Predicting donors

Hemal Agarwal

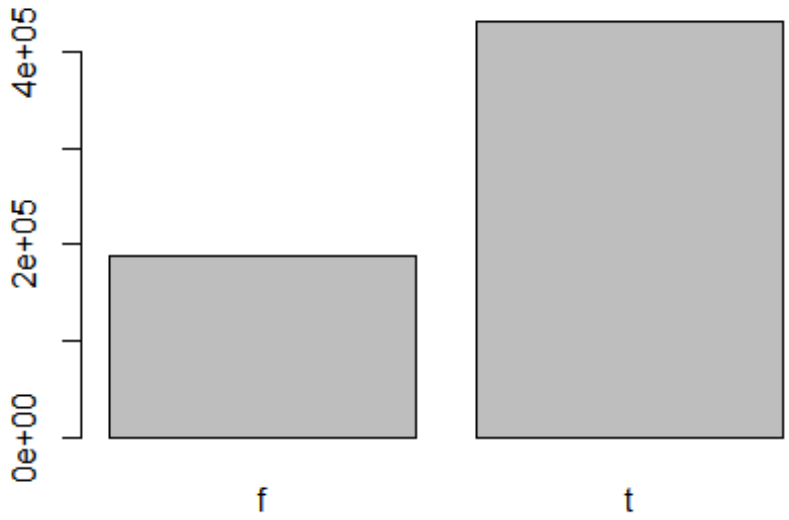
October 21, 2016

```
setwd("C:/Users/Adroit/Desktop/files_dssap")
data_outcomes <- read.csv("outcomes.csv")
data_projects<- read.csv("projects.csv")
#data_resources <- read.csv("resources.csv")
#data_essays <-read.csv("essays.csv")
data_donations <- read.csv("donations.csv")
#data_submissions <- read.csv("sampleSubmissions.csv")
```

Exploratory Analysis

1.Understanding the proportions of projects that have been fully funded

```
data_fullyFunded = data_outcomes$fully_funded
data_fullyFunded_table = (table(data_fullyFunded))
data_fullyFunded_freq <- as.data.frame((table(data_fullyFunded)))
#data_fullyFunded
#     f      t
#188643 430683
barplot(data_fullyFunded_table)
```

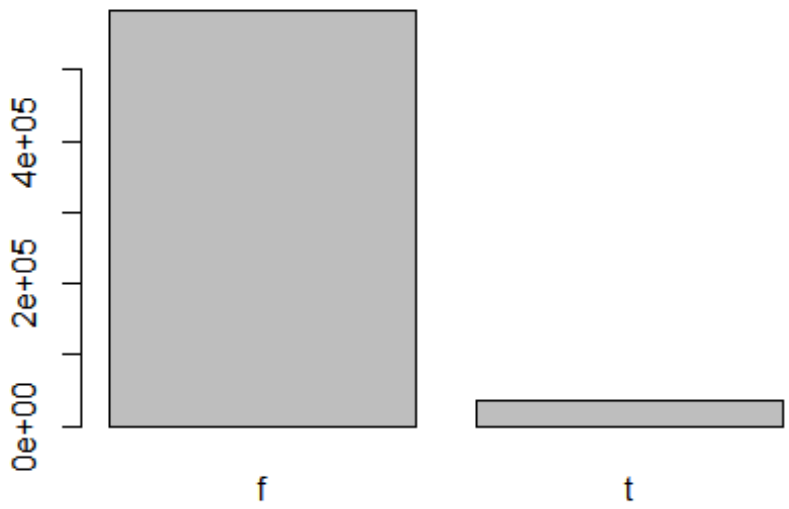



```
as.data.frame(table(data_fullyFunded))

##   data_fullyFunded   Freq
## 1                f 188643
## 2                t 430683

# x   freq
#1 f 188643
#2 t 430683

prop_fullyfunded =
(data_fullyFunded_freq$Freq[data_fullyFunded_freq$data_fullyFunded ==
"t"])/length(data_fullyFunded)
#0.695406
```

Analysis About 70% of the projects have been fully funded

2.Understanding the proportion of Exciting projets

```
data_excitingProjects = data_outcomes$is_exciting
data_excitingProjects.freq = as.data.frame(table(data_excitingProjects))
# data_excitingProjects   Freq
#1                     f 582616
#2                     t  36710
barplot(table(data_excitingProjects))
```




```
#proportion of exciting projects
prop_exciting<-
data_excitingProjects.freq$Freq[(data_excitingProjects.freq$data_excitingProj
ects=="t")]/ length(data_outcomes$is_exciting)
prop_exciting
```

```
## [1] 0.05927411
```

```
 #0.05927411
```

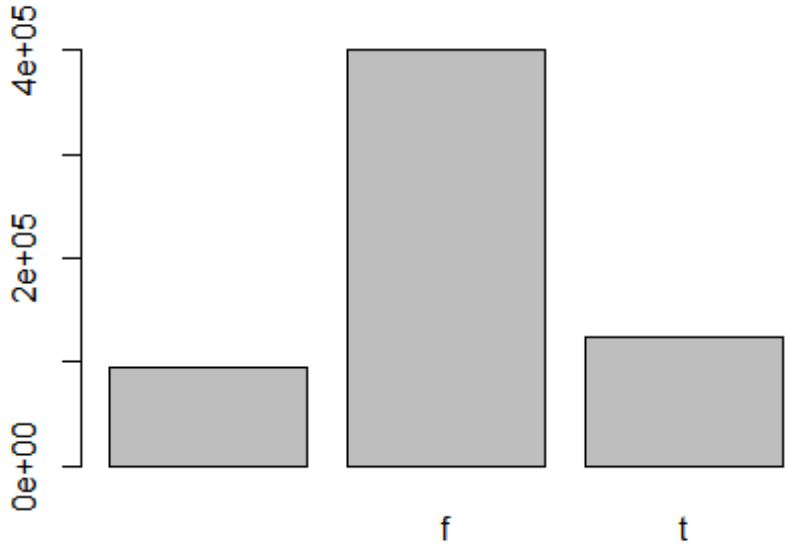
Analysis

Even though about 70% of the projects are fully funded, only 0.059% of the total projects can be deemed exciting from a business stand point. This is worrying.

3.Examinining the factors which form criteria for an exciting projects

```
# the number of projects that had atleast one teacher referred donor

data_atLeastOneTeacherDonor <-
data_outcomes$at_least_1_teacher_referred_donor
data_atLeastOneTeacherDonor.freq <-
as.data.frame(table(data_atLeastOneTeacherDonor))
# data_atLeastOneTeacherDonor   Freq
#1                          NA   94398
#2                           f 400268
#3                           t 124660
```

```
barplot(table(data_atLeastOneTeacherDonor))
```



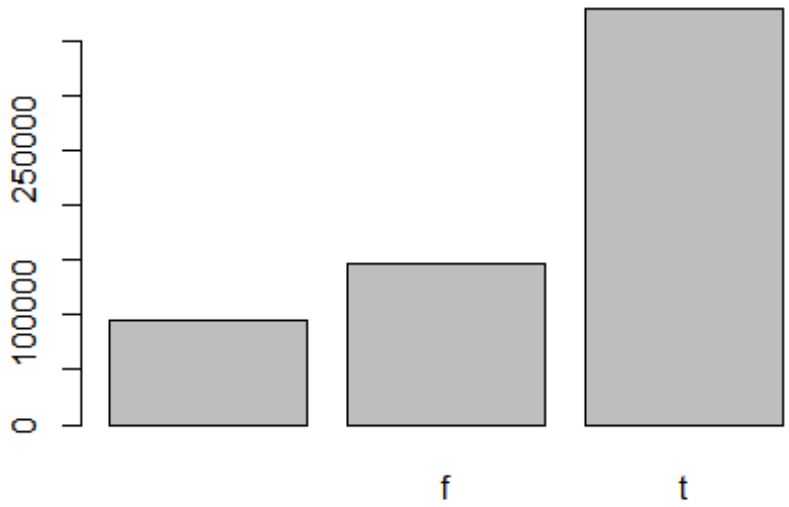

```
#proportion of projects that had atleast one teacher referred donor
a<-
data_atLeastOneTeacherDonor.freq$Freq[data_atLeastOneTeacherDonor.freq$data_a
tLeastOneTeacherDonor=="t"]
b<-
data_atLeastOneTeacherDonor.freq$Freq[data_atLeastOneTeacherDonor.freq$data_a
tLeastOneTeacherDonor=="f"]
prop<- a/(a+b)
prop

## [1] 0.2374802

# 0.2374802
```

Analysis:about 23% of the total projects(ignoring the projects for which the information is not available), are atleast one teacher referred donor

4.Projects with atleast one Green donation

```
data_atLeastOneGreenDonation <- data_outcomes$at_least_1_green_donation
data_atLeastOneGreenDonation.freq <-
as.data.frame(table(data_atLeastOneGreenDonation))
#  data_atLeastOneGreenDonation   Freq
#1                             94398
```

```
#2                              f 146235
#3                              t 378693

a<-
data_atLeastOneGreenDonation.freq$Freq[data_atLeastOneGreenDonation.freq$data
_atLeastOneGreenDonation== "t"]
b<-
data_atLeastOneGreenDonation.freq$Freq[data_atLeastOneGreenDonation.freq$data
_atLeastOneGreenDonation== "f"]
prop<-a/(a+b)
prop

## [1] 0.7214189

#[1] 0.7214189
barplot(table(data_atLeastOneGreenDonation))
```



About 72% of the projects have atleast one donor who pays via credit card/giftcard

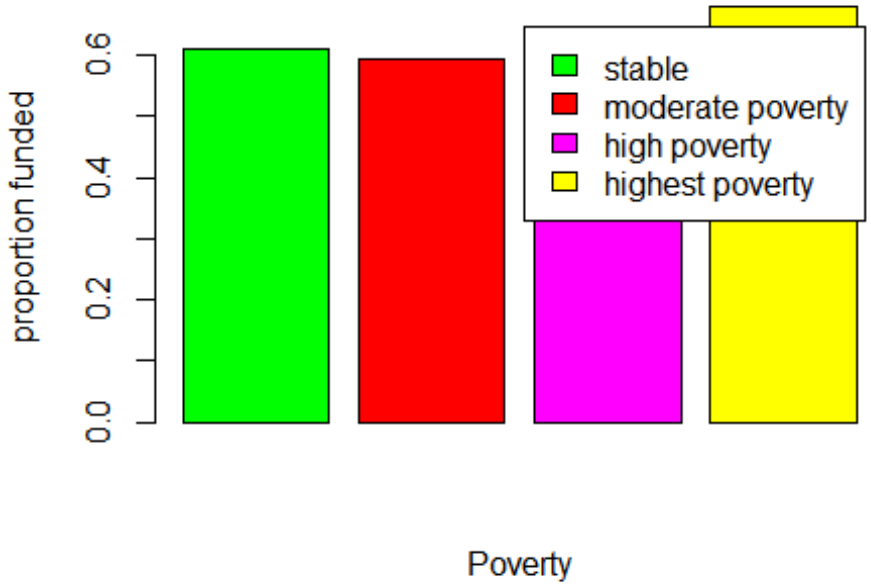
5Are schools with higher poverty levels likely to get funded any more than schools with lower poverty levels.

```
#analysing by poverty level
Df_outcomeProjects <- merge(data_projects, data_outcomes, by.x="projectid",
by.y="projectid", all.x=T)

Df_outcomeProjects$poverty<- 0
```

```r
Df_outcomeProjects$poverty[Df_outcomeProjects$poverty_level=="moderate poverty"] <- 1
Df_outcomeProjects$poverty[Df_outcomeProjects$poverty_level=="high poverty"] <- 2
Df_outcomeProjects$poverty[Df_outcomeProjects$poverty_level=="highest poverty"] <- 3

poverty_stats<- as.data.frame(table(Df_outcomeProjects$poverty))
# Var1   Freq
#1    0  16711
#2    1  90337
#3    2 173561
#4    3 383489
poverty_stats_funded<-
as.data.frame(table((Df_outcomeProjects$poverty[Df_outcomeProjects$fully_funded == "t"])))
poverty_stats <- merge(poverty_stats,poverty_stats_funded,by.x = "Var1",by.y = "Var1",all.x = TRUE)


poverty_stats["proportion"] <- (poverty_stats$Freq.y)/(poverty_stats$Freq.x)
poverty_stats
```

```
##   Var1 Freq.x Freq.y proportion
## 1    0  16711  10219  0.6115134
## 2    1  90337  53613  0.5934778
## 3    2 173561 106581  0.6140838
## 4    3 383489 260270  0.6786896
```

```r
barplot(poverty_stats$proportion, legend.text = c("stable","moderate poverty","high poverty","highest poverty"),col = c("green","red","magenta","yellow"),xlab = "Poverty",ylab = "proportion funded")
```



Analysis: Maximum number of projects are posted by schools with highest poverty levels. While greater proportion of projects from schools with highest poverty level are funded, however, it difference is not too significant. The povertly level of the school does not seem to matter all the much on the funding aspect of the project.This is a matter of concern. The next question to be asked is if donors look at the schools poverty while making a donation?

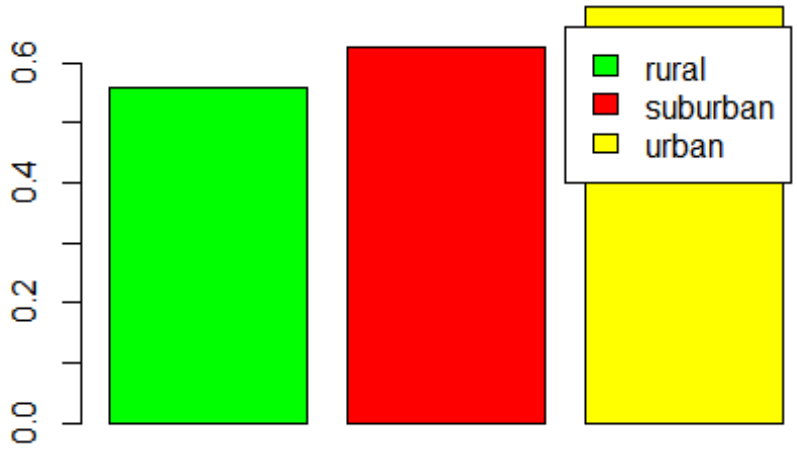
SECTION2 : DATA STORY

Are projects likely to get fully funded based on the type of location ?

Analysing by the type of location

```
levels(Df_outcomeProjects$school_metro)

## [1] ""         "rural"    "suburban" "urban"

Df_outcomeProjects$Metro_type<- 0
Df_outcomeProjects$Metro_type[Df_outcomeProjects$school_metro== "rural"] <- 1
Df_outcomeProjects$Grade[Df_outcomeProjects$school_metro== "suburban"] <- 2
Df_outcomeProjects$Grade[Df_outcomeProjects$school_metro== "urban"] <- 3

Metro_stat <- as.data.frame(table(Df_outcomeProjects$school_metro))
Metro_stats_funded <-
as.data.frame(table((Df_outcomeProjects$school_metro[Df_outcomeProjects$fully
_funded == "t"])))
```

```
Metro_stats <- merge(Metro_stat, Metro_stats_funded ,by.x = "Var1",by.y =
"Var1",all.x = TRUE)
Metro_stats["prop"] = Metro_stats$Freq.y/Metro_stats$Freq.x
barplot((Metro_stats$prop[2:4]),legend.text=
c("rural","suburban","urban"),col = c("green","red","yellow"))
```



Greater proportion of projects in urban areas are fully funded as compared to rural areas.

Does the location of the school really matter to the donors?

```
levels(data_projects$school_state)
```

```
##  [1] "AK" "AL" "AR" "AZ" "CA" "CO" "CT" "DC" "DE" "FL" "GA" "HI" "IA" "ID"
## [15] "IL" "IN" "KS" "KY" "La" "LA" "MA" "MD" "ME" "MI" "MN" "MO" "MS" "MT"
## [29] "NC" "ND" "NE" "NH" "NJ" "NM" "NV" "NY" "OH" "OK" "OR" "PA" "RI" "SC"
## [43] "SD" "TN" "TX" "UT" "VA" "VT" "WA" "WI" "WV" "WY"
```

```
states<- c(levels(data_projects$school_state))
states
```

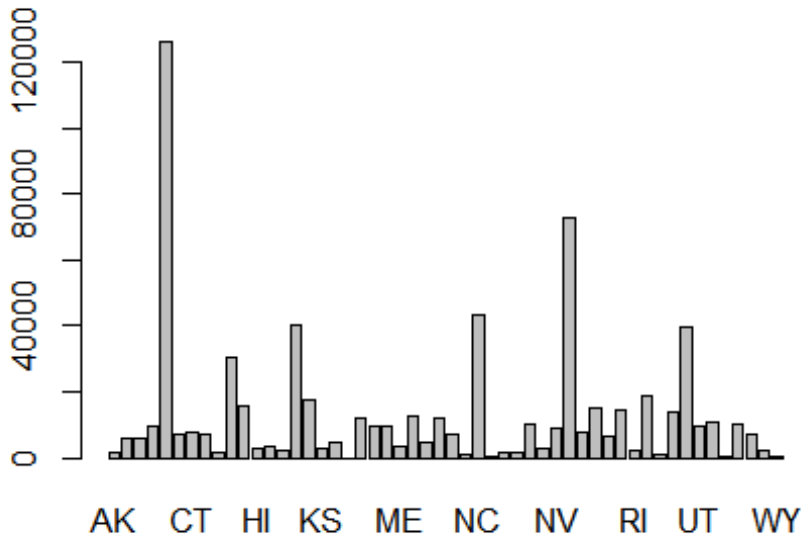
```
##  [1] "AK" "AL" "AR" "AZ" "CA" "CO" "CT" "DC" "DE" "FL" "GA" "HI" "IA" "ID"
## [15] "IL" "IN" "KS" "KY" "La" "LA" "MA" "MD" "ME" "MI" "MN" "MO" "MS" "MT"
## [29] "NC" "ND" "NE" "NH" "NJ" "NM" "NV" "NY" "OH" "OK" "OR" "PA" "RI" "SC"
## [43] "SD" "TN" "TX" "UT" "VA" "VT" "WA" "WI" "WV" "WY"
```

```
data_states_totalprojects = data_projects$school_sta
data_states_totalprojects_freq = table(data_states_totalprojects)
```

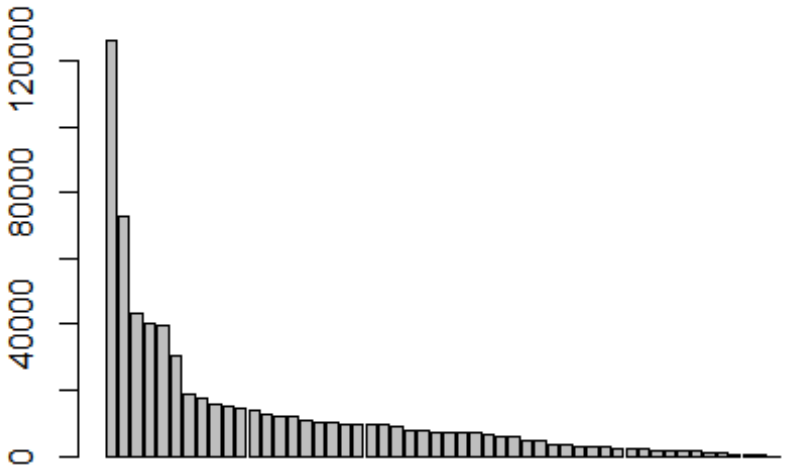
```
data_states_totalprojects_freq =
as.data.frame(table(data_states_totalprojects))
data_states_totalprojects_freq
```

```
##    data_states_totalprojects   Freq
## 1                         AK   1383
## 2                         AL   5650
## 3                         AR   5770
## 4                         AZ   9837
## 5                         CA 126242
## 6                         CO   7021
## 7                         CT   7728
## 8                         DC   6918
## 9                         DE   1605
## 10                        FL  30605
## 11                        GA  15403
## 12                        HI   2586
## 13                        IA   3186
## 14                        ID   2030
## 15                        IL  40167
## 16                        IN  17299
## 17                        KS   2829
## 18                        KY   4541
## 19                        La      3
## 20                        LA  12180
## 21                        MA   9403
## 22                        MD   9555
## 23                        ME   3413
## 24                        MI  12330
## 25                        MN   4519
## 26                        MO  12097
## 27                        MS   6930
## 28                        MT    819
## 29                        NC  43478
## 30                        ND    483
## 31                        NE   1542
## 32                        NH   1491
## 33                        NJ  10411
## 34                        NM   2649
## 35                        NV   8844
## 36                        NY  73182
## 37                        OH   7813
## 38                        OK  14853
## 39                        OR   6610
## 40                        PA  14379
## 41                        RI   2127
## 42                        SC  18615
## 43                        SD    990
## 44                        TN  14079
## 45                        TX  39661
```

```
## 46                          UT   9304
## 47                          VA  10716
## 48                          VT    555
## 49                          WA  10469
## 50                          WI   7027
## 51                          WV   2334
## 52                          WY    437

barplot(table(data_states_totalprojects))
```




```
project.state.orders<-
data_states_totalprojects_freq[rev(order(data_states_totalprojects_freq$Freq)
),]

barplot(project.state.orders$Freq)
```




```
sum(as.numeric(project.state.orders$Freq[1:10]))/sum(as.numeric(project.state
.orders$Freq))
```

```
## [1] 0.6316914
```

```
#0.6316914
sum(as.numeric(project.state.orders$Freq[1:5]))/sum(as.numeric(project.state.
orders$Freq))
```

```
## [1] 0.4859674
```

```
#0.4859674
sum(as.numeric(project.state.orders$Freq[1]))/sum(as.numeric(project.state.or
ders$Freq))
```

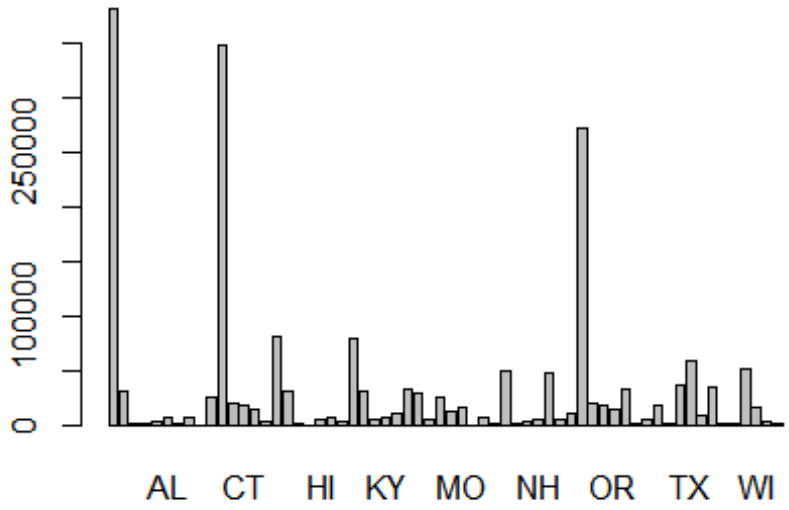
```
## [1] 0.1900954
```

```
#0.1900954
```

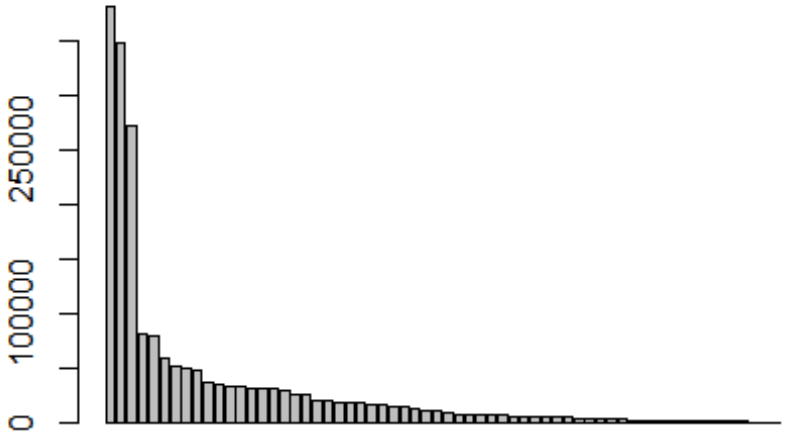
Analysis: Most of the projects come from specific states. Infact 48% of the total projects are posted from 5 states of the 52 :CA, NY, NC,IL, TX, and 19% come from CA alone.

```
#Merging dataframes
Df_outcomeProjects <- merge(data_projects, data_outcomes, by.x="projectid",
by.y="projectid", all.x=T)
```

```r
#analysing which states make maximi=um donation transactions
data_state_donations <- data_donations$donor_state
data_state_donations.freq <- as.data.frame(table(data_state_donations))
barplot(table(data_state_donations))
```




```r
data_state_donations.freq_order<-
data_state_donations.freq[rev(order(data_state_donations.freq$Freq)),]
barplot(data_state_donations.freq_order$Freq)
```




```
sum(as.numeric(data_state_donations.freq_order$Freq[2:11]))/sum(as.numeric(da
ta_state_donations.freq_order$Freq[2:61]))
```

```
## [1] 0.6732629
```

```
#0.6732629
sum(as.numeric(data_state_donations.freq_order$Freq[2:6]))/sum(as.numeric(dat
a_state_donations.freq_order$Freq[2:61]))
```

```
## [1] 0.5334914
```

```
#0.5334914
sum(as.numeric(data_state_donations.freq_order$Freq[2]))/sum(as.numeric(data_
state_donations.freq_order$Freq[2:61]))
```

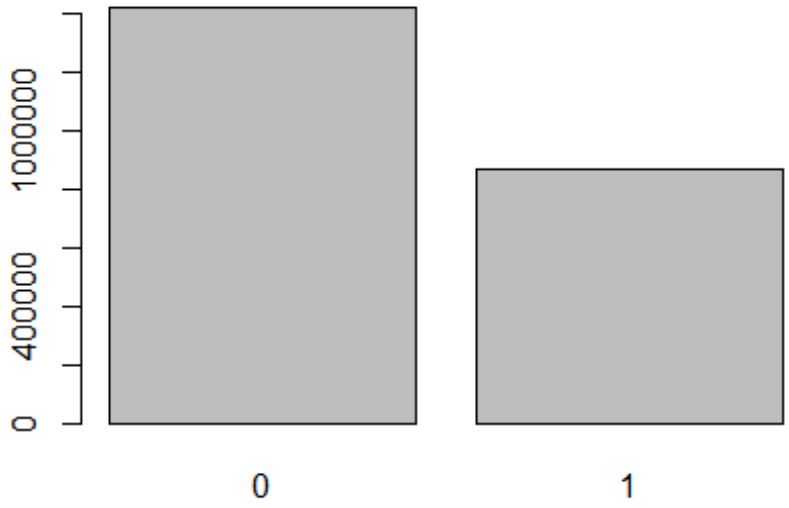
```
## [1] 0.2214089
```

```
#0.2214089
```

Analysis: About 53% of the donations (in terms of transcactions) come from 5 states. we also observe that ecept for FL, the states that make makinum number of donations are also the ones that post maximum number of projects. Question to answer: does it so happen that donors donate to schools that are located within their state ?

```
Df_outcomes.projects.donations <-
merge(Df_outcomeProjects,data_donations,by.x="projectid", by.y="projectid",
all.x=T)
```

```r
Df_outcomes.projects.donations$same.state <- 0
Df_outcomes.projects.donations$same.state[as.character(Df_outcomes.projects.d
onations$school_state)==
as.character(Df_outcomes.projects.donations$donor_state)]<-1
data_matchstates<-
as.data.frame(table(Df_outcomes.projects.donations$same.state))
# Var1    Freq
#1    0 1416875
#2    1  869490
donor_states_NA <- sum(data_state_donations.freq_order$Freq[1])

barplot(table(Df_outcomes.projects.donations$same.state))
```




```r
# examiniming the proportion of people who donated for projects of the
schools from the same state(Ignoring the NA values )
as.numeric(data_matchstates$Freq[data_matchstates$Var1==1])/(as.numeric(data_
matchstates$Freq[data_matchstates$Var1==0])+as.numeric(data_matchstates$Freq[
data_matchstates$Var1==1])-donor_states_NA)

## [1] 0.4564197

#0.4564197
```

KeyInsight

46% of the donation transactions were made for projects from the same state as donors state. It seems like donors like to make donations to schools from their own state.

Are some neighborhoods likely to make more donations than the other neighborhoods?

```
zip_codes<- as.numeric(data_donations$donor_zip)
zip_codes.freq = as.data.frame(table(zip_codes) )
zp.orders<- zip_codes.freq[rev(order(zip_codes.freq$Freq)),]
head(zp.orders,100)
```

```
##       zip_codes  Freq
## 19250     94102 23523
## 20375     98102  7545
## 2246      10018  7454
## 2229      10001  7402
## 13751     62715  6809
## 1934       8091  6271
## 10294     46077  4897
## 15280     73104  4682
## 2274      10065  4569
## 1             0  3769
## 2251      10023  3647
## 13277     60631  3195
## 2245      10017  2919
## 2250      10022  2911
## 2247      10019  2797
## 13301     60657  2703
## 1753       7661  2613
## 18844     92563  2497
## 2554      11215  2464
## 2296      10128  2458
## 2540      11201  2433
## 2244      10016  2365
## 18215     90048  2296
## 9000      38117  2285
## 2231      10003  2117
## 2252      10024  2011
## 4844      21214  1972
## 2241      10013  1960
## 2239      10011  1859
## 13261     60614  1778
## 2253      10025  1776
## 18627     92027  1755
## 2242      10014  1751
## 19371     94583  1624
## 18534     91711  1580
```

```
## 19258     94110  1558
## 13016     60069  1543
## 2576      11238  1522
## 10348     46220  1496
## 19406     94704  1452
## 2572      11234  1427
## 2238      10010  1409
## 6625      29536  1405
## 10339     46204  1390
## 5873      27516  1311
## 13272     60625  1310
## 1870       8003  1269
## 6391      28729  1246
## 19262     94114  1204
## 4570      20151  1204
## 18376     91007  1190
## 13292     60647  1190
## 10364     46240  1187
## 18192     90024  1177
## 19630     95383  1158
## 18172     90004  1146
## 2256      10028  1139
## 6829      30062  1124
## 19265     94117  1116
## 13260     60613  1076
## 13285     60640  1066
## 18827     92530  1063
## 6180      28306  1059
## 2249      10021  1052
## 2556      11217  1042
## 4525      20009  1041
## 16518     78245  1029
## 2264      10036  1029
## 20500     98370  1004
## 10371     46260   996
## 19255     94107   992
## 13265     60618   992
## 2331      10285   977
## 5870      27513   964
## 13269     60622   947
## 19253     94105   944
## 18843     92562   907
## 4517      20001   896
## 13107     60201   895
## 20277     97707   894
## 14037     64155   894
## 1711       7506   887
## 5963      27713   882
## 13257     60610   861
## 4524      20008   842
```

```
## 10813     48126   833
## 2237      10009   830
## 18216     90049   825
## 18427     91214   821
## 11454     50023   814
## 6966      30317   813
## 2569      11231   805
## 15083     72201   797
## 18184     90016   795
## 13258     60611   795
## 6935      30269   789
## 19270     94122   779
## 4516      20000   767
## 5926      27601   764
## 5017      22201   762
```

```
length(zp.orders$zip_codes)
```

```
## [1] 20822
```

```
sum(as.numeric(zp.orders$Freq[1:100]))/sum(as.numeric(zp.orders$Freq[1:length
(zp.orders$zip_codes)]))
```

```
## [1] 0.2326013
```

```
# 0.2326013
100/length(zp.orders$zip_codes)
```

```
## [1] 0.004802613
```

```
# 0.004802613
sum(as.numeric(zp.orders$Freq[1:50]))/sum(as.numeric(zp.orders$Freq[1:length(
zp.orders$zip_codes)]))
```

```
## [1] 0.178028
```

```
# 0.178028
```

Key Insight

Some neighborhoods make a lot more donation transactions than the other neighborhoods.

Infact 0.05% of the neighborhoods make 23% of donation transactions.

0.02% of the neighborhoods make 17.8% of total donation transactions.

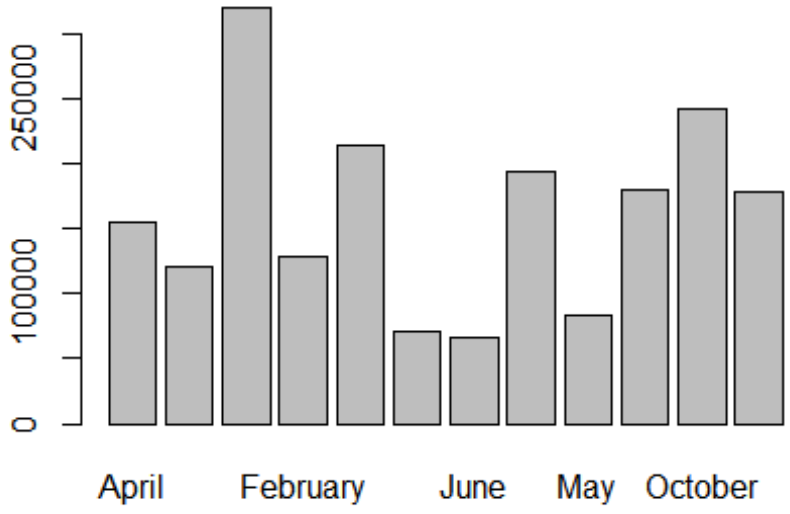
The next question to be asked is if the total donation made by these zipcodes is a lot more than the other zipcodes.

DATA STORY

2.Are peoply like to make more donations in certain months??

```
mon.donation<-c()
mon.donation<-months(
as.POSIXct(data_donations$donation_timestamp[1:length(data_donations$donation
_timestamp)]))
mon.donation.freq <- as.data.frame(table(mon.donation))
order

## function (..., na.last = TRUE, decreasing = FALSE, method = c("shell",
##     "radix"))
## {
##     z <- list(...)
##     if (missing(method)) {
##         ints <- all(vapply(z, function(x) is.integer(x) || is.factor(x),
##             logical(1L)))
##         method <- if (ints)
##             "radix"
##         else "shell"
##     }
##     else {
##         method <- match.arg(method)
##     }
##     if (any(unlist(lapply(z, is.object)))) {
##         z <- lapply(z, function(x) if (is.object(x))
##             as.vector(xtfrm(x))
##         else x)
##         if (method == "radix" || !is.na(na.last))
##             return(do.call("order", c(z, na.last = na.last, decreasing =
decreasing,
##                 method = method)))
##     }
##     else if (method != "radix" && !is.na(na.last)) {
##         return(.Internal(order(na.last, decreasing, ...)))
##     }
##     if (method == "radix") {
##         decreasing <- rep_len(as.logical(decreasing), length(z))
##         return(.Internal(radixsort(na.last, decreasing, FALSE,
##             TRUE, ...)))
##     }
##     if (any(diff((l.z <- lengths(z)) != 0L)))
##         stop("argument lengths differ")
##     na <- vapply(z, is.na, rep.int(NA, l.z[1L]))
##     ok <- if (is.matrix(na))
##         rowSums(na) == 0L
##     else !any(na)
##     if (all(!ok))
```

```
##         return(integer())
##     z[[1L]][!ok] <- NA
##     ans <- do.call("order", c(z, decreasing = decreasing))
##     ans[ok[ans]]
## }
## <bytecode: 0x00000000056bb4b8>
## <environment: namespace:base>

mon.order <- mon.donation.freq[rev(order(mon.donation.freq$Freq)),]
mon.order

##    mon.donation   Freq
## 3      December 319931
## 11      October 242150
## 5       January 214295
## 8         March 193442
## 10     November 179655
## 12    September 178941
## 1         April 154668
## 4      February 128979
## 2        August 120740
## 9           May  83838
## 6          July  71008
## 7          June  65697

barplot(table(mon.donation))
```



```r
sum(mon.order$Freq[1:3])/sum(mon.order$Freq[1:12])
```

```
## [1] 0.3974599
```

```r
#0.3974599
sum(mon.order$Freq[1])/sum(mon.order$Freq[1:12])
```

```
## [1] 0.1637863
```

```r
sum(mon.order$Freq[10:12])/sum(mon.order$Freq[1:12])
```

```
## [1] 0.1129054
```

KeyInsight

Highest number of donation is made in December, which accounts to almost 17% of the total donation.

Maximum number of donations are made in 3 months of December,October and january which account to 40% of the total donation transactions.

These are also the festive months. while the least donations are transacted in the summer months of May, June and July accounting to only 11% of the toal transactions.

PART3 : Some Questions to the partners

Q.What are the income levels of various neighbourhoods that form the donor audience for the site?

It may be interesting to understand how income levels effect the donations made to school projects.

Q. How is the audience made aware of this site? What mediums have been used by this website used for its own promotion?

Does the medium of promotion make an impact on the amount of donation for different projects?

Q.How does the company decide the order in which the projects are listed on the site?

Is it possible that projects that are published on the top of page get more funding than the rest?