Analysis Report

To implement K-means algorithm, I used continuous variables and a set of different clusters (2, 3, 4) to do the exploration.

With the cluster of 2, the dataset is separated into two clusters, the projects with “students_reached” greater than 0.027 are very likely to be clustered, which contains 8533 samles. And if the “total_price_including_optional_support” of projects are less than 0.104 and “students_reached” of projects are less than 0.027, they are likely to be clustered, which contains 116290 samples.

With the cluster of 3, the dataset is separated into three clusters. Even though the number of clusters has increased, the overall distribution of clusters is rarely varied. Whether or not the “students_reached” is greater than 0.027 and “total_price_including_optional_support” is less than 0.104 are still important dividing indicators. If the “students_reached” of project is greater than 0.027, it will be clustered with other 8539 samples. Meanwhile, if the “total_price_including_optional_support” of projects are less than 0.104 and “students_reached” of projects are less than 0.027, they are likely to be clustered, which contains 116290 samples.

With the cluster of 4, the dataset is separated into four clusters. In this case, the most important dividing indicator is whether or not the “students_reached” is greater than 0.011. Meanwhile, since there are four clusters, the subsets has increased and the clustering become more complex and detailed.