3.Data Processing

3.1 Collecting And Sorting out

The problem C provides data of 2977 candidate schools. The CollegeScorecardData has 7804 schools’ data. We sort out the data of candidate schools by the ID. We find that there are 41 school don’t have data. So we try to get these data from the IPEDS. The IPEDS provides various reports and the statistical data, which include year and statistical variable. Because of the types of the data classified difficultly, we download all the data of the IPEDS. Then we sort them out seriously.

We get 913 data from IPEDS eventually. We use ID to matching the 41 schools, which don’t have data. From the analysis result we can know that, the data is missing seriously. The analysis result is showed below:

Table 1

There are many “null” and “PrivacySuppressed” among the data the problem provides. We call them “missing data”. Through analyzing the data carefully, we know that this situation is result from the different institutions, which have different attributes such as the educational system respectively. We divide the schools into different types according to their attributes.

Table 2

3.2 Feature Extraction based on PCA(Principal component analysis)

We extract 28936 data in all about the candidate schools from the CollegeScorecardData. We exclude some schools which are not open according to the variable “CURROPER”. Eventually there are 2936 schools ranking. The data in the CollegeScorecardData are not able to be used for calculating. We exclude these indexes which is useless for us. We get 98 variables in the end,

For integrating the data, we use PCA algorithm to process them. The derailed process of PCA are as follows:

主成分分析过程

From the results, we can know that we could select 14 principal components, whose contribution rate reach to 99%.