Assignment 4 - Clustering Report
by Lilian Huang

We applied k-means clustering to data from DonorsChoose, which is an online charity that allows K-12 schoolteachers to propose projects to benefit their students' education, and to request logistical support for these projects through the donation of materials; donors can contribute any amount to the project of their choice. This clustering algorithm allowed us to identify patterns in the data.

We applied the standard pre-processing necessary to prepare data for clustering - filling in nulls, and converting categorical features to numeric features.
Rather than clustering by every feature in the dataset, we chose to focus on features that were especially of interest: whether the school is a charter school, whether the school is a magnet school, the grade level the school serves, where the school is located (an urban, suburban, or rural area), the level of poverty in the surrounding area, and whether the project qualifies for a match offer that will double the impact of an individual's donation. We believe that this narrower focus allows us to come up with more usefully descriptive clusters.
(We also wrote code that allows the user to interactively merge multiple clusters, recluster with a new desired number of clusters, and split a specific cluster into multiple new clusters. This is demonstrated in the accompanying Jupyter notebook.)

Characteristics of Overall Submitted Projects

We grouped all submitted projects into three clusters, with the following distinguishing characteristics:

Cluster 0 (with 60001 observations) consists of mostly urban schools, located in areas with high poverty levels, serving mostly younger children - from pre-K to grade 2, and some from grades 3 to 5.

Cluster 1 (29609 observations) consists of suburban and some rural schools, in areas with the lowest poverty levels, serving mostly younger children from pre-K to grade 2, and some from grades 3 to 5.

Cluster 2 (35366 observations) consists of primarily urban schools but has a larger proportion of suburban schools than cluster 0; these schools are located in areas with varying poverty levels, and serve older children from grades 6 to 12. This cluster also contains a larger proportion of charter and magnet schools than the others.

Characteristics of Top 5% Predicted Projects Unlikely to be Fully Funded

We then used a logistic regression model to predict which projects were unlikely to be fully funded, and identified the top 5% of those projects.
We grouped that subset of high-risk projects into three clusters as well, and found the following distinguishing characteristics:

Cluster 0 (with 642 observations) consists of non-charter and non-magnet schools in high-poverty rural areas, serving primarily younger children from pre-K to grade 2.

Cluster 1 (with 537 observations) consists of schools in lower-poverty rural areas, serving primarily younger children from pre-K to grade 2. This cluster also contains very few magnet and charter schools, whereas cluster 0 contained none at all.

Cluster 2 (with 383 observations) consists of schools in medium-poverty rural areas, serving primarily older children, mostly from grades 6 to 8, with some from grades 9 to 12.