Fall 2010

Multivariate Outlier Mining Using Cluster Analysis: Case Study - National Health Interview Survey

Md Monir Hossain Sharker

MULTIVARIATE OUTLIER MINING USING CLUSTER ANALYSIS:

CASE STUDY – NATIONAL HEALTH INTERVIEW SURVEY

A Thesis Submitted to the

McAnulty College & Graduate School of Liberal Arts

Duquesne University

In partial fulfillment of the requirements for

the degree of Master of Science

By

Md. Monir Hossain Sharker

August 2010

MULTIVARIATE OUTLIER MINING USING CLUSTER ANALYSIS:

CASE STUDY – NATIONAL HEALTH INTERVIEW SURVEY

By

Md. Monir Hossain Sharker

Approved August 23, 2010

____
Frank D'Amico, Ph.D.
Professor of Mathematics and Computer Science
[Committee Chair]

____
John Kern, Ph.D.
Associate Professor of Mathematics and Computer Science
[Committee Member]

____
John Fleming, Ph.D.
Assistant Professor of Mathematics and Computer Science
[Committee Member]

____
Donald Simon, Ph.D.
Associate Professor of Computer Science
[Graduate Director]

____
Christopher M. Duncan, Ph.D.
Dean, McAnulty College & Graduate School of Liberal Arts

____
Jeffrey Jackson, Ph.D.
Chair, Mathematics and Computer Science
Professor of Computer Science

ABSTRACT

MULTIVARIATE OUTLIER MINING USING CLUSTER ANALYSIS:

CASE STUDY – NATIONAL HEALTH INTERVIEW SURVEY

By

Md. Monir Hossain Sharker

August 2010

Thesis supervised by Dr. Frank D'Amico

Outlier mining is a fundamental issue in many statistical analyses, especially in multivariate cases. Outliers may exert undue influence on outcomes of the analysis. In most cases, it is a big challenge to reveal the pattern of the outliers and the "outlyingness". There are several approaches and methods to detect anomalous data points in data. But no single method is perfect for every data set especially when the data dimension and volume is high. In this thesis, I review *distance-based clustering* methods for multivariate outlier mining and demonstrate the usefulness of it in a medical setting. Specifically, I discuss *Hierarchical clustering* and the multivariate methods of determining appropriate cluster(s). After mining the multivariate outliers, I examine and describe the characteristics of the variables for those outliers. Finally, I demonstrate the application of these methods using the National Health Interview Survey (NHIS) 2008 database for the purposes of studying adolescent obesity.

ACKNOWLEDGEMENT

I would like to thank my thesis supervisor Dr. Frank D'Amico for his active support and guidance by all means to conduct this thesis. Along with the thesis itself, I have really learned something what, I do believe, will add true values in my future research work. I owe him my wholehearted gratitude.

I express my thanks to the committee members for approving this thesis idea and for their constructive comments. Special thanks go to my Graduate Director, Dr. Donald Simon, for his advices and prompt supports to me throughout this program. I am really grateful to the Chair, all the faculties & stuffs, and fellow students of Math and Computer Science Department to make my experience wonderful here at Duquesne.

Finally, I would like to thank my family for the love and happiness they bring into my life and the moral support to complete this journey. I truly appreciate their help.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

HC	: Hierarchical Clustering
HAC	: Hierarchical Agglomerative Clustering
NHIS	: National Health Interview Survey
CDC	: Center for Disease Control
NCHS	: National Center for Health Statistics
NN	: Nearest Neighbor
LOF	: Local Outlying Factor
BMI	: Body Mass Index
CAPI	: Computer-Assisted Personal Interviewing
SOM	: Self Organizing Map

Chapter 1

INTRODUCTION

Multivariate outlier mining is a key concern in sensitive statistical analysis especially for massive and high dimensional data. Detecting outlier, by sorting the residuals, for univariate case is comparatively easy. But in multivariate case, the residuals are also multivariate. In most of cases, it is a big challenge to reveal the pattern of the outliers and the degree of "outlyingness". So, mining outlier in the multivariate case requires special attention.

Cluster analysis is a process of re-organizing a set of data points (objects) into appropriate number of mutually exclusive *unknown* groups based on combinations of variables and the properties in common among the objects. It is comparatively easier to predict characteristics of objects based on group statistics, where the members share similar properties, as opposed to the individual case. The idea here is that the extreme cases (outliers) should be clustered together. Thus cluster analysis could be used in mining those outliers and reveal their characteristic pattern.

1.1 Specific Objective

There are several clustering approaches (such as *distribution-based, distance-based*, and *depth-based*) to detect anomalous data objects in a database. In this thesis, for multivariate outlier mining, I will first review clustering methods. Specifically, I am interested in *Hierarchical Clustering (HC),* one of the *distance-based* approaches. Second, I will discuss some of the methods presently being used for identifying multivariate outliers. Third, I will show how clustering, in particular HC, can be used to determine outlier(s). Then, I will describe a current database collected from the National Health Interview Survey (NHIS), 2008. The NHIS is a multi-purpose health survey conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). Finally, I will apply HC method and determine multivariate outlier(s) in the NHIS 2008 database. I will also examine if the clusters and the potential outliers have clinical significance in discriminating adolescent obesity.

1.2 Background

There are numerous data mining applications, where identifying exceptions or rare events (outliers) often lead to discovery of important knowledge. Example of such applications are, fraud detection, identifying network intrusions and causes of bottlenecks, criminal activities in E-commerce and/or online transaction, detection of suspicious activities in a database, structural defect detection, disease identification and many more. As a data mining tool, cluster analysis [32] of different approaches can play important role to detect multivariate outliers.

In **distribution-based** approaches [1, 2], data points are modeled using a stochastic distribution and the outliers are observations which deviate from the given distribution. But this is not suitable for moderately high dimension and expensive to determine proper model. In **distance-based** approach, *nearest neighbor (NN)-based* methods define outliers in different ways like a) data points for which there are less than a number of neighboring points within a specified distance [3], b) the top data points whose distance to the k^{th} nearest neighbor is the largest [4], c) data points whose average distance to the k^{th} nearest neighbor is largest [5] etc. But in *density-based* methods a Local Outlying Factor (LOF) [6] is used where the LOF of a sample is the average of the ratios of the density of the sample and the density of its nearest neighbors. The data points with the largest LOF are treated as outlier. **Depth-based** methods works on a quantitative measure, called *depth,* which measures the "degree of centrality" for a data point with respect to a data set [7, 8].

Hierarchical clustering [11] is a clustering method, under distance-based approach, that builds a hierarchy of clusters (of closely related objects). In the hierarchical structure, the hierarchy level increases as the similarity decreases between clusters. The similarity to consider could be measured for the multivariate dataset. So, the idea is that, if I apply hierarchical clustering on the multivariate similarity matrix for a dataset, the outliers in the dataset should be clustered together in one or some top level hierarchy.

1.3 Data and Methods

The database I use in this thesis is collected from the National Health Interview Survey (NHIS), 2008. This is the latest complete version released and publicly available for research. The NHIS is the principal source of information on the health of the civilian, non-institutionalized, household population of the United States. From each household family, in the NHIS 2008 [9], one Sample Child (if any under age 18 present) and one

Sample Adult were randomly selected and detail information in response to predesigned questionnaires on each was collected. The Sample Child and Sample Adult comprising the study data are used for illustration in this thesis.

Using exploratory methods, the Sample Child and Sample Adult databases will be aggregated into one database keeping statistically and clinically important predictor variables with respect to adolescent obesity. Then HC will be applied to each database in order to identify the clusters and potential outliers. Within each database, the candidate outlier cluster(s) will be examined for comparison with the widely used cluster analysis similarity measures (Mahalanobis distance, Jacknife technique, and T^2 statistic). Finally, characteristics of the clusters and the potential outliers will be described with respect to the adolescent obesity.

The definition I use for adolescent obesity throughout the thesis is obtained from physicians who examined the body mass index (bmi), gender of the child and the age for each child. They then took age-specific bmi growth curves and determined for each child in the database whether the child was either “normal”, “overweight” or “obese”. I create a new variable, “status”, to represent the obesity status.

1.4 Thesis Organization

The rest of the thesis is organized as follows. In Chapter 2, I present description of clustering, general classification of clustering algorithms, and description of Hierarchical Agglomerative Clustering (HAC) process with an illustrative example. Chapter 3 explains multivariate outliers, different similarity measures in multivariate outlier analysis, and how HAC could be used as a multivariate outlier mining tool. Chapter 4 is a description of NHIS 2008 database specifically, the dataset used in this thesis. Chapter 5 presents application of Hierarchical Clustering on NHIS 2008 database, analyses, demonstration of results, and discussion. Concluding remarks are given in Chapter 6.

Chapter 2

CLUSTERING

Clustering is an unsupervised learning method which assigns a set of comparatively related data points into subsets (*clusters*). The goal of clustering is to categorize similar (in some sense) data points together and dissimilar or extreme data points separately. The basic assumption in clustering is that objects in the same cluster behave similarly with respect to relevance to information needs [10].

2.1 Types of Clustering

There are several approaches of clustering namely a) distribution-based b) distance-based, and c) depth-based. The distance-based clustering is again classified, in general, as 1) Hierarchical clustering (HC) and 2) Partition-based clustering. There are two types of HC clustering based on the way it is implemented namely i) Agglomerative and ii) Divisive. The Hierarchical Agglomerative Clustering (HAC) is widely used and I will use it as the outlier mining tool in this thesis.

2.2 Hierarchical Clustering

In Hierarchical Clustering [11], agglomerative approach works by successively merging smaller clusters into larger one (bottom up) while the divisive approach works by splitting larger clusters to more related smaller clusters (top down). In either case, a tree structure of clusters called *dendrogram* is produced which shows the relationships among the clusters by means of similarity measure (distance). In HAC dendrogram, each merge of two clusters is represented by a horizontal line. The y-coordinate of the corresponding horizontal line is the similarity (called *combination similarity*) of the two clusters merged. Thus the smaller the distance implies the higher the similarity and vice versa.

2.2.1 Similarity (Distance) Measure

In general, to cluster N data objects, the agglomerative method works based on an NxN distance (similarity) matrix. Two commonly used methods to measure the distance (similarity), for any two K dimensional data vectors (**X**, **Y**) are explained below.

(1) **Euclidean distance**: It is simply the geometric distance in multidimensional space. So the Euclidian distance d(**X,Y**) between two vectors **X, Y** each having dimension K, is computed as:

$$\mathrm{d(X,Y)} = \sqrt{\sum_{\mathrm{i=1}}^{\mathrm{K}} \left(\mathrm{X_i} - Y_i\right)^2}$$

(2) **Squared Euclidean distance**: It is simply the square of the Euclidean distance. This measure is used in order to get greater weight on objects that are further apart. In the above case, the squared Euclidian distance is hence found as;

$$\mathrm{d(X,Y)} = \sum_{\mathrm{i=1}}^{\mathrm{K}} \left(\mathrm{X_i} - Y_i\right)^2$$

Usually, Euclidean (and squared Euclidean) distances are computed for data as is, not for standardized data. Here, the distance calculated between any two objects is not affected by introducing new objects into the scene. However, the resultant distances can be greatly affected by differences in scale among the dimensions. For example, if one of the dimensions states a measured weight in pounds (lb) and in distance calculation the measure is converted to ounces (lb x 16) then the final distance calculated from multiple dimensions along with this weight can be biased by the dimensions having larger scale. Consequently, the results of cluster analyses may vary a lot. So, before distance calculation the dimensions should be transformed into as similar scales as possible. In the continuous case, standardized units are often used.

2.2.2 HAC Algorithm

Assumption:

a) The merge operation is *monotonic*. That means, if d_1, d_2, d_3, ……., d_{N-1} are the distances of successive merges for a HAC to cluster N objects then,

$$d_1 \leq d_2 \leq d_3 \ldots\ldots\ldots\ldots\ldots\ldots, \leq d_{N-1}$$

b) The distances (similarities) between the clusters are assumed as the distances (similarities) between the objects of the clusters.

Having the NxN distance matrix, the algorithm steps work as follows;

1. Assign each item to a cluster initially (N clusters having 1 item each).
2. Merge the closest (most similar) cluster pair into a single cluster.

3. Re-compute distances (similarities) between the new cluster and each of the old clusters and update the distance matrix.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Step 2 and 3 are the key steps in the process. They extract the most similar (closest) cluster pair and merge them into a single cluster. Here the linkage rules come into effect.

2.2.3 Linkage rules

There are different ways (linkages) to find the closeness, such as Single-Linkage, Complete-Linkage, Average-Linkage. [Figure 2.1(a)-(c)].

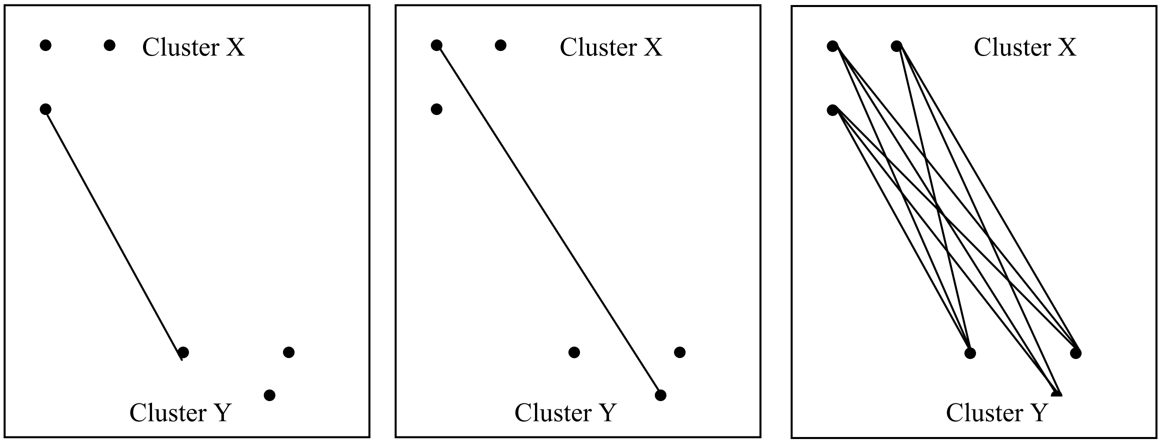


Figure 2.1(a): Single-Linkage Figure 2.1(b): Complete-Linkage Figure 2.1(c): Average-Linkage

In *single-linkage* clustering [12, 13, and later reinvented by [14] and [15]), the distance between clusters is found as the *shortest distance* between any member of one cluster to any member of the other cluster. That is,

$$\mathrm{d(X,Y)} = Min\{d(x,y); x \in X, y \in Y\}$$

In *complete-linkage* clustering, the distance between clusters is found as the *greatest distance* between any member of one cluster to any member of the other cluster (Milligan 1980). That is,

$$\mathrm{d(X, Y)} = Max\{d(x, y); x \in X, y \in Y\}$$

In *average-linkage* clustering, the distance is calculated as the average distance from any member of one cluster to any member of the other cluster. Average linkage tends to join clusters with small variances and is slightly biased toward producing clusters with the same variance [16]. The distance here is calculated as;

$$\mathrm{d(X, Y)} = \frac{1}{N_x \times N_Y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} d(x_i, y_i); x_i \in X, y_i \in Y$$

In this thesis, I have used Single-Linkage HAC algorithm [11, 14] as shown in Figure 2.2. The time complexity of single-linkage, complete-linkage, and average-linkage methods are $\Theta(N^2)$, $\Theta(N^2 \log N)$, $\Theta(N^2 \log N)$ respectively, where N is the sample size. The performance of using single linkage method is better [10] than the others. The reason for the difference in time complexity between single-linkage and complete-linkage is that, distance defined as the distance of the two closest members (single-linkage case) is a *local property* that is not affected by merging; distance defined as the diameter of a cluster (complete-linkage case) is a *non-local property* that can change during merging.

Single-LinkageHAC(D)

The $N \times N$distance matrix (*D)* contains all distances $d(i,j)$. The clusters are assigned sequence numbers 0,1,......, ($N - 1$) and $L(k)$ is the level of the kth clustering. A cluster with sequence number *m* is denoted (*m*) and the distance between clusters (x) and (y) is denoted *d*[(x),(y)].

1. Start with the disjoint clusters having level $L(0) = 0$ and sequence number m = 0.
2. Find the most similar pair of clusters in the current clustering, say pair (x), (y), according to *d*[(x),(y)] = Min *d*[(*i*),(*j*)] where the minimum is over all pairs of clusters in the current clustering.
3. Increment the sequence number: $m = m + 1$. Merge clusters (x) and (y) into a single cluster to form the next clustering *m*. Set the level of this clustering to $L(m)$ = *d*[(x),(y)]
4. Update the proximity matrix, *D*, by deleting the rows and columns corresponding to clusters (x) and (y) and adding a row and column corresponding to the newly formed cluster (x, y). The proximity between the new cluster, denoted (x,y) and old cluster (*k*) is defined as *d*[(*k*), (*x*, y)] = Min {*d*[(*k*),(x)], *d*[(*k*),(y)]}.
5. If all objects are in one cluster, stop. Else, repeat from step 2.

Figure 2.2: Single-Linkage Hierarchical Agglomerative Clustering algorithm.

2.3 How Many Clusters?

It is not necessary for HAC to pre-specify the number of clusters to generate. However, in some applications, clearly disjoint clusters may need to be separated. If so, the hierarchy (dendrogram) needs to be cut at some point of interest. There are several criterions [17] based on what cutting point could be determined:

a) The rule of thumb[18] in determining number of clusters K out of N observations is set as

$$K \approx \sqrt{\frac{N}{2}}$$

b) Set a threshold distance above which clusters should be separated and cut the dendrogram at that level of distance (similarity). The distance specification depends on the type of objects clustered. The higher the distance, the lower the number of cluster.

c) Choose the dendrogram cutting point such that the gap between two successive combination similarities is the largest. That is, cut the dendrogram where the distance between clusters starts to increase sharply. Adding more cluster decreases the quality of the clustering significantly at this point.

d) Educated guess. More about clustering validity methods is explained in [17].

2.4 Illustrative Example

I pick a simple example to demonstrate the Single-Linkage HAC application in clustering. Suppose we have the percent marks obtained by six(6) students of MathCS department in two different subjects Math and CS as listed in Table 2.1. We first calculate the distance matrix for the table and then apply the HAC algorithm to the distance matrix to cluster them and generate a dendrogram.

Table 2.1: MathCS student evaluation table

Name	**Math**	**CS**
Julie	90	90
John	92	94
Ryan	65	70
Bob	30	40
Ted	70	75
Sara	85	92

2.4.1 Distance Matrix Calculation

For a single variable, Math, the distance matrix (D_1) is calculated (shown in Table 2.2) using squared Euclidian distance measure. For example, distance between John and Bob is calculated as

$$d((\text{John}),(\text{Bob})) = (92-30)^2 = 3844$$

Here, distance for any student to itself is 0 as shown diagonally. Also, the upper part of the diagonal is just a mirror image of the lower.

Table 2.2: Distance matrix (D_1) for Math only

	Julie	John	Ryan	Bob	Ted	Sara
Julie	0	4	625	3600	400	25
John	4	0	729	3844	484	49
Ryan	625	729	0	1225	25	400
Bob	3600	3844	1225	0	1600	3025
Ted	400	484	25	1600	0	225
Sara	25	49	400	3025	225	0

Similarly, the distance matrix (D_2) for the CS variable is found as shown in Table 2.3.

Table 2.3: Distance matrix (D_2) for CS only

	Julie	John	Ryan	Bob	Ted	Sara
Julie	0	16	400	2500	225	4
John	16	0	576	2916	361	4
Ryan	400	576	0	900	25	484
Bob	2500	2916	900	0	1225	2704
Ted	225	361	25	1225	0	289
Sara	4	4	484	2704	289	0

For univariate clustering, I can apply HAC to any of the tables of interest. But I am interested in multivariate analysis. To do that, the distance matrix should be calculated using all the variables of interest (here Math and CS). So, to calculate the multivariate distance matrix D, I use the squared Euclidian distance calculation equation for multiple variables. In other words, I add all the single variable distance matrices. In this case, adding Table 2.2 and Table 2.3, I can find D (= D_1+ D_2) as shown in Table 2.4.

Table 2.4: Distance matrix (D) for Math and CS

	Julie	John	Ryan	Bob	Ted	Sara
Julie	0	20	1025	6100	625	29
John	20	0	1305	6760	845	53
Ryan	1025	1305	0	2125	50	884
Bob	6100	6760	2125	0	2825	5729
Ted	625	845	50	2825	0	514
Sara	29	53	884	5729	514	0

2.4.2 Clustering Process

I have the input distance matrix (6x6) evaluating 6 students of MathCS department in two subjects as shown in Table 2.4. I use hierarchical clustering method on it using single-linkage. I have 6 objects (students) and I assign each object into one cluster. So I have 6 clusters initially. According to the algorithm, these 6 clusters are grouped such that at the end of the iterations, it will produce one cluster consisting of the six objects. In each iteration, the closest distant pair is clustered. In the example, the closest clusters are Julie and John with the shortest distance of 20 between them as shown in Table 2.4. Cluster (Julie) and (John) are taken into cluster (Julie, John). The distance matrix is updated treating (Julie, John) as a single cluster in the matrix as shown in Table 2.5.

Table 2.5: Updated Distance matrix after merging (Julie) and (John)

	Julie, John	Ryan	Bob	Ted	Sara
Julie, John	0	1025	6100	625	29
Ryan	1025	0	2125	50	884
Bob	6100	2125	0	2825	5729
Ted	625	50	2825	0	514
Sara	29	884	5729	514	0

Distance among clusters not merged in the current iteration will not be changed in the original distance matrix. Now the concern is to calculate distances among the new cluster (Julie, John) and others. Here is exactly where the linkage rules come into play. Using single-linkage, minimum distance between original objects of the two clusters is determined. Using the input distance matrix, distance between cluster (Julie, John) and cluster (Ryan) is computed as

$$d((Joulie, John), Ryan) = Min(d(Julie, Ryan), d(John, Ryan)) = Min(1025,1305) = 1025$$

Distance between cluster (Julie, John) and cluster (Bob) is

$$d((Joulie, John), Bob) = Min(d(Julie, Bob), d(John, Bob)) = Min(6100,6760) = 6100$$

Distance between cluster (Julie, John) and cluster (Ted) is

$$d((Joulie, John), Ted) = Min(d(Julie, Ted), d(John, Ted)) = Min(625,845) = 625$$

Similarly, distance between cluster (Julie, John) and cluster (Sara) is

$$d((Joulie, John), Sara) = Min(d(Julie, Sara), d(John, Sara)) = Min(29,53) = 29$$

Now the next nearest pair of clusters are (Julie, John) and (Sara) having minimum distance (=29) as highlighted in Table 2.5. So merging them together and updating the distances accordingly, I have the new distance matrix as shown in Table 2.6.

Table 2.6: Updated Distance matrix after merging (Julie, John) and (Sara)

	(Julie, John), Sara	Ryan	Bob	Ted
(Julie, John), Sara	0	884	5729	514
Ryan	884	0	2125	50
Bob	5729	2125	0	2825
Ted	514	50	2825	0

Similarly merging (Ryan) an (Ted) who are the closest (distance=50) as shown in Table 2.6, I have the distance matrix as shown in Table 2.7.

Table 2.7: Updated Distance matrix after merging (Ryan) and (Ted)

	(Julie, John), Sara	Ryan, Ted	Bob
(Julie, John), Sara	0	514	5729
Ryan, Ted	514	0	2125
Bob	5729	2125	0

Merging the next closest cluster ((Julie, John), Sara) and (Ryan, Ted) I get the new distance matrix as in Table 2.8.

Table 2.8: Updated Distance matrix after merging (Julie,John),Sara) and (Ryan,Ted)

	(Julie, John), Sara), (Ryan, Ted)	Bob
(Julie, John), Sara), (Ryan, Ted)	0	2125
Bob	2125	0

Finally, merging the remaining two clusters ((Julie, John), Sara), (Ryan, Ted)) and (Bob) I get the single cluster (((Julie, John), Sara), (Ryan, Ted)), Bob) as in Table 2.9 which contains all the 6 objects (students). Thus, the clustering is completed.

Table 2.9: Final matrix after merging ((Julie, John), Sara), (Ryan, Ted)) and Bob

	(Julie, John), Sara), (Ryan, Ted), Bob
(Julie, John), Sara), (Ryan, Ted), Bob	0

2.4.3 Dendrogram

A dendrogram is a tree diagram that lists each observation, and shows which cluster it is in and when it entered into its cluster. The dendrogram for the above example is shown in Figure 2.3.

From the dendrogram, k clusters can be found by cutting it in k-1 level from the top. Here it is noticeable that there are 3 (cutting the dendrogram in level 2) distinct clusters in the MthCS students. This can be further seen from examining the clustering history (Figure 2.4). Here the distance is Euclidian distance.

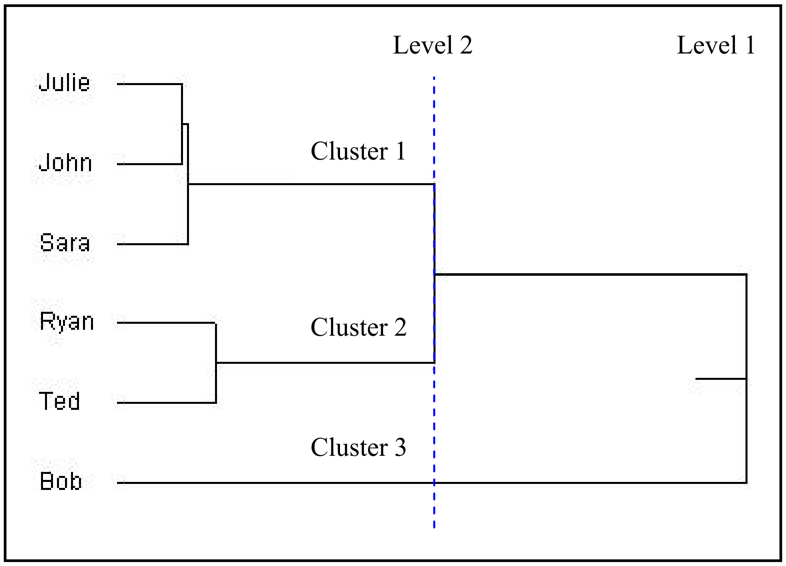
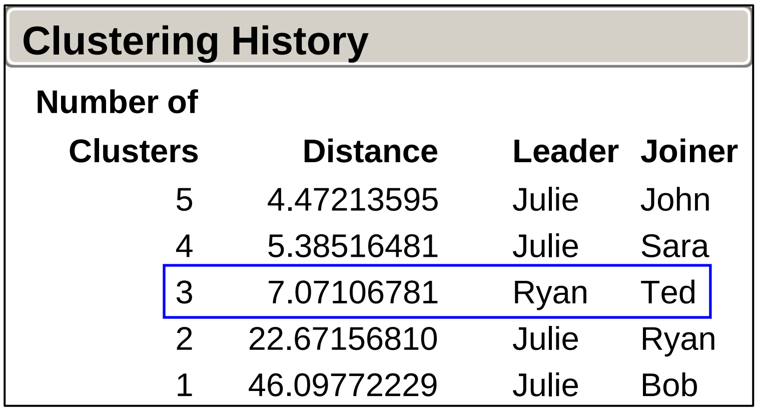


Figure 2.3: Dendrogram for the MathCS student clusters



Clustering History

Number of Clusters	Distance	Leader	Joiner
5	4.47213595	Julie	John
4	5.38516481	Julie	Sara
3	7.07106781	Ryan	Ted
2	22.67156810	Julie	Ryan
1	46.09772229	Julie	Bob

Figure 2.4: Clustering history of MathCS students

According to the history, when number of clusters changes from 3 to 2; that is, when the cluster ((Julie, John), Sara) is merged with (Ryan, Ted) cluster, there is a sharp change (as shown in Figure 2.5) in distance (7.07 to 22.67). So it is a good level to cut the dendrogram at distance 7.07. Thus three (3) distinct clusters are found.

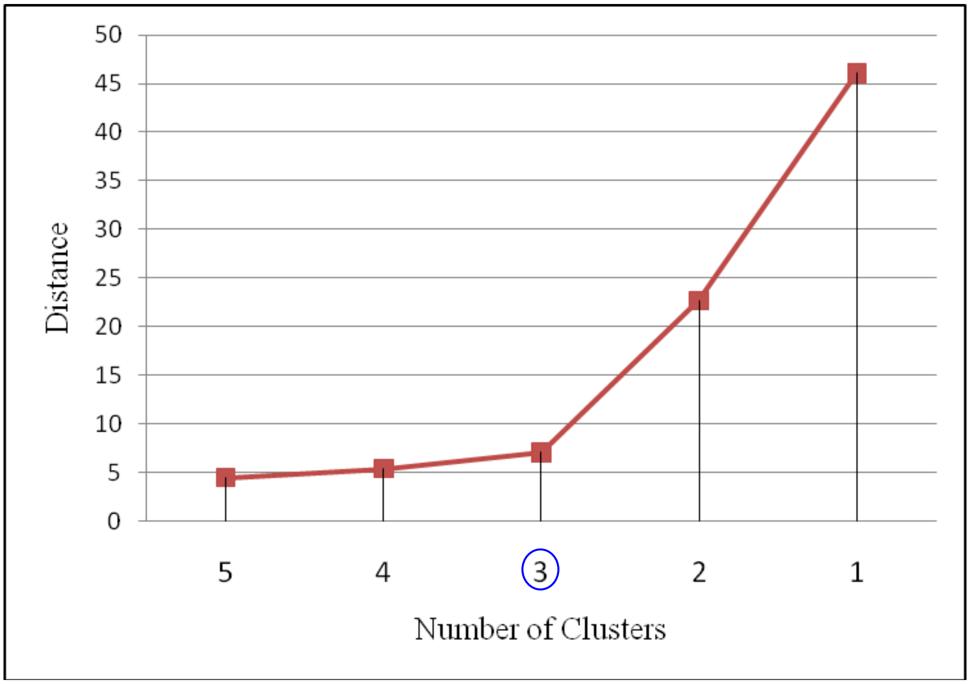


Figure 2.5: Change in distance vs number of clusters

Chapter 3

DETECTION OF MULTIVARIATE OUTLIERS

There are several definitions of outlier. It depends on the types and characteristics of data objects. Some of the common definitions are given here. "*An observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*" [2]. "*An observation which appears to be inconsistent with the remainder of that set of data*" [1]. Outliers in a database could be real value (extreme response) or it could be generated from data error, missing values or imported from a different population. Outliers can be an individual data point or a group of objects exhibiting considerable "outlyingness".

A value may seem to be normal in univariate consideration but outlier from multivariate points of view. For example, a child's weight of 65 lb may seem to be normal where the weight range is 10 -100 lb. But it would certainly be an outlier if it is for a 2 years old (considering weight with age). No matter what and how it is appeared, outliers may affect with undue influence on the analysis result. Outliers should be detected and taken proper care of using reliable method(s) before performing analysis. There are several approaches and methods to detect multivariate outliers. I use the distance-based approach, in particular hierarchical clustering method to identify potential candidates for multivariate outliers.

3.1 Distance-Based Definitions of Outlier and Methods

In general, distance-based methods treat an object as outlier if it is at least at a minimum distance away from a set percentage of objects in the dataset. This process usually needs detail domain knowledge [19]. According to [3], "*A point x in a dataset is an outlier with respect to the parameters k and d, if no more than k points in the dataset are at a distance d or less from x*". [4] defines outliers for high dimensional data where user does not required to specify the distance parameter. It is based on the distance of the Kth nearest neighbor of a point. An outlier score function is defined to measure the extremeness of a data point in the outlier-based association method. The more extreme a data point is, the higher outlier score it gets [20]. Also, there are various distance based similarity measure tools, such as *Mahalanobis distance*, *Jackknife technique*, and the *T^2 statistic,* which are widely used in multivariate outlier detection and analysis.

3.1.1 Mahalanobis Outlier Distance

In this outlier detection method, the Mahalanobis distance [21] of each point from the multivariate mean (centroid) is plotted. The Mahalanobis distance between two multidimensional vectors X, Y is found as,

$$\mathrm{D(X,Y)} = \sqrt{\left(\mathrm{X} - Y\right)^T S^{-1} \left(\mathrm{X} \;\; - Y\right)}$$ where, S is the covariance matrix.

The Mahalanobis distance is a **metric** (distance between two data points) which is better adapted than the usual Euclidian distance. The standard Mahalanobis distance depends on estimates of the mean, standard deviation, and correlation for the data. The distance is plotted for each observation number. Extreme multivariate outliers can be identified by highlighting the points with large distances.

3.1.2 Jackknife Technique

In **Jackknife technique** [22, 23], the distance for each observation is calculated with estimates of the mean, standard deviation, and correlation matrix that **does not** include the observation itself. It provides an alternative and robust method for determining the propagation of error from the data to the parameters.

Let there are N data points. Jackknife technique starts with N-1 re-sampled values. Suppose $X_{\mathbf{J}1}$ is the measured parameter for first the N-1 samples. Then a new re-sampling is done for another N-1 values sampled from N and may be the parameter this time is found as $X_{\mathbf{J}2.}$ In this process N parameter values ($X_{\mathbf{J}i}$, i = 1, 2, 3, ….., N) are found. Then the standard error is given by;

$$\sigma^2_{Jmean} = (N-1)\sum_{i=1}^{N}(X_{Ji} - \overline{X})^2 / N$$

The Jackknife distance can be calculated as,

$$\mathrm{D(X,Y)} = 1 - Min(\rho_{xy}^{(1)}, \rho_{xy}^{(2)}, \rho_{xy}^{(3)}, \ldots\ldots., \rho_{xy}^{(N)}), \textit{Jackknife Corelation}$$

$$\mathrm{D(X,Y)} = Min(d_{xy}^{(1)}, d_{xy}^{(2)}, d_{xy}^{(3)}, \ldots\ldots., d_{xy}^{(N)}), \textit{Jackknife Euclidian}$$

The jackknifed distances are useful when there is an outlier. In this case, the Mahalanobis distance is distorted and tends to disguise the outlier or make other points look more outlying than they are.

3.1.3 T^2 statistic

Hotelling's T^2 statistic [24] is a generalization of Student’s t-statistic that is used in multivariate hypothesis testing. T^2 statistic is simply the square of the Mahalanobis distance, that is

$$t^2 \approx (X - Y)^T S^{-1} (X \quad - Y)$$

It is preferred for multivariate control charts. The plot includes the value of the calculated T^2 statistic, as well as its upper control limit. Values that fall outside this limit are potential candidates for outlier.

3.2 HAC as a Multivariate Outlier Detection Tool

In **clustering-based** methods, the key assumption is that the normal data objects mostly belong to large and dense clusters, while outliers forms small distantly related cluster or even do not belong to any of the clusters. These methods cluster data into groups of different data density, takes points in small cluster as candidate outliers, and compute the distance between candidate points to other clusters. If candidate points are far from all other non-candidate points, they are treated as outliers. I explore hierarchical clustering to determine multivariate outlier(s) and cross-verify those outliers using other methods.

Hierarchical clustering method clusters data objects such that a cluster hierarchy is related to the distance (similarity) with the other clusters. So, the higher the hierarchy, lower the similarity. The highest distant cluster would be treated as the first candidate to be outlier. The data object that clustered last will have the largest distance and hence it is the top candidate for outlier. Usually this is found (if any) by cutting the dendrogram so that only two clusters are found containing a single data point in one and all other data points into the other. This single data point is the most potential candidate for outlier.

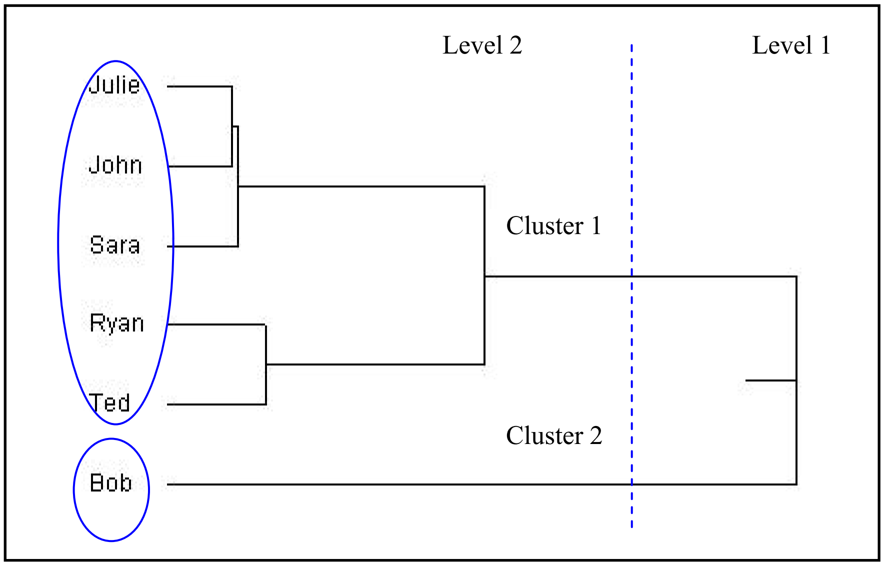


Figure 3.1: Showing outlier candidate (Bob here) in dendrogram

For example, in the MathCS student evaluation example, Bob as shown in Figure 3.1 is clustered as the last student and the distance between Bob and the cluster of all other students is the highest (46.09). So, Bob is a potential outlier in the database.

There are some cases possible where identifying the last element clustered is difficult. Because, may be there is no single data object found while cutting the resultant dendrogram into 2 clusters. In such cases, what happens is that, all the data points are already clustered into some sub-clusters with some other data points and later the final cluster is formed from those sub-clusters. In this scenario, the lastly clustered data object could be found out of a big cluster applying the procedures as shown in Figure 3.2. In this process, the first cluster found with single element in it is candidate for outlier. I can further verify this element using other methods if it is truly outlier or not. But it is also possible that a smaller cluster itself is a cluster of outliers.

```
Proc_HAC_Outlier(Dendrogram)
a) Cut the dendrogram so that only 2 clusters are found.
b) If the number of elements in any cluster is 1 STOP
           The potential outlier is this element
   Else, increase the number of clusters by 1 cutting the
   dendrogram in the next level and repeat step b)
c) STOP
```

Figure 3.2: Procedure to determine the most potential candidate in a dendrogram

Chapter 4

CASE STUDY: NATIONAL HEALTH INTERVIEW SURVEY

The National Health Interview Survey (NHIS) [9] is a multi-purpose health survey conducted by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). The NHIS is the principal source of health related data for civilian, non-institutionalized, household population of the United States. The survey has been being conducted since 1957. Microdata files for public use are released annually. The latest database available publicly is NHIS 2008 and that is the target database in this thesis.

Currently NHIS consists of a Basic Module and Core Module as well as variable Supplements. The Core Module consists of three components: 1) the Family Core, 2) the Sample Child Core, and 3) the Sample Adult Core. The Family Core takes information on everyone in the family in each household. One Sample Child (if any children under age 18 are present) and one Sample Adult are randomly selected, and information on each is collected with the Sample Child Core and the Sample Adult Core questionnaires. Because some health issues are different for children and adults, these two questionnaires differ in some items, but both collect basic information on health status, health care services, and behavior. In this thesis, the Sample Child and the Sample Adult data files are investigated.

4.1 Data Collection Procedures

The NHIS 2008 data was collected by the U.S. Census Bureau, as a data collection agent. Census interviewers collected the data through a personal household interview. About 600 interviewers were trained and directed by health survey supervisors in the 12 U.S. Census Bureau Regional Offices. The supervisors are career Civil Service employees selected through an examination and testing process. The NHIS provides training to interviewers annually in basic interviewing concepts and procedures.

A computer-assisted personal interviewing (CAPI) method was used by the interviewer for data collection. The CAPI presents the questionnaire on computer screens and guides the interviewer through the questionnaire. It automatically directs the interviewer to next appropriate questions based on answers entered to previous questions. Interviewers enter survey responses directly into the computer, and the CAPI program validates and saves the responses into a survey data file.

Response was provided by a knowledgeable adult member (18 years or older) residing in the household for children and for adults not present. For the Sample Child questionnaire, a knowledgeable adult residing in the household provided the information. And for the Sample Adult questionnaire, one adult for each family was randomly selected to response. If he/she was physically or mentally unable; a knowledgeable person was allowed to respond for the Sample Adult as proxy. I collect the data for this thesis from the NHIS data source [9] available publicly in different formats. Also a data retrieval tool, provided with the data files, is used to retrieve data in compatible format.

4.2 Data Details

The facts and figures about the NHIS 2008 data are given in Table 4.1.

Table 4.1: NHIS 2008 database at a glance

Entity	**Quantity**
Total number of households	28,790
Total number of families	29,421
Total number of persons	74,236
Total number of eligible sample children	10,303
Sample Child (age 12 to 17 years)	**8,815**
Total number of eligible Sample Adult	29,370
Sample Adults (18 years or older)	**21,781**
# Cases knowledgeable proxy answered for Sample Adult	257

In this thesis, I use Sample Child and the Sample Adult data set from the NHIS 2008. In the Sample Child data set, there are 8815 records and 195 variables. The Sample Adult data set has 21,781 records and 980 variables. I merge these two data sets using their house hold identifier to have a single dataset from where I can analyze the adolescent obesity frequencies and characterize the clusters and outliers.

4.3 Data Preprocessing and Variable Selection

In general, there are two major types of variables in both Sample Child and Sample Adult datasets, namely 1) Demographic variables and 2) Health related variables. To find the important variables and extract relevant records in each category, I use exploratory analysis for each variable. For example, in Sample Child dataset examining the age distribution and the BMI (recoded) distribution as shown in Table 4.2 and Table 4.3 respectively, I select only those children having BMI reported. Here there are 2716 such children and all of them are of age 12 to 17 years. Also, number of children within age range 12 to 17 years having Sample Adult relationship reported is 2488 as shown in Table 4.4. Again, I filter out those children with missing Sample Adult relationship. Finally I end up with 2168 children of age range 12 to 17 years having both BMI and Sample Adult relationship reported as shown in the Table 4.5.

.

Table 4.2: Distribution of Age for Child

Age(Yr)	Frequency	Percent
<1	480	5.4
1	520	5.9
2	514	5.8
3	488	5.5
4	485	5.5
5	490	5.6
6	404	4.6
7	463	5.3
8	451	5.1
9	445	5.0
10	446	5.1
11	444	5.0
12	461	5.2
13	505	5.7
14	514	5.8
15	568	6.4
16	569	6.5
17	568	6.4
Total	8815	100.0

Table 4.3: Status (Recoded BMI) Distribution for child

		Frequency	Percent
Valid	Normal	1743	19.8
	Overweight	516	5.9
	Obese	457	5.2
	Total	2716	30.8
Missing	System	6099	69.2
Total		8815	100.0

Table 4.4: Children(age 12 to 17) having adult relationship reported

		Frequency	Percent
Valid	12	360	14.5
	13	402	16.2
	14	398	16.0
	15	448	18.0
	16	456	18.3
	17	424	17.0
	Total	2488	100.0

Table 4.5: Status * Age cross-tabulation for those having adult relationship reported

Count								
		Age						Total
		12	13	14	15	16	17	
status	Normal	183	216	245	279	309	141	1373
	Overweight	71	74	75	76	73	50	419
	Obese	76	82	57	72	61	28	376
Total		330	372	377	427	443	219	2168

4.4 Final Variables

Following different preprocessing strategies and methods such as contingency tables, partition techniques and classification procedures in several steps, variables of interests in Sample Child dataset are filtered down as shown in Table 4.6.

Table 4.6: Final clustering variables for Sample Child

	Variable Name	**Description**	**Recoded values**
Demographic Variables	age	Age	0 - Under 1 Yr, 85 - 85+ yrs
	sex	Sex	1- Male, 2- Female
Health Related Variable	place	Place to go if sick recoded - for child	1- Doctor's office or HMO, 2- others
	days_missed	Days missed due to illness/injury, past 12 m	1- 3 or less, 2- more than 3/others
	health	Health better, worse, or about the same	1-Better, 2-About the same/others, 3-Worse
	asthma	Ever been told Sample Child had asthma recoded - for child	1-Yes, 2-No/others
	status	Sample Child Obesity Status	1-Normal, 2-Overweight, 3-Obese

Again, after exploratory analysis of the variables in Sample Adult dataset, the most relevant variables are extracted in Demographic and Health related categories as shown in Table 4.7. Initially there were 21,781 records and 980 variables in Sample Adult. After filtering, as described above, only those Sample Adult records are taken for which there is a child record in the filtered Sample Child dataset. That is how the number of records came down to 2168.

Table 4.7: Final clustering variables for Sample Adult

	Variable Name	**Description**	**Recoded values**
Demographic Variables	bmi	Body Mass Index (BMI) recoded adult	1-<25, 2-25 to <30, 3-30 to <35, 4-35 and up
	age	Age adult	1-18 to 29, 2-30 to 39, 3-40 to 49, 4-50 and up,
	activity	Freq light/moderate activity (times per wk) recoded	1-At least once a week, 2- Others
Health Related Variable	hypertension	Ever been told you have hypertension recoded	1-Yes, 2-No/others
	asthma	Ever been told you had asthma recoded	1-Yes, 2-No/others
	diabetes	Ever been told that you have diabetes recoded	1-Yes, 2-No/others
	depression	Ever had depression recoded	1-Yes, 2-No/others

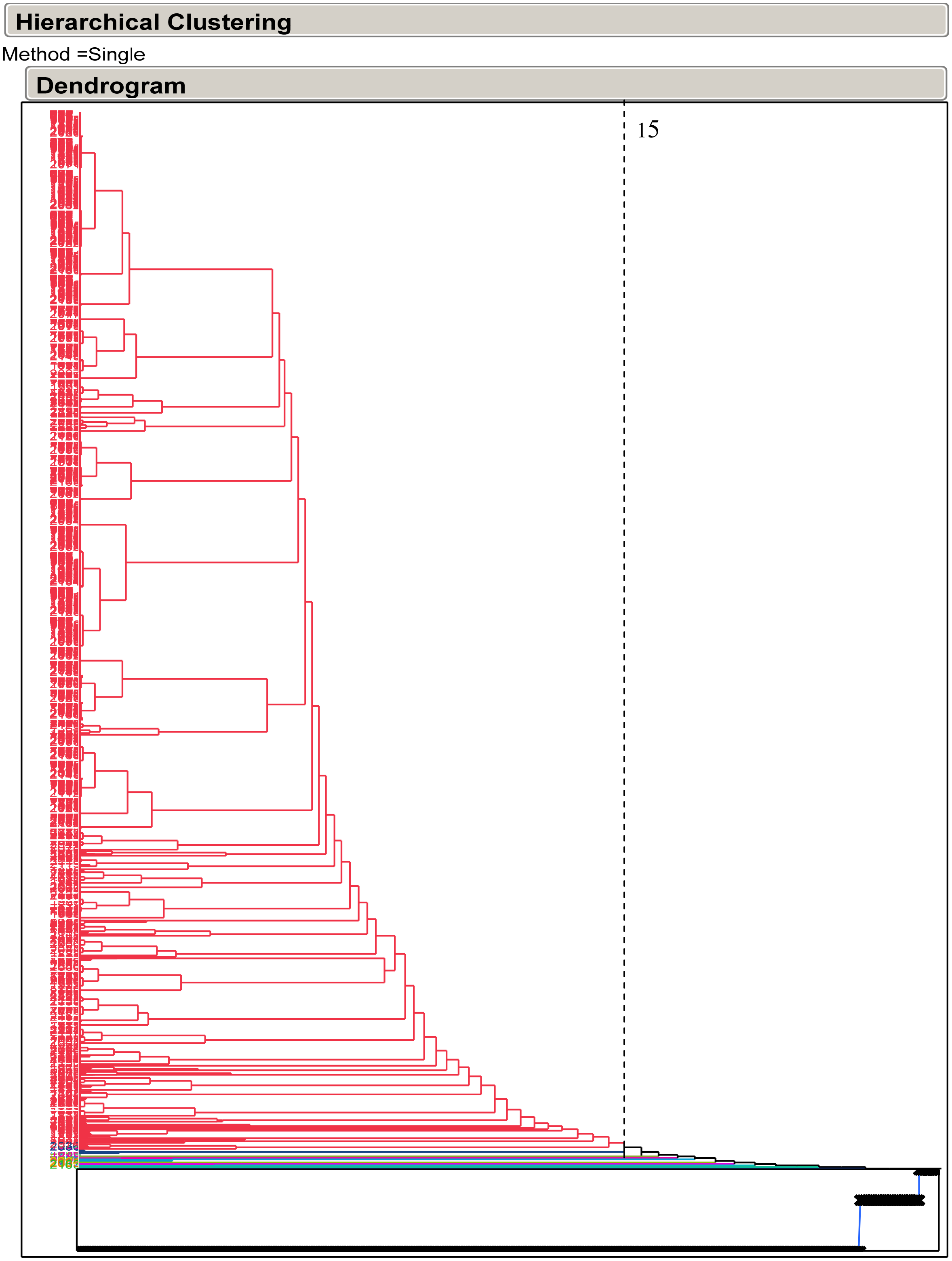
Chapter 5

RESULT AND DISCUSSION

I conducted the clustering, outlier mining and cluster analysis using HAC on Sample Child and Sample Adult datasets in three phases. First, Sample Child dataset is clustered and the characteristics of those clusters are analyzed. Then, HAC is applied on Sample Adult dataset and the resultant clusters are analyzed. Finally, cluster analysis is conducted on the combined Sample Child-Sample Adult dataset to mine the Child-Adult characteristics with respect to adolescent obesity. In the outlier detection part, HAC technique is applied taking all variables (demographic and health related) into consideration in each dataset. The possible outliers (sub cluster and/or single observations) are collected together into one group and compared with the general (non-outliers) group with respect to the distribution of adolescent obesity. It is important to mention that, I did not include the outcome variable "status" in the clustering process.

5.1 Result for Sample Child

Applying HAC on both the demographic variables (*age* and *sex*) and the health related variables (place, days_missed, health, asthma) together, I get the clusters as shown in the dendrogram (Figure 5.1). Then using exploratory analysis, I found an optimum number of clusters as 15, where I cut the dendrogram.



Hierarchical Clustering

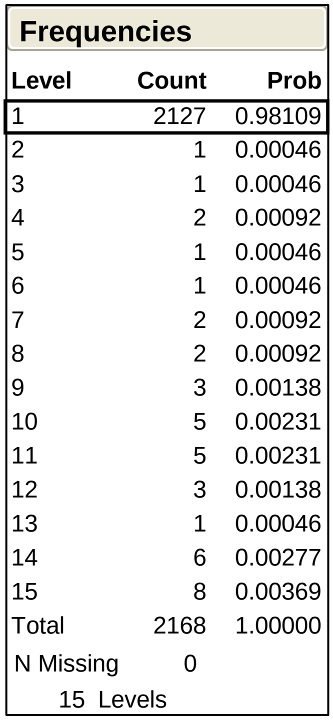
Method =Single

Dendrogram

Figure 5.1: Dendrogram for Sample Child clustering

Most of the observations (2127 out of 2168) are clustered together in cluster 1 (level 1) as shown in the distribution (Table 5.1). The rest of the observations are clustered either in a very small (of size 2 to 8) cluster or in a cluster with only a single observation. The observations in these smaller clusters and in the cluster with single observation are the candidates for outliers.

Table 5.1: Clustering frequency distribution for Sample Child



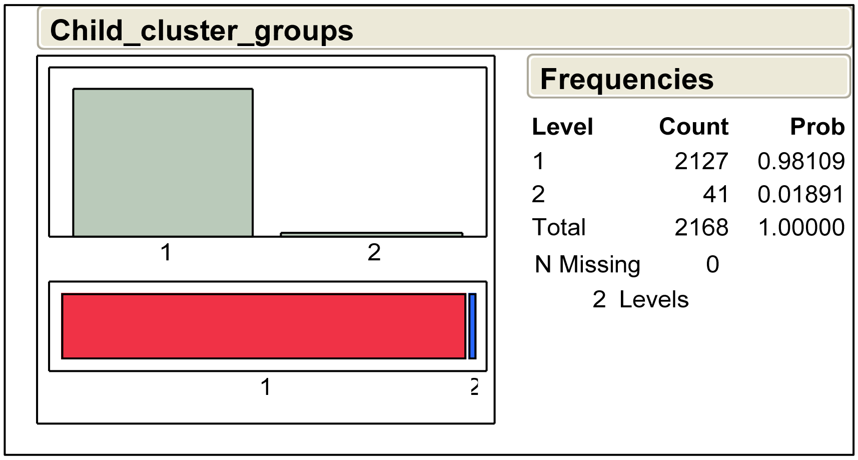
Frequencies

Level	Count	Prob
1	2127	0.98109
2	1	0.00046
3	1	0.00046
4	2	0.00092
5	1	0.00046
6	1	0.00046
7	2	0.00092
8	2	0.00092
9	3	0.00138
10	5	0.00231
11	5	0.00231
12	3	0.00138
13	1	0.00046
14	6	0.00277
15	8	0.00369
Total	2168	1.00000

N Missing 0

15 Levels

I group the outliers (level 2 to 15 in Table 5.1) together and separate the outlier group from the general group as shown in Figure 5.2 where level 1 is general and level 2 are the outliers. The distribution of obesity status for both the general and the outlier groups is shown in Figure 5.3. The obesity distribution for the general group (63.61% normal, 19.23% overweight and 17.16% obese) shows that the obesity rates are similar with the National rates [25].



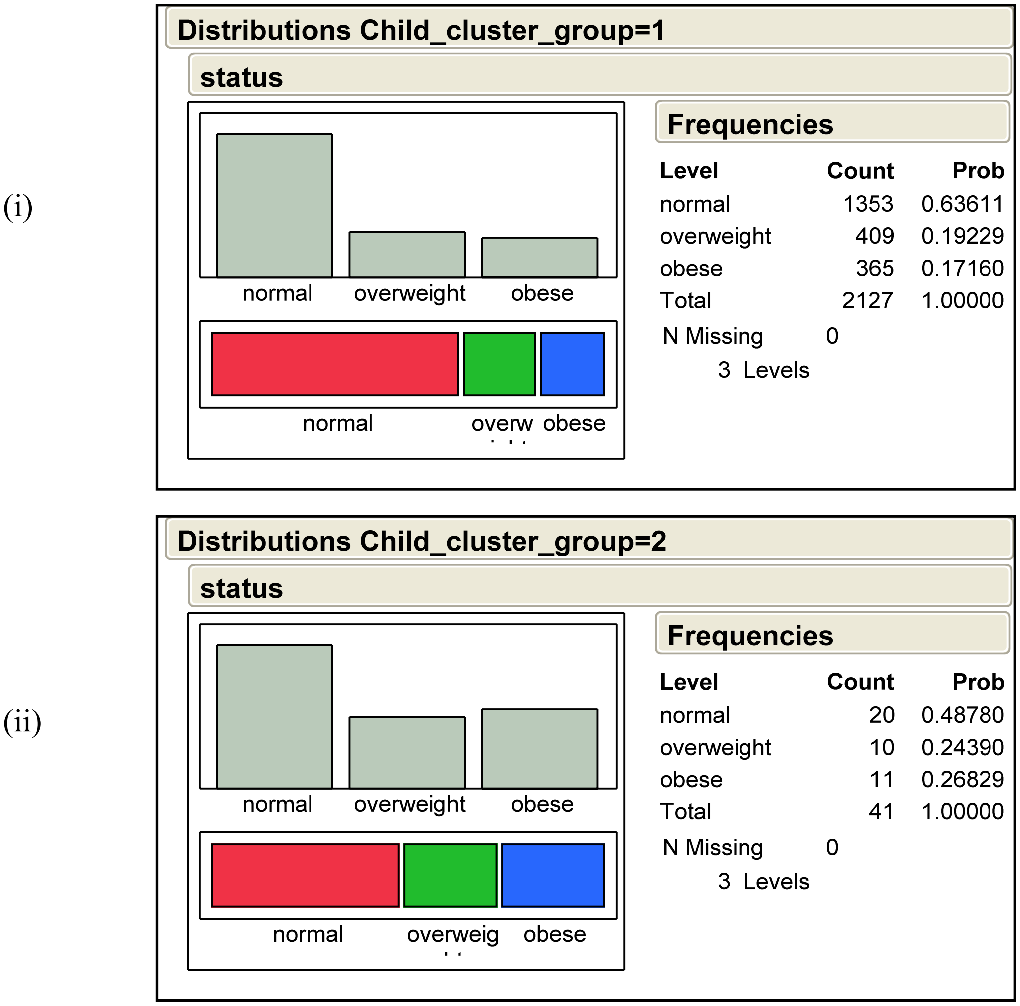
Child_cluster_groups

Frequencies

Level	Count	Prob
1	2127	0.98109
2	41	0.01891
Total	2168	1.00000

N Missing 0
2 Levels

Figure 5.2: Distribution of general (level 1) and outlier (level 2) groups for children



Distributions Child_cluster_group=1

status

Frequencies

Level	Count	Prob
normal	1353	0.63611
overweight	409	0.19229
obese	365	0.17160
Total	2127	1.00000

N Missing 0
3 Levels

Distributions Child_cluster_group=2

status

Frequencies

Level	Count	Prob
normal	20	0.48780
overweight	10	0.24390
obese	11	0.26829
Total	41	1.00000

N Missing 0
3 Levels

Figure 5.3: Distribution of status for children. (i) general(group 1), (ii) outlier(group 2)

But the obesity rates (48.78% normal, 24.39% overweight and 26.83% obese) in the outlier group are different than the National rates. There are 41 ouliers detected by the HAC method. In outlier group, more adolescents are overweight and obese compared to the general group.

5.2 Verification of the Outliers Result for Sample Child

I compare the obesity status distribution extracted by hierarchcal clustering with the obesity status distribution as determined by three widely used methods; namely, Mahalanobis distance measure, Jackknife Technique, and T^2 statistic. I consider approximately the same number of top outliers based on the distance generated by each method as shown in the distance plots (Figure 5.4). The effort here is to choose a distance in each method so that the number of outliers is as close as possible to the number of outliers generated by HAC. Table 5.2 shows the breakdown of the number of outliers by obesity status and methods. The obesity status distribution over these three outlier groups and corresponding general groups are shown in Figure 5.5(a)-(c). The result shows that the outlier extracted by hierarchical clustering exhibits almost the same obesity distribution as the outliers obtained by other methods.

Table 5.2: Obesity status breakdown over outliers and methods for Sample Child

Obesity status	Methods [frequency (%)]			
	Hieararchical Clustering	Mahalanobis Distance	Jackknife Technique	T2 Statistic
Normal	20 (48)	21 (53)	17 (43)	17 (43)
Overweight	10 (24)	7 (17)	12 (30)	12 (30)
Obese	11 (26)	11 (28)	10 (25)	10 (25)
Total	41	39	39	39

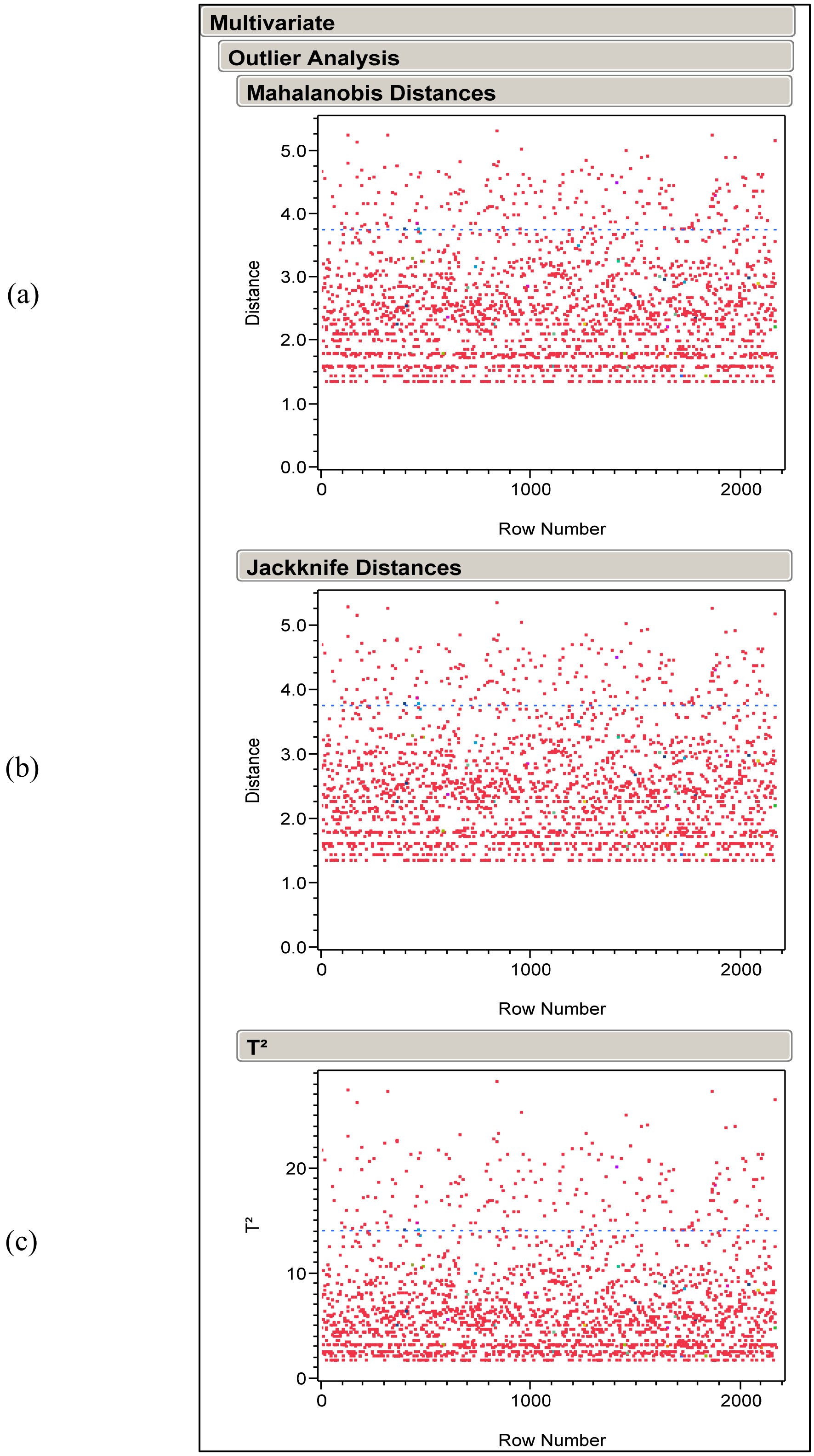
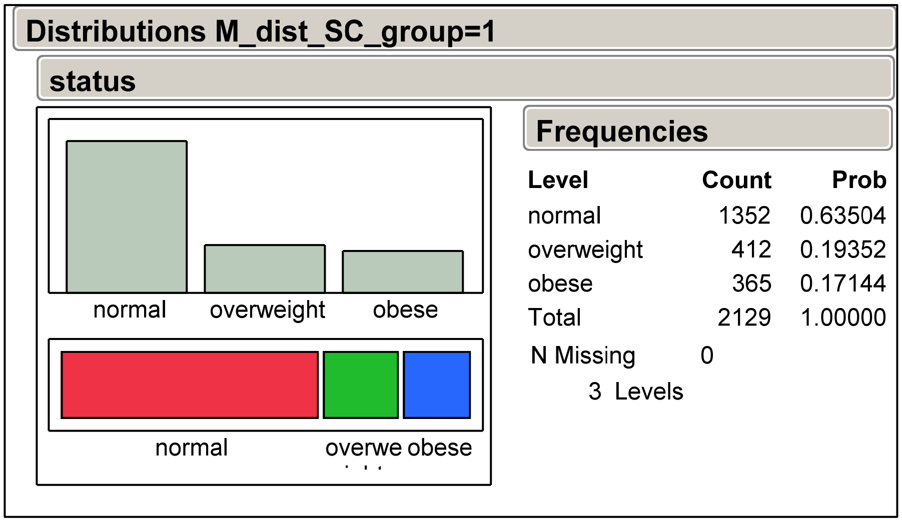


Figure 5.4: Distance plots for Sample Child using multivariate outlier detection methods. (a) Mahalanobis, (b) Jackknife, and (c) T^2



Distributions M_dist_SC_group=1

status

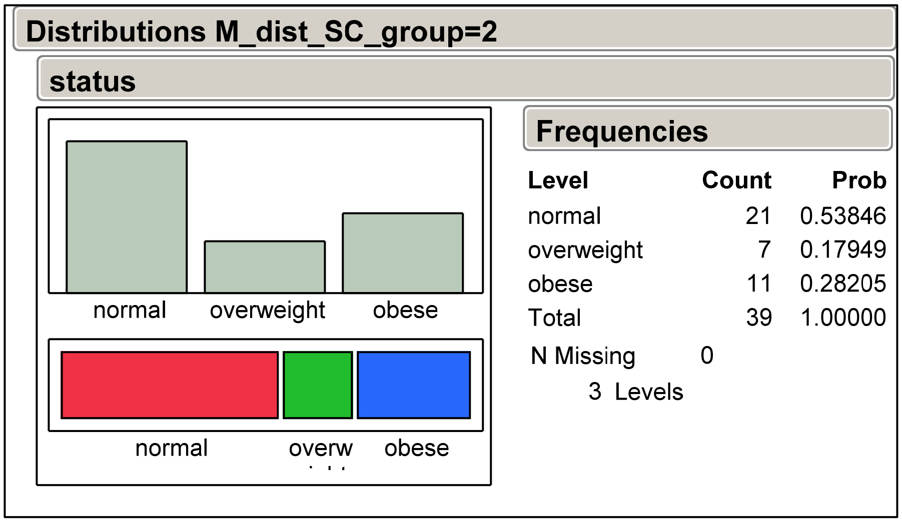
Frequencies

Level	Count	Prob
normal	1352	0.63504
overweight	412	0.19352
obese	365	0.17144
Total	2129	1.00000

N Missing 0

3 Levels

(i)



Distributions M_dist_SC_group=2

status

Frequencies

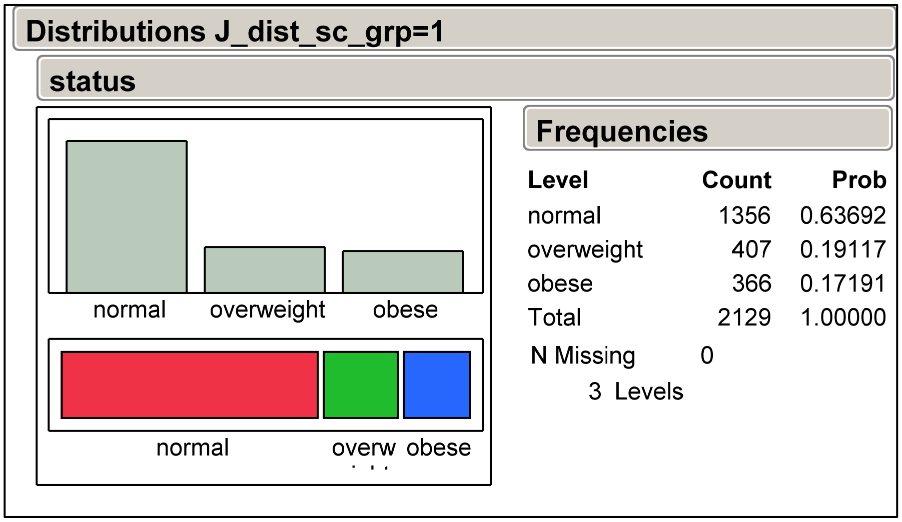
Level	Count	Prob
normal	21	0.53846
overweight	7	0.17949
obese	11	0.28205
Total	39	1.00000

N Missing 0

3 Levels

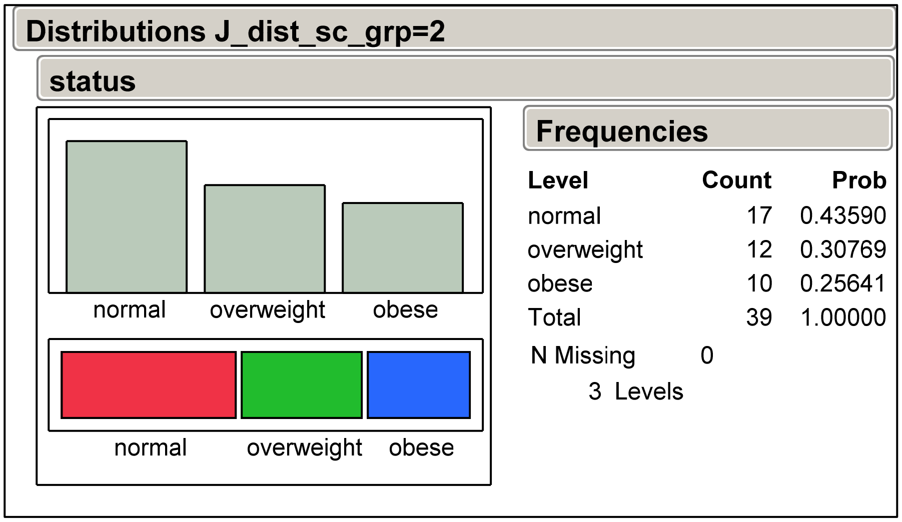
(ii)

Figure 5.5 (a): Distribution of status over outliers using Mahalanobis distance measure.
(i) for general group, (ii) for outlier group.



```
Distributions J_dist_sc_grp=1
  status
                        Frequencies
                        Level        Count     Prob
                        normal        1356  0.63692
                        overweight     407  0.19117
                        obese          366  0.17191
                        Total         2129  1.00000
                        N Missing    0
                            3  Levels
```

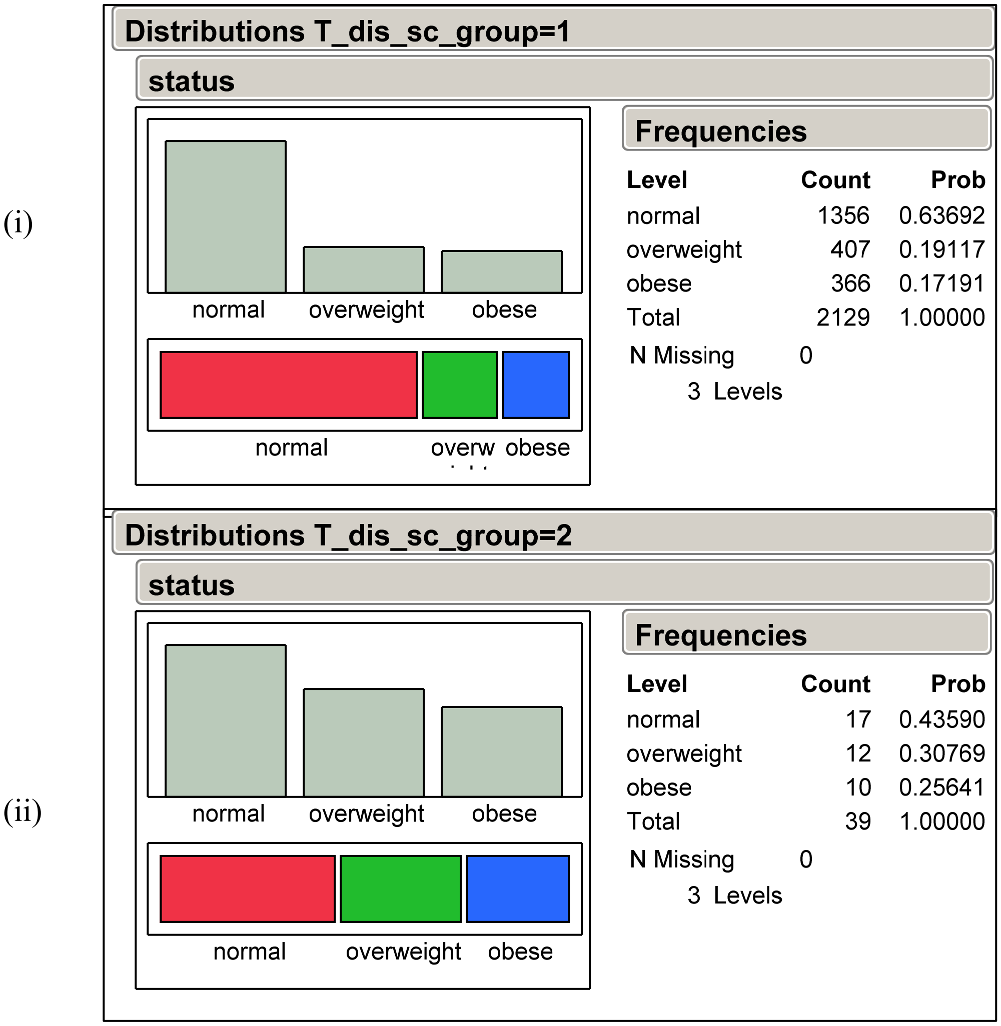
(i)



```
Distributions J_dist_sc_grp=2
  status
                        Frequencies
                        Level        Count     Prob
                        normal          17  0.43590
                        overweight      12  0.30769
                        obese           10  0.25641
                        Total           39  1.00000
                        N Missing    0
                            3  Levels
```

(ii)

Figure 5.5 (b): Distribution of status over outliers using Jackknife technique.
(i) for general group, (ii) for outlier group.



Distributions T_dis_sc_group=1

status

Frequencies

Level	Count	Prob
normal	1356	0.63692
overweight	407	0.19117
obese	366	0.17191
Total	2129	1.00000

N Missing 0
3 Levels

Distributions T_dis_sc_group=2

status

Frequencies

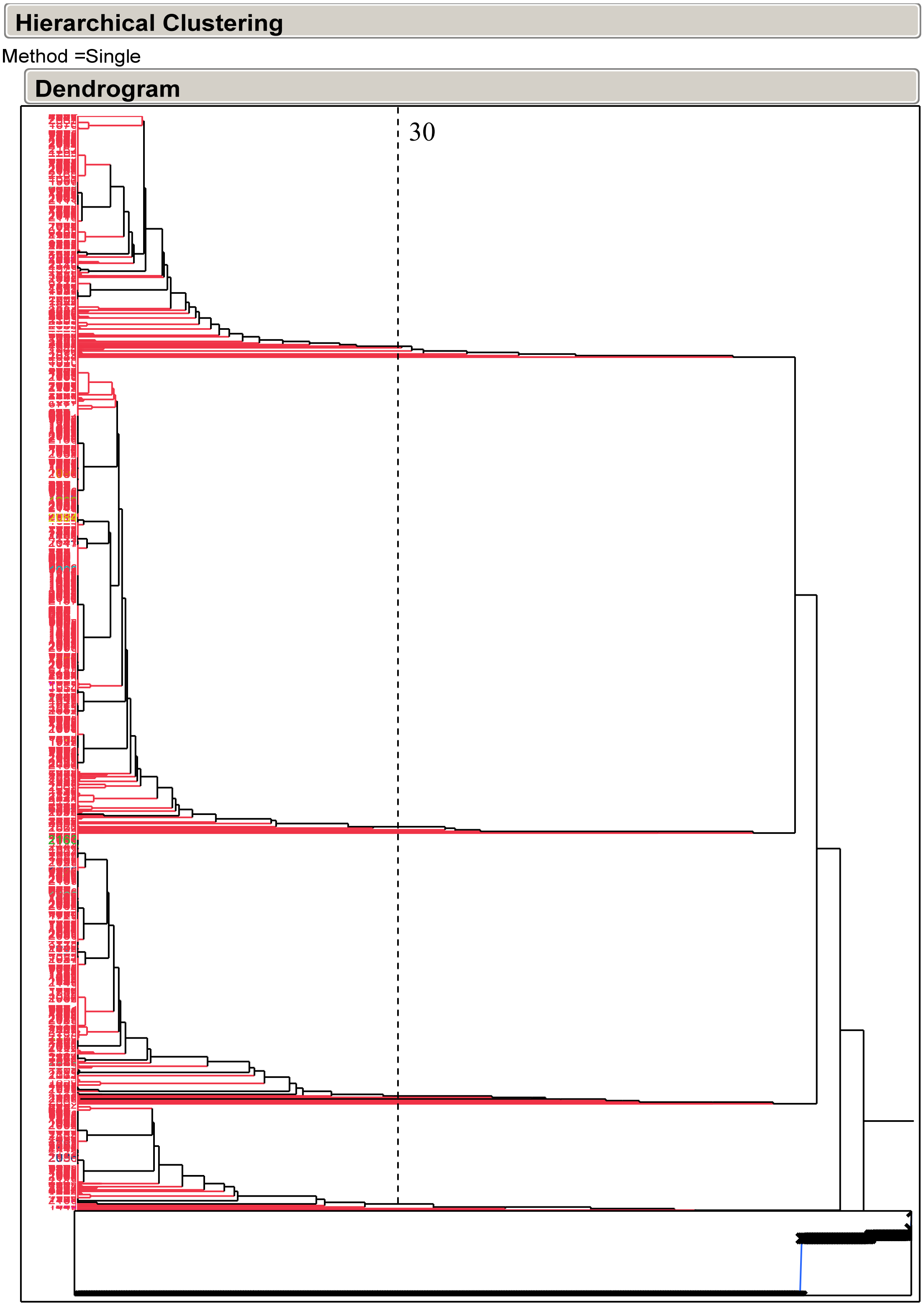
Level	Count	Prob
normal	17	0.43590
overweight	12	0.30769
obese	10	0.25641
Total	39	1.00000

N Missing 0
3 Levels

Figure 5.5 (c): Distribution of status over outliers using T^2 distribution.
(i) for general group, (ii) for outlier group.

5.3 Result for Sample Adult

I apply HAC on both the demographic variables (*bmi, age,* and *activity*) and the health related variables (hypertension, asthma, diabetes, and depression) together. I get the clusters as shown in the dendrogram (Figure 5.6). Then using exploratory analysis, I found the optimum number of clusters is 30, where I cut the dendrogram.



Hierarchical Clustering

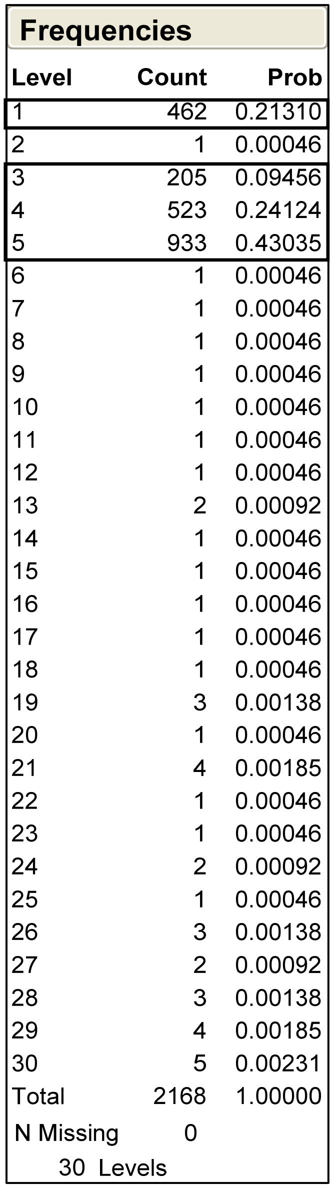
Method =Single

Dendrogram

Figure 5.6: Dendrogram for Sample Adult clustering

For Sample Adult, most of the observations (2123 out of 2168) are clustered into four major clusters (level 1, level 3, level 4, and level 5) as shown in the distribution (Table 5.3). The rest of the observations are clustered either in a very small cluster (of size 2 to 5) or in a cluster with only a single observation. The observations in these smaller clusters and in the cluster with single observation are the candidates for outliers in Sample Adult database.

Table 5.3: Clustering frequency distribution for Sample Adult



Frequencies

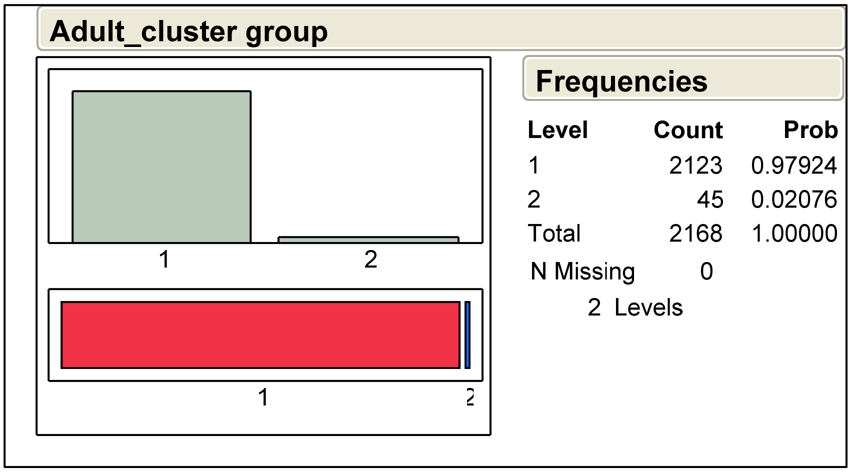
Level	Count	Prob
1	462	0.21310
2	1	0.00046
3	205	0.09456
4	523	0.24124
5	933	0.43035
6	1	0.00046
7	1	0.00046
8	1	0.00046
9	1	0.00046
10	1	0.00046
11	1	0.00046
12	1	0.00046
13	2	0.00092
14	1	0.00046
15	1	0.00046
16	1	0.00046
17	1	0.00046
18	1	0.00046
19	3	0.00138
20	1	0.00046
21	4	0.00185
22	1	0.00046
23	1	0.00046
24	2	0.00092
25	1	0.00046
26	3	0.00138
27	2	0.00092
28	3	0.00138
29	4	0.00185
30	5	0.00231
Total	2168	1.00000

N Missing 0

30 Levels

I collect the possible outliers together into one group and the major clusters (level 1, level 2, level 3, and level 4) together into general group. The distribution of the outlier group and the general group is shown in Figure 5.7. Here the level 1 is considered as the general group and level 2 is outlier group. There are 45 ouliers detected by the HAC method. The distribution of obesity *status* over both the normal and the outlier groups are shown in Figure 5.8. From the obesity distribution for the general group (63.78% normal, 19.12% overweight and 17.1% obese), it is seen that the obesity rates are approximately same as he National reates [25].

But the obesity status rates (42.22% normal, 28.9% overweight and 28.9% obese) in the outlier group are different than the National rates. In outlier group, more adolescents are overweight and obese compared to the general group.

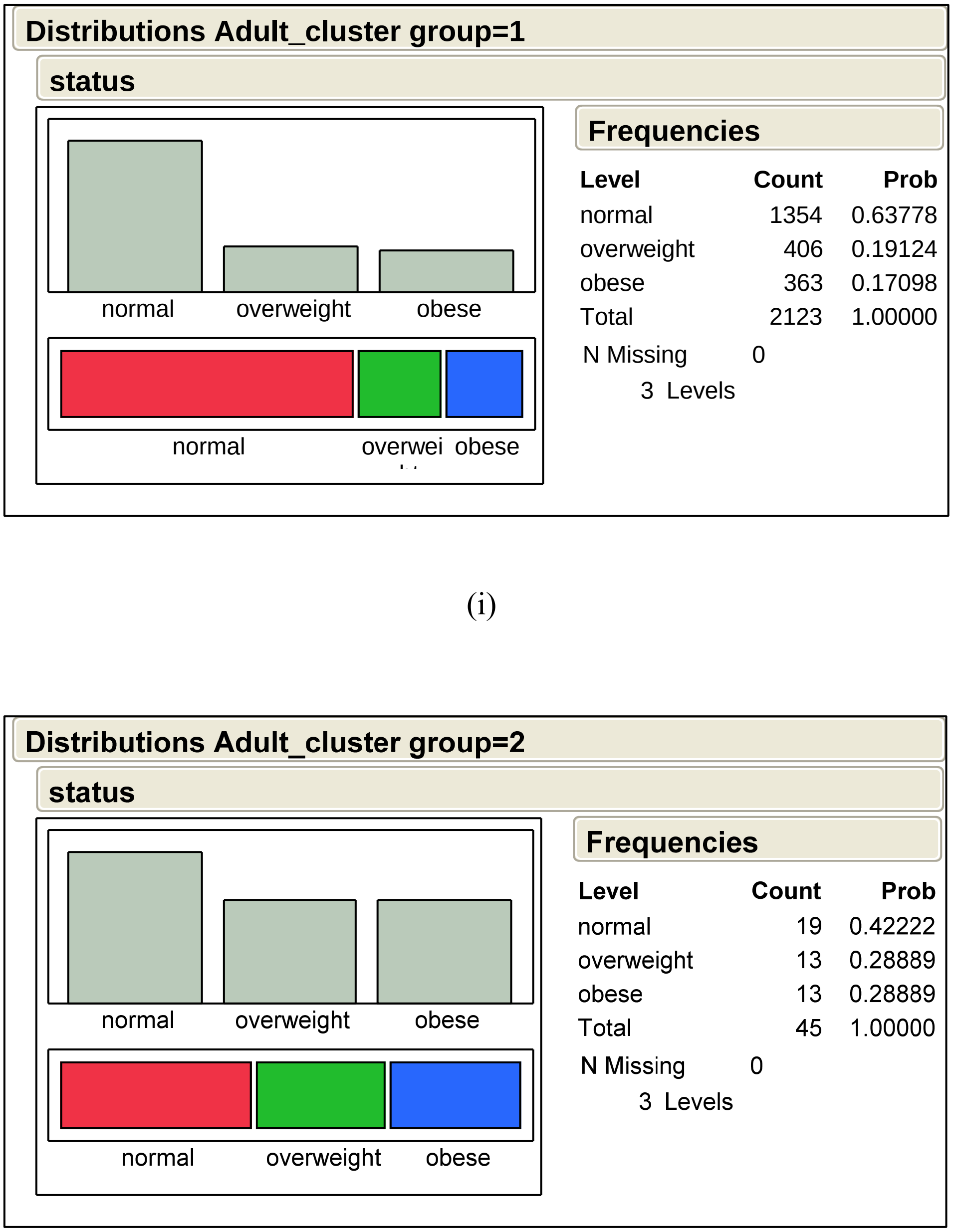


Frequencies

Level	Count	Prob
1	2123	0.97924
2	45	0.02076
Total	2168	1.00000

N Missing 0
2 Levels

Figure 5.7: Distribution of general (level 1) and outlier (level 2) groups for adults



Distributions Adult_cluster group=1

status

Frequencies

Level	Count	Prob
normal	1354	0.63778
overweight	406	0.19124
obese	363	0.17098
Total	2123	1.00000

N Missing 0

3 Levels

(i)

Distributions Adult_cluster group=2

status

Frequencies

Level	Count	Prob
normal	19	0.42222
overweight	13	0.28889
obese	13	0.28889
Total	45	1.00000

N Missing 0

3 Levels

(ii)

Figure 5.8: Distribution of status for Sample Adult.
(i) general (group 1), (ii) outlier (group 2)

5.4 Verification of the Outliers Result for Sample Adult

Again, I compare the obesity distribution over outlier group extracted by hierarchcal clustering with the obesity distribution over outlier groups as determined by Mahalanobis distance measure, Jackknife Technique, and T^2 statistic. I consider approximately the same number (43 for Mahalanobis, 44 for both Jackknife and T^2) of top outliers based on the distances from the distance plots of respective methods as shown in Figure 5.9.

Table 5.4 shows the breakdown of the number of outliers by obesity status and methods for sample adults. The obesity status distribution over the ouliers group generated by these three methods and corresponding general groups are shown in Figure 5.10(a)-(c). The result also shows that outlier extracted by hierarchical clustering exhibits almost the same obesity distribution as the outliers otained using other methods in case of Sample Adult dataset.

Table 5.4: Obesity status breakdown over outliers and methods for Sample Adult

Obesity status	Methods [frequency (%)]			
	Hierarchical Clustering	Mahalanobis Distance	Jackknife Technique	T^2 Statistic
Normal	19(42.22)	20(46.51)	21(47.72)	21(47.72)
Overweight	13(28.88)	12(27.91)	12(27.27)	12(27.27)
Obese	13(28.88)	11(25.58)	11(25.00)	11(25.00)
Total	45	43	44	44

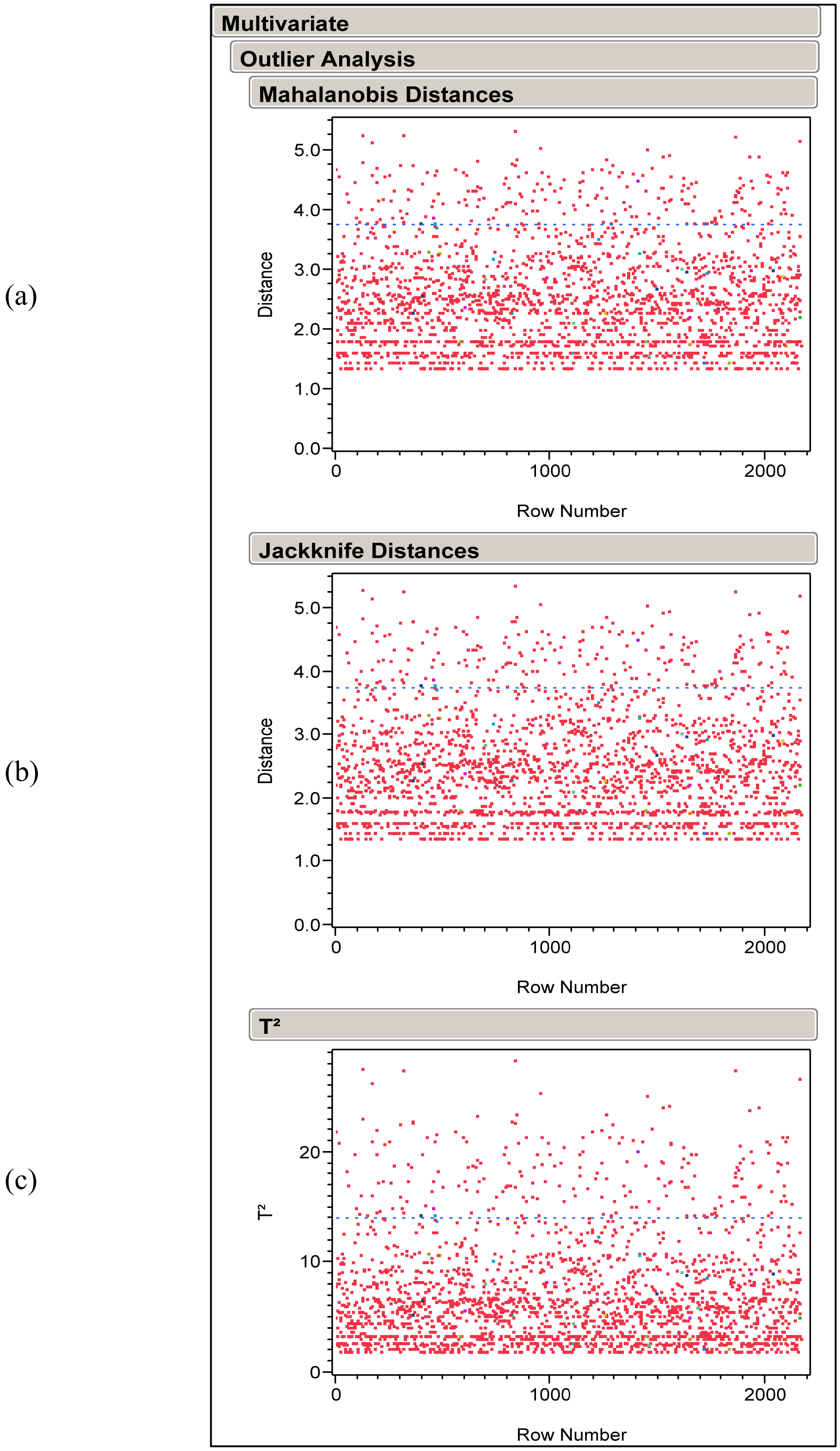
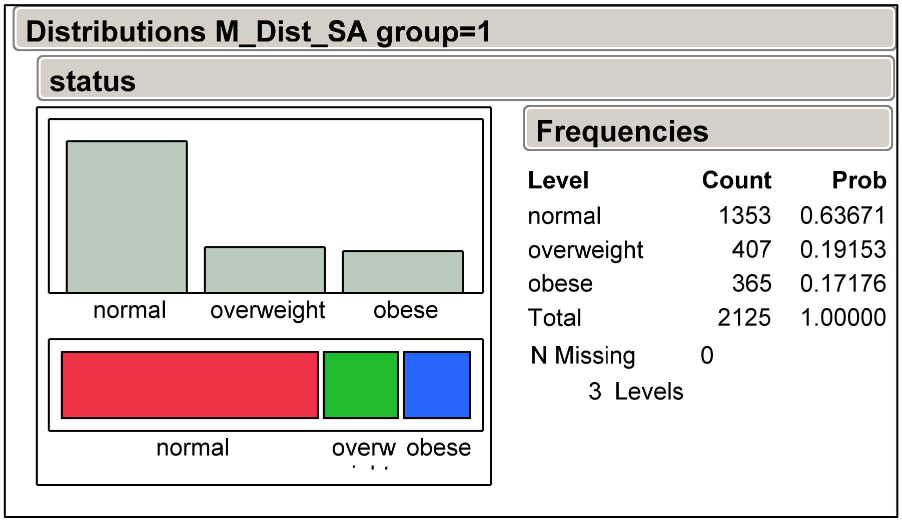


Figure 5.9: Distance plots for Sample Adult using multivariate outlier detection methods. (a) Mahalanobis, (b) Jackknife, and (c) T^2



Distributions M_Dist_SA group=1

status

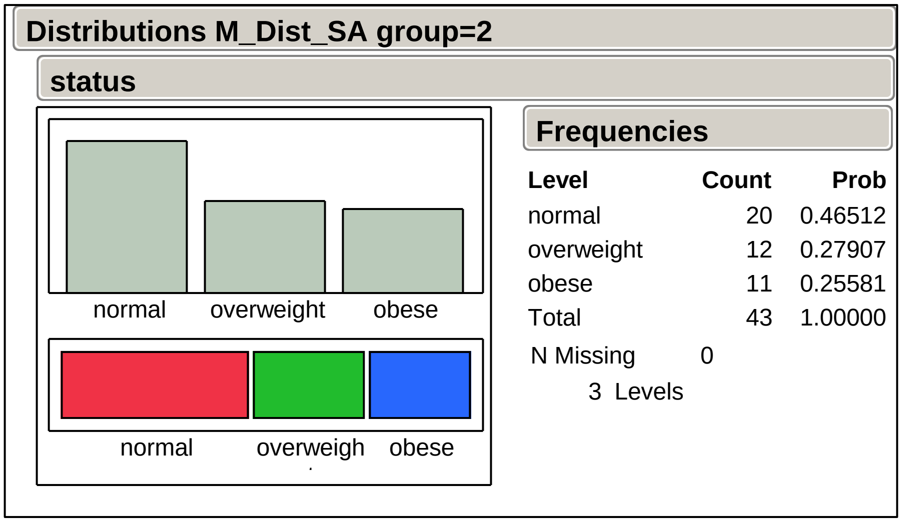
Frequencies

Level	Count	Prob
normal	1353	0.63671
overweight	407	0.19153
obese	365	0.17176
Total	2125	1.00000

N Missing 0

3 Levels

(i)



Distributions M_Dist_SA group=2

status

Frequencies

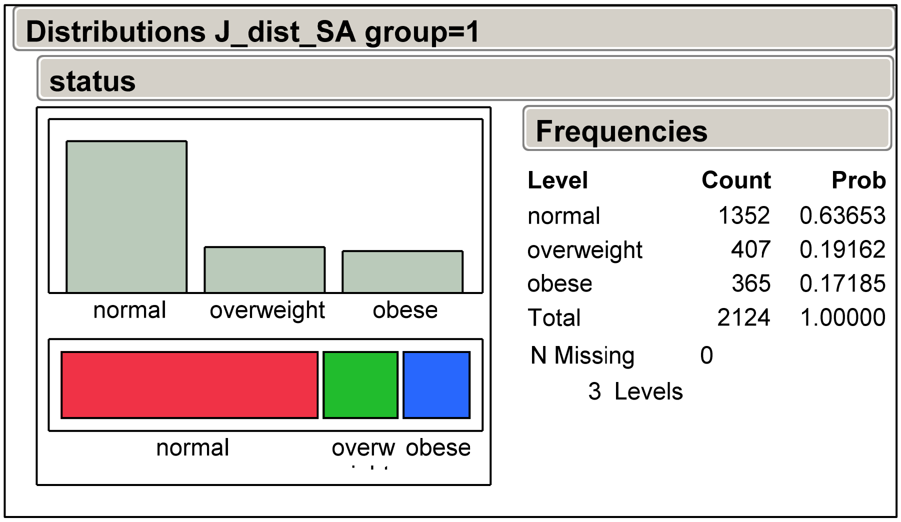
Level	Count	Prob
normal	20	0.46512
overweight	12	0.27907
obese	11	0.25581
Total	43	1.00000

N Missing 0

3 Levels

(ii)

Figure 5.10 (a): Distribution of status for Sample Adult over outliers using Mahalanobis distance measure. (i) for general group, (ii) for outlier group.



Distributions J_dist_SA group=1

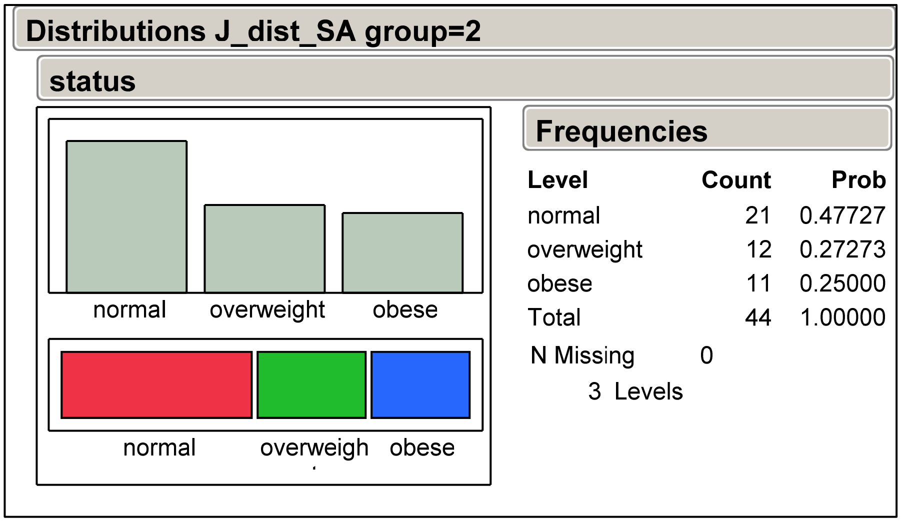
status

Frequencies

Level	Count	Prob
normal	1352	0.63653
overweight	407	0.19162
obese	365	0.17185
Total	2124	1.00000

N Missing 0
3 Levels

(i)



Distributions J_dist_SA group=2

status

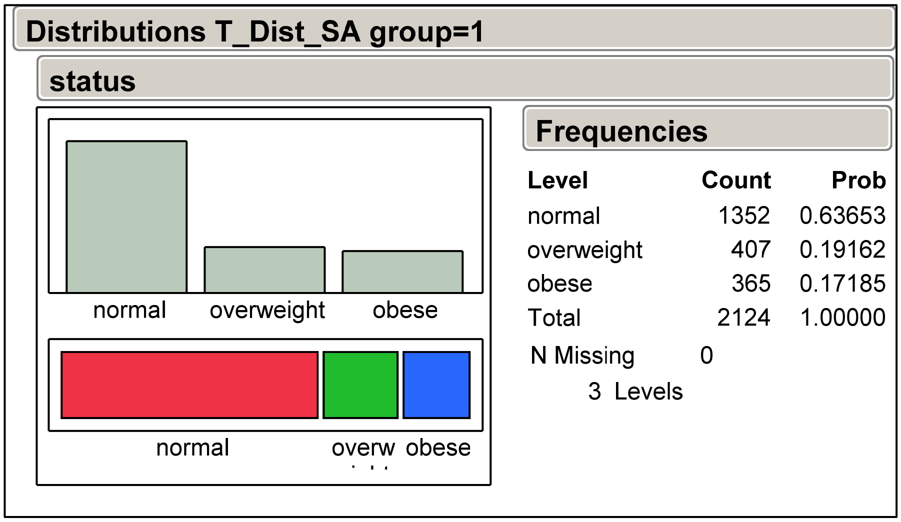
Frequencies

Level	Count	Prob
normal	21	0.47727
overweight	12	0.27273
obese	11	0.25000
Total	44	1.00000

N Missing 0
3 Levels

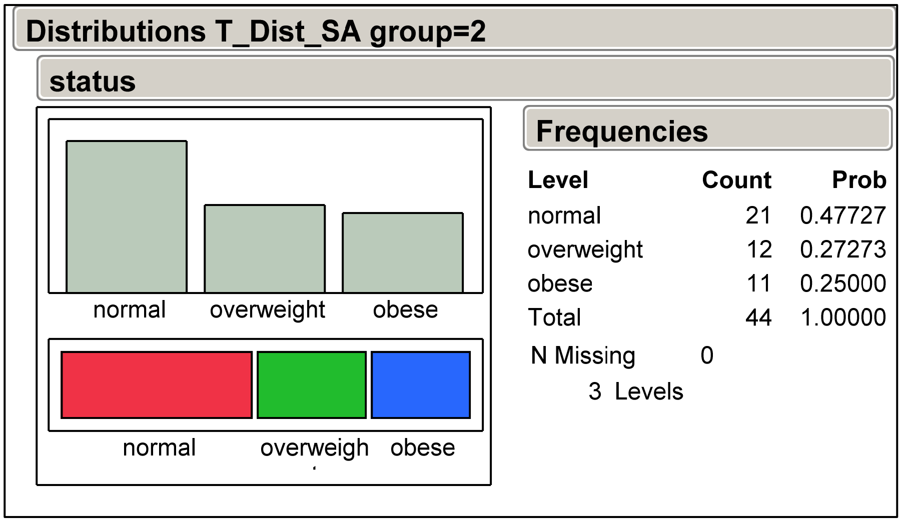
(ii)

Figure 5.10 (b): Distribution of status for Sample Adult over outliers using Jackknife technique. (i) for general group, (ii) for outlier group.



```
Distributions T_Dist_SA group=1
  status
                      Frequencies
                      Level          Count     Prob
                      normal          1352  0.63653
                      overweight       407  0.19162
                      obese            365  0.17185
                      Total           2124  1.00000
                      N Missing      0
                           3  Levels
```

(i)



```
Distributions T_Dist_SA group=2
  status
                      Frequencies
                      Level          Count     Prob
                      normal            21  0.47727
                      overweight        12  0.27273
                      obese             11  0.25000
                      Total             44  1.00000
                      N Missing      0
                           3  Levels
```

(ii)

Figure 5.10 (c): Distribution of status for Sample Adult over outliers using T^2 distribution. (i) for general group, (ii) for outlier group.

5.5 Characteristics Analysis of the Variables Used in Clustering

To explore the characteristics of child and adults with respect to their group (general or outlier), distribution of those variables used in clustering are examined over groups. This mines out the idea about variation in the distribution pattern of the attributes which might be responsible for being an outlier. I used the following legends while generating the distribution.

■ General group

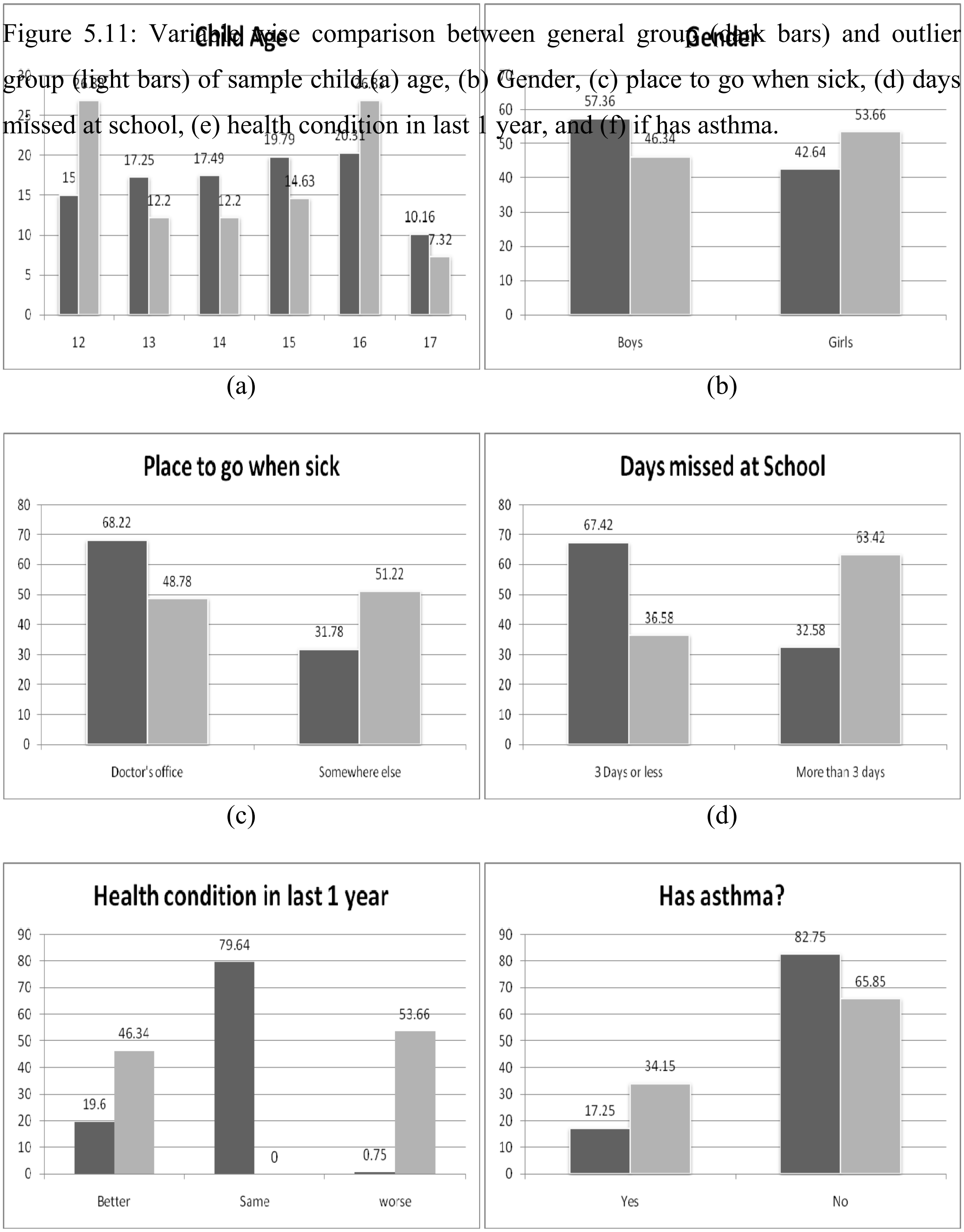
■ Outlier group

5.5.1 Sample Child Characteristics

The characteristics of the Sample Child variables for both the general group and the outlier groups are compared as plotted in Figure 5.11. It is shown (Figure 5.3) that the rates of overweight and obese kids in the outlier group is higher than that of in the general group. A large proportion of kids in the outlier group are *age* of 12 and 16. Gender seems to have little contribution to differentiate groups.

The *place* variable shows that 51.22% kids in the outlier group use to go to somewhere else than Doctor's office when sick. But 68.22% of general kids go to Doctor's office. Days missed at school for most (67.42%) of the general kids is 3 days or less but for the outlier group 63.42% kids missed more than 3 days at school. *Health condition* in last one year for general kids was mostly the same (79.64%) but it was mostly worse (53.66%) for the outlier group. Rate of asthma in the outlier group (34.15%) is more than that in the general kids (17.25%).

Figure 5.11: Variable wise comparison between general group (dark bars) and outlier group (light bars) of sample child (a) age, (b) Gender, (c) place to go when sick, (d) days missed at school, (e) health condition in last 1 year, and (f) if has asthma.



(a) (b)

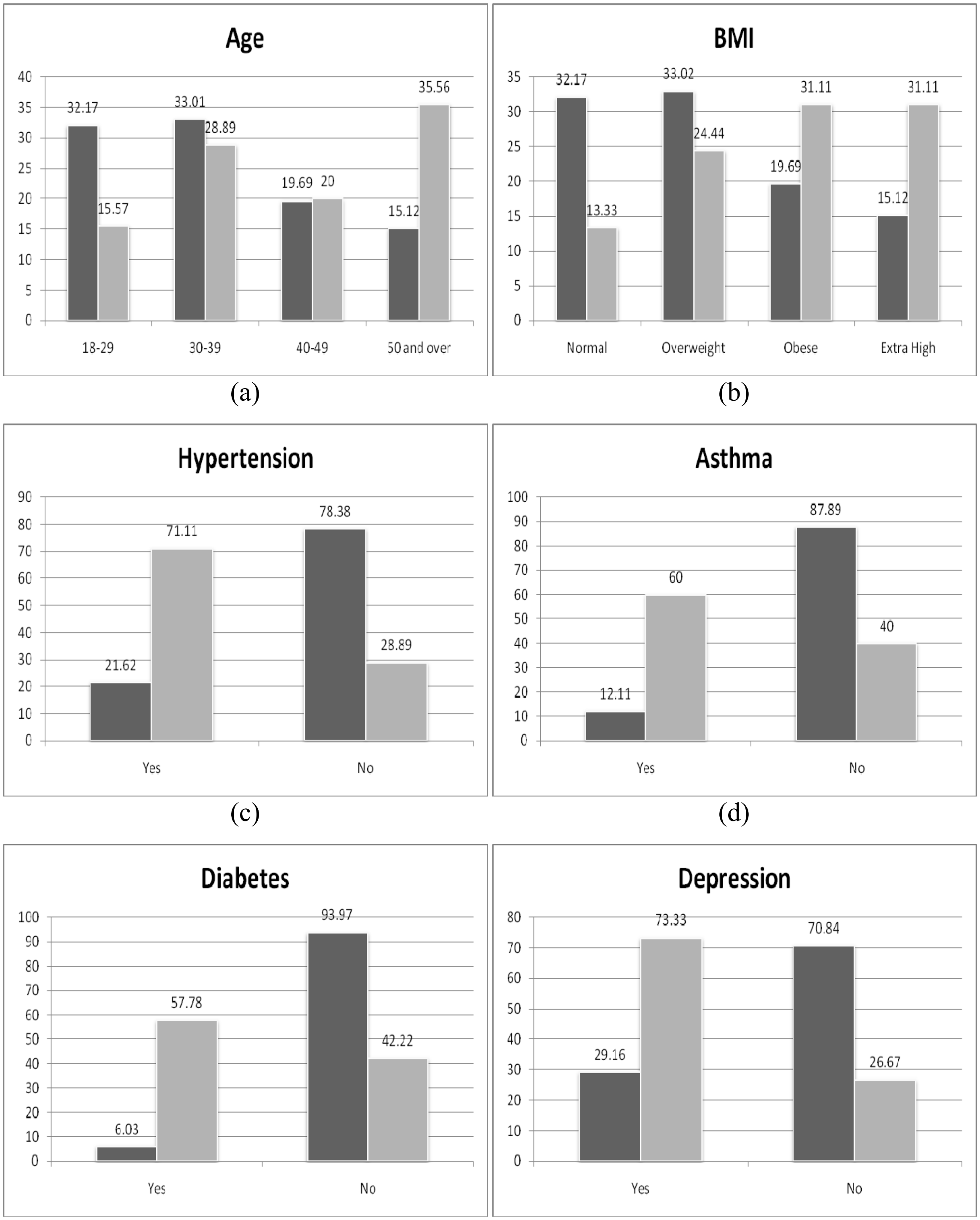
(c) (d)

(e) (f)

5.5.2 Sample Adult Characteristics

The Sample Adult variable's characteristics comparison between the general group and the outlier groups are shown in Figure 5.12. Again, as shown in Figure 5.8, the rates of overweight and obese kids for the adults in the outlier group are higher than that in the general group. Young adults are more in the general groups where the older adults are more in the outlier group. Rate of normal adult BMI (32.17%) in the general group is greater than that in the outlier group (13.33%). On the other hand, rates of obese adult (19.69%) and adult with extra high BMI (15.12%) are less in the general group than in the outlier group compared with 31.11% and 31.11% respectively. The difference between the general and the outlier groups with respect to *Activity* was minimal (data not shown).

With respect to health related variables, *hypertension* rates shows that only 21.62% adult has hypertension problem in general group where 71.11% adult in the outlier group got hypertension. Also, 87.89% general adult did not have *asthma* problem where 60% adults in the outlier group had asthma. Again, 93.97% general adult did not have *diabetes* problem where 57.78% adults in the outlier group had diabetes. And 70.84% general adult did not have *depression* problem where 73.33% adults in the outlier group had depression.



(a) (b)

(c) (d)

(e) (f)

Figure 5.12: Variable wise comparison between general group (dark bars) and outlier group (light bars) of sample adult. (a) age, (b) BMI, (c) hypertension, (d) asthma, (e) diabetes, and (f) depression.

5.6 Limitations

There are limitations of agglomerative hierarchical clustering methods. The time complexity of HAC algorithm is at least quadratic [$O(n^2)$], where n is the number of observations. It is not suitable for massive and high dimensional data. Selection of merge or split points, to generate optimum number of clusters, is critical. Once a group of data points are merged or split, it will continue operation on the new clusters and will not undo what was done previously. If the merge or split decisions are not well chosen, they may result in clusters which are not clearly different. I have chosen the number of clusters using exploratory analysis but it was not exhaustive. The clusters generated may not be optimum. Choosing different distance metrics and methods for measuring distances between clusters may generate different results. In that case it requires multiple experiments and comparison of the results.

Another limitation is that, I used single-linkage rule in my work which reduces the assessment of cluster quality to a single similarity between a pair of objects. A measurement based on one pair cannot fully reflect the distribution of the cluster pair. The merging criterion here is strictly local. A chain of points can be extended for a long distance and with no consideration about the overall shape of the emerging cluster. This effect is called chaining. The clusters generated in this case were not intensively examined for chaining. Multivariate distances are used for detecting outliers here. But, if the variables are highly correlated in a multivariate sense, then it is possible for a point to be ordinary if seen along one or two axes but still be an outlier by violating the correlation.

5.7 Future work

For more quality clustering in the purpose of multivariate outlier mining, hierarchical clustering method could be applied using other distance measure methods like complete-linkage, average-linkage, centroid [27], Ward [28], and fast Ward [29]. Then comparing the clustering results out of these different distance measure methods, the detected outliers might be more accurate.

I used simple Euclidian distance measure which is limited for categorical data. Since the data set here is mostly categorical, I could have used “mismatch value” for the distance matrix. The mismatch value is calculated simply as the number of variables for which the two objects have different values (mismatches), divided by the total number of variables.

Other clustering techniques like, K-means [29], Self Organizing Map (SOM) [30], and especially Tight clustering [31] can be tried for multivariate outlier detection. Tight clustering technique has a special feature that it clusters only with the closest (tight) observations together. The observation which are not functionally or structurally related or even distantly related, are clustered in a separate cluster named as “Noise”. Possibly, the observations that fall into this noise cluster are outliers.

Chapter 6

CONCLUSION

Hierarchical Clustering method always produces a grouping no matter how closely or distantly the objects are related. The group(s) produced may not be always useful for classifying objects. If the grouping discriminate between variables which are not used to do the grouping and if those discriminations are useful, then cluster analysis is useful. **In this thesis I did not use the outcome variable "status" to produce the cluster.** The clusters still distinguished the status between the normal weight, overweight, and obese. Thus the Hierarchical Clustering method was effective in obtaining valuable and meaningful result.

Cluster analysis methods are not clearly established. There are many options available to choose while performing the hierarchical clustering. I may mine the outlier trying different methods for computing the similarity matrix and linking groups until I "discover" a pattern. This may be a bias and could raise criticism.

In this thesis, analyzing the clustering results and the characteristics of variables, I can conclude that for adolescents who do not go to doctor's office when sick, miss more school days, have poor health condition and asthma, will have higher obesity rates. Also, the adults who are older, have higher BMI, hypertension, asthma, diabetes and depression have children with higher obesity rates.

References

[1] Barnett, V, & Lewis, T., “Outliers in statistical data” (3rd Ed.). Wiley and Sons, Inc., Hoboken, New Jersey, 1994

[2] Hawkins, D., “ Identification of Outliers”, Chapman and Hall, London, 1980

[3] E. Knorr and R. Ng, “A Unified Approach for Mining Outliers”. *In Proceedings of the Conference of the Centre for Advanced Studies on Collaborative Research*, pp. 11, 1997.

[4] S. Ramaswamy, R. Rastogi and K. Shim, “Efficient Algorithms for Mining Outliers from Large Data Sets”. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, vol. 29, Issue 2, pp. 427 – 438, 2000.

[5] Acuna E., Rodriguez C. A.,”Meta analysis study of outlier detection methods in classification,” Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, Retrived from academic.uprm.edu/eacuna /paperout.pdf, *In proceedings IPSI 2004*, Venice, 2004.

[6] Breunig M.M., Kriegel H.P., Ng R.T., Sander J.,”Lof: Identifying density-based local outliers,” *In Proc. ACMSIGMOD Conf. 2000*, 93–104, 2000.

[7] Preparata, F. and Shamos, M. “Computational Geometry: an Introduction.” *Springer Verlag*, 1988.

[8] Wang, J. & R. Serfling, “ Nonparametric multivariate kurtosis and tailweight measures”, *Journal of Nonparametric Statistics,* 441-456, 2005

[9] 2008 National Health Interview Survey (NHIS), Public Use Data Release, Centers for Disease Control and Prevention, U.S. Department of Health and Human Services, June 2009

[10] Prabhakar Raghavan & Hinrich Schütze, Christopher D. Manning “Introduction to Information Retrieval”, Cambridge University Press, 2008

[11] S. C. Johnson, "Hierarchical Clustering Schemes" *Psychometrika*, **2**, 241-254, 1967

[12] Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S., "Sur la Liaison et la Division des Points d'un Ensemble Fini," *Colloquium Mathematicae*, **2**, 282 -285. 1951a.

[13] Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S., "Taksonomia Wroclawska," *Przeglad Antropol.*, **17**, 193 -211, 1951b.

[14] McQuitty, L.L., "Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies," *Educational and Psychological Measurement*, **17**, 207 -229, 1957.

[15] Sneath, P.H.A., "The Application of Computers to Taxonomy," *Journal of General Microbiology*, **17**, 201 -226, 1957.

[16] Sokal, R.R. and Michener, C.D. "A Statistical Method for Evaluating Systematic Relationships," *University of Kansas Science Bulletin*, **38**, 1409 -1438, 1958.

[17] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "Cluster Validity Methods: part I". In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, vol. 31, Issue 2, pp. 40 – 45, June 2002

[18] Kanti Mardia, J.T. Kent, J.M. Bibby . *Multivariate Analysis*. Academic Press, 1979.

[19] V. Hautamäki, I. Kärkkäinen and P. Fränti, "Outlier Detection Using k-Nearest Neighbor Graph". In *Proceedings of the International Conference on Pattern Recognition*, vol. 3, pp. 430 – 433, Cambridge, UK, August 2004.

[20] S. Lin and D. Brown, "An Outlier-based Data Association Method". In *Proceedings of the SIAM International Conference on Data Mining*, San Francisco, CA, 2003.

[21] P C Mahalanobis, "On the generalised distance in statistics", *Proceedings of the National Institute of Sciences of India*, **2**, 49–55, 1936.

[22] Quenouille M., "Notes on Bias in Estimation", *Biometrika*, **43**, 353-360, 1956

[23] J.W. Tukey, "On the comparative anatomy of transformations", *Ann. Math. Statist.* **28**, 602-632, 1957.

[24] Hotelling, H., "The generalization of Student's ratio"., *Annals of Mathematical Statistics* **2** (3): 360–378, 1931.

[25] http://www.cdc.gov/HealthYouth/obesity/

[26] Wilks, D. S., "Statistical Methods in the Atmospheric Sciences", *Academic Press*, pp. 467, 1995.

[27] Ward, J. H., Hierachical grouping to optimize an objective function. *J. Am. Statist. Assoc.* **58**, 236-244., 1963.

[28] Hervada-Sala, C., Jarauta-Bragulat, E., “A program to perform Ward’s clustering method on several regionalized variables”. *Computers & Geosciences* **30**, 881–886., 2004.

[29] MacQueen, J., “Some methods for classification and analysis of multivariate observations”, In *Proceedings of Fifth Berkeley Symposium on Math. Statist. and Prob.*, pp. 281-297, 1967.

[30] Kohonen, T. “Self-organized formation of topologically correct feature maps”. *Biological Cybernetics*, 43:59-69, 1982.

[31] Tseng GC, Wong WH, **“**Tight clustering: a resampling-based approach for identifying stable and tight patterns in data”, *Biometrics*, **61:**10-16, 2005.

[32] Michael R. Anderberg, “*Cluster analysis for applications*”, 1973.