FINAL PROJECT: MATCHING DONORS TO PROJECTS

May 18, 2018

Chengyu Chen (A14051607), 2nd Year Applied Mathematics; Data Science, MATH 189
Chenyue Fang (A13686794), 2nd Year Probability and Statistics, MATH 189
Daniel Lee (A13726312), 2nd Year Probability and Statistics; Data Science, MATH 189
Xinran Wang (A13564644), 2nd Year Probability and Statistics, MATH 189
Yuqi Wang (A13532155), 2nd Year Applied Mathematics, MATH 189
Ning Xu (A92061610), 3rd Year Probability and Statistics; Economics, MATH 189

INTRODUCTION

DonorsChoose.org, which is founded by a high school teacher in 2000, is a nonprofit organization that helps raise money for American public school classroom projects. Teachers from public school request for what their students need on DonorsChoose.org, and individuals can choose projects that they are interested in to donate. Today, about 1.1 million projects on DonorsChoose.org have gotten donation from 3 million individuals, but many teachers still spend more than one billion dollars of their own money on classroom supplies. Thus, if DonorsChoose.org is able to inspire a fraction of donors who just made their first-time donation on the website to make another one, that will dramatically influence the number of fulfilled requests. For example, a 10% improvement in donor retention could yield up to a 200% increase in obtained donations ("The Ultimate Guide To Donor Retention").

In this project, in order to help DonorsChoose.org better connect donors with projects that most motivate them, we are going to build a recommender system that sends email campaigns recommending specific projects to prior donors. The recommender system will be the final step of our analysis; the preceding analysis will consist of exploratory data analysis relating to the provided datasets and will build up to the construction of the recommender system.

THE DATA

The six raw datasets used in this study come from DonorsChoose.org, which is a platform helping teachers raising money for their projects. The six datasets consist of information of donors, projects, schools and teachers who posted the project. A short description about what variables each datasets specifically present is attached in the chart at the end of the data section of this report. Along with that, one additional dataset listing the state population information from 2010-2017 from United States Census Bureau is listed to better assist us understanding the ratio of donor in that state to the state's population.

One significant feature about the datasets we used in our study is that these datasets are mainly constructed by categorical data. Among the 6 datasets, the only numerical data is the project cost, the donation amount made by one donor in a donation transaction, important date information for projects and donations, and the free lunch percentage per school. Since the ultimate goal of this study is to build a recommender system which analyzes the donations' pattern and connects donors to the the projects they most likely to donate to, we investigated on how each categorical data of one donation transaction influence donor's donation decision in this study and constructed a classification model to predict the match between donors and donations.

Specifically, we used variable "Donor Is Teacher" in donors.csv to investigate how the donation pattern of teachers and non-teachers differ; we used variable "Donor State" in donors.csv to discover how the total donation amount differs by states; we use variable "Teacher Prefix" in teachers.csv to infer how gender impacts donation amount. By combining donation.csv with donors.csv and projects.csv by Project.ID, we used variables "Donation Amount" and other categorical variables to investigate the donor interests and the change of donation pattern over time.

By merging almost all of the information from the six available datasets by the unique ID given to school, donor, donation, and project, we generated a combined dataset listing all transactions of donation with their related information. Those related information includes the topic of the project, the cost of the project that transaction went to, information about the teacher who initiated the project, school where the project benefited, donor who made the donation, etc.

By using the merged dataset and linking the background information of donors to the background of the school where the funds go to, we aim to use our data to get a better understanding on the differences among the US states, discuss about donors' psychological preferences on what to donate, and make a difference by connecting available resources to the needed.

Dataset	Description	Columns
Donations.csv	This dataset contains each donation from a citizen donor and is joined with the Projects.csv using the "Project ID" column.	Project ID Donation ID Donor ID Donation Included Optional Donation Amount Donor Cart Sequence Donation Received Date
Donors.csv	This dataset contains the information of each donor.	Donor ID Donor City Donor State

		Donor Is Teacher Donor Zip
Projects.csv	This dataset contains all the information about a specific project.	Project ID School ID Teacher ID Teacher Project Posted Sequence Project Short Description Project Need Statement Project Subject Category Tree Project Grade Level Category Project Resource Category Project Cost Project Posted.Date Project Expiration.Date Project Current Status Project Fully Funded Date
Schools.csv	More information at a school level. Each row represents a single school. This dataset is joined with the project data using the "School ID" column.	School ID School Name School Metro Type School Percentage Free Lunch School State School Zip School City School County School District
Teachers.csv	This dataset contains the information of each teacher that posted the project and is joined with the project data using the "Teacher ID" column.	Teacher ID Teacher Prefix Teacher First Project Posted Date
Resources.csv	Detailed description about the resources that is been requested	Project ID Resource Item Name Resource Quantity Resource Unit Price Resource Vendor Name
State_Population_Ratio.csv	Extra dataset which contains the state population totals	Title ID Notes Sources Release state

BACKGROUND

In the late 19th century, donations were mostly accomplished by direct interactions with

donors and people who were in charge of donations, which was very inconvenient for money transfer and was also very hard for donors to know specific donations. However, nowadays with the rise of social media, the donating process can be easily implemented online by using online crowdfunding platforms.

Online crowdfunding platforms, like DonorsChoose.org, allow projects and organizations to collect funding by millions of donors. On those platforms, anyone can become a donor and make contributions to specific projects, so the "crowd", which means a lot of donors, collectively contributes to the funding of the project. In *An Introduction To Crowdfunding*, it is shown that the arise of online crowdfunding platforms facilitates the transfer of money from people seeking to donate to people in need of the capital for donations. The number of crowdfunding platforms has been growing rapidly these years, so does the annual growth rate, which is shown by the fact that the worldwide annual growth rate of crowdfunding platforms has risen from 38% in 2008 to 60% in 2012. Therefore, crowdfunding platforms are very common at present, which provide a revolutionary avenue for funding transfer between donors and people in need of the capital for donations with ease.

One of the success of crowdfunding communities is continued engagement of donors, so a significant challenge for online crowdfunding platforms is the problem of donor retention, which is the topic our research paper intended to investigate by using statistical methods. According to Jepson, donor retention refers to "the number (or percentage) of donors that return to give another gift in a specific time period". The goal is to get connected with donors and never lose connection.

According to Jepson, a high retention rate indicates that a nonprofit has a healthy support system, which is important for nonprofits. Foremost, donor acquisition costs are high since the initial cost of finding first-time donors is very high, which could take 18-24 months for nonprofits to recruit a first-time donor. Moreover, it provides access to bigger networks with people fundraising, which is a commitment to build up relationships with different parties and increase promotion opportunities. Apart from those reasons, the online article "The Ultimate Guide To Donor Retention" also illustrates that higher retention rate contributes to bigger donations. Since majority of new donors don't make their first donations very large, larger amounts of donations will be resulted from builded relationships, which are guaranteed by donor retention. Furthermore, it enables useful feedback from loyal donors. Once fundraisers retain donors and the trust is built, donors grow more comfortable with fundraisers' organizations and thus will be more likely to provide feedback, which helps fundraisers to improve their fundraising strategies. Therefore, as those two articles indicated, a high retention rate is paramount and conducive for nonprofits, which saves cost and energy to find new donors as well as provides additional advantages for nonprofits.

In order to help DonorsChoose.org better connect with donors and obtain higher donor retention rate, in this research paper, we are going to investigate factors influencing donors to choose specific projects and how much to donate, which affects donor retention. The case study of DonorsChoose.org below serves as a good literary review and provides fundations and insights for our investigation.

In *Donor Retention in Online Crowdfunding Communities: A Case Study of Donors Choose.org*, researchers explore various factors influencing donor retention such like project success, project cost and response from teacher so that they can identify different groups of donors and quantify their inclination to make another donation. By plotting distribution of fraction of donors by the number of their donations, researchers find out that only 26% of first-time donors return for another donation and 76% of these donors return afterwards to show the overall state of donor return rates. Then, they investigate factors impacting donor retention from three aspects: projects, donors and teachers.

From project perspective, the researchers analyze how the first project positively and negatively influences the return rate, and the results show that donors whose first project is successful is 5% more likely to make another donation. So it can be concluded that donors who donate to a successful project are more likely to return. Furthermore, the researchers investigate the relationship between project costs and donor returns, which shows a negative relationship on the scatter-plot graph. They thus conclude that donors to small successful projects with less costs are much more likely to return (32%) than donors to large projects with higher costs (23%). This result provides hint to our research and we would further investigate how donation dollar amounts would affect the donor amounts in this paper.

From donor perspective, the researchers partition donors into groups by the distance between the school they funded in their first donation and their locations. By plotting the return rates for each group, researchers find that a U-shape pattern is formed on the plot, which reflects that donors who funded school far away or very close to their location are more likely to return for subsequent donation. This can be possibly explained by the fact that donors care more about the projects which they are familiar with, while donors that are very passionate about the community will even fund projects across the country. Moreover, the researchers consider the roles of donors within the project that they partition donors into three categories: *starters* that like to start off new projects with an initial donation, *closers* that like to finish off projects that are close to completion, and a third group that does not particularly follow any of the previous two behaviors. The U-shape plot shows that donors in the middle of the project's lifetime are less likely to return than starters, and closers, however, display the highest propensity to return for another donation. However, their research does not consider the gender of donors as well as

whether the donors are teachers or not, which will be further investigated in this paper.

From teacher perspective, the researchers investigate the relation between teacher response time, such as giving thanks to donors, and donor return rate to the same teacher. The result shows that response times in the first 24 hours after donations are correlated with significantly higher return rate. Since our dataset does not include the information about behaviors of projects' teachers in charge, we would not investigate in terms of this perspective.

Overall, their research investigates how three factors, including project, donor and teacher, would affect donor's return rates, which gives us good insights about how to improve donor retention based on those factors as well as build our own recommendation system. Apart from those factors, we will further investigate whether the distribution of states and time would also influence donor's amounts, and finally we will build a model to predict how likely for donors to donate for a project based on the features of projects and donors.

INVESTIGATIONS

Data Preprocessing

Before any analysis could be done on the data, the large datasets provided to us required much preprocessing and cleaning.

We first grabbed the unique column names across all six datasets, to help us identify which features are available to us for analysis, regression, and classification, and to facilitate feature section:

```
Donation Amount, Donation ID, Donation Included Optional Donation,
Donation Received Date, Donor Cart Sequence, Donor City, Donor ID, Donor
Is Teacher, Donor State, Donor Zip, Project Cost, Project Current Status,
Project Essay, Project Expiration Date, Project Fully Funded Date,
Project Grade Level Category, Project ID, Project Need Statement, Project
Posted Date, Project Resource Category, Project Short Description,
Project Subject Category Tree, Project Subject Subcategory Tree, Project
Title, Project Type, Resource Item Name, Resource Quantity, Resource Unit
Price, Resource Vendor Name, School City, School County, School District,
School ID, School Metro Type, School Name, School Percentage Free Lunch,
School State, School Zip, Teacher First Project Posted Date, Teacher ID,
Teacher Prefix, Teacher Project Posted Sequence
```

The remainder of our analysis will focus on the data contained in these columns.

Projects.csv contains a “Project Subject Category Tree” column which lists the subject categories that each project belongs to.

Project ID <chr>	Project Subject Category Tree <chr>
7685f0265a19d7b52a470ee4bac883ba	Applied Learning
f9f4af7099061fb4bf44642a03e5c331	Applied Learning, Literacy & Language
afd99a01739ad5557b51b1ba0174e832	Literacy & Language
c614a38bb1a5e68e2ae6ad9d94bb2492	Literacy & Language
ec82a697fab916c0db0cdad746338df9	Special Needs
563958074d7b12b48b939279eb59e6ca	Literacy & Language, History & Civics
717c7a01215d532d68f6fe9e666c88c3	Applied Learning
4202c4e251fe483dfd93520da022f987	Literacy & Language
49825532f85d0cdb569797df3ab8ec46	Literacy & Language
60dddb9495e5ed60c1f6c1b86fe9a7e4	Literacy & Language

1-10 of 1,110,017 rows

However, for accessing a project’s categories, for each project, we prefer for there to be a separate entry for each of its categories.

Project ID <chr>	Project Subject Category Tree <chr>
ce4e937227f53df90716c29404174040	Literacy & Language
7abd1ee32d657b764a19f5505d35e2ba	Music & The Arts
000832bc0aafcaf274db0db25e4f2bb0	Special Needs
5c40cf6b74dd35445b5d4d925cfe418d	Health & Sports
79ec4c73ac0f89eab1b6d9e42acf20f9	Music & The Arts
ec7e8033f28673316eb6c4fdbb5da8b9	Literacy & Language
dac56772afd1fc07ca8bae1f8cf71743	Literacy & Language
44b8b852492a0d4315d0d87c145e7718	Math & Science
12a9f616af0ff12f6f4b473352655027	Math & Science
60e47e4f23a9552381e82d52dfabe0bf	Music & The Arts

1-10 of 71,391 rows

Using this updated dataset, we can see that there are eight unique project subject categories, with Literacy & Language being the most popular.

Project Subject Category Tree <chr>	N <int>
Literacy & Language	25667
Math & Science	19682
Applied Learning	7160
Special Needs	6027
Music & The Arts	5294
Health & Sports	3634
History & Civics	3465
Warmth, Care & Hunger	462

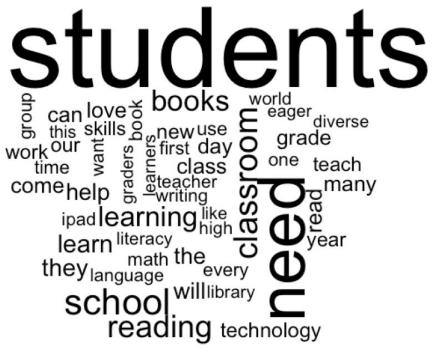
8 rows

Scenario 1: Text Mining

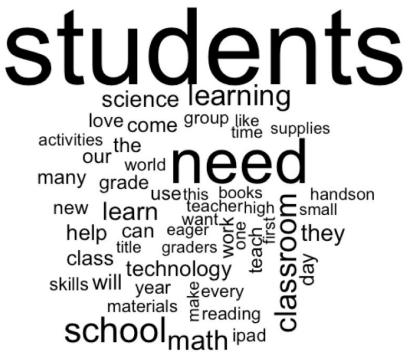
Before conducting any statistical tests or machine learning models, we wish to perform simple text mining, in particular sentiment analysis, on the projects, grouped by subject category, in order to get a good "sense" of the different subject categories.

Using Projects.csv, we take only the Project Subject Category Tree, Project Short Description, and Project Need Statement columns. We group this dataset by category, and perform text mining on Project Short Description (synopses of the projects) and Project Need Statement (what donations will be used for). The sentiment analysis results (word clouds) are shown below, grouped by the eight subject categories.

Literacy & Language

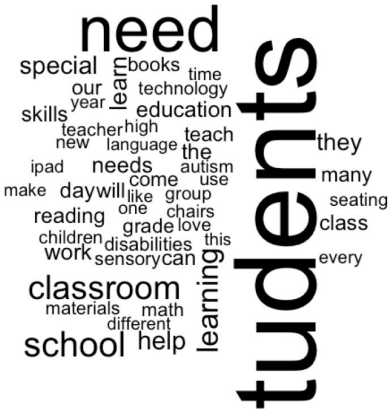


Math & Science



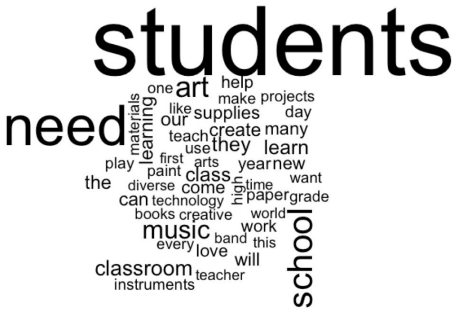
Applied Learning

Special Needs



Music & The Arts

Health & Sports



History & Civics

Warmth, Care & Hunger

From these results, we can observe that “students” and “need” are among the most popular words across all eight categories. This is as expected, since a purpose of a classroom donation request is to ask for funding to get what students need. We can see distinguish categories by observing differences in the remainder of the words in each word cloud. The popularity of the words for each category are as expected in terms of sentiment and vocabulary. For example, some popular words for each of the categories include:

- Literacy & Language: books, library, literacy, reading, writing
- Math & Science: math, science
- Applied Learning: learning
- Special Needs: autism, disabilities, help, sensory
- Music & The Arts: art, instruments, music, paint, play, projects
- Health & Sports: balls, fitness, equipment, healthy, physical, play, soccer
- History & Civics: history, social, world
- Warmth, Care & Hunger: backgrounds, basic, breakfast, children, families, food, free, healthy, help, hygiene, low income, lunch, poverty, snacks

Therefore, the sentiment analysis was able to produce concrete results and was able to match relevant words to project subject categories. We will keep this sentiment analysis in mind in future analyses.

Scenario 2: Exploratory Data Analysis

Prior to analyzing the data, we first explored the data as a whole to get some basic information of the data from numerical and graphical aspects.

Q1: What is the total amount that DonorsChoose have raised?

By adding up the total amount of donations made to donorschoose.org in donations.csv, we found out that DonorsChoose.org has raised 284,408,243 dollars in record.

Q2: What are the most expensive requested items among the projects?

From data in resources.csv, we learn that the most expensive requested item is handicapped accessible playground which worth 97085.5 dollars.

Q3: What is the highest amount of donation per transaction so far?

Looking at each donation transaction donors had made, the table is the numerical summary of how the those donation transactions are distributed. From the table below we can see that the the highest donation a donor has made per transaction by the time the dataset was created is 60,000 dollars.

Numerical Statistics for distribution of donation amounts

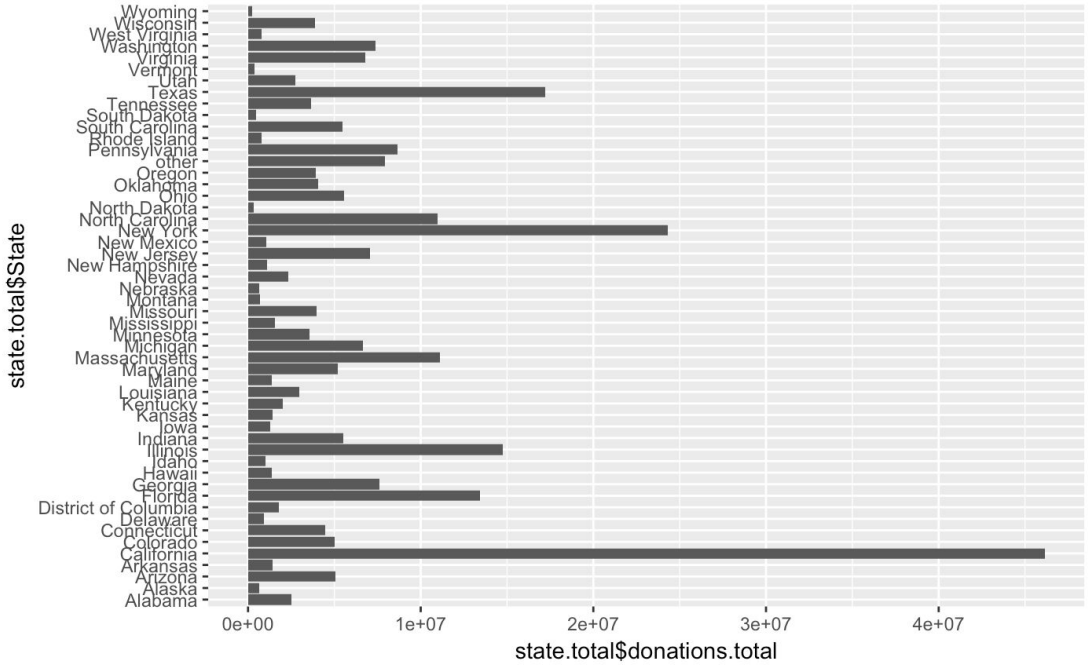
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0.01	14.82	25.00	60.67	50.00	60000.00

By looking up the donor who made this huge donation of 60,000 dollars, we gathered the information about the donors and the projects they donated to. The donor is from Anahola, Hawaii, who is not a teacher, donating to a teacher-led project under Health & Sports category. This project is designed for children that are from preK-2 to a unknown metro-type school where free lunch percentage is at 51%.

Q4: Which state has highest donated amount?

From donors.csv, we found that California has the highest total amount which is 46,140,356.5 dollars. The amount of donations made by each states is presented in the following bar graph.

Bar graph for the distribution of total donation amounts per state



Q5:Which project received the most amount of donations?

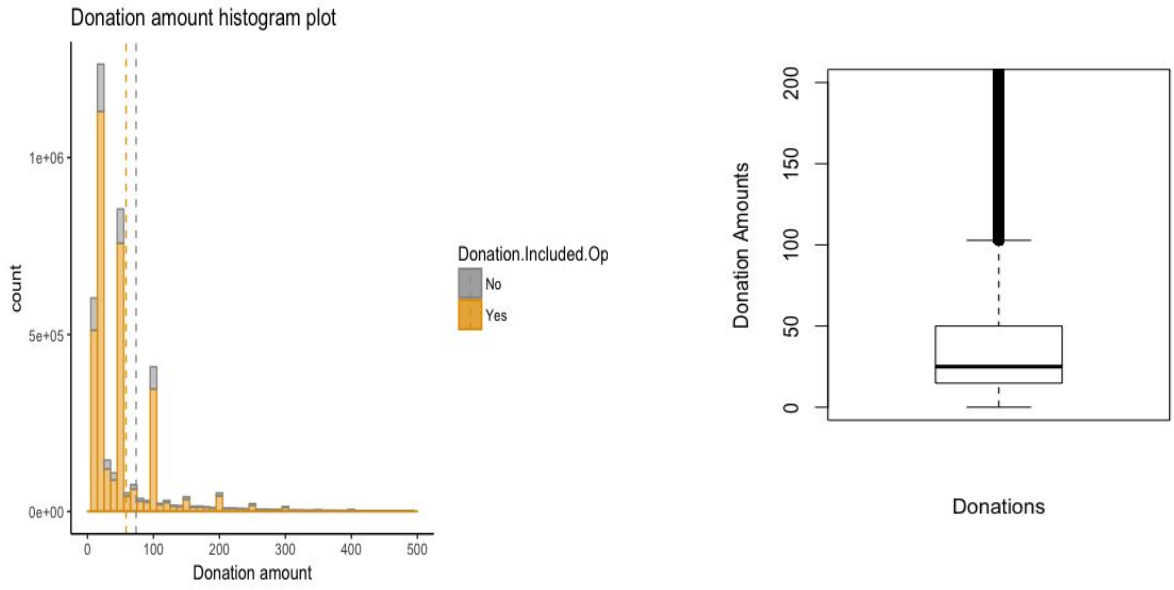
By looking up the project that received the most amount of donation which is 108,248.30 dollars, we get the information about this project and the donor who made this donation. It is a teacher-led project under Health & Sports category, which is designed for students from Grade 3 to 5 in a suburban school where the free lunch percentage is at 14%. The project resource is classified as supplies, and the current state shows that it is fully funded now.

Q6: Is there any difference between average received donations with 15% donation to the website included and non-included?

In order to further explore details of the distribution of donation amounts, several graphical methods are used here. In the figure below, the histogram which represents distribution of donation amounts, is asymmetric and heavily right-skewed. It shows that the majority of donations clusters is below 50 dollars, and donations that are higher than 200 dollars are too little to be shown on the graph. The total amount of these donations that are included in the graph is 284,189,102 dollars, which made up 99.93% of the donations donorschoose has collected.

The figure below displays the distribution of donations by their amounts as a boxplot. Since black lines here represent outliers, we can see that there are huge amounts of outliers in this distribution. Also, the majority of donations are lower than 100 dollars and the median value is clearly lower than 50 dollars.

Histogram and boxplot for the distribution of donation amounts per transaction



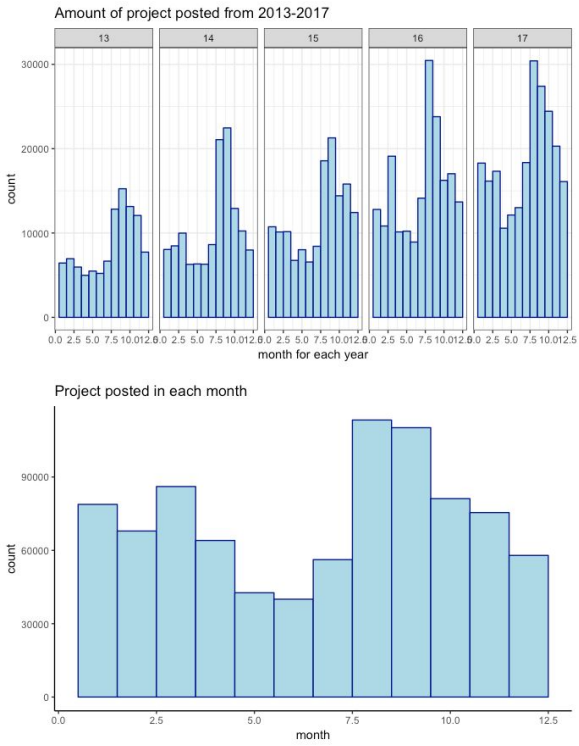
The average amounts of donations received when the 15% donation to the website is included and non-included are 58.455 and 73.608 dollars respectively. Here, the figure above displays the histograms for distributions of donations with 15% donation to the website (yellow) and those without (grey). By comparing two distributions, we find that both of these two distributions are asymmetric and skewed to right. The two distributions are nearly matched, but the number of donations which do not include 15% donation to the website is higher than the number of donations which include 15% donation to the website in every interval of donation amount. The two dash lines in yellow and grey represent means of two distributions, and we can see that the mean of the distribution of donations when 15% donation is included is lower than when 15% donation is not included.

Scenario 3: Seasonal Patterns of Projects

Question: Do the amount of fully funded projects grow over time? Is there a seasonal pattern of donations? For example, is there a peak month in the year that have a lot of projects posted annually? If yes, what do you think the reason would be? Are the dollar amount of received donations related to the date of those donations being made? Is there a peak month in the year that most of the teachers began post their first project? In other words, how does registered teachers grow throughout time?

1. Graphical display of the distribution of amount of project posted in each month.

Histograms of the amount of project posted in each month



We are motivated to see how the posted projects are distributed throughout months in the past five years. In order to investigate the seasonal pattern of donations, we make histograms with months being the horizontal axis and count being the vertical axis. From the histograms above, we discovered an increasing trend of projects being posted by teachers throughout the years. It also seems to be true that the peak time for teachers posting projects is between July and October.

2. Use unpaired Wilcoxon rank sum test to test whether the donation amount of two years is same.

```
Wilcoxon rank sum test

data:  merged_2014_sum_month$B and merged_2015_sum_month$B
W = 51, p-value = 0.2415
alternative hypothesis: true location shift is not equal to 0
```

In order to investigate the donation pattern of different years, we are going to compare the mean donation amount of years. Before applying any hypothesis testing, we divide each year into 12 months and get the mean donation amount of each month to be our input data. Then we choose to use unpaired Wilcoxon rank sum test because being as a kind of nonparametric test, Wilcoxon test does not require our data to be normally distributed. Our null hypothesis is the true

location shift of 2014 and 2015's mean donation amount is equal to 0, and the alternative is the true location shift is not equal to 0. The calculated p-value is 0.2415, which is greater than the significance level 0.05. Therefore, there is insufficient evidence at the alpha level of significance to reject the claim that the true location shift of 2014 and 2015's mean donation amount is equal to 0. We can infer that the donation pattern over the years does not change significantly.

3. Chi-squared test check if the distributions of two year-data are the same/ test if uniform

sub-interval = 3 Chi-squared test for given probabilities data: observed.counts X-squared = 4960, df = 2, p-value < 2.2e-16	sub-interval = 4 Chi-squared test for given probabilities data: observed.counts X-squared = 18772, df = 3, p-value < 2.2e-16
sub-interval = 6 Chi-squared test for given probabilities data: observed.counts X-squared = 18610, df = 5, p-value < 2.2e-16	sub-interval = 12 Chi-squared test for given probabilities data: observed.counts X-squared = 31102, df = 11, p-value < 2.2e-16
Sub-interval = 24 Chi-squared test for given probabilities data: observed.counts X-squared = 36475, df = 23, p-value < 2.2e-16	

It is stimulating for us to investigate the pattern of donation amount because we want to find the time period in the year when the donation amount is the highest. We divide the data into intervals with different number of days, and specifically we divide 365 days into 3, 4, 6, 12, 24 intervals. We want to check whether the donation amount for each time period is the same or not. Then, we use Goodness of Fit test to test whether the distribution of donation amount of each month is uniformly distributed. We use Goodness of Fit test because it is a kind of statistical model that tells how well a set of observation data fits the expected distribution. The null hypothesis is that the distribution of donation amount of each day throughout the interval is uniformly distributed, and the alternative hypothesis is the distribution of donation amount of each day throughout each interval is not uniformly distributed. As we see, the p-values are all less than 2.2e-16, which means the probability of being uniform is extremely small. Thus, we reject the null hypothesis, and conclude that the donation amount of each day throughout the interval is not uniformly distributed. Since the graph is not uniformly distributed, we can conclude that more people may post their projects in certain months. Furthermore, it can be

inferred from the result that in order to gain more donation amount, donation seekers can focus on posting donation projects on certain months when people tend to donate more.

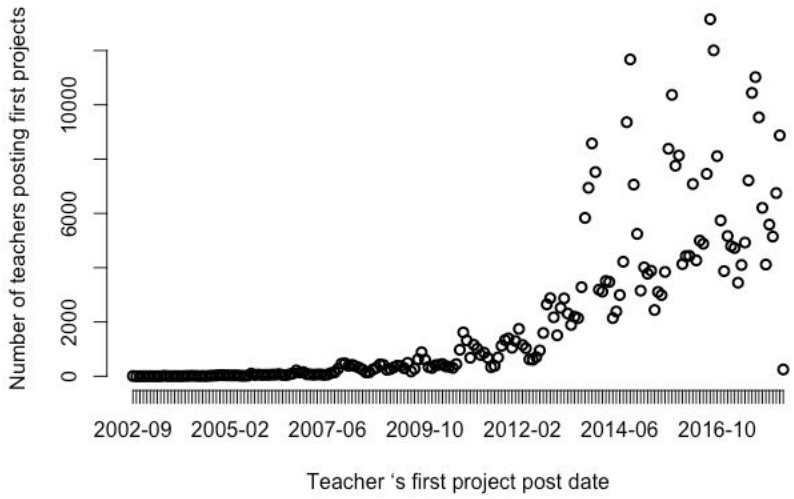
4. Use non-parametric Kolmogorov-Smirnov test to test if follows same distribution.

```
Two-sample Kolmogorov-Smirnov test

data:  difference_2014 and difference_2015
D = 0.039835, p-value = 0.6103
alternative hypothesis: two-sided
```

We are interested in researching how the distribution of donation amounts between each pair of years because we want to track the growth of donation pattern. We use Kolmogorov-Smirnov test, which is also called nonparametric Goodness of Fit test, to test the matchness of each pair of two distributions. KS test is appropriate to be used here since first of all the distribution of data is continuous, and second we do not need to assume the normality of data distribution. The null hypothesis is the distributions of each pair of two year's donation amount matches, and the alternative hypothesis is the distributions of each pair of two year's donation amount does not match. To get the test statistic, we first rank the date of year from small to large, and then we get the donation amount of each ranked date of year, and finally we use $\Delta amount/\Delta date\ of\ year$ ratio to be the input data for our KS test. The p-value calculated is 0.6103, which is greater than the significance level 0.05. Thus, there is insufficient evidence at the alpha level of significance to reject the claim that the donation amount of 2014 and 2015 follows the same distribution, which is consistent with our direct observation of the histograms above. We infer that people's donation pattern over the years does not alter significantly.

The figure below displays the distribution of number of teachers throughout time as a scatterplot. The pattern shows that the number of teachers posting projects on Donors Choose is increasing throughout time, from which we can interpret that more and more teachers attend Donors Choose to post their projects.

Scatter plot for growth of teachers throughout months



Scenario 4: Donor/ Teachers identity

Question: What is the difference in donation amount between donors who are teachers and those who are not? What is the difference in total donation amount each project received between projects led by female teachers and male teachers?

```
        Wilcoxon rank sum test

data:  teacher_money_amount and nonteacher_money_amount
W = 1.4444e+12, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 0
```

In this scenario, we plan to investigate which group donates more, teachers or non-teachers. We use the data from Donations.csv, which contains each individual teacher and non-teacher's donation amount. We plan to use the unpaired Wilcoxon signed rank test to test whether the mean donation amount of teachers equals the mean donation amount of non-teachers. Wilcoxon signed rank test is a kind of non-parametric test which does not require the basic assumption of normality, and the distribution of the population is unknown.

The null hypothesis is that the distributions of x and y differ by a location shift of $\mu = 0$, and the alternative is that true location shift is less than 0. In our model, P-value is a measure of the probability of having a location shift of $\mu < 0$ by chance. The decision rule is to reject the null hypothesis when P-value is smaller than the significance level $\alpha = 0.05$.

The P-value < 2.2e-16, which is significantly less than 0.05. Therefore, we reject the null hypothesis and conclude that Xi-Yi<0, which means the mean donation amount of teachers is less than the mean donation amount of non-teachers.

From this result, we further infer that in an attempt to achieve higher level of donation amount, donation seekers should promote donation projects to non-teachers, who have a higher mean of donation amount.

Numerical Statistics for distribution of donation amounts for non-teachers

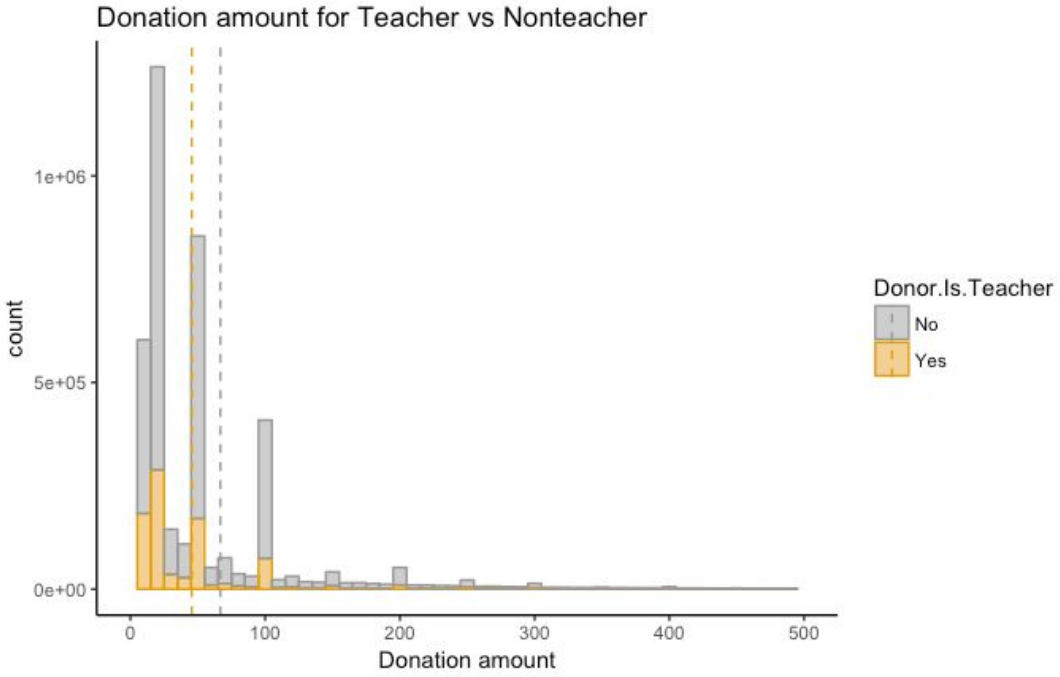
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0.01	20	30	66.7	62.1	60000.00

Numerical Statistics for distribution of donation amounts for teachers

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0.01	3	20	45.6	50.00	21299.95

In order to investigate the difference in donation amount for teachers and non teachers, we use graphical methods to compare the two distributions of donation amount. The figure below represents the two distributions as histograms. The yellow histogram and grey histogram, which display distributions of donation amount for teachers and non teachers respectively, are both asymmetric and skewed to right. Since the yellow dash line is clearly at the left side of the grey dash line, this implies that the mean of distribution of teachers is lower than mean of distribution of teachers. Also, the number of non teachers is higher than the number of teachers in every interval on the graph. The compared histograms show that there exists difference between distributions of donation amount for teachers and non teachers.

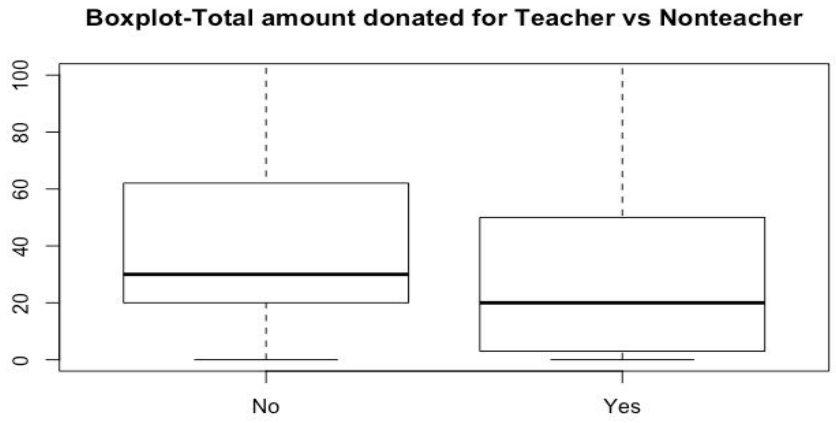
Histograms for distributions of donation amount for teachers and non teachers



The two boxplots in the figure below displays the distribution of donation amount among teachers and non teachers. From the boxplot, the 1st quartile, median and the 3rd quartile for the

donation amount for teachers are all lower than the non teachers. From figure 3 which also shows the distribution of donation amount among teachers and non teachers but in larger range, we can see that there exist many outliers in both distributions, and outliers in distribution for non teachers are more than those in the distribution for teachers.

Boxplots for distributions of total amount donated for teacher and non teacher



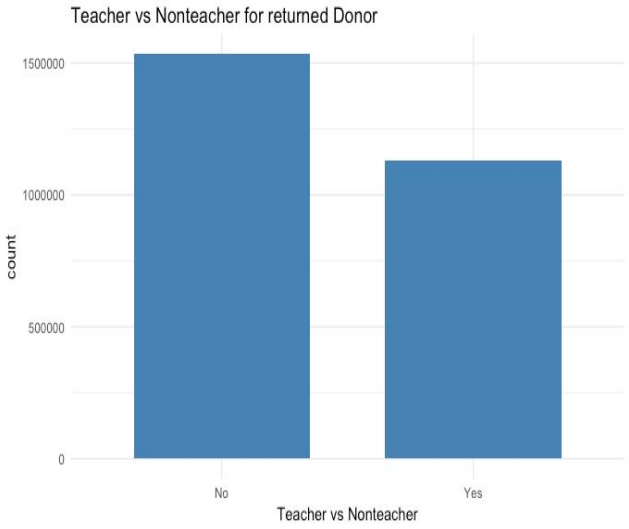
Two-sample Kolmogorov-Smirnov test

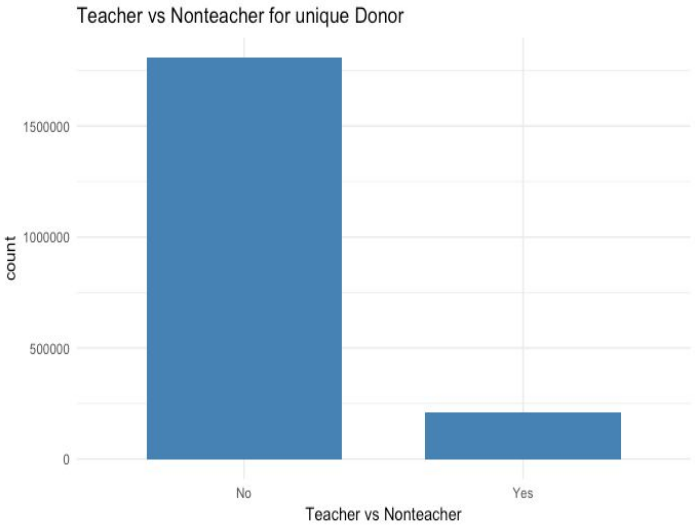
data: teacher_final$Donation.Amount and nonteacher_final$Donation.Amount
D = 0.28142, p-value < 2.2e-16
alternative hypothesis: two-sided

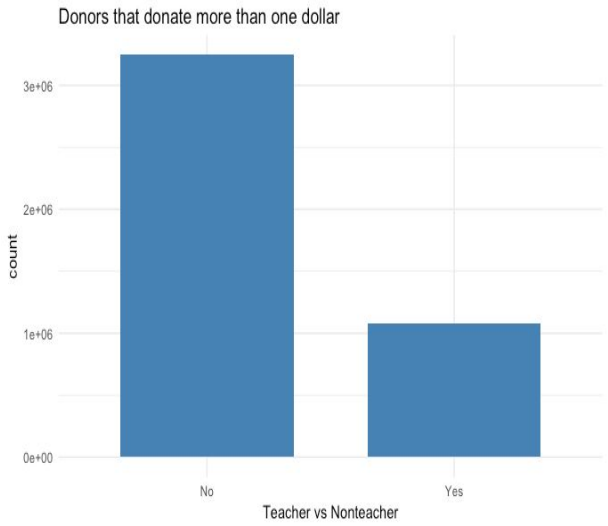
Building a recommendation system requires us to study the donation pattern of different groups of people. In order to do so, we choose to use Kolmogorov-Smirnov test to test whether the distribution of teachers' and non-teachers' donation amount are the same or not. Our input data is each teacher and non-teacher's donation amount, which is continuously distributed and satisfies the basic assumption for KS test. Our null hypothesis is the exact distribution of teachers' donation amount matches the distribution of non-teachers' donation amount. The calculated p-value is 2.2e-16, which is significantly smaller than the significance level 0.05. Therefore, there is enough evidence at significance level 0.05 for us to reject the null hypothesis and conclude that the exact distribution of teachers' donation amount does not match the distribution of non-teachers' donation amount.

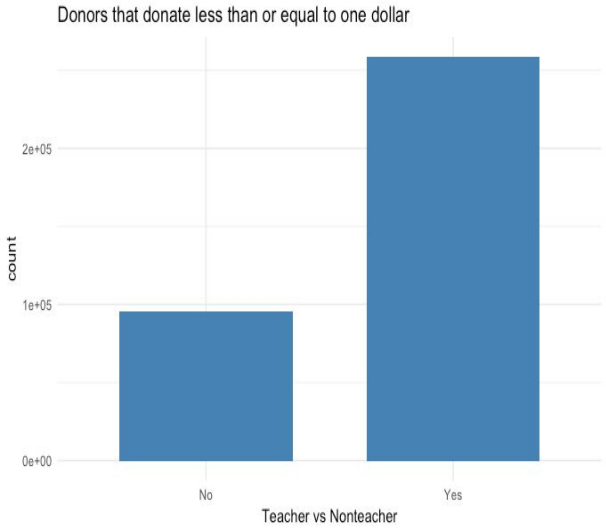
In addition, combined with our direct observation from the histogram and boxplot above, we infer that donation seekers should promote more donation projects to non-teachers who tend to donate statistically significantly more than teachers.

In order to see the difference between teacher and non-teacher more deeply, we split the data into two group, which are donors that donate only one time and donors that donate more than one times. After that we plot the bar graph for each of the two dataset, as you can see from the bar plot below, the 'Yes' means that the donors are teacher and 'No' means that donors are not teacher. I also split the data by donors that donate less than one dollar and by donors that donate more than a dollar. I again plot the barchart with respect to teacher and non-teacher.









Both barplots of "Teach vs Non-teacher for returned Donor" and "Teacher vs Non-teacher for unique Donor" shows the donation counts for non-teacher are greater than donation counts for teacher. However, those two plots differ by the number of donation counts for teacher. The donation counts for returned teachers are greater than donation counts for unique teacher donor.

From this observation, we may infer that teachers tend to differentiate with each other in terms of donation behavior.

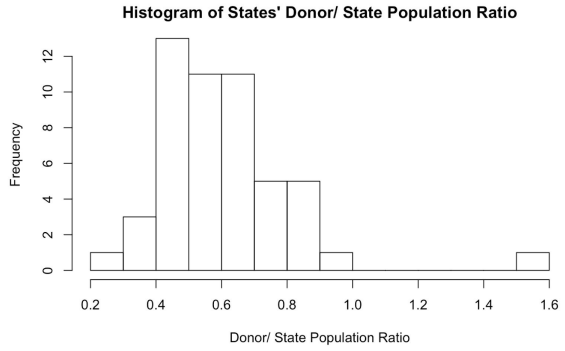
From the barplot above we notice that the number of teacher that donate less than or equal to one dollar is much larger than non-teacher that donate less than or equal to one dollar. Moreover, there are lots more non-teacher that donate more than one dollar and less teacher that donate more than one dollar. Thus we can infer that even though teacher tend to donate more than once, they tend to donate less than non-teacher. This might be due to the salary difference between teacher and non-teacher.

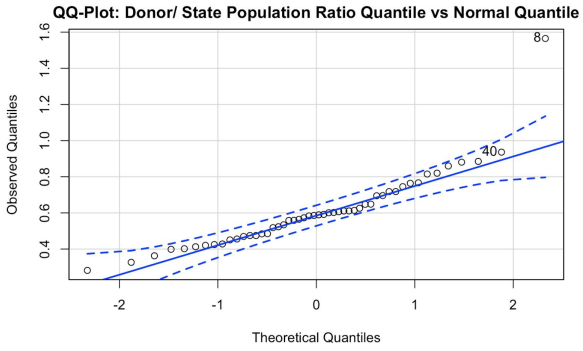
Scenario 5: Donor Population Ratio

Question: Which state has the highest donor to population ratio? Is the amount of donation correlated with the region donors live in?

1. Numerical Summary of donations amount in each region
2. Graphically display the distributions of donations

Q-Q Plot and Histogram of states' total donation and donor to population ratio

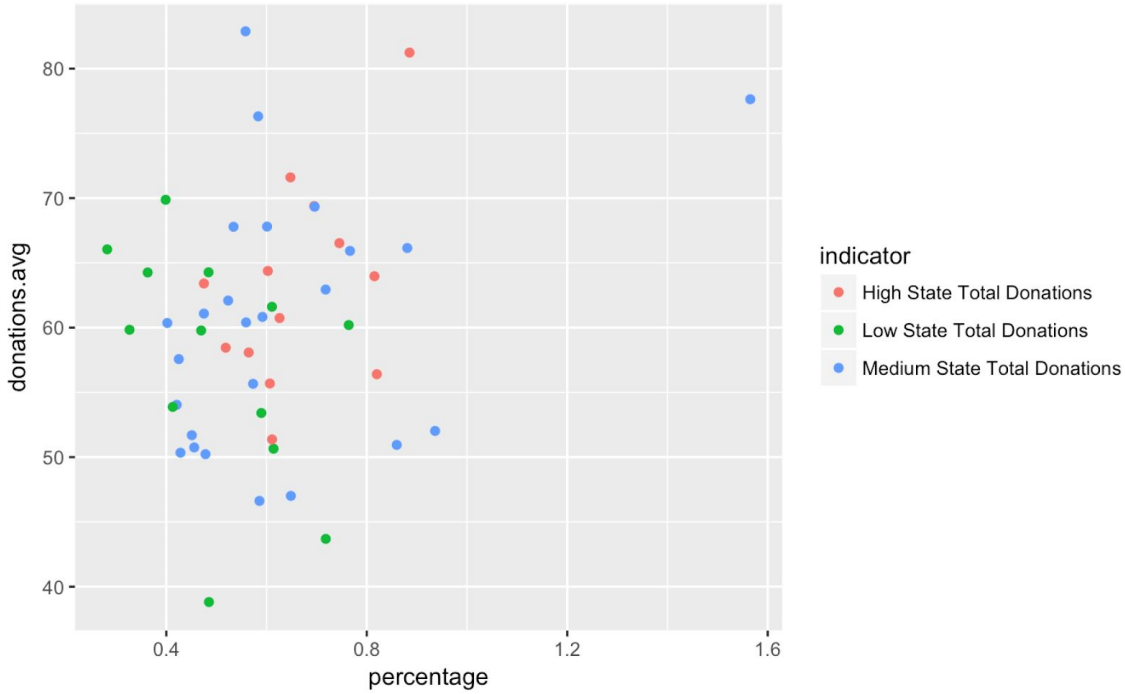




We are interested in finding the distribution of donor to state population ratio for each of the state. Then we use a histogram with frequency being the vertical axis and donor to state population ratio being the horizontal axis. From the graph, we can see the distribution of donor to state population ratio is approximately normal.

It is meaningful for us to find the distribution of donor to state population ratio for each of the state. We use a Q-Q plot with observed donor to population ratio quantile being the vertical axis and theoretically normal donor to state population ratio quantile being the horizontal axis. From the graph, we can see the white small circles fit the theoretical quantile line. Thus, we can conclude the distribution of donor to state population ratio is approximately normal.

Scatter plot of each state's total donation and donor to population ratio



Furthermore, we are going to discover each state's donation pattern, in terms of which state's donation amount per donor is higher or lower. Then we decide to use a scatter plot with each state's average donation amount being the vertical axis and each state's donor to population ratio being the horizontal axis, and we differentiate states into high/medium/low state total donations. From the scatter plot above, we can see majority of points are clustered. For example, for the state with medium state total donation, the average donation amount is medium, and the donor to population ratio is also medium. We consider those states as normal states, which means the donation amount per donor in each of those states is neither too low or too high, and the donation amount for each individual is similar. The right-most blue point has medium amount of total donation, but it has both high average donation and donor to population ratio. We infer it could be the case that some people donate quite a lot, while more people donate few, for instance 1 or 2 dollars. When this happens, the average donation amount can be boosted up by those who donate a lot, but the total donation amount is not high. This situation may possibly happen in states with high Gini coefficient with concaving up income distribution line, which means obviously most people are poor while a small portion of people are very rich.

Scenario 6: Regression Model Predicting

Question: If given certain feature of a donor and the project, is it possible to make a prediction on the amount that donor will donate?

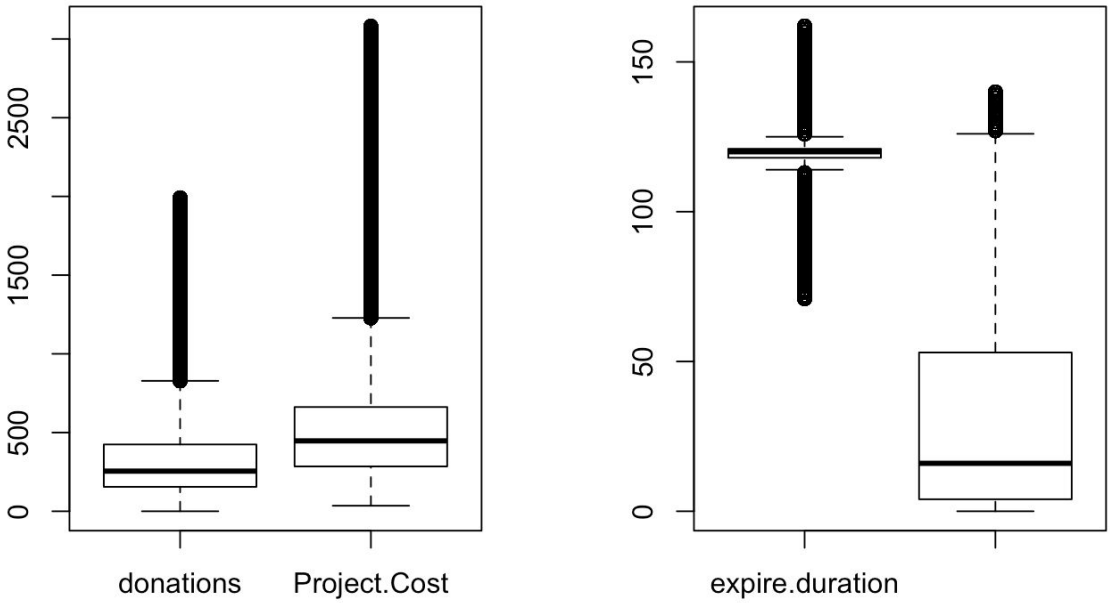
Multiple Regression Model:

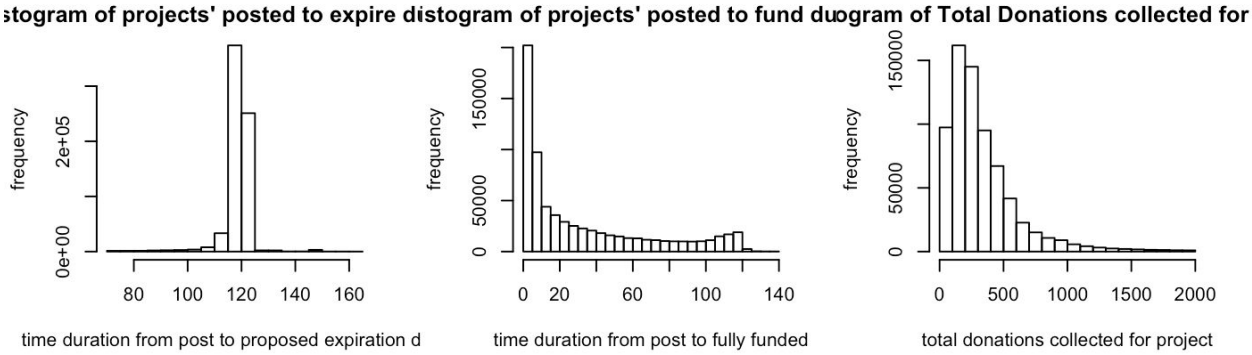
Having multiple variables in the dataset motivates a fitting of multiple regression to the data. Therefore, our model used categorical variables, such as Project Type, Project Grade level, and numerical variables Project Cost, and School Percentage Free Lunch. We performed backward elimination eliminating the variables that have small p-value in the prediction model and hurt the Adjusted R-squared data. As a result, the best model that covers the 68% explained variability of data we can get from the multiple regression is the following.

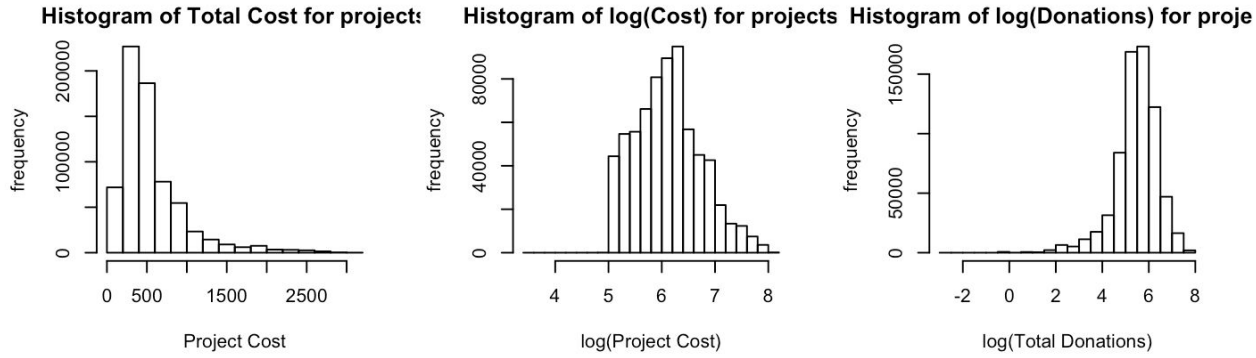
```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -2.634e+02  4.174e+01  -6.312 2.89e-10 ***
Project.TypeStudent-Led                  7.121e+01  5.272e+01   1.351   0.1768
Project.TypeTeacher-Led                  2.533e+02  4.040e+01   6.270 3.80e-10 ***
Project.Grade.Level.CategoryGrades 6-8   2.054e+00  1.028e+01   0.200   0.8416
Project.Grade.Level.CategoryGrades 9-12  2.353e+01  1.162e+01   2.026   0.0428 *
Project.Grade.Level.CategoryGrades PreK-2 1.995e+01 8.092e+00   2.465   0.0137 *
Project.Grade.Level.Categoryunknown     -1.223e+01  3.090e+02  -0.040   0.9684
Project.Cost                             6.134e-01  4.732e-03 129.623  < 2e-16 ***
School.Percentage.Free.Lunch            -5.148e-02  1.399e-01  -0.368   0.7128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 308.9 on 8221 degrees of freedom
Multiple R-squared:  0.6752,    Adjusted R-squared:  0.6749
F-statistic:  2136 on 8 and 8221 DF,  p-value: < 2.2e-16
```

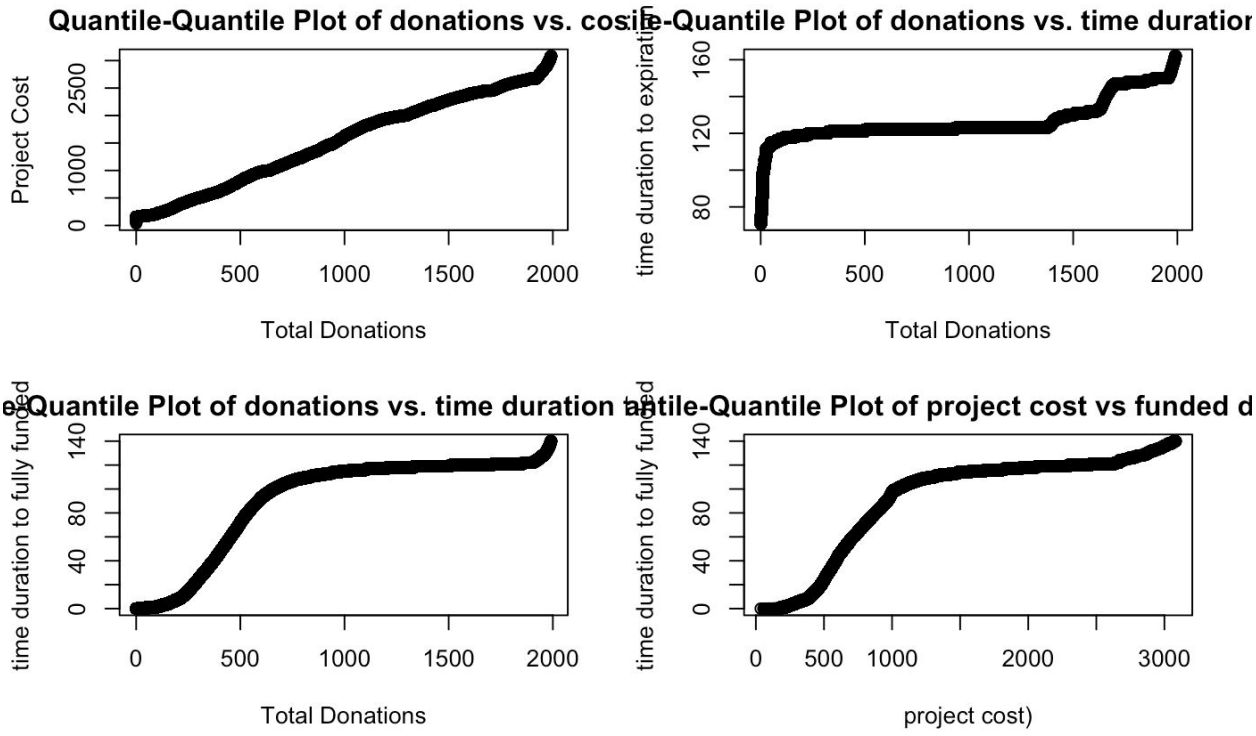
Least Squares Regression Model:

Upon looking at the variables that have an impact on the prediction of donations, we found out that the project cost is closely related to the donations amount projects received. Therefore, a least square regression model is performed upon the dataset to see how do projects' costs related to the donation amount a project can receive from donorschoose.org.


0
500
1500
2500
donations
Project.Cost
0
50
100
150
expire.duration


stogram of projects' posted to expire d
istogram of projects' posted to fund du
ogram of Total Donations collected for
frequency
0e+00
2e+05
80
100
120
140
160
time duration from post to proposed expiration d
0
50000
150000
0
20
60
100
140
time duration from post to fully funded
0
50000
150000
0
500
1000
1500
2000
total donations collected for project


Histogram of Total Cost for project
Histogram of log(Cost) for projects
Histogram of log(Donations) for proje
frequency
0
100000
200000
0
500
1500
2500
Project Cost
0
40000
80000
4
5
6
7
8
log(Project Cost)
0
50000
150000
-2
0
2
4
6
8
log(Total Donations)



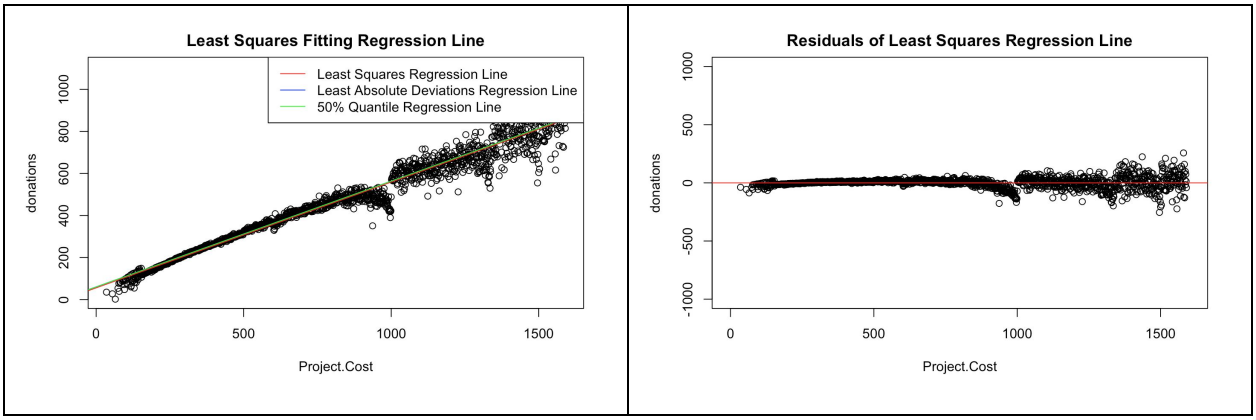
Upon first glance, it is obvious that the relationship between project cost and donation amount is strong, positive and linear, and this observation is supported by a very strong coefficient of determination (r^2) of 0.9559, which means 0.9559 of variability in response variable (donation) is explained by the linear model. Here, we mainly use least square regression model in our analysis, and we get the formula *donation amount* = 0.503(*project cost*) + 55.642 . We also use the other two methods: least absolute deviation regression line and 50% quantile regression line, and it turns out to be the same on the figure: three linear lines for project cost vs. donation. Furthermore, from the figure below, we can see that residuals of least square regression line model are all nearly 0, so this further supports that linear model fits the data well and this method turns out to be appropriate and robust.

Therefore, we can use this model to advice donors who proposed to have less than 1600 dollars as their project budget to make an estimate on how much donations they can possibly get from donors.choose.org by following the model *donation* = 0.503(*project cost*) + 55.642 .

```
Call:
lm(formula = donations ~ Project.Cost, data = test_avg)

Residuals:
     Min       1Q   Median       3Q      Max
-253.794  -13.233    4.784   17.807  257.371

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  55.642095   2.654947   20.96   <2e-16 ***
Project.Cost  0.503403   0.002801  179.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46.66 on 1490 degrees of freedom
  (2874 observations deleted due to missingness)
Multiple R-squared:  0.9559,    Adjusted R-squared:  0.9559
F-statistic: 3.231e+04 on 1 and 1490 DF,  p-value: < 2.2e-16
```

Three linear lines for project cost vs. donation

Residuals of Least Squares Regression Line



Scenario 7: Recommender System (Random Forest, Regression Trees)

Now we will begin constructing a recommender system, matching donors to projects. In particular, we will use projects, donations, donors, and schools attributes/features in order to predict project subject category.

To do this, we merged the Projects.csv, Donations.csv, Donors.csv, and Schools.csv datasets in order to extract features from multiple datasets. In particular, our merged dataset contains the column Project Subject Category Tree which we will predict using the other columns of Project Cost, Donation Amount, School Percentage Free Lunch, Donor is Teacher, School Metro Type, and Project Grade Level Category. We convert Donor is Teacher and School Metro Type columns to logical vectors of whether a donor is a teacher and whether a school metro type is suburban. In short, we use the numerical attributes Project.Cost, Donation.Amount, School.Percentage.Free.Lunch, the logical attributes Donor.Is.Teacher and School.Is.Suburban, and the categorical attribute Grade.Level in order to predict Subject.Category.

Subject.Category <fctr>	Project.Cost <dbl>	Donation.Amount <dbl>	School.Percentage.Free.Lunch <int>	Donor.Is.Teacher <lgl>	School.Is.Suburban <lgl>	Grade.Level <fctr>
Literacy & Language	178.55	25.00	70	FALSE	TRUE	Grades 3-5
Literacy & Language	178.55	50.00	70	FALSE	TRUE	Grades 3-5
Literacy & Language	178.55	15.77	70	FALSE	TRUE	Grades 3-5
Literacy & Language	178.55	50.00	70	FALSE	TRUE	Grades 3-5
Literacy & Language	178.55	25.00	70	FALSE	TRUE	Grades 3-5
Literacy & Language	178.55	10.00	70	TRUE	TRUE	Grades 3-5
Music & The Arts	522.02	345.00	82	FALSE	FALSE	Grades 9-12
Music & The Arts	522.02	50.00	82	FALSE	FALSE	Grades 9-12
Music & The Arts	522.02	10.00	82	FALSE	FALSE	Grades 9-12
Music & The Arts	522.02	50.00	82	FALSE	FALSE	Grades 9-12

1-10 of 276,715 rows

Previous 1 2 3 4 5 6 ... 100 Next

The base predictive model we will be using for our recommender system is a decision tree. To make our model more robust and less prone to overfitting, we will be using the random forest ensemble model. As mentioned earlier, our goal is to predict a project's subject category based off of projects, donations, donors, and schools attributes, as our ultimate goal is to match donors and their associated attributes with a project/project type.

We partitioned our dataset into training and testing datasets for the purposes of the random forest model.

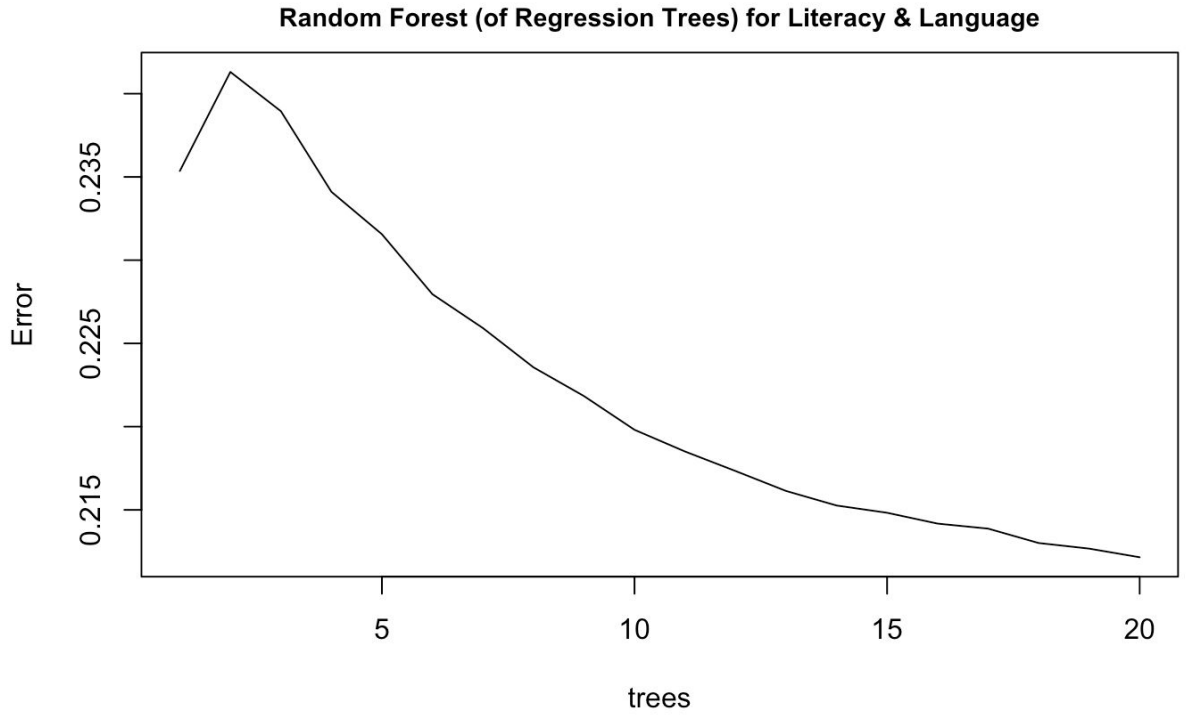
However, since the number of classes is large (eight), we will be using our model to indirectly predict a project. In particular, we will construct a random forest model (of regression trees) for each of the eight categories, denoting the probability of a project being of that project type. A sample regression tree of the random forest, the model error plots, and the variable importance plots for each of the eight categories are depicted below.

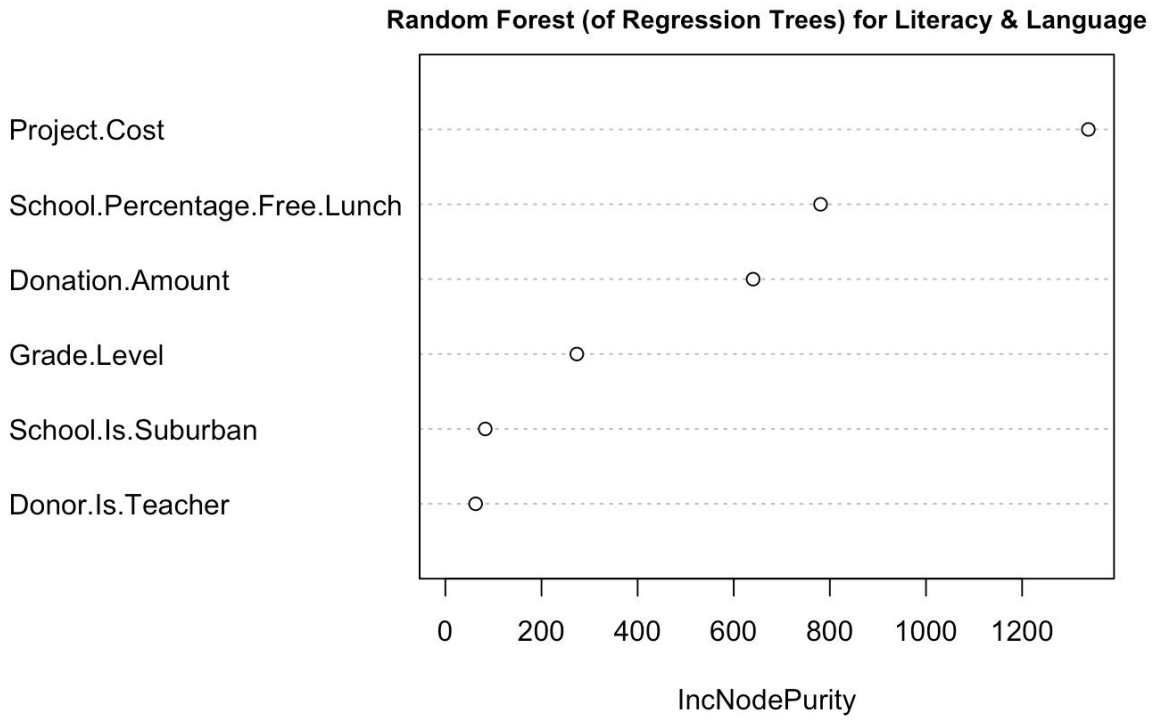
Literacy & Language

	left daughter <dbl>	right daughter <dbl>	split var <fctr>	split point <dbl>	status <dbl>	prediction <dbl>
1	2	3	Grade.Level	6.000	-3	0.35288000000000047107207
2	4	5	School.Percentage.Free.Lunch	61.500	-3	0.26289050652106815686437
3	6	7	Grade.Level	7.000	-3	0.38511898294033430900640
4	8	9	Project.Cost	1323.825	-3	0.23717059639389664882358
5	10	11	Donation.Amount	5.060	-3	0.27883552389141019745011
6	12	13	Donor.Is.Teacher	0.500	-3	0.34662170299891575941942
7	14	15	Project.Cost	2392.815	-3	0.41675739879242901242407
8	16	17	School.Percentage.Free.Lunch	50.500	-3	0.26497326203208981798554
9	18	19	Grade.Level	4.000	-3	0.15761285386381423956337
10	20	21	Grade.Level	2.000	-3	0.18954248366013382032236

1-10 of 4,299 rows

Previous 1 2 3 4 5 6 ... 100 Next



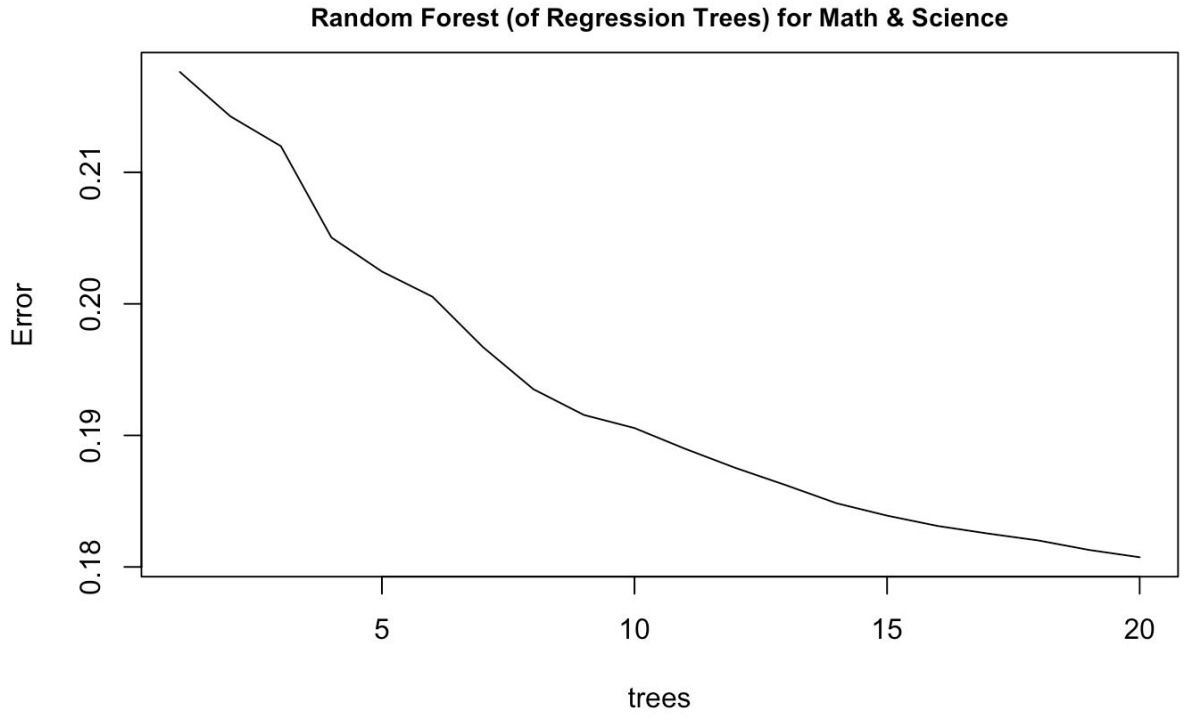


Math & Science

	left daughter <dbl>	right daughter <dbl>	split var <fctr>	split point <dbl>	status <dbl>	prediction <dbl>
1	2	3	Project.Cost	996.425	-3	0.25898000000000204190442
2	4	5	Donation.Amount	6.815	-3	0.24762271909534405711639
3	6	7	School.Percentage.Free.Lunch	35.500	-3	0.29882777276824956658174
4	8	9	Project.Cost	510.520	-3	0.20949761111915013445106
5	10	11	School.Percentage.Free.Lunch	22.500	-3	0.25585101396744047708154
6	12	13	Donor.Is.Teacher	0.500	-3	0.38125329120590040199446
7	14	15	Grade.Level	6.000	-3	0.28179741051028422971214
8	16	17	School.Is.Suburban	0.500	-3	0.17617526243723674594932
9	18	19	School.Percentage.Free.Lunch	53.500	-3	0.26732673267326706456615
10	20	21	School.Percentage.Free.Lunch	0.500	-3	0.30183255479698373191511

1-10 of 8,643 rows

Previous 1 2 3 4 5 6 ... 100 Next

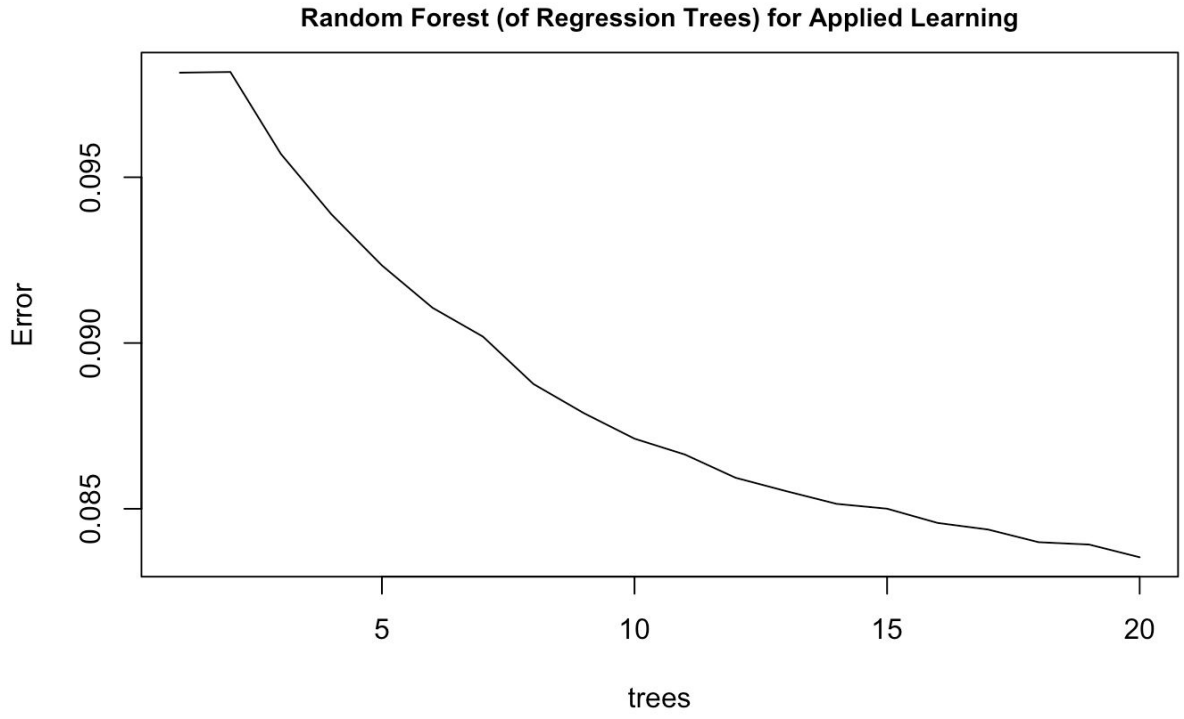


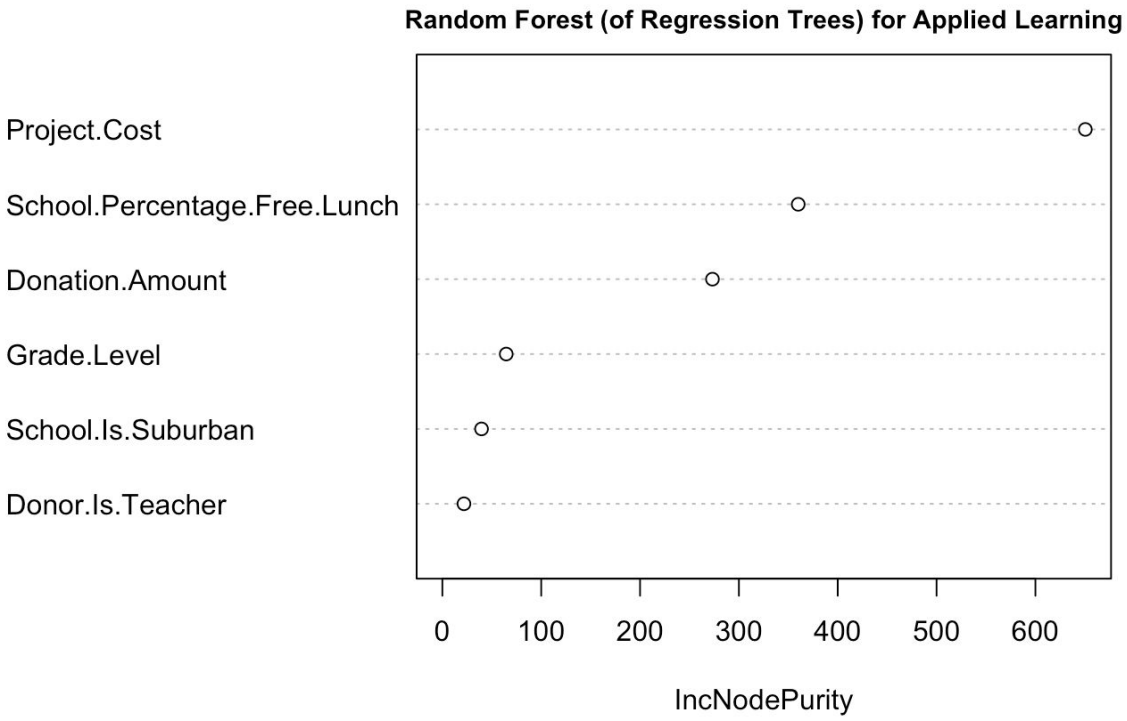
Applied Learning

	left daughter <dbl>	right daughter <dbl>	split var <fctr>	split point <dbl>	status <dbl>	prediction <dbl>
1	2	3	Project.Cost	4085.895	-3	0.10397999999999736620016
2	4	5	Grade.Level	3.000	-3	0.10190059965443623579695
3	6	7	Donor.Is.Teacher	0.500	-3	0.23105590062111580618165
4	8	9	Donor.Is.Teacher	0.500	-3	0.08395133738855900529252
5	10	11	Project.Cost	724.260	-3	0.11900130988767698703157
6	12	13	Project.Cost	4798.085	-3	0.23411371237458028504719
7	14	15	Project.Cost	4798.685	-3	0.22222222222222240417544
8	16	17	Project.Cost	2306.360	-3	0.07842426439490610134975
9	18	19	Project.Cost	1334.760	-3	0.09622869413501504431974
10	20	21	Project.Cost	166.100	-3	0.13144977852269665308249

1-10 of 5,055 rows

Previous 1 2 3 4 5 6 ... 100 Next



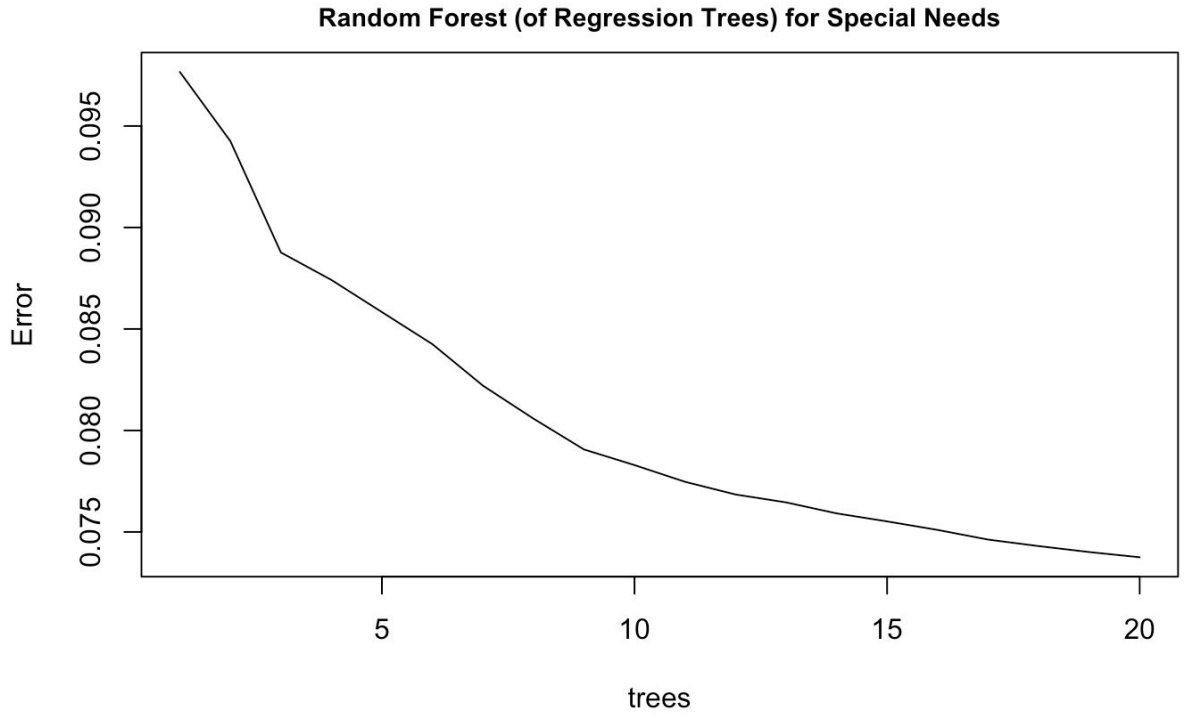


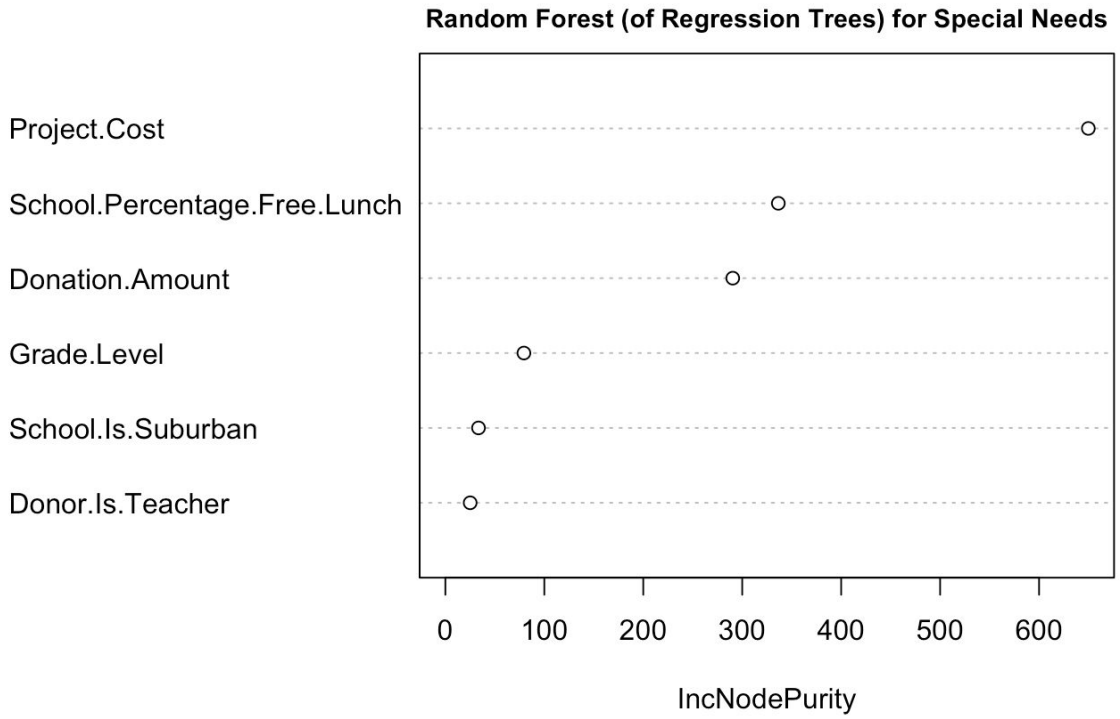
Special Needs

	left daughter <dbl>	right daughter <dbl>	split var <fctr>	split point <dbl>	status <dbl>	prediction <dbl>
1	2	3	Project.Cost	1162.090	-3	0.08863999999999894185976
2	4	5	Donor.Is.Teacher	0.500	-3	0.09246973365617640328029
3	6	7	School.Percentage.Free.Lunch	78.500	-3	0.07045977011494339303788
4	8	9	Project.Cost	1156.245	-3	0.08795811518324603284213
5	10	11	Project.Cost	447.885	-3	0.10268774703557484451188
6	12	13	School.Is.Suburban	0.500	-3	0.05585585585585721435375
7	14	15	School.Percentage.Free.Lunch	87.500	-3	0.09619047619047631225708
8	16	17	Grade.Level	13.000	-3	0.08748077729623943143977
9	18	19	Project.Cost	1158.025	-3	0.44736842105263163738016
10	20	21	School.Percentage.Free.Lunch	98.500	-3	0.08868243243243252582619

1-10 of 8,597 rows

Previous 1 2 3 4 5 6 ... 100 Next

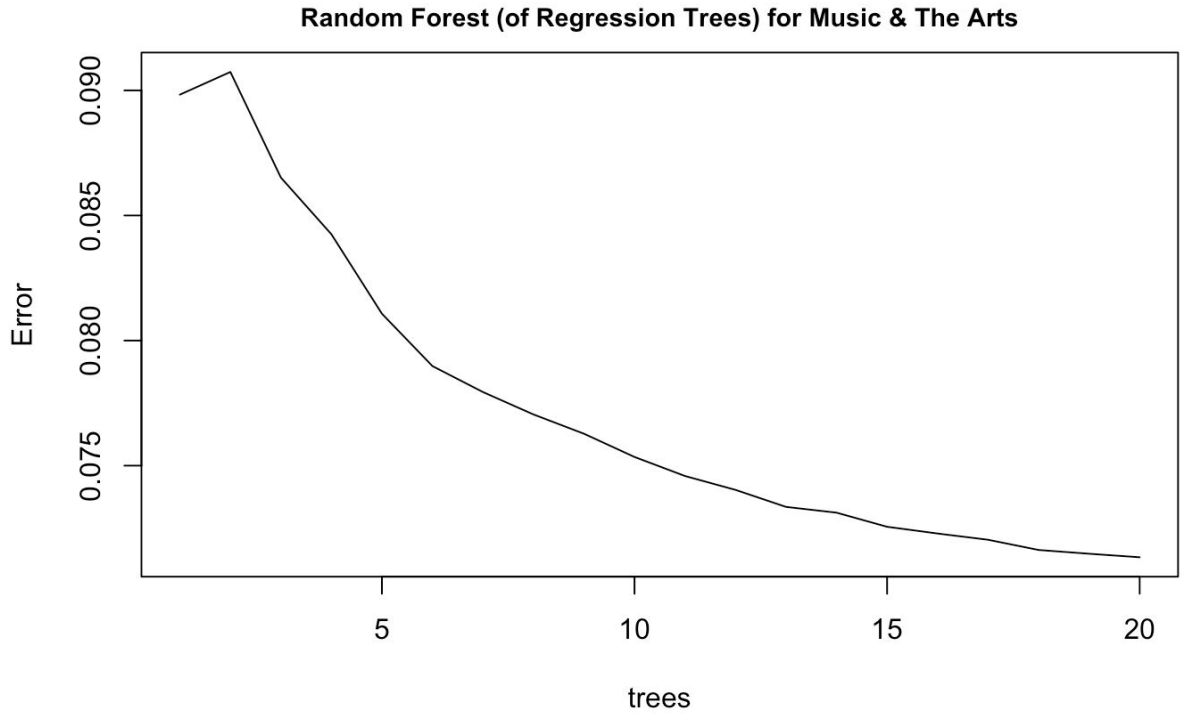

Random Forest (of Regression Trees) for Special Needs
Error
0.075
0.080
0.085
0.090
0.095
5
10
15
20
trees


Random Forest (of Regression Trees) for Special Needs
Project.Cost
School.Percentage.Free.Lunch
Donation.Amount
Grade.Level
School.Is.Suburban
Donor.Is.Teacher
0
100
200
300
400
500
600
IncNodePurity

	left daughter <dbl>	right daughter <dbl>	split var <fctr>	split point <dbl>	status <dbl>	prediction <dbl>
1	2	3	Grade.Level	8.000	-3	0.09385999999999958265384
2	4	5	School.Is.Suburban	0.500	-3	0.048484848484868673249884
3	6	7	School.Percentage.Free.Lunch	99.500	-3	0.12443923669235201767691
4	8	9	Project.Cost	156.055	-3	0.05155027092113349806990
5	10	11	Donation.Amount	1.450	-3	0.04253142356035843202333
6	12	13	Donation.Amount	4.005	-3	0.12204393312478467592541
7	14	15	Project.Cost	652.960	-3	0.62237762237762295214338
8	16	17	School.Percentage.Free.Lunch	53.500	-3	0.49999999999999944488885
9	18	19	School.Percentage.Free.Lunch	46.500	-3	0.05107729395811580475062
10	20	21	School.Percentage.Free.Lunch	89.500	-3	0.11241610738255036761046

1-10 of 4,375 rows

Previous 1 2 3 4 5 6 ... 100 Next

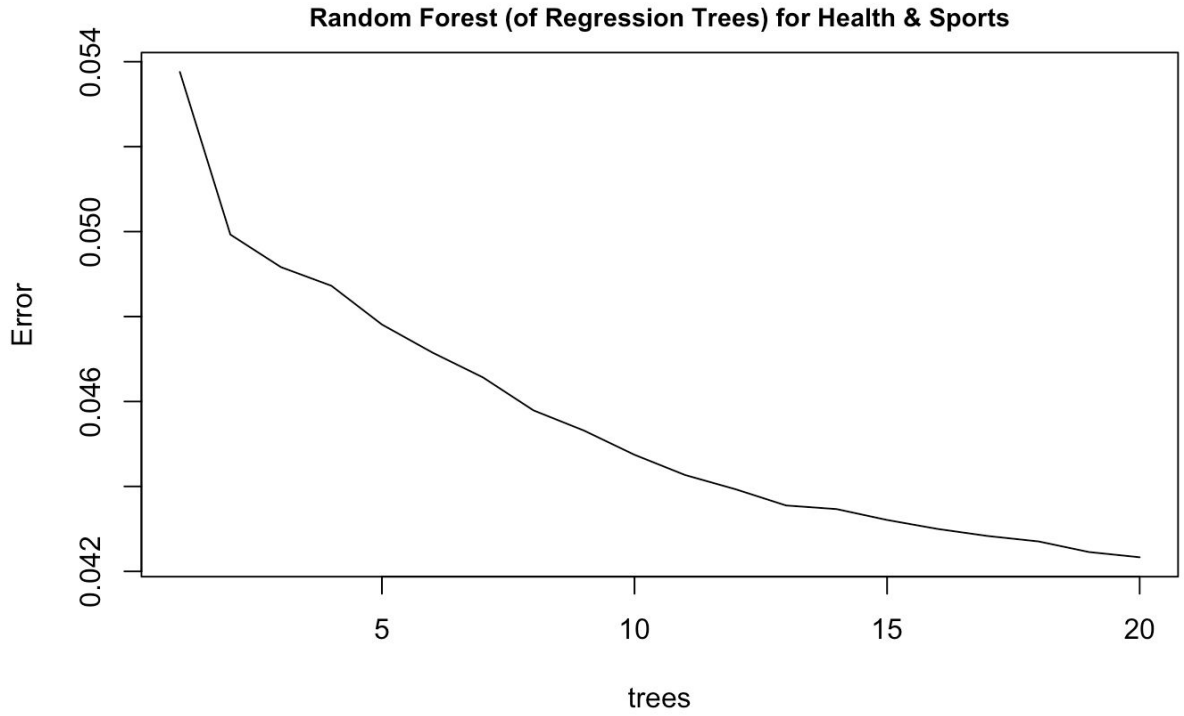


Health & Sports

	left daughter <dbl>	right daughter <dbl>	split var <fctr>	split point <dbl>	status <dbl>	prediction <dbl>
1	2	3	Grade.Level	10.000	-3	0.049039999999999438207166
2	4	5	Project.Cost	995.520	-3	0.038682206906513527666647
3	6	7	School.Is.Suburban	0.500	-3	0.061919504643969405788262
4	8	9	School.Percentage.Free.Lunch	81.500	-3	0.041036815107362081822551
5	10	11	Donation.Amount	7.470	-3	0.029556650246304706081135
6	12	13	School.Percentage.Free.Lunch	24.500	-3	0.057975826803027534495083
7	14	15	Donor.Is.Teacher	0.500	-3	0.070134181767879971713953
8	16	17	School.Percentage.Free.Lunch	71.500	-3	0.037605419604593431914807
9	18	19	Donor.Is.Teacher	0.500	-3	0.047600158667195142936368
10	20	21	School.Percentage.Free.Lunch	4.000	-3	0.005181347150258572287207

1-10 of 5,829 rows

Previous 1 2 3 4 5 6 ... 100 Next

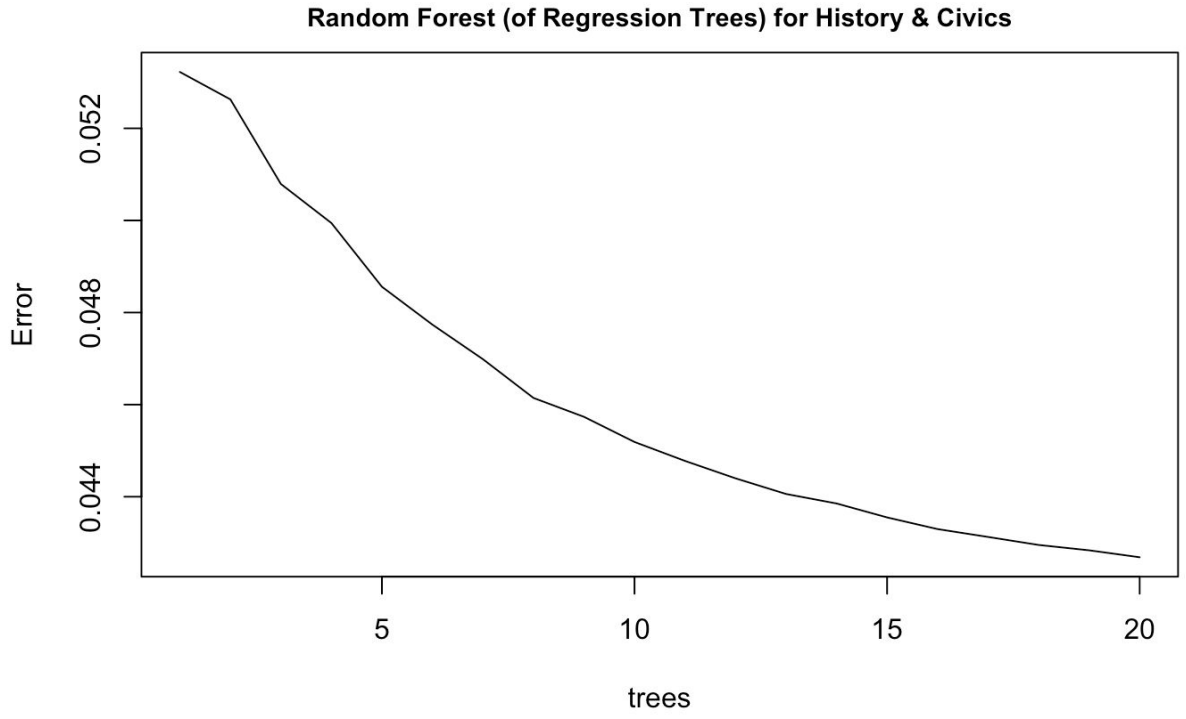


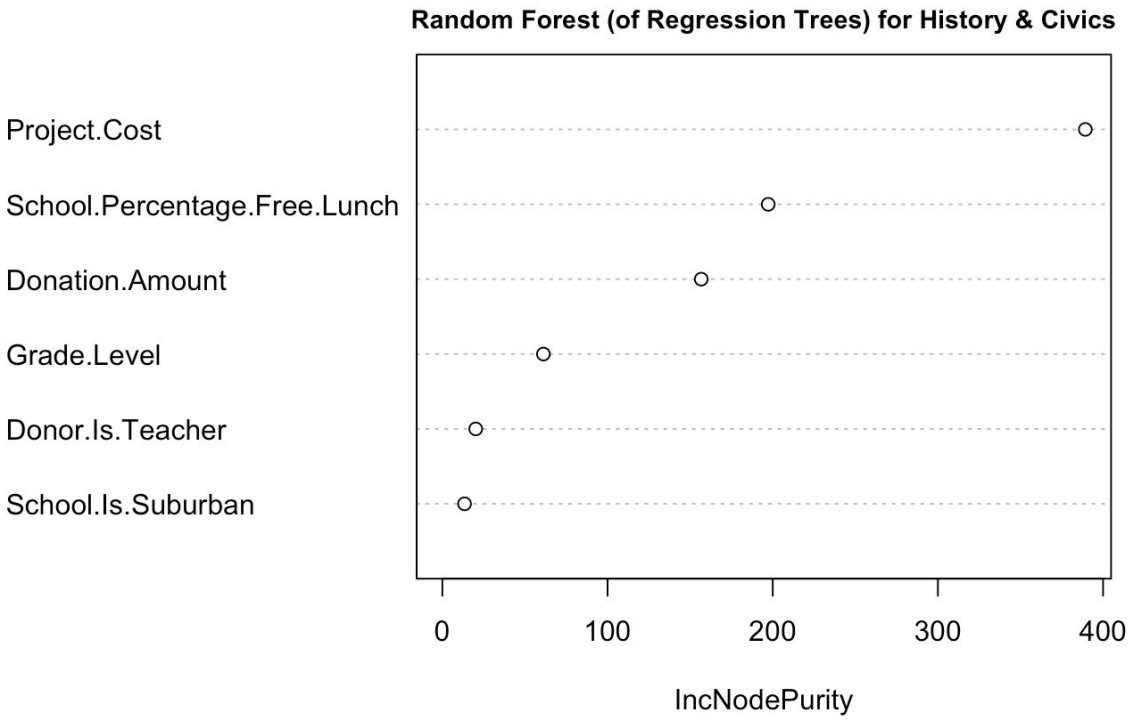
History & Civics

	left daughter <dbl>	right daughter <dbl>	split var <fctr>	split point <dbl>	status <dbl>	prediction <dbl>
1	2	3	Grade.Level	8.000	-3	0.049880000000000181137327
2	4	5	School.Percentage.Free.Lunch	77.500	-3	0.020922373521517219352450
3	6	7	Project.Cost	3337.340	-3	0.069382154093308723297184
4	8	9	School.Percentage.Free.Lunch	14.500	-3	0.017193155643296112455864
5	10	11	Donation.Amount	9.165	-3	0.026613105699218871785794
6	12	13	Project.Cost	155.270	-3	0.065294117647049704222795
7	14	15	Donation.Amount	72.050	-3	0.190184049079754530220399
8	16	17	School.Is.Suburban	0.500	-3	0.002183406113537178505624
9	18	19	Donation.Amount	116.310	-3	0.018416370106754433033647
10	20	21	School.Percentage.Free.Lunch	82.500	-3	0.056273764258555285244778

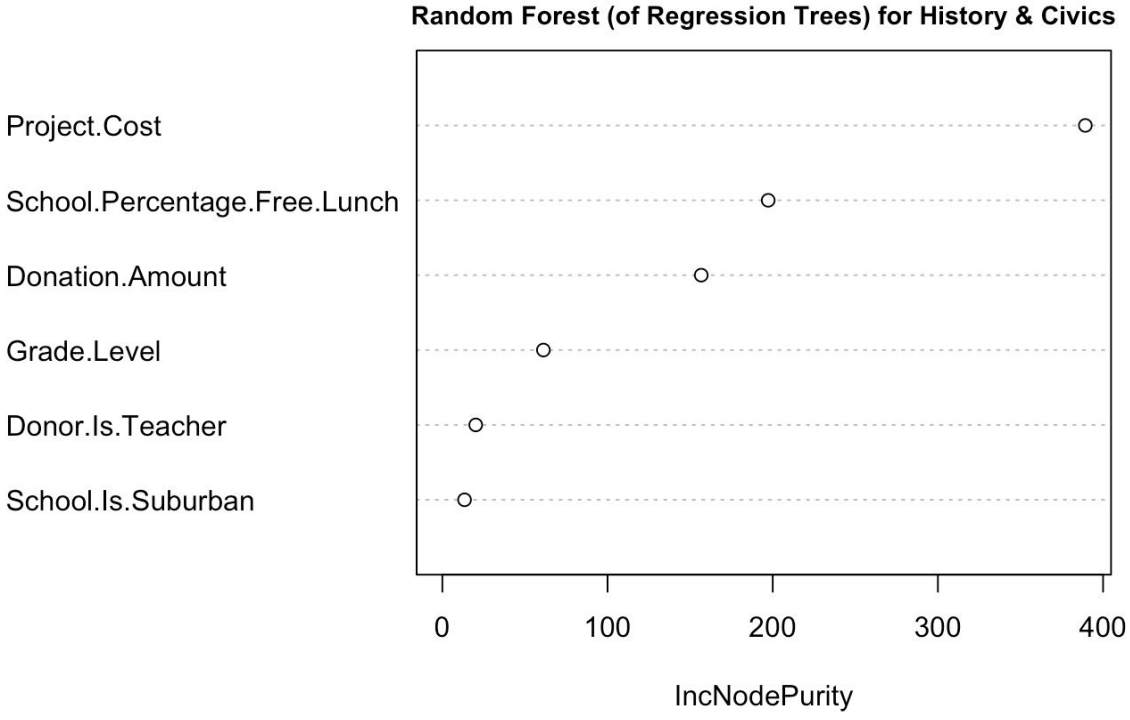
1-10 of 4,397 rows

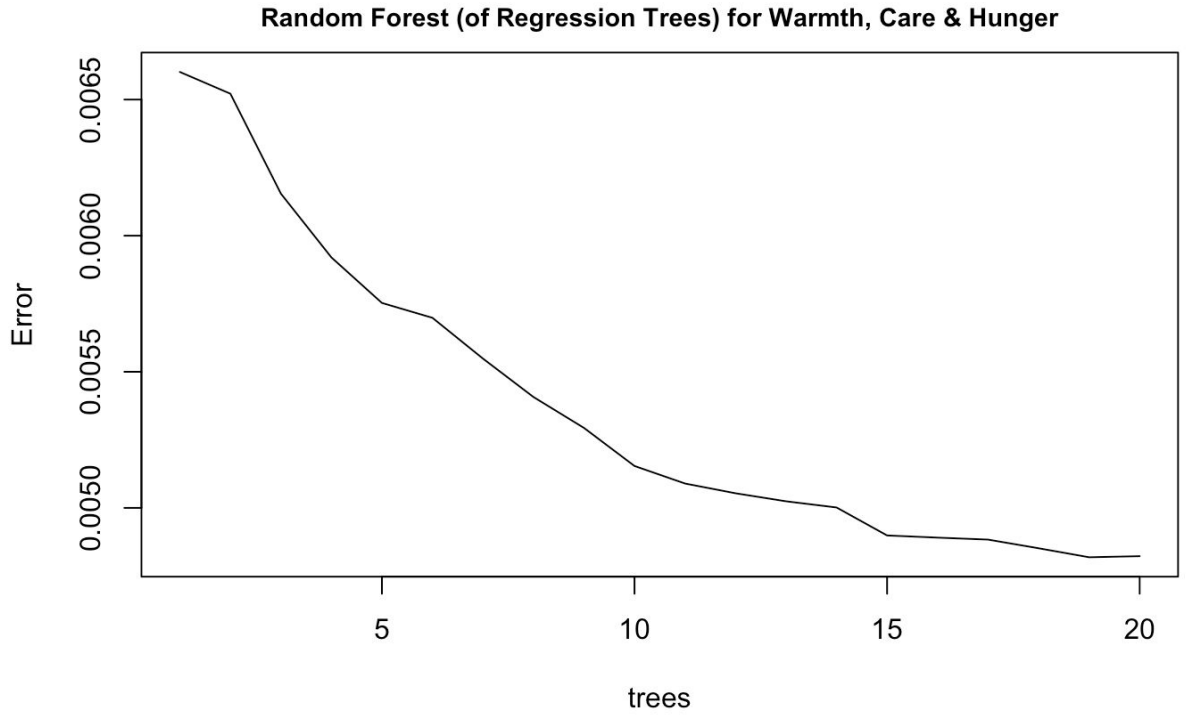
Previous 1 2 3 4 5 6 ... 100 Next

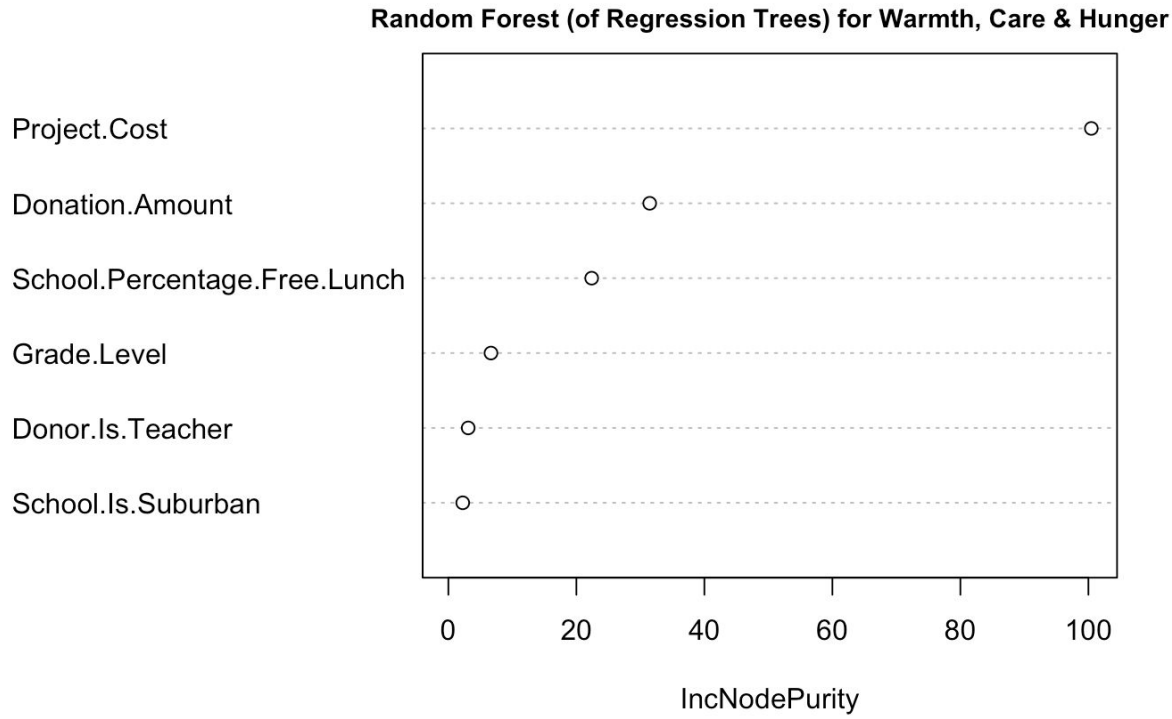




Warmth, Care & Hunger







From these results, we can observe that a good number of trees from bootstrapped samples in our random forest model is 20, before the error levels off. We can also observe that Project.Cost, Donation.Amount are the best predictors of our model, in terms of mean decrease

accuracy in the variable importance plots.

We then produce a table with the random forest model (regression tree) probabilities of being in each of the eight categories, then predict a category based off of the highest probability of being in a category in our testing data (excerpt below).

Subject.Category <fctr>	Predicted.Category <chr>	Literacy & Language <dbl>	Math & Science <dbl>	Applied Learning <dbl>	Special Needs <dbl>	Music & The Arts <dbl>
History & Civics	Math & Science	0.07015354	0.294330836	0.025269969	0.028869898	0.180701343
Music & The Arts	Literacy & Language	0.49967344	0.218545187	0.056829441	0.065908704	0.121619606
Music & The Arts	Literacy & Language	0.43776115	0.280754703	0.090634972	0.031852229	0.045282041
Health & Sports	Math & Science	0.19245000	0.354946619	0.093847985	0.016866300	0.020374182
Special Needs	Literacy & Language	0.33430329	0.188008512	0.037288160	0.257662781	0.043446726
Literacy & Language	Literacy & Language	0.33716764	0.188023768	0.169414433	0.039184107	0.155232640
Literacy & Language	Literacy & Language	0.43898081	0.243489464	0.084616892	0.055843054	0.039126815
Math & Science	Literacy & Language	0.44967121	0.188987412	0.083108038	0.102600370	0.036216325
History & Civics	Literacy & Language	0.32432709	0.169440786	0.163126274	0.046744394	0.095312765
Math & Science	Math & Science	0.30203234	0.399170036	0.054689858	0.077097039	0.124017024

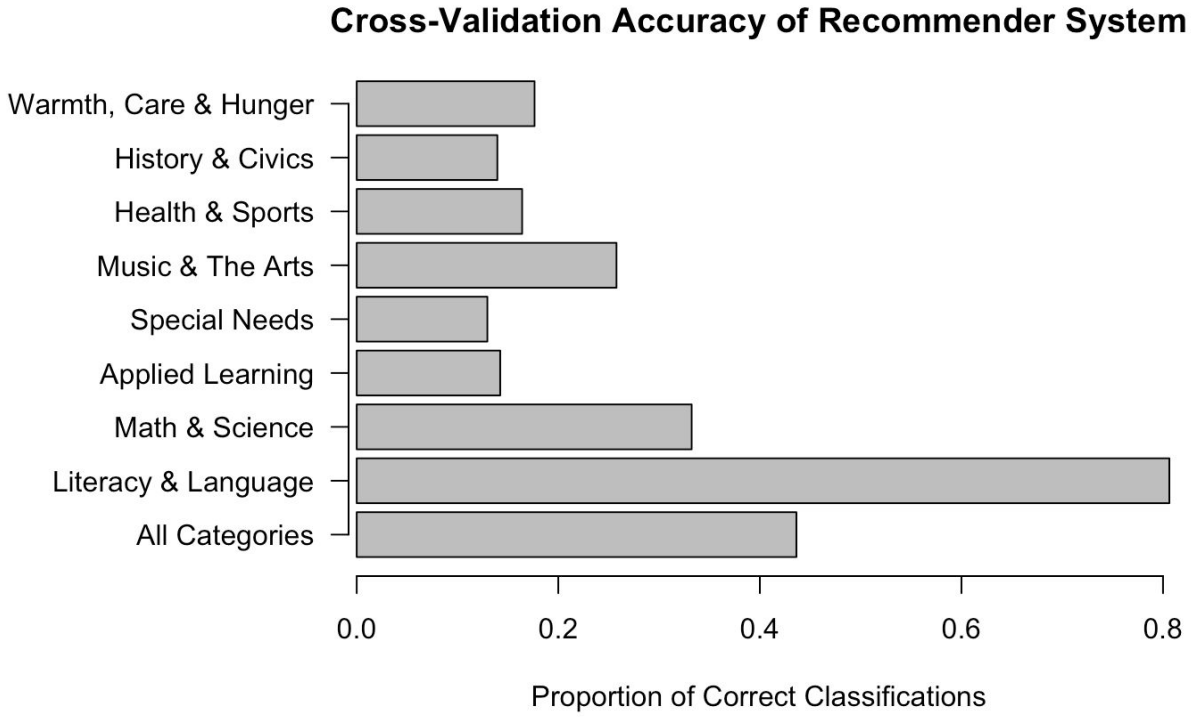
1-10 of 226,715 rows | 1-7 of 10 columns

Previous 1 2 3 4 5 6 ... 100 Next

We then use cross-validation to calculate the accuracy of our classifier on our testing data. A table and histogram of our cross-validation results are shown below.

Subject.Category <chr>	Correct.Predictions <int>	Total <int>	Accuracy <dbl>
All Categories	98905	226715	0.4362526
Literacy & Language	64424	79903	0.8062776
Math & Science	19452	58529	0.3323481
Applied Learning	3342	23473	0.1423763
Special Needs	2488	19190	0.1296509
Music & The Arts	5476	21246	0.2577426
Health & Sports	1853	11289	0.1641421
History & Civics	1660	11895	0.1395544
Warmth, Care & Hunger	210	1190	0.1764706

9 rows



In the end, we were able to construct a recommender system which matches donors (given their attributes) to a project type via a random forest model (of regression trees). DonorsChoose.org may use this classifier in order to effectively match donors to projects, increasing the total number of donations made.

CONCLUSION AND DISCUSSIONS

From Scenario 1, we are able to find keywords of projects across eight categories by using text mining and sentiment analysis, which helps us distinguish those different categories.

From Scenario 2, we delve into Donations.cvs and Donors.csv and use numerical and graphical methods to analyze donation amount, which has a total of 284,408,243 dollars and a median of 60.67 dollars. The distribution of donation amount is asymmetric and skewed to right. We also find the donor who donates the most, and the donor is from Anahola, Hawaii, who is not a teacher, donating to a teacher-led project under Health & Sports category, which is the project received the most amount of donations. Moreover, the amount of donations made by each states is calculated by bar graph, in which California has the highest total amount which is 46,140,356.5 dollars. Furthermore, the number of donations which do not include 15% donation to the website is higher than the number of donations which include 15% donation to the website in every interval of donation amount. Hence, in our recommendation system, we will recommend

that to guarantee more donations, 15% donation should not be included to the website.

In Scenario 3, we investigate seasonal pattern of posted projects and the result shows that the peak time for teachers posting projects is between July and October. Thus, we can predict a cluster of projects during these months and enforce advertisements in this time in order to attract as many customers as possible to satisfy these projects. Also, we observe an increasing trend of posted projects throughout the years, which means more and more teachers join Donors Choose for help, so Donors Choose should improve its website every year to satisfy expected increasing users.

From Scenario 4, we learn that there exists difference between donation patterns for teachers and non teachers, so in our recommendation system, we can divide donors into two groups: teachers and non teachers, and send distinct contents to different groups based on the difference in their donation patterns.

From Scenario 5, we focus on studying the donor population ratio. We first use histogram and Q-Q plot to find the distribution of donor to state population ratio is approximately normal. Then we use the scatter plot to investigate different states' different donation pattern. From it, we are able to infer state's economic situation in terms of distribution of income by analysing total donations, average donations and donor to population ratio, which are three variables represented in the scatter plot.

From Scenario 6, we use regression models to predict the amount that donor will donate. The multiple regression model covers the 68% explained variability of data. By using Least Square Regression Model, we find that the relationship between project cost and donation amount is strong, positive and linear, which is supported by a very strong coefficient of determination (r^2) of 0.9559. We then derive the formula for cost and donation amount, which is $donation = 0.503(project\ cost) + 55.642$. Therefore, we advice donors should have less than 1600 dollars as their project budget.

From Scenario 7, we use decision tree to construct recommendation system. In order to implement decision tree we use random forest. This is our last step and by constructing the recommendation system, we are able to connect the donors with the projects they favored in order to increase the chance of donating to that project. By using various features we are able to predict the Project Subject Category Tree and matches the donors to specific project type.

THEORY

Visualization

Box Plot - Box plot is a graphical display of the data based on its minimum, first quartile, median, third quartile, and maximum. The bottom and the top of the box describes the first quartile and third quartile respectively, and the line in the middle of the box shows the median. The whiskers above and below the box are maximum and minimum respectively. Dots outside the maximum and minimum are uncommonly far away from the mean, which are classified as the outliers.

Histogram - Histogram is a diagram displaying the frequency of data equally spaced numerical intervals. Histogram is a useful data visualization tool because it provides pivotal statistical information about the distribution of data.

Quantile-Quantile Plot - A Q-Q plot is a plot of the quantiles of one data set against the quantiles of another data set. Normally, it compares the quantiles of an observed dataset with quantiles of a theoretical distribution. If the two data sets come from the same distribution, the points should be located approximately along a 45-degree reference line. Information from the skewness and kurtosis of the Q-Q plot can also be used to examine how well the observed dataset fit the theoretical data distribution by looking at the shift in distribution and the fatness of tail.

Estimation

Bootstrap - Bootstrap is a resampling method that can be applied to finite samples to estimate for the population. The method can also build a confidence interval around the estimate via variance estimation. By Law of Large Number, the distribution of the sample generated by the simple random sample probability model looks roughy similar to that of the population, a new population with the same size based on the sample, defined as the bootstrap population, can be used to find the probability distribution of sampling average. To implement bootstrap, for every unit in the sample, we generate $\frac{N\ (population\ size)}{n\ (sample\ size)}$ units to produce a bootstrapped population with the same size as the population we are estimating and round the size off to the nearest integer. Once the bootstrap population is generated, the average can be calculated from taking a simple random sample of size n from the bootstrap population. This process is repeated for k iterations until k sample averages were generated. Then the histogram of those k bootstrap sample averages is generated to help evaluate the probability distribution of the sample average.

Normal Approximation and Confidence Intervals - By Central Limit Theorem, if sample size n is large and $x_1, ..., x_n$ are independent, identically distributed with mean μ

and variance σ^2 , then the probability distribution of sample average nearly follows normal distribution, that is, $Z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is approximately standard normal. Although $X_{I(j)}$ are not independent in simple random sampling, we can still apply normal approximation here if the sample size n is large enough but not too large comparing to the population size. Moreover, the normal distribution can give us confidence interval for the population parameter. For example, $(\bar{x}-\frac{\sigma}{\sqrt{n}}, \bar{x}+\frac{\sigma}{\sqrt{n}})$ and $(\bar{x}-\frac{2\sigma}{\sqrt{n}}, \bar{x}+\frac{2\sigma}{\sqrt{n}})$ are 68% and 95% confidence intervals respectively, which means that the probability that $\bar{x}$ is with one or two standard error of μ is about 68% or 95%. Thus, we can form a formula that $P(\bar{x}-\frac{2\sigma}{\sqrt{n}} \le \mu \le \bar{x}+\frac{2\sigma}{\sqrt{n}}) = P(\mu-\frac{2\sigma}{\sqrt{n}} \le \bar{x} \le \mu+\frac{2\sigma}{\sqrt{n}}) = P(2 \le \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \le 2) \approx 0.95$. Since σ is unknown in most cases, we replace σ by s in the formula, and the interval using s instead of σ is called approximate confidence interval.

Hypothesis Testing

Chi-Square Goodness of Fit Test - Chi-Square Goodness of Fit is operated when we have categorical variable in our dataset. The purpose of this test is to investigate if the sample data follows a hypothesized distribution. Prior to applying the test to the data, there are three assumptions of the dataset. First, the variable that is chosen should be categorical. Second, the sampling method of the sample needs to be simple random sampling. Third, the expected value of the observed data should be no less than 5 by rule of thumb. Any value lower than five should be grouped together into one bin. In general, the null hypothesis is that the data follows the hypothesized distribution. The decision rule for this statistical testing is to reject the null when p-value is smaller than 0.05, given the significance level is 95%.

F-test - An F-test is a statistical hypothesis test where its test statistic under the null hypothesis has a F-distribution. Prior to conducting a t-test, F-test is used to identify which version of two sample T-test is used (equal variance or unequal variance).
The test statistic is

$F = \frac{explained\ variance}{unexplained\ variance} = (\sum_{i=1}^{k} \frac{n_i(\overline{Y}_i-\overline{Y})^2}{K-1})/(\sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(Y_{ij}-\overline{Y}_i)^2}{N-K})$. $\overline{Y}_i$ is the sample mean of the ith group. $\overline{Y}$ is the overall mean of the sample. K is the number of groups. N is the sample size. The statistics will be large when the explained variance is larger than the unexplained variance. This is unlikely to take place when the population mean of all groups are the same.

Kolmogorov-Smirnov Test - Kolmogorov-Smirnov test is a kind of non-parametric test, which tests the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution, or to compare two samples. According to Method of Moment, the empirical cumulative distribution function is defined by

$Fn(t) = (\frac{1}{n}) \sum_{i=1}^{n} I\{Xi \leq t\}$. By Strong Law of Large Number, Fn(t) does estimate $Fn(t) \rightarrow F(t)$ as $n \rightarrow \infty$. By Fundamental Theorem of Statistics, $Sup|Fn(t) - F(t)| \rightarrow 0$ as $n \rightarrow \infty$. In the case where we compare a sample with a reference probability distribution, H0: $P = P0$, H1: $P \neq P0$.

The test statistic Dn =

$\sqrt{n}\, Sup|\, Fn(t) - F(t)\,| = \sqrt{n}\, Max\left\{\left|F(X_i) - \frac{i}{n}\right|,\ \left|F(X_i) - \frac{i-1}{n}\right|\right\}$ $for\ 1 \leq i \leq n$.

The rejection region $\Re = \{Dn \geq s\}$. In the two sample test of homogeneity case, $H0 : Px = Py$, $H1 : Px \neq Py$ Test statistic Dn,m = $\sqrt{\frac{nm}{n+m}} \times Sup|Fn(t) - Gm(t)|$.

Rejection region $\Re = \{Dn, m \geq s\}$.

Power - The power of any test of statistical significance is defined as the probability that it will reject an incorrect null hypothesis, which is related to type I error α and type II error β. Power = $1 - \beta = \alpha$. $1 - \beta = 1 - p(accept\ H0|H1\ is\ true)$, and $\alpha = p(reject\ H0|H0\ is\ true)$.

Robustness - Robustness can be described in different kinds of scenarios, the most common ones are robustness to outliers, robustness to non-normality, and robustness to non-constant variance.

In the case of tests, robustness usually refers to the test still being valid given such a change. In other words, whether the outcome is significant or not is only meaningful if the assumptions of the test are met. When such assumptions are relaxed (i.e. not as important), the test is said to be robust.

Student's t-test - A t-test is conducted to investigate if two sets of data are statistically significantly different from each other. It can be applied when the targeted data follows a normal distribution if the value of a scaling term in the test statistic was known. The null hypothesis suggests that the mean of the first data set equals the mean of the second data set, while the alternative hypothesis is the mean of data set 1 differs from the mean of data set 2. P-value of a t-test will be calculated through R. Then by comparing p-value with significance level, decision rule of the hypothesis testing can be drawn.

Test of Independence - Test of independence is a type of goodness of fit hypothesis testing, which measures the deviation of observed values and expected values. The null hypothesis is that two variables are independent, while the alternative hypothesis is two variables are not independent. By comparing p-value with significance level, decision rule of the hypothesis testing can be made.

Wilcoxon Signed-Rank Test - The Wilcoxon signed-rank test is a non-parametric test aiming to compare two related samples and to assess whether their population mean ranks differ. Unlike paired t-test, Wilcoxon test does not require the assumption of normality. It compares the difference of two signed rank, and distinguishes whether the difference between two sets are statistically significant. In general,
$Z_i = 1$ *if* $X_i - \mu_0 > 0$; $Z_i = 0$ *if* $X_i - \mu_0 < 0$. Ri is the rank of |Xi-Yi|.
Test statistic T = $\sum_{i=1}^{n} R_i Z_i$.
$T \sim P(T = t) \times (\frac{1}{2})^{n \times c(t)}$ *when* $n \leq 12$; $T \sim N(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24})$ *as* $n \to \infty$.
If n is large, $Z = \frac{T - n(n+1)}{\sqrt{n(n+1)(2n+1)/24}}$.
We can also apply one-sided test to further investigate the sign of the difference. The decision rule is as following:
a. To test H0:μ = μ0 versus H1:μ>μ0 at the α level of significance, reject H0 if $Z \geq Z_\alpha$.
b. To test H0:μ = μ0 versus H1:μ < μ0 at the α level of significance, reject H0 if $Z \leq -Z_\alpha$.
c. To test H0:μ=μ0 versus H1:μ≠μ0 at the α level of significance, reject H0 if z is $Z \leq -Z_{\frac{\alpha}{2}}$ *or* $Z \geq Z_{\frac{\alpha}{2}}$.

Prediction

Bootstrap Aggregating/Bagging - Bagging is a method used commonly in regression and classification to improve the accuracy and stability of prediction, for example to make decision tree. The basic ideas for bagging are a weak learning algorithm is given, and a training set is given; the accuracy of single weak learning algorithm is not high; the learning algorithm is used many times to get the prediction function sequence and vote; the accuracy of the final result will be improved.

The algorithm for Bagging:

1. Sampling from data set S (put samples back)
2. Training model Ht

3. For X classification of unknown samples, each model Ht gets a classification. The highest number of votes is the classification of unknown samples X.

In addition, the average value of the votes can also be used to predict continuous values.

CART (Classification and Regression Tree) - The outcome (dependent) variable is a categorical variable and predictor (independent) variables can be continuous or categorical variables.

Procedures:

1. Pick the variable that gives the best split (based on lowest Gini index).
2. Partition the data based on the value of this variable.
3. Repeat steps 1 and 2. Splitting stops when CART detects no further gain can be made, or some pre-set stopping rules are met.

Decision Tree - A decision tree is a graphical representation of possible solutions to a decision based on certain conditions. The three main components of a decision tree are:

- Root Node: The top most node that implies the best predictor (independent variable).
- Decision/Internal Node in which predictors are tested and each branch represents an outcome of the test.
- Leaf/Terminal Node: holds a class label. The advantages of decision tree are first it is easy to interpret, and works even if there is nonlinear relationships between variables. Lastly, it is not sensitive to outliers. Disadvantages of decision tree is it usually causes overfitting problem. It assumes all independent variables interact with each other.

Lasso (Least Absolute Selection Shrinkage Operator) Regression - Lasso is a kind of regression method that does variable selection and regularization. Lasso regression with outcome variable y_i, for cases i=1, 2, …, n, features x_{ij}, $j = 1, 2, ...p$ tries to minimize $\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right|$, which is equivalent to minimizing sum of squares with constraint $\sum \left|\beta_j\right| \leq s$. It is similar to ridge regression, which has constraint $\sum_j B^2{}_J \leq$ t.

Lasso does variable selection and shrinkage. The shrinkage process is defined as constraint on parameter that shrinks coefficients toward zero. Selection is defined as identifying the most important variables associated with the response variable. In statistics and machine learning, Lasso is used to enhance the prediction accuracy and interpretability of the statistical model it produces.

Lasso (ℓ1) penalties have powerful statistical and computational advantages, which provide a natural to enforce sparsity and simplicity in the solution. ℓ1 penalties are convex and the assumed sparsity can lead to significant computational advantages.

Pathwise coordinate descent works by holding all the coordinate directions constant except for one, and then minimizing the function subject to each of the coordinate directions, one at a time. Coordinate descent is guaranteed to provide a solution to the multivariate Lasso problem because the Lasso objective function is convex and the penalty term is separable according to each of the block coordinates. Its solution is the soft-thresholded estimate sign (βˆ)(|βˆ| − λ)+ where βˆ is usual least squares estimate. With multiple predictors, we can cycle through each predictor in turn and compute residuals $r_i = y_i - \sum_{j \neq k} x_{ij} . \beta_k$ and apply univariate soft-thresholding, pretending that our data is (x_{ij} , r_i). To find the pathwise coordinate descent for the lasso, start with large value for λ (very sparse model) and slowly decrease it. Most coordinates that are zero never become non-zero.

Pathwise coordinate descent can be generalized to many other models, such as logistic/multinomial for classification, graphical lasso for undirected graphs, fused lasso for signals. Its speed and simplicity are quite remarkable.

Least Squares Regression - Least squares linear regression is a method for predicting the value of a dependent variable Y, based on the value of an independent variable X. Linear regression finds the straight line, called the least squares regression line, that best represents observations in a bivariate data set. Suppose Y is a dependent variable, and X is an independent variable. The population regression line is $Y = \beta_0 + \beta_1 X$, where β_0 is a constant, β_1 is the regression coefficient, X is the value of the independent variable, Y is the value of the dependent variable.

The regression line has the following properties:

1. The line minimizes the sum of squared difference between observed values (y) and predicted values ($\hat{y}$).
2. The regression line passes through the mean of the X values and through the mean of the Y values.
3. The regression constant (β_0) is equal to the y intercept of the regression line.

4. The regression coefficient (β_1) is the average change in the dependent variable (Y) for a 1-unit change in the independent variable (X). It is the slope of the regression line.

The least squares regression line is the only straight line that has all of these properties.

Logistic Regression - Logistic regression is a type of generalized linear model which is used for binary outcome data, that is, Y can only be 0 or 1. The probability mass function of logistic regression is $P(Y=1,X=x)=\frac{exp(\beta' x)}{1+exp(\beta' x)}=\frac{1}{1+exp(-\beta' x)}$ and $P(Y=0,X=x)=1-P(Y=1,X=x)=\frac{1}{1+exp(\beta' x)}$ where $X_0=1$ and β_0 is the intercept.

There are two functions used for mapping. The logit function logit(x)=log(x/(1-x)) maps the unit interval onto the real line. The expit function, which is the inverse of logit function, expit(x)= $\frac{e^x}{(1+e^x)}$ maps the real line onto the unit interval. In logistic regression, the logit function is used to map linear predictor $\beta' X$ to a probability. Here, the linear predictor is computed as $\beta' X=log[\frac{P(Y=1,X)}{P(Y=0,X)}]$. Thus, one unit increase in X_j results in a change of β_j in the conditional log odds.

We can view the binary outcome Y as a dichotomization of a latent continuous outcome Y_c, so $Y=I(Y_c\geq 0)$. Suppose $Y_c|X$ follows a logistic distribution with CDF: $F(Y_c|X)=\frac{exp(Y_c-\beta' X)}{1+exp(Y_c-\beta' X)}$, so Y|X follows the logistic regression model $P(Y=1|X)=P(Y_c\geq 0|X)=1-\frac{exp(0-\beta' X)}{1+exp(0-\beta' X)}=\frac{exp(\beta' x)}{1+exp(\beta' x)}$.

We have mean and variance for logistic regression as following:
$E(Y|X)=P(Y=1|X)=\frac{exp(\beta' x)}{1+exp(\beta' x)}$,
$var(Y|X)=P(Y=1|X)*P(Y=0|X)=\frac{1}{2+exp(\beta' X)+exp(-\beta')}$
Since the variance depends on X, logistic regression models are always heteroscedastic.

Suppose we sample people based on their disease status D where D=1 is a case and D=0 is a control. We are interested in a binary marker M $\in \{0,1\}$ that may predict a person's disease status. The prospective log odds $log[\frac{P(D=1|M=m)}{P(D=0|M=m)}]$ measures how informative the marker is. We want to model M|D using logistic regression, so $P(M=1|D)=\frac{exp(\alpha+\beta D)}{1+exp(\alpha+\beta D)}$ and $P(M=0|D)=\frac{1}{1+exp(\alpha+\beta D)}$. The prospective log odds can be computed as $log[\frac{exp(M*(\alpha+\beta))/(1+exp(\alpha+\beta))}{exp(M*\alpha)/(1+exp(\alpha))}\times\frac{P(D=1)}{P(D=0)}]$, and it will have the form $\theta+\beta M$. Thus, the

coefficient β when we use logistic regression to regress M on D using case-control data is the same as that we obtain from regression D on M in a prospective study.

Assuming independent cases, the log-likelihood for logistic regression is $L(\beta|Y,X) = log\prod_i \frac{exp(Y_i*\beta' X_i)}{1+exp(\beta' X_i)}$. The logistic regression model is fit using maximum likelihood estimation. The score function, which is the gradient of the log-likelihood function is $G(\beta|Y,X) = \sum_{i:Y_i=1} X_i - \sum_i \frac{exp(\beta' X_i)}{1+exp(\beta' X_i)} * X_i$. The Hessian of the log-likelihood is $H(\beta|Y,X) = -\sum_i \frac{exp(\beta' X_i)}{(1+exp(\beta' X_i))^2} \times X_i X_i'$. The Hessian is strictly negative definite, so $L(\beta|Y,X)$ is a concave function of β and has a unique maximizer which means the maximum likelihood estimation is also unique. The Fisher information $I(\beta) = -[EH(\beta|Y,X)|X]^{-1} = -H(\beta|Y,X)^{-1}$ is the asymptotic sampling covariance matrix of the MLE $\widehat{\beta}$. If the model is correctly specified, I(β) is consistent, asymptotically unbiased and asymptotically normal.

Multiple Observation - Suppose we have m distinct values of explanatory variables X. For each Xi, we have k replicate measurements, then we have the formula Y_{ij} =a+ bX_i + ε_{ij} where i=1,2,…,m and j=1,2,...,k. Here, X is the explanatory variable, and Yij is the dependent variable which is jth measurement taken at Xi. We will assume that εij is uncorrelated for all i and j. We use replicate measurement to estimate σ^2 that does not rely too much on the model. If the residual is filed incorrectly, the residuals include measurement error from εij as well as model misfit. Suppose $Y_{ij} = c+d X_i + e\mu_i + \varepsilon_{ij}$ but simple linear model is fit by least squares. Then, $Y_{11}, Y_{12} \ldots Y_{1k}$ are k uncorrelated variables. $E[Y_{11}] = E[c + dX_1 + e\mu_1 + \varepsilon_{11}] = c + dX_i + e\mu_i$ and $var(Y_{11}) = \sigma^2$. From $(S_1)^2 = \frac{1}{k-1}\sum_{j=1}^{k}(Y_{1j} - \overline{Y})^2$ and $\overline{Y} = \frac{1}{k}\sum_{j=1}^{k} Yij$, we get $S_1{}^2$, $S_2{}^2$ …, $S_m{}^2$ (no regression), so $S_{pooled}{}^2 = \frac{1}{m}\sum_{i=1}^{m} Si^2$.

Random Forest - Random forest refers to a classifier trained and predicted by multiple trees, and the type of output is determined by the number of classes exported by individual trees.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners:

1. N is used to represent the number of training cases, and M represents the number of features.
2. The input characteristic number m is used to determine the decision result of a node on the decision tree, and the m should be much less than M.
3. From the N training use case, sample N is sampled, and a training set (bootstrap sampling) is formed, and the error is evaluated with the use case that is not drawn.
4. For each node, m features are randomly chosen, and the decisions of each node in the decision tree are determined based on these characteristics. According to these m characteristics, the best way of splitting is calculated.
5. Each tree will grow without pruning, which may be adopted after building a normal tree classifier.

The advantages of random forests are:

1. For many kinds of data, it can produce a classifier with high accuracy.
2. It can handle a large number of input variables.
3. When building a forest, it can generate an estimate of the internal error for the general error.
4. It can assess the importance of variables when deciding categories.
5. It contains a good method to estimate the missing data, and if a large part of the data is lost, it can still maintain accuracy.
6. It provides an experimental method to detect variable interactions.

Ridge Regression - Ridge regression is an estimation regression method which is used for collinear data analysis. Similar to the ordinary least squares method, ridge regression aims to minimize the squares error. However, ridge regression implements a constraint for $\sum_{j=1}^{p} \beta_j^2 \leq c$. In essence, it is an improved ordinary least squares estimation method since ordinary least squares estimation may have the issue of over-fitted or under-fitted. With a penalty term added to the equation, the formula greatly reduced model overfitting our data. To improve the ordinary least squares estimation, a regularization term can be included in the minimization problem: $||Ax - b||^2 + ||\Gamma x||^2$ for some suitably chosen Tikhonov matrix Γ, which is chosen as a multiple of the identity matrix ($\Gamma = \alpha I$). The explicit solution, denoted by $\hat{x}$, is given by $\hat{x} = (A^T A + \Gamma^T \Gamma)^{-1} A^T b$.

Miscellaneous

Cross-Validation - Cross-validation is a technique used to assess how accurate a predictive model will perform in practice by dividing the original sample into training set

(known data) on which model is trained and testing set (unknown data) on which model is tested. In cross-validation, we first leave a set of data points out for validation, then perform our model on the remaining data to provide an estimation interval for the data point that we just left out. Finally, we check where the true data point falls to assess accuracy of our model. In our linear regression model, we have n real response values $y_{1,} \ldots y_n$ corresponding to $x_1, \ldots x_n$. Suppose we have the function $y = \alpha + \beta x$ to fit the pattern. Then, we can assess the model by calculating mean squared error (MSE), where $\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_{actual} - y_{predicted})^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$.

Homogeneous Poisson Process - The homogeneous Poisson process is a process that arises naturally from the notion of points randomly distributed on a line without obvious regularity. Many random phenomena can be modeled by the homogeneous Poisson process such like arrival times of calls at an exchange, the decay times of radioactive particles and positions of stars in parts of the sky. The homogeneous Poisson process has three major characteristics: 1. The rate λ at which points occur does not change with location. 2. The number of points falling in different regions are independent. 3. No two points can hit at the same place.

Counts of the number of points in different regions follow Poisson distribution with rate λ. P(k points in a unit interval)= $\frac{\lambda^k * e^{-\lambda}}{k!}$ and the expected number of hits per unit interval is λ.

Since in most cases λ is unknown, there are two methods of estimation: the method of moments uses the empirical average number of hits per unit interval as an estimate, and the other one is maximum likelihood method. These two methods result in the same estimator in Poisson distribution.

Nonparametric Statistics - Nonparametric statistics is a subfield of statistics that does not only depend on parameterized groups of probability distributions. Nonparametric statistics depends on either distribution-free or having a specific distribution with parameters being unspecified. Nonparametric tests are used for studying a population with a ranked order. In addition, nonparametric tests rely on much fewer assumptions than parametric tests, so non-parametric tests are more robust. Nonparametric models differ from parametric models in the way that the model is not specified a priori but is instead determined by data.

WORKS CITED

Althoff, Tim and Leskovec, Jure. *Donor Retention in Online Crowdfunding Communities: A Case Study of Donors Choose.org. Acm Digital Library,* 2015, pp. 34-44. https://dl.acm.org/citation.cfm?id=2741120

"An Introduction To Crowdfunding", *Nesta*, https://www.em-a.eu/fileadmin/content/REALISE_IT_2/REALISE_IT_3/IntroToCrowdfunding.pdf

Jepson, Tina. "How To Actually Calculate Donor Retention (The Right Way) & 8 Essential Tips For Effective Donor Retention." *causevox*, 11 April, 2017, https://www.causevox.com/blog/measure-donor-retention/

The Ultimate Guide To Donor Retention. *FUNDLY*, https://blog.fundly.com/ultimate-guide-donor-retention/

State Population Totals and Components of Change: 2010-2017, https://www.census.gov/data/tables/2017/demo/popest/state-total.html#par_textimage