TERADATA University Network

DATA CHALLENGE

Teradata 2018 Data Challenge KDD Group Project

Author(s)

Name *University* *Email address* *Major* *Level*	Sudhakar Nagarajan University of North Carolina - Charlotte snagar13@uncc.edu Data Science and Data Analytics Master	Vasanth Balasubramanian University of North Carolina - Charlotte vbalasu7@uncc.edu Data Science and Data Analytics Master

Faculty Advisor

Name: Dr. Pamela Thompson

Email: plthomps@uncc.edu

Phone: 980-234-0042

Summary

1) Abstract

Bike MS is the fundraising cycling series for the National Multiple sclerosis foundation. Multiple sclerosis is an unpredictable, often disabling disease of the central nervous system that disrupts the flow of information within the brain and between the brain and body. According to Bike MS website, Bike MS cyclists, volunteers, and donors have raised more than $1 billion so people affected by MS can live their best lives as we stop MS in its tracks, restore what's been lost and, end MS forever. Even though Bike MS is the biggest cycling fundraiser, over the years, many other organizations have also trying to attract the same audience. Because of this, there is a steady decrease in revenue for Bike MS since 2012.

The objective of this project is to identify opportunities and the impact on Bike MS fundraising. In addition to identifying key factors, explore the greatest opportunities for digital marketing investments and the greatest ROI.

Problem and Motivation

We wanted to answer the following business problems as part of our analysis

- Greatest growth opportunities for the corporate teams
- What are the common denominators of the top performing teams i.e. connection to MS, gender, prior year participation, digital communication preference.
- Once someone is registered, what are the tactics and behaviors that drive fundraising?

Problem Approach

As part of this exercise, we wanted to analyze the behavior of corporate teams by understanding the factors that affect the gift contribution. We thought each independent variable may have some influence on the outcome variable. The bottom line for National Multiple Sclerosis Society (NMSS) is to increase the donations and contributions by controlling the attrition of the teams and its participants and reach out to the new corporate teams. The dataset was very huge. Due to infrastructure constraints, we wanted to pick specific variables which we think will have a greater impact on the outcome, specifically from corporate sponsored teams to align our analysis with the business challenges.

We took three methods in our problem approach and they're explained in detail below.

The first and the second methods involved the combination of teams and participants datasets. In our first method, 'ParticipantContribution' was the target variable whereas in the second method, 'Non-ParticipantContribution' was the target. The predictor variables are the same in both these methods.

We considered "CaptainAcceptEmail", because if the captain preferred digital communication, there could be more awareness of this social cause and the donation requests could be circulated to his/her team members, to many non-participants, and eventually generate more revenue from non-participants to NMSS.

We considered "Team-PriorYearParticipation", because we wanted to understand the impact of the team participation in the previous years and see if this plays an important role in the retention and fetch more donations. Similarly, another variable "PriorParticipant", the participant being part of Bike MS events in the previous years was also considered for more revenue opportunity.

We wanted to analyze how gender plays a key role in predicting the $ contribution. There were instancesThat is why we included "GenderMale", "GenderFemale" and "GenderNotSpecified".

Furthermore, we thought more the participant or the team goes digital more would be the fund raising when compared with the traditional methods. To validate and understand our assumption, we have included "RegOnline" and "RegOffline" variables.

One important variable we thought we should include is the participant connection to MS. we wanted to understand how much impact that creates in terms of bringing non-participant $ revenue.

Datasets

The following CSV files provided at the competition URL were imported into a relational database (SQL) for cleanup and transformation.

2013-2017 Bike Teams.csv

2013-2017 Bike MS Participants.csv

2013-2017 Bike MS Participants.csv

2013 Bike Donations.csv (Similar files for the years 2014 through 2017).

Data Preparation

This was the most labor intensive process for us in order to prepare the final dataset for our analytics and modeling tools. This included the cleaning of raw, dirty data and selection of the cases and variables that were appropriate for our analysis. We also performed transformations on certain variables. Some of the activities performed in this process are explained below.

1) Cleanup of the spell errors in the values of some of the attributes.

Team Division
Frien`s and Family

- [ ] ofbline
- [ ] offline
- [ ] online

- [x] Feiale
- [x] Female

- [x] Acpive
- [x] Active
- [x] Inactive

2) In some of the attributes, the values were represented with different spellings. This cleanup or correction did not change the original meaning of the values.

In the Gender attribute, Female values were also represented as F. In order to maintain the consistency, the values were transformed from their raw format.

- [x] (Select All)
- [x] F
- [x] Female
- [x] Male

- [x] M
- [x] Male

- [ ] Bike Shop
- [ ] Bike Shops

- [ ] Corporate
- [ ] Corporation

- [ ] Family and Friends
- [ ] Family/Friends
- [ ] Frien`s and Family
- [ ] Friend and Family
- [ ] Friends and Family

- [x] Organization (Clubs, Civic Groups, etc.)
- [x] Organization (Clubs; Civic Groups; etc.)

3) In some of the attributes, the values did not match with the attribute. Instead, they contained the values of the adjacent or neighboring attributes in the original CSV files.

Offline Status	Soft Credit Type	Is Registration?	Donor ConsID	Donor Member ID	Donor Affiliate Code	Donor Gender
Teamraiser Participant Gift	FALSE	11567220	43323309	CAS	Male	FALSE
Teamraiser Participant Gift	FALSE	12582469	82311058	KSG	Female	TRUE

Participant Member ID	Participation Type Name	Registration Active Status	Registration Date	Participant Goal($)	Is Prior Participant	Is Team Captain	Additional Gift Amount($)
Participant - No Bus	Active	6/14/2012 13:42	2000	N/A	TRUE	0	DAs Against MS
Event Participant	Active	9/13/2012 10:41	200	Yes	FALSE	0	
Event Participant - Cyclist	Active	7/7/2013 19:52	300	N/A	FALSE	0	Team Mellen
62540741(Event Participant	Active	2/14/2012 16:52	300	Yes	FALSE	50	

4) The values in some of the attributes were transformed especially when the attribute contained several classes of information but conveying similar information for our research and analysis.

Example: ConnectionMS

We believed that a donor or team member would contribute more to the fundraising campaign if there is any connection to the MS disease regardless of who in his or her social network. All the below classes would be transformed in the attribute as the **person having a connection to MS**.

Care Manager of Person with MS
Caregiver of Person with MS
Child has MS
Friend has MS
I have a Friend or Co-worker with MS
I have MS
Other
Parent has MS
Possible MS
Relative has MS
Relative: Child of person with MS
Relative: Other
Relative: Parent of person with MS
Relative: Sibling of person with MS
Relative: Spouse of person with MS
Sibling has MS
Spouse has MS

Final Datasets

First, we cleaned the datasets and combined Teams and Participants csv files and created a new dataset called **Team-Participants-Merge.csv** with the following fields and the definitions.

Attribute	Definition	Type
CaptainAcceptEmail	Each Team will have a single captain. This attribute indicates if the captain of the team will accept an email.	Binary
PriorYearParticipation	It indicates if the team had participated in any of the events in the previous years.	Binary
GenderMale	# of Male participants in the team.	Numeric; Continuous

GenderFemale	# of Female participants in the team.	Numeric; Continuous
GenderNotSpecified	# of participants who have not specified their gender but part of the team.	Numeric; Continuous
RegOnline	# of team members who registered online.	Numeric; Continuous
RegOffline	# of team members who registered offline.	Numeric; Continuous
PriorParticipant	# of team members who participated in the any of the events in previous years.	Numeric; Continuous
NotPriorParticipant	# of team members who have not participated in the any of the events in previous years.	Numeric; Continuous
ConnectionMS	# of team members who have some connection to MS.	Numeric; Continuous
NoConnectionMS	# of team members who do not have some connection to MS.	Numeric; Continuous
RegActive	# of team members whose registration status is active.	Numeric; Continuous
RegInActive	# of team members whose registration status is not active.	Numeric; Continuous
TotalFromParticipants	Overall # of the Participants	Numeric; Continuous
TotalFromNonParticipants	Overall # of the non-participants	Numeric; Continuous

Second, we combined Teams and Donors csv files and created a new dataset called **Team-Donors-Merge.csv** with the following fields and the definitions.

Attribute	Definition	Type
GenderMale	# of Male participants in the team.	Numeric; Continuous

GenderFemale	# of Female participants in the team.	Numeric; Continuous
GenderNotSpecified	# of participants who have not specified their gender but part of the team.	Numeric; Continuous
RegisteredDonors	# of team members who registered online.	Numeric; Continuous
NonRegisteredDonors	# of team members who registered offline.	Numeric; Continuous
PriorDonor	# of team members who participated in the any of the events in previous years.	Numeric; Continuous
NotPriorDonor	# of team members who have not participated in the any of the events in previous years.	Numeric; Continuous
ConnectionMS	# of team members who have some connection to MS.	Numeric; Continuous
NoConnectionMS	# of team members who do not have some connection to MS.	Numeric; Continuous
Gift Amount	Amount contributed as a gift to NMSS	Numeric; Continuous

Finally, we combined the three datasets i.e. Teams, Participants and Donors to create a third dataset **Team-Donor-Participant-Merge.csv**. This dataset contained all the attributes from the earlier two datasets illustrated above.

Tools and Analytics

One of our major limitations is the use of our Ubuntu laptops for our analysis purposes. These datasets were extremely large and we had challenges even in opening them in VI editor. In addition to that, one of the working machines being on Linux OS, we can not use many of the visualization tools, like Tableau for exploratory analysis.

To overcome these challenges, we had created following an account in Amazon Web Services (AWS) cloud computing services. This enabled the group perform their data analysis remotely and concurrently without having to mess up with the distribution or transfer of files.

Server type	Purpose	Server hostname	Size
RDS	mysql server	bikems.chgkkfljddq9.us-east-2.rds.amazonaws.com	1-vCPU

db.t2.micro			1GB memory 20GB HDD
EC2 server m5.xlarge		ec2-52-15-142-133.us-east-2.compute.amazonaws.com	2-vCPU 16GB memory 450 GB HDD
Rstudio Server,	To write R code.	Installed RServer, running on port 8787 http://ec2-18-188-80-129.us-east-2.compute.amazonaws.com:8787/	
RAW Graphics Server	For Visualization	Installed RAW Graphics, running on port 4000 http://ec2-18-188-80-129.us-east-2.compute.amazonaws.com:4000/	
WebServer		Python webserver running on port 4001	
VI Editor	Data Cleanup		

Secondly, the selection of appropriate modeling techniques is critical to meet our research objectives.

1) Multiple regression modeling provided an elegant method of describing the relationships as explained in our problem definition.
2) Similarly, segmentation of teams based on the size and donation amount was also performed.

2) Results

2.1 - Team Donation Analysis - Analysis on the Total donations collected by the team.

The total donations consists of two components. 1) The donation and fees from the participant and 2) The additional support or contribution from non participant. We wanted to find out which variables have more influence on the team's fundraising contribution to National MS Society. We believed this would help us understand the potential opportunities of a corporate team. We considered the following two datasets with reference to corporate teams.

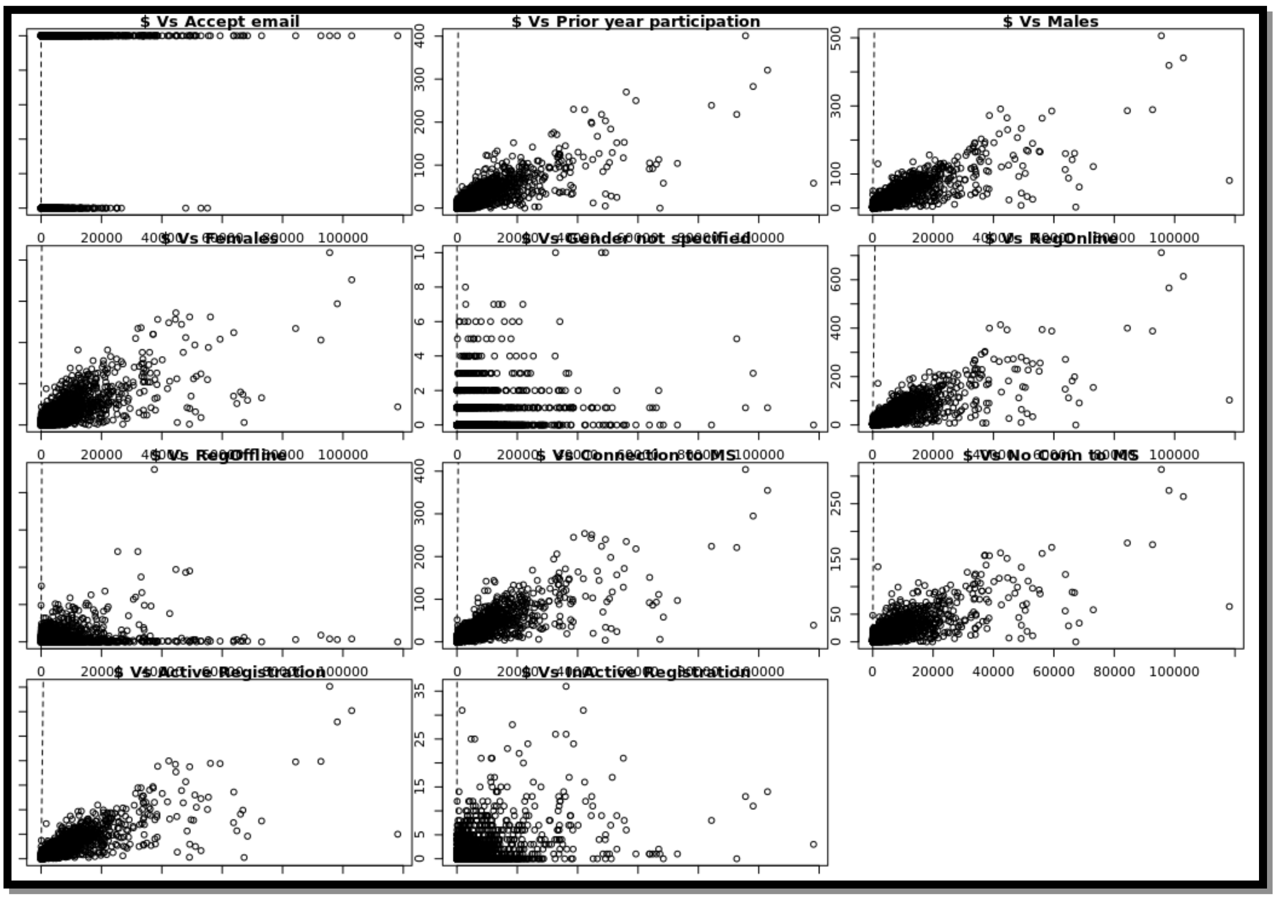
We analyzed the significance of the above multiple input variables (a to m) and their impact on output variables (n or o).

The updated dataset is available both github as well as AWS server at
ec2-52-15-142-133.us-east-2.compute.amazonaws.com:4001

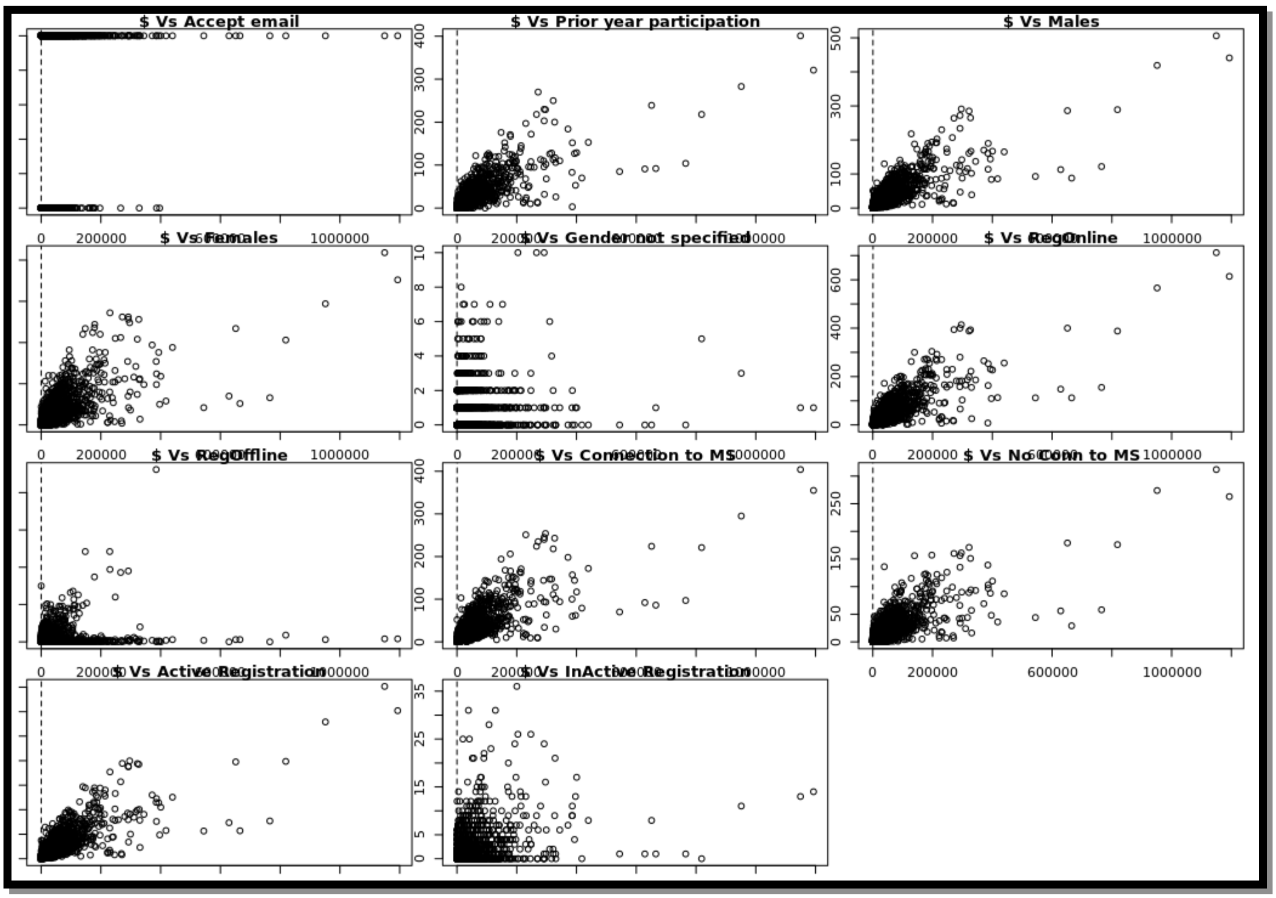
2.1.1 Correlation between Independent variable and output variable

This is done by doing a simple plot to explore the relationship between the input and output variables. Two sets of plots were created

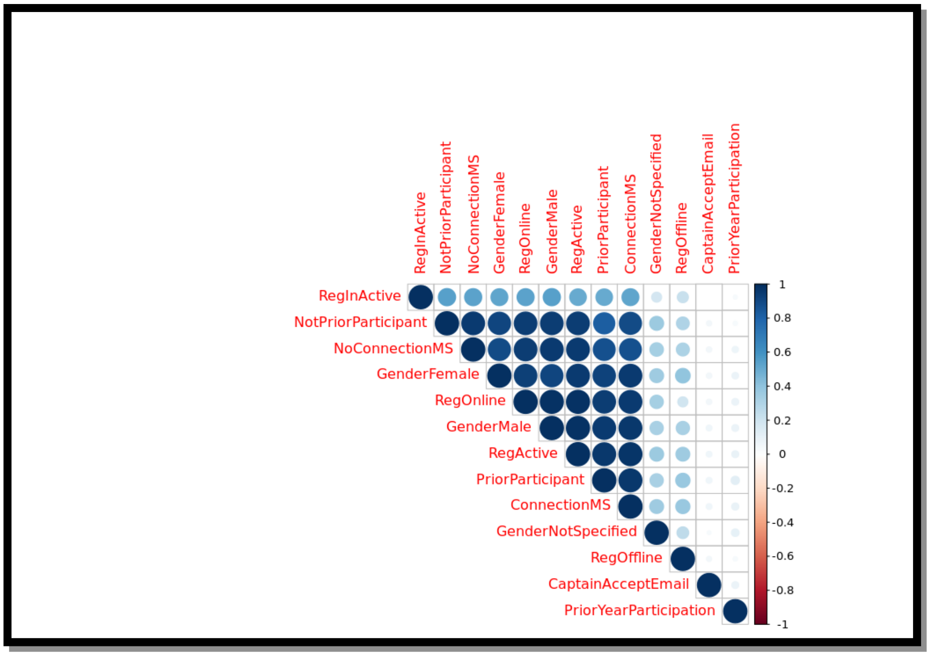
a) Total $ from participants as an output variable



b) Total $ from non-participants as an output variable



2.1.2 Correlation between the variables



Positive correlations are displayed in blue and **negative correlations** in red color. Color intensity and the size of the circle are proportional to the **correlation coefficients**. In the right side of the **correlogram**, the legend color shows the **correlation coefficients** and the corresponding colors.

2.1.3 Linear regression

2.1.3.1 Analysis for the total contribution from Participants.

```
fit_Participants <- lm(TotalFromParticipants ~ CaptainAcceptEmail+PriorYearParticipation +
GenderMale + GenderFemale + GenderNotSpecified + RegOnline + RegOffline +
PriorParticipant + NotPriorParticipant + ConnectionMS + NoConnectionMS +  RegActive +
RegInActive)
> summary(fit_Participants)

Call:
lm(formula = TotalFromParticipants ~ CaptainAcceptEmail + PriorYearParticipation +
    GenderMale + GenderFemale + GenderNotSpecified + RegOnline +
    RegOffline + PriorParticipant + NotPriorParticipant + ConnectionMS +
    NoConnectionMS + RegActive + RegInActive)

Residuals:
   Min      1Q Median      3Q     Max
```

```
-21908   -694    -53    345 100983

Coefficients: (4 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -518.508     80.878  -6.411 1.53e-10 ***
CaptainAcceptEmail     261.367     88.956   2.938 0.003312 **
PriorYearParticipation 207.558     86.905   2.388 0.016950 *
GenderMale              75.441     23.123   3.263 0.001109 **
GenderFemale          -104.553     24.339  -4.296 1.76e-05 ***
GenderNotSpecified    -120.280     66.929  -1.797 0.072354 .
RegOnline               24.107      6.760   3.566 0.000365 ***
RegOffline                  NA         NA      NA       NA
PriorParticipant        82.892      8.231  10.070  < 2e-16 ***
NotPriorParticipant         NA         NA      NA       NA
ConnectionMS            20.609     10.101   2.040 0.041347 *
NoConnectionMS              NA         NA      NA       NA
RegActive               53.122     22.549   2.356 0.018507 *
RegInActive                 NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3289 on 7532 degrees of freedom
  (119 observations deleted due to missingness)
```

Multiple R-squared: 0.7217,   Adjusted R-squared: 0.7214
F-statistic: 2171 on 9 and 7532 DF, p-value: < 2.2e-16

2.1.3.2 Analysis for the total contribution from Non Participants.

```
fit_NonParticipants<-lm(TotalNotFromParticipants ~ CaptainAcceptEmail+PriorYearParticipation + GenderMale + GenderFemale + GenderNotSpecified + RegOnline + RegOffline + PriorParticipant + NotPriorParticipant + ConnectionMS + NoConnectionMS +  RegActive + RegInActive)
> summary(fit_NonParticipants)

Call:
lm(formula = TotalNotFromParticipants ~ CaptainAcceptEmail + 
    PriorYearParticipation + GenderMale + GenderFemale + GenderNotSpecified + 
    RegOnline + RegOffline + PriorParticipant + NotPriorParticipant + 
    ConnectionMS + NoConnectionMS + RegActive + RegInActive)

Residuals:
    Min      1Q  Median      3Q     Max 
-188268   -4570     936    3813  550459 
```

```
Coefficients: (4 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            -5304.75    593.38  -8.940  < 2e-16 ***
CaptainAcceptEmail      1662.12    652.64   2.547   0.0109 *
PriorYearParticipation   414.05    637.60   0.649   0.5161
GenderMale               -18.02    169.65  -0.106   0.9154
GenderFemale           -2026.04    178.57 -11.346  < 2e-16 ***
GenderNotSpecified     -2995.48    491.04  -6.100 1.11e-09 ***
RegOnline                394.48     49.60   7.953 2.08e-15 ***
RegOffline                   NA        NA      NA       NA
PriorParticipant         721.83     60.39  11.953  < 2e-16 ***
NotPriorParticipant          NA        NA      NA       NA
ConnectionMS             -51.69     74.10  -0.697   0.4855
NoConnectionMS               NA        NA      NA       NA
RegActive               1013.32    165.44   6.125 9.52e-10 ***
RegInActive                  NA        NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24130 on 7532 degrees of freedom
  (119 observations deleted due to missingness)
```

Multiple R-squared: 0.7082, Adjusted R-squared: 0.7079
F-statistic: 2031 on 9 and 7532 DF, p-value: < 2.2e-16

Based on the analysis, the following are the models to predict the contribution from participant and non-participant.

Predicted contribution from participant	Capitan Accepts email	Participant participated in prior year	−49.583+75.441(GM)-104.553(GF)-120.280(GNS)+24.107(ROL)+82.892(PP)+20.609(CMS)+53.122(RIA)
		Participant did not participate in prior year	-257.141+75.441(GM)-104.553(GF)-120.280(GNS)+24.107(ROL)+82.892(PP)+20.609(CMS)+53.122(RIA)
	Caption does not accept email	Participant participated in prior year	-310.95+75.441(GM)-104.553(GF)-120.280(GNS)+24.107(ROL)+82.892(PP)+20.609(CMS)+53.122(RIA)
		Participant did not participate in prior year	-518.508+75.441(GM)-104.553(GF)-120.280(GNS)+24.107(ROL)+82.892(PP)+20.609(CMS)+53.122(RIA)
Predicted contribution from Non participant	Capitan Accepts email	−3642.63-2026.04(GF)-2995.48(GNS)+394.48(RO)+721.83(PP)+1013.32(RA)	
	Caption does not accept email	-5304.75-2026.04(GF)-2995.48(GNS)+394.48(RO)+721.83(PP)+1013.32(RA)	

2.2 - Team Donor Analysis - Analysis on the Total donations from Donor.

We have combined Team table and donor table to analyze Donations from donors based on other factors. The combined dataset is available both github as well as AWS server at http://ec2-18-188-80-129.us-east-2.compute.amazonaws.com:4001/Team-Donor-Participant-Merge.csv

2.2.1 Linear regression - Analysis for the total contribution.

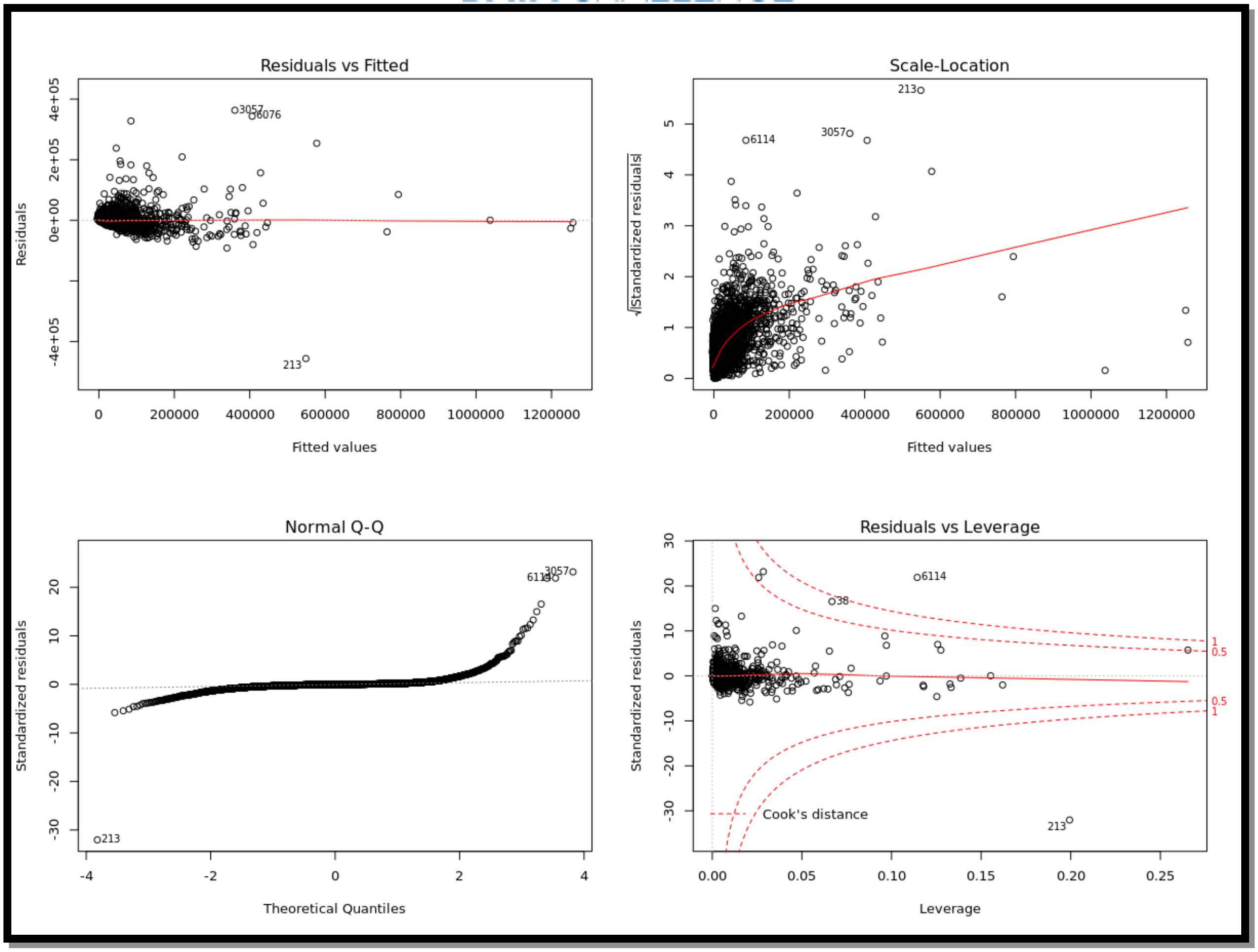
```
fit_donar<-
lm(GiftAmount~CaptainAcceptEmail+PriorYearParticipation+GenderMale+GenderFemale+GenderNotSpecified+RegOnline+RegOffline+PriorParticipant+NotPriorParticipant+ConnectionMS+NoConnectionMS+RegActive+RegInActive+DonorGenderMale+DonorGenderFemale+DonorGenderNotSpecified+RegisteredDonors+NonRegisteredDonors+DonorConnectionMS+DonorNoConnectionMS+DonorPriorParticipant+DonorNotPriorParticipant)
```

```
summary(fit_donar)

Call:
lm(formula = GiftAmount ~ CaptainAcceptEmail + PriorYearParticipation +
    GenderMale + GenderFemale + GenderNotSpecified + RegOnline +
    RegOffline + PriorParticipant + NotPriorParticipant + ConnectionMS +
    NoConnectionMS + RegActive + RegInActive + DonorGenderMale +
    DonorGenderFemale + DonorGenderNotSpecified + RegisteredDonors +
    NonRegisteredDonors + DonorConnectionMS + DonorNoConnectionMS +
    DonorPriorParticipant + DonorNotPriorParticipant)

Residuals:
    Min      1Q  Median      3Q     Max
-456112  -2808    -297    1419  363387

Coefficients: (6 not defined because of singularities)
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              -1291.68     396.14  -3.261  0.00112 **
CaptainAcceptEmailTRUE     979.64     430.65   2.275  0.02295 *
PriorYearParticipation    2299.71     424.48   5.418 6.22e-08 ***
GenderMale                 518.15     118.80   4.362 1.31e-05 ***
GenderFemale               895.50     125.69   7.124 1.14e-12 ***
GenderNotSpecified        -735.23     329.19  -2.233  0.02555 *
RegOnline                  -97.06      33.98  -2.857  0.00429 **
PriorParticipant          -926.08      63.62 -14.555  < 2e-16 ***
ConnectionMS               157.72      52.74   2.990  0.00280 **
RegActive                  -70.83     109.92  -0.644  0.51935
DonorGenderMale            252.18      36.06   6.994 2.91e-12 ***
DonorGenderFemale          -16.13      37.03  -0.435  0.66323
DonorGenderNotSpecified    146.39      36.52   4.008 6.18e-05 ***
RegisteredDonors          -268.22      24.29 -11.042  < 2e-16 ***
```

```
DonorConnectionMS          -12.42       4.12  -3.015  0.00258 **
DonorPriorParticipant       36.05      35.79   1.007  0.31386
DonorNotPriorParticipant   -80.13      36.46  -2.198  0.02801 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15910 on 7525 degrees of freedom
  (119 observations deleted due to missingness)
```

Multiple R-squared: 0.8985,   Adjusted R-squared: 0.8983
F-statistic: 4163 on 16 and 7525 DF, p-value: < 2.2e-16

```
coef(fit_donar)
             (Intercept)   CaptainAcceptEmailTRUE   PriorYearParticipation              GenderMale
             -1291.68410              979.63672             2299.70610               518.14603
            GenderFemale      GenderNotSpecified               RegOnline              RegOffline
               895.50230             -735.22620              -97.06298                      NA
        PriorParticipant     NotPriorParticipant            ConnectionMS          NoConnectionMS
              -926.08398                      NA             157.71676                      NA
               RegActive             RegInActive         DonorGenderMale       DonorGenderFemale
               -70.82647                      NA             252.17895               -16.12740
 DonorGenderNotSpecified        RegisteredDonors     NonRegisteredDonors
DonorConnectionMS
               146.38924             -268.22113                      NA               -12.41979
     DonorNoConnectionMS   DonorPriorParticipant DonorNotPriorParticipant
                      NA                36.05088              -80.12730
> layout(matrix(1:4,2,2))
> plot(fit_donar)
```



Based on the analysis, the following are the model to predict the contribution from participant and non participant.

If capitan accepted email:

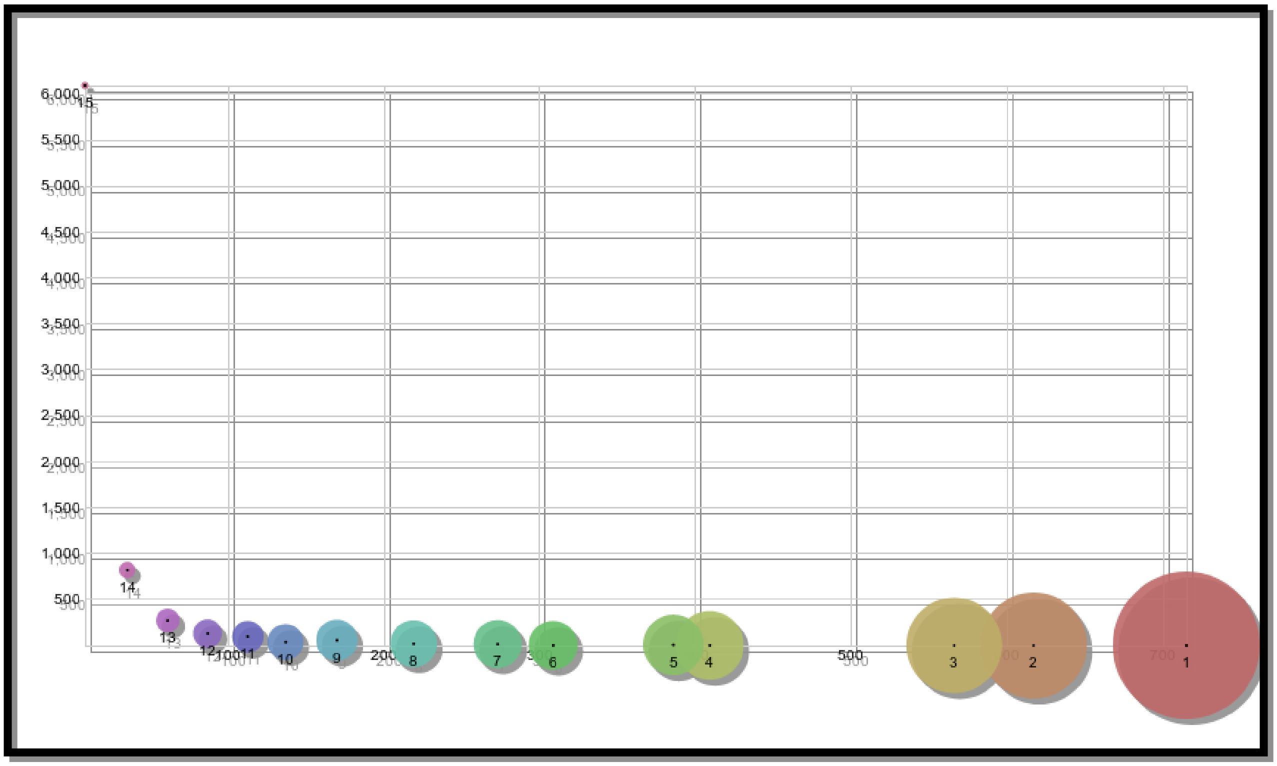
Predicated_donation=-312.04+2299.71(PriorYearParticipation)+518.15(GenderMale)
+895.50(GenderFemale)-735.23(GenderNotSpecified)-97.06(RegOnline)-926.08(PriorParticipant)
+157.72(ConnectionMS)-70.83(RegActive)+252.18(DonorGenderMale)-16.13(DonorGenderFemale)
+146.39(DonorGenderNotSpecified)-268.22(RegisteredDonors)-12.42(DonorConnectionMS)
+36.05(DonorPriorParticipant)-80.13(DonorNotPriorParticipant)

If not

Predicated_donation=-1291.68+2299.71(PriorYearParticipation)+518.15(GenderMale)
+895.50(GenderFemale)-735.23(GenderNotSpecified)-97.06(RegOnline)-926.08(PriorParticipant)
+157.72(ConnectionMS)-70.83(RegActive)+252.18(DonorGenderMale)-16.13(DonorGenderFemale)
+146.39(DonorGenderNotSpecified)-268.22(RegisteredDonors)-12.42(DonorConnectionMS)
+36.05(DonorPriorParticipant)-80.13(DonorNotPriorParticipant)

2.3 - Team size and Donation

We created the cluster based on the team size and number of teams. The bin size was 25. Here is the representation based on our analysis



X - Axis represents the number of participants
Y - Axis represents the total number of teams.
The circle size represents the amount collected under each bin.

The data set is available at
http://ubuntu@ec2-18-188-80-129.us-east-2.compute.amazonaws.com:4001/Bucket.csv

Observation: Based on the observation, few teams with large number of participants collect lot of money compared to ton of teams with small participants per team.

3) Presentation:

Please refer the Presentation.ppt enclosed. The same is also available at github as well as AWS server at http://ec2-18-188-80-129.us-east-2.compute.amazonaws.com:4001/Presentation.ppt

4) References

Git hub

https://github.com/sudhakaren/BikeMS

AWS server

http://ec2-18-188-80-129.us-east-2.compute.amazonaws.com