

Module #2 – SPSS Output

SA8903 – Applied Spatial Statistics

Shakeeb Tahir

Dr. Brian Ceh

Introduction

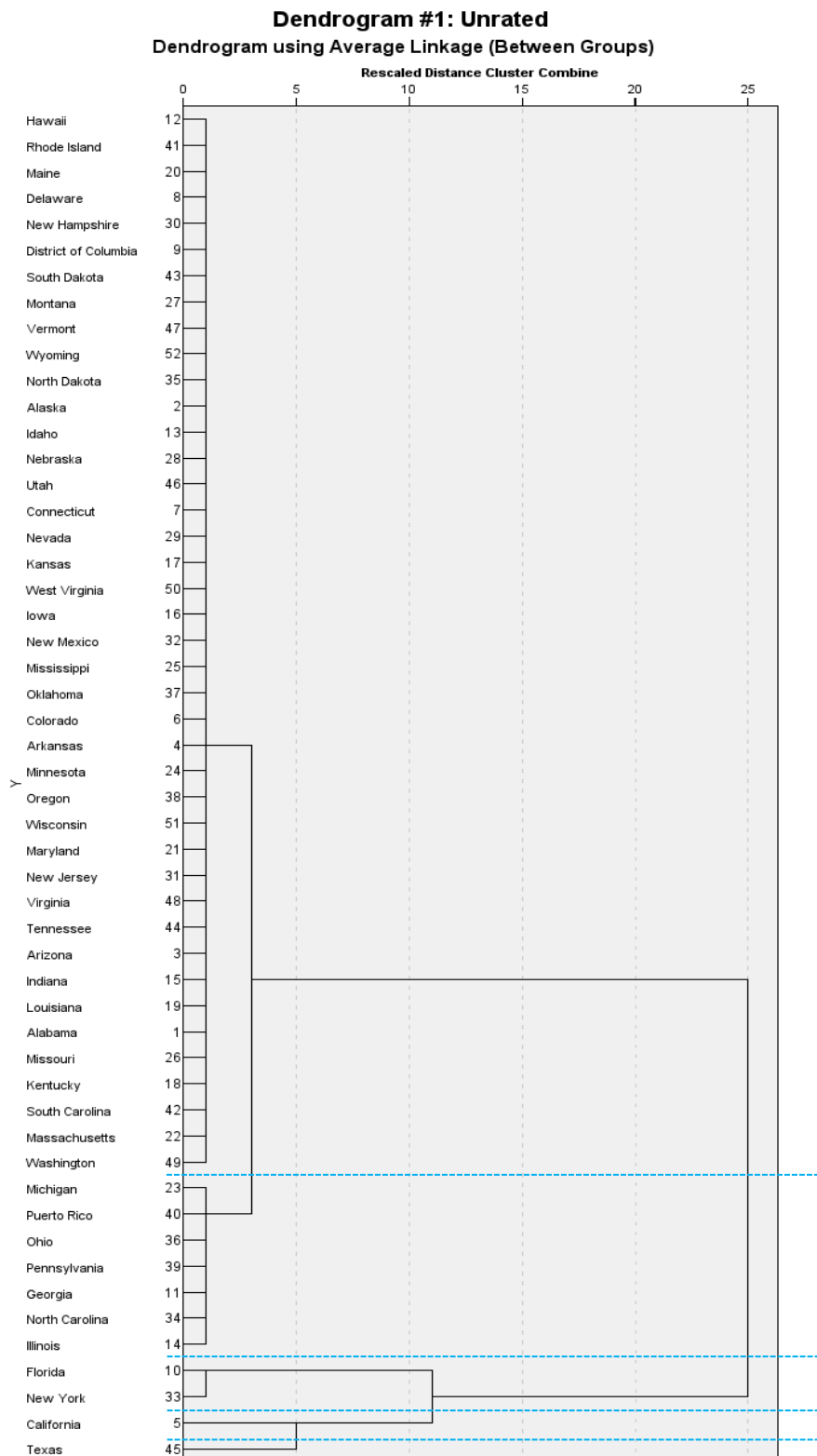
Cluster analysis is a multivariate procedure for detecting groupings within data sets. In this report, we will be comparing the results of a hierarchical and k-means clustering analysis. Hierarchical clustering finds the closest pair of objects according to a distance measure and combines them to form a cluster. This analysis gives us a dendrogram, which is a tree-like structure where the leaves represent individual data points, and the internal nodes are used to represent clusters of varying sizes. Once two objects or clusters are joined, they remain together until the final step. K-Means clustering divides the data set into a pre-determined number of clusters by continuously assigning data points to the nearest center and updating them. The goal of k-means is to minimize the within cluster variance and provide clusters that are as similar to each other as possible. The variables we will be using in this cluster analysis comparison come from the 2020 U.S Census data and all help to measure levels of poverty, which will be the focus of the analysis. They are the following:

- 1) # Educational Attainment | Less than high school diploma, 2021 [Education2]
- 2) # Employment Status | In Civilian Labor Force, Unemployed, 2021 [Employment5]
- 3) # Poverty Status by Age | In Poverty, 2021 [Poverty2]
- 4) # Computers in Household | No Computer, 2021 [Computer3]
- 5) # Internet Subscriptions In Household | No Internet access, 2021 [Internet4]

These variables were chosen since they all give a measure of the level of poverty within a state and as such will be useful in finding clusters of poverty levels throughout the United States. First, a hierarchical model will be run to see how the groups cluster based on the variables being unrated, then rated (by dividing variable by total population and multiplying by 10000), then z-scored, and finally rated and z-scored. Next, the k-means analysis will be run by repeating the same process for the variables through unrated, rated, z-scored, rated and z-scored based on 4 cluster groups and comparing how the clusters change. A two-step cluster analysis will then be run to determine what is the optimal number of groups to be used for the final model (rated & z-score). A two-step cluster provide us with a cluster quality score which will tell us the quality of the clusters that were formed. Once we have determined the number of clusters to use based on the two-step, we can run the k-mean analysis for the final model (rated & z-score) and compare how it differs or is similar to the final (rated & z-score) hierarchical model.

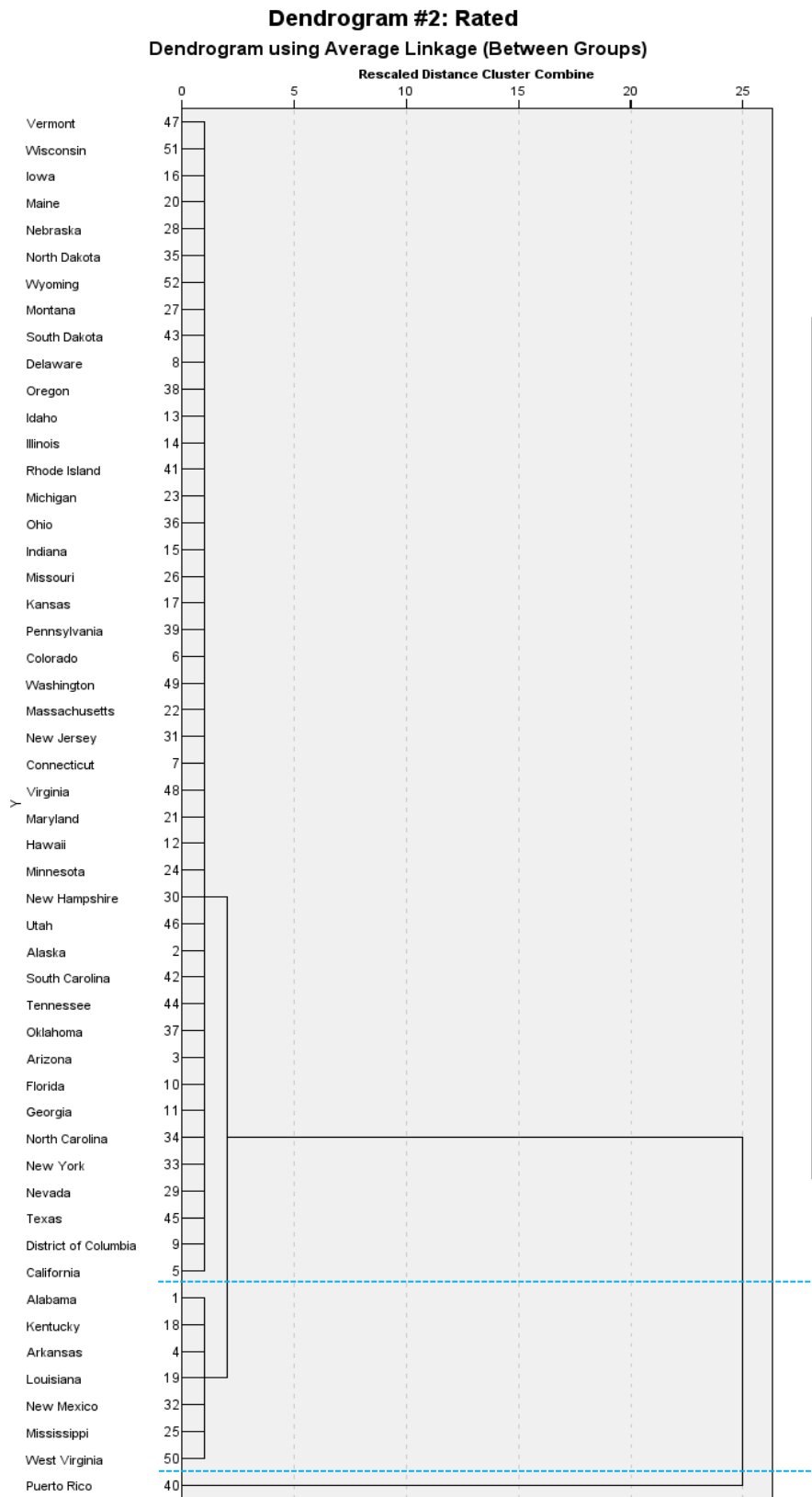
Hierarchical SPSS Output

1. Unrated



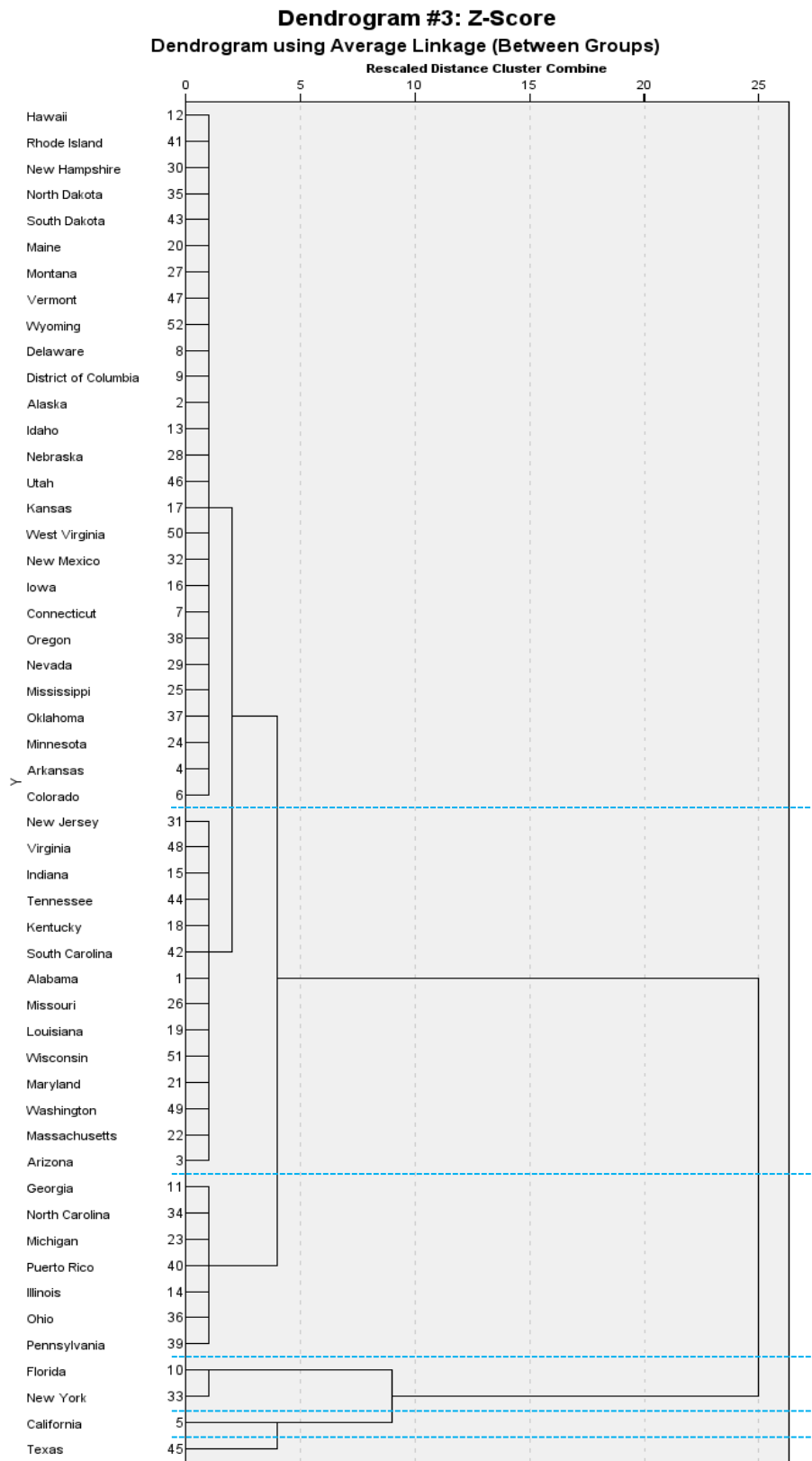
In the dendrogram for the unrated variables we can see that we have 5 cluster groups. These clusters mostly seem to be largely influenced by the size of the state since they we are using unrated variable. We know this because at the bottom of the dendrogram we can see the largest / most populous U.S states such as Texas and California which are grouped on their own and New York and Florida which are grouped together. This is likely occurring since they are the most populous states and as such likely have the most amount of people in poverty which is why the dendrogram separates them from the rest of the clusters. We can also see other populous states such as Pennsylvania, Illinois, Georgia in a group together. These are the secondary level of most populous states (Source: U.S Census Data 2020) after the major states like California, Texas, etc. Since we are using unrated variables, the dendrogram seems to be clustering the groups based on the level of poverty in each state and states with the higher populations will tend to have a higher raw number of people in poverty.

2. Rated



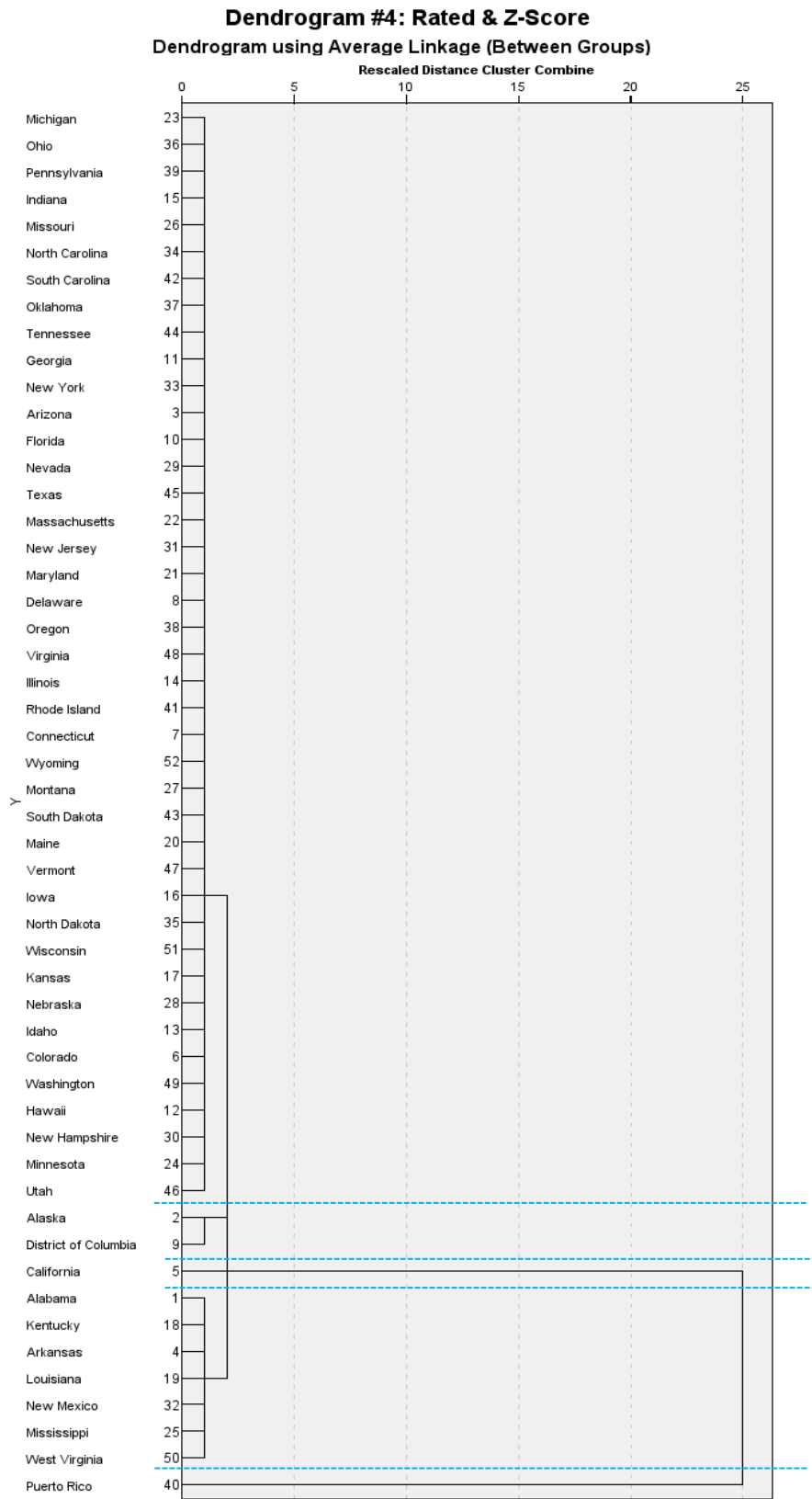
In the dendrogram for the rated variables, we can see that there are now 3 groups and that the previous problem where the unrated dendrogram had the largest states being their own groups since they have higher raw numbers of poverty is also solved. Since the variables are now rated, the level of poverty is measured relative to the number of people in the state. As such, we can see that states which we would expect to have a higher ratio of poverty such as Alabama, Kentucky, New Mexico, and other southern states are now in their own group. We can also see that Puerto Rico is in its own group, this is due to the high level of poverty in the state relative to the population as most people in Puerto Rico live in poverty which is why it is clustered into its own group as it is such an extreme case.

3. Z-Score



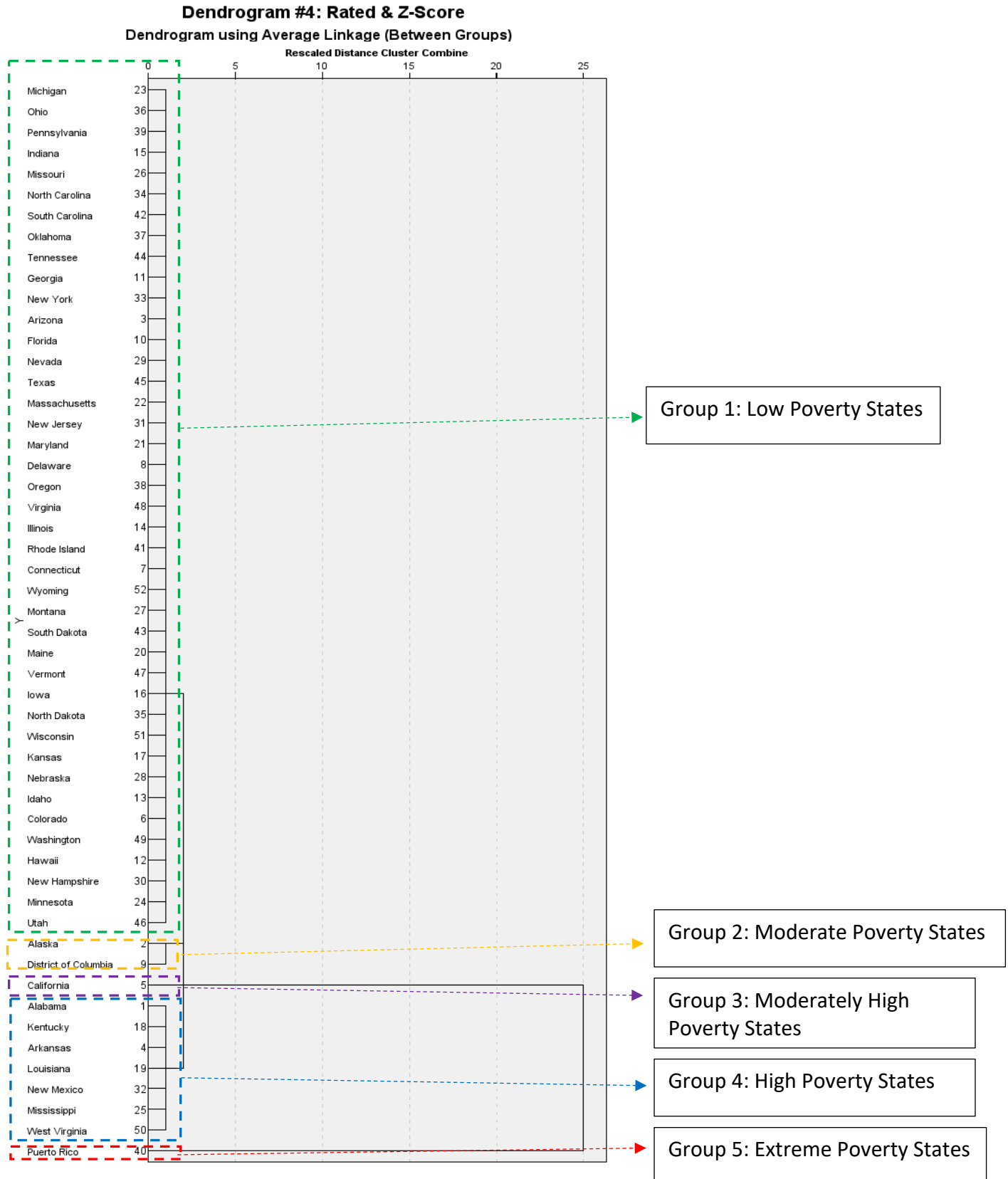
In the dendrogram for the z-score variables, we can see that we now have 6 cluster groups. Once again, we see large populous states such as California and Texas being in their own clusters. This is because they have a high standard deviation away for the average level of poverty compared to the rest of the states. Since these are the most populous states, their level of poverty will be higher than the other states and as such they will have higher standard deviations away from the mean of the data which is why they become separate clusters. We can see how compared to the unrated and rated dendrograms, the clusters are much larger for the other non-major populous states since they all have lower z-scores and are closer to the mean level of poverty of all the states.

4. Rated & Z-Score

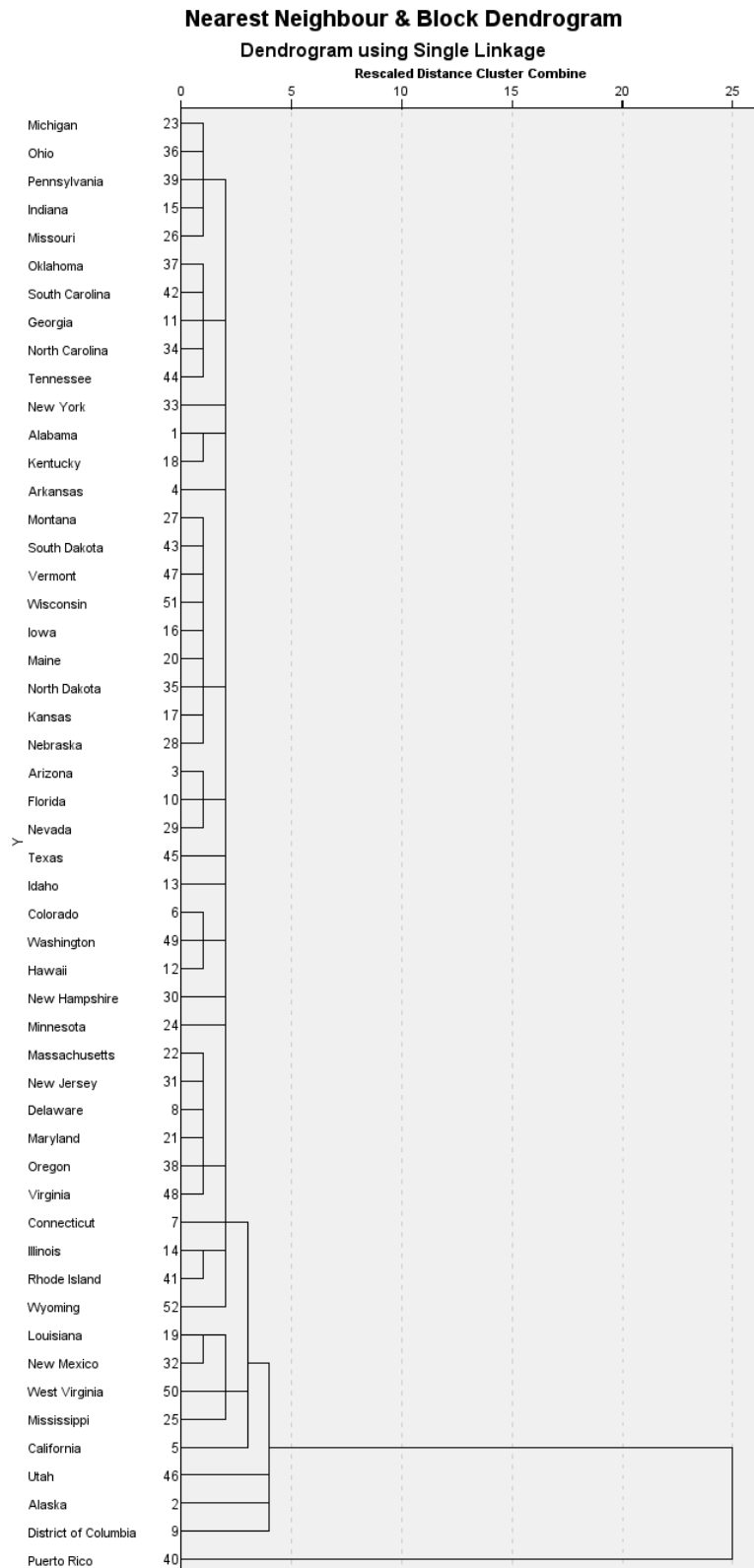


In the dendrogram for the rated & z-scored variables, we can see that there are now 5 cluster groups. Since these are both rated and z-scored, they give us the best idea of the clusters of states and their level of poverty. We can see that Puerto Rico is once again on its own, which makes sense as majority of the people living there are in poverty. Southern states like Alabama, Kentucky, Mississippi which also have high amounts of the population in poverty are in their own cluster which also makes sense. Another interesting case is California, which is grouped on its own, however this also makes sense as California is known to have a high level of poverty and homelessness problem. DC and Alaska are their own cluster, which is understandable as they are both high cost of living states so there is likely a moderate amount people living in poverty within them. The rest of the states (mostly northern) are all grouped, likely because they have a lower ratio of people living in poverty compared to the other states.

Final Model (Rated & Z-Score) Descriptors:



What happens to the Dendrogram when you change Methods to **Nearest Neighbor** and Interval to **Block**?


















After changing the methods from “between group linkage” to “nearest neighbor” and the interval from “squared Euclidean distances” to “block”, we can see that there are now far more clusters within the dendrogram. Nearest neighbor calculates the distance between two clusters as the minimum distance between the pairs of data points. Block makes sure not to apply any scaling to the data points so that the original distances are used in the clustering process. As a result, this leads to many more clusters and small groups of clusters since clusters are joined together based on the closest data points. Since block does not apply any scaling to the distances and uses the original distance, small differences between the data points leads to the formation of more clusters. We can see this in the dendrogram, as states that are nearby one another (Such as Alabama and Kentucky, or Michigan and Ohio) also tend to have similar data points, and as such we can see they are mostly clustered together.

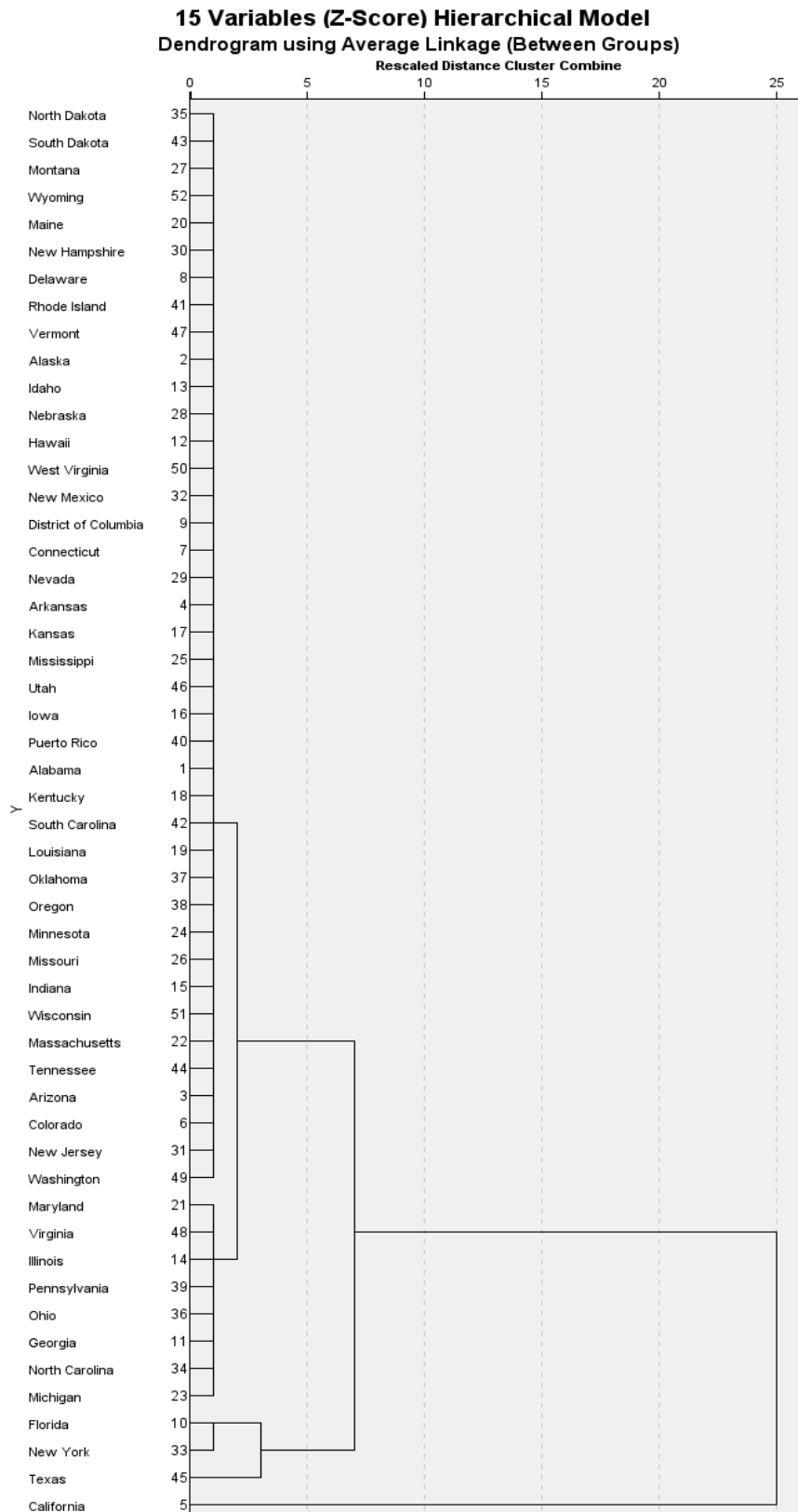
Run a Hierarchical Model on 15 variables of your choice where the variables have been standardized using z scores. (Variables do not need to be rated). Which variables show to be most important on your dendrogram. Why do you think they are the most important?

Variables Chosen:

Variables(s):

-  Healthcare, 2022 [Health1]
-  Healthcare | Drugs, 2022 [Health3]
-  Healthcare | Drugs | Nonprescription drugs, 2022 [Health5]
-  Healthcare | Drugs | Nonprescription vitamins, 2022 [Health7]
-  Healthcare | Drugs | Prescription drugs, 2022 [Health9]
-  # Industry | Agriculture, forestry, fishing and hunting, and mining, 2021 [Estimated] [Labour1]
-  # Industry | Construction, 2021 [Estimated] [Labour2]
-  # Industry | Manufacturing, 2021 [Estimated] [Labour3]
-  # Occupation | Farming, fishing, and forestry occupations, 2021 [Estimated] [Occupation12]
-  # Occupation | Transportation occupations, 2021 [Estimated] [Occupation16]
-  # Occupation | Building and grounds cleaning and maintenance occupations, 2021 [Estima...]
-  # Employment Status | In civilian labor force, Employed, 2021 [Estimated] [Employment4]
-  # Class of Worker | Local government workers, 2021 [Estimated] [Worker6]
-  # Class of Worker | State government workers, 2021 [Estimated] [Worker7]
-  # Class of Worker | Federal government workers, 2021 [Estimated] [Worker8]

15 Z-Score Variables Dendrogram:



There seem to be 5 cluster groups based on the dendrogram. The variables that seem to be the most important based on the dendrogram are the healthcare variables and potentially the class of worker variables. This is because the states which spend a lot more on healthcare than average and also have a higher-than-average number of government workers tend to be clustering together. This is states such as North Dakota, South Dakota, Montana, Wyoming (Source: <https://www.forbes.com/advisor/health-insurance/most-and-least-expensive-states-for-health-care-ranked/>)

<https://www.businessinsider.com/percentage-workforce-employed-by-government-every-us-state-2019-1>)

These states spend a lot more than the national average on healthcare while also having many more government employees than the national average. Other clusters have states such as Florida, New York, Texas, and California which is due to the population size which is why they will almost always veer away from the standard deviation. The other cluster group with states like Michigan, Pennsylvania, Ohio spend a lot less on healthcare than the mean and also have much less government workers (Same sources as above). These variables are likely the most important since they are the best for grouping similar states into clusters. Other variables like Occupation Type and Employment Status were probably harder to make clusters for as there is likely a lot of variation, so it is hard to make clusters than for something like healthcare and the amount of government workers.

K-Means SPSS Output

Employ the same 5 variables and region that you used above in a K means analysis. How do the clusters change from unrated, rated, z score, to rated-z score (based on 4 cluster groups)?

1. Unrated

Initial Cluster Centers

	Cluster			
	1	2	3	4
# Educational Attainment Less than high school diploma, 2021 [Estimated]	3237138	31221	4750478	984923
# Employment Status In civilian labor force, Unemployed, 2021 [Estimated]	690964	14079	991697	246082
# Poverty Status by Age In Poverty, 2021 [Estimated]	4057889	63839	4911186	1469961
# Computers In Household No Computer, 2021 [Estimated]	604768	14615	603378	269272
# Internet Subscriptions In Household No Internet access, 2021 [Estimated]	1096773	22810	988589	449954

Iteration History^a

Iteration	Change in Cluster Centers			
	1	2	3	4
1	.000	357523.293	.000	182795.478
2	900424.723	31612.272	.000	37920.036
3	348016.255	.000	.000	140445.938
4	.000	15501.897	.000	37592.335
5	.000	16134.644	.000	33365.420
6	.000	16583.430	.000	31337.955
7	.000	33745.206	.000	54629.328
8	.000	18147.348	.000	24273.186
9	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 9. The minimum distance between initial centers is 1766482.316.

Final Cluster Centers

	Cluster			
	1	2	3	4
# Educational Attainment Less than high school diploma, 2021 [Estimated]	2397827	179101	4750478	664406
# Employment Status In civilian labor force, Unemployed, 2021 [Estimated]	548804	54119	991697	192482
# Poverty Status by Age In Poverty, 2021 [Estimated]	3177748	291355	4911186	1063861
# Computers In Household No Computer, 2021 [Estimated]	539801	66524	603378	238209
# Internet Subscriptions In Household No Internet access, 2021 [Estimated]	888125	99489	988589	345164

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		3787434.195	2958050.853	2825950.364
2	3787434.195		6648164.358	970186.789
3	2958050.853	6648164.358		5716990.962
4	2825950.364	970186.789	5716990.962	

Number of Cases in each Cluster

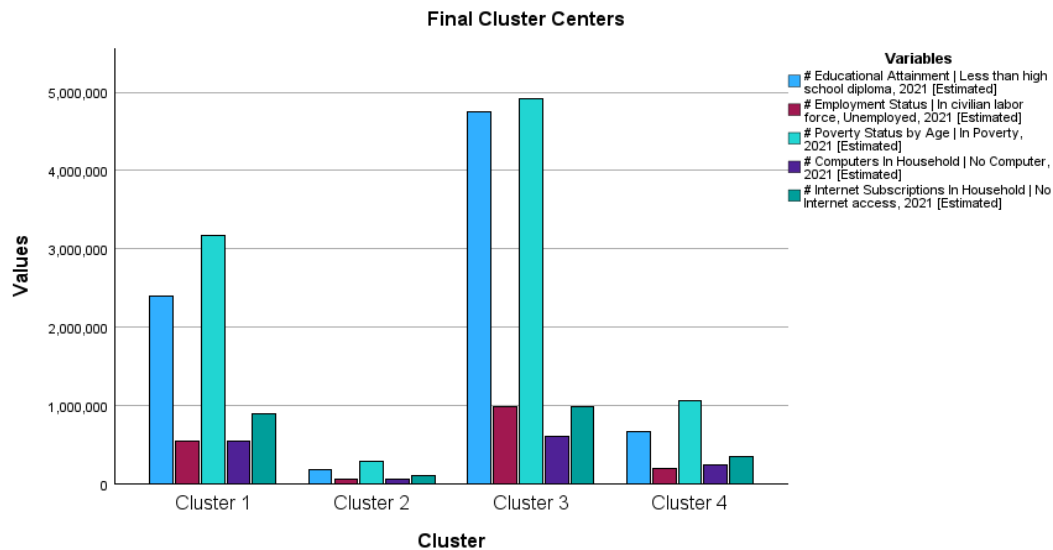
Cluster	1	2	3	4
1	3.000			
2	29.000			
3	1.000			
4	19.000			
Valid	52.000			
Missing	.000			

Cluster Membership

Case Number	Name	Cluster	Distance
1	Alabama	4	347060.542
2	Alaska	2	269434.737
3	Arizona	4	141920.256
4	Arkansas	2	245412.324
5	California	3	.000
6	Colorado	2	332036.021
7	Connecticut	2	116184.878
8	Delaware	2	225259.493
9	District of Columbia	2	240765.615
10	Florida	1	562764.453
11	Georgia	4	531476.618
12	Hawaii	2	201066.192
13	Idaho	2	110178.273
14	Illinois	4	636083.901
15	Indiana	4	242216.353
16	Iowa	2	82021.926
17	Kansas	2	53726.945
18	Kentucky	4	419033.793
19	Louisiana	4	254846.235
20	Maine	2	183046.851
21	Maryland	2	403518.671
22	Massachusetts	4	436917.349
23	Michigan	4	298404.765
24	Minnesota	2	276388.905
25	Mississippi	2	341668.707
26	Missouri	4	359652.816
27	Montana	2	213787.241
28	Nebraska	2	110516.775
29	Nevada	2	153923.071
30	New Hampshire	2	232762.079
31	New Jersey	4	197273.589
32	New Mexico	2	112910.812
33	New York	1	696032.511
34	North Carolina	4	479403.463
35	North Dakota	2	267010.751
36	Ohio	4	575433.015
37	Oklahoma	2	360689.724
38	Oregon	2	255271.402
39	Pennsylvania	4	637927.091
40	Puerto Rico	4	330739.200
41	Rhode Island	2	203325.569
42	South Carolina	4	408784.792
43	South Dakota	2	232987.679
44	Tennessee	4	100839.676
45	Texas	1	1243807.125
46	Utah	2	59075.954
47	Vermont	2	284674.151
48	Virginia	4	184413.545
49	Washington	4	393420.638
50	West Virginia	2	31686.132
51	Wisconsin	2	415973.858
52	Wyoming	2	289498.558

ANOVA						
	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
# Educational Attainment Less than high school diploma, 2021 [Estimated]	1.070E+13	3	45274482629	48	236.333	<.001
# Employment Status In civilian labor force, Unemployed, 2021 [Estimated]	4.955E+11	3	3153483266.0	48	157.117	<.001
# Poverty Status by Age In Poverty, 2021 [Estimated]	1.421E+13	3	84904167204	48	167.424	<.001
# Computers In Household No Computer, 2021 [Estimated]	3.322E+11	3	4340639762.4	48	76.523	<.001
# Internet Subscriptions In Household No Internet access, 2021 [Estimated]	8.653E+11	3	8924450332.7	48	96.961	<.001

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.



Based on the k-means SPSS output of the unrated variables, we can see that it took 9 iterations to create the final clusters, so there is sufficient clustering within our data set. All the variables are also significant and important to the cluster as we can see from the ANOVA table. We can see that the large states (Cluster 1: Florida, New York, Texas) are grouped together and that California is in a cluster of its own (Cluster 3). Cluster 2 seems to be mostly northern states which tend to have less poverty such as Delaware, Colorado, Minnesota, and Cluster 4 seems to be mostly southern states such as Kentucky, Louisiana, Missouri, which tend to have more poverty. The cluster centers graph also shows significant variation in the variables between the clusters.

2. Rated

Initial Cluster Centers

	Cluster			
	1	2	3	4
LessThanHighSchoolRate	1193.89	565.31	1745.51	1111.02
d				
UnemployedRated	249.23	179.35	603.08	318.32
InPovertyRated	1234.28	717.35	4306.27	1903.15
NoComputersRated	151.64	209.00	883.12	427.64
NoInternetRated	248.45	274.15	1066.63	693.44

Iteration History^a

Iteration	Change in Cluster Centers			
	1	2	3	4
1	388.125	334.418	.000	199.267
2	23.662	16.517	.000	.000
3	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 819.245.

Final Cluster Centers

	Cluster			
	1	2	3	4
LessThanHighSchoolRate	883.64	631.96	1745.51	1033.50
d				
UnemployedRated	242.32	207.20	603.08	269.04
InPovertyRated	1332.11	1043.11	4306.27	1749.75
NoComputersRated	271.09	247.81	883.12	404.39
NoInternetRated	422.17	355.81	1066.63	608.60

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		391.204	3241.684	500.127
2	391.204		3599.140	867.656
3	3241.684	3599.140		2755.591
4	500.127	867.656	2755.591	

Number of Cases in each Cluster

Cluster	1	19.000
	2	25.000
	3	1.000
	4	7.000
Valid		52.000
Missing		.000

Cluster Membership

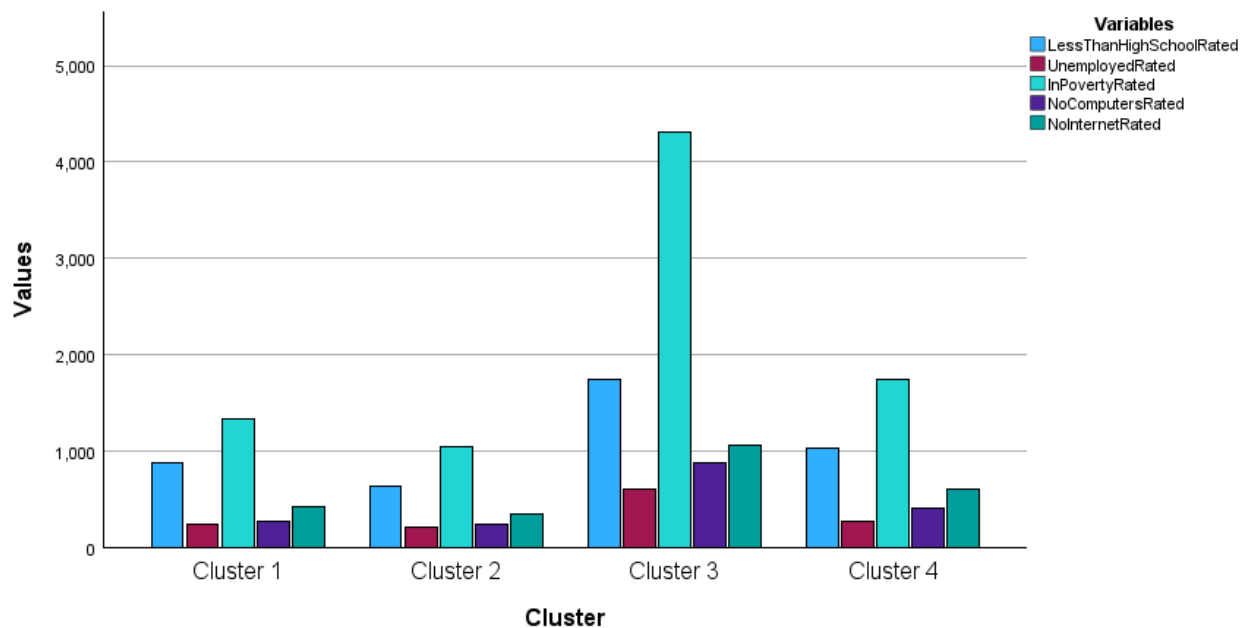
Case Number	Name	Cluster	Distance
1	Alabama	4	167.715
2	Alaska	2	218.522
3	Arizona	1	136.225
4	Arkansas	4	176.079
5	California	1	387.714
6	Colorado	2	147.001
7	Connecticut	2	151.940
8	Delaware	2	180.696
9	District of Columbia	1	276.559
10	Florida	1	103.486
11	Georgia	1	73.860
12	Hawaii	2	174.121
13	Idaho	2	200.568
14	Illinois	1	161.325
15	Indiana	1	116.890
16	Iowa	2	149.059
17	Kansas	2	130.343
18	Kentucky	4	138.574
19	Louisiana	4	154.542
20	Maine	2	120.067
21	Maryland	2	203.620
22	Massachusetts	2	117.495
23	Michigan	1	195.395
24	Minnesota	2	181.288
25	Mississippi	4	199.267
26	Missouri	1	161.747
27	Montana	2	281.006
28	Nebraska	2	49.994
29	Nevada	1	167.987
30	New Hampshire	2	345.700
31	New Jersey	2	168.147
32	New Mexico	4	113.944
33	New York	1	127.678
34	North Carolina	1	53.343
35	North Dakota	2	191.746
36	Ohio	1	161.369
37	Oklahoma	1	188.408
38	Oregon	2	202.642
39	Pennsylvania	1	223.952
40	Puerto Rico	3	.000
41	Rhode Island	1	181.611
42	South Carolina	1	130.825
43	South Dakota	2	278.833
44	Tennessee	1	174.754
45	Texas	1	246.864
46	Utah	2	303.174
47	Vermont	2	125.617
48	Virginia	2	150.033
49	Washington	2	163.034
50	West Virginia	4	70.104
51	Wisconsin	2	135.516
52	Wyoming	2	120.409

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
LessThanHighSchoolRate	704827.630	3	11320.667	48	62.260	<.001
UnemployedRate	54627.437	3	1657.033	48	32.967	<.001
InPovertyRate	4040254.728	3	14104.511	48	286.451	<.001
NoComputersRate	164889.486	3	3621.278	48	45.534	<.001
NoInternetRate	255696.170	3	5502.545	48	46.469	<.001

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Final Cluster Centers



Based on the k-means SPSS output of the rated variables, we can see that it only took 3 iterations to create the final clusters, so there is sufficient clustering within our data set. All the variables are also significant and important to the cluster as we can see from the ANOVA table. We can see that the only outlier cluster (Cluster 3) is Puerto Rico, which makes sense as majority of the population in Puerto Rico is in poverty. The high poverty ratio states which have a large portion of the population in poverty such as Alabama, Kentucky, Louisiana are in a cluster (Cluster 4), and the other clusters seem to vary with states between medium and low levels of poverty. The cluster centers graph also shows significant variation in the variables between the clusters, however the bars are closer than before, which is why the groups have a bit more states balanced out in each group rather than 1 or 3 states in a single cluster.

3. Z-Score

Initial Cluster Centers

	Cluster			
	1	2	3	4
Zscore: # Educational Attainment Less than high school diploma, 2021 [Estimated]	.14449	-.66009	5.09675	1.70449
Zscore: # Employment Status In civilian labor force, Unemployed, 2021 [Estimated]	.49359	-.76543	4.69000	1.87298
Zscore: # Poverty Status by Age In Poverty, 2021 [Estimated]	.53166	-.79942	4.26508	1.89434
Zscore: # Computers In Household No Computer, 2021 [Estimated]	.85971	-.99068	2.83989	2.64912
Zscore: # Internet Subscriptions In Household No Internet access, 2021 [Estimated]	.73537	-.94055	3.02540	2.21950

Iteration History^a

Iteration	Change in Cluster Centers			
	1	2	3	4
1	.649	.589	1.343	.450
2	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 3.200.

Final Cluster Centers

	Cluster			
	1	2	3	4
Zscore: # Educational Attainment Less than high school diploma, 2021 [Estimated]	.10112	-.47236	4.17372	1.71492
Zscore: # Employment Status In civilian labor force, Unemployed, 2021 [Estimated]	.20780	-.52737	3.85090	1.82186
Zscore: # Poverty Status by Age In Poverty, 2021 [Estimated]	.23733	-.55642	3.81931	1.99420
Zscore: # Computers In Household No Computer, 2021 [Estimated]	.48713	-.66807	2.84441	2.21491
Zscore: # Internet Subscriptions In Household No Internet access, 2021 [Estimated]	.40006	-.63672	3.24753	2.18444

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		1.977	7.507	3.803
2	1.977		9.345	5.751
3	7.507	9.345		3.876
4	3.803	5.751	3.876	

Number of Cases in each Cluster

Cluster	1	19.000
	2	29.000
	3	2.000
	4	2.000
Valid		52.000
Missing		.000

Cluster Membership

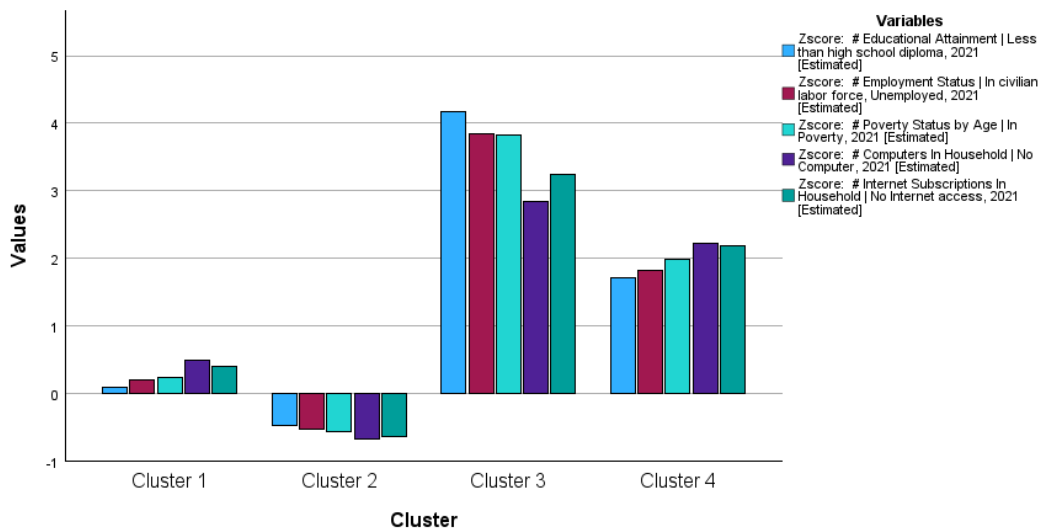
Case Number	Name	Cluster	Distance
1	Alabama	1	.677
2	Alaska	2	.563
3	Arizona	1	.746
4	Arkansas	2	.554
5	California	3	1.343
6	Colorado	2	.579
7	Connecticut	2	.348
8	Delaware	2	.491
9	District of Columbia	2	.518
10	Florida	4	.450
11	Georgia	1	.809
12	Hawaii	2	.405
13	Idaho	2	.297
14	Illinois	1	1.443
15	Indiana	1	.415
16	Iowa	2	.335
17	Kansas	2	.160
18	Kentucky	1	.909
19	Louisiana	1	.595
20	Maine	2	.343
21	Maryland	2	.847
22	Massachusetts	1	.851
23	Michigan	1	.649
24	Minnesota	2	.695
25	Mississippi	2	.730
26	Missouri	1	.678
27	Montana	2	.404
28	Nebraska	2	.216
29	Nevada	2	.219
30	New Hampshire	2	.455
31	New Jersey	1	.435
32	New Mexico	2	.219
33	New York	4	.450
34	North Carolina	1	.896
35	North Dakota	2	.514
36	Ohio	1	1.293
37	Oklahoma	2	.682
38	Oregon	2	.372
39	Pennsylvania	1	1.897
40	Puerto Rico	1	.440
41	Rhode Island	2	.409
42	South Carolina	1	.861
43	South Dakota	2	.464
44	Tennessee	1	.225
45	Texas	3	1.343
46	Utah	2	.274
47	Vermont	2	.567
48	Virginia	1	.310
49	Washington	2	1.006
50	West Virginia	2	.144
51	Wisconsin	1	.968
52	Wyoming	2	.589

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore: # Educational Attainment Less than high school diploma, 2021 [Estimated]	15.796	3	.075	48	209.840	<.001
Zscore: # Employment Status In civilian labor force, Unemployed, 2021 [Estimated]	15.061	3	.121	48	124.280	<.001
Zscore: # Poverty Status by Age In Poverty, 2021 [Estimated]	15.726	3	.080	48	197.436	<.001
Zscore: # Computers In Household No Computer, 2021 [Estimated]	14.482	3	.157	48	92.006	<.001
Zscore: # Internet Subscriptions In Household No Internet access, 2021 [Estimated]	15.145	3	.116	48	130.615	<.001

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Final Cluster Centers



Based on the k-means SPSS output of the z-score variables, we can see that it only took 2 iterations to create the final clusters, so there is sufficient clustering within our data set. All the variables are also significant and important to the cluster as we can see from the ANOVA table. We can see that the large states are once again grouped together in their own two clusters. Cluster 3 has California and Texas, and Cluster 4 has Florida and New York. This is because these states have the most deviation the mean amount of poverty and the ones grouped together are close enough to have been clustered. Surprisingly, Puerto Rico is not in a cluster of its own, and is instead with the other known low poverty states now (Cluster 1: Alabama, Kentucky, Missouri, etc.). The cluster centers graph also shows significant variation in the variables between the clusters.

4. Rated & Z-Score

Initial Cluster Centers

	Cluster			
	1	2	3	4
Zscore (LessThanHighSchoolRate d)	1.72811	4.14445	-1.22787	1.36510
Zscore(UnemployedRated)	.19195	5.31373	-1.01838	1.19190
Zscore(InPovertyRated)	-.14433	5.98816	-.55977	1.19092
Zscore (NoComputersRated)	-1.20510	5.18398	.38322	1.20562
Zscore(NoInternetRated)	-1.26097	4.49286	.34576	1.86840

Iteration History^a

	Change in Cluster Centers			
Iteration	1	2	3	4
1	1.668	.000	1.182	1.056
2	.146	.000	.208	.299
3	.099	.000	.164	.138
4	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 4. The minimum distance between initial centers is 3.934.

Final Cluster Centers

	Cluster			
	1	2	3	4
Zscore (LessThanHighSchoolRate d)	.19103	4.14445	-.78197	.84563
Zscore(UnemployedRated)	.24789	5.31373	-.63695	.31044
Zscore(InPovertyRated)	-.22319	5.98816	-.40832	.70503
Zscore (NoComputersRated)	-.56059	5.18398	-.08361	.77837
Zscore(NoInternetRated)	-.53755	4.49286	-.20834	1.06332

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		11.755	1.449	2.377
2	11.755		12.266	9.747
3	1.449	12.266		2.673
4	2.377	9.747	2.673	

Number of Cases in each Cluster

Cluster	1	20.000
	2	1.000
	3	21.000
	4	10.000
Valid		52.000
Missing		.000

Cluster Membership

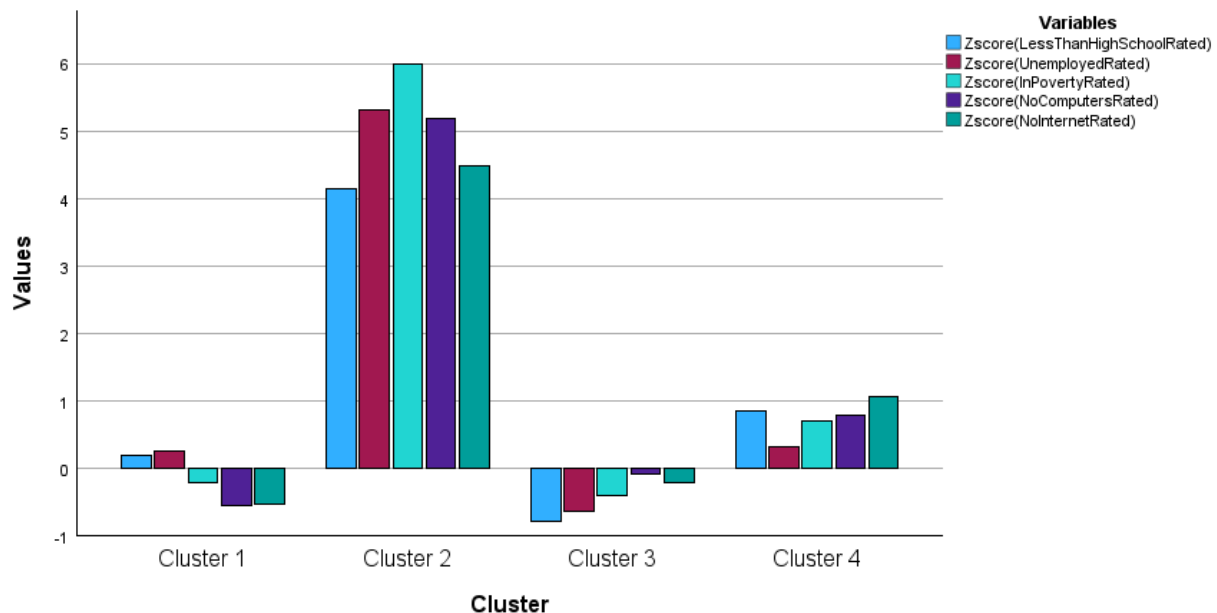
Case Number	Name	Cluster	Distance
1	Alabama	4	.401
2	Alaska	1	2.159
3	Arizona	1	.644
4	Arkansas	4	.719
5	California	1	1.820
6	Colorado	3	1.356
7	Connecticut	1	.861
8	Delaware	1	.489
9	District of Columbia	1	1.848
10	Florida	1	.693
11	Georgia	1	.842
12	Hawaii	3	.940
13	Idaho	3	.838
14	Illinois	1	.591
15	Indiana	3	1.200
16	Iowa	3	.634
17	Kansas	3	.300
18	Kentucky	4	.507
19	Louisiana	4	.873
20	Maine	3	.434
21	Maryland	1	.831
22	Massachusetts	1	.781
23	Michigan	3	.932
24	Minnesota	3	.862
25	Mississippi	4	1.454
26	Missouri	3	.871
27	Montana	3	1.079
28	Nebraska	3	.341
29	Nevada	1	.760
30	New Hampshire	3	1.352
31	New Jersey	1	.533
32	New Mexico	4	.755
33	New York	1	1.053
34	North Carolina	1	1.100
35	North Dakota	3	.945
36	Ohio	3	1.014
37	Oklahoma	4	1.243
38	Oregon	1	.813
39	Pennsylvania	3	1.153
40	Puerto Rico	2	.000
41	Rhode Island	1	.826
42	South Carolina	4	1.038
43	South Dakota	3	.921
44	Tennessee	4	.862
45	Texas	1	1.277
46	Utah	3	2.375
47	Vermont	3	.597
48	Virginia	1	.649
49	Washington	1	1.421
50	West Virginia	4	.952
51	Wisconsin	3	.755
52	Wyoming	3	.851

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore (LessThanHighSchoolRated)	12.633	3	.273	48	46.282	<.001
Zscore(UnemployedRated)	12.983	3	.251	48	51.708	<.001
Zscore(InPovertyRated)	15.109	3	.118	48	127.821	<.001
Zscore (NoComputersRated)	13.121	3	.242	48	54.129	<.001
Zscore(NoInternetRated)	12.728	3	.267	48	47.665	<.001

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Final Cluster Centers



Based on the k-means SPSS output of the rated & z-score variables, we can see that it took 4 iterations to create the final clusters, so there is sufficient clustering within our data set. All the variables are also significant and important to the cluster as we can see from the ANOVA table. We can see that Puerto Rico is back to being in its own cluster (Cluster 2), likely because of the extreme amounts of poverty, and also this time we can see that we have much more balanced clusters for the other 3 than previously where most states were in 2 clusters. Cluster 4 with 10 states has the known high poverty southern states again like Kentucky, Louisiana, Alabama. Cluster 1 seems to be mostly states with moderate amounts of poverty like Oregon, Georgia, Washington. All the large states (New York, Texas, California, Florida) have seemed to move into this cluster since they likely have moderate amounts of their population in poverty than high amounts like the southern states. Cluster 3 seems to be mostly prosperous states with low amounts of poverty like Maine, Minnesota, Michigan.

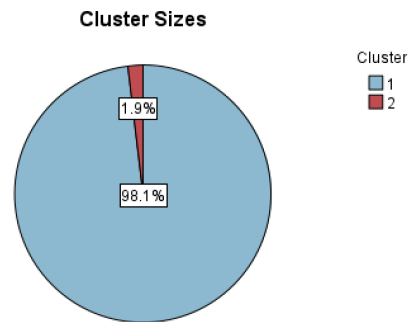
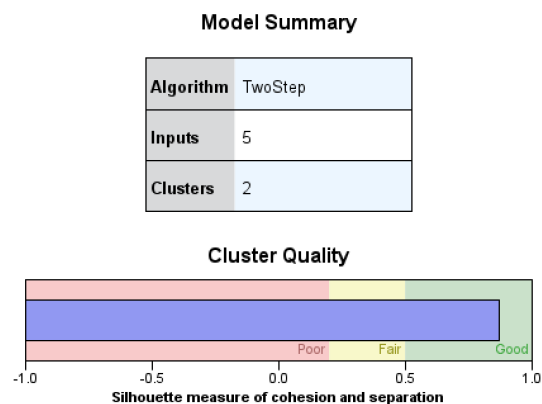
Determining Groups

Based on your final Hierarchical Model, based on rated-z scores, how many groups should be used in K Means?

Based on the final Hierarchical model, 5 groups should be used in K means. This is because when creating the group breaks on the dendrogram, 5 groups were identified.

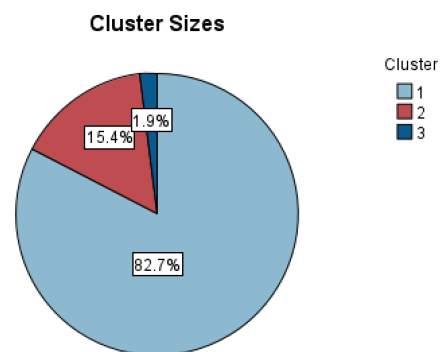
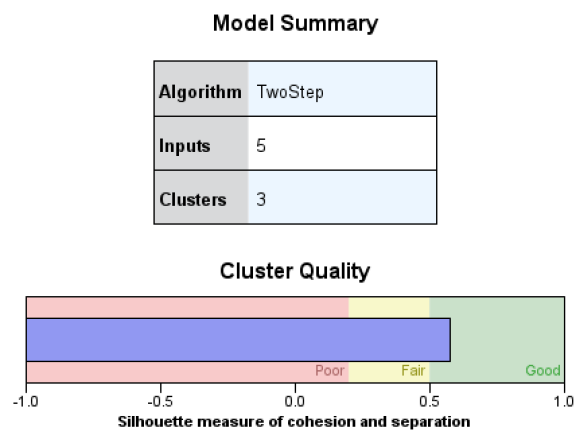
Based on your final model (rated-z scores) What is the optimal number of groups to be used? Show how you determined this using 2 step cluster analyses? And show bar graph for your clusters.

2 Clusters:



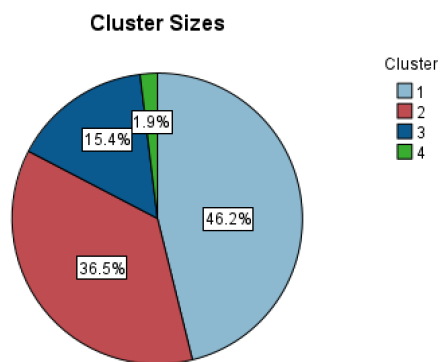
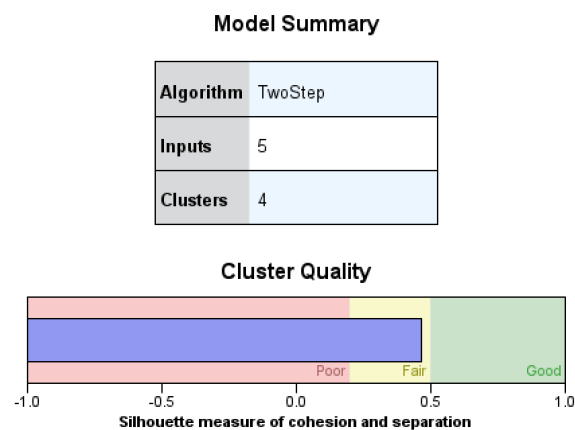
Size of Smallest Cluster	1 (1.9%)
Size of Largest Cluster	51 (98.1%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	51.00

3 Clusters:



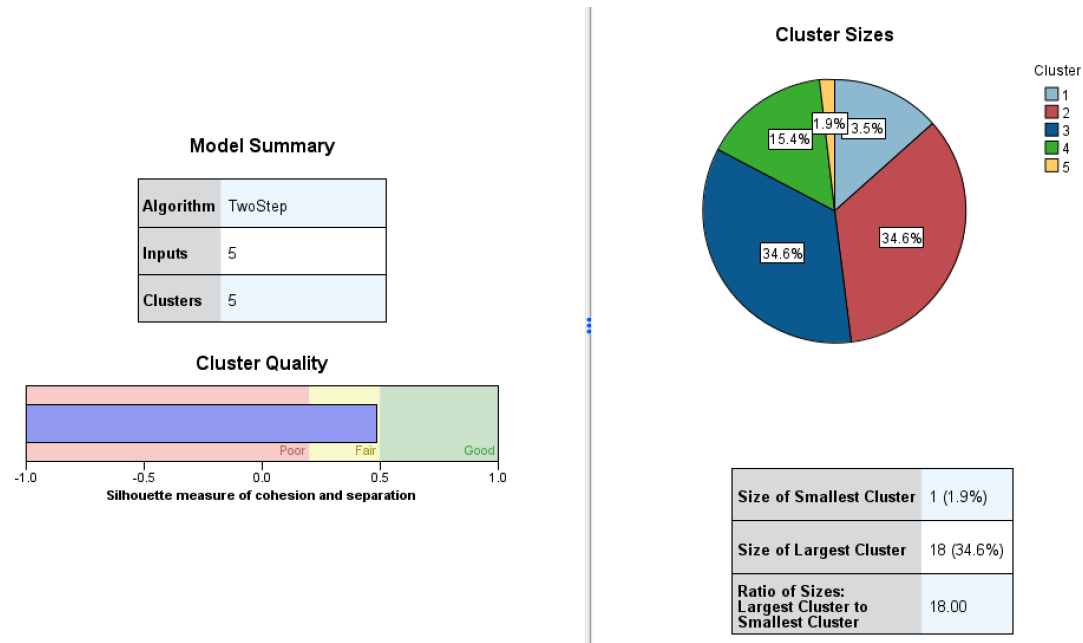
Size of Smallest Cluster	1 (1.9%)
Size of Largest Cluster	43 (82.7%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	43.00

4 Clusters:

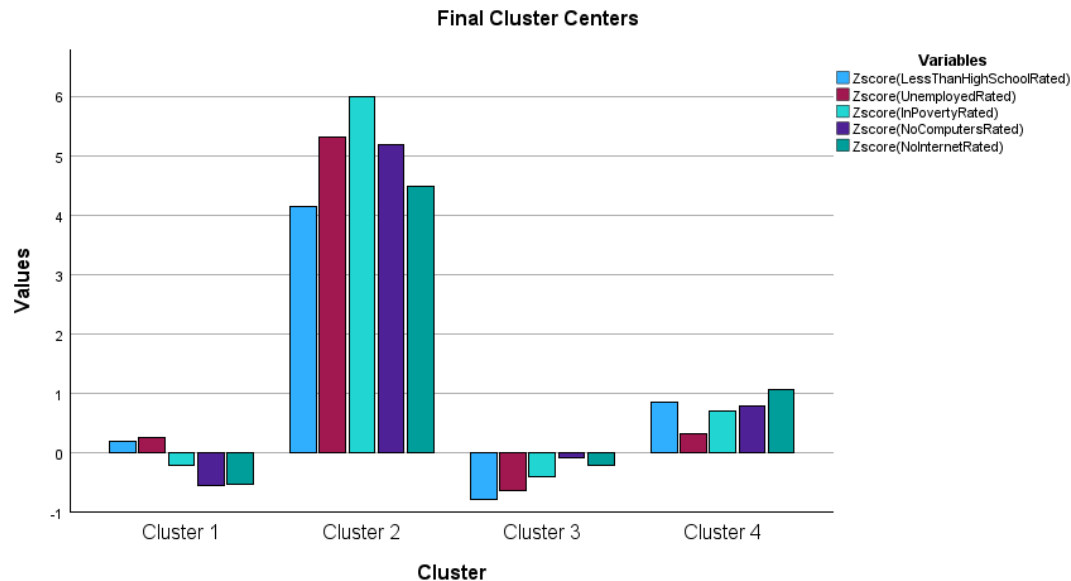


Size of Smallest Cluster	1 (1.9%)
Size of Largest Cluster	24 (46.2%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	24.00

5 Clusters:



Based on the 2-step cluster analysis, the optimal number of groups to use is 4. Although the 2-step cluster analysis showed that 2 clusters and 3 clusters had higher cluster quality, this was because most of the states were within one group. This is not optimal as there are too many states in the group to properly give a descriptor too it which is why we need more cluster groups. The difference between the 2-step for 4 and 5 clusters is negligible and either can be chose. The reason 4 was chosen in this case is because it is easier to give descriptors to the cluster. 2-step analysis was not done for more than 5 clusters since it is recommended that more than 5 clusters should not be used as the groups begin to become too small and the clustering becomes too granular and doesn't convey enough meaningful information. We can see if the bar graph below that there is plenty of variation in the variables between the cluster groups, so using 4 clusters is sufficient.



Articulation & Discussion of Rated & Z-Score Model SPSS Output:

Initial Cluster Centers

	Cluster			
	1	2	3	4
Zscore (LessThanHighSchoolRate d)	1.72811	4.14445	-1.22787	1.36510
Zscore(UnemployedRated)	.19195	5.31373	-1.01838	1.19190
Zscore(InPovertyRated)	-.14433	5.98816	-.55977	1.19092
Zscore (NoComputersRated)	-1.20510	5.18398	.38322	1.20562
Zscore(NoInternetRated)	-1.26097	4.49286	.34576	1.86840

Initial Cluster centers show variations of each of the variables in each of the clusters

Iteration History^a

	Change in Cluster Centers			
Iteration	1	2	3	4
1	1.668	.000	1.182	1.056
2	.146	.000	.208	.299
3	.099	.000	.164	.138
4	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 4. The minimum distance between initial centers is 3.934.

Iteration history shows that it only took 4 iterations to create the clusters. Since it took such few iterations, this shows that there is sufficient clustering in the data set and that we chose good variables which can be clustered together.

Cluster Membership

Case Number	Name	Cluster	Distance
1	Alabama	4	.401
2	Alaska	1	2.159
3	Arizona	1	.644
4	Arkansas	4	.719
5	California	1	1.820
6	Colorado	3	1.356
7	Connecticut	1	.861
8	Delaware	1	.489
9	District of Columbia	1	1.848
10	Florida	1	.693
11	Georgia	1	.842
12	Hawaii	3	.940
13	Idaho	3	.838
14	Illinois	1	.591
15	Indiana	3	1.200
16	Iowa	3	.634
17	Kansas	3	.300
18	Kentucky	4	.507
19	Louisiana	4	.873
20	Maine	3	.434
21	Maryland	1	.831
22	Massachusetts	1	.781
23	Michigan	3	.932
24	Minnesota	3	.862
25	Mississippi	4	1.454
26	Missouri	3	.871
27	Montana	3	1.079
28	Nebraska	3	.341
29	Nevada	1	.760
30	New Hampshire	3	1.352
31	New Jersey	1	.533
32	New Mexico	4	.755
33	New York	1	1.053
34	North Carolina	1	1.100
35	North Dakota	3	.945
36	Ohio	3	1.014
37	Oklahoma	4	1.243
38	Oregon	1	.813
39	Pennsylvania	3	1.153
40	Puerto Rico	2	.000
41	Rhode Island	1	.826
42	South Carolina	4	1.038
43	South Dakota	3	.921
44	Tennessee	4	.862
45	Texas	1	1.277
46	Utah	3	2.375
47	Vermont	3	.597
48	Virginia	1	.649
49	Washington	1	1.421
50	West Virginia	4	.952
51	Wisconsin	3	.755
52	Wyoming	3	.851

Group membership shows which cluster each state belongs to. We can see Puerto Rico is an outlier as it is the only state in Cluster 2.

Final Cluster Centers

	Cluster			
	1	2	3	4
Zscore (LessThanHighSchoolRate d)	.19103	4.14445	-.78197	.84563
Zscore(UnemployedRated)	.24789	5.31373	-.63695	.31044
Zscore(InPovertyRated)	-.22319	5.98816	-.40832	.70503
Zscore (NoComputersRated)	-.56059	5.18398	-.08361	.77837
Zscore(NoInternetRated)	-.53755	4.49286	-.20834	1.06332

Final Cluster Centers shows significant amounts of variations in the variables between each of the clusters. Since there is lots of variation in the variables between each of the clusters, we know that it is clustering appropriately as it is able to identify the key difference that cluster the states together.

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		11.755	1.449	2.377
2	11.755		12.266	9.747
3	1.449	12.266		2.673
4	2.377	9.747	2.673	

We can see large differences between the final cluster centers, so we know each cluster is unique. We can see cluster 1 (which is the moderate poverty states) is most similar to the cluster 3 (low poverty) and then cluster 4 (high poverty states) which makes sense. It is very distant to cluster 2 (extreme poverty) which makes sense as Puerto Rico has extreme levels of poverty.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore (LessThanHighSchoolRate d)	12.633	3	.273	48	46.282	<.001
Zscore(UnemployedRated)	12.983	3	.251	48	51.708	<.001
Zscore(InPovertyRated)	15.109	3	.118	48	127.821	<.001
Zscore (NoComputersRated)	13.121	3	.242	48	54.129	<.001
Zscore(NoInternetRated)	12.728	3	.267	48	47.665	<.001

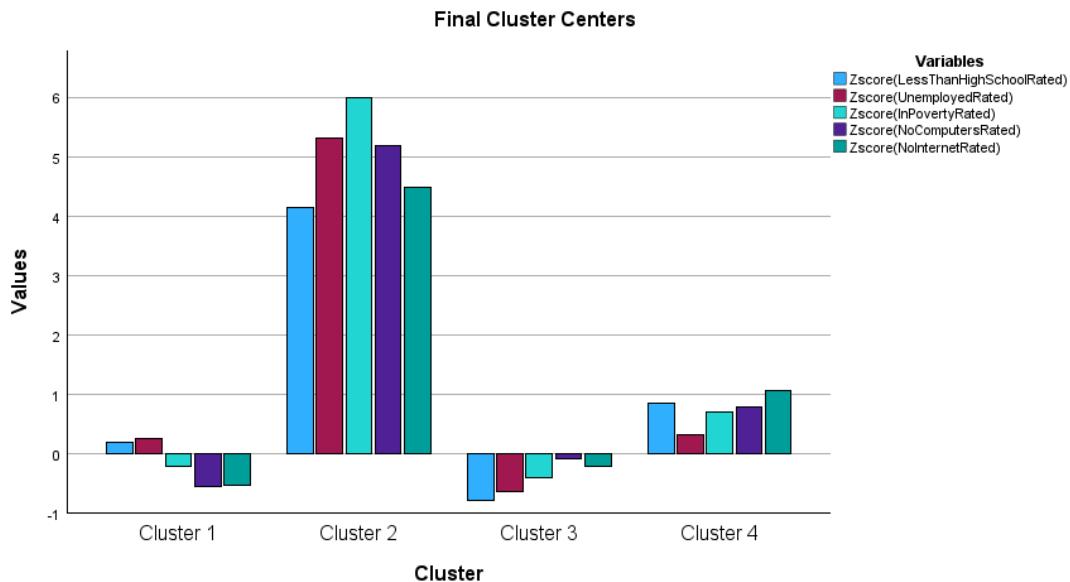
The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

ANOVA shows that all our variables are significant, so they are all important in the clustering analysis.

Number of Cases in each Cluster

Cluster	1	20.000
	2	1.000
	3	21.000
	4	10.000
Valid		52.000
Missing		.000

We can see the number of cases in each cluster. Cluster 1 is moderate poverty states, Cluster 2 is extreme poverty states, Cluster 3 is low poverty states, Cluster 4 is high poverty states.



We can see that there is sufficient variation between the variables in each of the clusters which means are clusters are not too similar.

How do your results differ from the hierarchical model?

The results are somewhat different from the hierarchical model as we have 4 cluster groups in the k-means compared to 5 cluster groups in the hierarchical. In both the hierarchical model and the k-means model, we had a lone cluster for Puerto Rico. This is likely because Puerto Rico has such an extreme level of poverty that both of the models identified it as being too different from the other clusters so it was separated. The hierarchical model also had California as a lone cluster, which fell between high and moderate levels of poverty. In the k-means analysis, California was moved into the moderate cluster, likely because it was more similar in distance to the moderate cluster than the high poverty cluster. The hierarchical model had 7 states in low poverty, while the k-means has 10 states in low poverty. All of the high poverty states in the hierarchical model were also in the 10 low poverty states of the k-means model (Alabama, Kentucky, Louisiana, New Mexico, Arkansas, Mississippi, West Virginia). Oklahoma, South Carolina, Tennessee were the additional states added into the high poverty category by the k-means model which previously were in the low poverty category in the hierarchical. This is likely because hierarchical had more groups and because the groups tend to stick together when formed while k-means is based on distance these states likely had less distance in the k-means to the high poverty group while in the hierarchical they were already formed and initially put into the low poverty cluster.