MS -Business Intelligence & Analytics

Fall 2015

**BIA – 652 C**

*December 14, 2015*

**Mohit Ravi Ghatikar**

**CWID - 10405877**

Multivariate Data Analytics – HW7

**Ethics Statement**

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment /examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature \_Mohit Ravi Ghatikar\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: 14/01/2015\_\_\_

**Question:**

**16.4**

For the accompanying small, hypothetical data set, plot the data by using methods given in this chapter, and perform both hierarchical and *K*-means clustering with *K* = 2.

Cases *X*1 *X*2

1 11 10

2 8 10

3 9 11

4 5 4

5 3 4

6 8 5

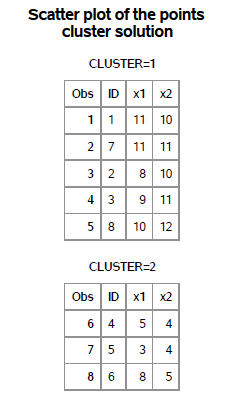
7 11 11

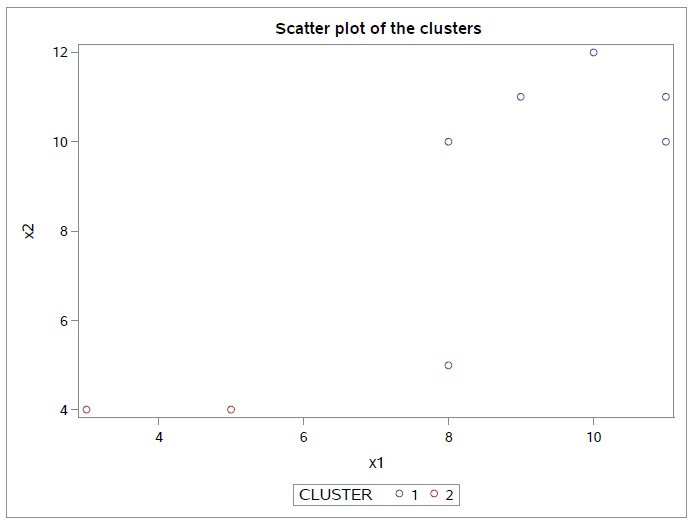
8 10 12

**Solution:**

Hierarchical Clustering:

We run Hierarchical clustering for the dataset containing 7 points. With the condition of nclusters=2, we see that the points 1,7,2,3,8 are in cluster 1 and points 4,5,6 are in cluster 2. This can be shown by the following:





K – Means Clustering:

After running K-means clustering with the value of k=2 and 100 iterations, we arrive at the same result as Hierarchical Clustering. Points 1,7,2,3,8 are in cluster 1 and points 4,5,6 are in cluster 2.

**Question:**

**16.9**

Create a data set from the family lung function data described in Appendix A as follows. It will have three times the number of observations that the original data set has — the first third of the observations will be the mothers’ measurements, the second third those of the fathers, and the final third those of the oldest children. Perform a cluster analysis, first producing three groups and then two groups, using the variables AGE, HEIGHT,WEIGHT, FEV1, and FVC. Do the mothers, fathers, and oldest children cluster on the basis of any of these variables?

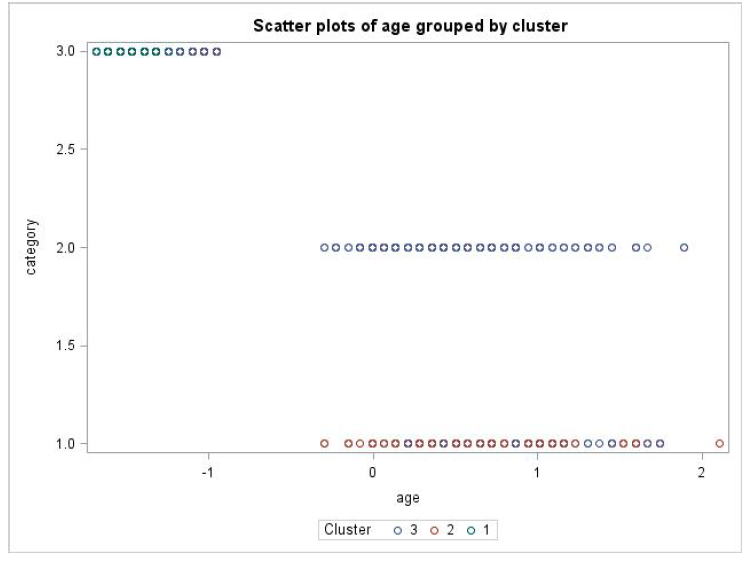
**Solution:**

After creating a new dataset which contains 450 observations, K-means clustering was run on the variables age, weight, height, FVC, FEV1.

The new dataset has an additional column called category to distinguish among father, mother and oldest child. The first 150 observations are data points for father with category=1. Similarly the next 150 observations are data points for mother with category=2 and last 150 observations are data points for oldest child with category=3.

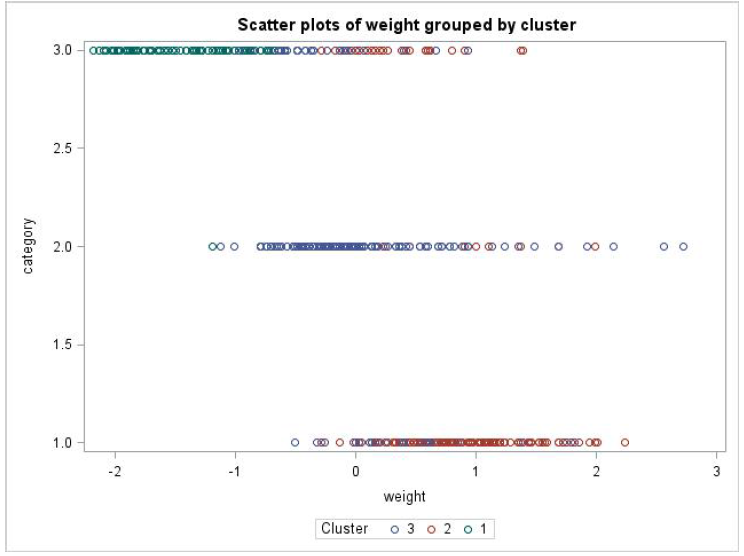
For K=3, we need to know if mothers, fathers and oldest child cluster based on any of these variables. To do this, we run a scatter plot for each variable against its clusters.

Age:



We can see that cluster 2 is predominant for fathers and cluster 3 is predominant for mothers but a mixture of cluster 3 and cluster 1 is present for oldest child. So we cannot say that there is strictly categorized based on age.

Weight:

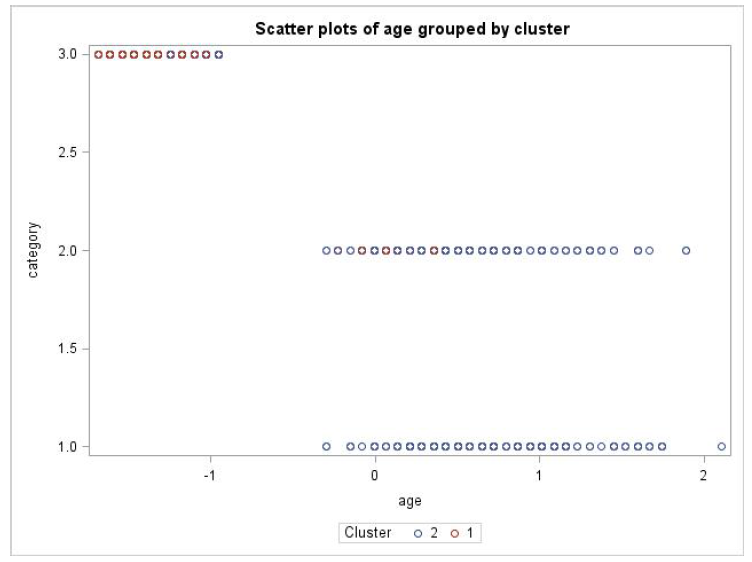


Similarly, for weight variable cluster 2 is predominant for fathers and cluster 3 is predominant for mothers, but for oldest child there is a mix of cluster 1, cluster 2 and cluster 3.

The same pattern is visible for height, FVC and FEV1. Therefore none of the variables can be clustered strictly for mothers, fathers and oldest child.

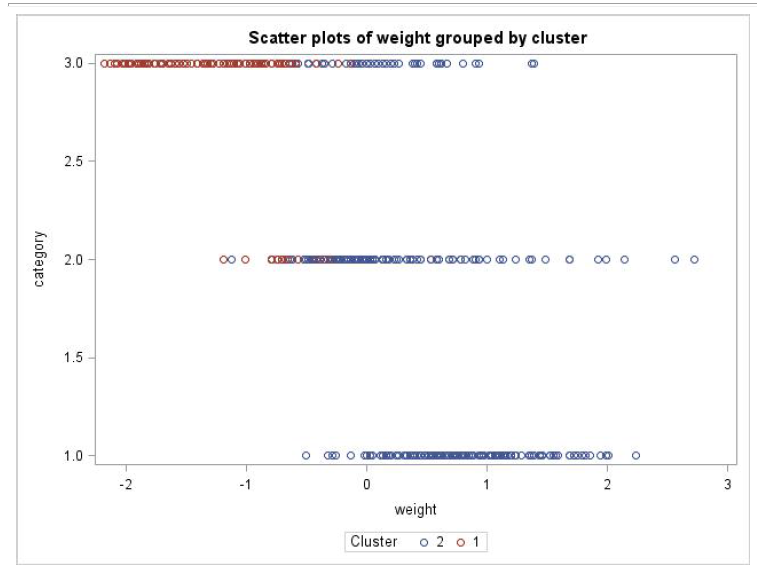
For k=2, same procedure was followed.

Age:



Cluster 2 is predominant for fathers and mothers and cluster 1 is predominant for oldest child with a few exceptions. So we can say age is clustered pretty well for fathers, mothers and oldest child.

Weight:



Similarly, for weight variable, cluster 2 is predominant for fathers and mothers but for oldest child there is a mix of cluster 1 and cluster 2.

The same pattern is followed for height, FVC and FEV1. Therefore we can say that expect for age, the others variables are not clustered strictly into fathers, mothers and oldest child.