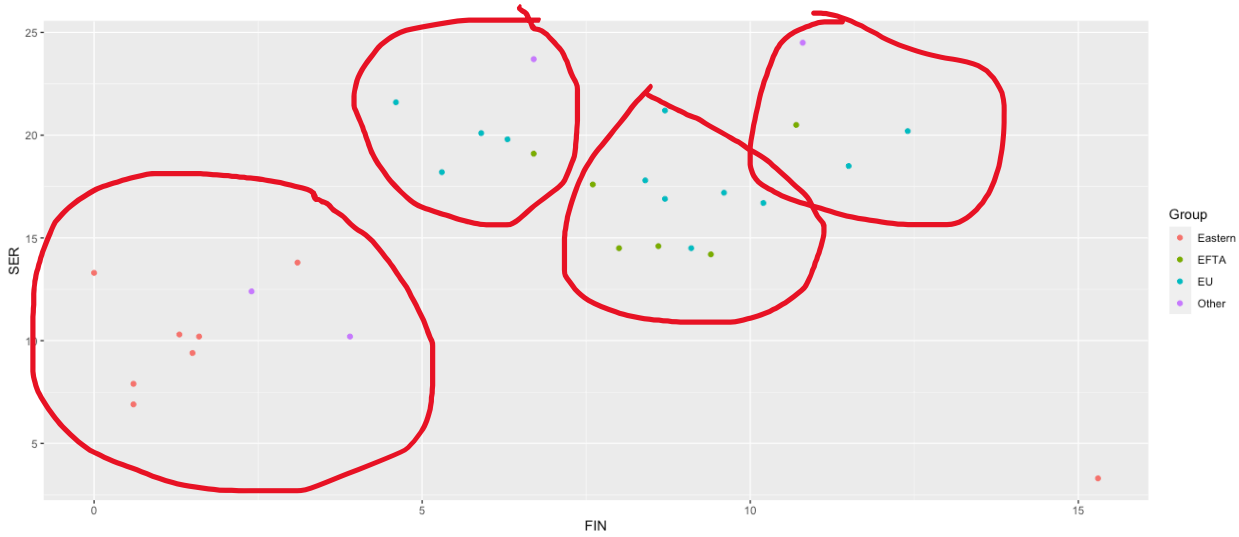


Do you see any interesting 2D views of the data? What would be 'interesting'? Remember, we are interested in applying cluster analysis so 2D plots that show clusters are the plots that would be interesting. Why don't we consider MAN versus SER and SER versus FIN? Do these 2D views look interesting?

- In the context of this analysis, "interesting" would be visible groupings of countries in a pairwise scatterplot. What would not be interesting in this analysis, but might be interesting in another analysis would be something like the clear linear relationship between AGR and SPS, because there are no clear groupings. We just see a continuous linear relationship.

  The pairwise plots with the clearest groupings are MAN versus SER and FIN versus SER. Plotted below I added in rough circles around the clusters that I feel the observations most neatly fall into, although I think argument could be made for consolidating some groups, especially in the FIN versus SER plot.

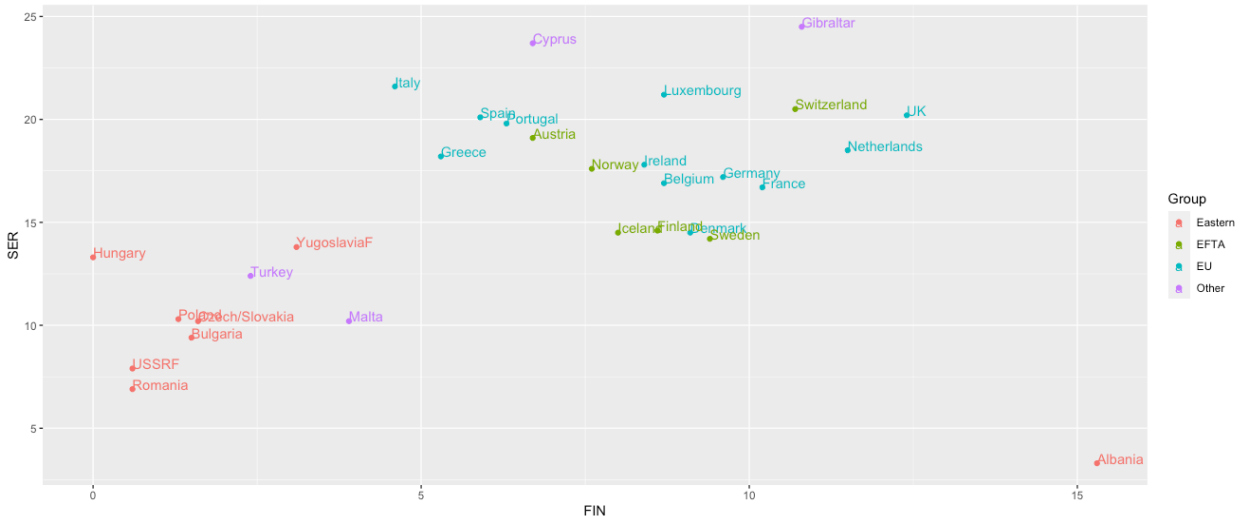




(2) Visualizing the Data with Labelled Scatterplots: While the pairs plot allows us to scan all pairwise scatterplots easily and efficiently, it is not the ideal visualization of the data. After we have homed in on some interesting dimensions we can create more specialized plots for those dimensions. Specialized plots should always include labels and color. The objective is to compress more than two dimensions of information into a two dimensional plot.

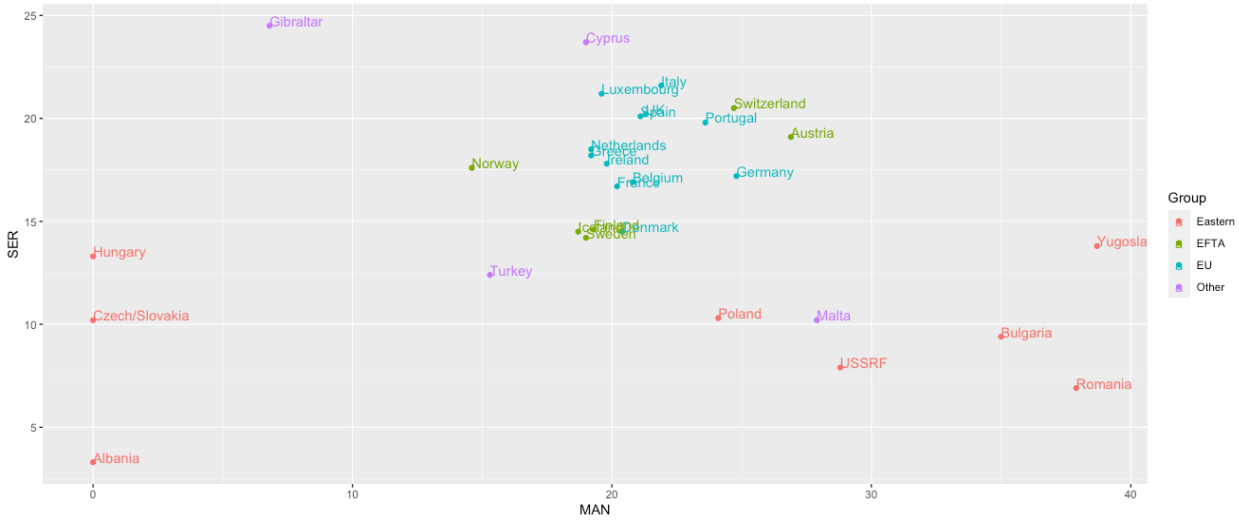
a) Plot FIN versus SER. Do we see some clusters in this plot? How many clusters do we have? How many clusters would you have if you were creating a segmentation?

- There are a couple distinct clusters here. In the previous answer I had said four clusters. I think the labels might be contributing to the perception but there are definitely two very distinct groups, while the upper right cluster could conceivably be broken down into smaller clusters. But for the purposes of this analysis, I'll say two clear clusters: the Eastern + Turkey and Malta vs the EU/EFTA + Cyprus and Gibraltar.



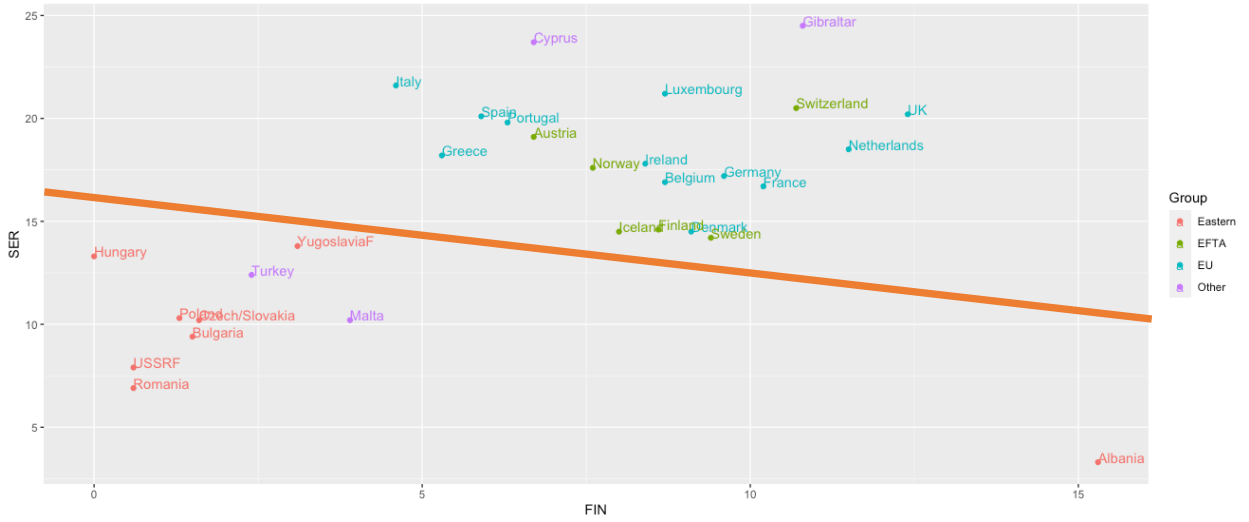
b) Plot MAN versus SER. Do we see some clusters in this plot? How many clusters do we have? Are they the same clusters as we saw in the previous plot? How many clusters would you have if you were creating a segmentation?

- I would likely revise my previous answer again and say there are three clusters here and they largely overlap with the alignment of groups from part A in that the EU/EFTA countries are clustered together. The difference in this chart compared to the previous is that the Eastern bloc countries are stratified into those with no manufacturing versus those that have a lot of manufacturing.



c) Of the two 2D views of the data which one do you think would be the better view for supervised clustering, i.e. using a clustering algorithm to create a classifier that will assign the countries to the correct class/label? Why?

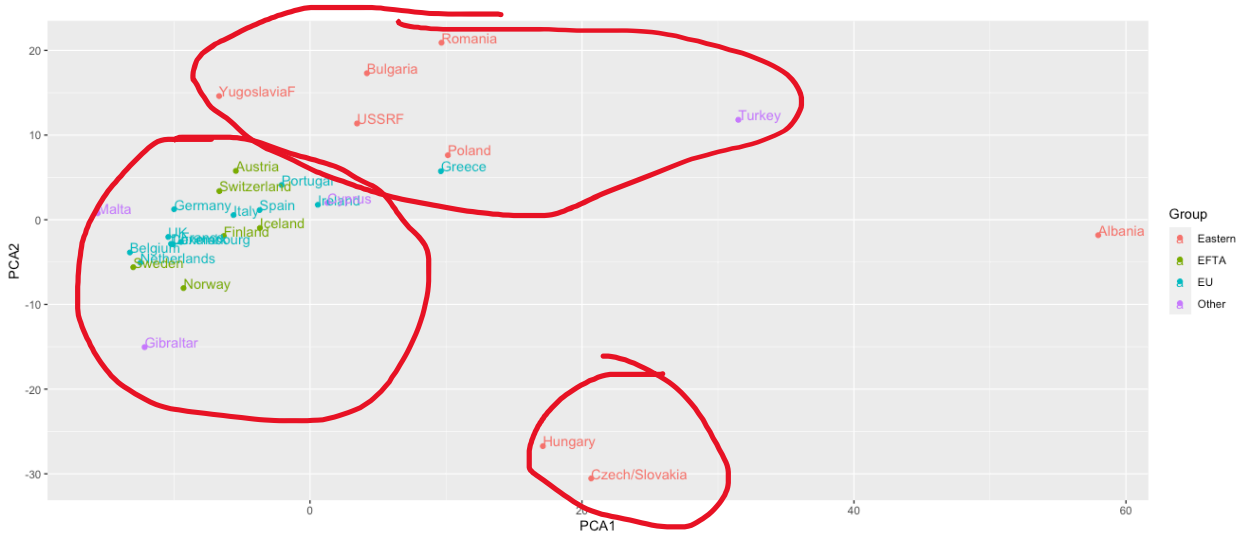
- The FIN versus SER plot would seem to be more well suited for a supervised clustering. It is very clear that you can basically draw a line that would perfectly separate all the Eastern countries from the EU and EFTA countries. It is possible that a more sophisticated machine learning algorithm could take the MAN versus SER plot, correctly predict the Eastern countries based on their distinct clusters and then predict either EFTA or EU based on the distance from the center of the cluster where MAN is ~21 and SER is ~19 due to the way the EU countries seem to be closer to the center of that cluster. But this seems like a bit of a stretch.



(3) Creating a 2D Projection Using Principal Components Analysis:

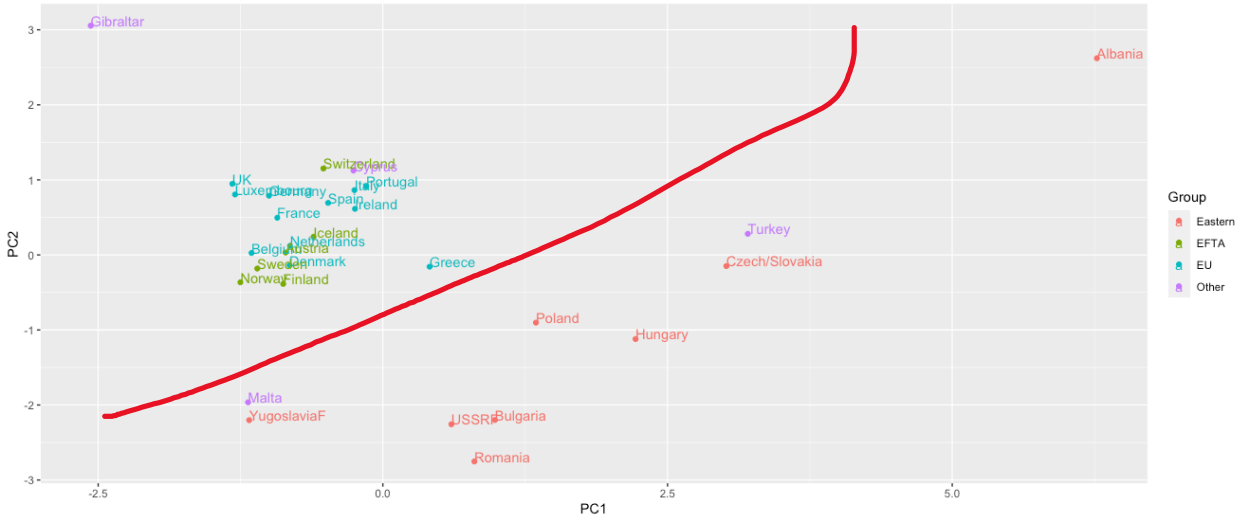
a) Use the raw data and conduct a PCA. Plot the first two principal components. How does this 2D projection of the data compare to the two other views of the data that we are considering? How many clusters does this 2D projection have? Clearly, our data can have different degrees of separation in different 2D profiles, and hence some low dimension representations will be better clustered than others.

- The plot below has three rough clusters as depicted below. These do not feel as clear as the clusters that we had previously observed. The separation between the EFTA/EU cluster and the upper Eastern cluster does not feel as distinct as it appeared in the FIN versus SER plot. Greece, which was clearly clustered with the other EU countries before is now mixed in with the Eastern countries. Albania is still an outlier but does not have a clear rule for classification as it did previously.



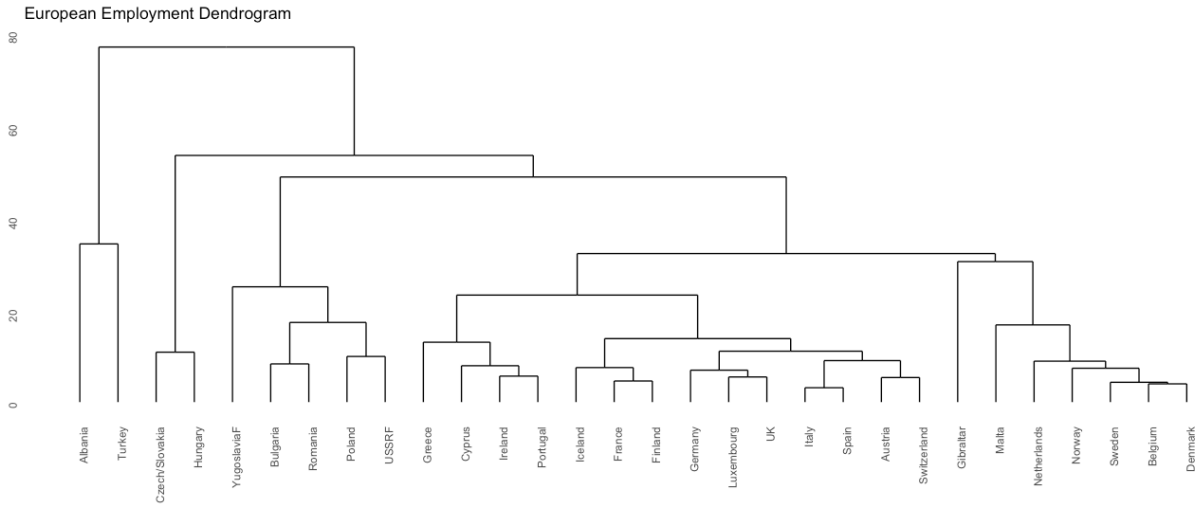
b) Usually, one is supposed to standardize the data to be mean zero with unit variance before performing a PCA. Standardize the data and run a second PCA on the standardized data. Compare the two results. Does standardizing have much of an effect here?

- There does seem to be some effect from standardizing the data. You can now see a fairly clear dividing line between the Eastern countries and the EU/EFTA countries.



(4) Hierarchical Clustering Analysis:

a) Perform a hierarchical cluster analysis and obtain a dendrogram.



Use the cutree() function to force an assignment of the observations to a particular number of clusters. Use k=3 and k=6 and compare the classification accuracy of two cluster tree cuts. Which set of clusters is more accurate?

K	Between Sum of Squares Pct

3	0.5889724
6	0.8428839

- The tree with k = 6 is more accurate than the k = 3 tree.

b) 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.

K	Between Sum of Squares Pct
3	0.5889724
6	0.8428839
3 (PCA)	0.5048075
6 (PCA)	0.8037453

- The most accurate model is the k = 6 from the full dataset without principal components. I was expecting the PCA versions to perform better, but it makes sense that the models that are using a subset of the total variance would lose some of its ability to correctly classify.

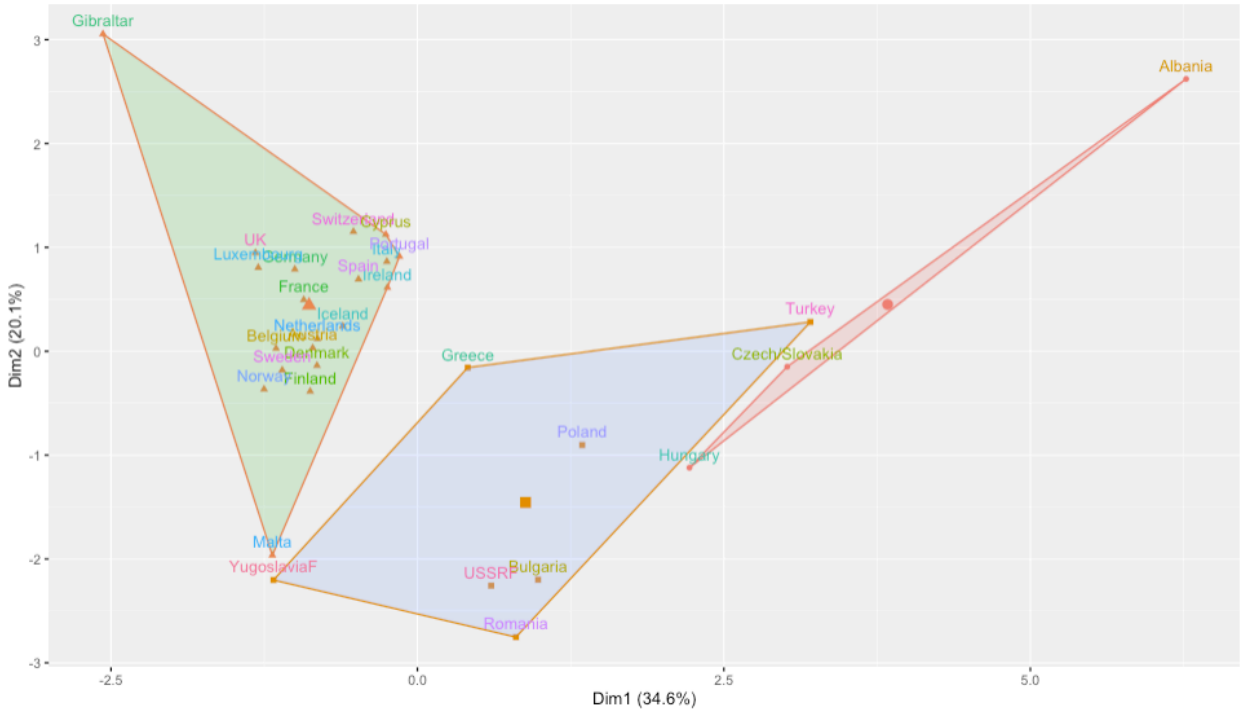
(5) k-Means Clustering Analysis:

a) Conduct a K-Means Cluster Analysis on the European Employment data for k=3 and k=6. Compare the classification accuracy of these models with the hierarchical models obtained in task (4).

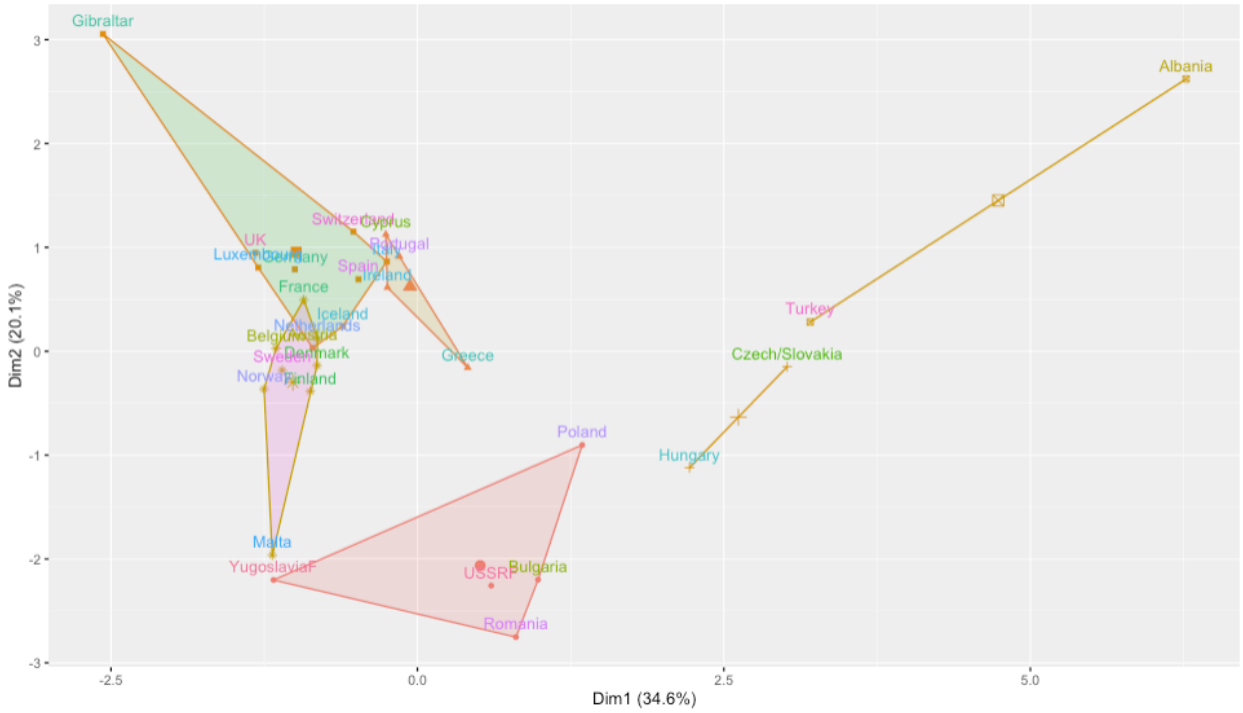
K	Between Sum of Squares Pct
3 - heriarchical	0.5889724
6 - heriarchical	0.8428839
3 (PCA) - heriarchical	0.5048075
6 (PCA) - heriarchical	0.8037453
3 - k means	0.5792964
6 - k means	0.8274197

b) For the k-Means Cluster Models obtain a plot that includes the original labels, their assigned clusters, and the cluster centers. What do you see in these two graphics?

- K = 3 plot:



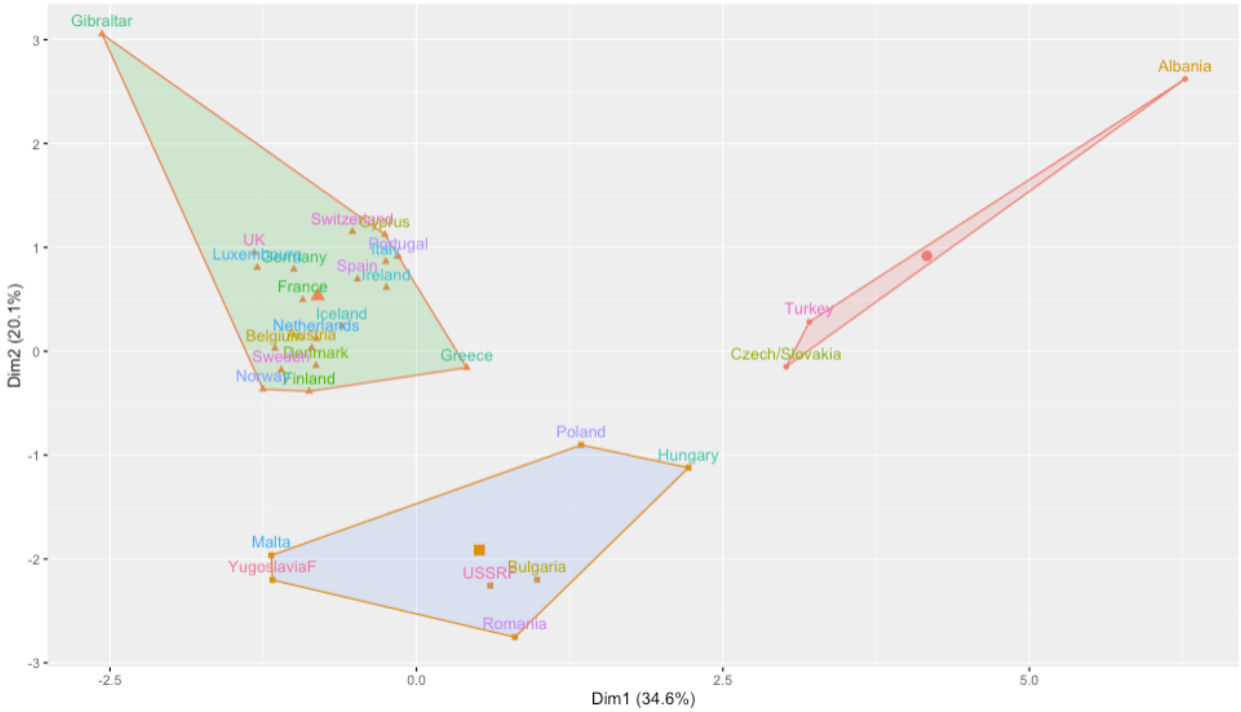
- K = 6 plot:



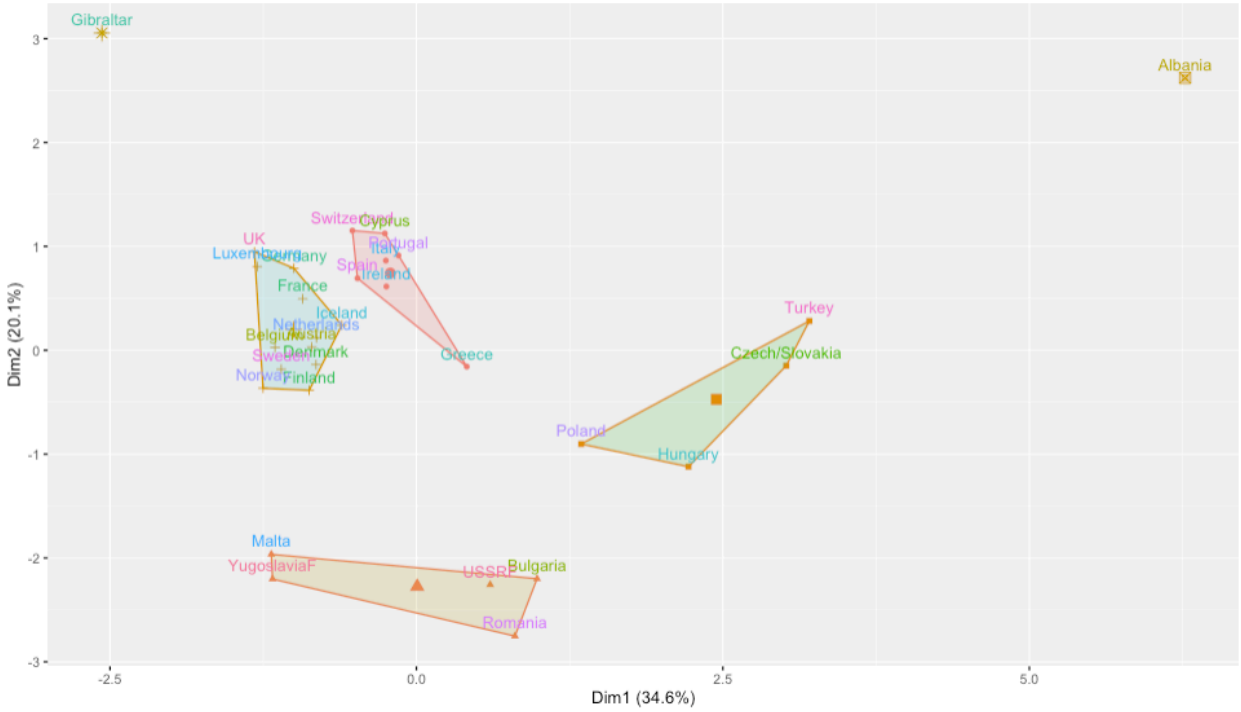
- The actual plotting of the points are the same. The only difference in the two plots is where the boxes are drawn around the clusters. When increasing k, all we are doing is finding the new optimal center points for the given number of clusters, not making any changes to the relative distribution or locations of the points themselves.

c) Conduct a K-Means Cluster Analysis for k=3 and k=6, but use the Principal Components space.

- K = 3:



- K = 6



- These plots are the same as their non-PCA counterparts since the plotting function uses the first two principal components. The only difference is the shapes and centers of the clusters.

d) What happens as we increase the number of clusters from k=3 to k=6?

- The between sums of squares percentage increases as the number of clusters increases. We get smaller clusters with the points more closely located to their cluster's center point.

e) Of these eight cluster models which is the most accurate? Make a table summarizing the eight models and their accuracy.

- The k means model using principal components where k = 6 is the most accurate model.

K	Between Sum of Squares Pct
3 - heriarchical	0.5889724
6 - heriarchical	0.8428839
3 (PCA) - heriarchical	0.5048075
6 (PCA) - heriarchical	0.8037453
3 - k means	0.5792964
6 - k means	0.8274197
3 (PCA) – k means	0.7051841
6 (PCA) – k means	0.907592

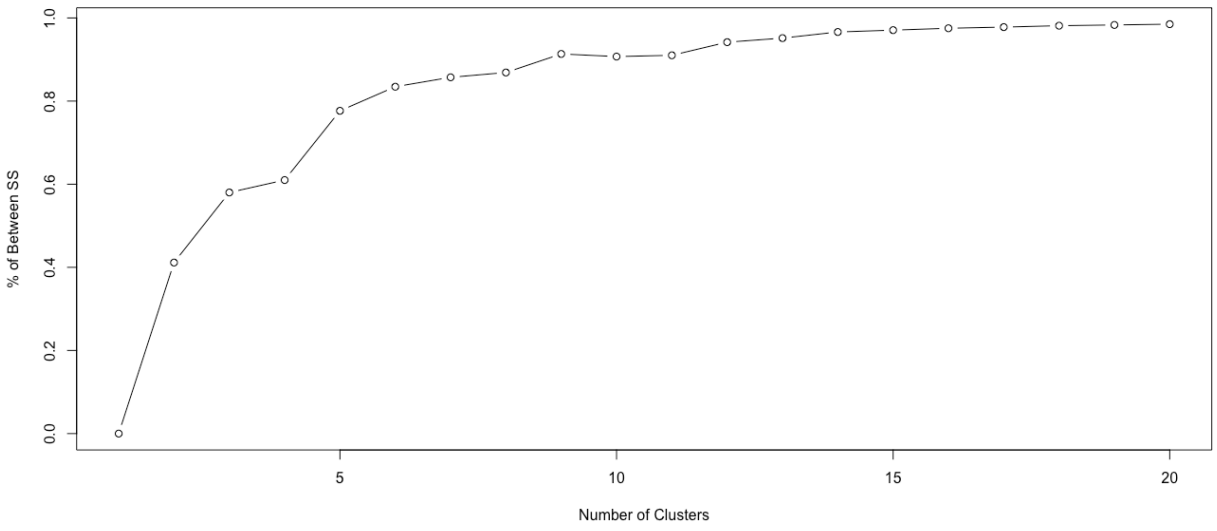
f) How do the clusters compare with the original *labels* (EU, EFTA, Eastern, or Other)?

- The clusters still seem to group together the EU and EFTA countries, but seems to accurately separate out the Eastern and Other countries.

(6) Computing the 'Optimal' Number of Clusters by Brute Force:

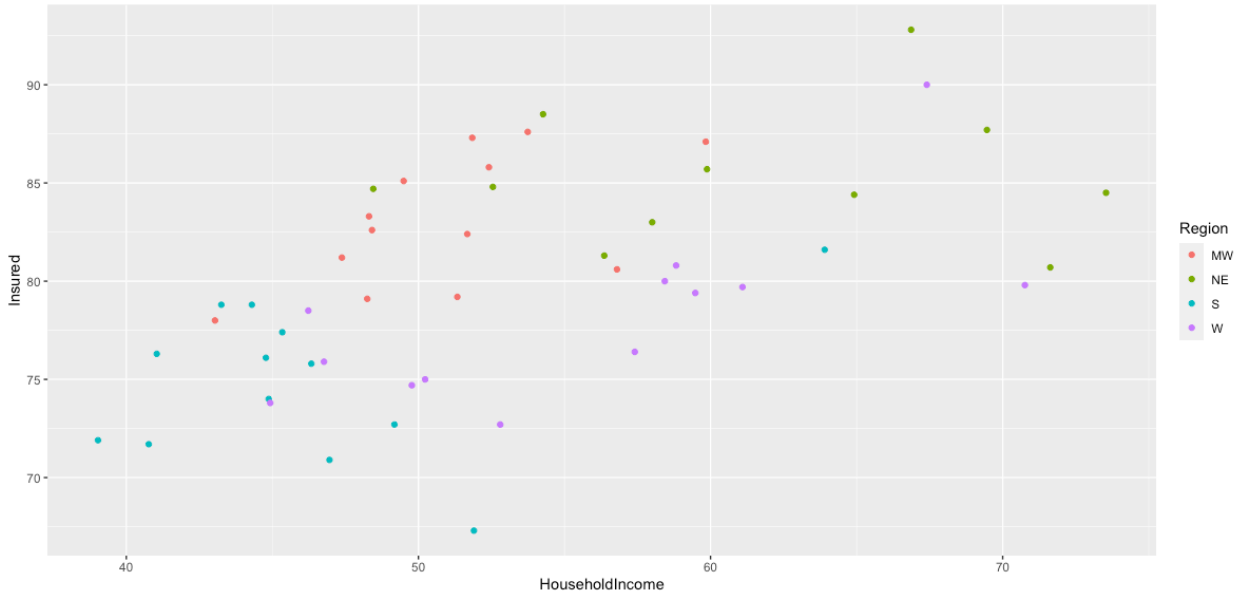
a) Obtain and plot the classification accuracy for k=1 to k=20 for both hierarchical and k-means clustering algorithms. What can you conclude based on this graph?

- In this graph you can see that there are significant increases in the accuracy of the model early on. Once you get to 5 clusters, you don't get as large of increases in the model accuracy. Based on this graph I think 5 clusters would be the ideal number. After you get past five, there is a diminishing return to adding any more, which signals to me that we're adding more clusters but not making any significant progress in improving the accuracy when doing so. Its likely that there are some pretty good clusters that are getting broken up into a couple smaller clusters, which is not going to be very useful in most applications.

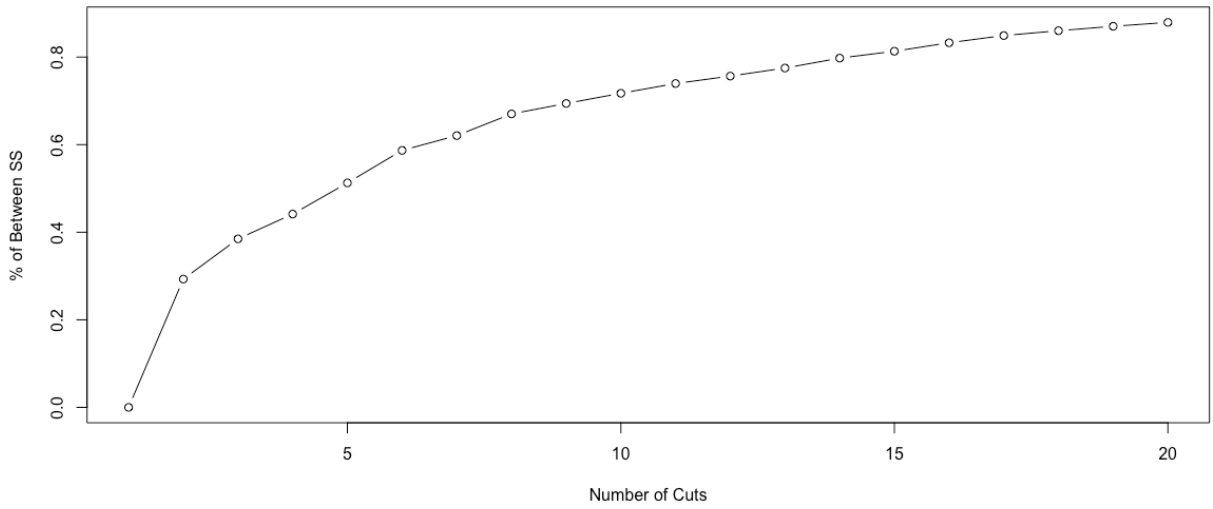


(7) On Your Own Modeling 1: The USSTATES dataset is a 12 variable dataset with n=50 records that you used briefly in MSDS 410. The data, calculated from census data, consists of state-wide average or proportion scores for the non-demographic variables. As such, higher scores for the composite variables translate into having more of that quality. There is no other information available about this data. Use this data set and conduct a hierarchical cluster analysis. Decide on the total number of clusters to retain and describe the differences amongst the clusters.

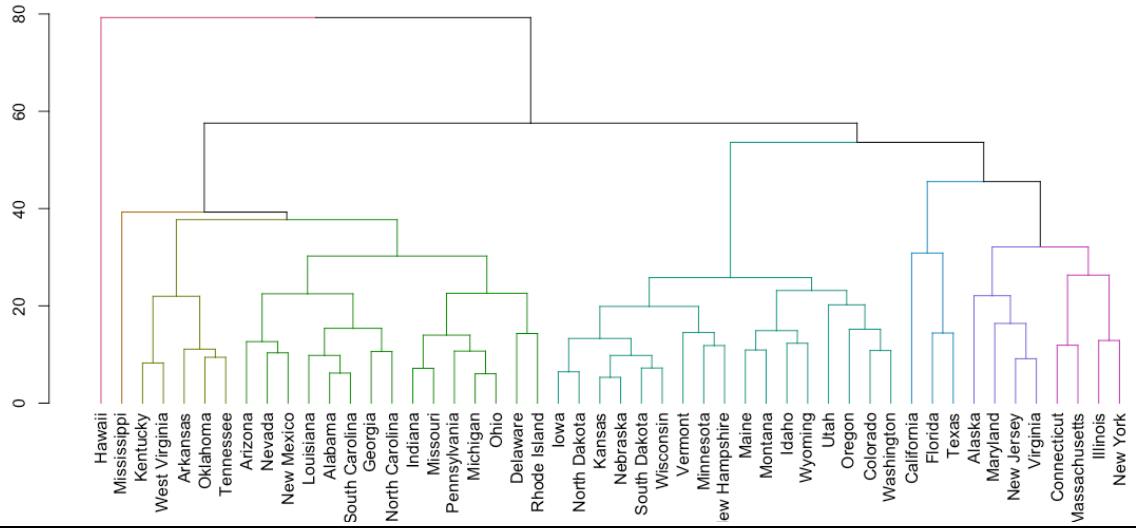
- An initial EDA didn't yield any obvious pairwise clusters. One that seemed potentially interesting was HouseholdIncome versus Insured. While I was hoping there would be some clear regional differences, it looks like the regions are fairly interspersed, although the southern states are mainly centered in the bottom left quadrant.



- In determining the number of clusters to retain, I used the following plot of the between sum of squares percentage. 8 clusters appears to be the right number to retain based on this plot as that is when we see the slop of the line begin to level off. 8 clusters leaves us with a percentage of 68.83% of the between sum of squares.



- Using 8 cuts gives us the following dendogram.



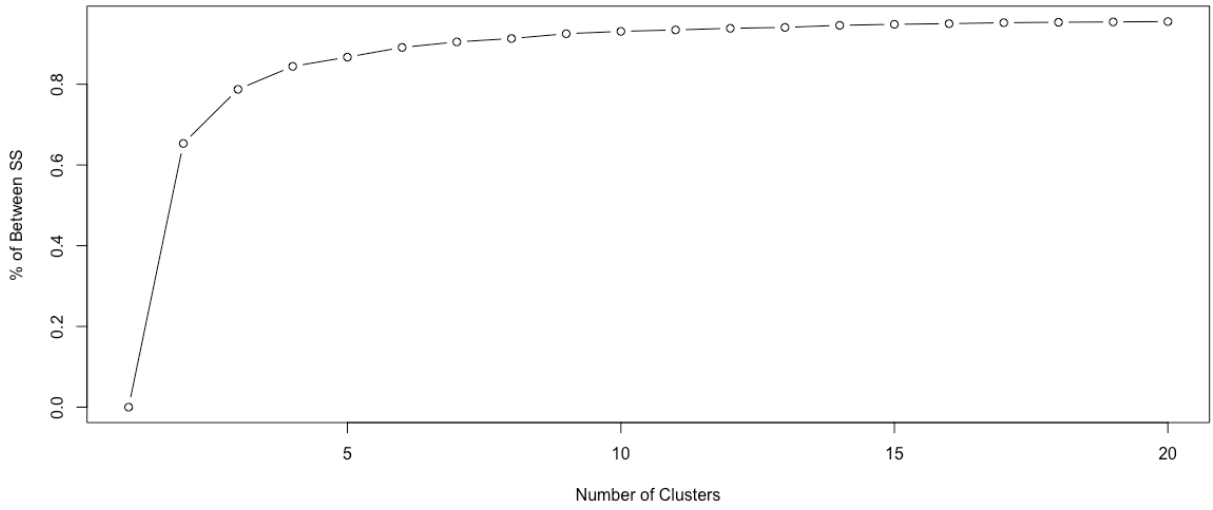
- Here is the breakdown of the 8 clusters:
    1. Hawaii: Hawaii is a very unique state. It has high values for HighSchool, but relatively low College. It also has high values for PhysicalActivity, Insured, and TwoParents. The most significant value that sets Hawaii apart is its huge non-white population of 75.0, an 80% increase over the next closest state.
    2. Mississippi: Another unique state, Mississippi has the 3rd highest non-white population. It is much less educated, poorer, and more obese than almost every other state.
    3. KY, WV, OK, TN, AR: These mostly Southern states have low end household incomes, education, insurance, and heavy drinking rates, while having high smoking rates.

4. AZ, NV, NM, LA, AL, SC, GA, NC, IN, MO, PA, MI, OH, DE, RI: A broad range of states that are pretty middle of the road in most metrics
5. IA, ND, KS, NE, SD, WI, VT, MN, NH, ME, MT, ID, WY, UT, OR, CO, WA: Like the previous cluster but with higher rates of Physical Activity
6. CA, FL, TX: These states are all very large and have low insured rates
7. AK, MD, NH, VA: Wealthy states with smaller populations
8. CT, MA, IL, NY: Relatively wealthy states with larger populations and more diversity

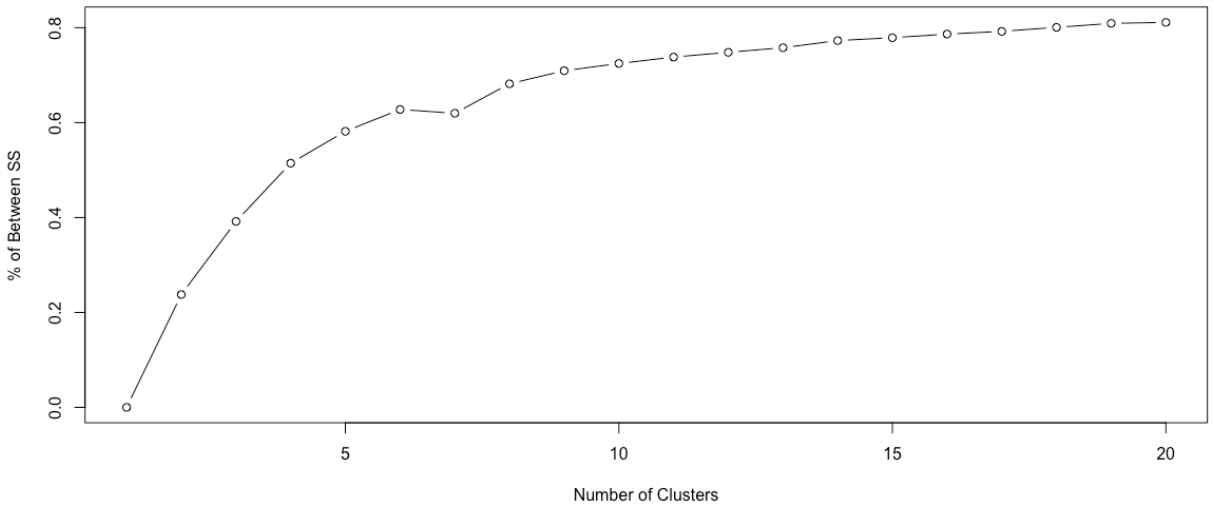
(8) On Your Own Modeling 2: The RECIDIVISM dataset is an 18 variable dataset with n=1445 records. Please see the data description file for the variable definitions and additional information about the dataset. The data consists of a random sample records on convicts released from prison during 1977/1978. Use this data set and conduct a kmeans cluster analysis. Decide on the total number of clusters to retain and describe the differences amongst the clusters.

- There are no real obvious clusters in the pairwise data. Since so many of the variables are binary, many of the pairwise plots have clusters in each corner of the chart.

- Based on the following chart, I have decided to use four clusters in the kmeans analysis. After four clusters we hit diminishing returns on additional clusters. Four clusters gives us about 80% of the between sums of squares.



- I decided against doing the kmeans clustering on any principal components because the results of the non-PCA graph above are much more clear than the PCA results.



- We end up with the following four clusters:
    1. Cluster 1 is most likely to be black of the four clusters and most likely to have committed a property crime. They are the least likely to have committed a crime against a person. Almost all have priors. On average this group served the most time with an average of 24.02 months. With an average age of 22, they are the youngest cluster. On average they recidivate after 17.88 months.
    2. Cluster 2 is the most educated cluster and serves the shortest sentence on average. 92% of the people in cluster 2 do no recidivate. They are a mix of black and non-black and have the lowest levels of both drug and alcohol problems
    3. Cluster 3 is another mix of black and non-black, but with elevated levels of drug and alcohol problems. This cluster is the most likely to have committed a felony and much more likely to have committed a crime against a person than cluster 2. They also served an average of 6.5 months more than cluster 2. With an average age of 34, they are much older than cluster 2. 35% of those in cluster 3 recidivated.
    4. Cluster 4 has the highest proportion of non-black people. They are the most likely to have problems with alcohol and drugs and also the most likely to be married. They have low felony rates, which likely contributes to lower sentences than Clusters 1 and 3. Cluster 4 has an average age of almost 52 years old, making them far and away the oldest cluster. They are almost the least educated and have the most priors. However, they also have the fewest rules violations of all four clusters, which likely contributes to lower time served that I would have anticipated. 31% of those in cluster 4 recidivated.

(9) Please write a reflection on your cluster modeling experiences.

- This was probably my favorite assignment yet. The different cluster analyses felt really intuitive and applicable. It feels like doing a clustering analysis requires a level of familiarity with the data that a lot of other modeling exercises do not require, especially in supervised learning contexts. Since there is no automated way to select the number of clusters like we would be able to do with variables in a regression, there is a level of discretion that goes

into it on the part of the analyst. Doing a really thorough EDA on all the datasets before diving into the clustering helped when it was time to determine the number of clusters to use. The sums of squares charts were also useful guides and good for citing on the reasoning for making certain decisions, but it felt like a decision that I had to come to holistically instead of just relying on one metric.

The other thing that I liked about doing this analysis was how interpretable the results were. It wasn't difficult to trace back the values of countries and states in any given cluster and compare their values across different variables to see why certain clusters formed together. It felt really informative to see in the European Employment dataset that the EU and EFTA countries were very similar to each other, while the Eastern bloc countries had very different economies. And that was all really highlighted by the clustering results.