

# **Project Report**

Group 14

**Submitted Date: 05/07/2017**

**Prepared by:**

Dipendu Chanda  
Piyush Kulkarni  
Ramesh Kumar Nunna  
Ramkumar Sreeram  
Shiva Abhishek Akella

**CONTENTS:**

Executive Summary:.....	2
Demographic Factor Analysis for Donations:.....	3
Donation Behaviour: .....	4
Appeal Analysis .....	5
Donation Analysis.....	7
Behavioral factor analysis - RFM analysis .....	9
Donor RFM Analysis .....	9
Zip code Level RFM Analysis .....	12
Scope for Further Analysis .....	14
Appendix .....	15

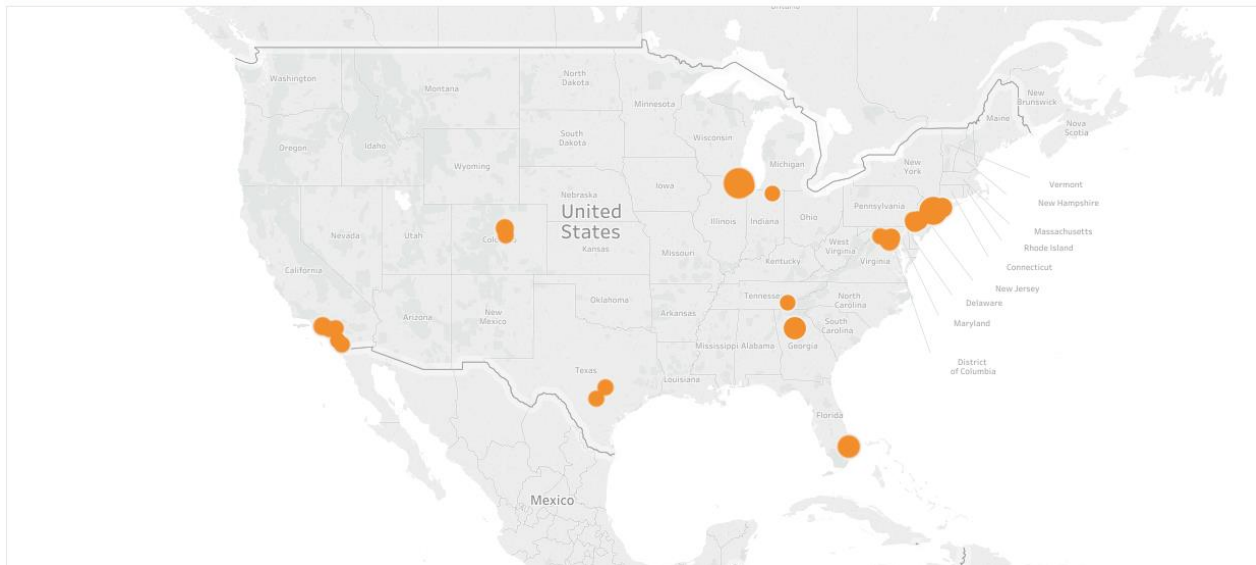
# Executive Summary

- We have analyzed that the donations are directly related with appeals and it strikes an emotional chord with the sentiments of the donor.
- We captured other demographic data points at source to predict future donor behavior with better accuracy using our models.
- There is a consistent one month window between appeal and donation which is a trend the marketing campaign manager of the NGO can seriously consider into account before taking any future action, also there is no relationship between appeal amount and gift amount.
- There is also an increasing trend in donation during holiday season and the overall donations are gradually dwindling, therefore the marketing campaign manager has to aggressively send campaigns to get more donations or breakeven with campaign costs.
- Although emotional factors drive the donor behavior also average household income, the socio-economic strata of the individual and his dependents affect the donor spending, also the based-on location of the donor (rural, urban, semi-urban) the donor behavior differs despite the same type of appeals being sent in the above-mentioned case.
- The substantial part of the donation comes from people who are loyal and are regularly giving donation and the donation that comes from new donors per year is insignificant at 2%. NGO should try to send appeals to new people to get more donors. NGO should target zip code present in segment 2 of our zipcode-rfm clusters to acquire new donors.

# Demographic Factor Analysis for Donations:

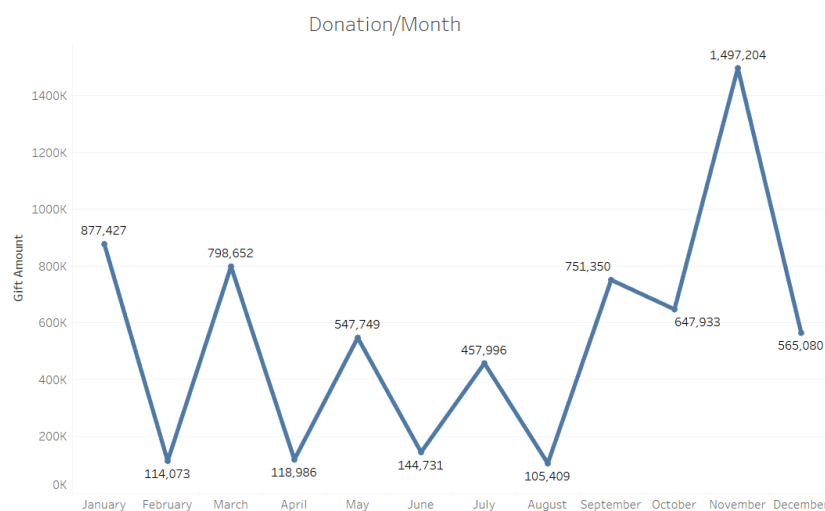
- As we can see from the above Map, most of the top donations are coming from East and West coast.

Places with the most donation amounts.



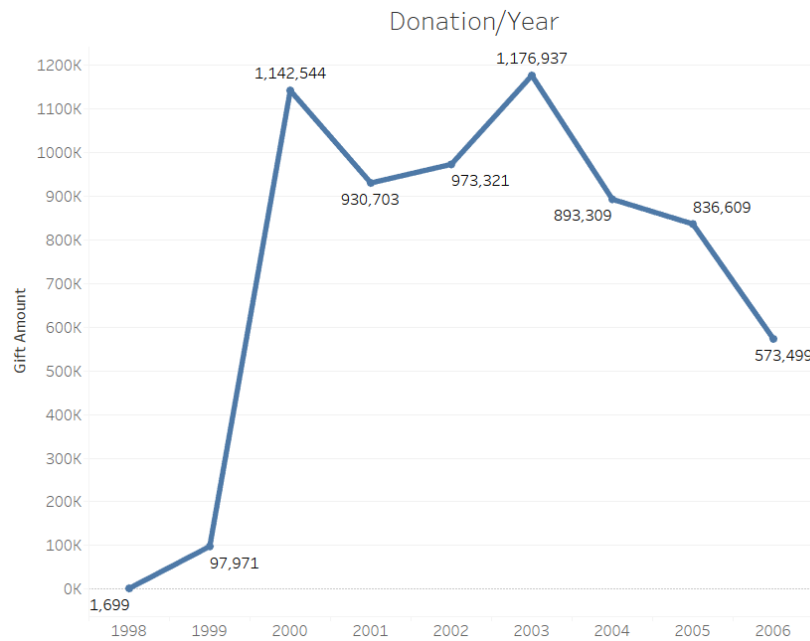
- Seasonality Analysis:** Analysis around the number of donation across months presents us the following picture about the market (shown below).

As we can clearly see, the last few months of the year make up what is commonly called the Giving Season for the nonprofit community. As the holidays near, people may feel encouraged to give more generously than during the rest of the year.



- Trend Analysis:** Analysis around the number of donation across years presents us the following picture about the market.

As we can see from the chart below, the donations from year 2004 and 2005 decreased. It may be due to churning of donors further explained in RFM analysis.

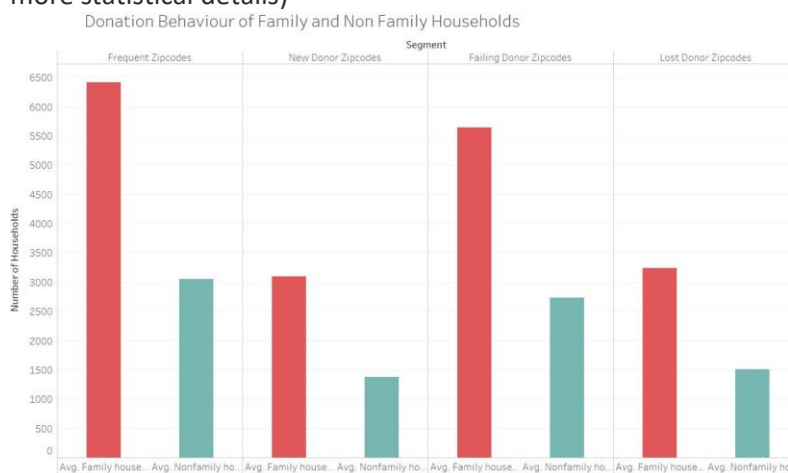


- **For New Donor Acquisition:** We should target those donors who falls in the geographical regions and have demographical factors similar to the donors in segment 1 and 3 of our donor RFM analysis. We should target:
  1. Person having average household income greater than \$65,000
  2. Married couple and family households having kids
  3. Household in urban areas

## Donation Behaviour:

1. Based on our statistical analysis (\*See appendix 4), donations are highly affected by the appeals sent
2. For most of the data we see that there is a gap of more than a month between appeal date and gift date, this could be because of the time taken for enrolment in to a donation plan and actual donation (\*See appendix 5).
3. Donation behaviour may not be fully determined by just the demographic factors. It has an emotional factor to it which cannot be captured.

The factors which boost the effectiveness of appeal are Average household income, type of household and the presence of children in the house and marital status. (\* refer appendix 4 for more statistical details)



# Appeal Analysis

## 1. Male versus Female:

According to the zip code data Male tends to respond more to our appeals and donate greater number than females.

We compared the population on zip code level and saw that for places with more men than female population the appeal conversion rates are more.

## 2. Appeals are more effective on Zip code with more Urban population. It is explained in detail in RFM analysis.

## 3. We merged demographic level data with donation's zip code level data to determine top zip codes based on appeals and donation and found that 70% of the zip codes match. Which suggest that our appeals are effective. (\*see appendix 6)

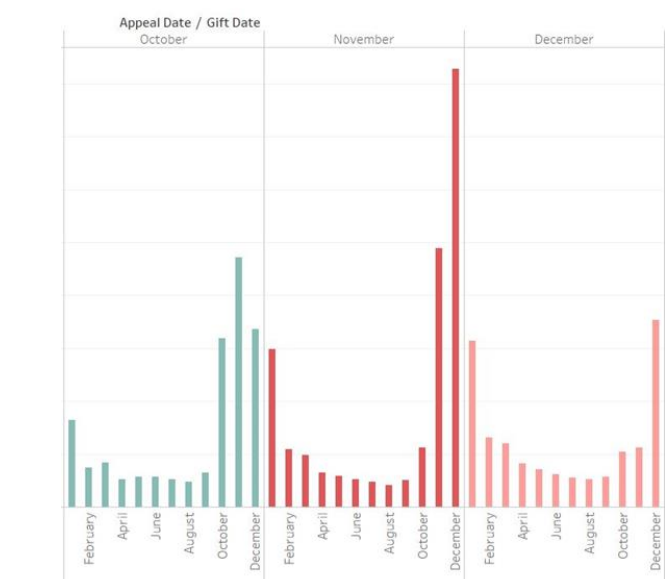
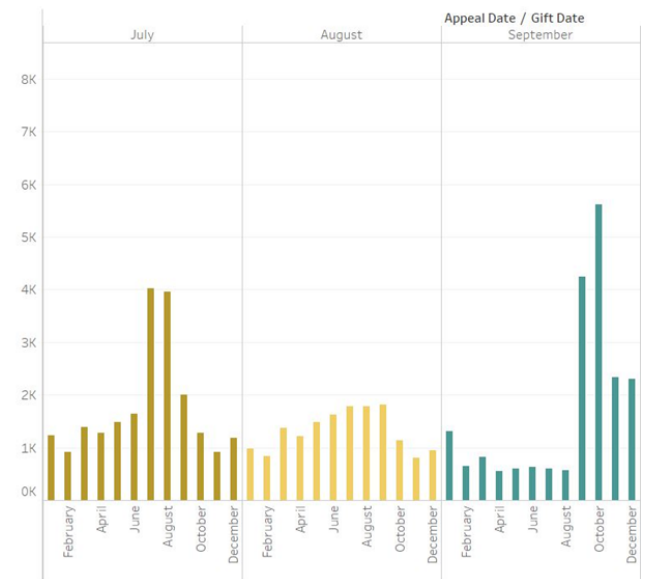
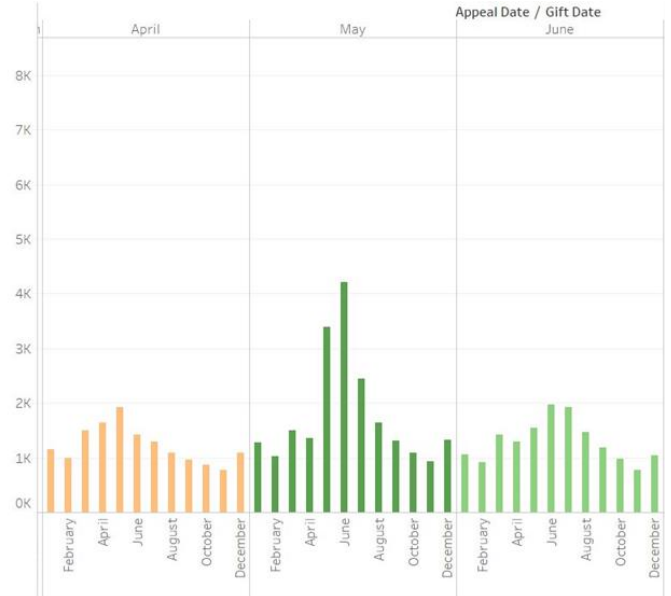
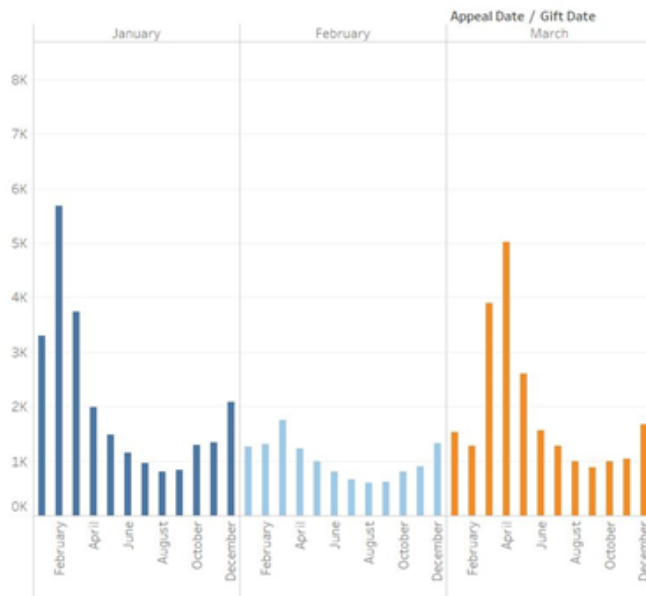
## 4. **Impact of appeals on Donation Propensity:**

1. Appeals has an emotional factor associated with it which effects the donation behaviour. And therefore, appeals are the most important factor responsible for donations. (\*see appendix 4)

2. The cost per appeal has almost no effect on the gift amount associated with that appeal. (\*see appendix 10)

3. Gift date and donation date are highly related which gives few key insights:

There is a successive monthly relationship between appeals and donation. That is the maximum response for an appeal occurs in the next month of that appeal.

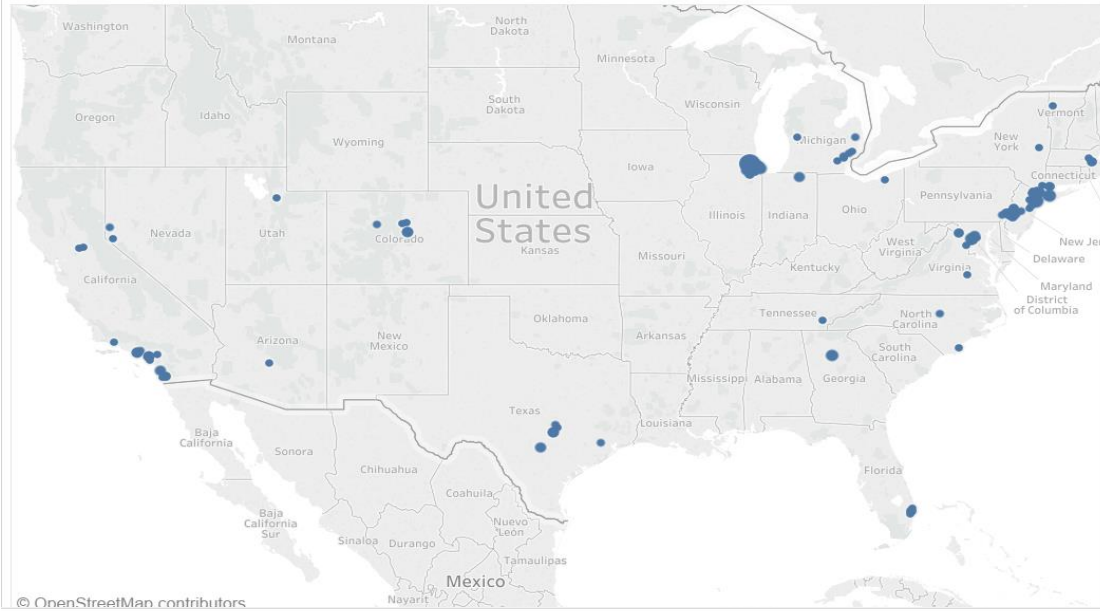


# Donation Analysis

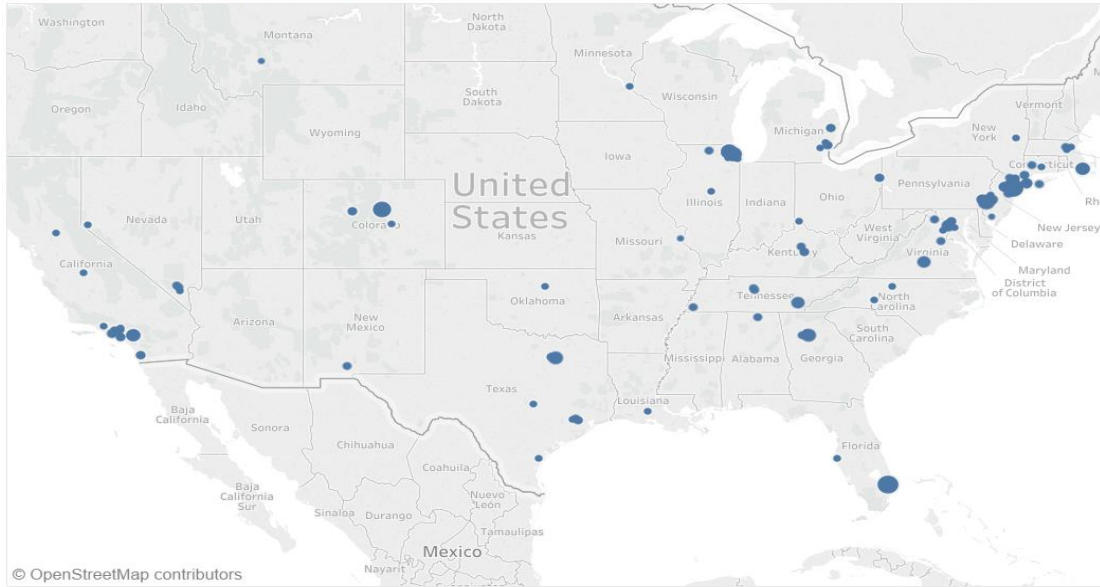
1. Analysis of train data: Using multiple linear regression of uncorrelated demographic data, we came to know that average household income, number of urban housing units, number of semi urban housing units are all the significant parameters affecting the total amount donated. The results can be seen in Appendix 7.
2. Analysis of future (test) data: Average household income and the total number of housing units have a significant effect on the total amount donated. The results can be seen in Appendix 8.
3. Customer lifetime value: We analysed the data for with assumption that the donors average age is 33 years and according to the centre for disease control the average expectancy is 65 years, also the average 32 year old donor contributes twice an year with a mean contribution of 15\$, therefore the average customer life time value is  $33 \times 2 \times (65 - 33) \times 15 = \$31,680$ .
4. The past donation amount of a donor is a good predictor of the future donation amount. The donor who are in segment 1 of our RFM analysis of donors will donate higher amount in near future.
5. And donors which are in segment 3 of our RFM analysis will keep donating small amounts near future also.
6. Average income and urban households has a positive relationship with the total amount donated.
7. We can see from the graphs below that the zip codes which gave most donations in past data (till 2004 Dec 31) were nearly same for the future data too.



Donations by Zipcode (train data)



Donations by zipcode (test data)



8.

## Behavioral factor analysis - RFM analysis

### Objective

Our Marketing strategy is to increase the donations and to do that one approach is to segment the donors by performing RFM Analysis. Recency (R), Frequency (F) and Monetary (M) describe donor's historical donation behavior.

RFM Analysis can help us

- Decide which all donors to appeal based on the likelihood and find ways to increase their donation; appeals can be made more directed to convince that particular cluster characteristics.
- Target lost donors or retain donors by adjusting the frequency of appeals to them. Also, one other reason to perform RFM Analysis is to find that minority of donors who are responsible for most our donation.

### Steps followed:

- Step1, did RFM analysis on Panel Donations data. We decided to do RFM analysis of consumers as well as zip code, as we have the demographic data on zip code level and not on Individual customer level.
- Step2, divided the donor data in 8 segments and
- zip code RFM scores and then performed cluster analysis on it. the zip code rfm data into 4 segments.
- Step3, we then joined demographic data with the

### Why Donor level RFM?

- To identify those Existing donors who are likely to respond to appeals
- Targeted approach specific to donor's characteristic

### Why Zip code level RFM?

- To Find future donors in accordance to their zip code
- How likely to get new donors from a specific zip code based on the Zip code characteristic

## Donor RFM Analysis

We performed RFM analysis on the dataset to segment donors based on RFM and to identify those donors who are likely to respond to our appeals. Our Marketing strategy is to increase the donations and to do that one approach is to segment the customers by performing RFM Analysis. Recency (R), Frequency (F) and Monetary (M) describe customer's historical donation behavior.

RFM Analysis can help us

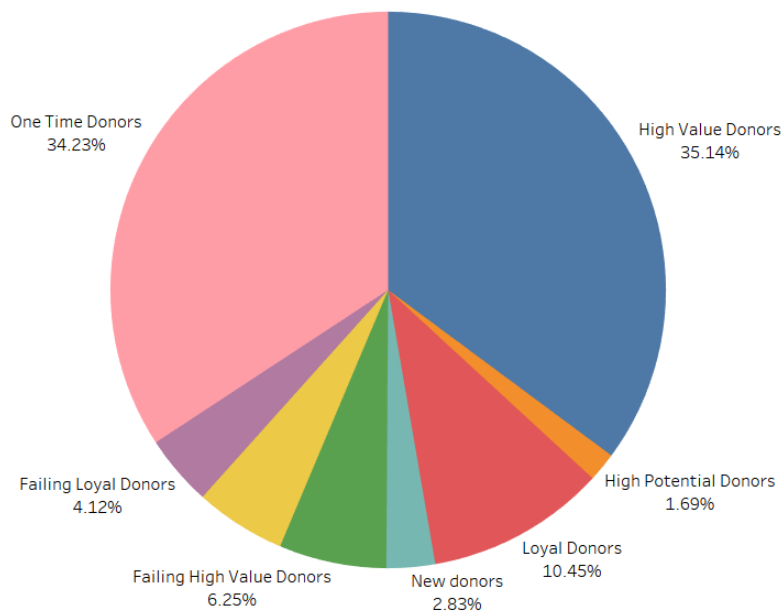
- Decide which all donors to appeal based on the likelihood and find ways to increase their donation. Appeals can be made more directed to convince any particular cluster characteristics.
- Target lost donors or retain donors by adjusting the frequency of appeals to them. Also, one other reason to perform RFM Analysis is to find that minority of donors who are responsible for most our donation.

We sorted donations data by donors and by their most recent donation date, number of donations and the total amount they donated. PROC RANK was used to rank these customers on three variables: Recency, Frequency and Monetary. Following 8 segments were created on different combination of R F M ranks.

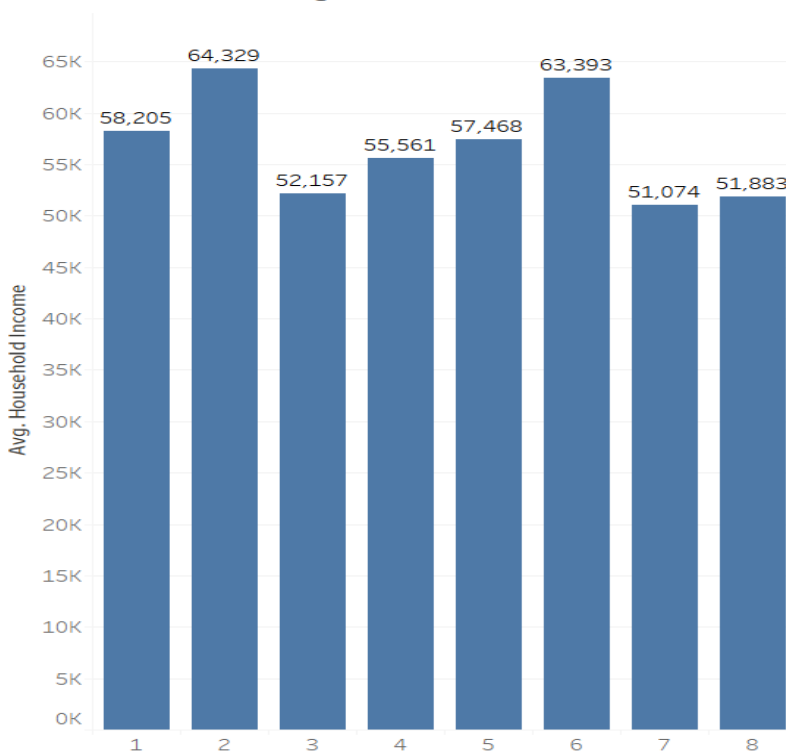
R	F	M	Segment	Label
0	1	1	1	<b>High Value Donors</b>
0	0	1	2	New donors with high donation amount <b>High Potential Donors</b> <i>"Big fish"</i>
0	1	0	3	Frequent donors who donate small amount regularly <b>Loyal Donors</b>
0	0	0	4	New donors who have just started to donate <b>New donors</b>
1	1	1	5	High value donors who have not donated recently. Could be possible that they have churned. Try to bring them back <b>Failing High Value Donors</b>
1	0	1	6	Donors who donated large amount in past but failed to donate regularly <b>Failing Starters</b>
1	1	0	7	These are the customers who used to donate often but stopped donating <b>Failing Loyal Donors</b>
1	0	0	8	Donors who donated very few times in past but didn't continue <b>One Time Donors</b> <i>"OK, as you asked nicely"</i>

The below graph shows the percent of donors per segments. As we can see most of the donors fall in segment 1 or in segment 8.

RFM- Donors per Clusters



Cluster's Average Household Income



#### Key Points:

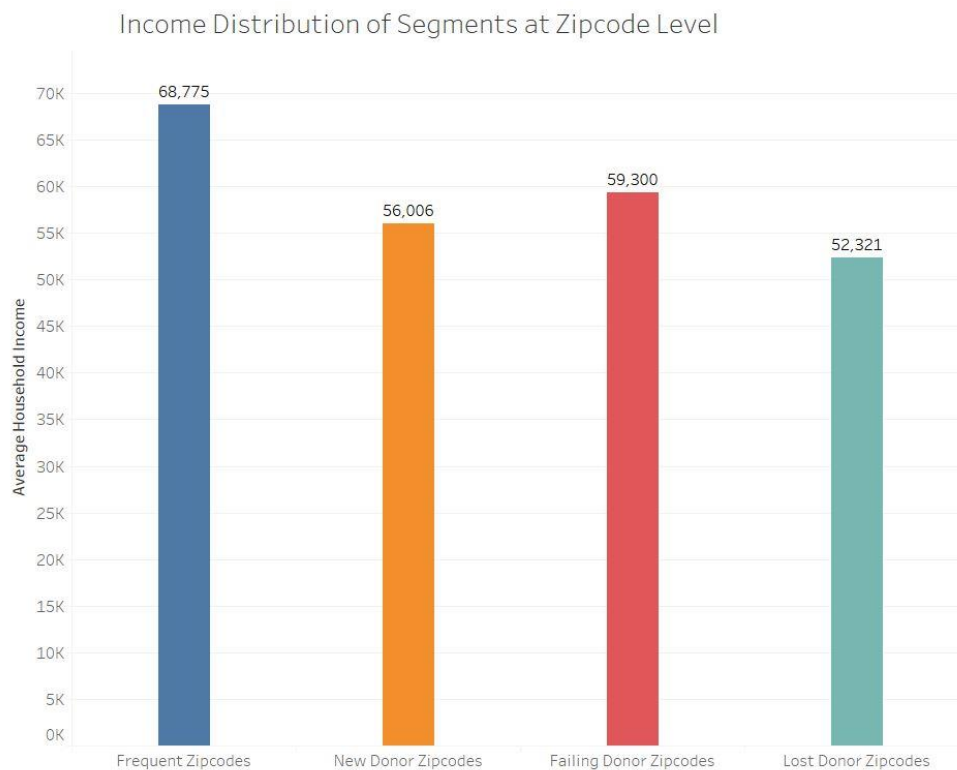
- As there are only 2.83% new donors the NGO be sending appeals to reach new donors
- Segment 2 donors (High Potential Donors) which are the Big Fish in our study should be given with more targeted appeals to increase their donation frequency.

- High Value Donors are our loyal customer which comprises of 35.14% of our total customers. These should not be neglected and assumed to keep donating. But not to bog them with many appeals.
- Failing High Value Donors, Failing Starters, Failing Loyal Donors are the donors who used to donate regularly in past but have stopped donating recently. Further investigation needs to be done for the reasons.
- One Time Donors which comprises of 34.23% are the donors who just donated few times in the past.

## Zip code Level RFM Analysis

Based on the Recency and Frequency, we divided the zip codes 4 Different Segments

R	F	Segment	Label
0	1	1	Loyal Donors Zip Codes
0	0	2	New donors zip code (High Potential)
1	1	3	Were frequent but churned Failing Donor Zipcodes
1	0	4	Lost customers. Why they were lost Lost donor Zipcodes

**Segment 1: Frequent Donor Zip codes**

Share: 41.63%

Brings the highest monetary value, they are also frequent donors. here are a few key characteristics.

- Average Household Income: **\$68,777**
- Mostly consist of urban households but have fair share of semi urban as well as rural households.

**Segment 2: New Donor Zip codes**

Share: 9.56%

The share of this segment is very less as compared to other segments. This means that we should explore new zip codes with characteristics as segment 1 zip codes to get more new donors. These are potential future loyal donors.

- Average household income: **\$56,000**
- Mostly consist of Semi urban and Rural households

**Segment 3: Failing Donor Zip codes**

Share: 11.41%

These are the zip codes from which we used to receive frequent donations but the donations have stopped recently.

- The segment is largely dominated by urban and rural households.

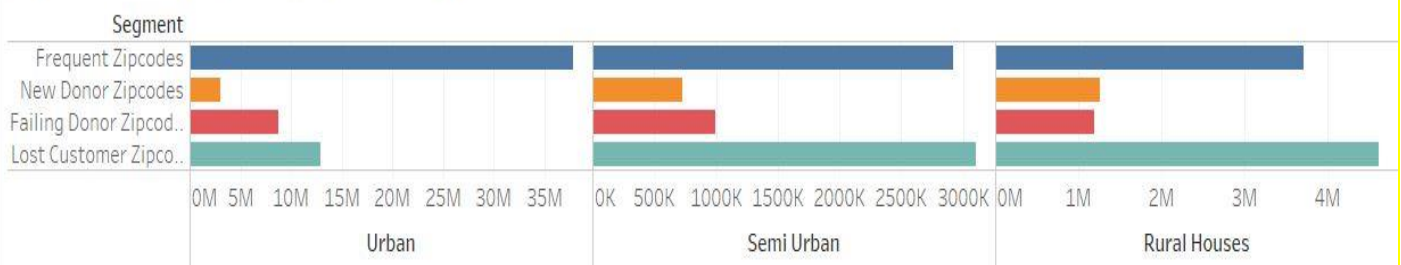
**Segment 4: Lost Donor Zip codes**

Share: 37.41%

This segment consists of donor zip codes who donated very few times in the past and have not donated after that. The large percentage of donors in this segment could be because the donors were not satisfied with their experience.

- This segment is also dominated by Urban households but Semi Urban and Rural Households also have sizable share.

## Household Type in Different Segments



## Scope for Further Analysis

- Presently the demographical data available is on zip code level and it creates biasing if we mirror it on donor level. So, to do more detailed analysis, donor level demographic data is needed.
- So, if the donor's demographic data is available, it should be incorporated with the donations data. Else we need to design a field study with different segments of donors. And their demographics could be treated as base characteristics for that segment.
- Further, some more analysis should be done as in why a particular donor donates and what factor influence him to donate. These factors could be used in sending donor specific appeals.
- We also need to track the type of appeal are sending to the donor. Different type of appeals has different emotional factor associated with it. These could help us to judge the success of appeals better.
- Appeal can have a feedback mechanism to track the success of the appeal directly. We can use tools like google analytics to track the success of emails. Effectiveness of telephonic appeals and the conversation status can be tracked by the executives and they can then gauge the effectiveness of the publicity plan.
- We need to find out the churn ratio based on the response of the appeal campaign under different broadcast medium like telephone, email, in person media. Based on this, we can find out the significant driving factors for the churn and we can deploy suitable marketing campaigns/ donor relationship programs to mitigate churn and engage the donor for future donation.
- We need to identify which marketing campaign resonates emotionally with what age group of people while considering other geopolitical/cultural factors so that we can send out more appeals to people via the medium they are most receptive to.

# Appendix

## Appendix 1:

### Data Set Information:

In this project, we are using Donation and Appeal dataset.

Number of Observations:

Appeal: 1,730,598 observation

Donation: 192,840 observation

For Logistic regression: 269100 Observation after cleaning

## Appendix 2:

### Correlation Between all Demographic Variables

Pearson Correlation Coefficients, N = 32038																
	Zip	Total_Households	IncomeAbv60	Median_HH_Income	AvgIncome	Total_Households_	Family_households_	MarriedWithKid	Nonfamily_households_	Total_population	validMale2244	Female2244	Total_Housing_Units	Urban	Semi_Urban	Rural
Zip	1.00000	-0.00995	-0.00514	-0.10511	-0.10033	-0.00995	-0.00625	0.02512	-0.01552	0.00745	0.02381	0.00650	-0.01379	-0.01102	0.04375	-0.08137
Total_Households	-0.00995	1.00000	0.87655	0.20585	0.25864	1.00000	0.97717	0.91300	0.92140	0.98184	0.95936	0.97469	0.99554	0.92298	0.21281	0.11160
IncomeAbv60	-0.00514	0.87655	1.00000	0.45781	0.50437	0.87655	0.86316	0.88347	0.78458	0.85559	0.84436	0.86444	0.85974	0.83679	0.09871	0.02257
Median_HH_Income	-0.10511	0.20585	0.45781	1.00000	0.94860	0.20585	0.23630	0.31400	0.12538	0.20691	0.19551	0.21059	0.19004	0.21189	-0.04358	-0.03329
AvgIncome	-0.10033	0.25864	0.50437	0.94860	1.00000	0.25864	0.28605	0.35931	0.17739	0.25684	0.24701	0.26133	0.24349	0.25650	-0.02699	-0.01051
Total_Households_	-0.00995	1.00000	0.87655	0.20585	0.25864	1.00000	0.97717	0.91300	0.92140	0.98184	0.95936	0.97469	0.99554	0.92298	0.21281	0.11160
Family_households_	-0.00625	0.97717	0.86316	0.23630	0.28605	0.97717	1.00000	0.96226	0.81780	0.98961	0.94766	0.97065	0.97040	0.88926	0.21863	0.14348
MarriedWithKid	0.02512	0.91300	0.88347	0.31400	0.35931	0.91300	0.96226	1.00000	0.71404	0.94930	0.92239	0.93829	0.89645	0.82609	0.19250	0.13248
Nonfamily_households_	-0.01552	0.92140	0.78458	0.12538	0.17739	0.92140	0.81780	0.71307	1.00000	0.84943	0.86530	0.86477	0.92168	0.87357	0.17655	0.03984
Total_population	0.00745	0.98184	0.85559	0.20691	0.25684	0.98184	0.98961	0.94930	0.84943	1.00000	0.97549	0.98718	0.97484	0.90544	0.20390	0.10723
validMale2244	0.02381	0.95936	0.84436	0.19551	0.24701	0.95936	0.94766	0.92239	0.86530	0.97549	1.00000	0.98436	0.94959	0.89709	0.17011	0.07110
Female2244	0.00650	0.97469	0.86444	0.21059	0.26133	0.97469	0.97065	0.93829	0.86477	0.98718	0.98436	1.00000	0.96482	0.91611	0.15984	0.06756
Total_Housing_Units	-0.01379	0.99554	0.85974	0.19004	0.24349	0.99554	0.97040	0.89845	0.92168	0.97484	0.94959	0.96482	1.00000	0.91696	0.23103	0.13807
Urban	-0.01102	0.92298	0.83679	0.21189	0.25650	0.92298	0.88926	0.82609	0.87357	0.90544	0.89702	0.91611	0.91696	1.00000	-0.12828	-0.17370
Semi_Urban	0.04375	0.21281	0.09871	-0.04358	-0.02699	0.21281	0.21863	0.19250	0.17655	0.20390	0.17011	0.15984	0.23103	-0.12828	1.00000	0.37234
Rural	-0.08137	0.11160	0.02257	-0.03329	-0.01051	0.11160	0.14348	0.13248	0.03984	0.10723	0.07110	0.06756	0.13807	-0.17370	0.37234	1.00000

## Appendix 3:

Correlation between cleaned demographic variables used for testing the effectiveness of the appeals.

Pearson Correlation Coefficients Number of Observations						
	effAppeal	Zip	Median_HH_Income	MarriedWithKid	Semi_Urban	Rural
effAppeal	1.00000 269100	0.01086 264214	-0.12373 264214	0.00337 264214	0.03845 264214	0.03295 264214
Zip	0.01086 264214	1.00000 264214	-0.15685 264214	0.08463 264214	0.07630 264214	0.02760 264214
Median_HH_Income	-0.12373 264214	-0.15685 264214	1.00000 264214	0.16514 264214	-0.21963 264214	-0.23531 264214
MarriedWithKid	0.00337 264214	0.08463 264214	0.16514 264214	1.00000 264214	-0.04894 264214	-0.09126 264214
Semi_Urban	0.03845 264214	0.07630 264214	-0.21963 264214	-0.04894 264214	1.00000 264214	0.43653 264214
Rural	0.03295 264214	0.02760 264214	-0.23531 264214	-0.09126 264214	0.43653 264214	1.00000 264214

## Appendix 4:



## Checking the effectiveness of Appeals

### Data Cleaning:

#### Divide the dataset:

We started by dividing our dataset into Train, Validation and Test set

#### Preprocessing:



1. We created a set with equivalent number of effective appeals values, which is our final binary Y variable to check the effectiveness of our appeals. (Total of 192,840 observations)
2. We created new data frame with the useful demographic columns
3. We Imputed few of the missing values in 2 of the columns we got from demographic data
4. We found the correlations of all the input variables and removed correlated column from our models.

We were left with 269100 Observation after cleaning.

### Models:

A sample run of the following models were made on the dataset:

1. Linear Regression Model along with Robust Regression
2. Logistic Regression Models
3. Probit Model
4. Logit Model

 <b>'Linear Probability Model'</b> 22:17 Tuesday, May 2, 2017					 <b>'Linear Probability Model - Robust Standard Errors'</b> 22:17 Tuesday, May 2, 2017				
The REG Procedure Model: MODEL1 Dependent Variable: effAppeal					The REG Procedure Model: MODEL1 Dependent Variable: effAppeal				
Number of Observations Read 269100 Number of Observations Used 264214 Number of Observations with Missing Values 4886					Heteroscedasticity Consistent Covariance of Estimates				
<b>Analysis of Variance</b>					<b>Variable</b>				
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	Intercept	Zip	Median_HH_Income	Married_WithKid
Model	6	1252.06497	208.67749	850.82	<.0001	0.0000165743	-6.12968E-11	-1.6722E-10	1.0012208E-9
Error	264207	64801	0.24527			-6.12968E-11	1.139801E-15	3.101777E-16	-4.1895E-15
Corrected Total	264213	66053				-1.6722E-10	3.101777E-16	2.524598E-15	-1.98311E-14
						1.0012208E-9	-4.1895E-15	-1.98311E-14	8.82485E-13
						Urban	-4.87217E-10	5.041199E-15	-1.78662E-13
						Semi_Urban	-5.53964E-10	7.066376E-15	-1.52871E-13
						Rural	-8.18998E-10	7.024012E-15	-1.32421E-13
						Heteroscedasticity Consistent Covariance of Estimates			
Root MSE		0.49525	R-Square	0.0190		Variable	Urban	Semi_Urban	Rural
Dependent Mean		0.50005	Adj R-Sq	0.0189		Intercept	-4.87217E-10	-5.53964E-10	-8.18998E-10
Coeff Var		99.04002				Zip	6.637337E-16	-3.1646E-16	8.470745E-16
						Median_HH_Income	5.041199E-15	7.066376E-15	7.024012E-15
						Married_WithKid	-1.78662E-13	-1.52871E-13	-1.32421E-13
						Urban	5.62866E-14	4.775634E-14	4.584467E-14
						Semi_Urban	4.775634E-14	2.840256E-13	-4.75713E-14
						Rural	4.584467E-14	-4.75713E-14	2.214648E-13
						Analysis of Maximum Likelihood Estimates			
						Parameter	DF	Estimate	Standard Error
						Intercept	1	0.8939	0.0176
						Zip	1	-1.22E-6	1.381E-7
						Median_HH_Income	1	-0.00002	2.296E-7
						Married_WithKid	1	0.000121	3.999E-6
						Urban	1	-0.00003	1.023E-6
						Semi_Urban	1	-0.00001	2.229E-6
						Rural	1	-0.00002	1.925E-6
						Odds Ratio Estimates			
						Effect	Point Estimate	95% Wald Confidence Limits	
						Zip	1.000	1.000	1.000
						Median_HH_Income	1.000	1.000	1.000
						Married_WithKid	1.000	1.000	1.000
						Urban	1.000	1.000	1.000
						Semi_Urban	1.000	1.000	1.000
						Rural	1.000	1.000	1.000
						Association of Predicted Probabilities and Observed Responses			
						Percent Concordant	56.8	Somers' D	0.146
						Percent Discordant	42.2	Gamma	0.147
						Percent Tied	1.0	Tau-a	0.073
						Pairs	17452259305	c	0.573

Based on our finding  $R^2$  values comes to be insignificant.

It shows that the appeals are not directly determined by the demographic factors only. It has an emotional factor to it which cannot be captured by demographics variables.

**Appendix 5:**

The below table is just a small subset of data which contains the minimum difference between the appeal date and the Gift date. The delay is normally more than a month.

Appeal Id	Totalappea...	Totalgiftam...	Gift Date	Appeal Date	difference i... ٣
00AD2AA228	3.928	165.00	1/3/2000	11/5/1999	59
00AD2AA190	5.331	40.00	1/3/2000	11/5/1999	59
00AD2AA126	50.789	485.00	1/3/2000	11/5/1999	59
00AD2AA156	5.893	65.00	1/3/2000	11/5/1999	59
00AD2AA149	40.687	1,035.00	1/3/2000	11/5/1999	59
00AD2AA120	4.209	10.00	1/3/2000	11/5/1999	59
00AD2AA118	23.009	322.00	1/3/2000	11/5/1999	59
00AD2AA085	17.397	1,295.00	1/3/2000	11/5/1999	59
00AD2AA082	11.785	235.00	1/3/2000	11/5/1999	59
00AD2AA064	34.233	677.00	1/3/2000	11/5/1999	59
00AD2AA017	10.382	170.00	1/3/2000	11/5/1999	59

**Appendix 6:**Donation and Appeal Relations:

Top Zip codes from where we are getting the maximum donation is matching 70% with the Top zip codes where maximum appeals were sent.

```

176 PROC SQL;
177 CREATE TABLE project.topDonationAppeal as
178 SELECT b.zipcode, b.TOTALDONORS, a.TOTALAPPEALS
179 FROM project.mTopZipappeals as a, project.mTopZipDonation as b
180 WHERE a.zipcode = b.zipcode;
NOTE: Table PROJECT.TOPDONATIONAPPEAL created, with 70 rows and 3 columns.

181 QUIT;
NOTE: PROCEDURE SQL used (Total process time):
      real time          0.09 seconds
      cpu time           0.03 seconds

```

Correlation between Total donation and Total Appeals comes to be **0.6334**.

Pearson Correlation Coefficients, N = 37858 Prob >  r  under H0: Rho=0			
	donor_id	TOTALDonations	TOTALAppeals
donor_id	1.00000	-0.10859 <.0001	-0.16798 <.0001
TOTALDonations	-0.10859 <.0001	1.00000	0.63339 <.0001
TOTALAppeals	-0.16798 <.0001	0.63339 <.0001	1.00000

**Appendix 7:**

First we separated the past data as the data that has all the gift dates before Jan 1 2005 to analyze the past donation behavior. To analyze the effects of demographical factors like average income, number of household unit on the

amount of total donation made in each zip code we merged the demographic data with zip code data and then ran a multiple linear regression with uncorrelated demographic data as our independent variables and the total amount donated from each zip code as our dependent variable.

**The SAS System**

**The CORR Procedure**

3 Variables: Avg\_HH\_Income Urban Semi\_Urban


Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Avg_HH_Income	10428	60324	27955	629062034	0	1063163
Urban	10428	5973	6761	62283880	0	69229
Semi_Urban	10428	736.24607	2107	7677574	0	28160

Pearson Correlation Coefficients, N = 10428 Prob >  r  under H0: Rho=0			
	Avg_HH_Income	Urban	Semi_Urban
Avg_HH_Income	1.00000	0.06072 <.0001	-0.15575 <.0001
Urban	0.06072 <.0001	1.00000	-0.30437 <.0001
Semi_Urban	-0.15575 <.0001	-0.30437 <.0001	1.00000

## Appendix 8

To analyze the effects of demographical factors on the future data (Jan 1 2005 to 2006), we segregated the data with gift date after Jan 1 2005 and ran a multiple linear regression to predict the total amount donated.

 **The REG Procedure**  
Model: MODEL1  
Dependent Variable: TOTDONATION

Number of Observations Read	10428
Number of Observations Used	10428

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1743786941	581262314	1002.89	<.0001
Error	10424	6041639195	579589		
Corrected Total	10427	7785426136			

Root MSE	761.30765	R-Square	0.2240
Dependent Mean	462.74173	Adj R-Sq	0.2238
Coeff Var	164.52107		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	-472.72262	19.94028	-23.71	<.0001
Avg_HH_Income	1	0.01100	0.00027002	40.74	<.0001
Urban	1	0.04099	0.00116	35.40	<.0001
Semi_Urban	1	0.03674	0.00375	9.79	<.0001

## Appendix 9:

To understand how good customer lifetime value (total donations over the duration of the dataset) are correlated with demographics we did a correlation test between some of the demographic factors and the customer lifetime value. Then we ran a regression model with the total lifetime value of a customer (total amount donated) to find the significant uncorrelated variables.

The CORR Procedure

5 Variables: TotaalAMount Avg\_HH\_Income Total\_Households Urban Semi\_Urban

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
TotaalAMount	41399	146.11224	394.86169	6048901	0	23600
Avg_HH_Income	41399	71238	30764	2949179816	0	1063163
Total_Households	41399	10359	6418	428840147	0	61898
Urban	41399	9717	7497	402285413	0	69229
Semi_Urban	41399	582.85777	2141	24129729	0	28160

Pearson Correlation Coefficients, N = 41399 Prob >  r  under H0: Rho=0					
	TotaalAMount	Avg_HH_Income	Total_Households	Urban	Semi_Urban
TotaalAMount	1.00000	0.09544 <.0001	0.00558 0.2566	0.00982 0.0457	-0.01513 0.0021
Avg_HH_Income	0.09544 <.0001	1.00000	-0.06996 <.0001	0.01025 0.0371	-0.18170 <.0001
Total_Households	0.00558 0.2566	-0.06996 <.0001	1.00000	0.92378 <.0001	-0.00790 0.1079
Urban	0.00982 0.0457	0.01025 0.0371	0.92378 <.0001	1.00000	-0.34876 <.0001
Semi_Urban	-0.01513 0.0021	-0.18170 <.0001	-0.00790 0.1079	-0.34876 <.0001	1.00000

The REG Procedure  
Model: MODEL1  
Dependent Variable: TotaalAMount

Number of Observations Read	41399
Number of Observations Used	41399

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	59773337	29886669	193.47	<.0001
Error	41396	6394826922	154479		
Corrected Total	41398	6454600259			

Root MSE	393.03860	R-Square	0.0093
Dependent Mean	146.11224	Adj R-Sq	0.0092
Coeff Var	268.99772		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	50.20559	5.96410	8.42	<.0001
Avg_HH_Income	1	0.00124	0.00006295	19.64	<.0001
Total_Households	1	0.00075758	0.00030174	2.51	0.0121

## Appendix 10:

Correlation of donation date - donation date and appeal cost and gift amount

Pearson Correlation Coefficients, N = 224095 Prob >  r  under H0: Rho=0				
	TOTALAPPEALSCOST	TOTALGIFTAMOUNT	gift_date	appeal_date
TOTALAPPEALSCOST	1.00000	0.08505 <.0001	0.12939 <.0001	-0.15838 <.0001
TOTALGIFTAMOUNT	0.08505 <.0001	1.00000	-0.05349 <.0001	-0.04753 <.0001
gift_date	0.12939 <.0001	-0.05349 <.0001	1.00000	0.98249 <.0001
appeal_date	-0.15838 <.0001	-0.04753 <.0001	0.98249 <.0001	1.00000

## Appendix 11:

### Demographic Analysis Steps:

#### Data pre- processing

**Centering and Scaling/Normalization:** The dataset was reasonable balanced and there was no need for up sampling and down sampling or transforming the data using any mathematical transformation like Log, Box & Cox transformation.

**Omitted and Missing value computing and Imputing:** The dataset was first checked for missing values by the proc means procedure, there wasn't any trace of missing values and hence there was no necessity for imputing the missing values.

**Feature Selection:** Data about the zip code level information about the population, income groups, family and non-family households, age, Urban and rural households was imported and consolidated with the original dataset using the proc Import step in SAS.

**Demographical Zone sorting of the postal codes:** The postal codes were divided into four zones according to the data obtained from the US census repository, and this was then used as sorting algorithm to divide the zones into four principal zones South, West, Northeast, Midwest.

### Analysis of Variance – ANOVA

The P value is significant as it is  $< .001$  and has a F value of 2.59 ( $< 3$ ), Hence we can safely reject the null hypothesis for the alternative hypothesis that the average value of Region and gift\_amount are unequal and donation is not static within different regions. Hence it makes sense for the marketing campaigner to target differently to different regions

### Regression Analysis

Ordinary linear regression using all numeric and factor level variables

The model seemed underperforming with the dependent variable as Gift\_amnt\_sum and yielded some abnormal and inconsistent results. The number of observations with missing values is 2710 as the data that was merged had spurious and missing merge fields when consolidated with the earlier dataset.

Also the F value is 92.74, and the coefficient of Variance is 256.23 and the model performance is abysmal  $R^2$  value is at 0.0283. Hence we need to go for further analysis for modifying the regression equation for interpretable results.

The partial least square model also is inefficient with number of extracted factors as '1' and the current effect and total effects being almost one hundred percent correlated hence the PLS model accounts for almost nothing in the PLS model and is also inefficient.

We have to use a logit model to explore the investigation between explanatory variables further.

### Region

Census_region_name
South
West
Northeast
Midwest

### Missing Values & Imputation

The selected features from the significant results of the regression are chosen (\*\*\*) at 0.001 significance and rest are dropped for an ordinary linear regression.

## Appeals

## Donations

## The MEANS Procedure

Variable	N	N Miss
appeal_date	1730598	0
appeal_cost	1730598	0

## The MEANS Procedure

Variable	N	N Miss
gift_amount	192840	0
gift_date	192840	0
first_gift_date	192840	0