Hierarchical and k-Means Cluster Analyses

Serra Uzun, MSDS_411 FALL 2020
11/02/2020

Introduction

The similarities and differences between the collection of observations within a dataset give us an essential outlook on the patterns and common characteristics we may see. We will experiment with Hierarchical and k-Means Clustering for the Cluster Analysis study, performed both on the European Employment dataset. Hierarchical Clustering is an excellent method to see the overall big picture, identify possible clusters that we may see and subgroups that live within the dataset. On the other hand, k-Means Clustering is a useful tool to distinctly identify the groups and visualize with the observations that live in them. We are hoping to conduct both Hierarchical and k-Means Clustering through our analysis, compare the results as well as plots, and determine the most accurate option.

The Dataset

The European Employment consists of 30 observations; all represent a unique country and 11 variables. Except for the two categorical variables, Country and Group, all of the remaining nine variables present employment in various industry segments reported as a percent of 30 European nations. Aside from 30 unique countries, the numeric variables are grouped per Group variables, which have values such as EU (European Union), EFTA, European Free Trade Association, Eastern (Eastern European Nations), and Other. All list of the variables with their type and description is listed below:

Variable	Type	Description
Country	Categorical	Name of the Country
Group	Categorical	The Group Country falls in
AGR	Numeric	Agriculture Employment Percentage
MIN	Numeric	Mining Employment Percentage
MAN	Numeric	Manufacturing Employment Percentage
PS	Numeric	Power and Water Supply Employment Percentage
CON	Numeric	Construction Employment Percentage
SER	Numeric	Services Employment Percentage
FIN	Numeric	Finance Employment Percentage
SPS	Numeric	Social and Personal Services Employment Percentage
TC	Numeric	Transport and Communications Employment Percentage

Table 1: European Employment Data Variables, Types and Description

Exploratory Data Analysis *(Question #1 & #2)*

(1) Each of the 30 observations has unique country names associated with them, whereas these observations are grouped under four groups presented under the Group variable. There are 12 countries in the EU group, eight countries in the Eastern group, six countries under the EFTA group, and four countries in the Other group. While we will be using the categorical variables in the following steps, we will only be using the numeric variables for our exploratory data analysis plots. Through our numeric variables, we are likely to be able to explore our dataset further. We will start by plotting a pairwise scatter plot. Our dataset's size allows us to visualize the relationship between numeric variables in our European Employment dataset by plotting them against each other.

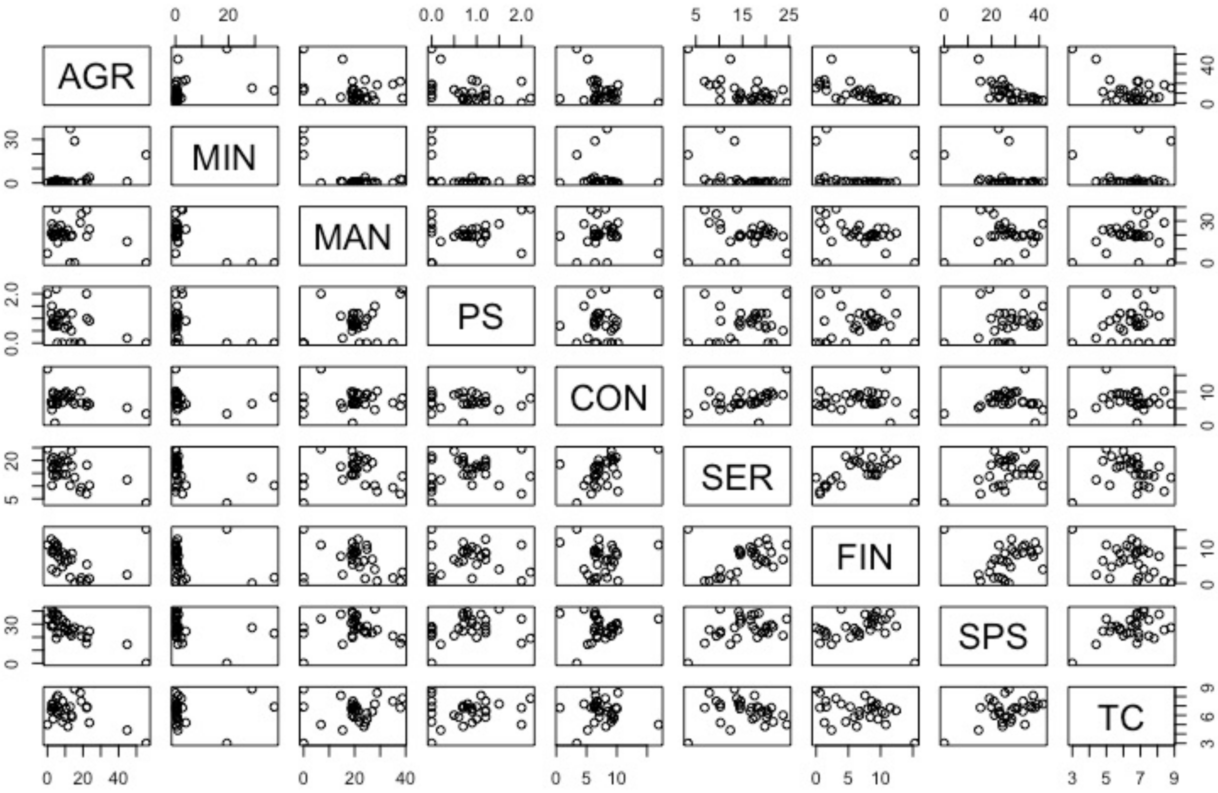


Figure 1: Pairwise Scatter Plot

Through our pairwise scatter plots presented in Figure 1, we do not see any pairwise plots indicating a close association between two variables. As generally, the data points are spread out and do not distinctly follow a path or a pattern, we can't point to a relationship that will be likely to be grouped into a small cluster in our cluster analysis. One variable to point out that gave us interesting scatter plots against all other variables is MIN (mining). We see a linear cluster and pattern opposite to the MIN axis in all scatter plots, which suggests that MIN remains low and insignificant against all variables, except for few outliers; the scatter plots against MIN seem to be dominated by the opposite variable. Another variable to look at would be CON (construction). CON variable plotted against other variables show clustering, generally in the mid-portion of the plot area. All these plots mentioned seem to have scattered yet have two or three clusters. Finally, explicitly focusing on the individual plots, FIN vs. SER plot gives us a relatively clearly visible two clusters. Aside from that specific plot, we can see that MAN plotted against PS, CON, SER, FIN, SPS, and TC also show a large central cluster with smaller, more scatter clusters on the sides.

(2a & 2b) To get a more granular look at some of the variables in the dataset, we plot FIN (Finance) vs. SER (Services) and MAN (Manufacturing) vs. SER (Services) separately on a 2D scatter plot, which is presented below in Figure 2 and Figure 3.

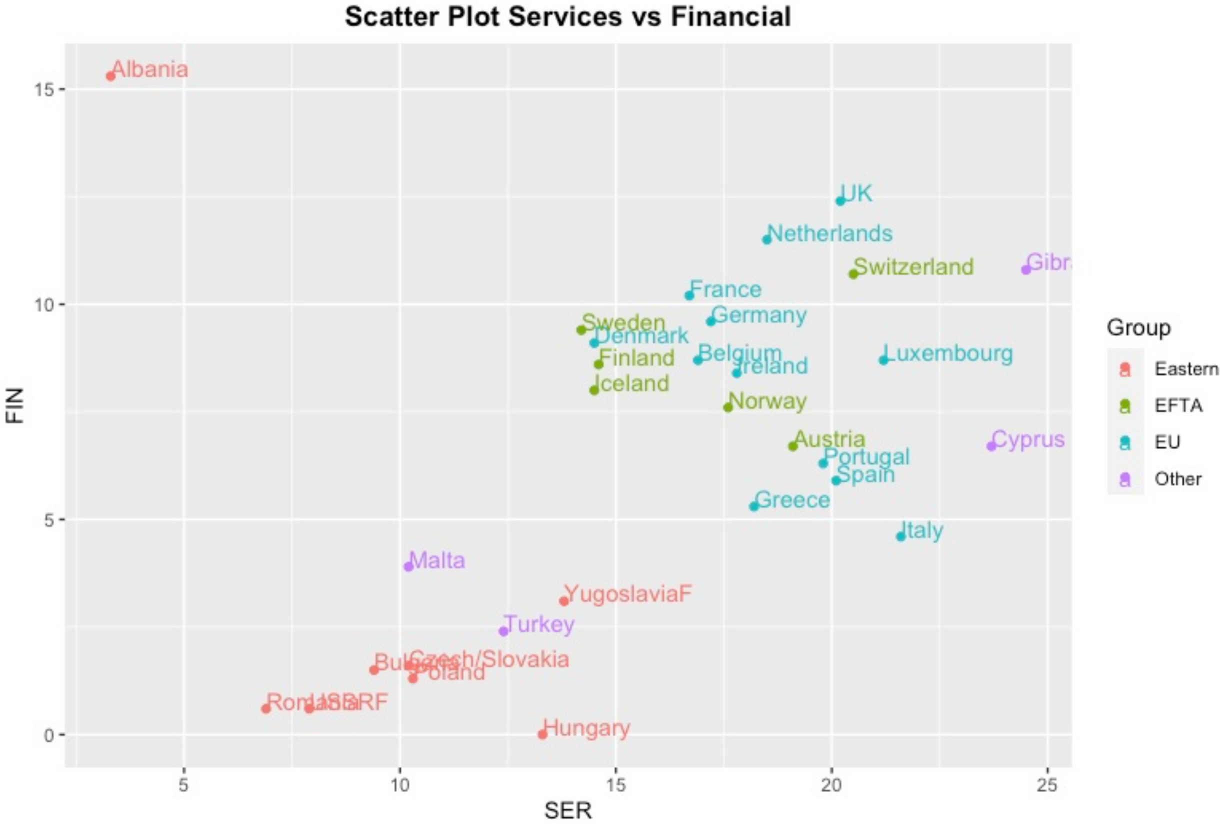


Figure 2: Scatter Plot FIN vs. SER

When we plot FIN against SER, the scatter plot presents us with data points in two clusters. The scatter plot that has the country names printed as labels show that the countries that are located on the eastern part of Europe make the smaller cluster that is on the left lower portion, whereas the relatively western European countries make the larger cluster that is on the upper portion of our plot. The countries in the 'Other' group are in both clusters, while country Albania, which is in the Eastern group, is apart from both clusters. This plot suggests that the eastern European countries are mainly in the range of 5-15 on the SER axis and 0-5 on the FIN axis. On the other hand, EU and EFTA countries are mainly in the range of 14-22 on the SER axis and 4-12 on the FIN axis. Overall Eastern European countries and half of the Other European Countries are lower in the SER vs. FIN plot than EU, EFTA, and other half of the 'Other' countries.

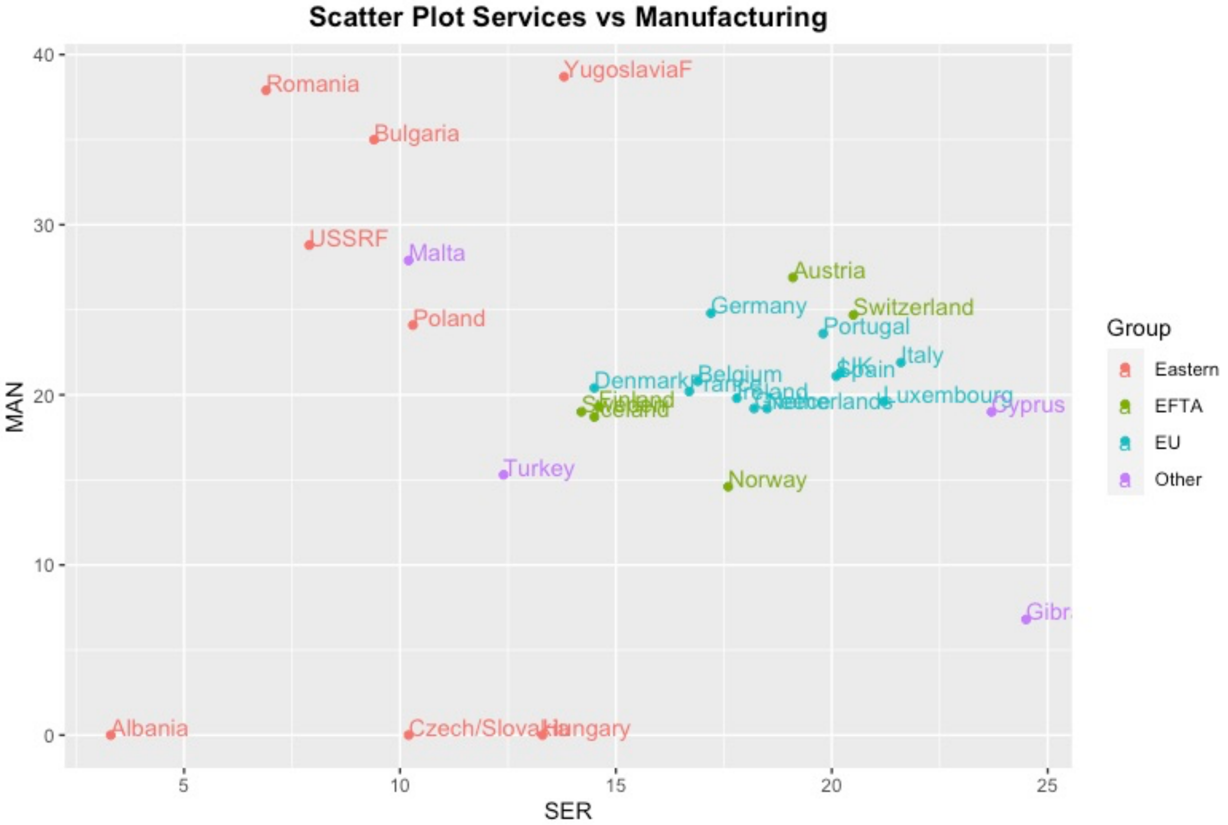


Figure 3: Scatter Plot MAN vs. SER

MAN vs. SER scatter plots show a different spread and clustering compared to FIN vs. SER presented and discussed above. From the Figure 3 plot, we see that EFTA and EU countries are forming a cluster between 14-22 on the SER axis, similar to and as expected from the FIN vs. SER plot, yet the Eastern and Other European countries are much more spread out and not forming any cluster(s). This shows us that while EU and EFTA countries show similar characteristics with each other based on MAN vs. SER, Eastern and Other European countries show rather various features than each other, having their data points more randomly laid on the scatter plot.

The two scatter plots presented and examined above, MAN vs. SER plot in Figure 2, are the better view for supervised clustering where the clustering algorithm would create a classifier that will assign the countries to the correct class/labels. FIN vs. SER plot shows us that the clusters are better and more correctly defined in the MAN vs. SER plot. Therefore, we can assume that FIN vs. SER does not require more supervision, yet MAN vs. SER could use supervision in assigning countries to correct class/label due to some groups and countries not showing obvious patterns and clusters that we would have expected to see.

Principal Component Analysis *(Question #3)*
(3a) Principal Component Analysis (PCA) conducted on European Employment Dataset will help us reduce the dimension from 9D (the nine numeric variables) to 2D, PC1, and PC2. Firstly, we conduct the PCA using the raw European Employment dataset then plot PC1 against PC2 to get a general idea of how the data points are spread within the scatter plot with components that hold the most amount of information.

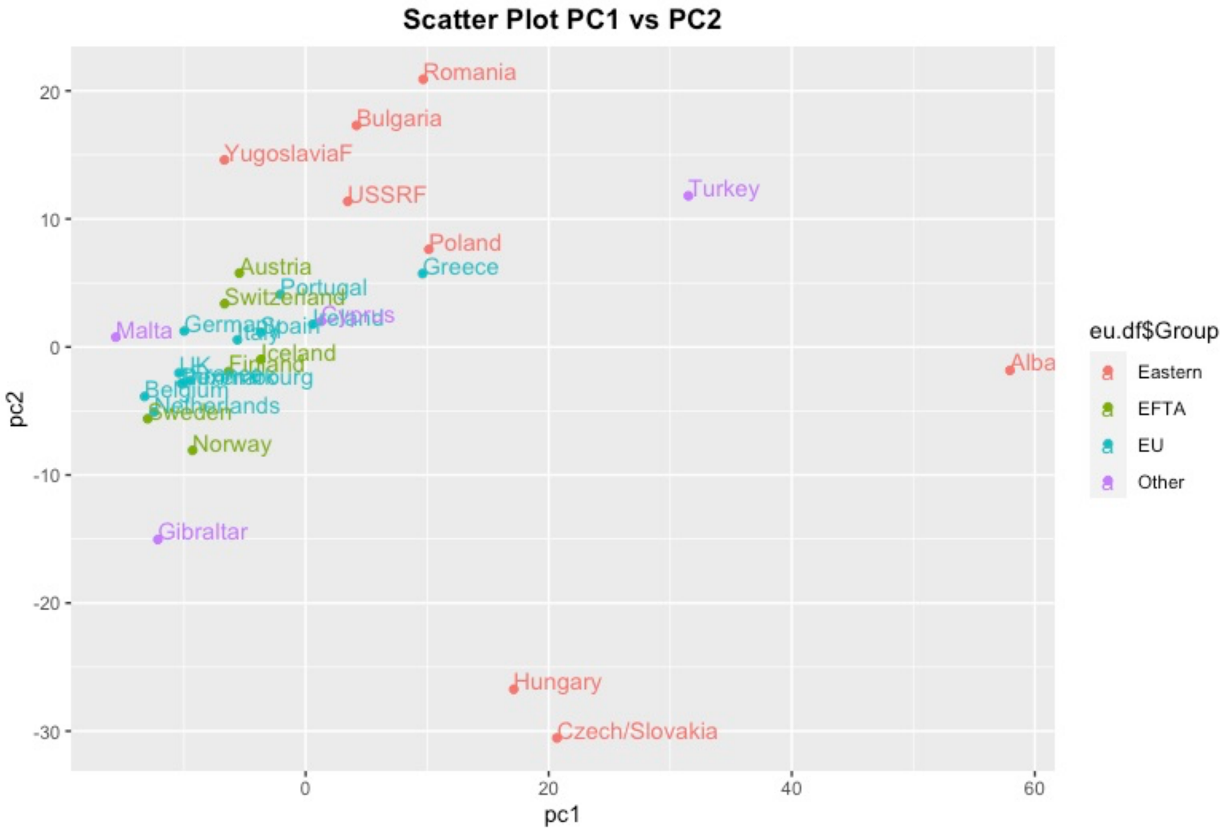


Figure 4: PC1 Scores vs. PC2 Scores

The PC1 vs. PC2 scatter plot suggests that we should expect to see the data on EU, EFTA, and the majority of the 'Other' European countries similar to each other. These groups are in the range of -20 – 10 PC1 and -15 – 10 PC2. While PC1 vs. PC2 presents a rather distinct cluster of EFTA, EU, and almost all 'Other' European Countries, The Eastern European countries are shown to be scattered in a more spread out way that the other groups were. The Eastern European countries do not show common characteristics in the PC1 vs. PC2 context and are located in the lower and higher and of axes of both PC1 and PC2.

When we standardize the EU Employment dataset using the scale() function in R, we see a change in our PC1 vs. PC2 plot yet no significant improvements on the results. The cluster formed by the EFTA and EU countries remains. In contrast, Other European countries are now more spread out, and the Eastern European countries are relatively closer to each other than they were with the non-standardized data was used for the PCA.

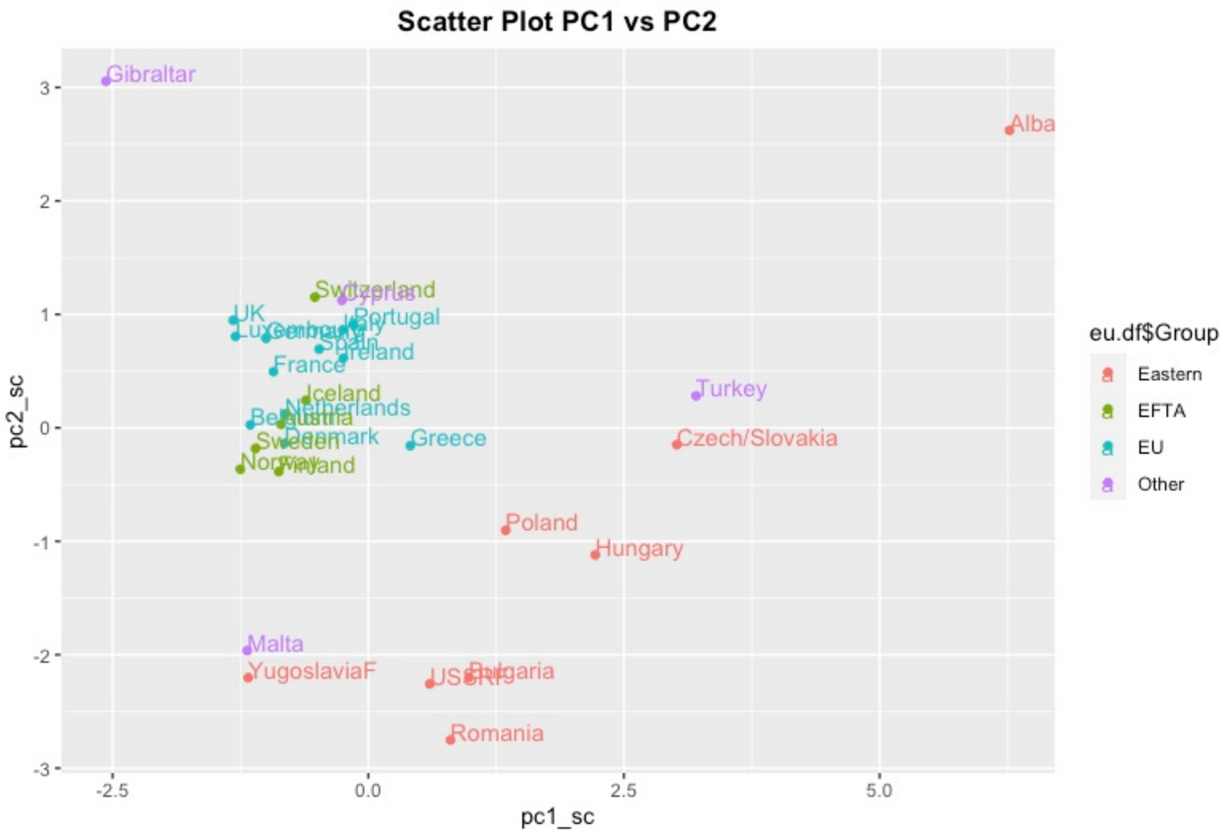


Figure 5: PC1 Scores vs. PC2 Scores with Standardize Dataset

Plots presented in Figure 4 and Figure 5 suggest that while standardizing did affect our PC1 vs. PC2 yet not at a level that significantly impacted the outcome or shows any significant improvements in our results. The positive effect of standardizing our dataset is seen in regard to the increased proximity of some of the previously spread points. Contrary to this improvement, we see that there are data points in our plot that are even further out and isolated than before. We can see that Eastern European countries are not densely and closely clustered through both plots, and Albania is outside in both approaches. The reason for not seeing a significant impact on the dataset due to standardization is our dataset's nature, which consists of 'percentage' value, meaning they all range between 0 and 100 even though they all continuous variables.

Hierarchical Clustering Analysis *(Question #4)*

(4a) The hierarchical clustering analysis will generate a dendrogram, which gives us the tree of clusters where clusters and variables in them are hierarchically distributed down. At the end of these branches of the tree, like plot the dendrogram will give us, we will have the chance to see the variables with similarities branching down from the same tree. The dendrogram plot provides us with all sizes of clusters and gives us a broader view of the tree of similarities within the dataset.

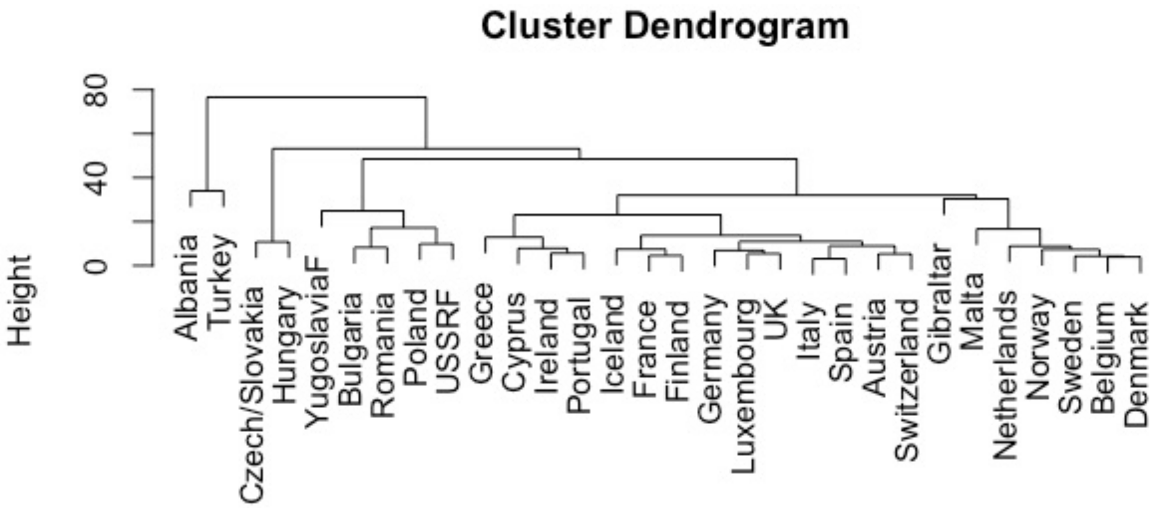


Figure 6: Complete Type Hierarchical Clustering Dendrogram Plot

The dendrogram plot of the EU Employment dataset shows about five main clusters that then sub-branch to smaller clusters within the same group. For example, if we look at the third cluster that includes Yugoslavia, Bulgaria, Romania, Poland and USSRF, we can see that even though all these countries are in the same cluster, there are multiple sub-clusters that are made of 1-2 countries. So hierarchical clustering and dendrogram plot can provide us with an overall as well as a granular level clustering of the dataset. We will conduct two hierarchical clustering models to experiment with the hierarchical clustering analysis, one with three clusters and another with 6 clusters. The total sum of squares (TSS) is the same for both k=3 and k=6; thus, we will compare these two models using the Between Cluster Sum of Squares (BCSS). We are looking to get a high BCSS that indicates the distance between clusters, and a higher BCSS essentially suggests a more distinct and well-defined cluster.

	Between Cluster Sum of Squares Percentage
Hierarchical Cluster Analysis with k=3	0.5893
Hierarchical Cluster Analysis with k=6	0.8421

Table 2: Complete Type Hierarchical Clustering Dendrogram Plot

The BCSS presented in Table 2 for both iterations of the hierarchical model k=3 and k=6 suggest that the model with six clusters instead of three has a higher distance/difference between its clusters. We can conclude that six hierarchical cluster models show less overlapping characteristic similarities between variables and can group variables with similar characteristics more distinctly. As more defined clusters that are visible distant from each other are what we aim to get, we can conclude that the six-cluster model is a better model than the three cluster models.

Next, we will use the PC1 and PC2 that we generated in the Principal Component Analysis previously for the hierarchical clustering model. Once again, try a cut tree of three and six, which we will then compare with the results we obtained with our raw dataset. Below in Figure 7 is the dendrogram of the hierarchical model conducted with PC1 and PC2.

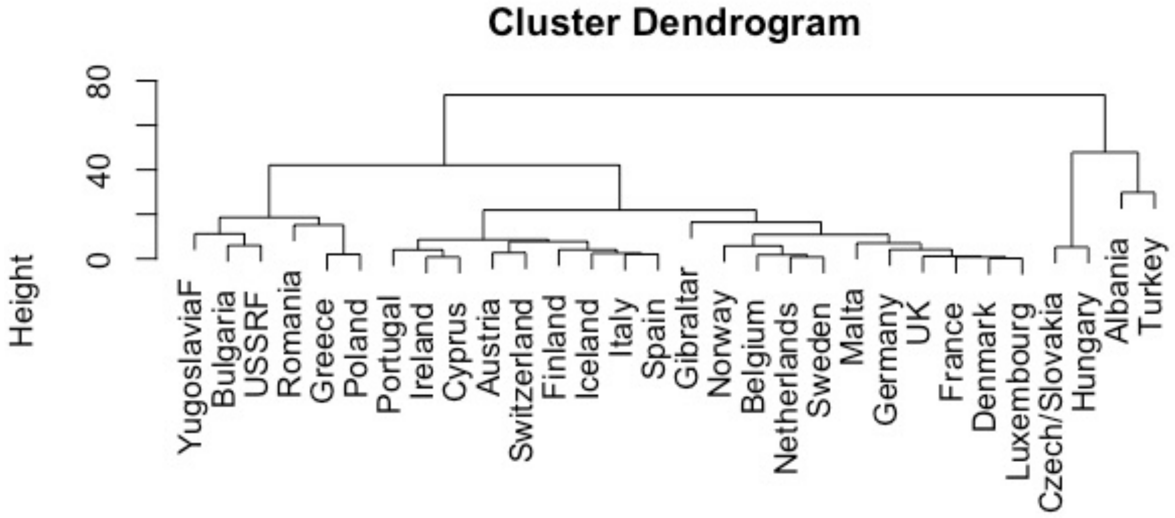


Figure 7: Complete Type Hierarchical Clustering Dendrogram Plot with PCA Components

Firstly, we can visibly see a positive impact of using PC1 and PC2 over the raw dataset on our dendrogram exhibited in Figure 7. The clusters are shown to be not only denser and more organized but more precisely separated. We can see four main clusters through the plot, and as the tree reaches the sub-clusters, we can see an even branching within almost all clusters.

	Between Cluster Sum of Squares Percentage
Hierarchical Cluster Analysis with k=3	0.5893
Hierarchical Cluster Analysis with k=6	0.8421
Hierarchical Cluster Analysis with PC1 and PC2 and with k=3	0.7373
Hierarchical Cluster Analysis with PC1 and PC2 and with k=6	0.9426

Table 3: Comparison of HC Analyses

The hierarchical clustering model results improvement has also been reflected in the BCSS. As we compare the BCSS of all the HC models we ran so far. We can see that the highest BCSS results are obtained through the HC model conducted with PC1 and PC2 and six clusters. Per this result, we can determine that while the dendrogram plot showed four clusters, we can increase the distance between clusters even further and generate better results through six clusters.

k-Means Clustering Analysis *(Question #5)*

(5a & 5b) Following the Hierarchical Clustering Analysis another crucial clustering analysis, k-Means clustering is performed on the European Employment dataset. The k-Means model will identify clusters and centroids within those clusters to group data points accordingly. Alike Hierarchical Clustering, this method aims to identify observations with similar features and are likely to fall under same subcategory. Below is the table with BCSS results of k-Means clustering models with k=3 and k=6, as well as the hierarchical model results for comparison.

	Between Cluster Sum of Squares Percentage
Hierarchical Cluster Analysis with k=3	0.5893
Hierarchical Cluster Analysis with k=6	0.8421
Hierarchical Cluster Analysis with PC1 and PC2 and with k=3	0.7373
Hierarchical Cluster Analysis with PC1 and PC2 and with k=6	0.9426
k-Means Clustering Analysis with k=3	0.5793
k-Means Clustering Analysis with k=6	0.8342

Table 4: Comparison of Hierarchical and k-Means Clustering Model BCSS

Table 4 shows us that k-Means clustering that we conducted with 3 and 6 clusters, both didn't prove to have better results of our hierarchical clustering model results presented in the previous section. Both BCSS models with k=3 and k=6 have dropped 1% between Hierarchical clustering and k-Means clustering, and while this is not a significant difference, it indicates that k-means clustering method using raw European Employment data didn't present itself to be the better option.

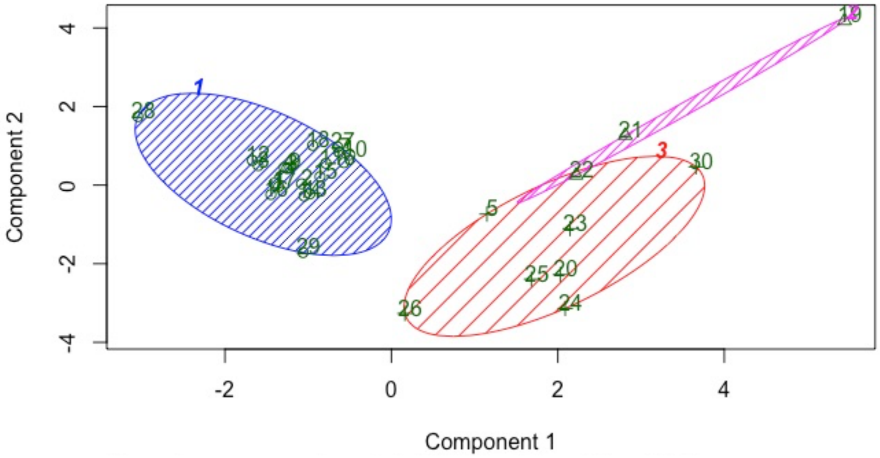


Figure 8: k-Means Clustering Plot with k=3

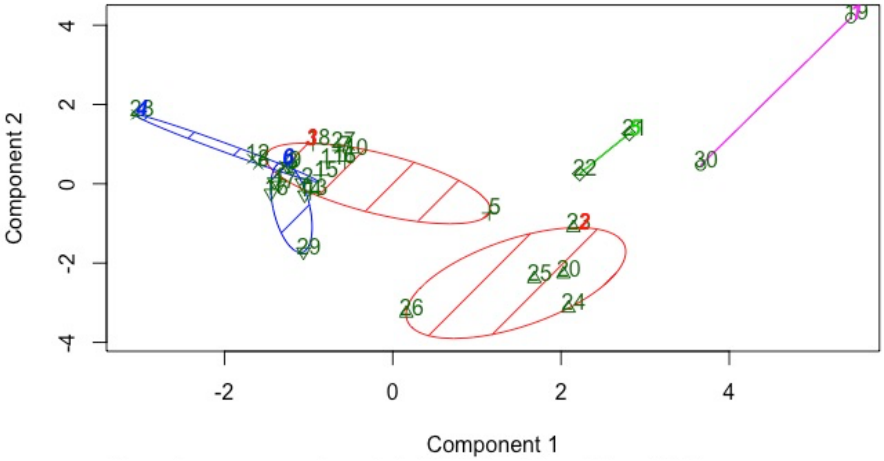


Figure 9: k-Means Clustering Plot with k=6

In both of the above plots in Figure 8 and Figure 9, we see the similar patterns and clustering we have seen in in all our clustering plots for European Employment dataset. The k-Means clustering model and plots show that the EU and EFTA countries are clustered together, while Eastern and Other European countries are not densely clustered, the data points are not as close to each other and overall much more spread out. The plots show parallel results as our BCSS percentage presented in Table 4, suggesting a better performing k-Means clustering model with six clusters compared to three clusters.

(5c & 5d & 5e) Similar to the step we have taken with hierarchical clustering analysis, we will use the PC1 and PC2 as our variables and conduct the k-Means clustering once again. We aim and expect that using principal components will improve BCSS and give us better defined clusters than the other models.

	Between Cluster Sum of Squares Percentage
Hierarchical Cluster Analysis with k=3	0.5893
Hierarchical Cluster Analysis with k=6	0.8421
Hierarchical Cluster Analysis with PC1 and PC2 and with k=3	0.7373
Hierarchical Cluster Analysis with PC1 and PC2 and with k=6	0.9426
k-Means Clustering Analysis with k=3	0.5793
k-Means Clustering Analysis with k=6	0.8342
k-Means Clustering Analysis with PC1 and PC2 with k=3	0.6872
k-Means Clustering Analysis with PC1 and PC2 with k=6	0.9018

Table 5: Comparison of All Hierarchical and k-Means Clustering Model BCSS

(5d & 5e) Through all the clustering models we conducted so far where we started our model with 3 clusters and then raised the number of clusters to six, we saw that increasing the count of clusters showed an improvement in BCSS results in all the models. As better-defined clusters are likely to give us more accurate results, we can conclude that in case of the European Employment dataset, 6 clusters, and specifically when principal components are used, are better than 3 when running cluster analysis of any kind.

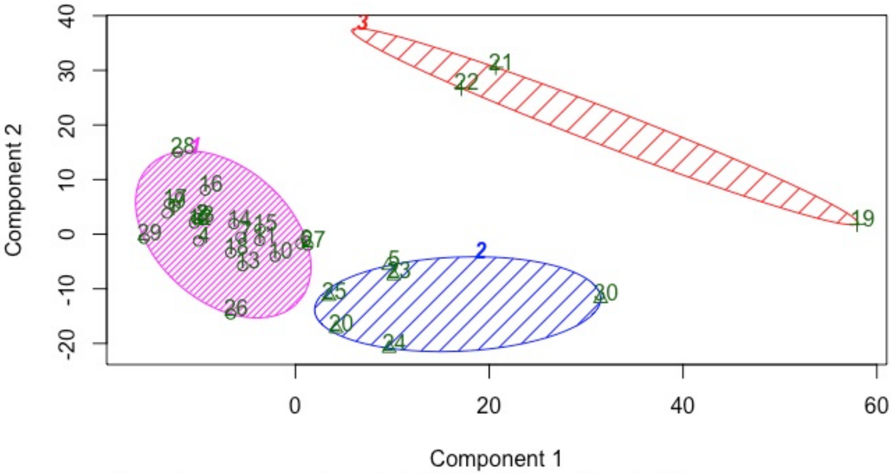


Figure 10: k-Means Clustering Plot with PC1 and PC2 and with k=3

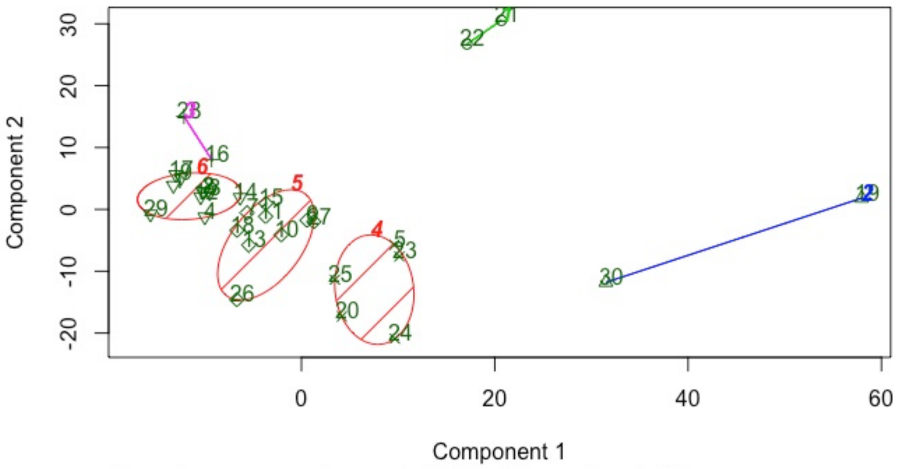
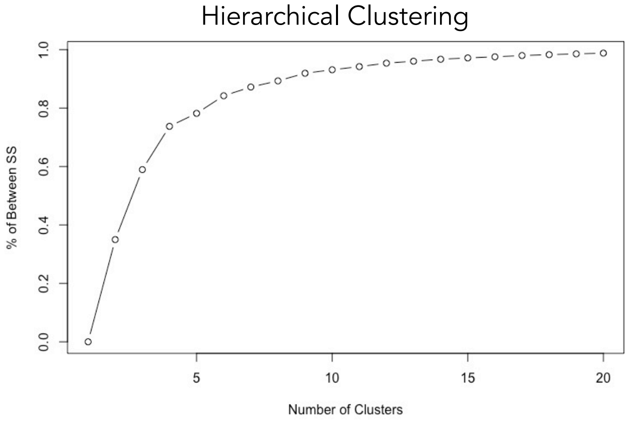


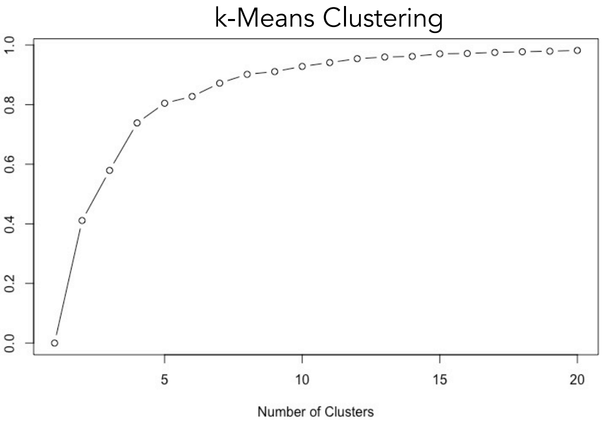
Figure 11: k-Means Clustering Plot with PC1 and PC2 and with k=6

Finally, when we plot the k-Means clusters with PC1 and PC2 with number of clusters set as 3 and 6, we see a noticeable improvement between k=3 and k=6, which was also indicated through the BCSS results. The plot continuously suggests clusters of EU and EFTA countries, and other, less dense clusters of Eastern and Other European countries. If we observe our best performing k-means model plot in Figure 1, we can conclude that EU and EFTA countries are very likely to fall in the same clusters, data points proximate to each other, whereas Eastern and Other European countries are likely to have characteristics that are not quite similar to each other and other European countries.

Optimal Number of Clusters by Brute Force *(Question #6)*

Although we have tested k=3 and k=6 through all the clustering models we conducted earlier, in real-life situations selecting the number of clusters to use in these models are not so simple. The below plots that provide us with the BCSS percentage that both hierarchical and k-Means clustering models would obtain using various number of clusters.





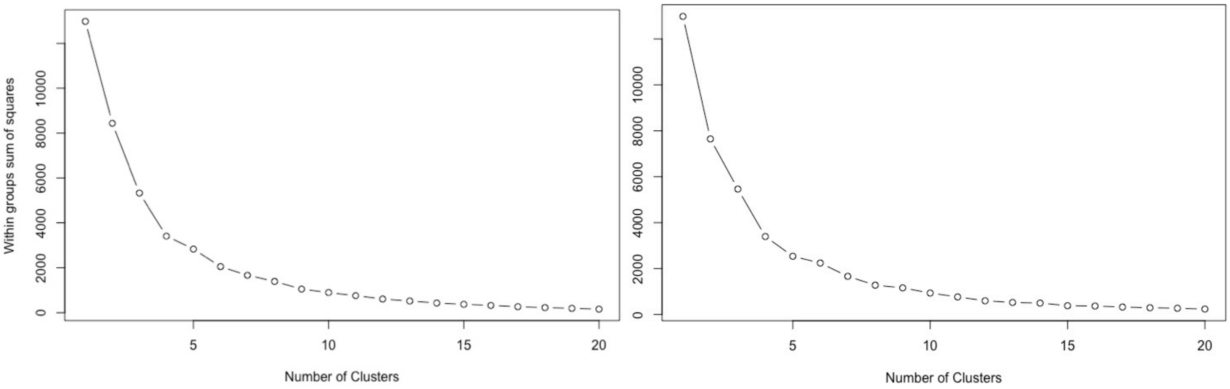


Figure 12: BCSS and WSS with Various Number of Clusters

Presented above is the Between Cluster Sum of Squares (BCSS) and Within Cluster Sum of Squares (WSS) results per each iteration of cluster count, from k=1 to k=20. These graphs show us that with k=3 we remained below 0.8 in BCSS and above 4000 in WSS, yet both these results have improved when we increased the cluster count from 3 to 6. If we look further, we can see that the BCSS and WSS continue to get better as we increase the number of clusters, and with about 10 clusters we would be able to achieve 0.9 BCSS and below 1000 WSS. So, if we were to take this study further, and were asked to pick the optimal number of clusters we would definitely assess k=10 for a better performing model and more accurate outputs.

Hierarchical Clustering Analysis with US States Data *(Question #7)*

The USSTATES dataset consists of 12 variables (two character and 10 numeric) and 50 observations. The data holds the information on state-wide average or proportion scores for the non-demographic variables. For each observation the higher the score, the more of that quality of that variable is present for the specific row/observation. The list of the variables are as follows: State (chr), Region (chr), Population (num), Household Income (num), High School (num), College (num), Smokers (num), Physical Activity (num), Obese (num), Non-White (num), Heavy Drinkers (num), Two Parents (num), Insured (num).

As the initial step of the Hierarchical Clustering exercise, we will start by generating a dendrogram plot for 'complete' clustering to visually see the layout and distribution of data by state. Figure 13 below shows the dendrogram plot of the complete hierarchical clustering of USSTATES data.

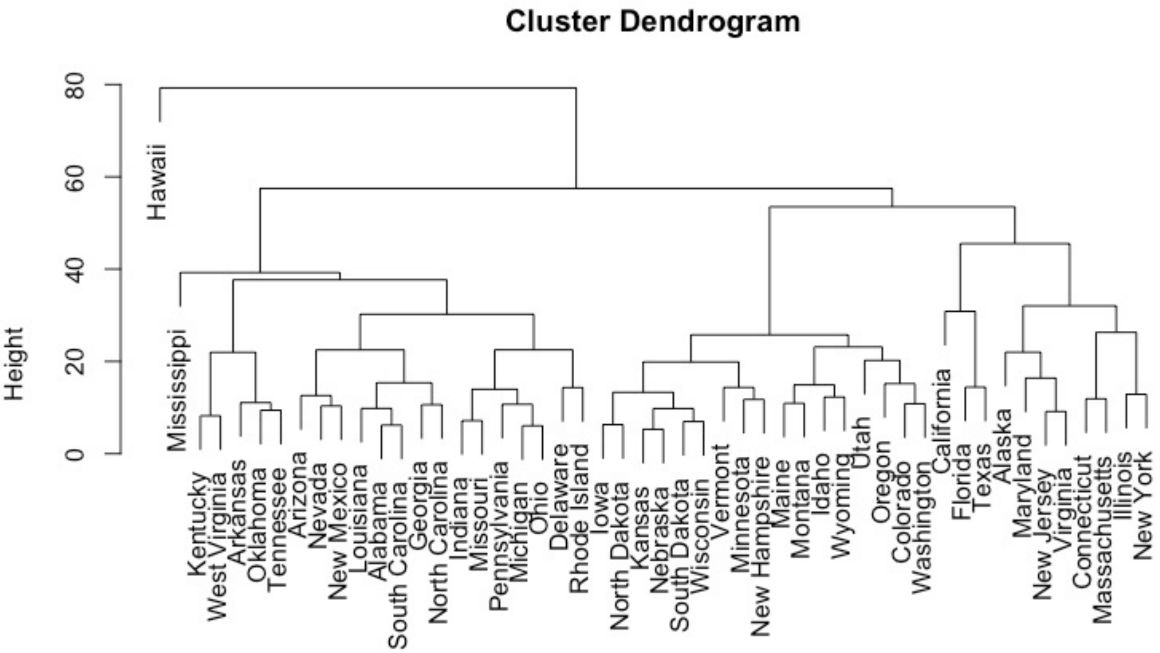


Figure 13: Complete Type Hierarchical Clustering Dendrogram Plot of USSTATES data

Overall when we look at the dendrogram we can see 8 clusters in the USSTATES dataset. Another characteristic we see through dendrogram is how each cluster holds states that are either geographically close to one another or have similar characteristics such as climate, lifestyle, infrastructure, etc. While at first look we are able to determine the clusters within the USSTATES dataset as 8, we need to conduct a classification accuracy plot to see the optimal number of clusters.

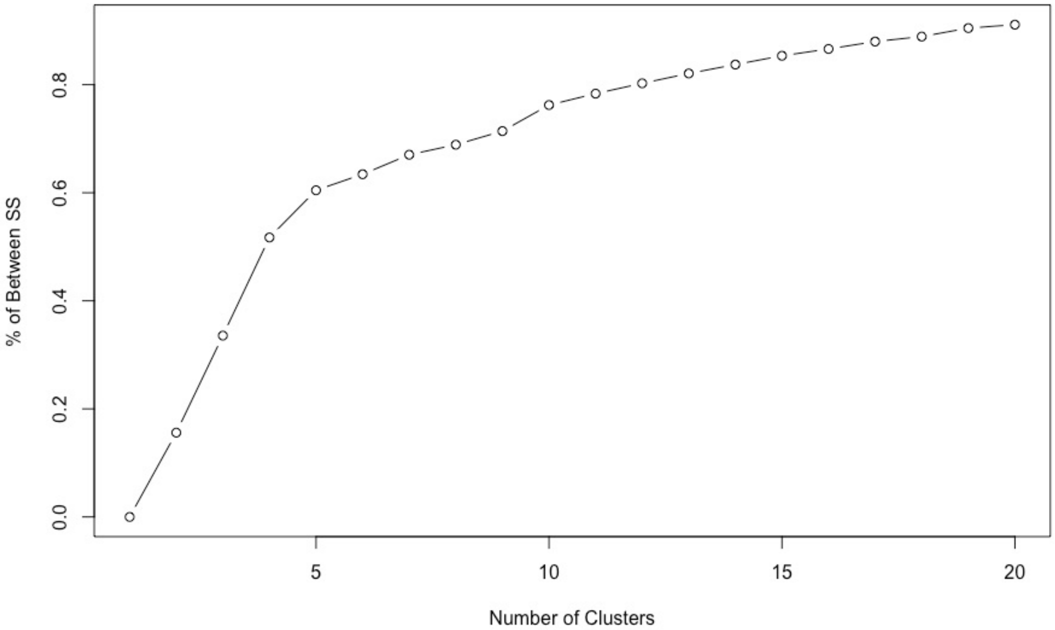


Figure 14: Optimal Number of Clusters with USSTATES data

Per Figure 14, we can see that at around 13 clusters, we are able to achieve BCSS above 80%. This result suggests that the 50 observations in the USSTATES dataset are rather varied and are not distinctly separated from each other through their similarities and/or differences. While the dendrogram plot presented in Figure 14 showed about 8 clusters, the optimal clusters plot in Figure 14 suggests that a more granular approach with 13 clusters would give us a much more accurate model.

k-Means Clustering Analysis with Recidivism Data *(Question #8)*
Recidivism dataset consists of random sample records of convicts released from prison between the years of 1977 and 1978. The dataset has 18 variables, all numeric, and 1445 observations in total. We will start our k-Means clustering analysis by plotting the classification accuracy plot that will give us an idea on how many clusters are likely to be the optimal for our Recidivism dataset.

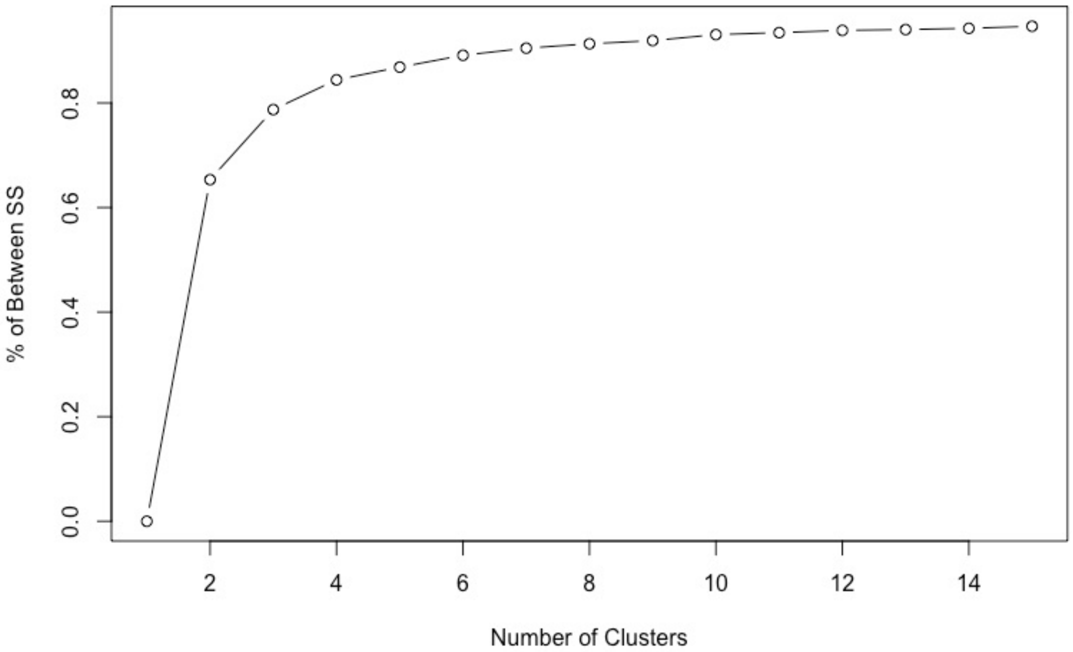


Figure 15: Optimal Number of Clusters with Recidivism Data

Per Figure 15, we see that by 4 clusters we pass beyond 80% accuracy, where it starts to slowly and incrementally towards 90%+ between 4 and 15 clusters. Our plot shows us that 4 clusters are a good starting point for the k-means clustering analysis of Recidivism data, and that we can take it potentially to about 9-10 clusters to obtain a decent 90%+ accuracy if needed.

Figure 13 below presents us with the Recidivism data in 4 clusters. Overall the clusters are vertically separated in a quite distinct pattern. The division between clustering is only seen throughout PC1 (x-axis) with cluster #2 being the densest, then followed by cluster #3, cluster #4 and finally cluster #1, which is shown to be the most spread out. The separation between the clusters occur around -50 PC1 for clusters 2 and 3, around 75 PC1 for cluster 3 and 4 and around 200 PC1 for cluster 4 and 1.

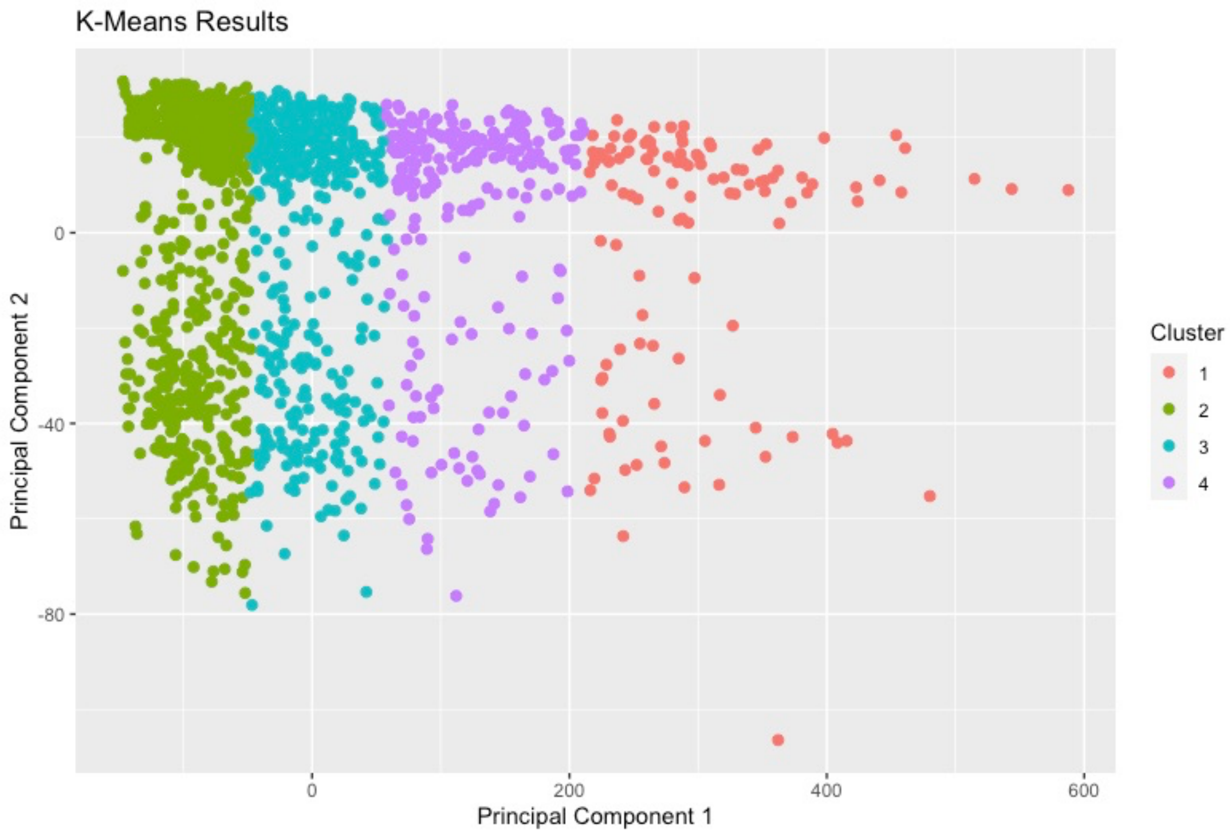


Figure 13: k-Means Clustering Plot with k=4

When we divide BSS (between sum of squares) to TSS (total sum of squares) we obtain an 84.4% accuracy. The sizes of our clusters are also as we predicted them. Cluster #2, which we identified as the densest has 671 data points, followed by cluster #3 with 430, cluster #4 with 228 and finally cluster #1 with only 116 data points. The mean of each variable within each clusters is presented in the Table 6 below.

	black	alcohol	drugs	super	married	felon	workprg	property	person	priors
1	0.353	0.371	0.276	0.621	0.422	0.241	0.397	0.138	0.112	4.293
2	0.466	0.124	0.228	0.705	0.124	0.291	0.334	0.265	0.022	0.484
3	0.553	0.219	0.221	0.698	0.344	0.388	0.600	0.307	0.067	1.481
4	0.478	0.364	0.303	0.693	0.390	0.281	0.632	0.184	0.088	2.671
	educ	rules	age	tserved	follow	durat	cens			
1	7.043	0.388	646.966	19.241	74.474	58.638	0.690			
2	9.899	1.484	255.529	17.463	74.925	52.241	0.565			
3	10.330	1.226	340.170	22.065	75.016	56.698	0.637			
4	9.294	0.636	466.553	18.772	74.732	60.425	0.702			

Table 6: Mean of Each Variable in Every Individual Cluster

Upon reviewing the above table, we can have an idea on the mean values within each cluster for each variable. For example, if we look at the 'priors' variable in Table 6, we can see that a large value within 'priors' is very much likely to be in Cluster #1, which as we mentioned before, is the least dense cluster that falls in the area 200+ in PC1. Per these results we can predict in which cluster any new data may fall through looking at its similarities with the means that we obtained through our k-means clustering analysis on Recidivism dataset.

Reflection *(Question #9)*
Going through both Hierarchical and k-Means clustering in a single module has been demanding, challenging, and rewarding. Even though neither of the datasets we experiment with was extensive, it was interesting to see how clustering unsupervised learning methods helps us visualize similarities within the dataset. After completing this week's tasks, I can state that I will be starting future cluster analysis with the 'Optimal Number of Clusters' plot to overview the effect number of clusters have on model performance. I found the plot informative and helpful in drawing a course plan for the next steps. One final note would be on plotting. Personally, I found it quite challenging to plot k-Means clusters with labels. While with clusplot I was able to get a good enough plot that allowed for interpretation, I wanted to have a more audience-friendly, relatable plots, which I was unable to obtain.