UNI_final

NAME

2016-5-5

Q 1

1.1Visualizing categorical data

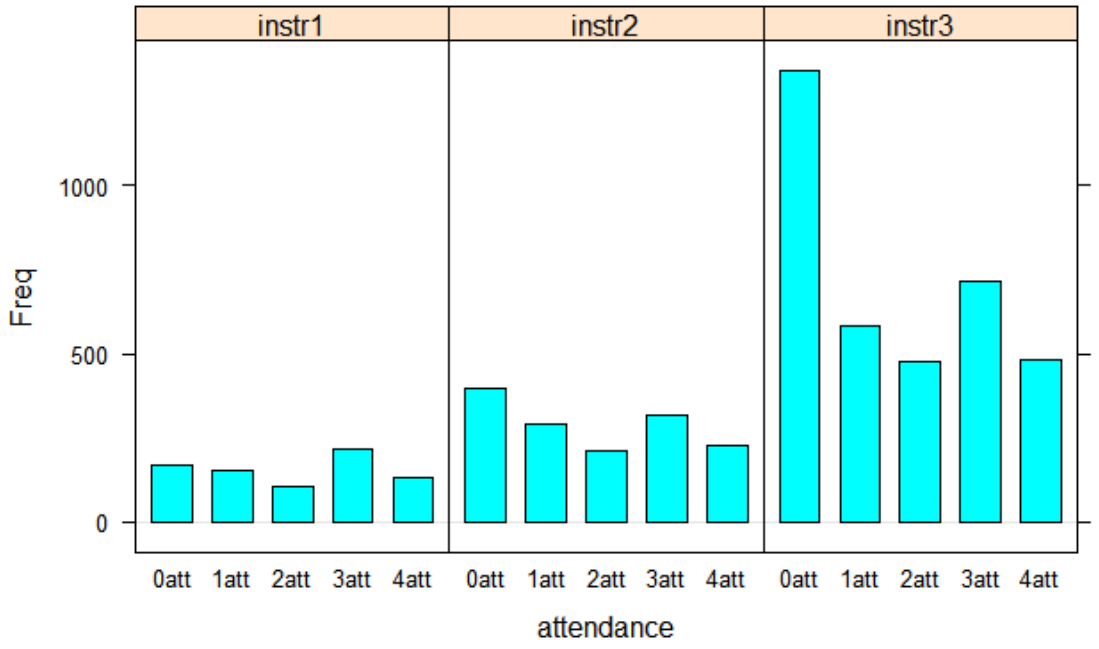

Figure 1.1 Bar chart displaying the frequency distribution of the level of attendance, by Instructor's ID. The *x*-axis represents the level of attendance, and can be viewed as a prior achievement variable. The different panels represent subsets of the instructors, and may be viewed as the response.Obviously,the students taking the third Instructor's courses are more.

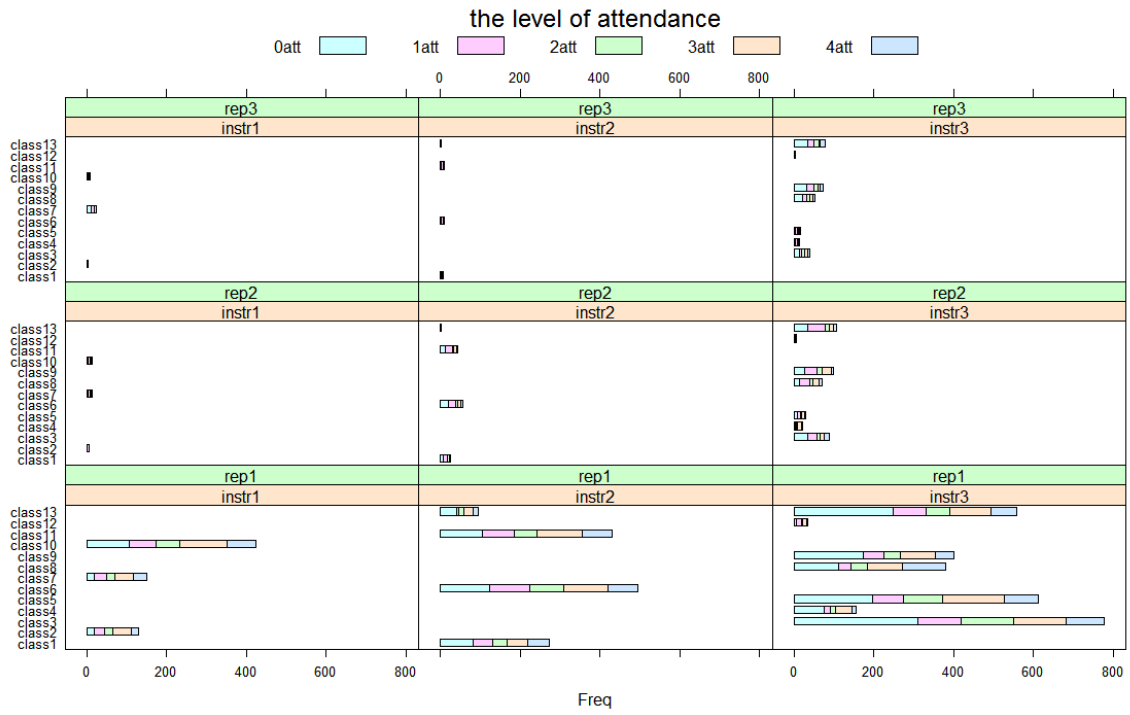

Figure 1.2 The level of attendance among different subgroups of students, with a different horizontal scale in each panel. This emphasizes the proportion of level of attendance within each subgroup, rather than the absolute numbers. The proportion of 0 level of attendance is biggest among the 13 courses, and the students taking the courses one times are most.

1.2Visualizing Univariate Distributions

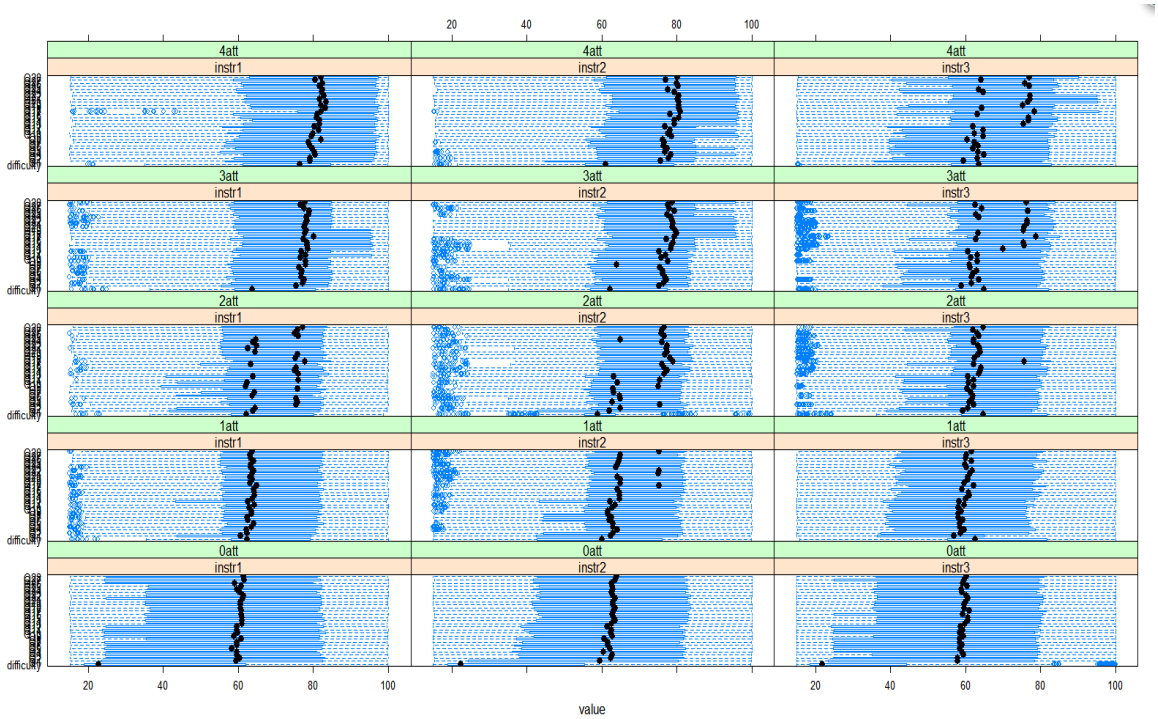

Figure1.3 Comparative box-and-whisker plots of the evaluation scores, representing the subsets in a slightly different layout.It highlights a pattern that the evaluation scores on Level of difficulty of the course among the students of 0 level of attendance is the lowest about 20,although the scores is not too low compared to the other levels of attendance.

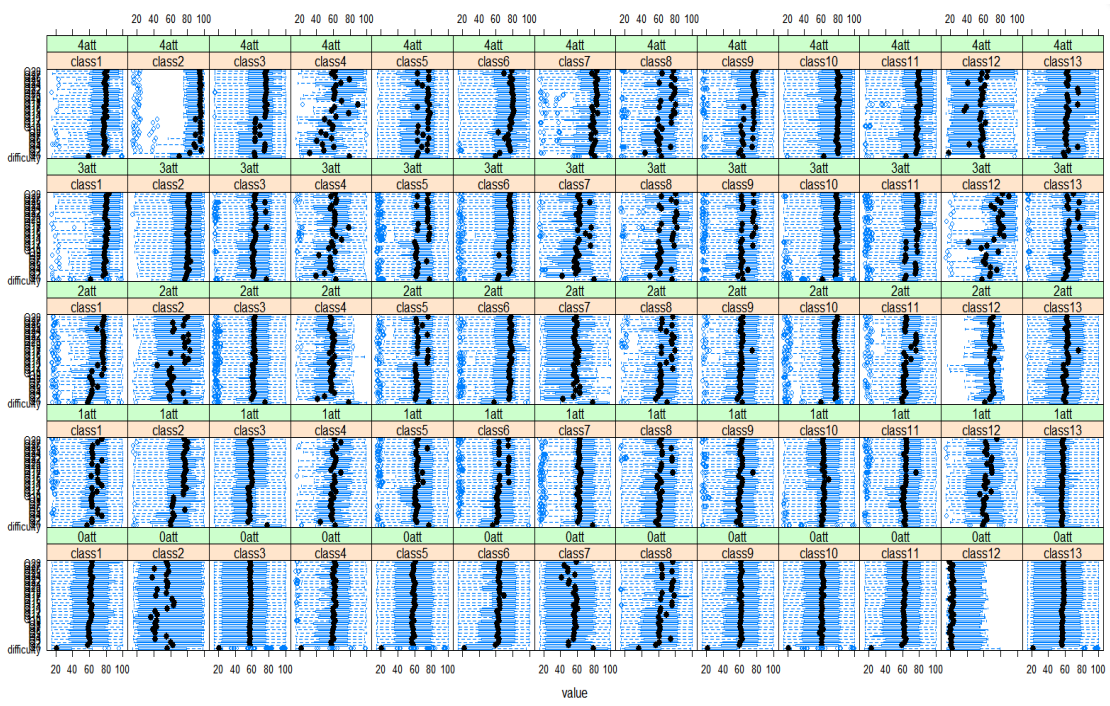

Figure 1.4 Comparative box-and-whisker plots of the evaluation scores among different subgroups, displaying the distribution of the score, by class and the level of attendance.

Box-and-whisker plots summarize the data using a few quantiles, and possibly some outliers. This summarizing can be important when the number of observations is large.

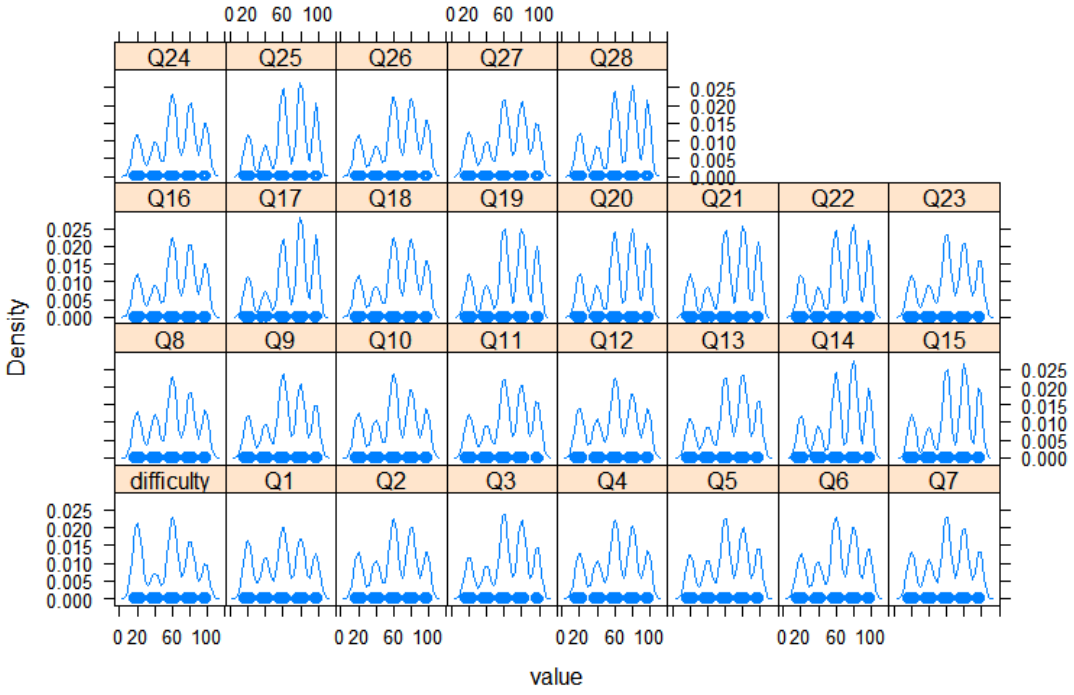

Figure 1.5 A kernel density plot of the evaluation scores of the 29 variables.The plot show the density distribution of the 29 variables is multimodal.

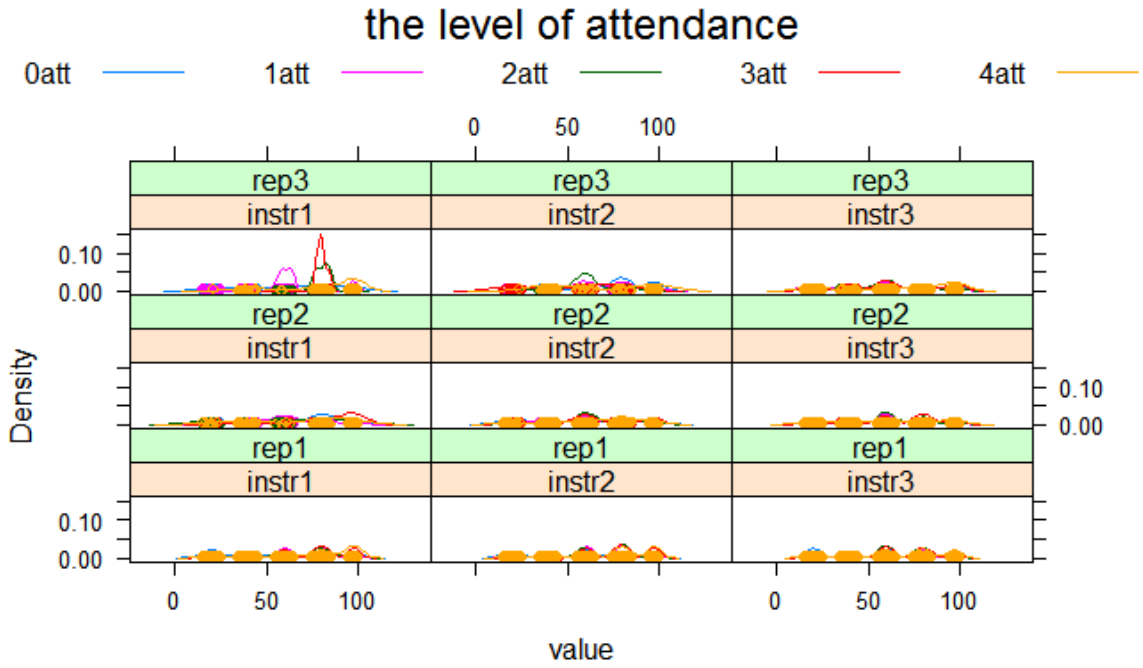

Figure 1.6 Kernel density plots of evaluation scores, combined with all of the 29 variables, representing the distribution among different subgroups(Instructors and

number of times the student is taking this course) of students,with a different scale in each panel. It is not difficult to find that the numbers are distribute in four interval.

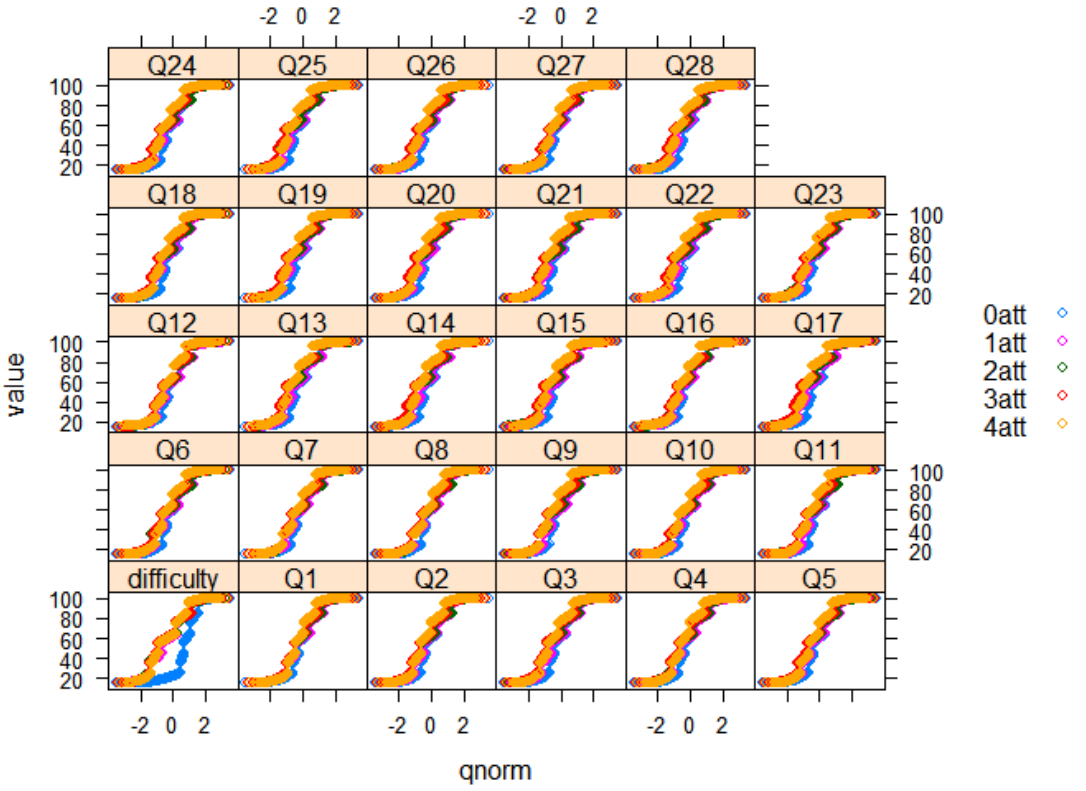

Figure 1.7 Normal Q–Q plots of for evaluation scores from different variables, grouped by the level of attendance.There are a lot of the systematic departures from normality and homoscedasticity.

1.3 Multivariate Displays

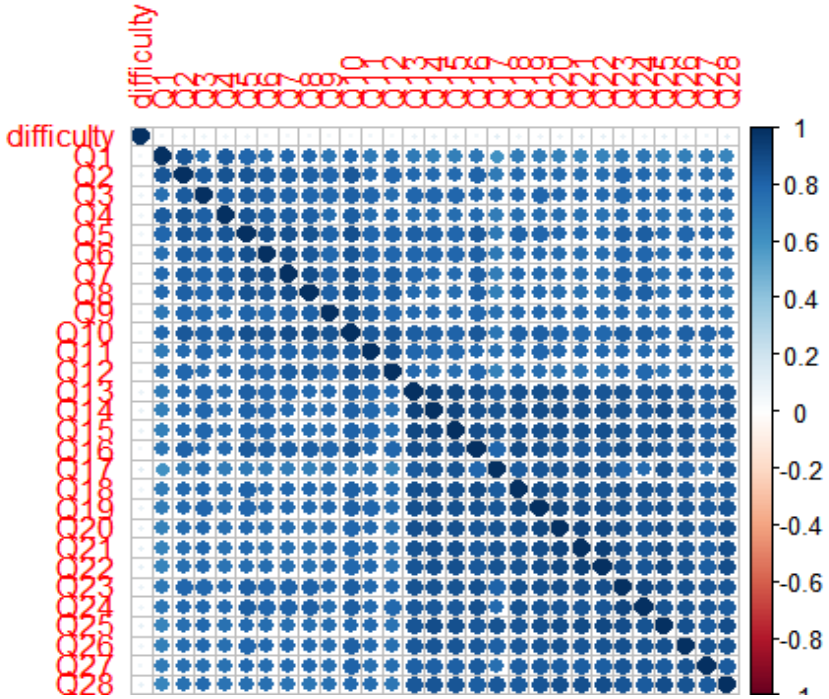

Figure 1.8 Correlation matrix of the numerical variables.The color stand for the correlationship in each pair of the variables,displaying the high correlation between Q1~Q28.

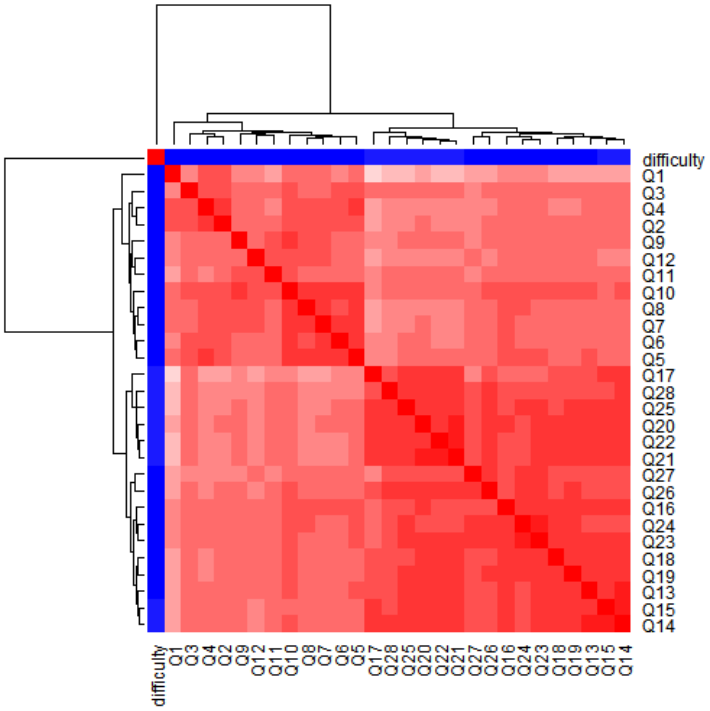

Figure 1.9 A heatmap created with the heatmap function along with a legend representing a hierarchical clustering. The thin strip at the root of the dendrogram represents a grouping of the variables based on evaluation scores.

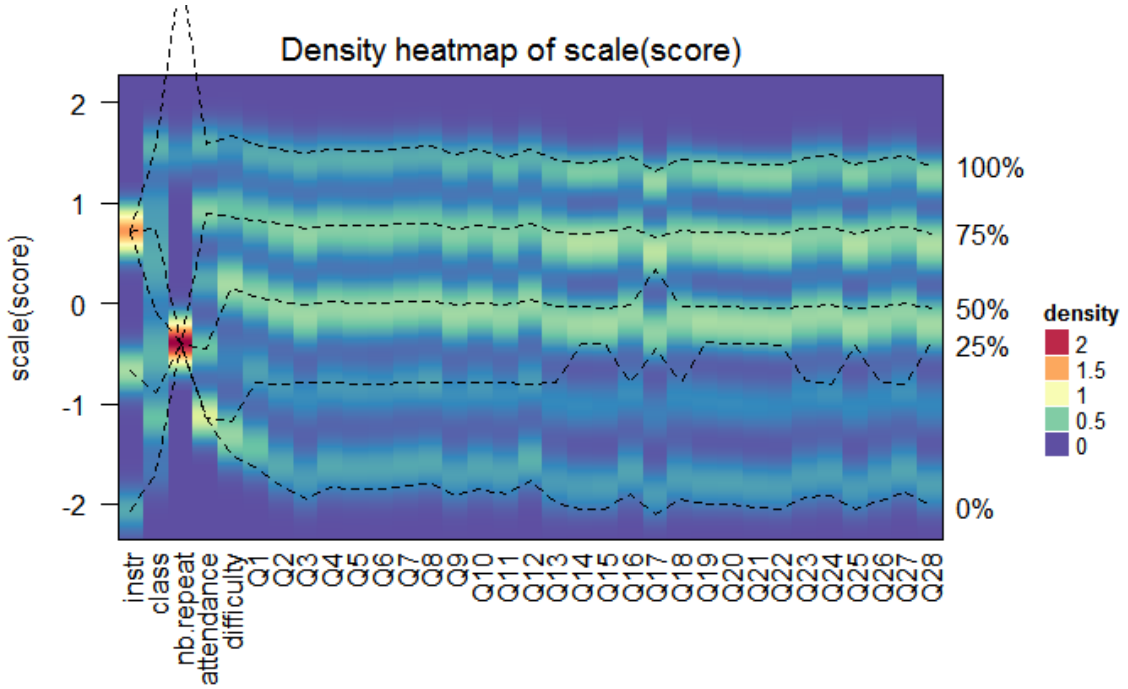

Figure 1.10 Density heatmap of scaled score data.The dashed lines on the heatmap correspond to the five quantile numbers. The text for the five quantile levels are added in the right of the heatmap.

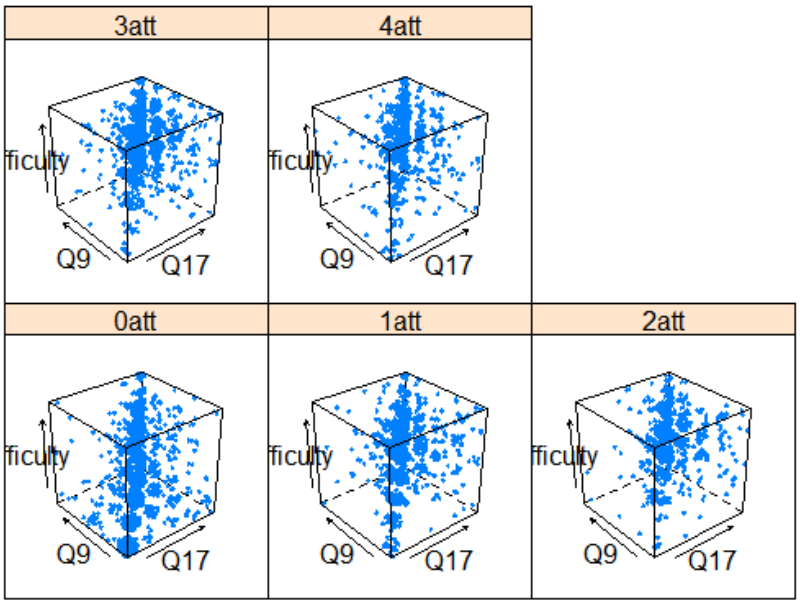

Figure 1.11 A three-dimensional scatter plot of evaluation scores in terms of Q9,Q17, and difficulty,by attendance. Arrows indicate the direction in which the axes increase;the data points mainly gather near the oblique diagonal line.

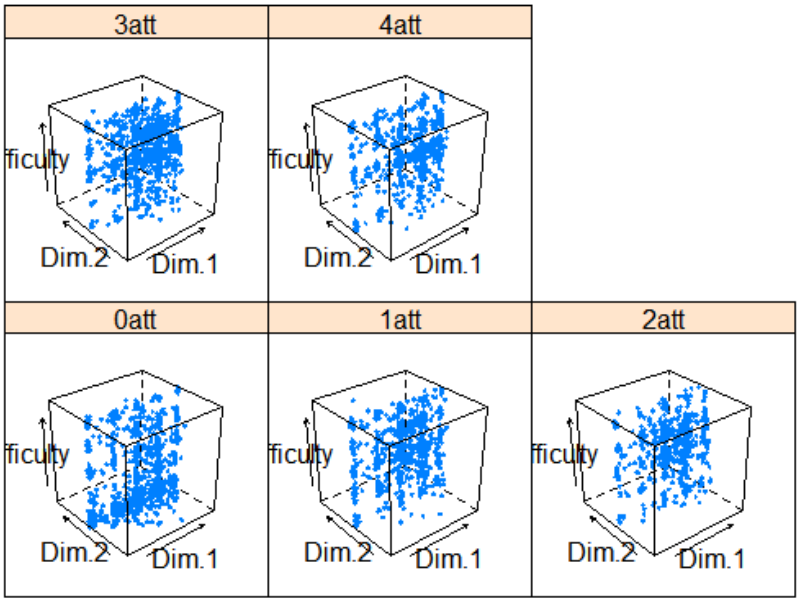

Figure 1.12 A three-dimensional scatter plot of the first two PCA data and difficulty. The point mainly distribute in the middle of Dim.2.

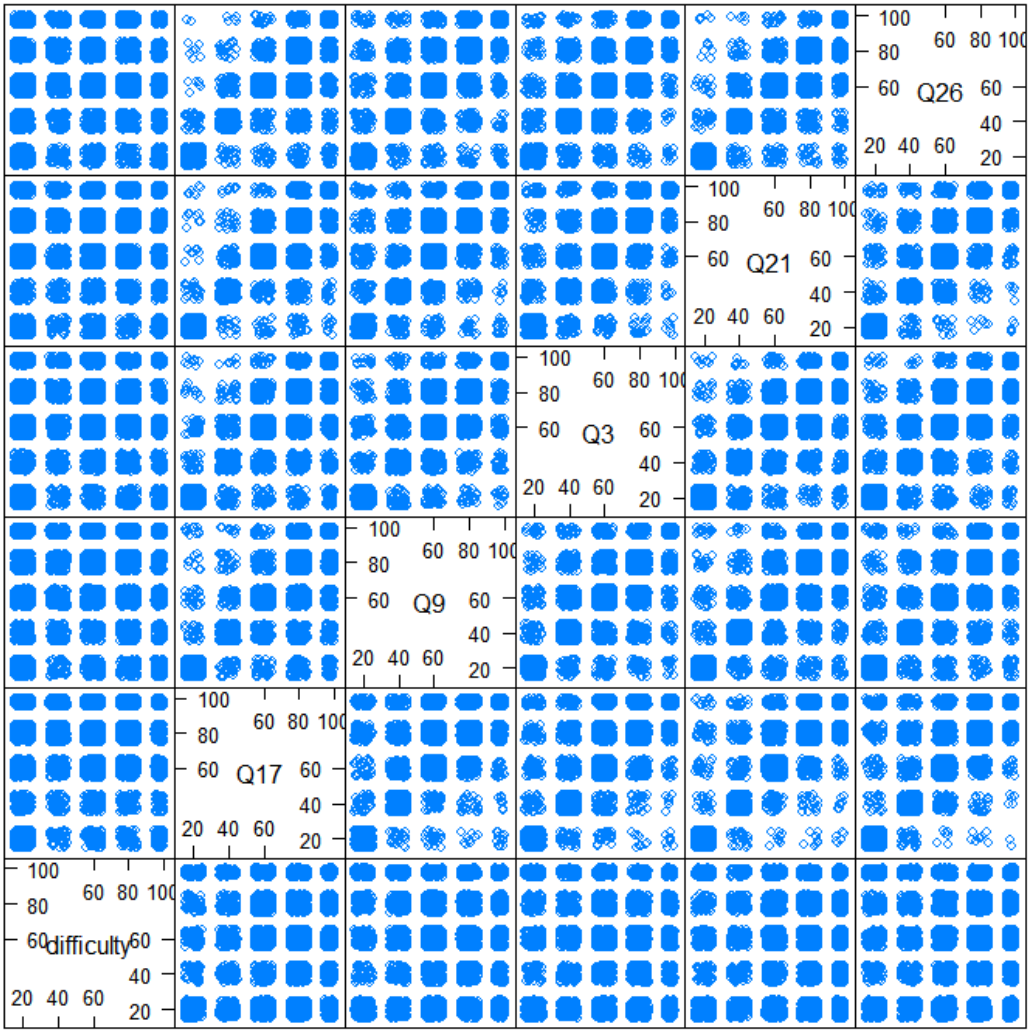

Figure 1.13 A scatter-plot matrix of a part of score data. They are distributed in data grid, representing the numerical variable is discrete.

1.4 outlier detection based on cluster

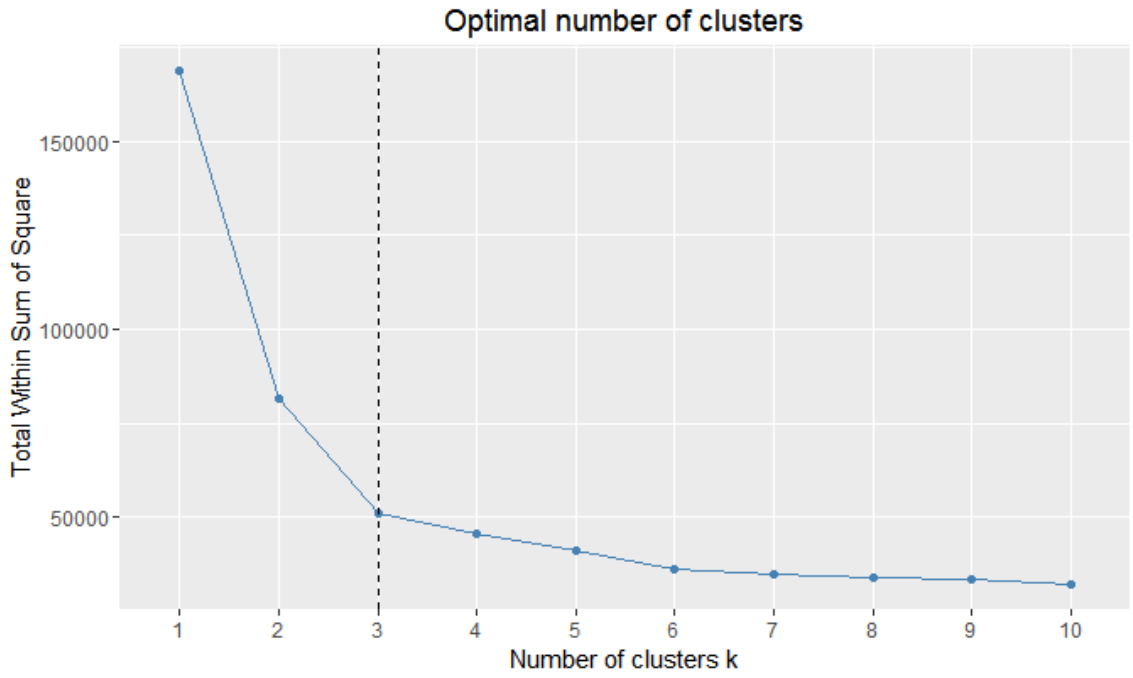

Figure 1.14 Optimal number in the Elbow method for k-means clustering.The optimal cluster number is 3 .

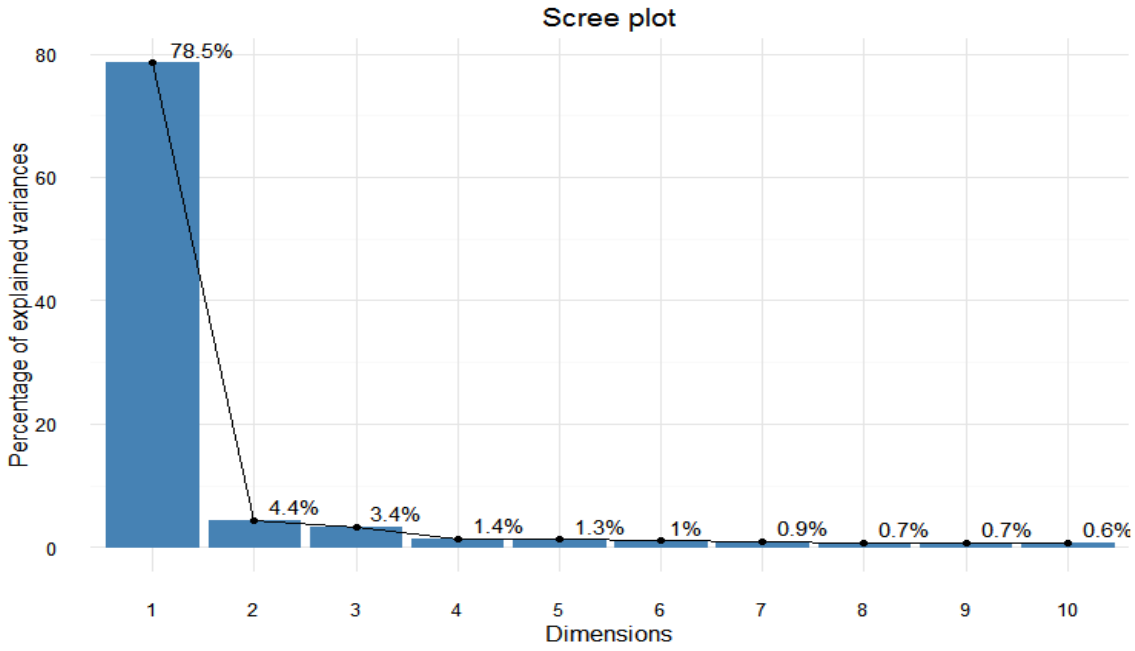

Figure 1.15 The scree plot of principal component analysis.PCA method makes the dimension reduced in the score dataset. The x axis contains the Principal Components sorted by decreasing fraction of total variance explained, the y axis contains the fraction of total variance explained.

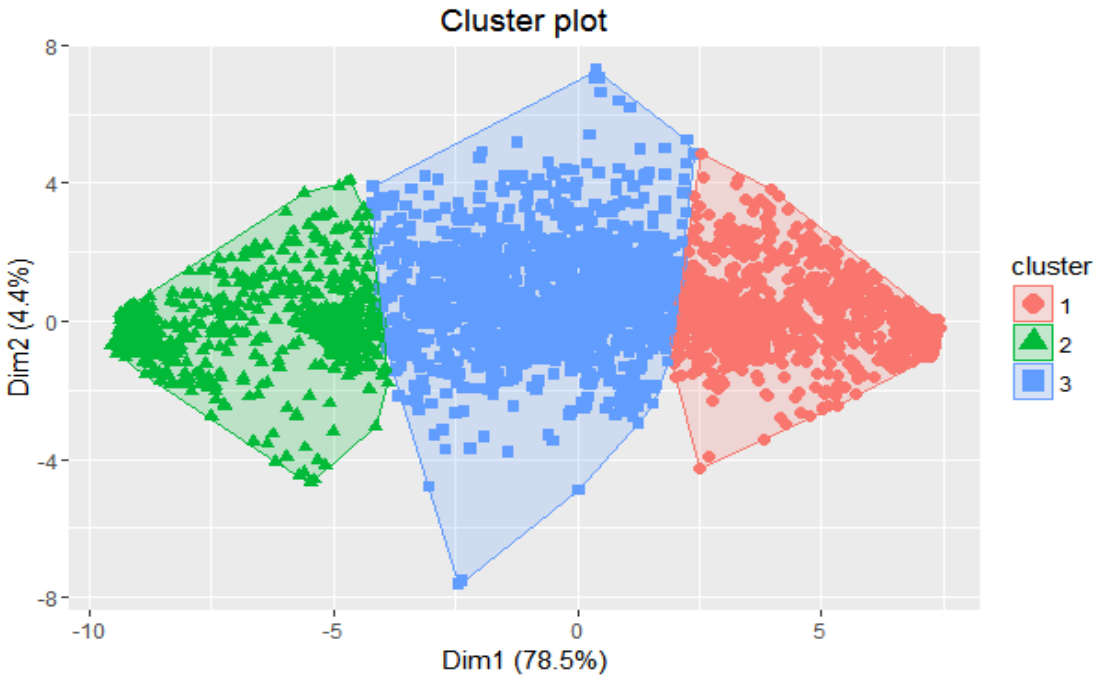

Figure 1.16 Partitioning Clustering Plot based on k-means algorithm.The data is divided into three categories.

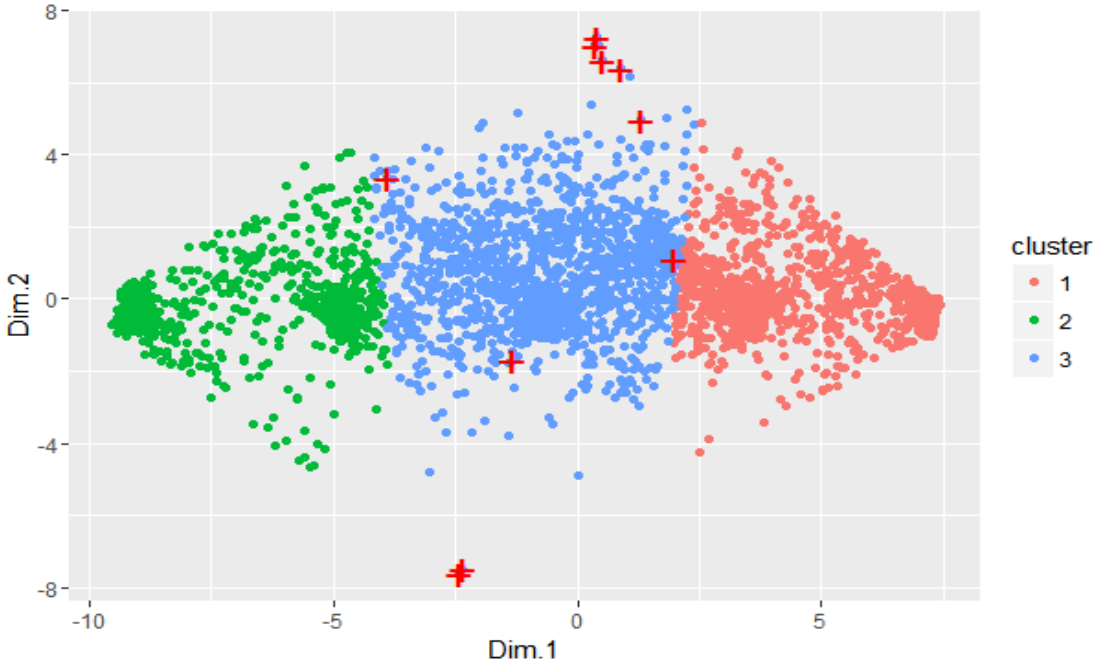

Figure 1.17 Outlier detection based on k-means cluster, showing outliers with a biplot of the first two principal components where outliers are labeled with "+" in red. The *x*-axis represents the first principal component, *y*-axis represents the second principal component.

1.5 Variable importance

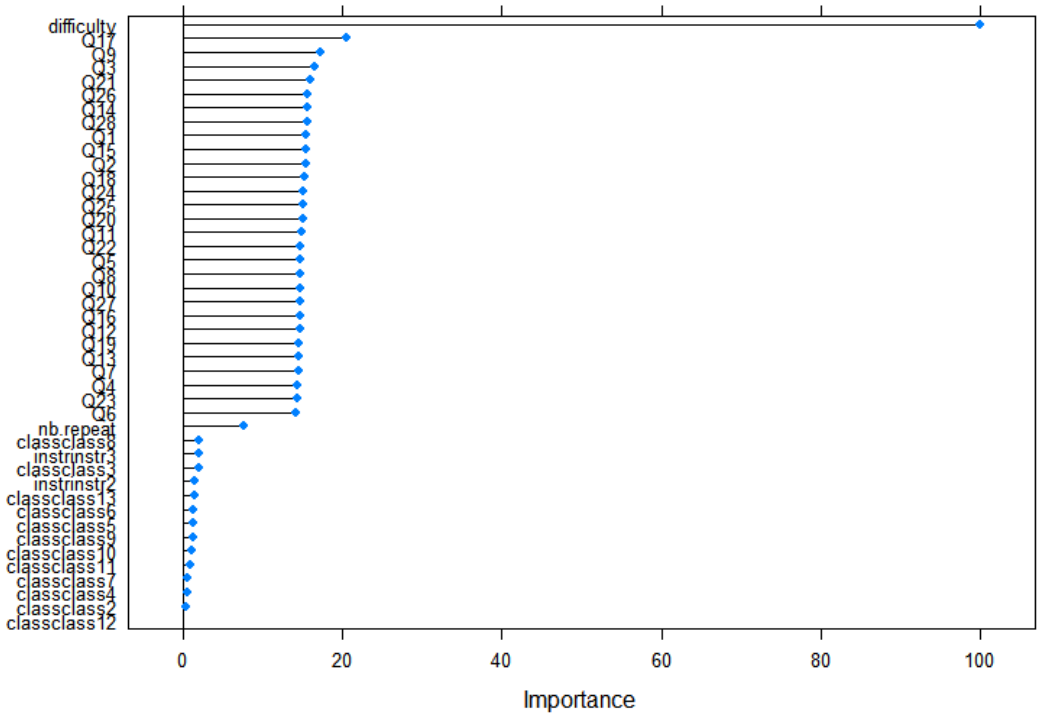

Figure 1.18 Feature importance based on randomForest method with raw data , the level of attendance is regard as response variables.As showed,the level of difficulty of the course have a major impact on the level of attendance.

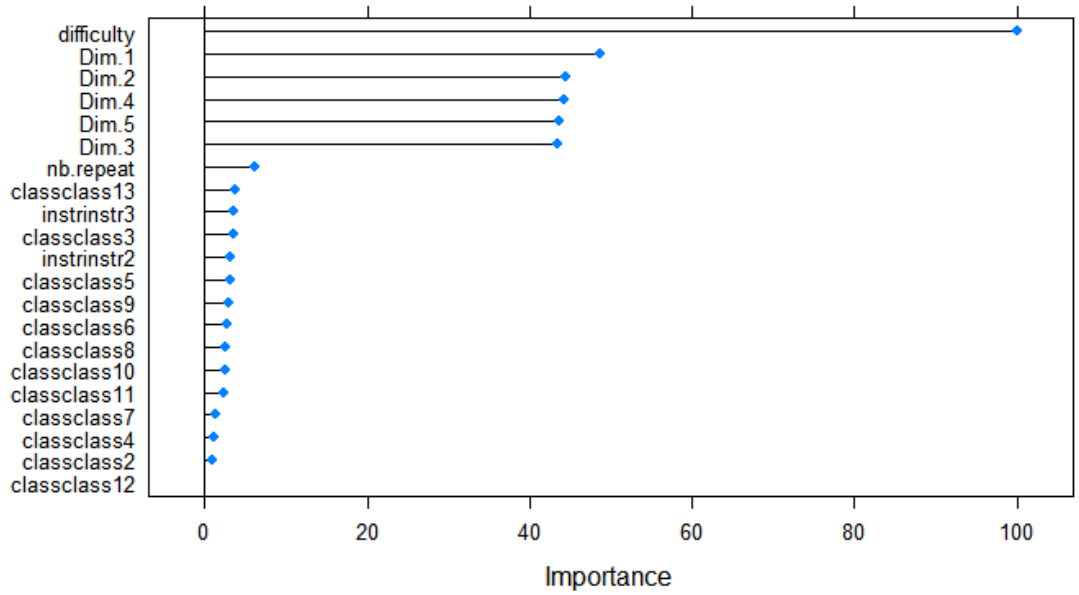

Figure 1.19 Feature importance based on randomForest method with the Q1~Q28 data after principal component analysis , the level of attendance is regard as response variables.As showed,the level of difficulty of the course have a major impact on the level of attendance.

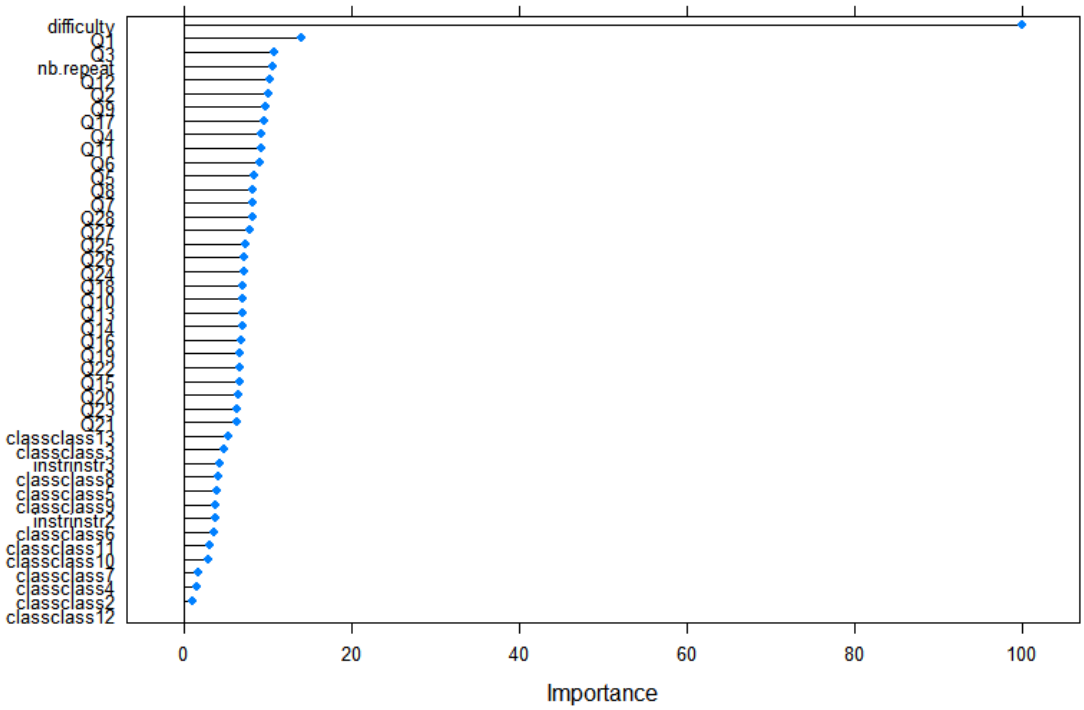

Figure 1.20 Feature importance based on randomForest method with the discreted data on the variables (Q1~Q28).

Q 2

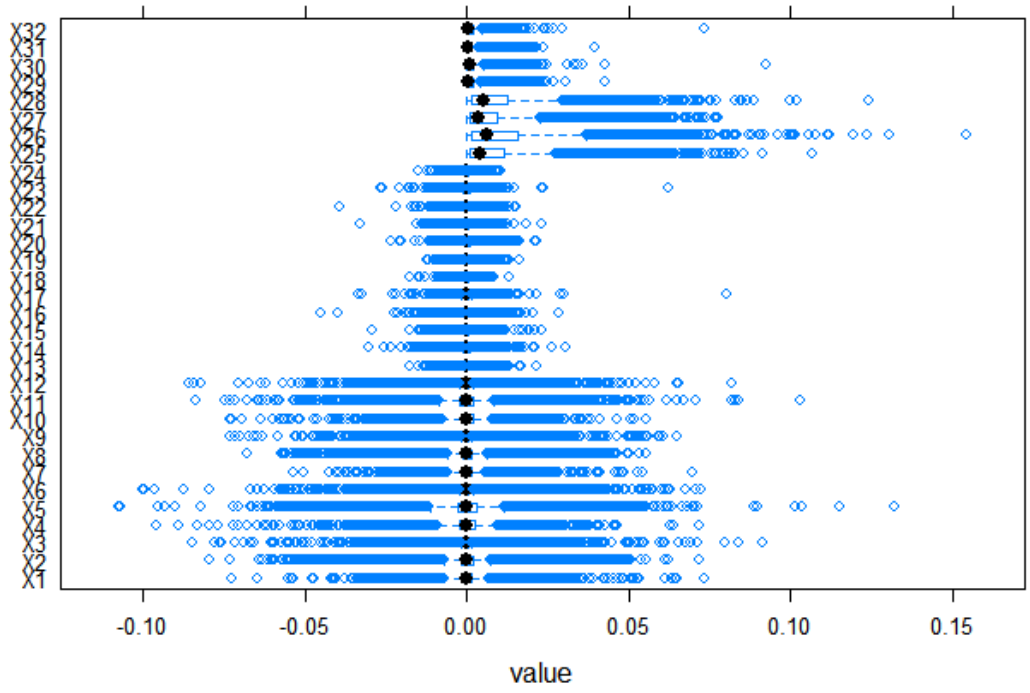

Figure 2.1 Comparative box-and-whisker plots of processed data composed by 7 videos. The average value of each variable is about 0. The plot summarize the data using a few quantiles, and possibly some outliers.The variables can be preliminary divided into three types.(Q1~Q12,Q13~Q24,Q25~Q32)

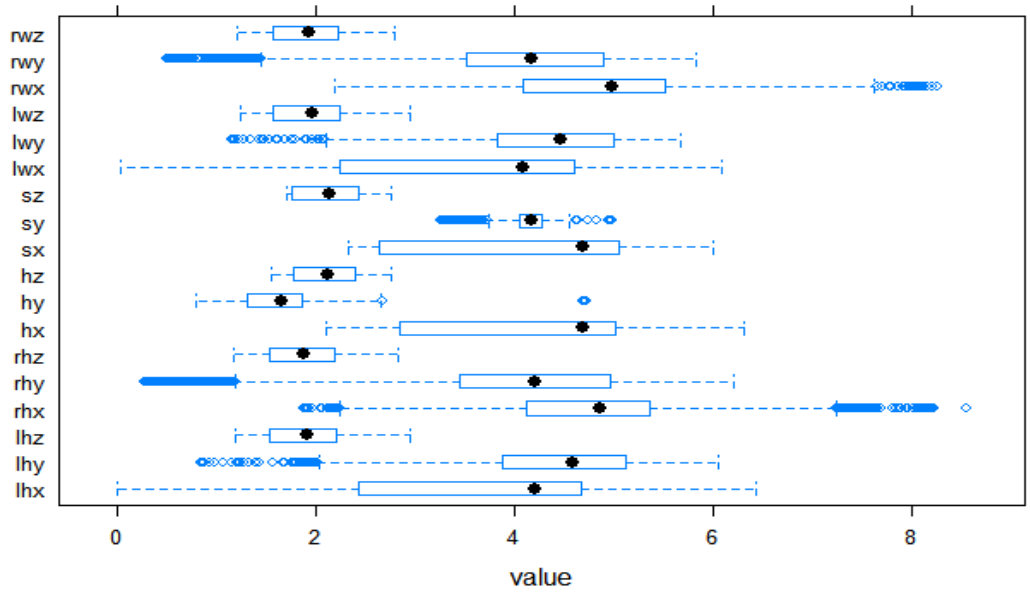

Figure 2.2 Comparative box-and-whisker plots of raw data composed by 7 videos.The means of z coordinate of six articulation points is nearly 2. The means of x and y coordinate of six articulation points is nearly 4, In addition to the y coordinate of Position of head.

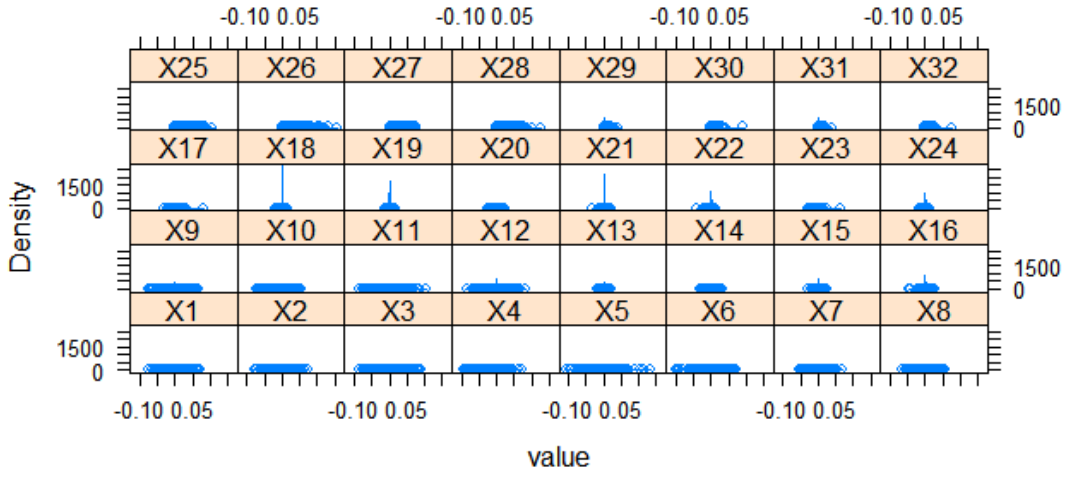

Figure 2.3 A kernel density plot of the processed data composed by 7 videos.

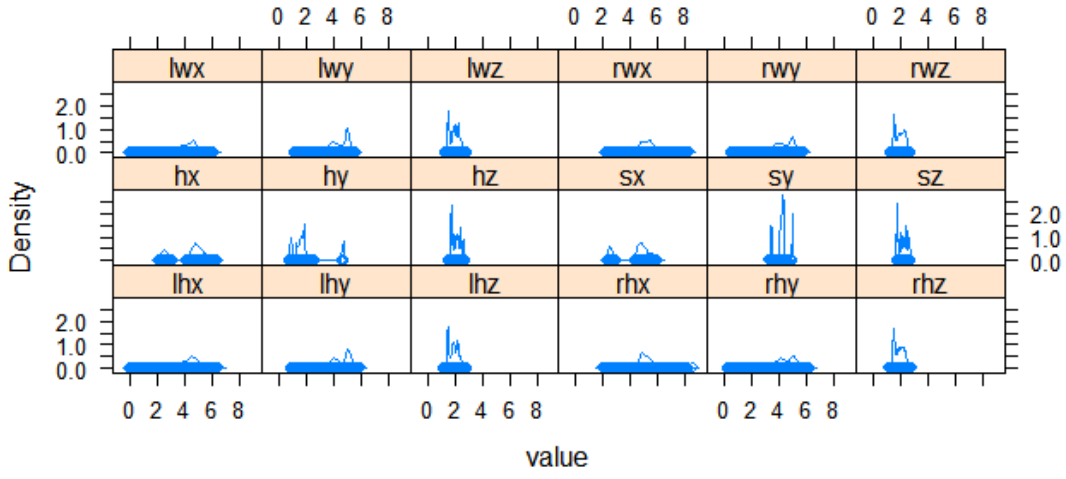

Figure 2.4 A kernel density plot of the raw data composed by 7 videos.

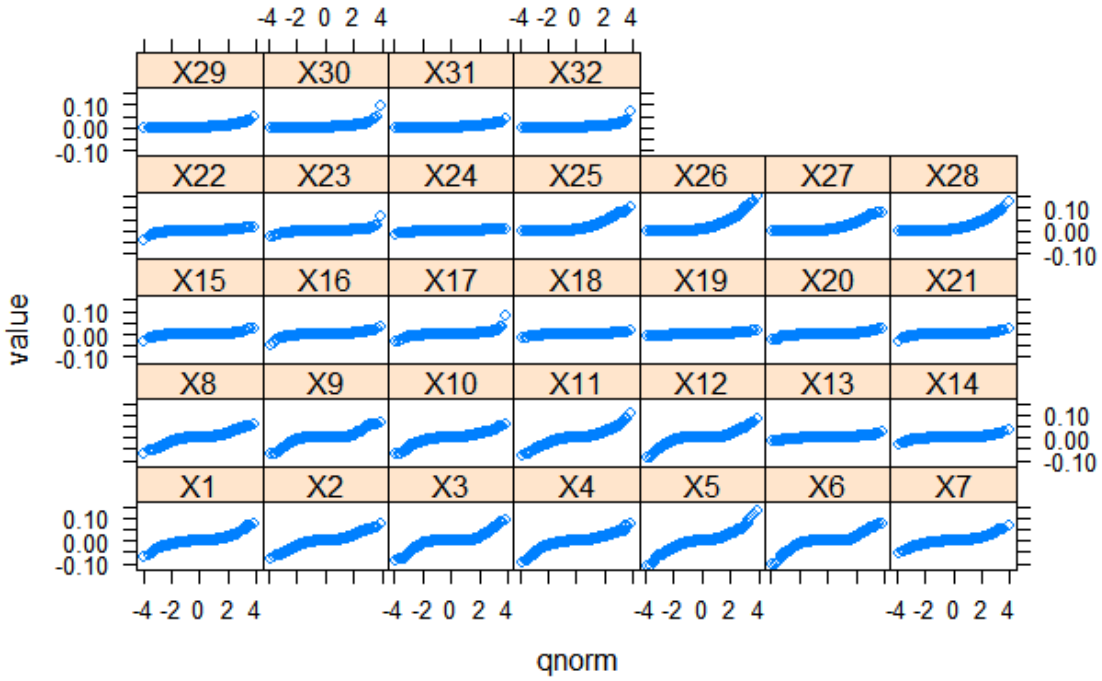

Figure 2.5 Normal Q–Q plots of for processed data composed by 7 videos.

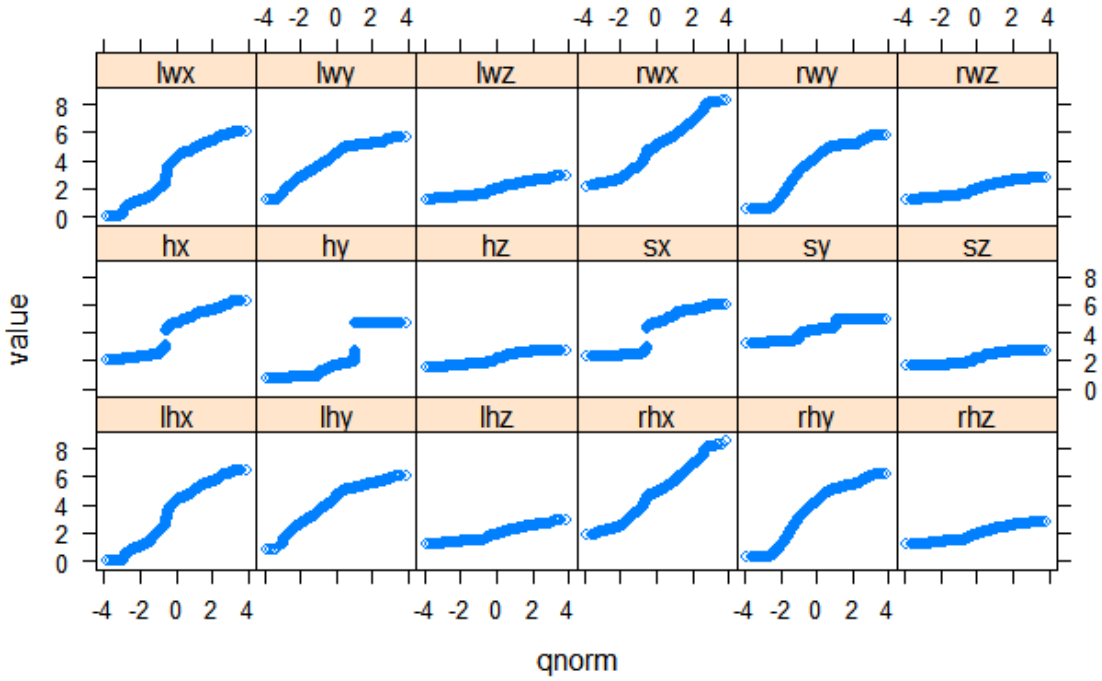

Figure 2.6 Normal Q–Q plots of for raw data composed by 7 videos.

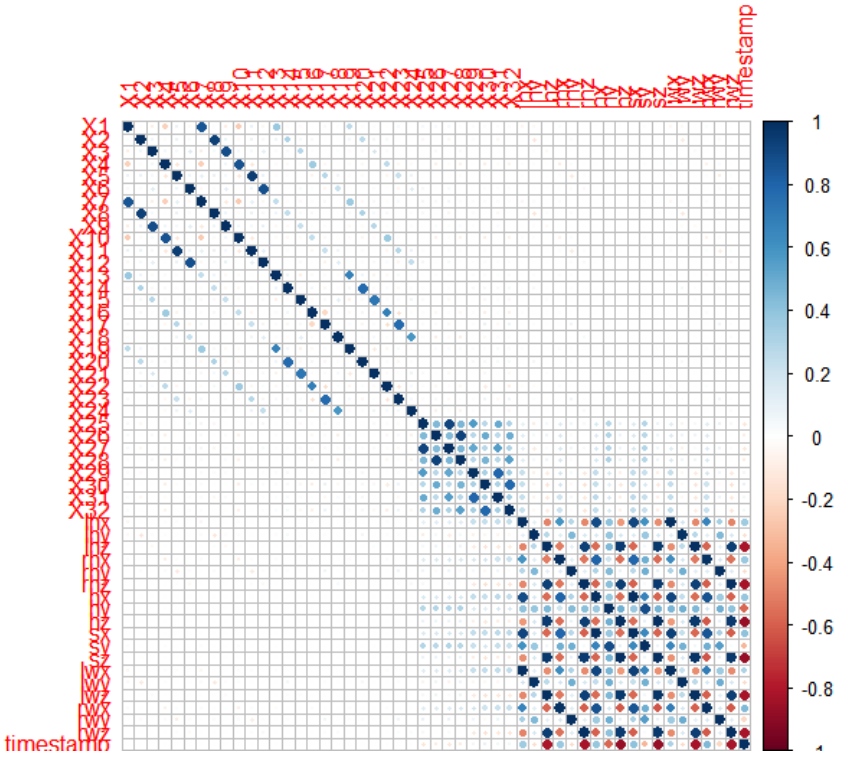

Figure 2.7 Correlation matrix of 51 numeric attributes. It displays a high negative correlation between z coordinate of six articulation points and timestamp,although a high positive correlation between z coordinate of six articulation points. Besides,there are high positive correlation between some variables.

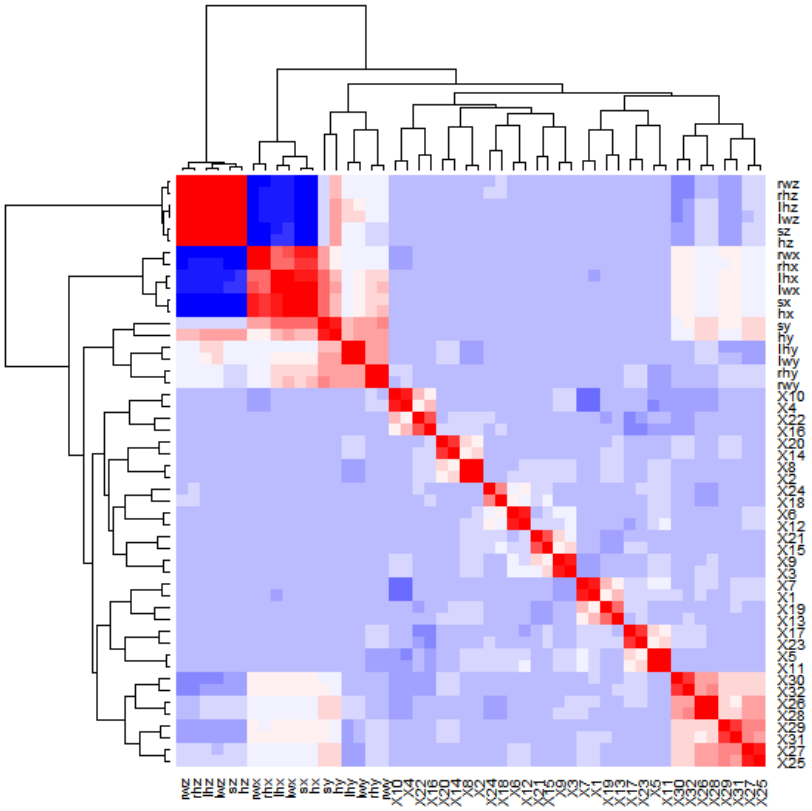

Figure 2.8 A heatmap created with the heatmap function along with a legend representing a hierarchical clustering. The thin strip at the root of the dendrogram represents a grouping of the variables.

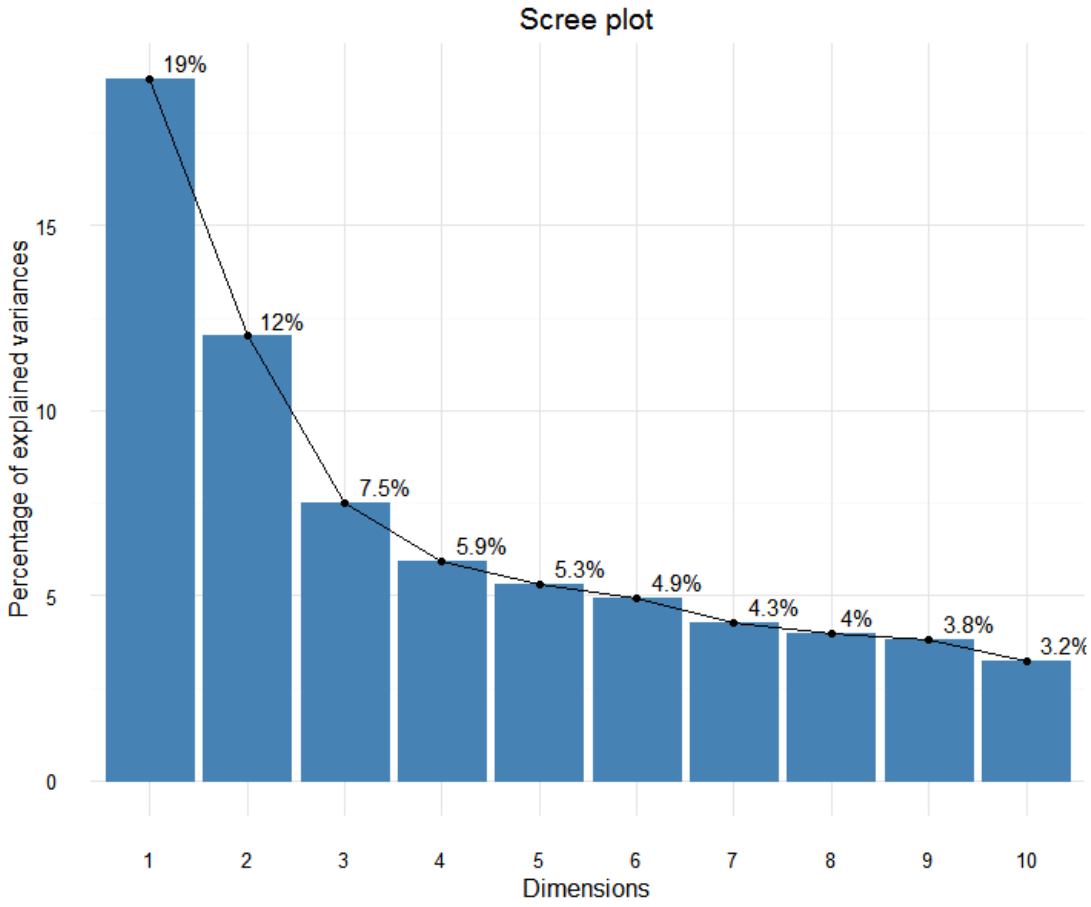

Figure 2.9 Scree plot of principal component analysis, showing the proportion of variance for each principal component The x axis contains the principal components sorted by decreasing fraction of total variance explained, the y axis contains the fraction of total variance explained.

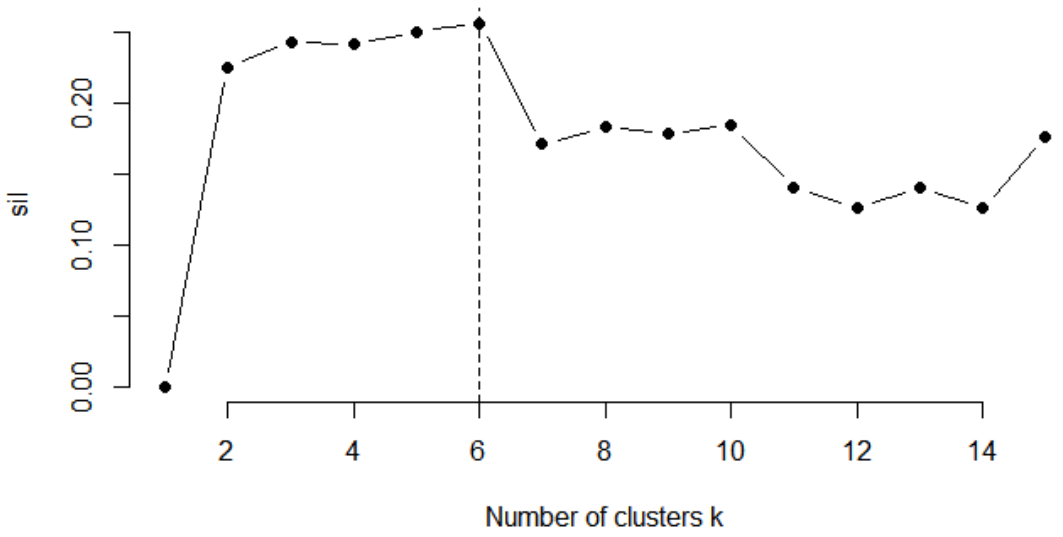

Figure 2.10 Optimal number in the Average silhouette method for k-means clustering.The optimal cluster number is 6 ,although there are totally 5 different kinds of gestures said in the README_gest.txt.

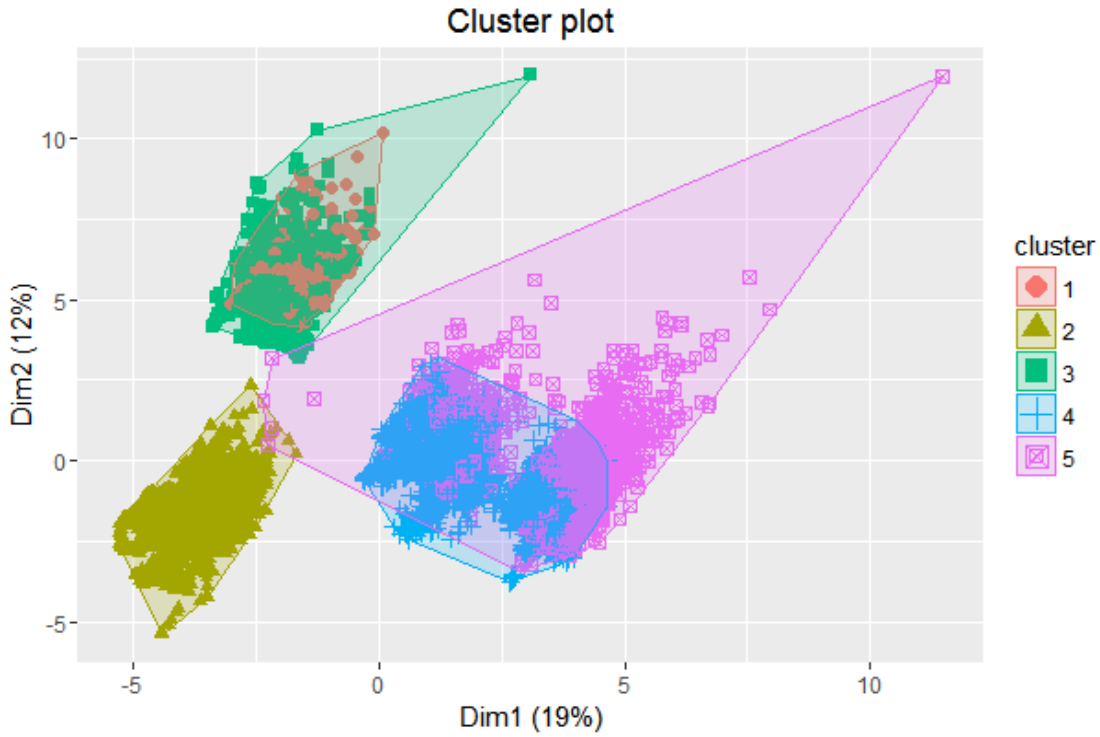

Figure 2.11 Kmeans clustering plot, drawing the result of partitioning by color.

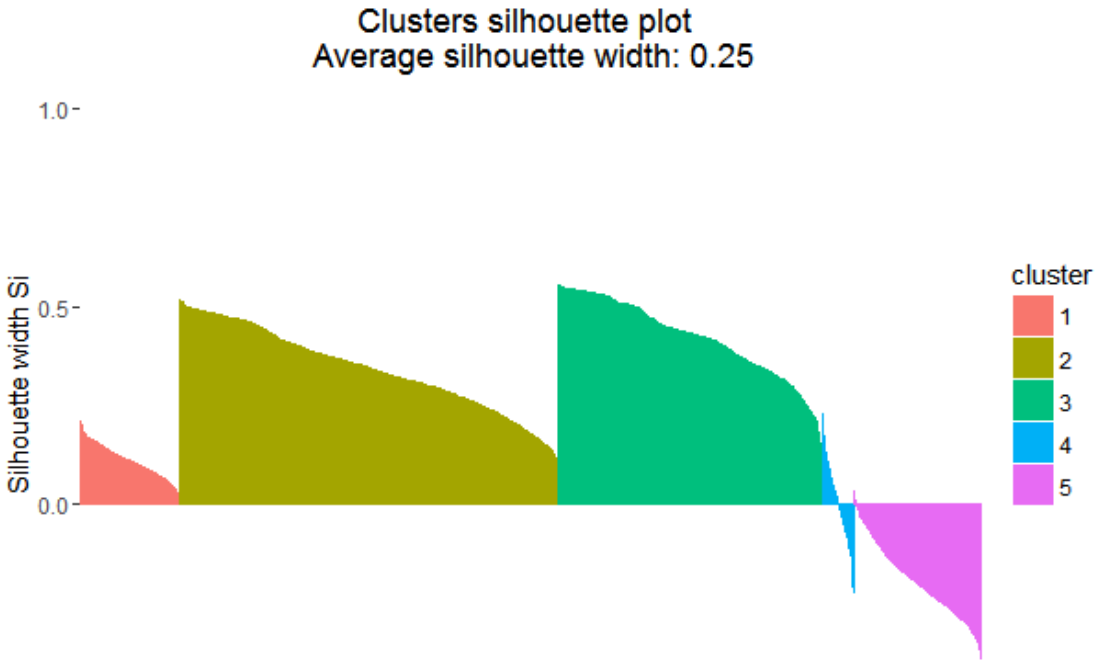

Figure 2.12 Silhouette plot of the kmeans clustering.

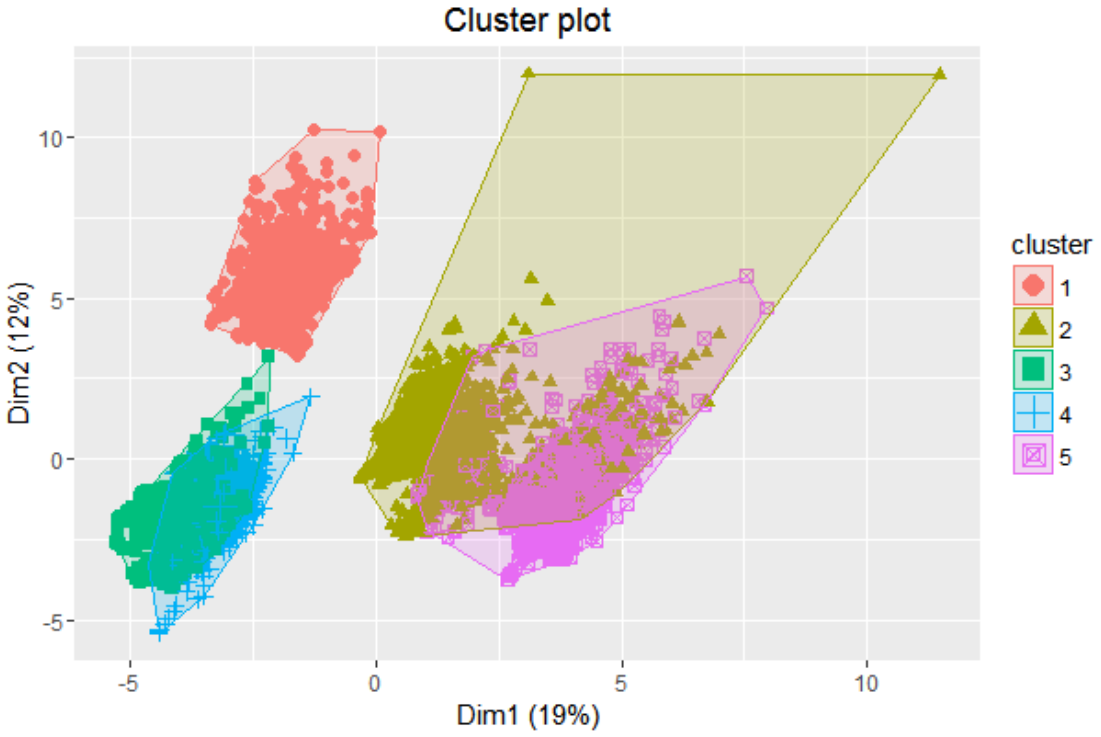

Figure 2.13 PAM clustering Plot .

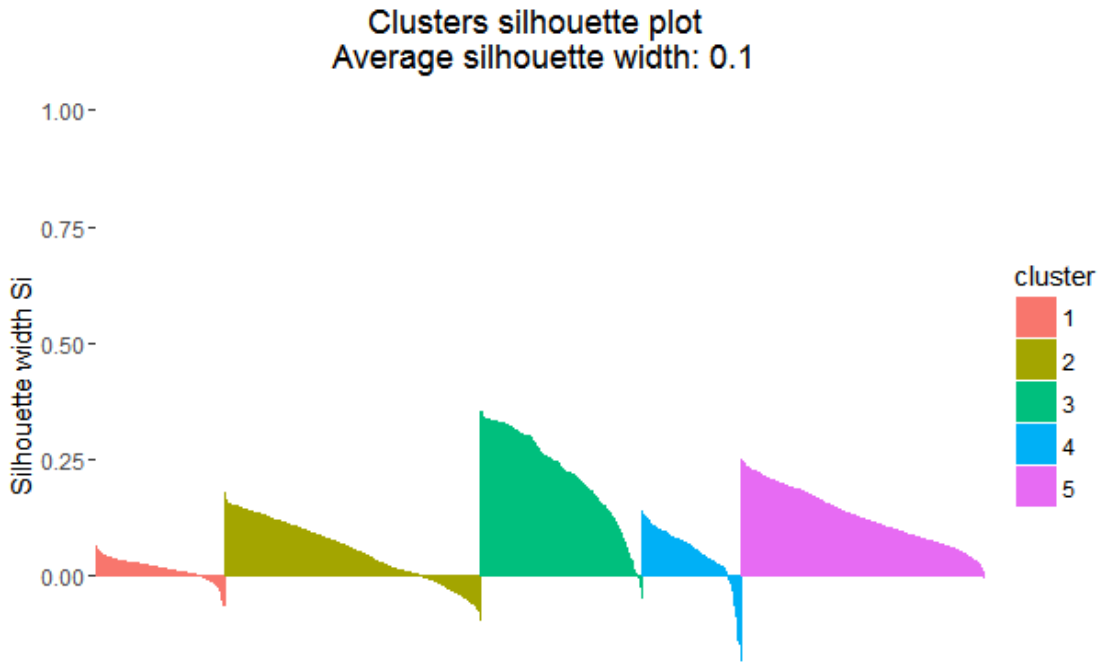

Figure 2.14 Silhouette plot of thePAM clustering.

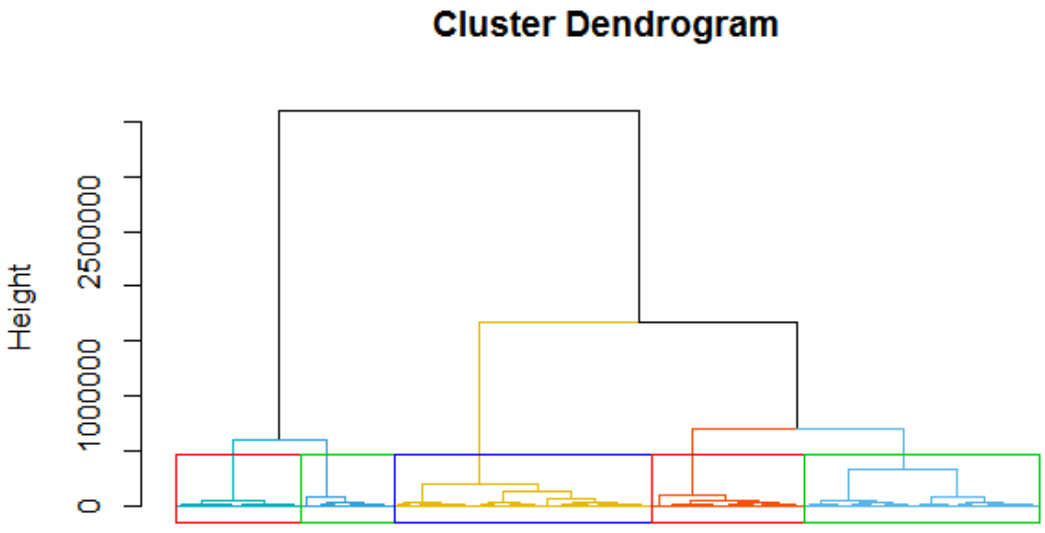

Figure 2.15 Hierarchical clustering Plot .

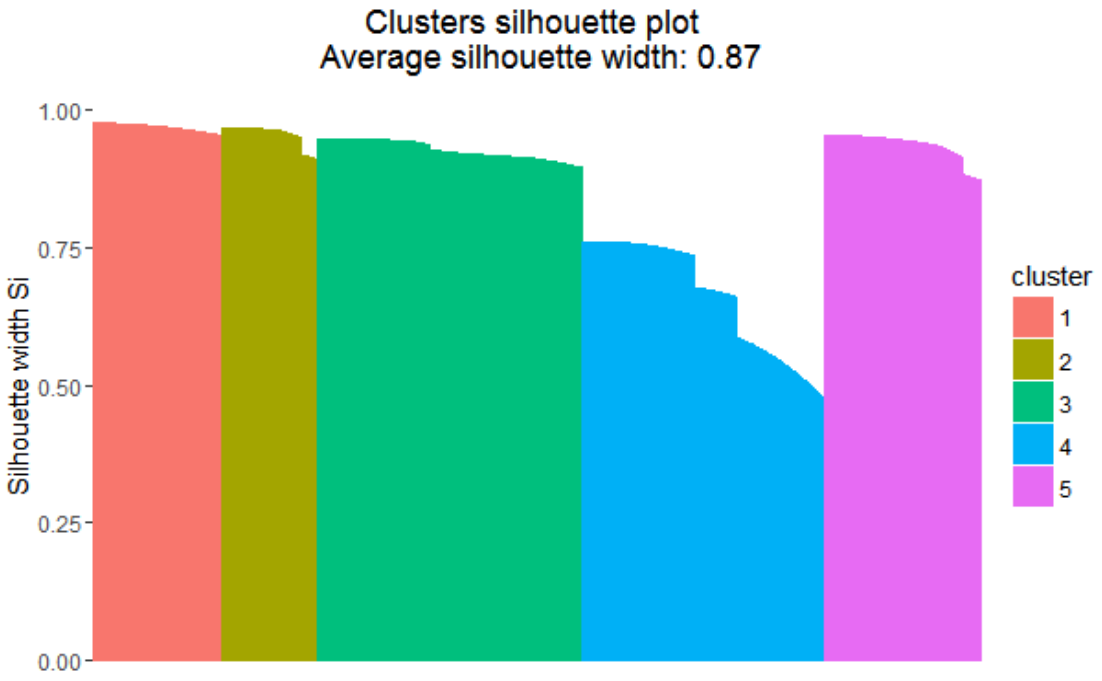

Figure 2.16 Silhouette plot of the Hierarchical clustering.

The silhouette value ranges from -1 to +1. A high silhouette value indicates that it is well-matched to its own cluster, and poorly-matched to neighboring clusters. If most points have a high silhouette value, then the clustering solution is appropriate. If many points have a low or negative silhouette value, then the clustering solution may have either too many or too few clusters. In the Figure 2.12, Figure 2.14,we can observe some negative silhouettes that probably mean wrong assignations, on the contrary,in the Figure 2.16 all the silhouettes are positive and higher than 0.5, thus the best configuration is that with Hierarchical clustering.