IRS Form 990 Data Project Pad

We are very excited to have you join us for the weekend to do good and drive social change with data. This is a fully-editable community wiki-hack-pad with general information that you might find useful to help you orient yourself.
The shortlink to this document is http://bit.ly/dd_form990

Overview

Form 990 is a required IRS filing for active nonprofits in the United States. It details the financial activities and income sources for a nonprofit. At the last DataDive, attendees reviewed a trove of form 990 filings to understand how different non profits solicited donations. They found that certain organizations rely on vastly different donation methodologies--an important insight with implications for #GivingTuesday. This effort will pick up where the previous one left off.
Specifically, this effort seeks to take on the following next steps:

- Helping nonprofits find mission partners
- Tagging organizations using SDG and other philanthropic giving categories
- Identify specific organizations for donors

More information can be found in the Project Brief

Objective(s)

Objective 1: Giving Story Profiler Tool

Objective 2: Increase Donations to Nonprofit Organizations #GivingTuesday

For additional objectives, and more detail, please refer to the Project Brief

Logistics

Slack (Communication): #datadive_0817_p3vols (please put your name and email under the "team members" heading below to receive an invitation to join the slack team)

Data:

- 990 Efile Operational Data 2009 - 2015 - Link to download file
- Data Dictionary - Link to download file
- Sample Program Service Data - Link to download file
- NTEE source: http://nccs.urban.org/classification/national-taxonomy-exempt-entities
- NTEE Classification Data - Link
- Sample Twitter Data - Link

Organizations with gross receipts normally < $50,000 must file Form 990-N (but may choose to file a complete Form 990 or Form 990-EZ). In prior years only organizations with gross receipts normally < $25,000 could file the Form 990-N ("e-postcard").
Organizations with gross receipts > $50,000 and < $200,000 and total assets < $500,000 must file Form 990-EZ or a complete Form 990.
Organizations with gross receipts > $200,000 or total assets > $500,000 must file Form 990.
Private foundations must file Form 990-PF.

GitLab (Code & Project Tracking):

- https://github.com/datakind/datadive-gates92y-proj3-form990.git

Please see your DA about gaining access to the GitHub repo

Project Brief (Project Overview):
https://docs.google.com/document/d/1JXZxXWZkM3YmXpgkBSSaQzyQgN-dQDC0k1ZBMYP2hpk/edit

Links

Tax Form 990 Wiki

Contacts

Project Champion

- Nathan Dietz, Urban Institute, ndietz@umd.edu

Data Ambassadors

- Elizabeth Walsh, eliz.walsh@gmail.com
- Miranda Tao, ttc1994@uw.edu

Data Expert

- Woodrow Rosenbaum, *92Y,* woodrowr@with-intent.com
- Jessica Schneider, *92Y,* woodrowr@with-intent.com

Team Members (name, email, GitHub handle)

- Miranda Tao, ttc1994@uw.edu, mirandattc
- Patricia Decker, pdecker@habitat.org,
- Kelly Domico, kdomico@gmail.com, ongk
- Brett Bejcek, bejcek.2@osu.edu
- Amy LaSota, amy.lasota@libertymutual.com
- Mark Conrad, marktconrad@gmail.com
- Jay Cheng , watchtheblur@gmail.com
- Francis Duplessis, fduplessis21@gmail.com
- Pricilla Chooy, workingpris@gmail.com
- Jenny Hung, jehung@me.com
- Carlton Bonney, bonneyc@gmail.com, bonneyc
- Dan Taber, dtaber47@icloud.com
- Matt Henricks, matthenricks@gmail.com, matthenricks
- Weijie Gao, gweijie01@gmail.com, balla121
- Yedong Wei, wei.yedong@gmail.com
- Rui Han, rui.han.rh@gmail.com
- Dave Langer, david_langer@hotmail.com, EasyD

Name	Email	GitHub Handle
Lara Haase	llhaase@gmail.com	

karthick chandrasekran	karthick12mar@gmail.com	
Ijaz	ijazahamed@gmail.com	
Adi	adithya.1111@gmail.com	
Iris Blackburn	iris.blackburn@gmail.com	justforwhimsy
Cynthia	CyntheSYS@gmail.com	CyntheSYS
Dan Taber	dtaber47@icloud.com	dantaber

Results

(Make sure to capture all of your work!)

Data Cleaning Recommendations from Financial Team (Rui, Yedong, Iris, Andreas, Amy)

a. Data cleaning for volunteer & employee counts
   i. There were a few rows with negative total salary
   ii. The 99th percentile of Salaries / Total Employees was $175k
b. About 8000 true duplicates in the full set should be dropped
c. Assets, total revenue, sub-categories of revenue, expenses, and sub-categories of expenses are never supposed to be negative
d. Calculation of Total Revenue

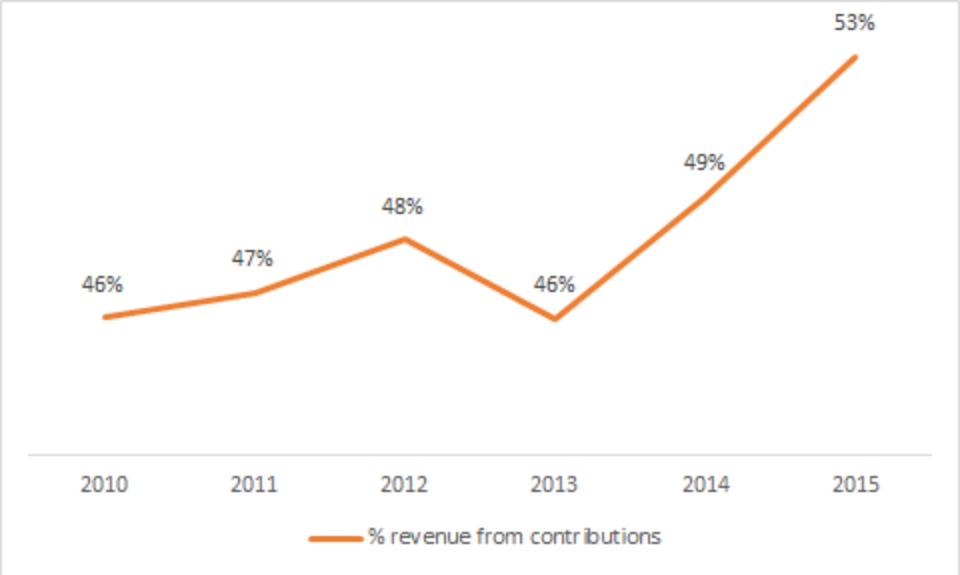
   For 20% of the data (only EZ forms), the Total Revenue field doesn't equal the sum of Revenue from Current, Program service, Investment, and Other Revenue. We think this is because of the revenue fields #3-7 on the EZ Form: Net Sales from Inventory & Non-Inventory Assets, and Net Sales from Gaming. These all should be included in "Other" Revenue in the full 990 form. We are not able to add these variables to verify calculation of Total Revenue because they are missing from the "Sample" dataset.
e. Exclusion of certain NTEE Categories that may skew financial results due to asset size
   i. Initial recommendations: B (Education), E (Health Care)

- By excluding NTEE Categories “B”, which includes large university endowment funds and “E”, which includes hospitals, the largest average asset size (5th quartile) drops by almost $45 billion

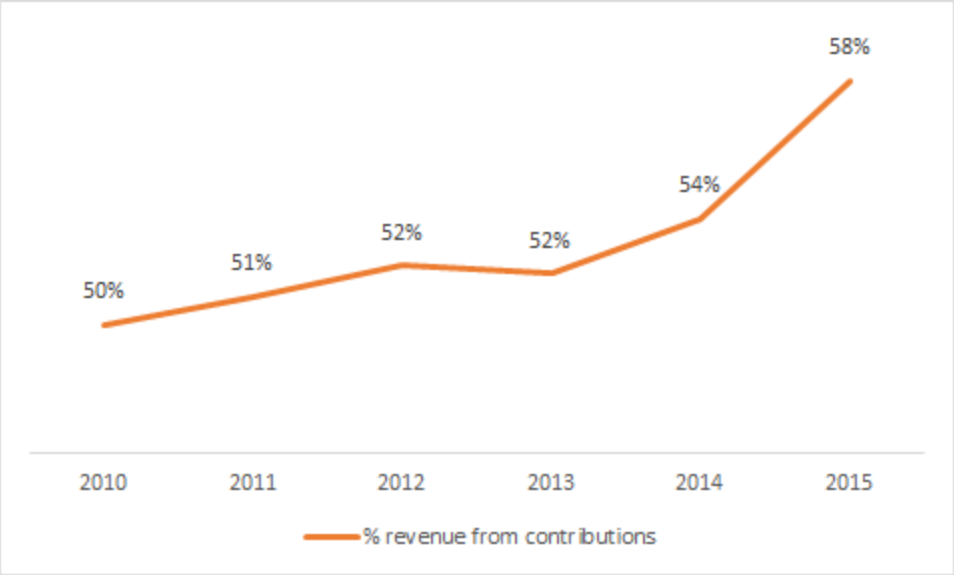
Size Quartile (Asset)	Including NTEE Categories “B” & “E”	Excluding NTEE Categories “B” & “E”
1	$99,700	$93,650
2	$311,022	$278,353
3	$1,271,054	$987,798
4	**$54,000,000,000**	**$9,480,000,000**

We included Member Dues within Current Contributions for Form EZ, which allowed us to tie out the calculation of Total Revenue for 6% of rows. We think the additional fields will make a big difference in the %. Compare the following 2 charts

<u>% of Revenue from Contributions over time, comparing inclusion of Membership Dues</u>



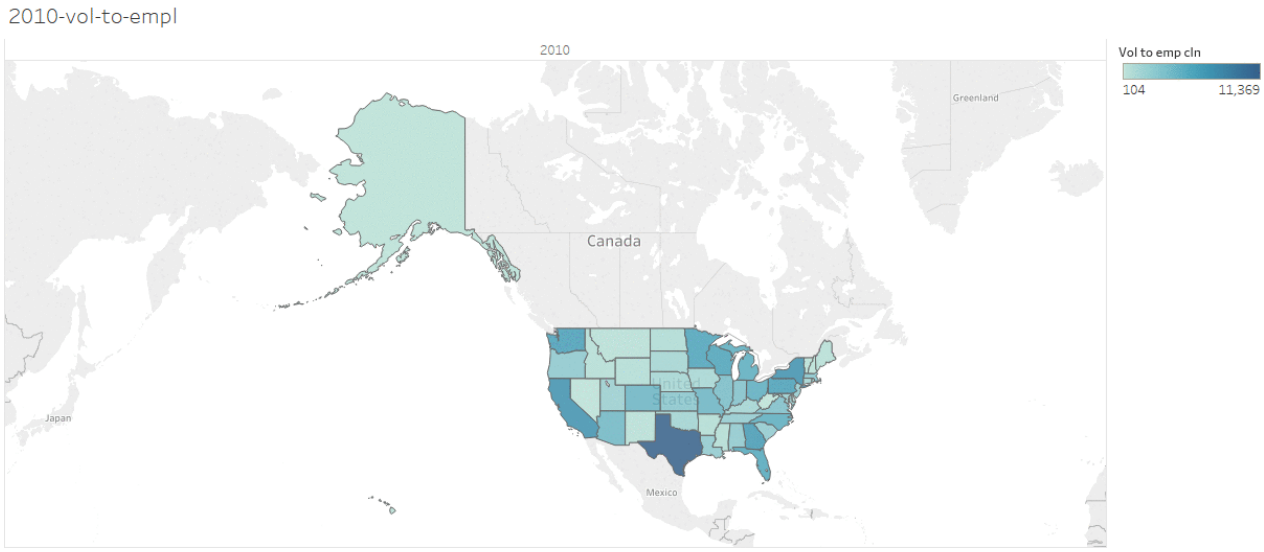
% of Revenue from Contributions over time, excluding Membership Dues



% of Revenue from Contributions over time, including Membership Dues

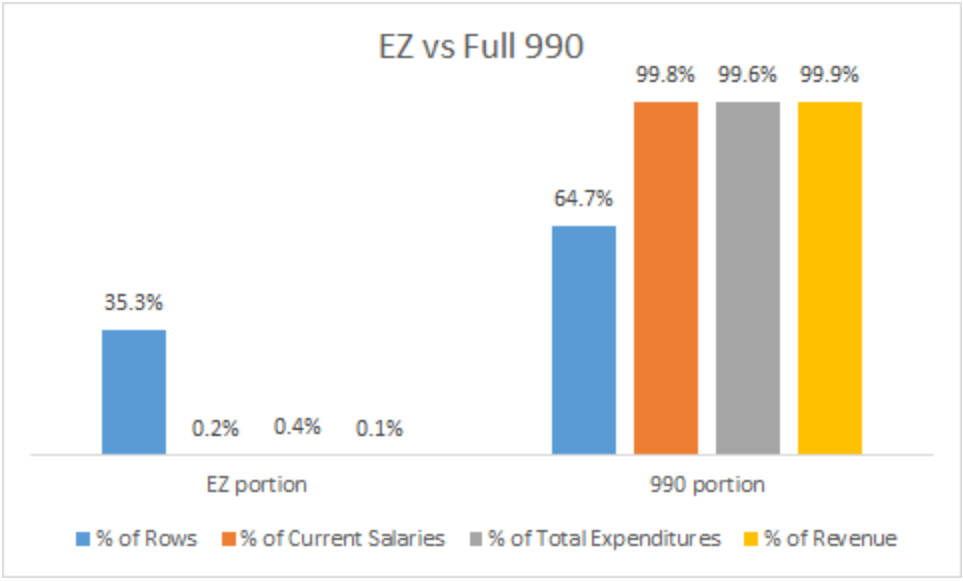
Objective 1: Exploratory Data Analysis (EDA) of Revenue Metrics
Members: Amy, Iris, Andreas, Jenny H, PJ D, Yedong Wei, Rui Han
Data source: Financial Data Sub-Set 2010 to 2015.csv

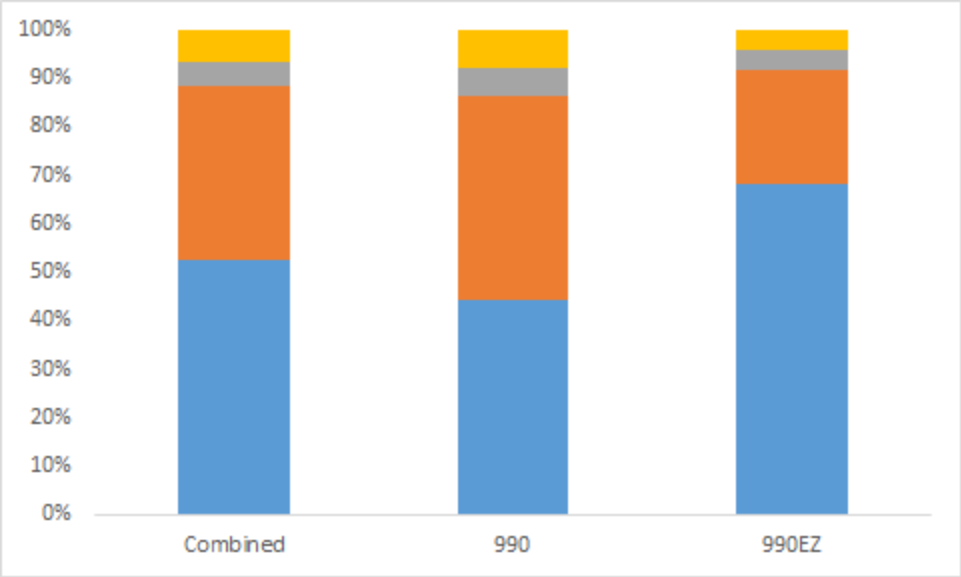
Volunteer to Employee ratios year-over-year
The following images are the Volunteer to Employee ratios for each state. We cut out the 99th percentile over volunteers and 99.9th percentile for employees to remove fat finger mistakes.
We did not get a chance to dig deeper into what caused changes year over year.
Data can be found in Tableau.



Revenue Source Distribution:

Most of our analysis focuses only on the 990 form because the 990 EZ form contains less information, and most of the revenue and expenditures come from the full 990.

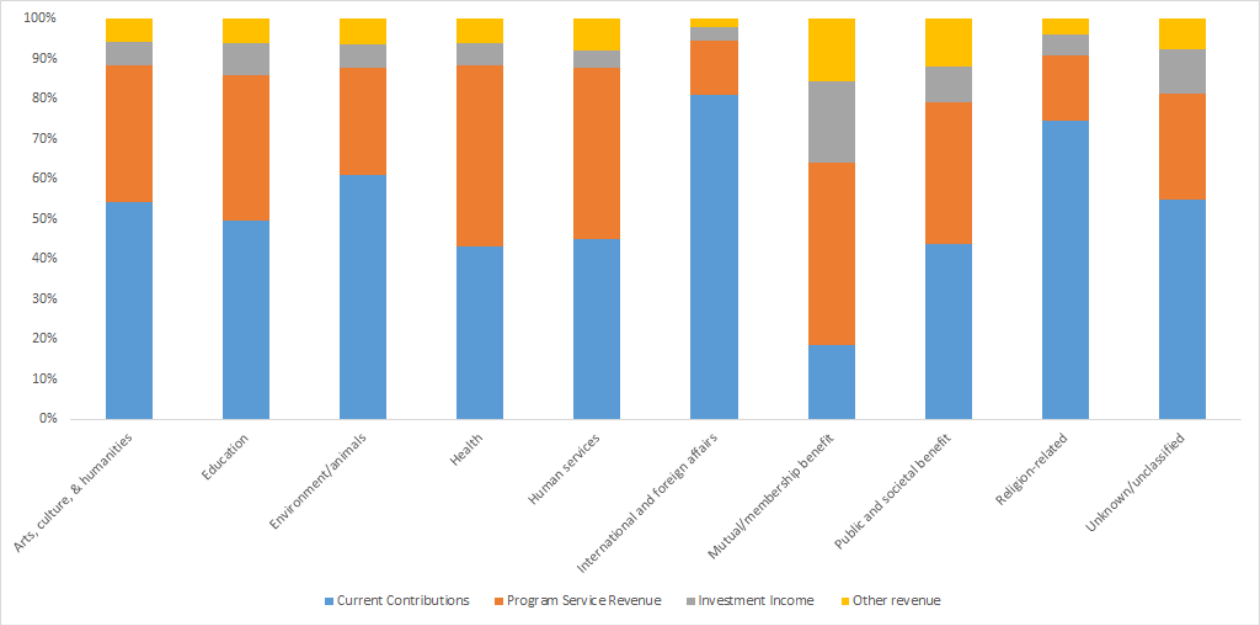




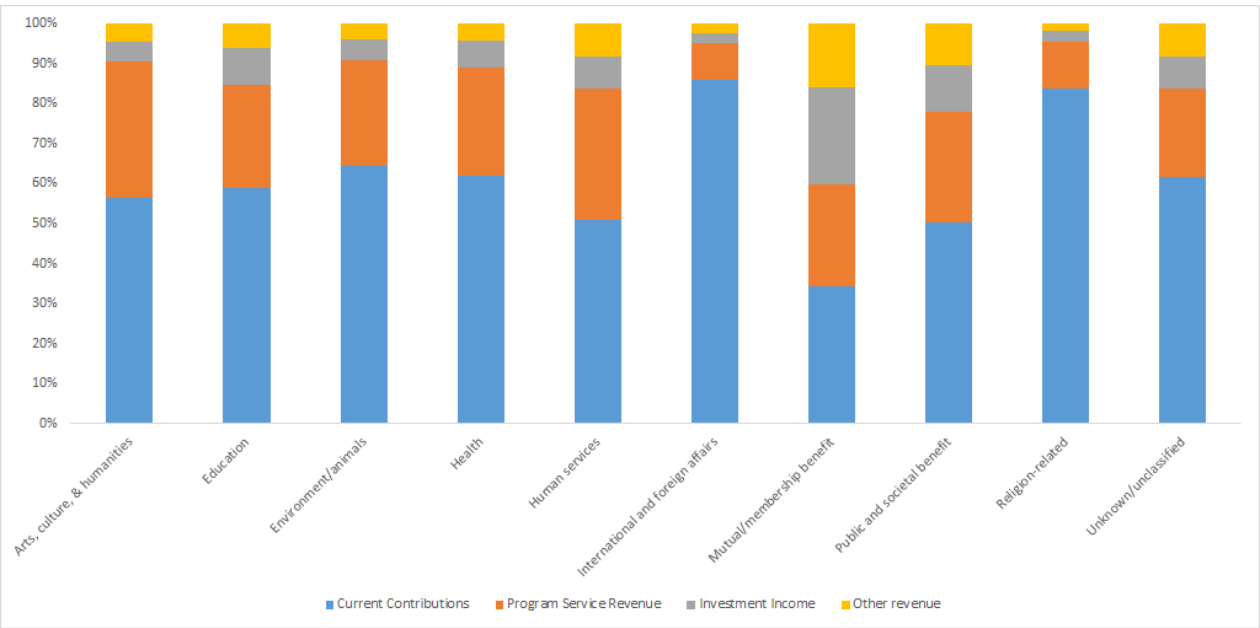
Revenue Source Distribution

Revenue Source	Overall	990	990EZ
Current Contributions	53%	44%	68%
Program Service Revenue	36%	42%	24%
Investment Income	5%	6%	4%
Other revenue	7%	8%	4%

% of Revenue Source using NTEE code classification



% of Revenue Source using NTEE code classification, all forms combined



% of Revenue Source using NTEE code classification, EZ form only

Gross Revenue Bins

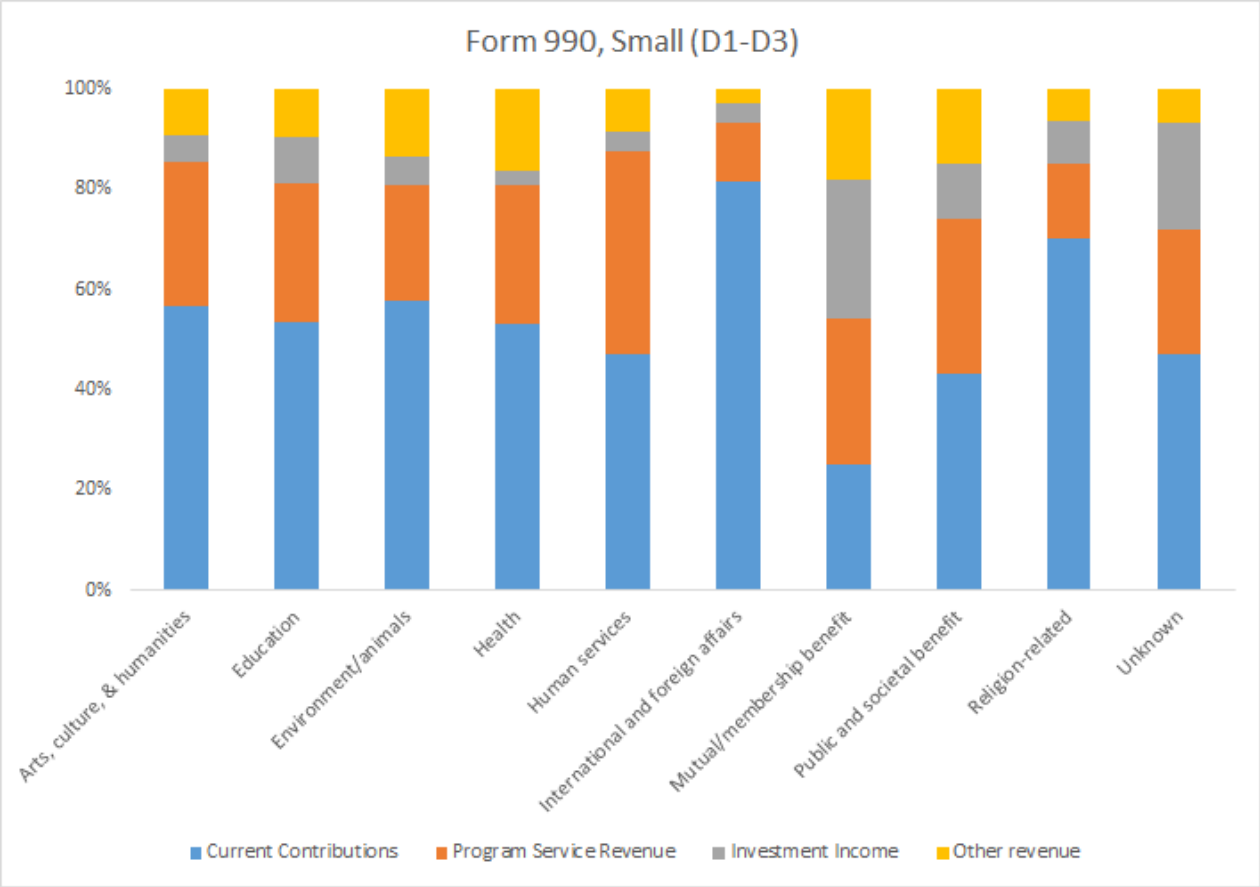
Methodology:

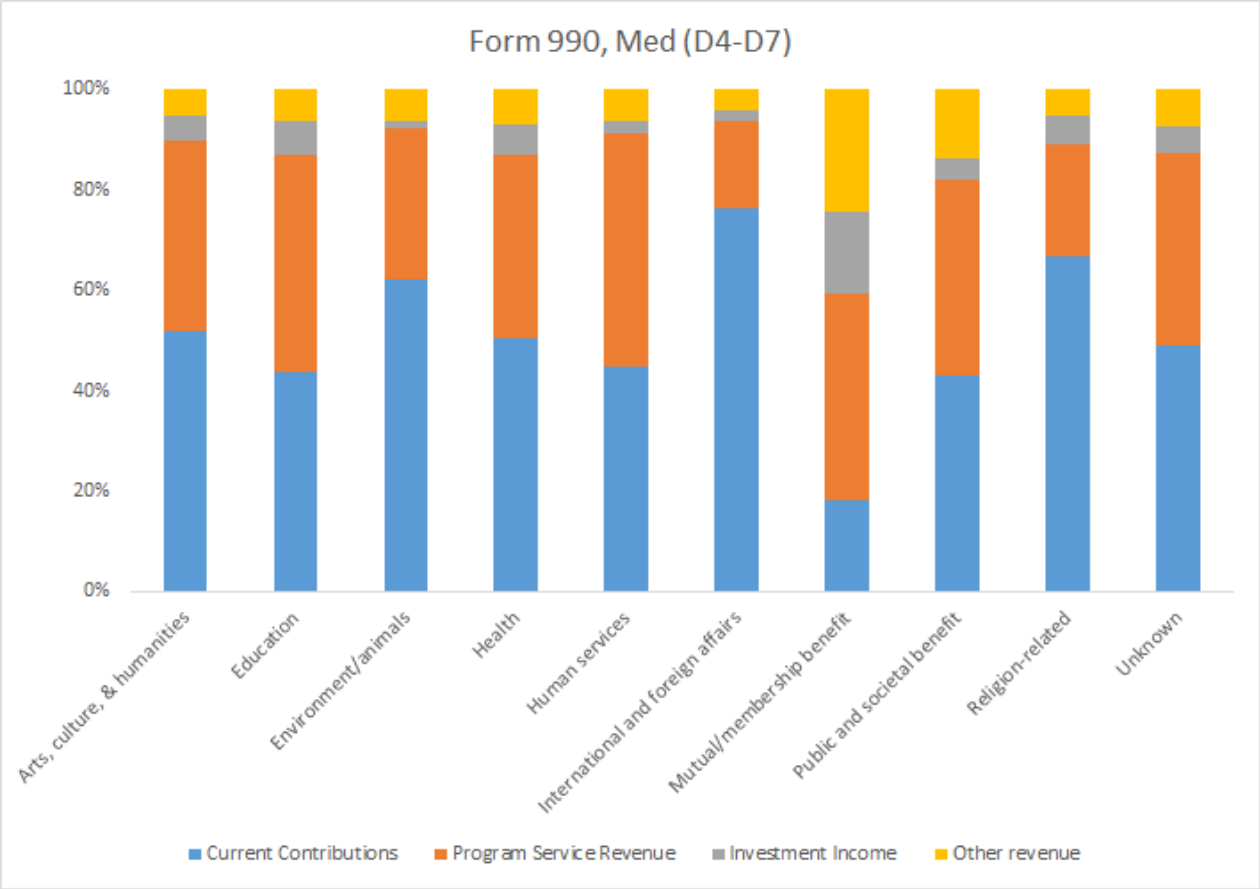
a. Only look at Regular 990 forms. Exclude 990EZ. The 990EZ is only for GROSS RECEIPTS less than $250k.
b. For Regular 990, split into deciles by Gross Receipts.
    i. Small =$0-$275k (Deciles 1,2,3)

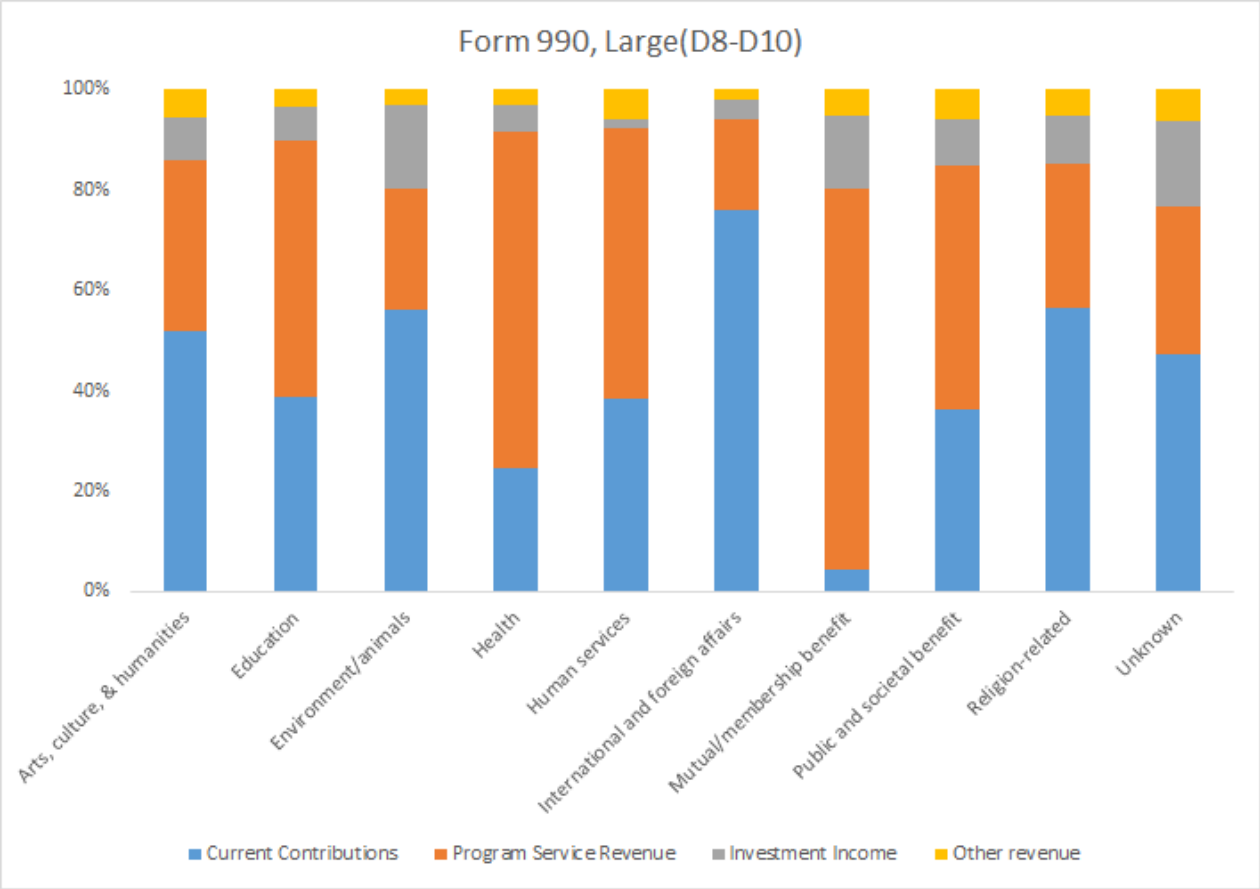
ii. Medium =$275k to $1.6M (Deciles 4,5,6,7)
iii. Large = $1.6M to $86B (Deciles 8, 9, 10)

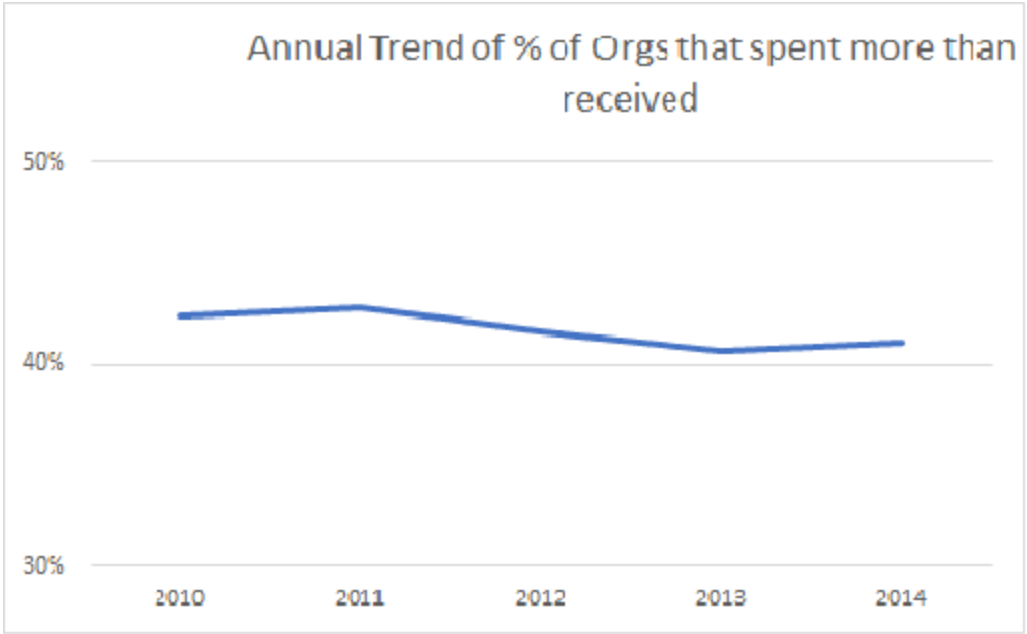
c. Some NTEEs, like Mutual/Membership Benefits, have wide variance in business model by revenue size. Others, like arts, culture, and humanities, are fairly similar across size buckets.
d. We included entities with Gross Receipts < $250k in the Small Bucket, even though they are not required to file 990 (they are eligible to file 990 EZ). We think these entities are self-selecting to file the 990 because their gross receipts may fluctuate over and under the threshold.

Group Size	Overall	Small	Medium	Large
Current Contributions	44%	50%	47%	36%
Program Service Revenue	42%	32%	40%	54%
Investment Income	6%	7%	4%	6%
Other revenue	8%	11%	8%	5%


Form 990, Small (D1-D3)
100%
80%
60%
40%
20%
0%
Arts, culture, & humanities
Education
Environment/animals
Health
Human services
International and foreign affairs
Mutual/membership benefit
Public and societal benefit
Religion-related
Unknown
Current Contributions
Program Service Revenue
Investment Income
Other revenue


Form 990, Med (D4-D7)
100%
80%
60%
40%
20%
0%
Arts, culture, & humanities
Education
Environment/animals
Health
Human services
International and foreign affairs
Mutual/membership benefit
Public and societal benefit
Religion-related
Unknown
Current Contributions
Program Service Revenue
Investment Income
Other revenue


Form 990, Large(D8-D10)
100%
80%
60%
40%
20%
0%
Arts, culture, & humanities
Education
Environment/animals
Health
Human services
International and foreign affairs
Mutual/membership benefit
Public and societal benefit
Religion-related
Unknown
Current Contributions
Program Service Revenue
Investment Income
Other revenue

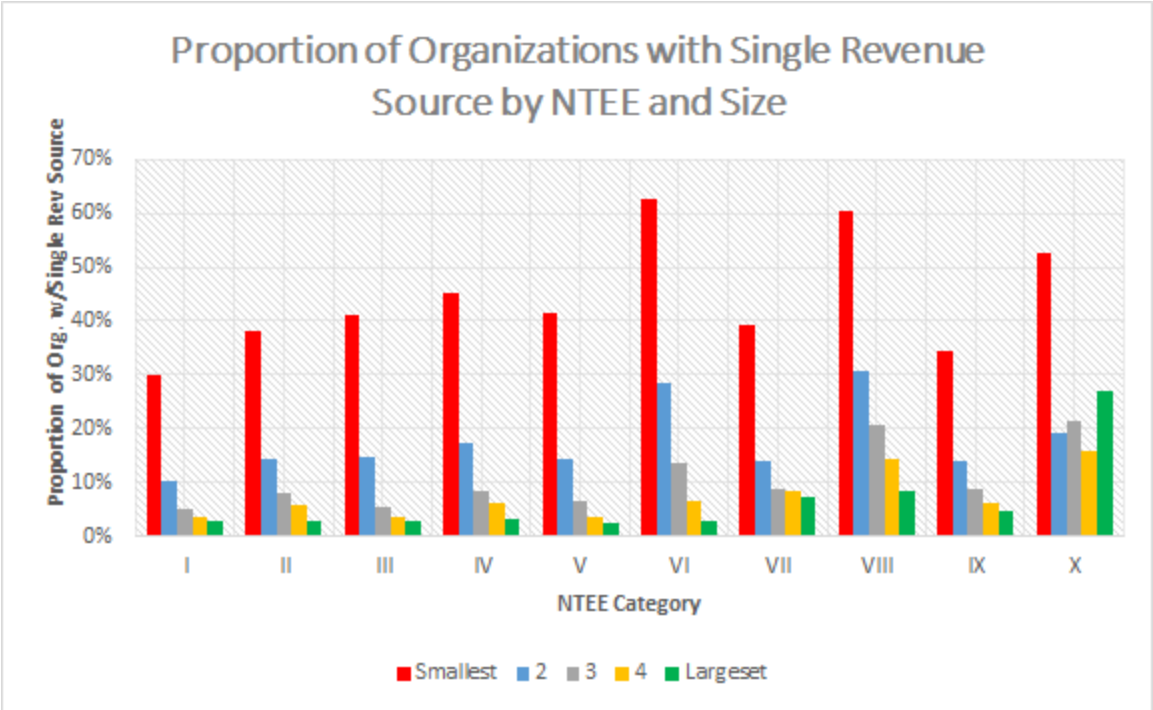

Annual Trend of % of Orgs that spent more than received
50%
40%
30%
2010
2011
2012
2013
2014

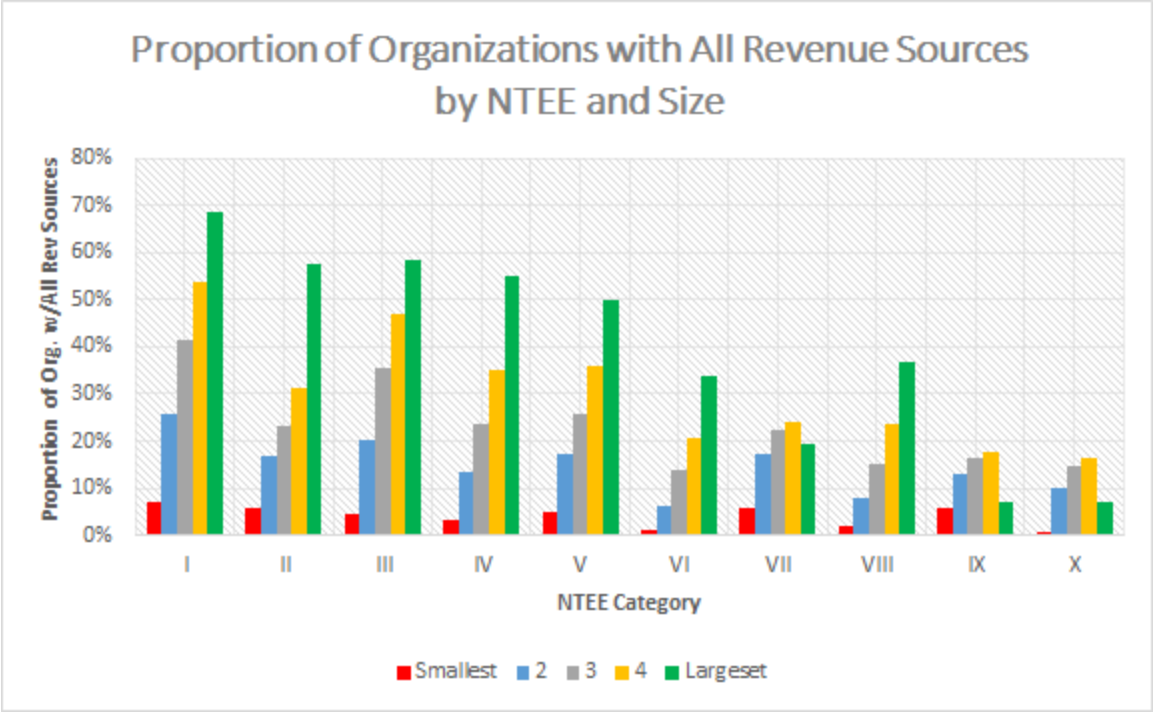
This graph displays the % of organizations with - Net income over the time periods 2010 and 2014. Surprisingly this graph shows that consistently the orgs range between ~40% per year.

Revenue Source Concentration

Form990 list four potential revenue sources. Revenue source concentration may be a cause for concern if an organization relies solely on a single funding source. If there is a shock to the system, organizations that has only one funding source may lose a significant portion of their revenue stream and thus must cut their program expenses or may not be able to fund them at all.

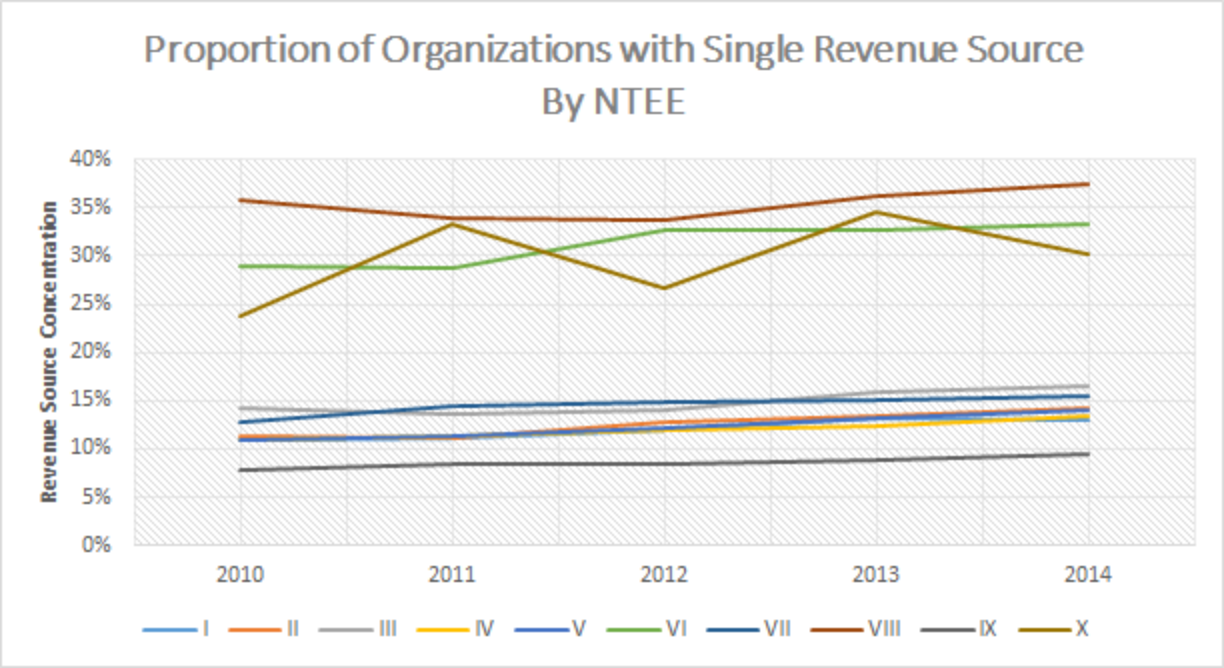
- Organizations are divided into quintiles by total assets (size) for each fiscal year
- Proportion of organizations with only one revenue source is significantly higher for the smallest size quintile compared to other quintiles
- Proportion of organizations with revenue from all four sources is significantly higher for the largest size quintile compared to other quintiles
- This pattern shows up across most if not all 10 NTEE categories
- Takeaway: the lack of certain organization's ability to tap into multiple revenue sources may be constrained by the amount of resources available (which is captured by size here)





The graph below plots a time series of the proportion of organizations with just one funding source and breaks it out by NTEE (fiscal year 2015 is excluded because it only has about 1/3 number of observations as other years). Three of the ten NTEE categories (VI International, Foreign Affairs; VIII Religion Related; X Unknown) stand out because the proportions of organizations with only one revenue source is significantly higher than that of the other NTEE categories.

The cross-sectional difference across NTEE categories are persistent over time. This could mean that the differences in business model across NTEE categories are due to fundamental constraints.



The Ratio of Volunteers to Employees

Using the sample set, the ratio of volunteers to employees had a significant rise in 2012. This could be due to factors such as the great recession and individuals wanting to be productive, natural disasters and recovery efforts, etc. More detail highlights states with higher than average ratios.

Row Labels	Sum of TOTEMPLOYEE	Sum of TOTVOLUNTEERS	Sum of ratio vol/emplee
⊞2010	3833642	15250748	4.0
⊞2011	5112408	15810003	3.1
⊞2012	4338225	50331220	11.6
⊞2013	4322390	21523946	5.0
⊞2014	4228719	14862625	3.5

Creating Indices

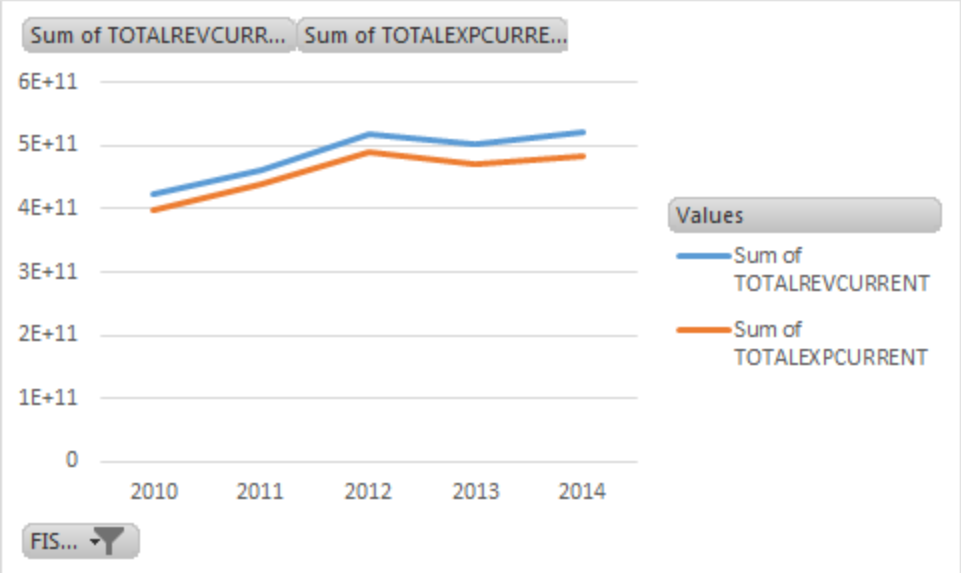
This Python model reflects an index comparison for each entity over comparable periods for the full data set. The model can assist in entity selection based upon entity self-reported financial information.

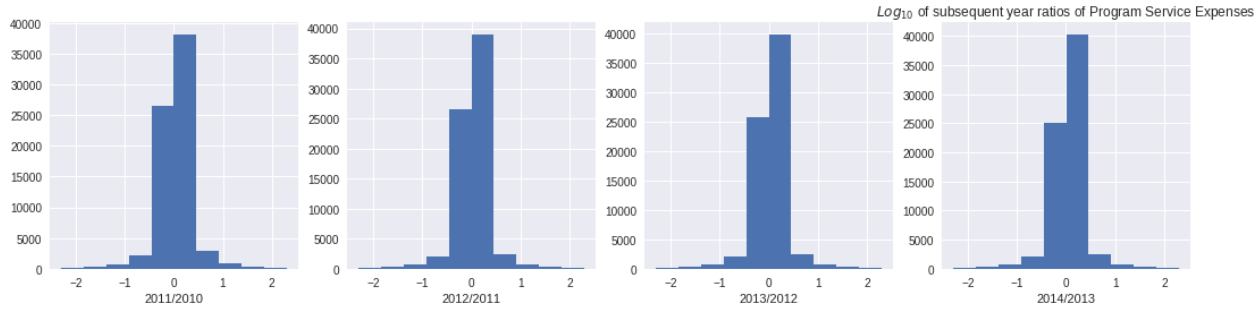
The model uses fiscal year 2010 as the basis with all entities starting with 1.0 as the index. This table includes the EIN, fiscal year, the Net Income Total (NIT) and the full data pool as Broad Spectrum (BR). The Net Income Total (NIT) is calculated on the IRS Form 990 as Total Revenue (line 12) minus Total Expenses (line 18) as reported on line 19 (Revenue less expenses). On the IRS Form 990EZ it is calculated as Total Revenue (line 9) minus Total Expenses (line 17) as reported on line 18 (Excess or deficit for the year).

The core table is supplemented with Broad Spectrum (BR) analysis is by State; including the EIN and fiscal year. Other attributes can be substituted in the variable field “State”. Suggested variables may be NTEE, employees, volunteers, etc.

The graph shows all entities at the same index of 1.0 for 2010. In 2011 there was a big dip and then a recovery and less volatility in recent years. The dip is attributed to the Great Recession.

The next analysis determines how entities recovered their Net Income Total (NIT) level after 2011 – either by increasing revenue or decreasing expenses. This incorporates the blend of the past five years to level the revenue and expense patterns of each entity.

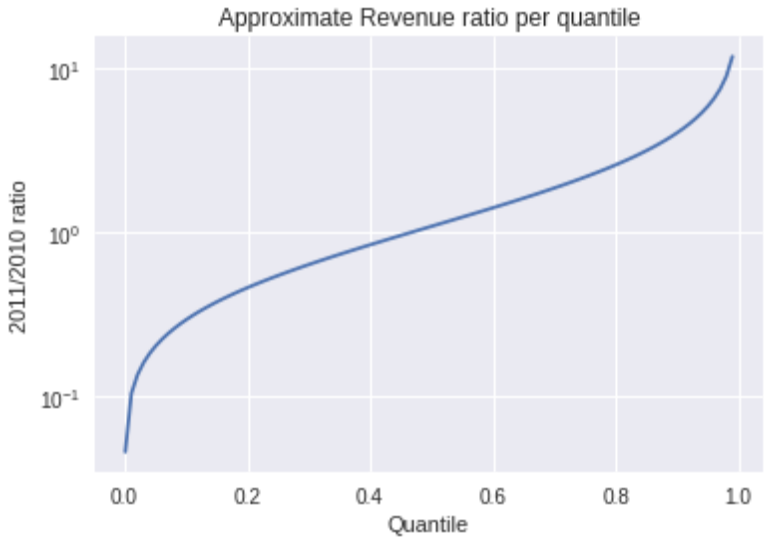




Log of subsequent years ratios of Program Service Expenses

With this dataset, we built a model by looking at the mean ratio and its standard deviation throughout the years. Plotting these shows that the log of their value looks fairly Gaussian. Hence we will assume the log(ratio) to be a Gaussian variable with mean and standard deviation determined from this sample.

Now we run the query function over the subset of data we've been playing and use its quantile output to make a new variable in the dataset.



Objective 2: Twitter Analysis

Success:

- We have a list of ~6,500 agency names, EIN, domain name, and twitter handle
- We have automated processes that have successfully automated the discovery of approximately 1300 Twitter usernames.

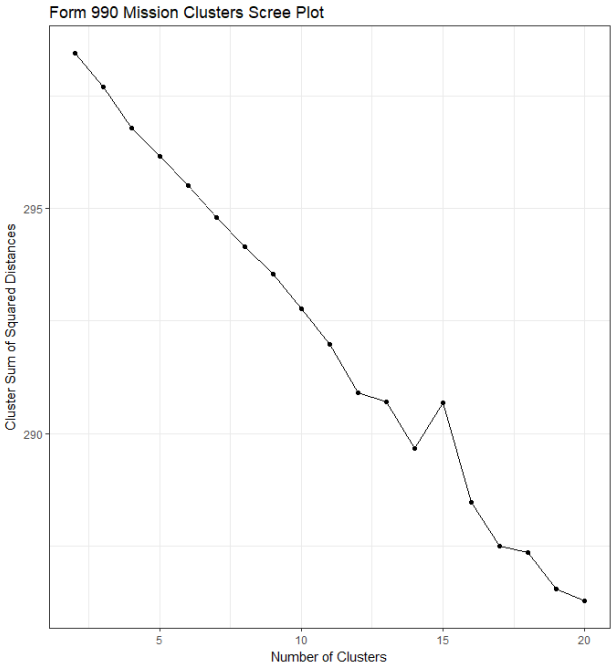
- We have associated the 1300 Twitter usernames that we have with Giving Tuesday tweets from 11/14/2016.
- We performed clustering of the Form 990 Mission, Project 1, Project 2, and Project 3 descriptions fields using NLP techniques.
- In the spirit "fail fast, learn fast" we POC'ed the use of Twitter's advanced search page as a possible method of scraping Twitter user names for Form 990 organizations. Our efforts showed this is not a viable approach.
- We created a python beautifulsoup web scraping script to retrieve link information within non-profit listed websites in the 990. It grabs social connections, including twitter handles, to eventually target additional analysis/scraping. Still has issues in cleansing and can be better scaled/batched.

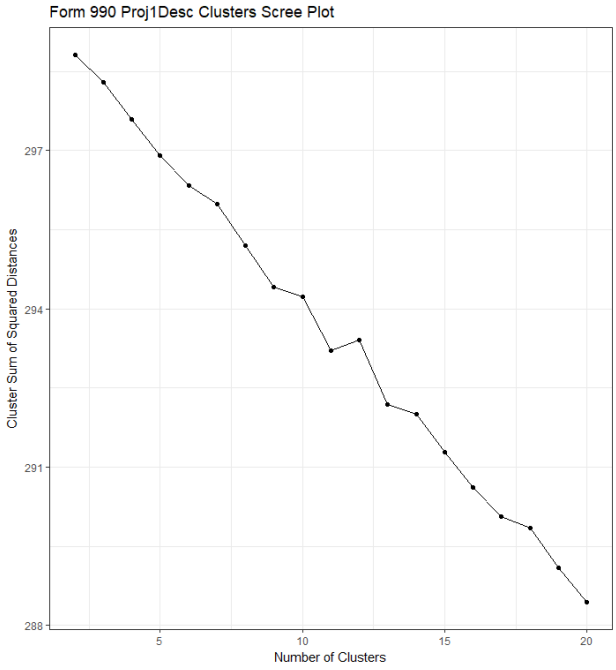
EIN	Name	TwitterHandle	Social Links
237172077	100 CLUB OF ARIZONA	100ClubAZ	
366158087	The Hundred Club of Cook County	Chicago100Club	https://www.facebook.com/100clubofchicago https://www.facebook.com/100clubofchicago https://twitter.com/Chicago100Club https://twitter.com/Chicago100Club https://www.linkedin.com/company/the-100-club-of-ch https://www.linkedin.com/company/the-100-club-of-ch https://www.instagram.com/100clubchicago/
455195419	100PLUSANIMALRESCUE INC	100plusrescue	https://www.facebook.com/ABANDONEDDOGSEVERC https://twitter.com/100plusrescue http://instagram.com/100plusevergladesrescue
900702671	100REPORTERS	100Reporters	https://facebook.com/100Reporters https://twitter.com/100Reporters

- We are building a database of organizations' history of using #GivingTuesday on Twitter every year from 2012 to 2016

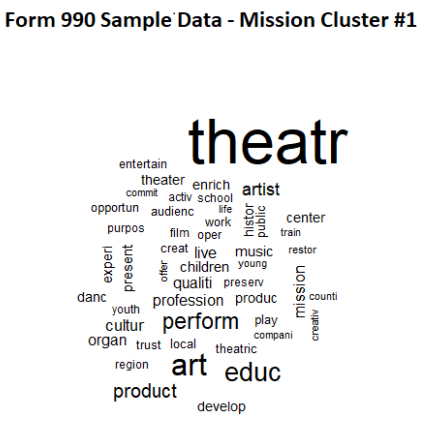
	user	text	year	2012	2013	2014	2015	2016
0	99balloonsorg	On #GivingTuesday GIVE REST! rEcess provides ...	2016	0	0	0	0	1
1	99balloonsorg	Help us change the story of disability with a ...	2014	0	0	1	0	0
2	6thSP	For #GivingTuesday we're offering $20 tickets ...	2016	0	0	0	0	1
3	AAA1C	Celebrate #GivingTuesday with The Senior Allia...	2016	0	0	0	0	1
4	AAA1C	If you are looking for a worthy cause to donat...	2015	0	0	0	1	0
5	AAA1C	You can be a part of #GivingTuesday by donatin...	2014	0	0	1	0	0
6	AAA1C	http://t.co/BPvepWIXLu\n^Purchasing a holiday ...	2014	0	0	1	0	0
7	AAA1C	RT @GivingTues: Thank you for celebrating #Giv...	2013	0	1	0	0	0
8	AAA1C	#GivingTuesday is tomorrow.The Holiday Card Pr...	2013	0	1	0	0	0
9	AAA1C	#GivingTuesday is a new day for giving back. T...	2013	0	1	0	0	0
10	AASummerFest	https://t.co/xhrSwD88tN #GivingTuesday If you ...	2016	0	0	0	0	1
11	AASummerFest	❤ Help Us Make Magic Happen on #GivingTuesday ...	2016	0	0	0	0	1
12	AASummerFest	https://t.co/TrbA81UIxq Wishing you a happy #G...	2015	0	0	0	1	0
13	AASummerFest	http://t.co/vG4uMtzaAD Thanks to everyone who ...	2014	0	0	1	0	0
14	AASummerFest	http://t.co/zjytVFYCgT Sluggo likes #unselfie ...	2014	0	0	1	0	0
15	AASummerFest	Happy #GivingTuesday http://t.co/vG4uMtzaAD Gi...	2014	0	0	1	0	0
16	AASummerFest	http://t.co/TM4pNGgJgt Join A2SF one week from...	2014	0	0	1	0	0
17	100Reporters	Rubies have blood, too. Learn more @100Reporte...	2016	0	0	0	0	1
18	100Reporters	@100Reporters revealed Kimberley offs. launder...	2016	0	0	0	0	1
19	100Reporters	RT @aschweig: .@100Reporters joins professiona...	2014	0	0	1	0	0
20	100Reporters	Just 80 minutes to go: Invest in accountabilit...	2013	0	1	0	0	0

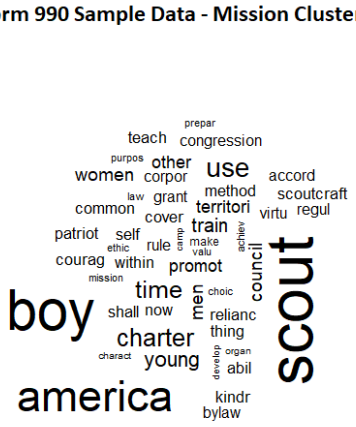
- We performed a cluster analysis of the Mission vs. Proj1Desc texts in the Form 990 sample dataset.
  - The data was preprocessed using the following pipeline:
    - Tokenization
    - Stop word removal using a custom stop word list.
    - Stemming
    - The unigram document-term matrices were scored using TF-IDF
    - Latent semantic analysis (LSA) was applied to each matrix.
  - K-means was selected as the clustering mechanism to allow the use of the "elbow" method for determining an interesting number of clusters to use for analysis. The following images illustrate the use of 12 clusters for the Mission texts and 9 clusters for the Proj1Desc texts:

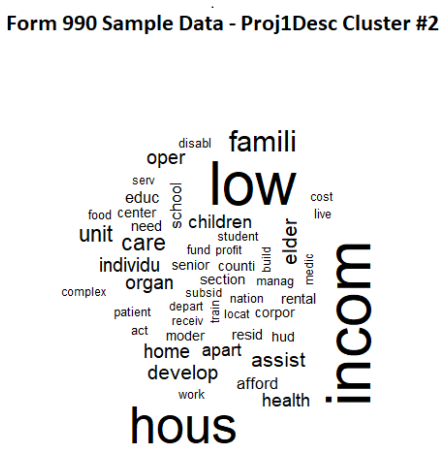


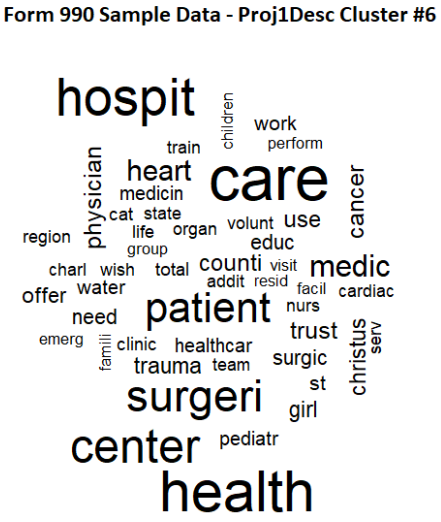


- The cluster assignments were then used to create word cloud visualizations to understand the potential themes/topics of each cluster. The following are two examples of Mission and two examples of Proj2Desc wordclouds:

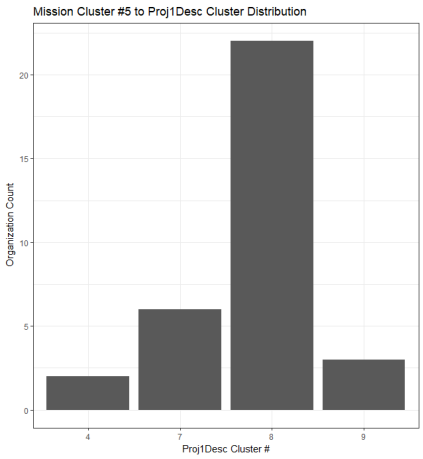


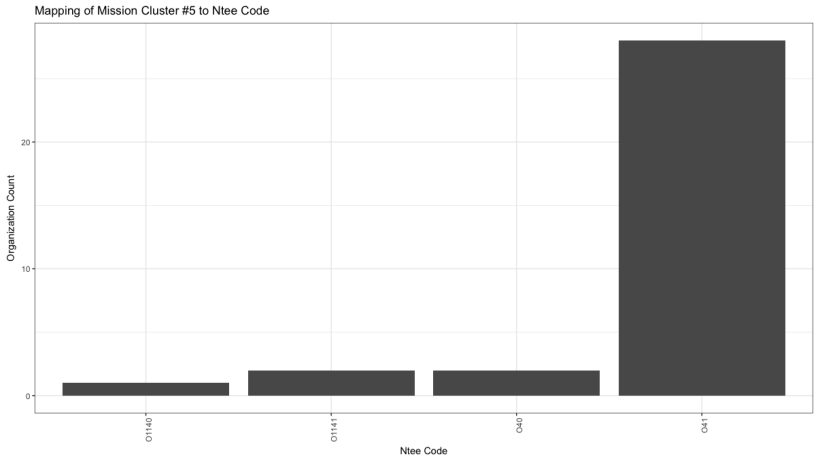






- With the above data, mappings between various data (e.g., Proj1Desc clusters by Mission Clusters, Ntee code distributions by Mission clusters are posible). The following are two examples:

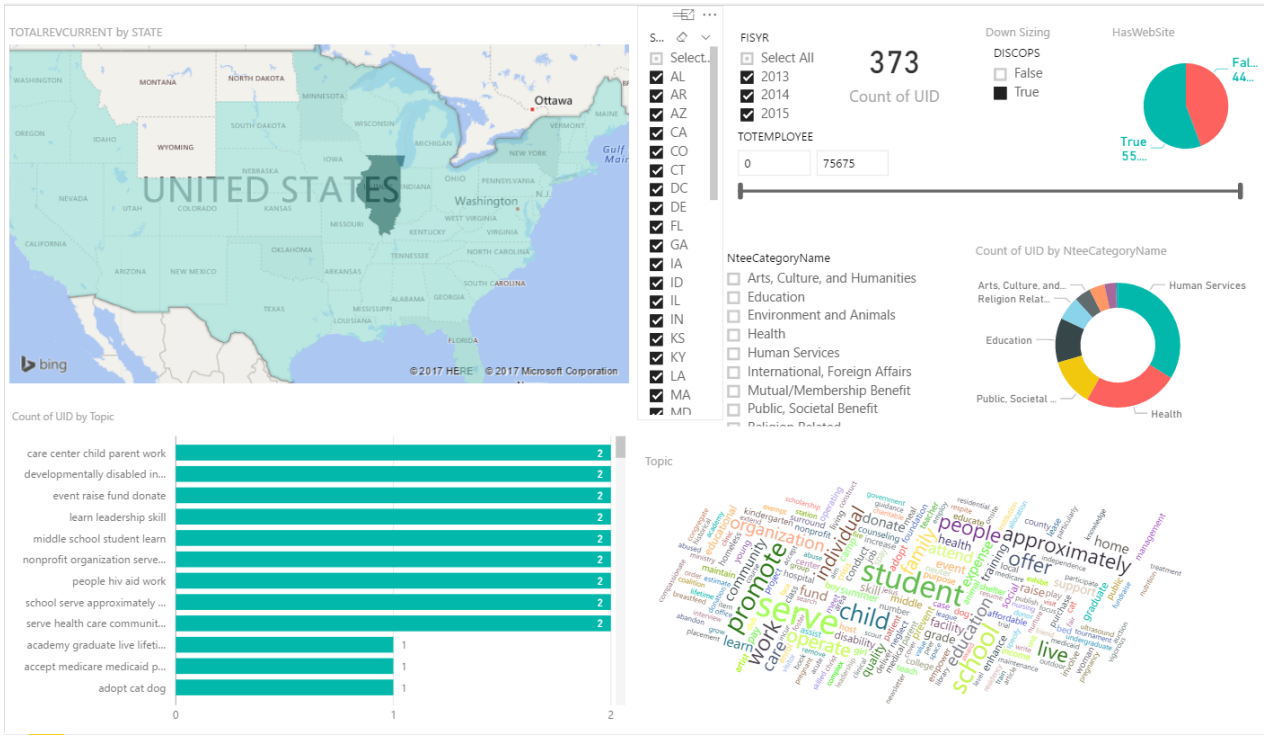


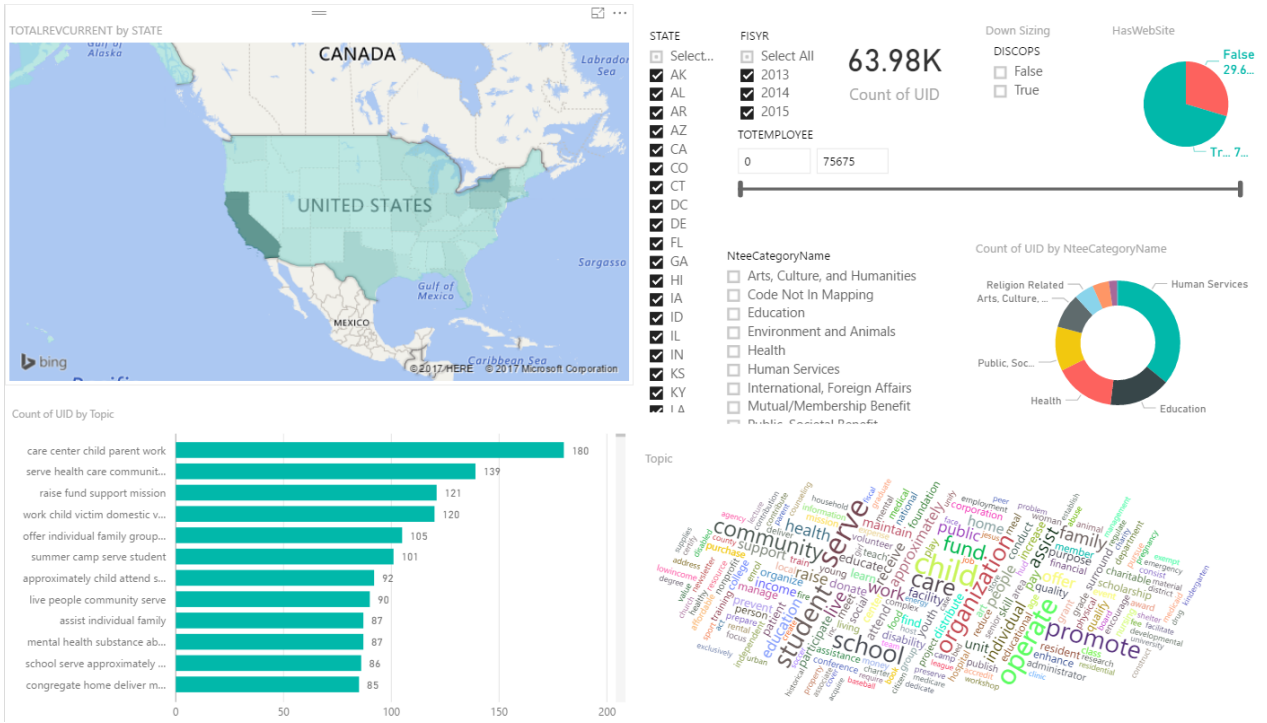


In Progress:

- We are building a Power BI dashboard to provide visual insights into the clustering results as well as promote further exploratory analysis. Current progress is below - integration with the updated clustering output is necessary.
- We hope to have 16,000 additional Twitter usernames by EOD tomorrow.
- We are in the process of using Twitter usernames to look up historical Giving Tuesday tweets prior to November 2016.

- Created a web script to pull all social accounts to link EIN & web link from 990 to social presence for further data analysis and marketing of #givingtuesday.

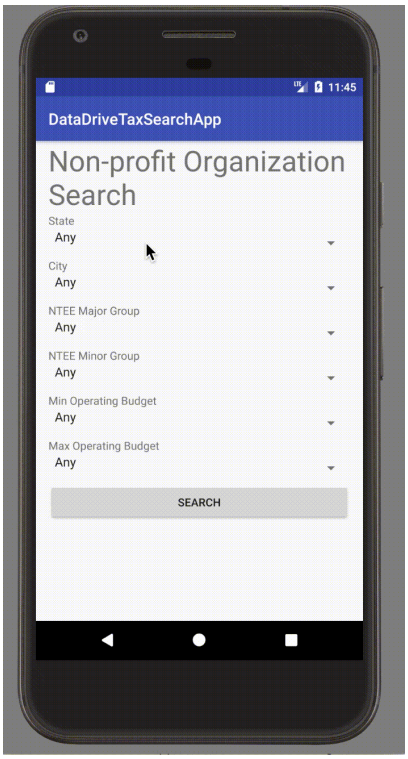




Objective 3: App for Individuals

Initial Android App with search functionality.

User can filter organizations based on similar fields(city, state, NTEE main group and subgroup, and operating budget) as the dashboard.

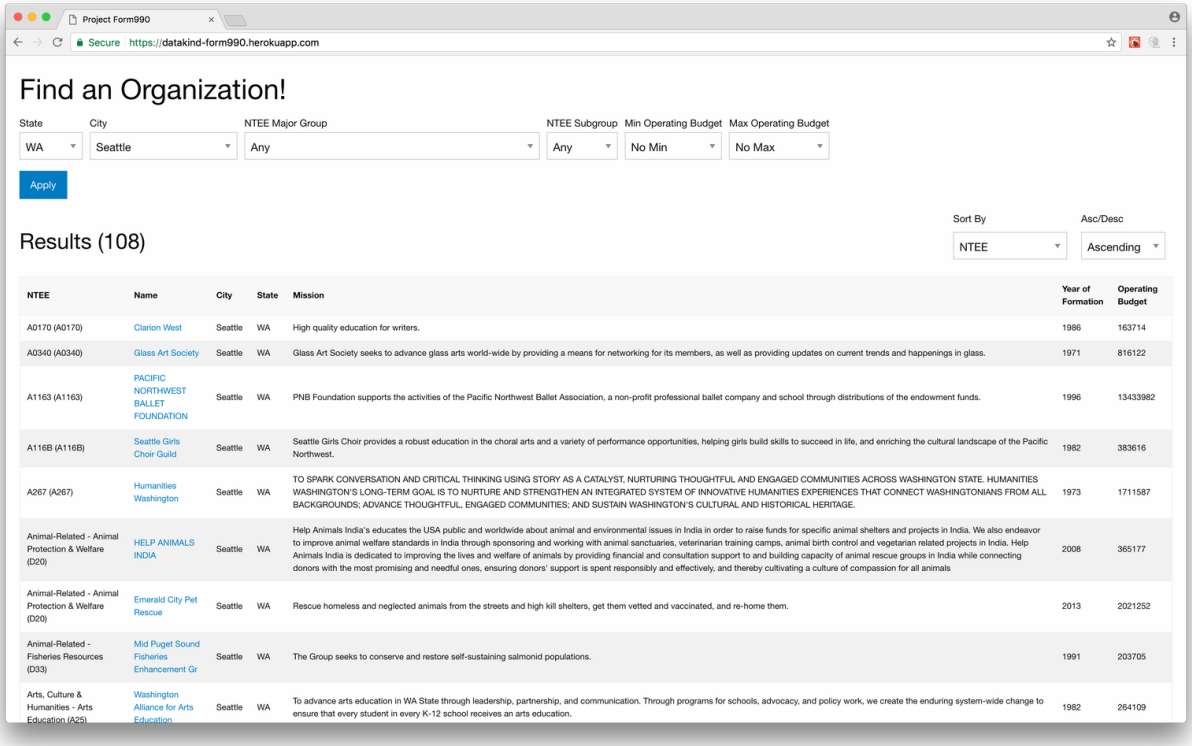


Android app demo

Web Version of Dashboard (Kelly)

User can filter by city, state, NTEE main group and subgroup, and operating budget (a determinant for the size of the organization). Table can be sorted by any of the columns displayed.

Temporary Link: https://datakind-form990.herokuapp.com



NTEE	Name	City	State	Mission	Year of Formation	Operating Budget
A0170 (A0170)	Clarion West	Seattle	WA	High quality education for writers.	1986	163714
A0340 (A0340)	Glass Art Society	Seattle	WA	Glass Art Society seeks to advance glass arts world-wide by providing a means for networking for its members, as well as providing updates on current trends and happenings in glass.	1971	816122
A1163 (A1163)	PACIFIC NORTHWEST BALLET FOUNDATION	Seattle	WA	PNB Foundation supports the activities of the Pacific Northwest Ballet Association, a non-profit professional ballet company and school through distributions of the endowment funds.	1996	13433982
A116B (A116B)	Seattle Girls Choir Guild	Seattle	WA	Seattle Girls Choir provides a robust education in the choral arts and a variety of performance opportunities, helping girls build skills to succeed in life, and enriching the cultural landscape of the Pacific Northwest.	1982	383616
A267 (A267)	Humanities Washington	Seattle	WA	TO SPARK CONVERSATION AND CRITICAL THINKING USING STORY AS A CATALYST, NURTURING THOUGHTFUL AND ENGAGED COMMUNITIES ACROSS WASHINGTON STATE. HUMANITIES WASHINGTON'S LONG-TERM GOAL IS TO NURTURE AND STRENGTHEN AN INTEGRATED SYSTEM OF INNOVATIVE HUMANITIES EXPERIENCES THAT CONNECT WASHINGTONIANS FROM ALL BACKGROUNDS; ADVANCE THOUGHTFUL, ENGAGED COMMUNITIES; AND SUSTAIN WASHINGTON'S CULTURAL AND HISTORICAL HERITAGE.	1973	1711587
Animal-Related - Animal Protection & Welfare (D20)	HELP ANIMALS INDIA	Seattle	WA	Help Animals India's educates the USA public and worldwide about animal and environmental issues in India in order to raise funds for specific animal shelters and projects in India. We also endeavor to improve animal welfare standards in India through sponsoring and working with animal sanctuaries, veterinarian training camps, animal birth control and vegetarian related projects in India. Help Animals India is dedicated to improving the lives and welfare of animals by providing financial and consultation support to and building capacity of animal rescue groups in India while connecting donors with the most promising and needful ones, ensuring donors' support is spent responsibly and effectively, and thereby cultivating a culture of compassion for all animals	2008	365177
Animal-Related - Animal Protection & Welfare (D20)	Emerald City Pet Rescue	Seattle	WA	Rescue homeless and neglected animals from the streets and high kill shelters, get them vetted and vaccinated, and re-home them.	2013	2021252
Animal-Related - Fisheries Resources (D33)	Mid Puget Sound Fisheries Enhancement Gr	Seattle	WA	The Group seeks to conserve and restore self-sustaining salmonid populations.	1991	203705
Arts, Culture & Humanities - Arts Education (A25)	Washington Alliance for Arts Education	Seattle	WA	To advance arts education in WA State through leadership, partnership, and communication. Through programs for schools, advocacy, and policy work, we create the enduring system-wide change to ensure that every student in every K-12 school receives an arts education.	1982	264109