Explore Public Education Donation Data

ykang84@gatech.edu

ABSTRACT

This research aims to help better understand public education donation by investigating data provided by DonorsChoose.org. The goal is to investigate the relationships between donors and the projects that can potentially motivate them. Exploratory data analyses were implemented to provide insights on donations, donors, teachers, schools, projects and resources. Content-based recommendation systems were applied to discover similar projects that can attract donors.

Author Keywords

DonorsChoose; public education; natural language processing; data analysis

INTRODUCTION

DonorsChoose.org, founded in 2000 by a history teacher, is a website that allows crowdfunding to make a difference in public education. It has raised $685 million for America's classrooms so far [1]. Teachers at 79% of public schools have posted projects that need help on DonorsChoose. Stephen Colbert, the well-known talk show host, has endorsed the platform, become a board member, and funded all the classroom projects in South Carolina in 2015 [2].

On this platform, classroom causes can be sorted by location, materials requested or greatest need. Profiles of each classroom include itemized requested supplies, statement proposed by the instructor and information of the school. Once a user selects a classroom, Donors Choose facilitates the donation process, easily and efficiently connecting needy teachers and students with gracious benefactors. This is similar to building a recommendation system such as Netflix, YouTube and Amazon [3,4].

History data between 2013 and 2018 has been made publicly available. The data included information about projects, schools, teachers, resources, donations and donors. This research-oriented work aims to investigate a problem proposed by DonorsChoose: how to efficiently connect donors to the projects that motivate them. The analysis results can be utilized to enhance the performance of this platform. Improving public awareness of a platform like this will also substantially contribute to public education.

The paper is organized as follows. The next section summarizes previous research in regards to education data analysis. Section 3 presents the exploratory data analysis and visualization. Section 4 introduces the details of recommendation approaches. Section 5 discusses the findings, and concludes the paper.

LITERATURE REVIEW

Education data can be studied in various ways to help improve the quality of education, and provide people a great opportunity to consider where we are headed. However, there was limited literature specifically targeted at public education donations in particular. Instead, existing studies focused on education-related fundraising were reviewed.

Liang used various machine learning approaches, including Gaussian Naive Bayes, random forest, and support vector machine algorithms, to study fundraising success in higher education [5]. These algorithms were able to distinguish promising donors from non-promising donors, at an overall accuracy of 97%.

Wastyn did a qualitative analysis based on his 14 years of expertise in fundraising [6]. Her study concluded that where donors and non-donors differed was in the ways in which they socially constructed their college experiences to create their own realities. However, such research can be biased and subjective depending on the researcher. Hence, its value may be limited to a specific circumstance.

Wesley and Christopher used statistical logit regression analysis to predict the individuals who would give higher (e.g., $100,000) or lower ($1,000) donations based on the data from the alumni database as well as the geo-demographic information [7]. The models were developed from the alumni database at Northwestern University for both major gifts and annual fund prospects.

To summarize, there is a difference between the existing studies and the current study. Previous studies were mainly focused on identifying promising donors, while the current study aims to explore relationships between multiple groups.

EXPLORATORY DATA ANALYSIS AND VISUALIZATION

The data was 3.86 GB, including 4.68 million donations, 7.21 donated resources, 1.11 million projects and etc., as shown in Table 1. Unique IDs were created for each teacher, school, donor, donation and project, respectively. The data only included projects that received at least one donations.

Data	Instances	Features
Donations	4,687,884	7
Donors	2,122,640	5
Resources	7,210,448	5
Schools	72,993	9
Teachers	402,900	3
Projects	1,110,017	18

Table 1. Data structure

The first step is check for missing values. Considering the size of the data, missing values are not issues for most columns. "Project Fully Funded Date" has 25.52% of missing rows. This could be because of expired projects or projects that have not been funded yet. So it is reasonable to keep these rows for now. The second is to check for duplicates. Some donations that share donation ID, project ID and donor ID but had difference received date were excluded. For the purpose of presentation, these IDs were replaced with indices. Subsequently, the exploratory data analyses are presented in the following section.

Teachers

In total there are 402,900 teachers registered. The data has the format of:

Teacher ID	Teacher Prefix	Teacher First Project Posted Date
1	Mrs.	2013-08-21

Table 2. Example of teacher

According to National Center for Education Statistics, about 76% of public school teachers were female in 2012 [8]. In the users of DonorsChoose, 87.99% of the teachers were female, as shown in Figure 1.

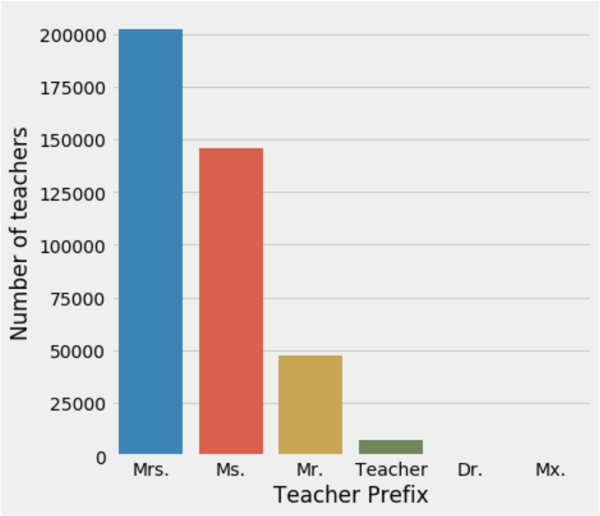

Figure 1. Teacher Prefix

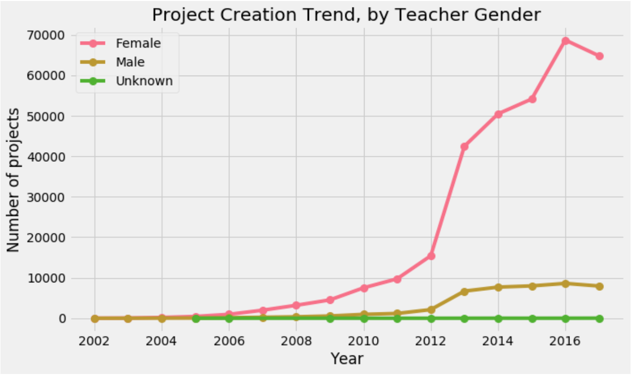

Figure 2. Project Creation by Gender over Time

The increased number of female teachers is also much higher than the male teachers, while the reason remains unknown based on the limited information. As for the date that a teacher post project for the first time, it is related to the need of drawing the teachers to ask for help. Figure 3 below shows an increasing trend in the fall semester, especially from August to November.

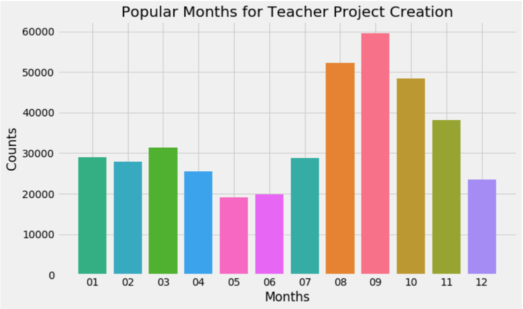

Figure 3. First Project Creation Date

Schools

There are 72,993 schools that were involved with at least one posted project or teacher. The table has the format of:

School ID	School Name	School Metro Type	School Percentage Free Lunch	School State	School Zip	School City	School County	School District
1	Capon Bridge Middle School	rural	56.0	West Virginia	26711	Capon Bridge	Hampshire	Hampshire Co School District

Table 2. Example of School

Figure 4 shows the distribution of school metro types. Suburban and urban schools consist of the majority of the schools.

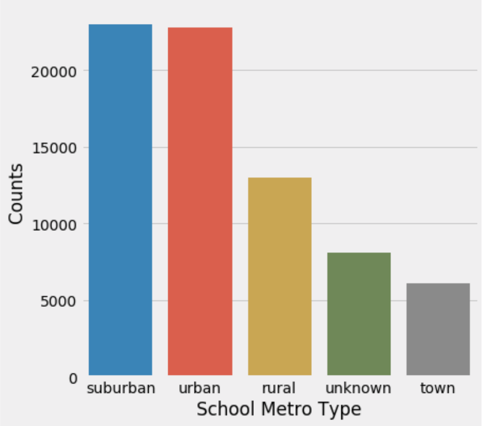

Figure 4. School Metro Type

The feature "School Percentage Free Lunch" could be related to the funding status of the school. To make comparisons between states, the top five states that have the highest average percentages of free lunch are District of Columbia, Mississippi, Louisiana, New Mexico and Oklahoma, as summarized in Table 4 below.

State	Count	Mean	Median	Stand Dev.
District of Columbia	155	87.85%	95%	18.80
Mississippi	833	77.82%	81%	17.70
Louisiana	1192	74.33%	79%	18.56
New Mexico	538	73.83%	75%	20.64
Oklahoma	1323	67.53%	70%	20.17

Table 4. Statistics for Free Lunch Percentage

The columns "State", "City", "County" and "District" are redundant since there is "School Zip". Considering the scope of this research, it is more convenient to make inferences using "State". Examples will be demonstrated in the projects section.

Resources

Resources are what the project is for from a certain vendor. The format of resources is:

Project ID	Resource item name	Quantity	Unit price	Vendor name
1	Chair move and store cart	1	350	NaN
2	Sony mdr zx100 blk headphones	40	12.86	CDW-G
3	Gaiam kids stay-n-play balance ball, grey	4	19	Amazon Business

Table 5. Examples of Resources

The table of resources record all requested items, and the quantity, unit price and the vendor name. As shown in Figure 5, Amazon Business received a massive amount of orders. Deals with Amazon Business can be beneficial to both the platform and the vendor.

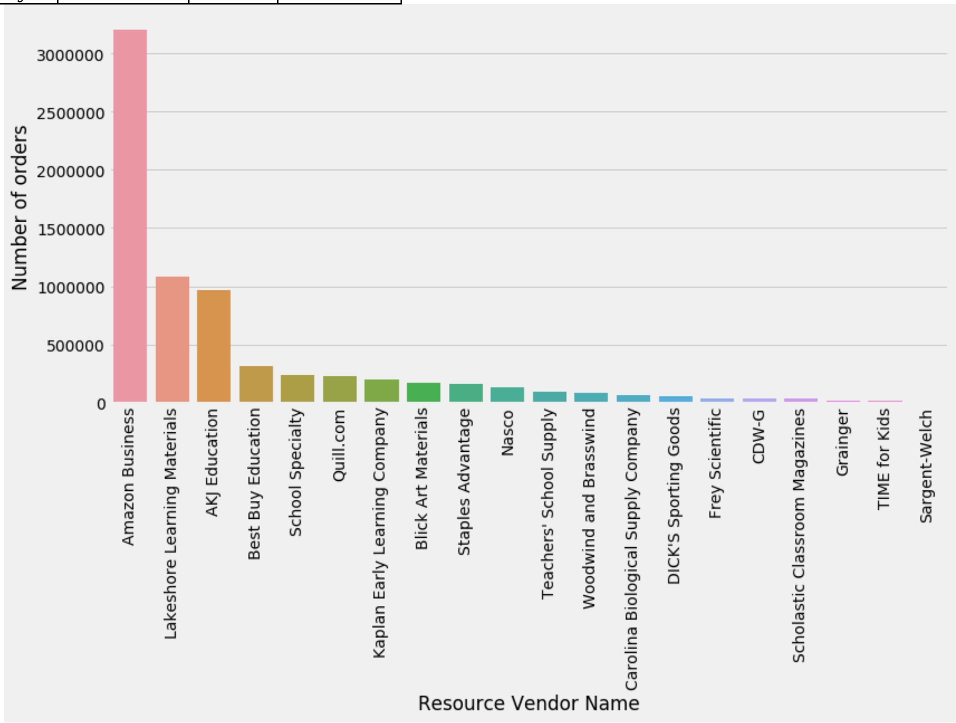

Figure 5. Resource Vendors

Donors

There were more than 2 million donors that have contributed to public education through this platform. This table has the format of:

Donor ID	Donor City	Donor State	Donor is teacher	Donor Zip
1	Evanston	Illinois	No	602

Table 6. Example of Donor

The data allows to identify if a donor is also a teacher since the platform may be more known to the community. Figure 6 shows the majority of the donors are in fact not teachers, which indicates the public awareness and the influence of this particular platform.

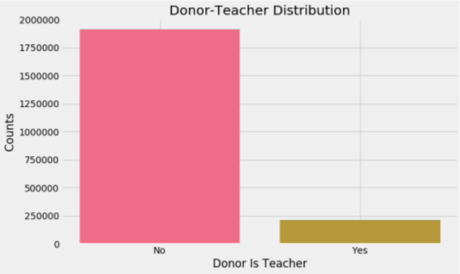

Figure 6. Donor is Teacher or Not

By aggregating the donors by each state, Figure 7 shows the population of donors in each state. It did not consider the base population.

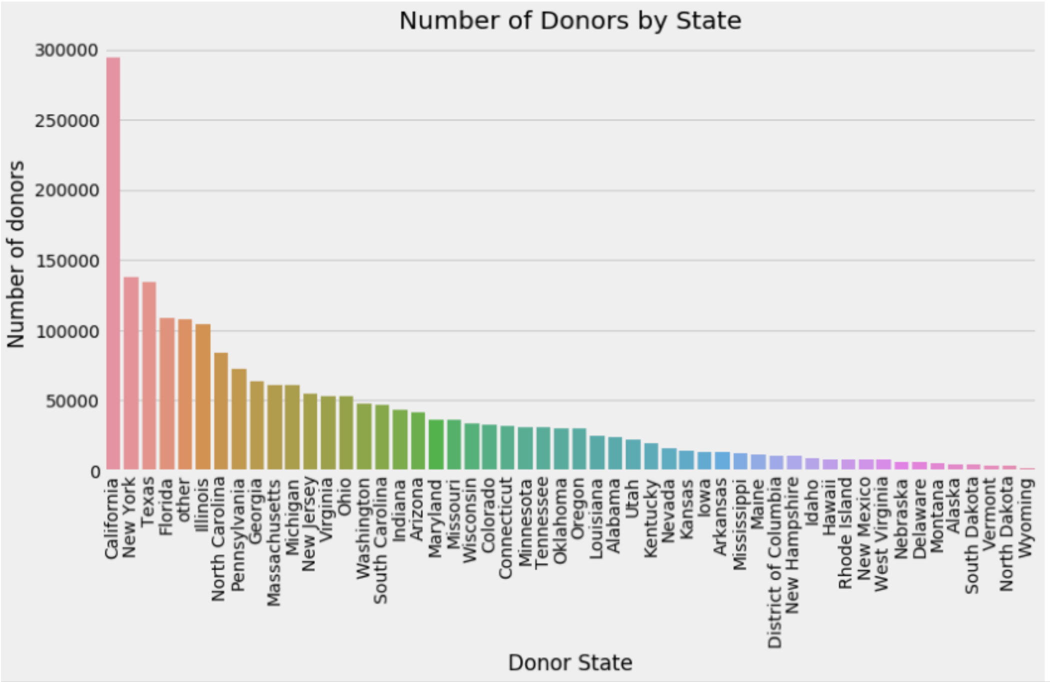

Figure 7. Donor Population by State

The top states in Figure 7 have great advantages in the base population. Hence the population of year 2017 was included and used to calculate the donor rate per 100,000 people [9]. Considering the base population, the top five states that have the highest donor rate are District of Columbia, South Carolina, Massachusetts, Connecticut and Maine, as shown in Table 7 and Figure 8 below.

Donor State	Number of Donors	2017 population	Donor per 100,000
District of Columbia	10862	693972	1565.19
South Carolina	47043	5024369	936.30
Massachusetts	60730	6859819	885.30
Connecticut	31604	3588184	880.78
Maine	11486	1335907	859.79

Table 7. Top 5 States by Donor Rate

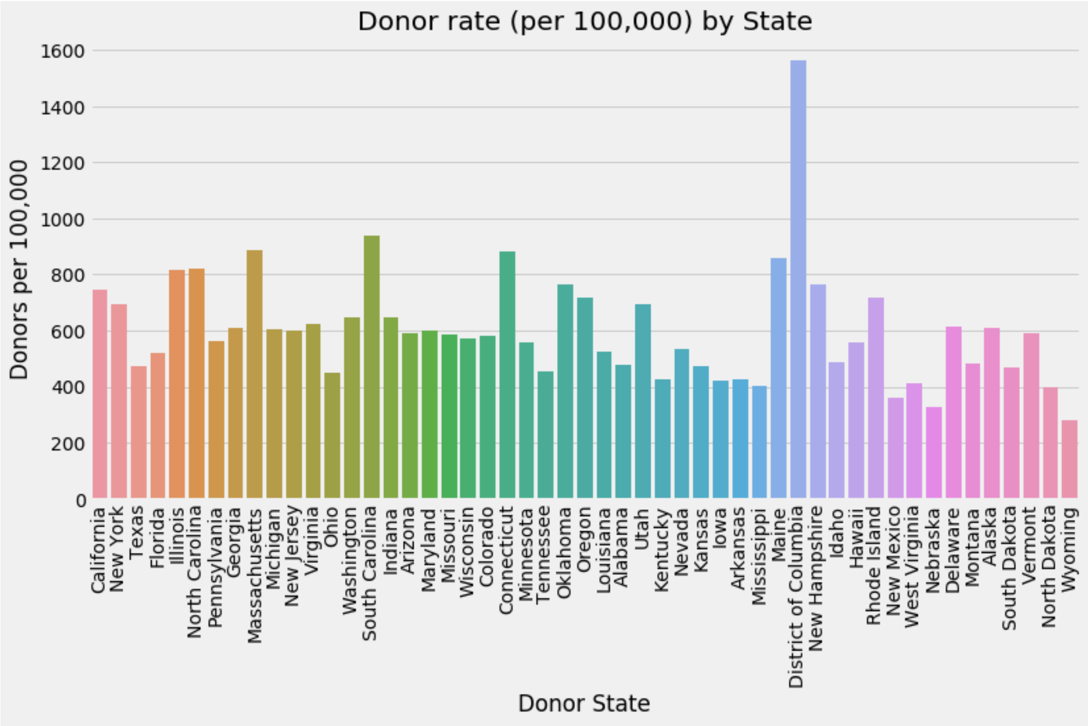

Figure 8. Donor Rate

Donation

Every single donation has a donor ID, project ID and etc., which allows to identify returning donors. The format of the data is:

Project ID	Donation ID	Donor ID	Included optional donation	Donation Amount	Donor Cart Sequence	Received date
1	1	1	No	178.37	11	2016-06-23 13:15:57

Table 8. Example of Donation

Increasing the percentages of returning donors is as important as attracting new donors. Among the 2,122,640 donors, 552,936 donors (26.04%) have made more than one donation. Figure 9 shows more than 85% of the donations have included optional donation to DonorsChoose, as a gesture of encouragement and appreciation to the platform. These signs show a great potential of developing more repeat donors.

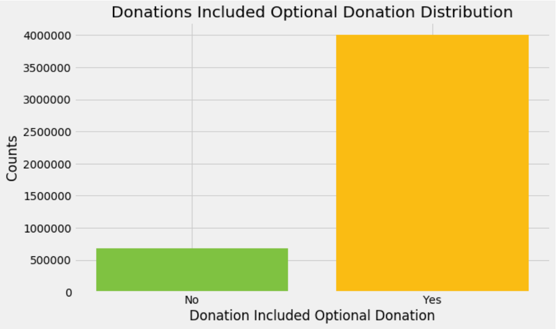

Figure 9. Optional Donation

In all the donations, we found that: the donated amount had a minimum of $ 0.01, a maximum of $ 60000, a mean of $ 60.65 and a $ 25.0 median. It is possible the $25 is the default donation amount provided to donors.

The amount of donations received annually has been steadily increasing since 2013. The top 5 states that donated the most amount are California, New York, Texas, Illinois and Florida.

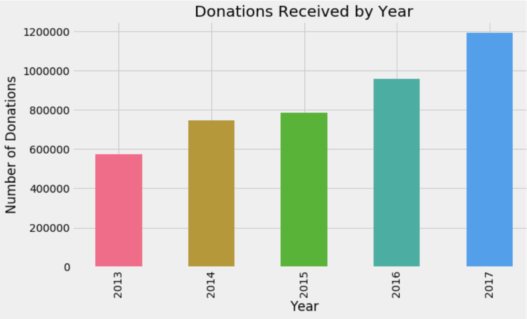

Figure 10. Donations by Year

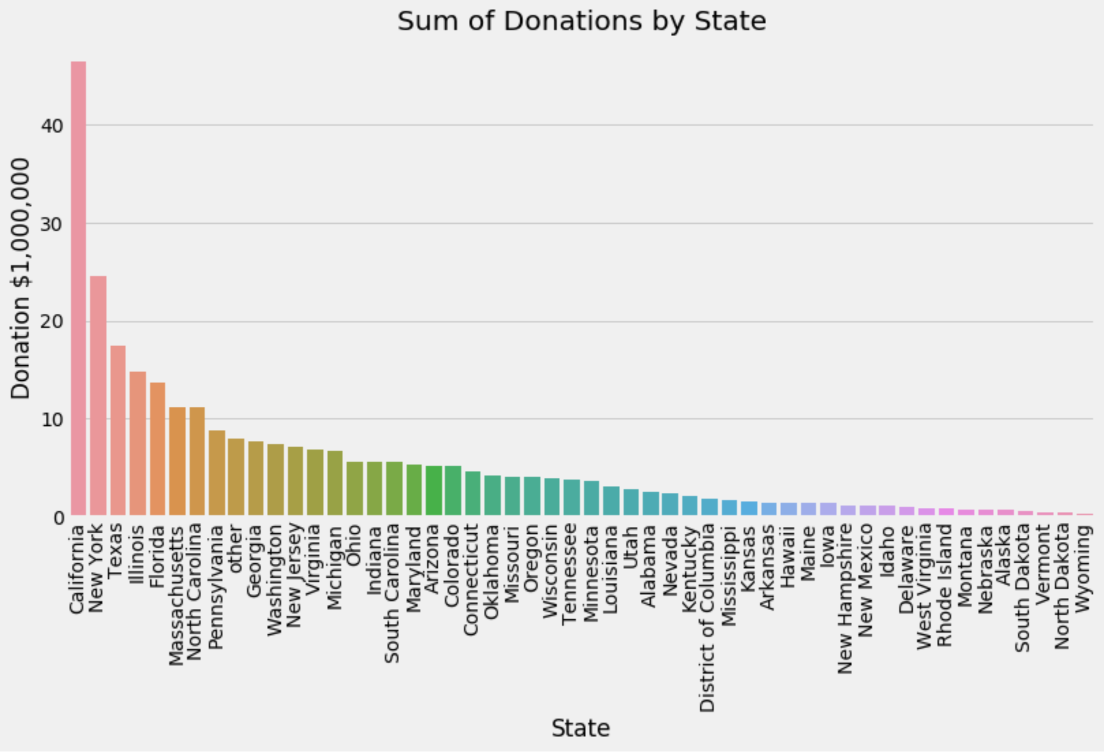

Figure 11. Sum of Donation by State

Another interesting aspect is to examine the relationship between the state of the donor and which state his/her donation went to. This intuitive idea can be tested by implementing Sankey diagrams. Considering the large size of the data, flows of donations that were above $100,000 and $200,000 are presented in the plots below. The states on the left and right are the origins and destinations, respectively.

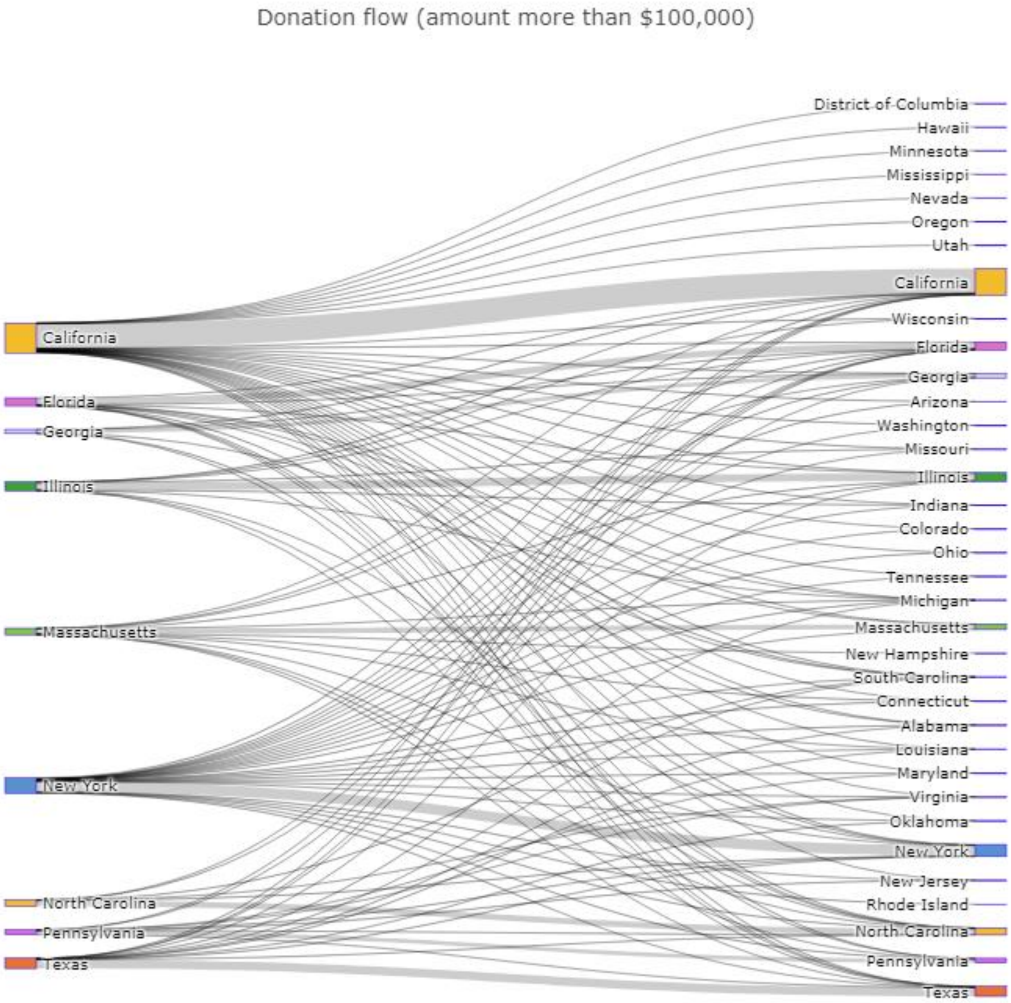

Figure 12. Donations over $100,000

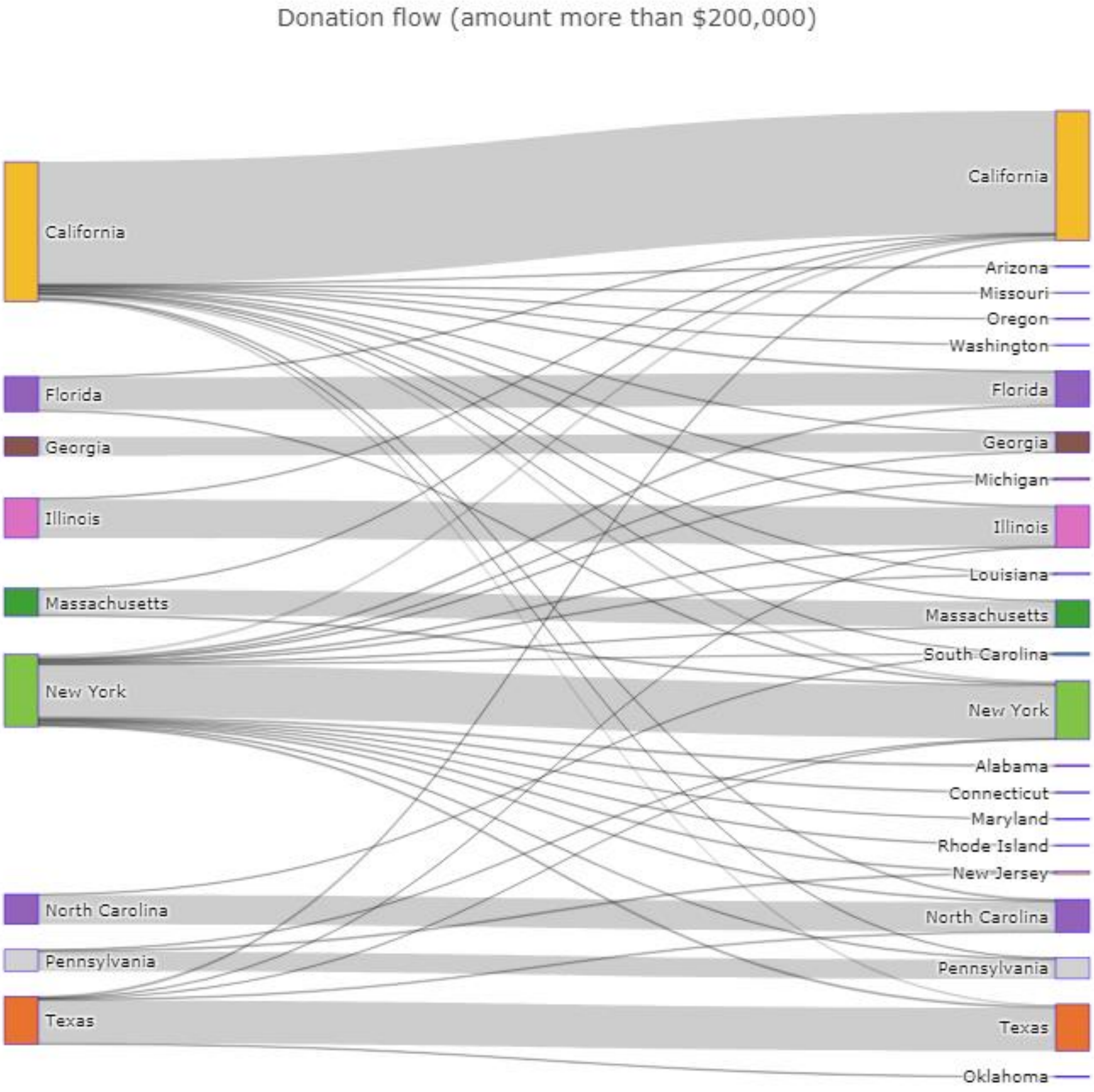

Figure 13. Donations over $200,000

Compared to donations over $100,000, Figure 13 shows an apparent trend of having the same origin state and destination state, indicating that donors tended to donate to the states they reside in.

Projects

Each project has columns of type, title, essay, short description, need statement, subject category, subject subcategory, resource category, grade level, project cost, posted date, expiration date, current status and fully funded date.

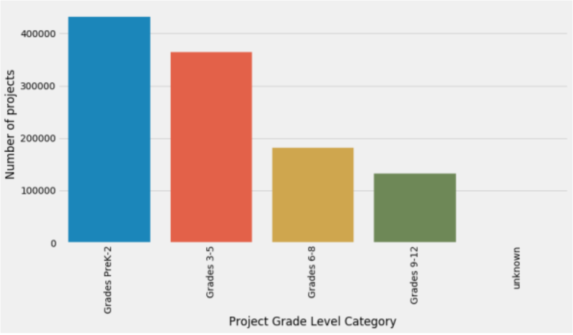

Figure 14. Projects by Grade Level

As shown in Figure 14, the number of projects decreases along with the grade level of the classrooms, which could be due to the education resources leaning towards younger population. As for the categories of the posted projects, Figure 15 presents literacy, math and science as the most focused subjects.

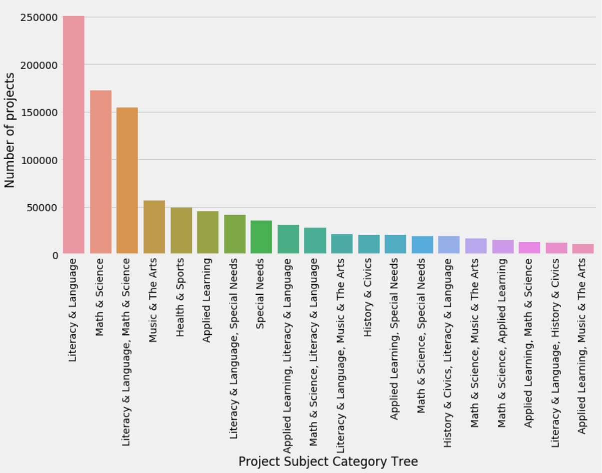

Figure 15. Top Subject Categories

The textual information of a project is critical and visual to users. For example, the title could be very important to provide the first impression to potential donors. Thus, WordClouds have been created for project titles.

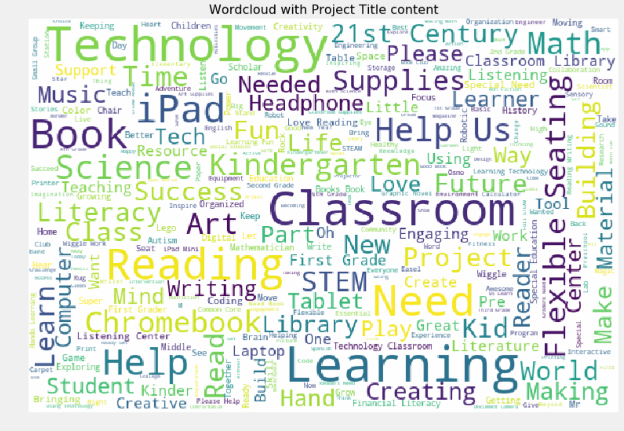

Figure 16. Project Title WordCloud

Some of the keywords represent a large need of technology items, such as iPad, Chromebook, headphones and etc. Some are related to the category of the projects, such as art, literacy, STEM and etc. There are also a few positive words, such as “love”, “future” and “fun”.

In comparison, the WordClouds of need statement and short descriptions are shown below in Figures 17&18. According to the key words shown in these figures, the need statement column is more focused on the requested resource item, while short description tends to provide a detailed context for the project, e.g. “eager”, “low income” and “high poverty”.

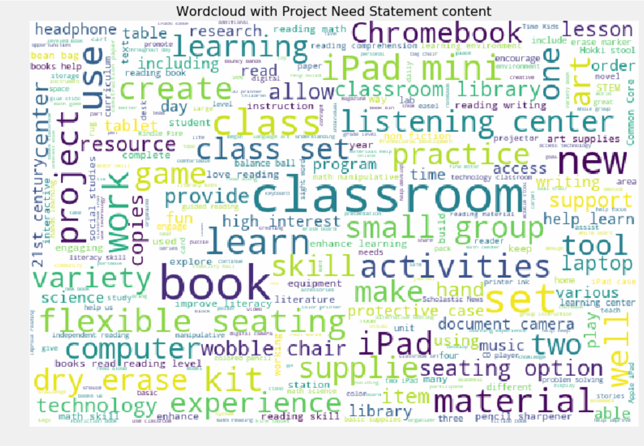

Figure 17. Need Statement WordCloud

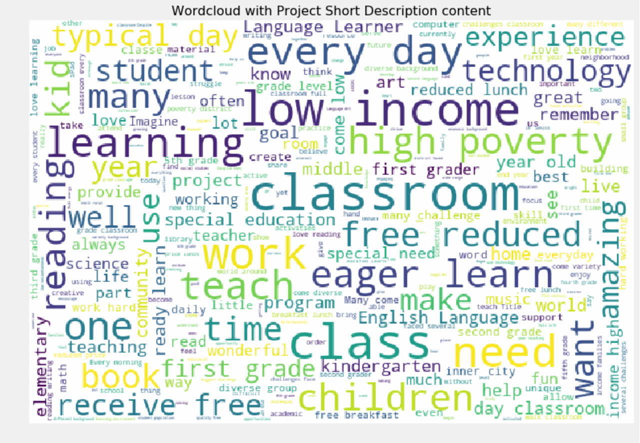

Figure 18. Short Description WordCloud

Cluster Analysis

In this section, the attempt to implement K-means cluster analysis on donation data is described, as each donation instance can be joined to its corresponding project, school and donor. The merged table eventually yielded columns including donation amount, donor state, school state, school metro type, optional donation, project category, grade level and project cost. Using K-means clustering may be helpful to find “similar” donations, and feed similar donations to donors.

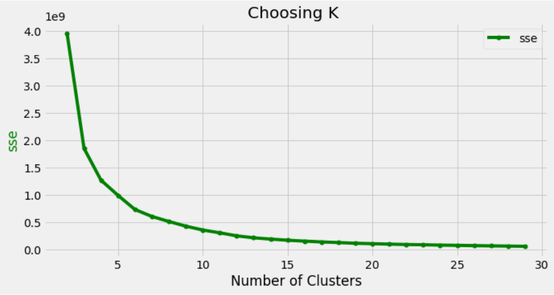

Figure 19. Choosing K

The plot above shows using elbow method to choose an appropriate number of clusters for K-means. Y-axis indicates the sum of squared error (SSE). An ideal K value would be a level-off point for SSE while allowing for good interpretability. As a result, K values from 3 to 6 were all examined. Then, T-distributed Stochastic Neighbor Embedding (TSNE) was utilized to reduce the data dimensionality to two dimensions, the K-means results can be visualized, as shown in Figure 20.

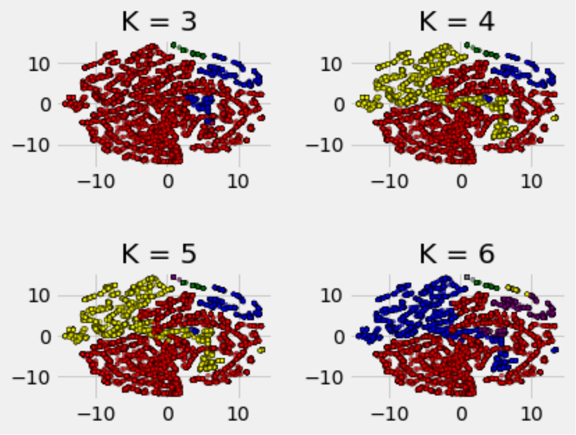

Figure 20. K-means Results Visualized

However, the results are showing a clear-cut between different clusters. Moreover, increasing the number of clusters did not seem to yield distinct results. Due to the limitation of lacking a “true label” for the donation data, it was also difficult to validate the clustering results. Thus, a content-based recommendation mechanism is proposed in the following section.

CONTENT-BASED RECOMMENDATION

Recommending similar projects to donors that have made at least one donation can improve the efficiency of feeding the right projects to potential donors. Content-based recommendation system is a widely used mechanism to help identify instances that are of particular interest to users [10].

Based on the short description of a project, representation of the projects that are amenable to comparison was created. Term frequency-inverse document frequency (TF-IDF) vector was computed for the subsequent analysis. TF-IDF evaluates the importance of words to a document in a collection of corpus, and has been applied heavily in natural language processing [11]. Then, cosine similarity was applied to measure the similarities between projects.

Due to the large amount of data, only 10000 projects were included in this section.

Recommending projects to users are proposed in two ways: 1) match projects by project title; 2) matching projects based on donor’s profile.

Matching by project title

For example, if a user has visited, created or made a donation to a single project with the title “Stand Up to Bullying: Together We Can!” The top ten similar projects listed are:

“Stand Up to Bullying: Together We Can!”	
ID	**Similar Projects**
8327	Carving Tools for Block Prints
2771	Reward Student Memoirists
45	Writing Against Bullying!
1948	Create an Anti-Bullying Culture
8867	Putting Technology in the Hands of First Graders
868	A "Bully Free" Class in Math Class!
2154	Today's Students Are Tomorrow's Leaders
1297	Button Making Machine
7155	Health, Fitness, and Wellness: Wii Are On the...
166	Just Say No To Bullies

Table 9. Finding Similar Projects I

Another example to find similar projects to "Our Old Computer Can't Keep Up With Our Learning".

“Our Old Computer Can't Keep Up With Our Learning”	
ID	**Similar Projects**
1071	We Need Programs To Make Our Smart Board Smarter!
5749	Mobile Research Technology! Help Us Get Info!
8044	Let’s Get Crabby About Science
6591	Media Center In First Grade
2516	Technology in the Classroom!
4622	Help Make my Classroom Technologically Advanced!
9072	High School Visual Arts
8140	iPad Learning Fun!
217	Communication Is Key For Our Students With Aut...
5011	Laptops for the Classroom

Table 10. Finding Similar Projects II

Matching based on donor’s profile

Another perspective is to examine a donor’s donation history. By assigning weights to the past projects based on the donation amount, the recommendations can be ranked by the weighted similarities. For instance, one of the donors’ history is as shown in Table 11.

Some Donor		
Donation ID	**Project Title**	**Donation Amount**
326	Our Learning Is On Fire!	$ 1.0
8421	Our Learning Is On Fire!	$ 1.0
8463	Fight the Summer Brain Drain With Books	$ 1.0

Table 11. One Donor’s Donation History

The donor has made three donations, two of which went to the same project. Each donation had the same amount. Hence, projects that are similar to “Our Learning Is On Fire!” will be ranked higher. This allows feeding personalized projects to each donor and can potentially improve the efficiency of the recommendation system. The recommended projects for this donor are listed in Table 12 below.

Personalized Similar Projects		
Project ID	**Title**	**Weighted Similarity**
721	Our learning is on fire!	0.85
245	Fight the summer brain drain with books	0.57
8819	Summer slump no more for children with autism	0.26
1273	Leveled books to help us read!	0.20
662	¡en sus marcas, listos... a leer!	0.19
2355	kindle fire tablets for ms. c.'s class	0.19
4573	What do they want? Books! When? Now!	0.19
38	Kindle this!	0.18
279	Reader’s workshop = instilling a love of readi..	0.17
273	Can you hear it now?	0.17

Table 12. Finding Similar Projects III

CONCLUDING REMARKS

There has been little research that focused on public education donation. This work looked at data provided by DonorsChoose and provided insights on schools, teachers, donors, projects and etc. The research uncovered several interesting findings, such as: 1) the number of donors and donation amount has been steadily growing since 2012; 2) 85.3% of the donations included extra donation to DonorsChoose, showing appreciation for the platform; 3) 26.04% of the donors were repeat donors; 4) donors tend to donate to projects from the state where they reside in; 5) technology, science and literacy are the most popular subjects for projects. According to these analyses, DonorsChoose as a platform has been performing well, and still has great potential to make significant impacts.

By evaluating TF-IDF vectors and computing the similarities between projects, content-based recommendation was implemented to discover projects based on a single project or donor’s donation history. This can improve the recommendation system by feeding projects to users based on their preferences.

The current work on DonorsChoose is a starting point for studying public education donation data analysis. There are still a few limitations of the research. For instance, the available projects did not include projects that did not get fully funded. Inclusion of these projects can allow identifying key factors that contribute to attract donations. Another limitation is the performance of machine learning techniques. Clustering analysis did not yield an interpretable result. For future studies, more feasible approaches like spatial clustering can be explored. In terms of alternative recommendation systems, collaborative filtering and hybrid recommender systems could also be considered.

REFERENCES

1. DonorsChoose. (2018, June). Retrieved from https://www.donorschoose.org/
2. CBS. (2016, Mar 10). *Stephen Colbert on $14M DonorsChoose.org funding by public figures.* Retrieved from https://www.youtube.com/watch?v=I4GMX0MJNYw
3. Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., ... & Sampath, D. (2010, September). The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems* (pp. 293-296). ACM.
4. McDonald, D. W., & Ackerman, M. S. (2000, December). Expertise recommender: a flexible recommendation system and architecture. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 231-240). ACM.
5. Liang, Y. (2017). *A Machine Learning Approach to Fundraising Success in Higher Education* (Doctoral dissertation, Department of Computer Science, University of Victoria).
6. Wastyn, M. L. (2009). Why alumni don’t give: A qualitative study of what motivates non-donors to higher education. *International Journal of Educational Advancement*, 9(2), 96-108.
7. Lindahl, W. E., & Winship, C. (1992). Predictive models for annual fundraising and major gift fundraising. *Nonprofit Management and Leadership*, 3(1), 43-64.
8. U.S. Department of Education. (2018). *Digest of Education Statistics.* (NCES 2017-094).
9. U.S. Census Bureau. (2018). *State Population Totals and Components of Change: 2010-2017*. Retrieved from https://www.census.gov/data/datasets/2017/demo/popest/state-total.html
10. Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web* (pp. 325-341). Springer, Berlin, Heidelberg.
11. Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, pp. 133-142).