Drishti Bansal
Dr. Douglas MacDonald
DATA 5070

"Cluster analysis"

The aim for this paper is to determine whether the data collected by the instructor (unknown) for his primary research identification of latent constructs using factor analysis can be clustered into discrete groups of people who share similarities in how they express. The data consists of thirty-nine variables and a sample of 247 undergraduate students on several standardized questionnaires designed to measure such concepts as spirituality and well-being. There are five demographic variables such as sex, age, ethnic (Ethnicity), marital (Marital status) and religaff (Religious Affiliation), and the rest are variables of wellbeing concepts.

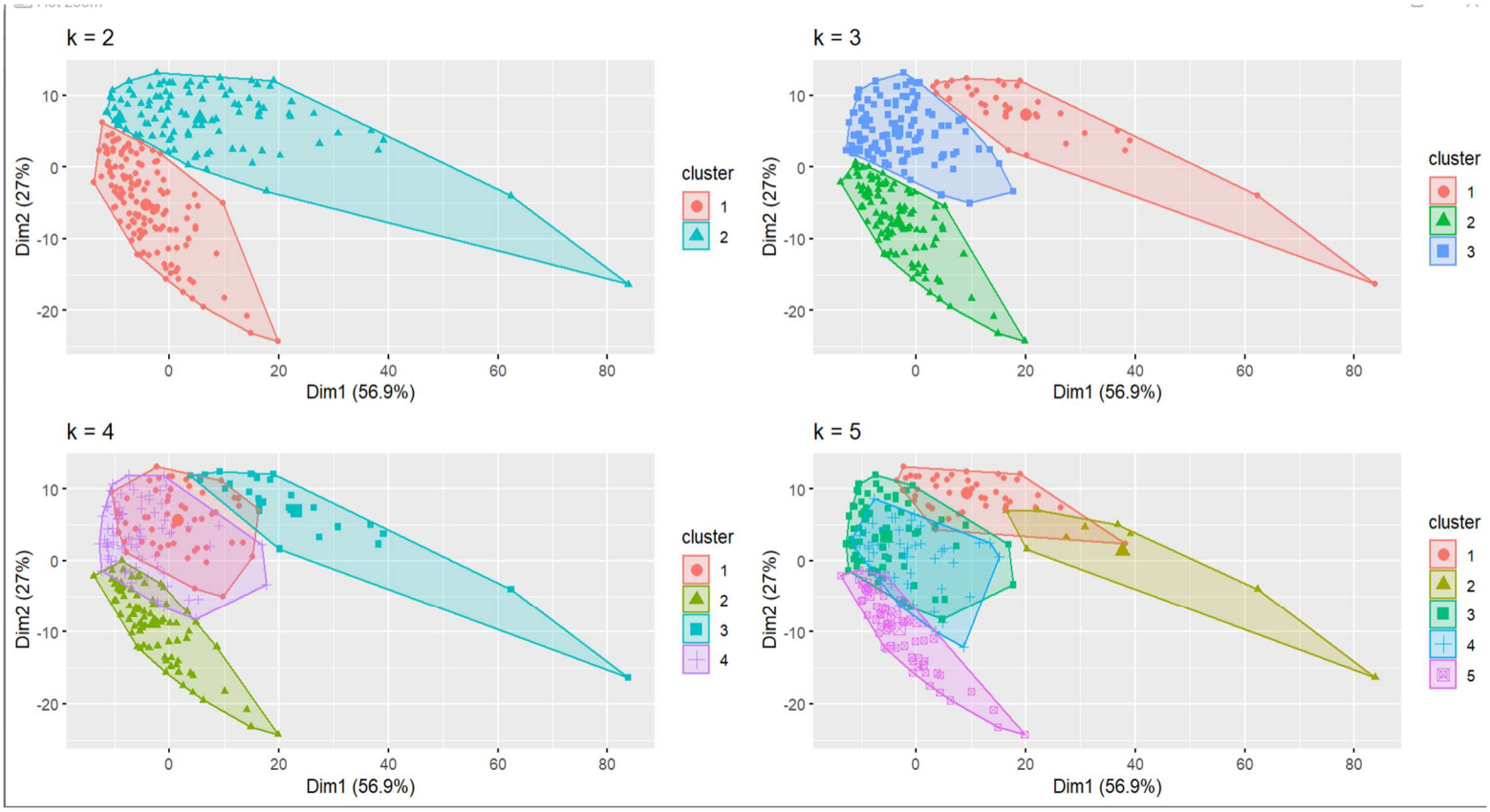
Cluster analysis is a data mining tool to cluster rows into groups on the basis of how they are closely related. Cluster analysis groups 'rows' whereas factor analysis 'clusters' columns. Suppose we have this dataset that contains 247 rows and 39 columns and we need to classify the objects in the data. The clustering process itself contains three distinctive steps:

- Calculating dissimilarity matrix
- Clustering methods (ward hierarchical clustering, or K-means)
- Assessing clusters

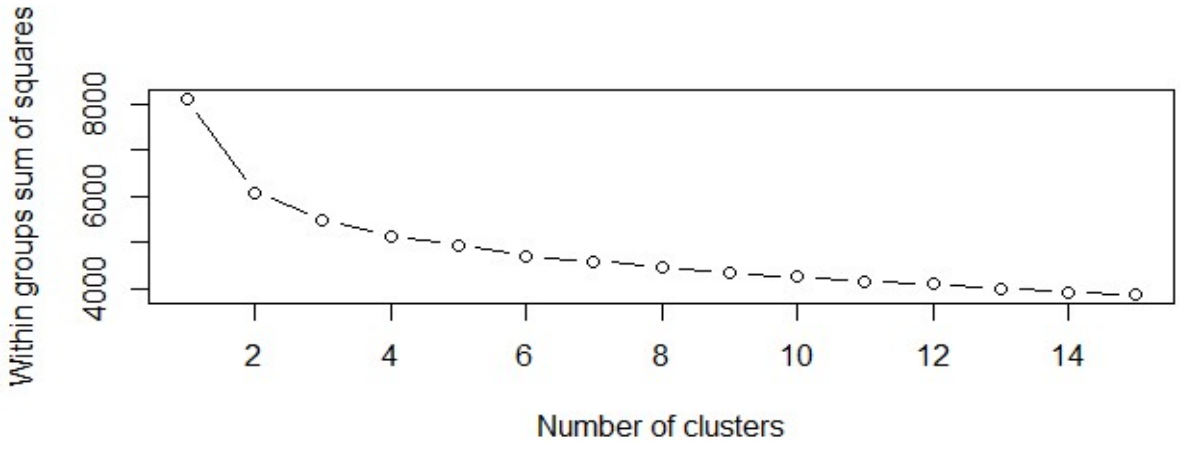
There were nominal data types for ESI-R items, Life Attitude Profile Goal Seeking, Life Attitude Profile Future Meaning to Fulfill, Spirituality Assessment Scale Meaning and Purpose in Life variables which don't make sense as they are in scale measure. Therefore, we must change the data type to scale measure in SPSS. Religaff variable is a string variable which does not deal with k-means algorithm, so we took that out of the data for the analysis. We now have thirty-eight variables for a sample of 247 algorithms (including a new variable named age_1). There were two missing values in age, we can replace those values with the average of age named age_1 and eventually, the original age variable must be invisible for the analysis to work. Ethnic variable has one missing value which can be replaced with the mean, and marital variable has two missing values, replaced with means.

For this cluster analysis, we will start with K-Means algorithm which helps with the presence of clusters by finding their centroid points. A centroid point is the average of all the data points in the cluster. By iteratively assessing the Euclidean distance between each point in the dataset, each one can be assigned to a cluster. The centroid points are random to begin with and will change each time as the process is conducted. K-means is commonly used in cluster analysis, but it has a limitation in being mainly useful for scalar data. The reason we think Euclidean is more appropriate than any algorithms is because that this algorithm will tell us the shortest path that has a connection of two coordinates.

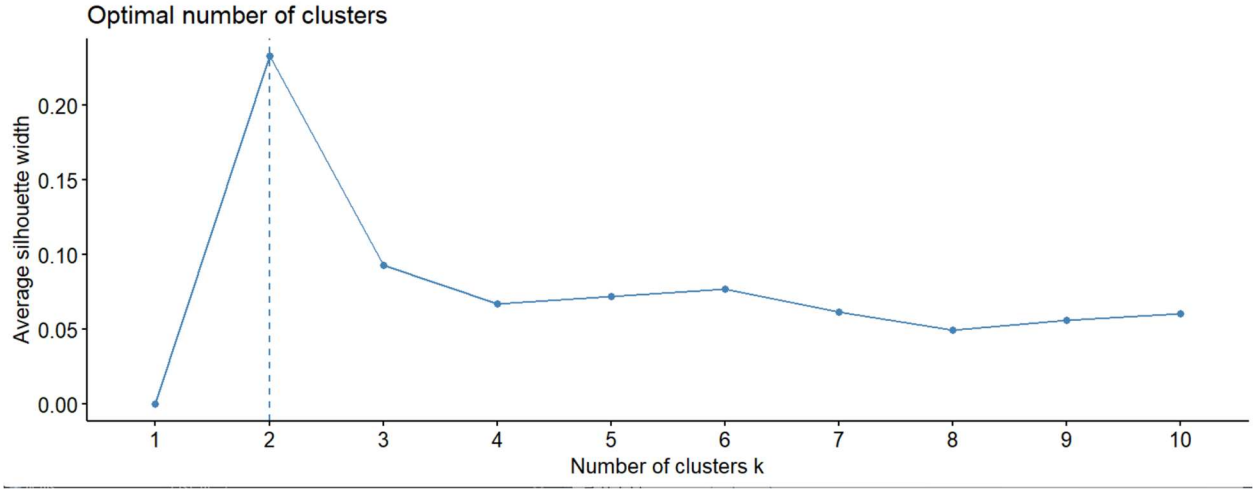
We will use Rstudio for this project instead of SPSS (although the outputs from SPSS are displayed in Appendix). We install necessary packages that are needed for this clustering analysis such as hclust, mclust, fcp, haven, cluster, dplyr, pvclust, psych. We start by standardizing the data for having the standard mean and variance which are 0 and 1 respectively because we want to avoid distorting the difference in the range of values. We compute Euclidean distance matrix with "ward" method as an input for the clustering algorithm. Ward's minimum variance criterion minimizes the total within-cluster variance. As we do not know how many clusters are optimal for the data and the number of clusters should be determined before we start the algorithm, we can use several different values of k and examine the differences in the results. We execute the kmeans() function for 2, 3, 4, and 5 clusters, and fviz_cluster () function transforms the initial set of variables into a new set of variables thorough principal component analysis (PCA); the result is shown below:



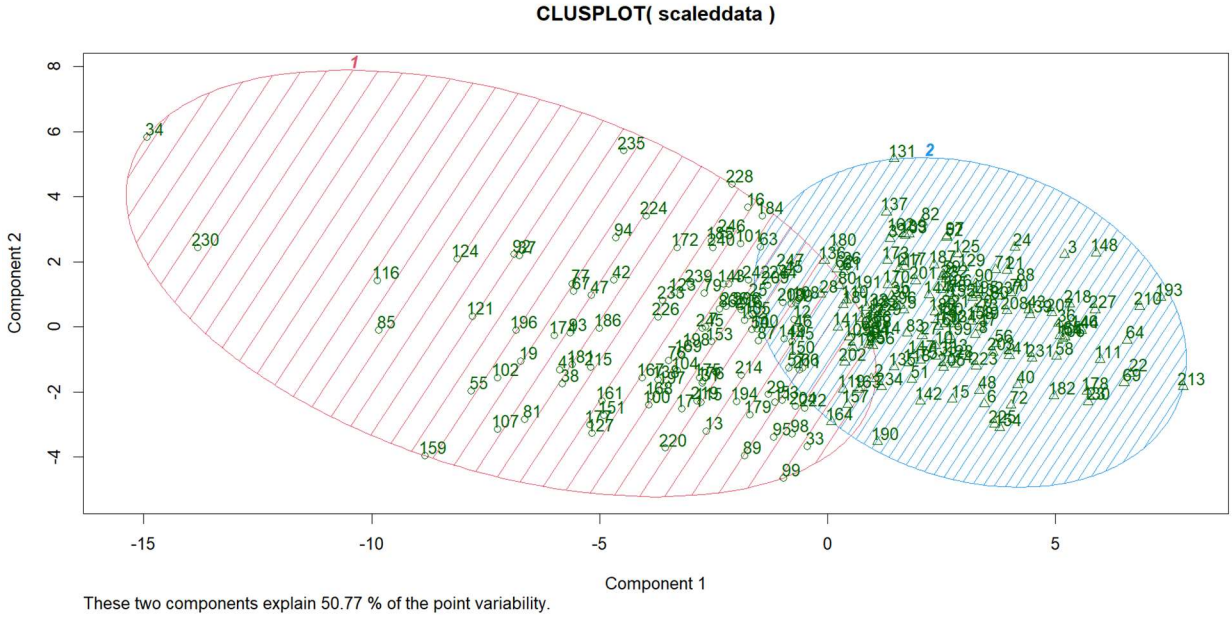
Each dimension represent a certain amount of the variation/information contained in the original data set. In this case, Dim1 and Dim2 are 27% and 56.9% respectively. When plotted, this lower-dimensional picture can be difficult to interpret. Dims together account for 83.9% (27% + 56.9%) of the information. The new centers are shown in respective cluster symbols. The higher the similarity level, the more similar the datapoints are in each cluster. So we get that from k= 2 clusters since other clusters are rather dissimilar in datapoints. However, we still have to assess the clusters, so we can use the elbow method which suggests the number of clusters within groups sum of squares.



So, we have produced the “elbow” graph. It illustrates how the within groups sum of squares as a measure of closeness of observations: the lower it is the closer the observations within the clusters are… changes for the different number of clusters. In the case of a graph above, we should go for something around two.



The figure above is the average silhouette method which basically computes the average silhouette of records for different values in k. In other words, it measures the quality of clustering. The optimal number of clusters is the value that maximizes the average silhouette over a range of possible values of k.

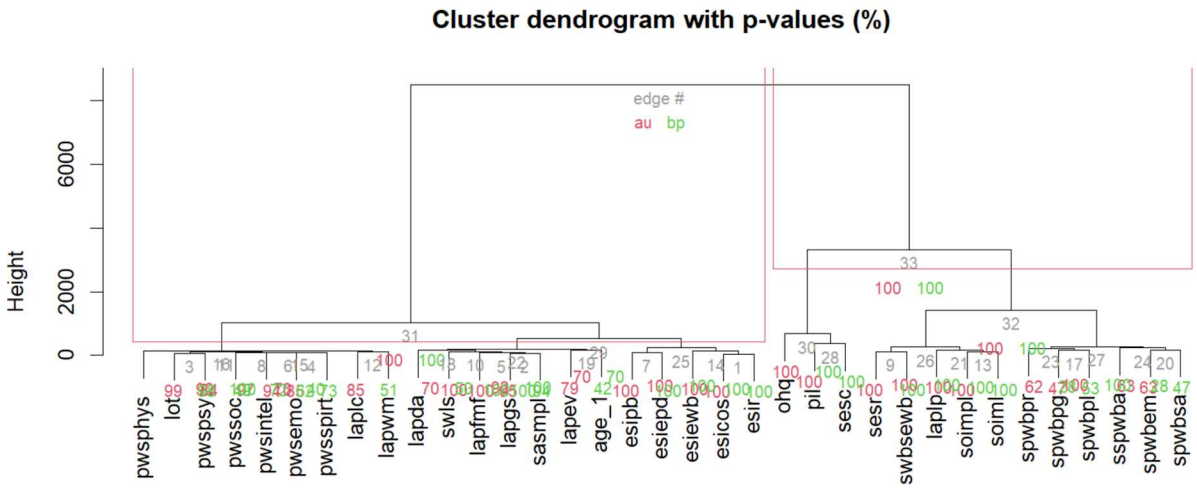


This function was executed for explanatory of analysis although this is not necessary for the analysis. We know that we cannot read this thoroughly, but we obtain the valuable information out of it, at the bottom of the figure above displays that these two components explain 50.77% of the point variability. This figure is easier to interpret than fviz_cluster figure above. Since PCA is to reduce the dimensions of variable, with this plot, almost 51% of the information about the data is explained with component 1 and 2. It is similar to the fviz_cluster for k = 2 cluster, so this is performed well.

Now, we start with another method of clustering which is Hierarchical clustering using complete linkage with Bootstrapped p-values. Here is the coding for the hierarchical clustering:

```
#Ward hierarchical clustering
df <- dist(scaleddata, method = "euclidean")
fit <- hclust(df, method = "ward.D")
plot(fit, cex = 0.4)

#ward hierarchical clustering with Bootstrapped p values
fit <- pvclust(scaleddata, method.hclust = "ward", method.dist = "euclidean")
plot(fit)
pvrect(fit, alpha = 0.95)
```



So now, we understand that this is the final dendrogram which is a compact visualization of the dissimilarity matrix. We can move to the last step of assigning clusters to the data points, this can be done using clusplot () function as resulting from hclust () into several groups by specifying the desired number of groups (k). Following our demo, we can see ohq which is Oxford Happiness Questionnaire, joins later at least height of 50. This means that the cluster joins is closer before other variables join. We can see that there are many pairs of samples that are fairly close but it is also hard to see which pairs. There are two cutting trees in red box upside down, we see that in one cluster, it represents perceived well-being, life experiences/attitudes like aha moment; and in second cluster on the right red box, it represents spiritual well-being in general.

In conclusion, this study used two clustering methods, which are hierarchical and k-means clustering using different techniques. The result from the original data suggests that classifying categorical and continuous variables into those subgroups varied to some extent [Appendix 4], some functions in Rstudio were easier to use and the interpretability of their visualization of findings also varied. The result from the scale measures of variables data indicates that the clustering methods showed the number of clusters are obtained.

Appendix

1. This picture below shows the summary of the data without the string variable which is religious affiliation.

```
> summary(data)
      sex              age            ethnic          marital           pil             esipb            esiepd           esicos      
 Min.   :1.000   Min.   :17.00   Min.   :1.000   Min.   :1.000   Min.   : 56.0   Min.   :  0.00   Min.   :  0.00   Min.   : 0.0  
 1st Qu.:2.000   1st Qu.:19.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:100.5   1st Qu.:  7.00   1st Qu.:  6.00   1st Qu.:14.0  
 Median :2.000   Median :21.00   Median :5.000   Median :1.000   Median :109.0   Median : 11.00   Median : 10.00   Median :18.0  
 Mean   :1.781   Mean   :23.36   Mean   :3.618   Mean   :1.241   Mean   :107.6   Mean   : 10.02   Mean   : 10.12   Mean   :17.2  
 3rd Qu.:2.000   3rd Qu.:24.00   3rd Qu.:5.000   3rd Qu.:1.000   3rd Qu.:117.0   3rd Qu.: 13.00   3rd Qu.: 14.00   3rd Qu.:21.0  
 Max.   :2.000   Max.   :55.00   Max.   :6.000   Max.   :5.000   Max.   :138.0   Max.   : 22.00   Max.   : 23.00   Max.   :24.0  
                 NA's   :2       NA's   :1       NA's   :2                                                                      
     esiewb           esir            swls           sspwba          spwbem          spwbpg          spwbpr          spwbpl     
 Min.   : 1.00   Min.   : 1.00   Min.   : 5.00   Min.   :37.00   Min.   :20.0   Min.   :52.00   Min.   :25.00   Min.   :41.00  
 1st Qu.:13.00   1st Qu.:15.00   1st Qu.:19.00   1st Qu.:66.00   1st Qu.:62.0   1st Qu.:76.00   1st Qu.:68.00   1st Qu.:73.00  
 Median :18.00   Median :17.00   Median :25.00   Median :74.00   Median :71.0   Median :83.00   Median :76.00   Median :80.00  
 Mean   :16.36   Mean   :17.08   Mean   :23.38   Mean   :73.25   Mean   :68.8   Mean   :81.77   Mean   :75.49   Mean   :78.65  
 3rd Qu.:20.00   3rd Qu.:20.50   3rd Qu.:27.50   3rd Qu.:82.00   3rd Qu.:77.5   3rd Qu.:87.00   3rd Qu.:84.50   3rd Qu.:86.00  
 Max.   :24.00   Max.   :24.00   Max.   :35.00   Max.   :98.00   Max.   :96.0   Max.   :98.00   Max.   :98.00   Max.   :98.00  

     spwbsa          sesr            ohq             lot             laplp           lapev           laplc           lapda      
 Min.   :32.00   Min.   :26.00   Min.   : 58.0   Min.   :12.00   Min.   : 9.00   Min.   : 7.00   Min.   :10.0   Min.   : 7.00  
 1st Qu.:67.00   1st Qu.:52.00   1st Qu.:139.0   1st Qu.:26.00   1st Qu.:40.50   1st Qu.:20.00   1st Qu.:29.0   1st Qu.:21.00  
 Median :76.00   Median :58.00   Median :154.0   Median :30.00   Median :47.00   Median :24.00   Median :33.0   Median :25.00  
 Mean   :73.86   Mean   :56.84   Mean   :150.8   Mean   :29.43   Mean   :45.35   Mean   :24.18   Mean   :32.3   Mean   :24.81  
 3rd Qu.:82.00   3rd Qu.:63.00   3rd Qu.:165.0   3rd Qu.:34.00   3rd Qu.:51.00   3rd Qu.:29.00   3rd Qu.:36.0   3rd Qu.:29.00  
 Max.   :98.00   Max.   :70.00   Max.   :194.0   Max.   :42.00   Max.   :63.00   Max.   :48.00   Max.   :42.0   Max.   :42.00  

     lapwm           lapgs           lapfmf          sesc            sasmpl          soimpl          soiml      
 Min.   :11.00   Min.   : 5.00   Min.   : 5.00   Min.   : 50.0   Min.   : 4.00   Min.   :17.00   Min.   :22.00  
 1st Qu.:27.00   1st Qu.:21.00   1st Qu.:24.00   1st Qu.:111.0   1st Qu.:21.00   1st Qu.:48.00   1st Qu.:43.00  
 Median :30.00   Median :23.00   Median :26.00   Median :127.0   Median :24.00   Median :53.00   Median :48.00  
 Mean   :30.19   Mean   :22.83   Mean   :25.74   Mean   :123.9   Mean   :22.46   Mean   :52.49   Mean   :47.69  
 3rd Qu.:34.00   3rd Qu.:24.00   3rd Qu.:28.00   3rd Qu.:138.0   3rd Qu.:25.00   3rd Qu.:58.00   3rd Qu.:53.00  
 Max.   :42.00   Max.   :28.00   Max.   :35.00   Max.   :175.0   Max.   :28.00   Max.   :70.00   Max.   :63.00  

    swbsewb          pwspsy          pwsemo          pwssoc          pwsphys         pwsspirt        pwsintel    
 Min.   :14.00   Min.   :12.00   Min.   :13.00   Min.   :16.0   Min.   :12.00   Min.   :13.00   Min.   :17.00  
 1st Qu.:52.00   1st Qu.:27.00   1st Qu.:29.00   1st Qu.:29.0   1st Qu.:26.00   1st Qu.:28.00   1st Qu.:29.00  
 Median :57.00   Median :31.00   Median :33.00   Median :34.0   Median :31.00   Median :32.00   Median :32.00  
 Mean   :55.61   Mean   :30.68   Mean   :32.12   Mean   :33.3   Mean   :30.15   Mean   :31.84   Mean   :31.61  
 3rd Qu.:61.00   3rd Qu.:34.00   3rd Qu.:36.00   3rd Qu.:37.0   3rd Qu.:35.00   3rd Qu.:36.00   3rd Qu.:35.00  
```

2. You can see below, which data points assign to which group of clustering in fit.cluster column, though it displays the first five observations of the dataset.

```
> head(scaleddata)
         pil       esipb      esiepd     esicos     esiewb        esir        swls     sspwba     spwbem     spwbpg
1  0.32240978  0.92689770 1.12420958 -0.6741390  0.7833381 -0.63283316 -0.39343018  0.1498255 0.33601789 -0.6615809
2 -0.40685044 -1.17136461 1.69792811  1.0120619  0.5683553  0.18984995 -0.39343018 -0.4484385 0.17610186  0.4843551
3  1.05167001  0.22747693 0.74173057 -0.8849142  1.6432692 -0.22149161  1.09651069  0.9190221 1.29551404  0.4843551
4 -0.40685044 -0.00566332 -0.78818551 -1.5172395  0.5683553 -0.83850393 -0.22788119  1.0899546 0.65584994 -1.2345489
5 -0.04222033  1.39317821 -0.02322747 -1.0956893  0.3533725 -1.66118704 -0.06233221  0.9190221 0.49593391  0.3697615
6  0.90581797  0.46061719 0.35925155  0.5905117 -0.2915758 -0.01582083  0.26876576 -0.1065733 0.01618583  0.5989487
       spwbpr      spwbpl      spwbsa       sesr        ohq        lot      laplp      lapev      laplc      lapda      lapwm
1  0.12449812  0.03344807  0.17252278 0.4647152  0.1542058  0.8288939  0.1897736  0.5205176 -0.55899496 -1.0208811 -0.41715982
2  0.37182548 -0.06261790 -0.15030393 0.4647152  0.5374897 -0.6218540  0.4201130  0.1119333 -0.05137821 -0.2708584  2.05424274
3  1.27869249  0.12951405  1.30241627 1.2470837  1.4477888  1.7356113  1.1111312 -1.3862090  0.62544412 -1.1708856 -0.60726771
4  0.12449812 -0.35081583 -0.06959726 0.2411813 -0.1332571 -1.1658844 -0.5012447  0.3843229  0.62544412 -1.3208901 -0.03694404
5 -0.53504152  0.99410784  0.41464281 0.4647152  0.1062953  0.1035200  0.5352827 -0.9776247  0.62544412 -0.5708675  0.15316385
6 -0.04038679 -0.44688181  0.81817620 0.8000160  0.8728630  1.3729243  0.8807918  1.2014914  0.79464970 -0.2708584  2.05424274
       lapgs      lapfmf       sesc     sasmpl     soimpl      soiml    swbsewb      pwspsy      pwsemo      pwssoc     pwsphys
1  0.05203572 -0.19077224 -0.643198082 -0.8731454 -0.4292344 -0.7458343 -0.1869309  0.05611954 -0.3774205 -0.2447325  0.7570445
2  0.35805531  0.06813294 -0.095885464  0.6420788  0.5547183  0.0408749  0.1610988 -0.29930420  0.3342045  0.1307267  0.6009594
3  0.35805531  0.32703813  1.546052388  1.1471535  0.3087301 -1.0080707  1.2051878  1.65552634  1.4016419  1.6325636  0.6009594
4 -0.86602305  0.06813294 -0.543686697 -0.1155333  0.3087301 -0.8769525 -0.1869309 -1.72099914 -0.3774205 -0.8079214 -0.6477217
5  0.35805531  0.32703813  0.899228386  0.3895414  0.3087301  0.8275841  0.2771087  0.41154327 -0.1995143 -0.6201918  0.2887891
6  0.35805531  0.58594332  0.003625921  1.3996909  0.9237006  1.4831751  0.7411482  1.30010261  1.4016419  0.8816452 -1.2720622
    pwsspirt   pwsintel fit.cluster
1  0.2084791 -0.7499279           1
2 -0.1503665 -0.5422685           2
3  0.3879019  1.1190067           2
4 -0.5092121 -1.3729061           1
5 -0.5092121  1.1190067           2
6  0.5673247  1.5343255           2
>
```

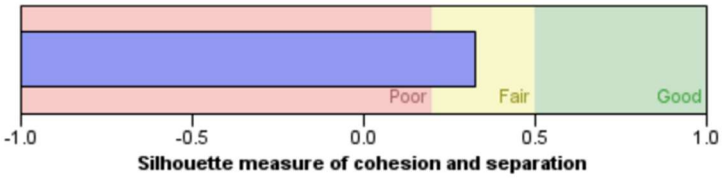
3. Model View Summary for Two-step Clustering in SPSS

TwoStep Cluster

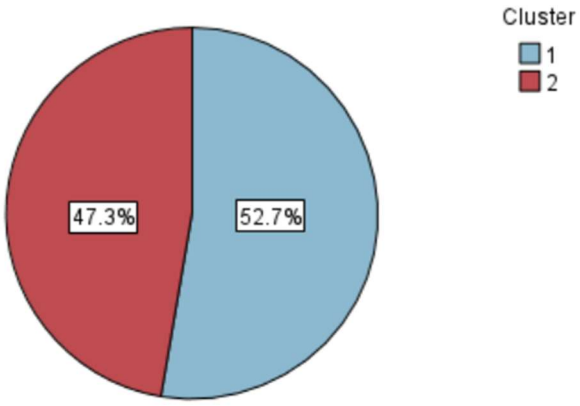
Model Summary

Algorithm	TwoStep
Inputs	39
Clusters	2

Cluster Quality



Cluster Sizes



Size of Smallest Cluster	116 (47.3%)
Size of Largest Cluster	129 (52.7%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1.11

4. Subgroup of clusters

```
> # Number of mem
> table(sub_grp)
sub_grp
  1   2
151  96
>
```

5. Here is a sample of Euclidean distance by variables instead of cases.

Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
247	100.0%	0	0.0%	247	100.0%

Proximity Matrix

Euclidean Distance

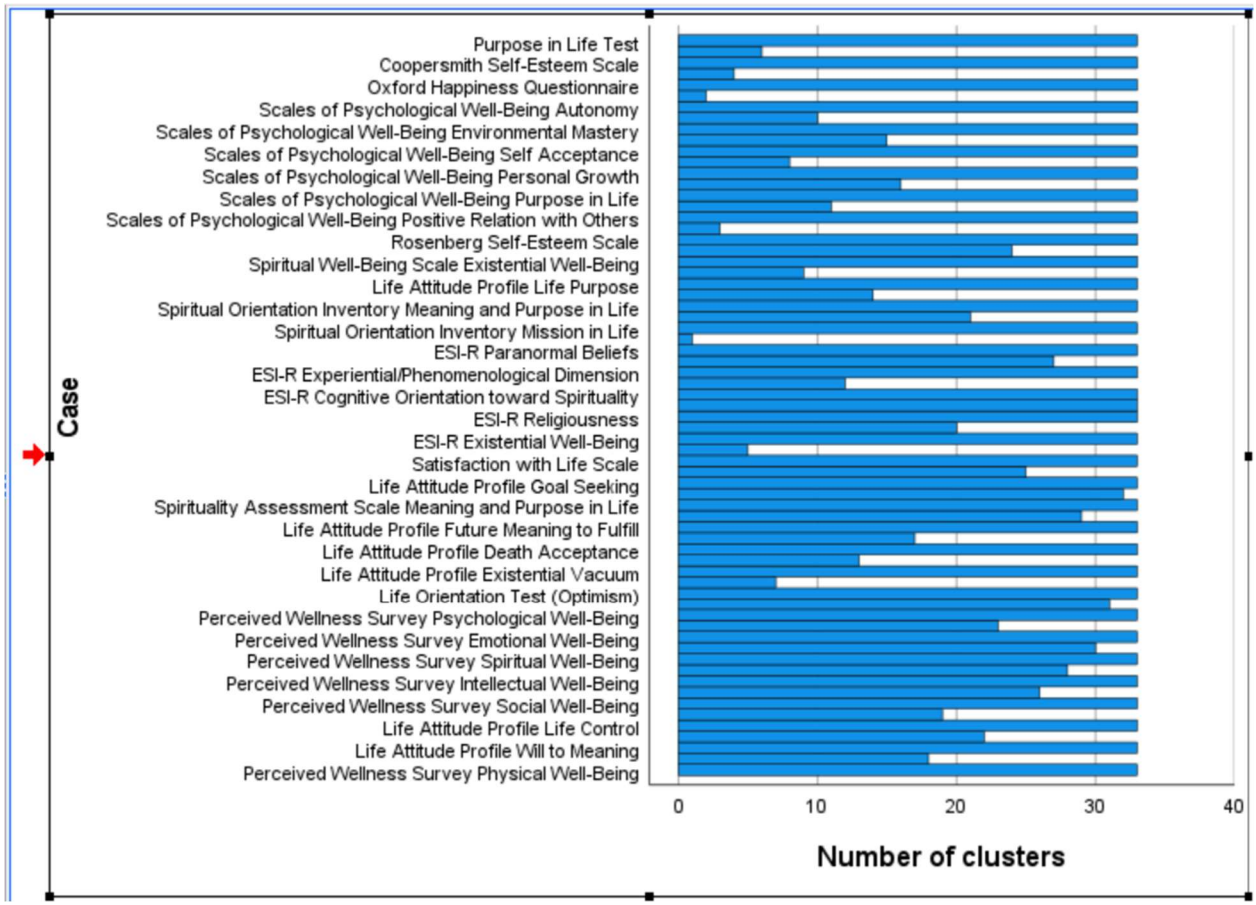
	Purpose in Life Test	ESI-R Paranormal Beliefs	ESI-R Experiential/Phenomenological Dimension	ESI-R Cognitive Orientation toward Spirituality	ESI-R Existential Well-Being	ESI-R Religiousness	Satisfaction with Life Scale
Purpose in Life Test	.000	1550.135	1548.632	1435.831	1445.033	1438.464	1335.248
ESI-R Paranormal Beliefs	1550.135	.000	85.849	148.054	142.092	147.811	242.091
ESI-R Experiential/Phenomenological Dimension	1548.632	85.849	.000	140.307	150.672	144.499	244.422
ESI-R Cognitive Orientation toward Spirituality	1435.831	148.054	140.307	.000	102.489	43.955	151.901
ESI-R Existential Well-Being	1445.033	142.092	150.672	102.489	.000	105.811	143.136
ESI-R Religiousness	1438.464	147.811	144.499	43.955	105.811	.000	157.474
Satisfaction with Life Scale	1335.248	242.091	244.422	151.901	143.136	157.474	.000
Scales of Psychological Well-Being Autonomy	591.403	1011.514	1011.382	902.280	911.488	904.786	807.183

6. Agglomerative Clustering Schedule

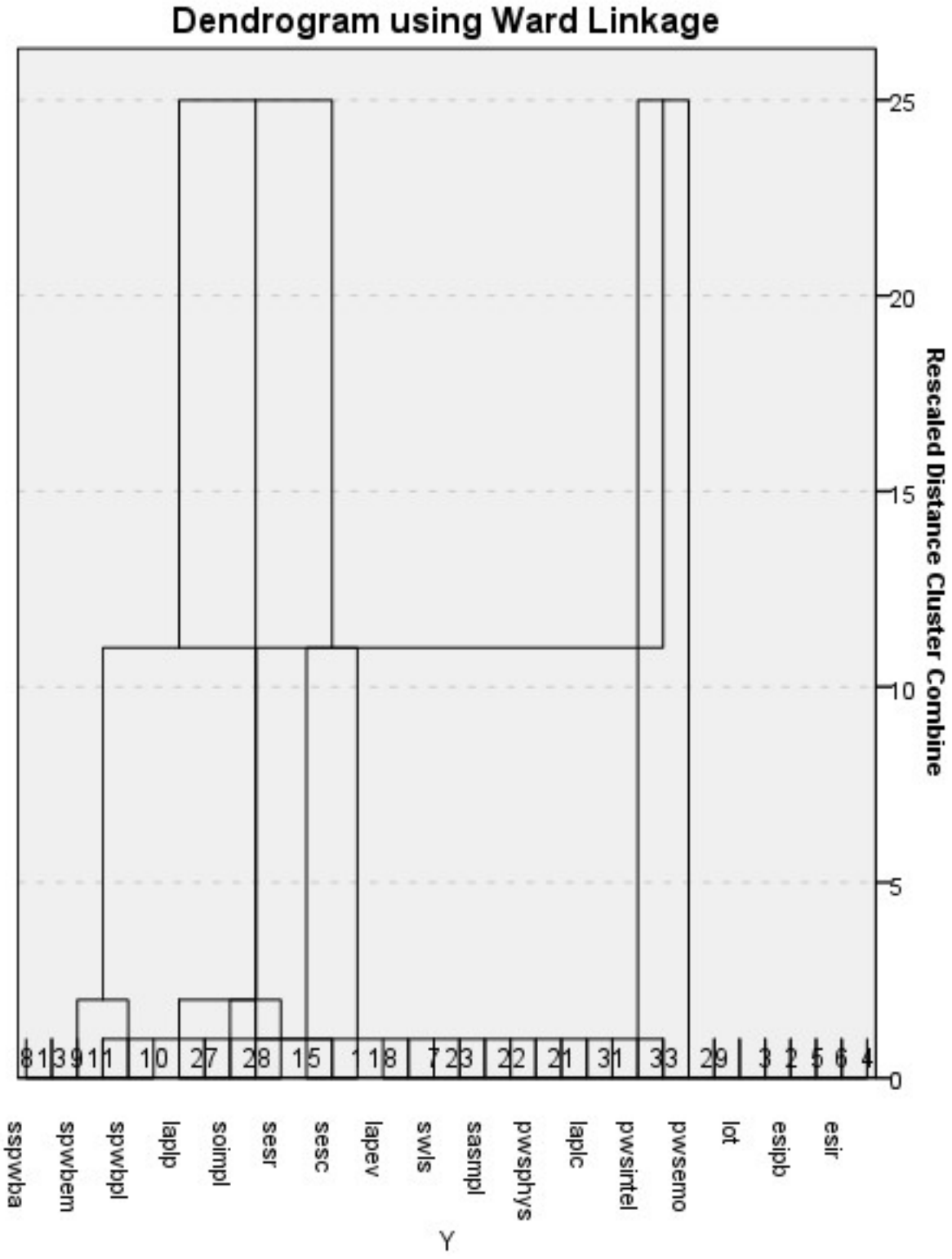
Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	6	966.000	0	0	14
2	22	25	2997.000	0	0	5
3	16	29	5066.000	0	0	11
4	30	33	7795.000	0	0	6
5	22	23	11064.667	2	0	9
6	30	34	14568.333	4	0	8
7	2	3	18253.333	0	0	22
8	30	31	22387.667	6	0	11
9	7	22	28221.250	0	5	17
10	14	28	34087.250	0	0	25
11	16	30	39969.583	3	8	15
12	19	21	46392.083	0	0	15
13	26	27	52845.083	0	0	20
14	4	5	59756.417	1	0	22
15	16	19	67572.458	11	12	16
16	16	32	76178.472	15	0	27
17	7	20	87003.422	9	0	21
18	10	12	98257.922	0	0	23
19	9	13	112600.422	0	0	24
20	17	26	127270.089	0	13	25
21	7	18	142411.222	17	0	27
22	2	4	162494.489	7	14	29
23	10	11	184844.656	18	0	26
24	8	9	210811.489	0	19	26
25	14	17	239742.022	10	20	31
26	8	10	274887.689	24	23	31
27	7	16	333998.600	21	16	29
28	1	24	400350.600	0	0	30
29	2	7	596994.667	22	27	33
30	1	15	823647.333	28	0	32
31	8	14	1222545.558	26	25	32
32	1	8	3587205.300	30	31	33
33	1	2	9449444.971	32	29	0

7. Cases for number of clusters



8. Dendrogram in SPSS


Dendrogram using Ward Linkage
Rescaled Distance Cluster Combine
0
5
10
15
20
25
esir
esipb
lot
pwsemo
pwsintel
laplc
pwsphys
sasmpl
swls
lapev
sesc
sesr
soimpl
laplp
spwbpl
spwbem
sspwba
Y

9. Aggregating the scaling data to get the cluster means

```
> aggregate(scaleddata, by = list(fit$cluster), FUN = mean)
  Group.1         sex      ethnic     marital         pil       esipb      esiepd      esicos      esiewb        esir        swls      sspwba
1       1 -0.03399942  0.02823087  0.02269108  0.4655099 -0.06724755  0.05735773  0.1451001  0.3966395  0.08248722  0.3114545  0.2438922
2       2  0.06143078 -0.05100805 -0.04099866 -0.8410918  0.12150409 -0.10363499 -0.2621695 -0.7166554 -0.14903940 -0.5627416 -0.4406688
      spwbem     spwbpg     spwbpr     spwbpl     spwbsa       sesr        ohq        lot      laplp      lapev      laplc      lapda      lapwm
1  0.4134615  0.3755273  0.3604184  0.5192157  0.5090544  0.4794768  0.4907845  0.4126022  0.4983415 -0.3660332  0.2125399 0.04235863  0.2296853
2 -0.7470497 -0.6785095 -0.6512104 -0.9381283 -0.9197687 -0.8663275 -0.8867583 -0.7454972 -0.9004124  0.6613554 -0.3840210 -0.07653434 -0.4149995
       lapgs      lapfmf       sesc     sasmpl     soimpl      soiml     swbsewb     pwspsy     pwsemo     pwssoc     pwsphys    pwsspirt    pwsintel
1  0.2117818  0.2439931  0.4999311  0.4721323  0.2979004  0.3567131  0.5047507  0.4093079  0.4449763  0.4117308  0.2298890  0.5255722  0.4111363
2 -0.3826512 -0.4408511 -0.9032847 -0.8530572 -0.5382519 -0.6445157 -0.9119928 -0.7395450 -0.8039912 -0.7439226 -0.4153677 -0.9496135 -0.7428486
        age_1
1  0.06367349
2 -0.11504642
>
```

10. D

```
> summary(fit)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust XXX (ellipsoidal multivariate normal) model with 1 component:

 log-likelihood   n  df       BIC       ICL
      -9772.725 247 819 -24057.64 -24057.64

Clustering table:
  1
247
```

11. This picture below is the clustering dendrogram without boxes representing the group of clusters.